TUM Data Innovation Lab

Munich Data Science Institute (MDSI)

Technical University of Munich

&

TUM Chair of Data Science in Earth Observation

Final report of project:

Integrating Microwave Remote Sensing Data for Enhanced Monitoring using Deep Learning

Authors	Aziz Banna, Yahya Hefnawi, Linhan Li, Julian Schmitt
Mentor(s)	M.Sc. Yueli Chen and Prof. Xiaoxiang Zhu
TUM Mentor	M.Sc. Yueli Chen
Project lead	Dr. Ricardo Acevedo Cabra (MDSI)
Supervisor	Prof. Dr. Massimo Fornasier (MDSI)

Jul 2025

Acknowledgements

We extend our sincere gratitude to M.Sc. Yueli Chen for her exceptional mentorship and tireless dedication, without which our work would not have been possible. We also thank Prof. Xiaoxiang Zhu for supporting us, providing the necessary resources, and making this project possible as head of the TUM Chair of Data Science in Earth Observation. Additionally, we are grateful to Dr. Ricardo Acevedo Cabra and Prof. Dr. Massimo Fornasier for organizing the TUM Data Innovation Lab.

Abstract

Microwave Earth observation satellites face a trade-off between temporal resolution and spatial resolution. This contrast is evident in the capabilities of *SMAP*, which provides global coverage every 2–3 days at a coarse resolution of approximately 9 km, and *Sentinel-1*, which offers much finer detail at around 10 m but revisits the same location only every 6–12 days. We bridge this gap with a cross-sensor super-resolution scheme that turns SMAP's frequent L-band low-resolution images into images that resemble Sentinel-1's C-band high-resolution products. Two models are explored: a denoising diffusion model for high-fidelity texture and a much faster Vision Transformer (ViT) model for lightweight inference.

Tests on six regions in Nunavik, Canada show that the denoising diffusion model produces highly detailed and sharply defined images, particularly along coastlines and lakes. The ViT model yields slightly lower reconstruction quality but offers an approximately 100 times faster runtime. Both models cope when asked to predict an unseen time period, yet quality drops, highlighting the need for longer training records. Transferring to a completely new region is harder; quick ViT fine-tuning recovers most of the loss, whereas denoising diffusion still struggles. Finally, extending the denoising diffusion model to jointly predict co-polarized (VV) and cross-polarized (VH) SAR backscatter maintains high reconstruction quality while reducing runtime by around one half compared to using two separate networks.

Overall, the denoising diffusion model yields the best images and handles dual polarization naturally, but its heavy compute cost and limited portability point to future work on lighter diffusion variants and better transfer-learning strategies.

Contents

1 Introduction

1.1 Motivation

Earth observation satellites play a crucial role in monitoring environmental change on a global scale. They support a wide range of applications, including tracking land cover transformations, vegetation dynamics, soil moisture, and hydrological events such as floods and snowmelt [1]. A fundamental limitation of satellite imagery, however, is the inherent trade-off between spatial and temporal resolution: images with fine spatial detail are typically acquired infrequently, whereas observations with high revisit frequency often lack the necessary spatial detail. This trade-off poses a significant challenge for applications that require both frequent updates and high-resolution information. Super-resolution (SR) techniques [2] offer a promising solution by reconstructing high-resolution (HR) images from their low-resolution (LR) counterparts. As illustrated in Fig. 1, SR can conceptually enable the generation of satellite imagery that combines fine spatial detail with high temporal frequency. Recent advances in deep learning have made it possible to learn complex mappings from LR to HR, leveraging spatial patterns, multi-scale textures, and domain-specific cues that traditional interpolation methods cannot capture. This promise of SR is especially attractive for Earth observation, where improving the spatiotemporal resolution of data can significantly enhance environmental monitoring capabilities.

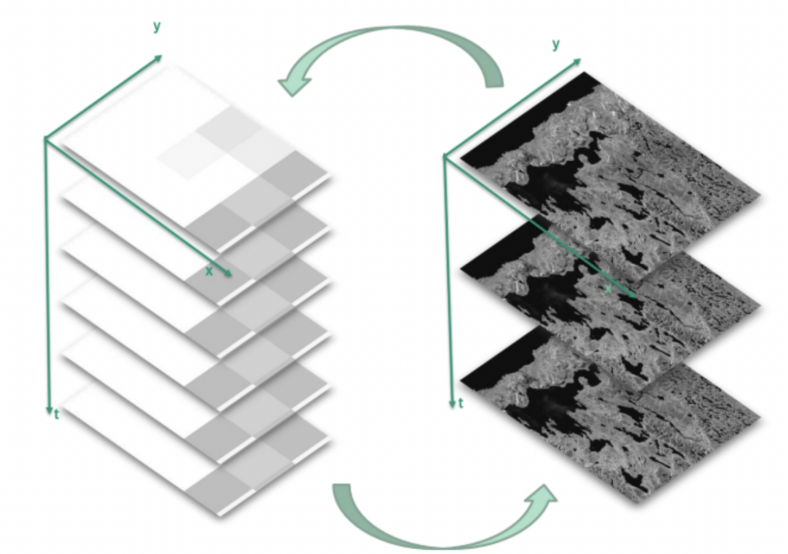

Figure 1: Conceptual illustration of super-resolution in the spatiotemporal domain [3]. The left side represents a sequence of low spatial resolution images with high temporal density, shown in the spatial grid defined by axes x and y. The right side illustrates the corresponding high spatial resolution sequence with low temporal density, highlighting the resolution trade-off addressed by super-resolution techniques.

1.2 Problem Statement

In this work, we focus on applying super-resolution to microwave remote sensing data, a domain that exemplifies the spatial-temporal resolution trade-off. Different satellite sensors operating in the microwave spectrum offer complementary capabilities. Notably, microwave sensors (such as radar and radiometers) can penetrate clouds and operate day-and-night, making them invaluable for continuous monitoring in regions with frequent cloud cover or polar night conditions [4]. Despite these advantages, microwave missions face the same fundamental trade-off: a sensor like a radiometer provides frequent observations but at coarse spatial resolution, while a Synthetic Aperture Radar (SAR) yields fine spatial details but revisits any given location only after several days. This imbalance limits our ability to capture rapid surface dynamics at high detail.
Cross-sensor super-resolution offers a way to mitigate this limitation by fusing data from multiple microwave instruments. In particular, the high revisit rate of passive radiometers can be combined with the rich spatial detail of active SAR imagery to produce reconstructed datasets with both improved temporal coverage and fine spatial resolution. We implement this fusion using state-of-the-art deep learning frameworks: specifically, we explore Denoising Diffusion Models [5, 6] for their ability to generate high-fidelity textures, and Vision Transformers (ViT) [7] for their strength in modeling long-range dependencies and global context. Both types of models are trained to translate low-resolution microwave observations into super-resolved outputs that faithfully recover dynamic surface features.

1.3 Related Work

Super-Resolution in Computer Vision. Image super-resolution has been an active research area for decades, evolving from traditional interpolation and filtering approaches to modern deep learning-based methods [2]. Early CNN models (e.g., SRCNN and its successors) demonstrated substantial improvements in upscaling quality by learning mappings from LR to HR images. Subsequent breakthroughs introduced generative models to further improve visual fidelity. For example, Ledig et al. [8] proposed SRGAN, which was among the first to use a Generative Adversarial Network for single-image SR, producing more photo-realistic details than previous methods optimized purely for pixel-wise accuracy. In recent years, denoising diffusion models have emerged as a powerful alternative to GANs for image synthesis. Saharia et al. [9] presented SR3, a diffusion-based super-resolution approach that achieves state-of-the-art perceptual quality on natural images by iteratively refining random noise into a high-resolution output. Alongside these advances, the ViT architecture [7] has opened new avenues for image restoration tasks. By leveraging self-attention mechanisms to capture long-range dependencies, transformer-based models (e.g., SwinIR and other ViT variants) have shown competitive performance in SR and denoising, complementing and surpassing convolutional networks on certain benchmarks.
Super-Resolution in Remote Sensing. Driven by the successes in computer vision, super-resolution techniques have been increasingly applied to Earth observation data. Lanaras [10], for instance, demonstrated that deep learning models can super-resolve coarse Sentinel-2 multispectral imagery to a finer resolution, achieving a globally applicable per-

formance across diverse regions. A recent survey [11] provides a comprehensive overview of SR methods for Earth observation, covering both classic approaches and modern deep networks, and discussing applications to multispectral, hyperspectral, and radar imagery. These efforts illustrate the potential of SR beyond natural RGB images. However, many of the state-of-the-art SR models (including GAN- and diffusion-based methods) have so far been developed and evaluated primarily on natural image datasets, and their adaptation to non-RGB remote sensing domains remains an emerging area of research. Unique challenges in these domains - for example, the presence of speckle noise in SAR images or the different radiometric characteristics of satellite data - necessitate careful tuning of models originally designed for photographs. To date, only limited work has explored advanced generative SR models for microwave remote sensing data. Our work addresses this gap by integrating diffusion models and transformer architectures into a cross-sensor SR pipeline for radar and radiometer data, pushing the frontier of super-resolution in the geospatial context.

1.4 Contribution

The main contributions of this work are summarized as follows:

- Novel SR Framework: We develop a cross-sensor super-resolution approach leveraging two advanced deep learning paradigms. In our framework, a DDPM-based model [5] is used to produce highly detailed and realistic HR images, while a ViT-based model [7] provides a faster alternative by capturing global context for efficient SR inference. To our knowledge, this is among the first applications of diffusion models and transformers for SR in the microwave remote sensing domain.

- Spatial and Temporal Transferability: We rigorously evaluate the generalization capabilities of the proposed models across different times and regions. The models are trained on certain time periods and geographic areas and then tested on unseen timestamps and locations to assess how well the learned super-resolution mapping transfers beyond the training conditions. This addresses the crucial question of temporal and spatial transferability for real-world deployment.

- Per-Region Performance Analysis: We conduct a detailed region-wise analysis to identify where each model performs best and where it struggles. By comparing super-resolution results across diverse geographic and environmental contexts, we gain insights into the conditions under which the diffusion and ViT models excel or face challenges. This analysis helps reveal the robustness of each approach and guides future improvements.

- Dual-Channel Reconstruction: We extend our super-resolution approach to jointly predict both co-polarized (VV) and cross-polarized (VH) SAR backscatters in a single model forward pass. This dual-channel SR yields consistent reconstruction of the two polarization bands while roughly halving the inference time compared to running two separate models for VV and VH. The unified prediction ensures that the model leverages cross-polarization information, resulting in more efficient and coherent recovery of the high-resolution dual-channel signal.

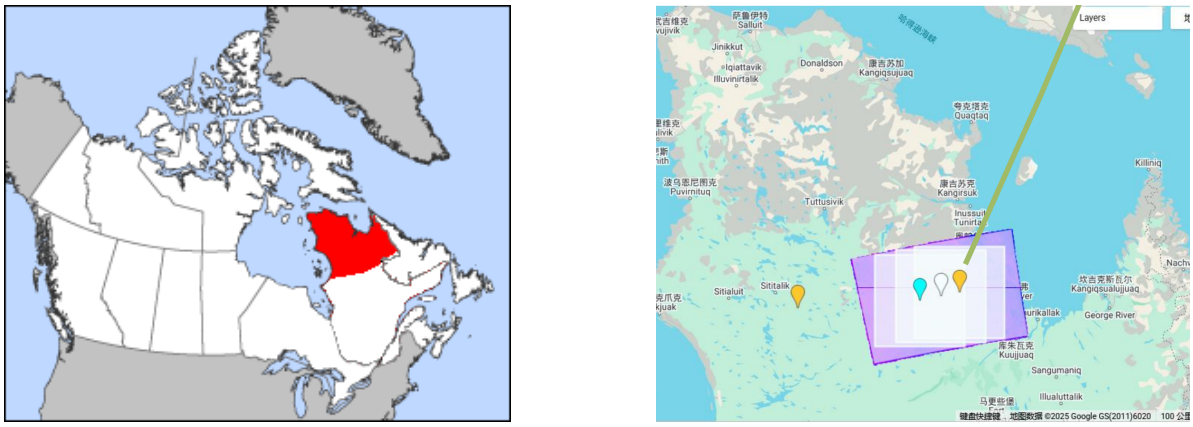

Figure 2: (left) Nunavik region within Canada highlighted in red[1]; (right) Zoomed-in view of the Nunavik region showing the three patches of the Nu912 area [3].

2 Study Area and Database

2.1 Nunavik

Nunavik is located in northern Québec, Canada, and spans a vast area above the 55th parallel. The region is home to several Inuit communities and is characterized by subarctic and arctic climatic zones. The area features a variety of permafrost types, ranging from continuous to sporadic and isolated zones, and exhibits significant spatial heterogeneity in vegetation and terrain. These environmental characteristics make Nunavik particularly suitable for remote sensing applications, especially for tasks involving land cover transitions and moisture retrieval [12].
This study focuses on six geographical regions within Nunavik, each subdivided into three center points, resulting in a diverse and spatially distributed dataset. In our work, we used the following subregions: Nu621PX, Nu622PX, Nu623PX, Nu911PX, Nu912PX, Nu912PX, Nu913PX, where *Nu* represents the region name (in this case Nunavik), the first two digits are the path number, then the slice number, and at the end the center point of each region represented by X, which can take the values 1, 2, and 3. From each region, we acquired HR images, LR images and conditional observations, which provide additional information about the region as well as the images themselves. Fig. 2 shows a view of Nunavik and particularly the region Nu912PX.

2.2 Remote Sensing Data

The dataset includes satellite-based Earth observation data from multiple sensors, captured between the years 2017 and 2021. These observations span several spatially diverse regions across Nunavik and form the basis of our supervised learning task. In this project, we consider two primary sources of observations: LR images from the Soil Moisture Active Passive (SMAP) mission and HR images from synthetic aperture radar (SAR) provided by the Sentinel-1 satellite. The SMAP satellite employs an L-band radiometer (1.4 GHz)

[1]Left image source: `https://upload.wikimedia.org/wikipedia/commons/f/fd/Nunavik-Qu%C3%A9bec.PNG`

to provide radiometric measurements with high temporal frequency (every 2–3 days) but limited spatial resolution (9 km). L-band signals penetrate vegetation and soil effectively, making them particularly suitable for retrieving surface moisture information [13]. These LR images serve as the input in our task and are used to construct time series that match the Sentinel-1 acquisition dates.
In contrast, Sentinel-1 uses C-band SAR (5.4 GHz) to produce finer and more detailed imagery with resolutions up to 10 m, albeit at lower temporal frequencies, requiring between 6 and 12 days per acquisition. The C-band exhibits greater sensitivity to surface roughness and vegetation structure, offering higher spatial detail but reduced penetration capability. HR images from Sentinel-1 are available in two different backscatter-polarization modes: vertical transmit and vertical receive (VV) and vertical transmit and horizontal receive (VH). VV images tend to be more sensitive to surface roughness and are useful for observing bare soil and urban areas, while VH polarization is more responsive to volume scattering, making it particularly suitable for detecting vegetation and snow cover [14, 15].These HR timeseries images, including both VV and VH, are available across all regions and serve as the ground truth to be reconstructed from the temporally aligned LR observations. The dataset was built around Sentinel-1 acquisition dates to enable proper pairing with LR SMAP inputs. Ultimately, we aim to generate denser time series at the SMAP revisit frequency, surpassing the temporal resolution of Sentinel-1. Both SMAP and Sentinel-1 datasets are freely available and widely used in remote sensing applications. Besides the main SAR and radiometric inputs, additional conditioning datasets are used to provide extra knowledge and information to our model, which are helpful for training and evaluation. These include:

- **Digital Elevation Model (DEM):** Provides topographical information at a resolution of 30 m. The DEM data used in this work is taken from two sources: the SRTMGL1 product [1] and the Copernicus GLO-30 product [2].
- **Land Cover (LC) Classification:** Directly taken from the Copernicus Global Land Cover 100 m product. This categorical map distinguishes between 11 classes with specific pixel values in the image, where 10: tree cover, 20: scrubland, 30: grassland, 40: cropland, 50: build-up, 60: bare/sparse vegetation, 70: snow and ice, 80: permanent water bodies, 90: herbaceous wetland, 95: mangroves, 100: moss and lichen.
- **Incidence Angle Map:** Collected from Sentinel-1 metadata, this map describes the angle at which the radar signal interacts with the surface, which strongly affects the magnitude and interpretation of backscatter values.

All three additional information are of high importance as they give additional knowledge to our models to reconstruct the HR images from their corresponding LR counterparts. These are all temporally and spatially aligned per sample.

[1] https://developers.google.com/earth-engine/datasets/catalog/USGS_SRTMGL1_003?hl=de
[2] https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_DEM_GLO30?hl=de

2.3 Data Preprocessing

To facilitate supervised learning, all input modalities were resampled and preprocessed to share a consistent spatial resolution and image dimension of 256×256 using channel-wise interpolation. The HR observations consist of two separate channels representing VV and VH polarizations, which are stacked together to form a dual-channel representation. Each LR-HR pair is associated with DEM, LC and incidence angle conditioning inputs that provide additional spatial context (see Section 2.2).
The DEM and incidence angle maps are normalized to ensure numerical stability, and each one consists of a single channel. To preprocess the LC data, we employ a spatial one-hot encoding approach. Each LC map is a single-band raster image, where pixel values denote discrete land cover categories (e.g., 10 = Tree cover, 20 = Scrubland; see Section 2.2). These values are normalized by dividing by 10 and then mapped to integer indices $\{0, \ldots, 8\}$, representing the 9 land cover classes. The resulting indices are converted into a spatial one-hot encoded tensor of shape $9 \times H \times W$, with each channel corresponding to a specific class. This allows the network to learn spatial interactions between semantic regions and terrain features, as each land cover class is represented as an independent binary channel. Such one-hot encoding is widely used in semantic segmentation and deep learning tasks involving categorical raster data [16], as it enables the model to treat each class independently without imposing any ordinal relationships.

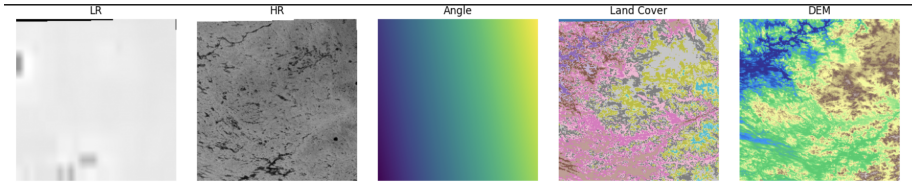

Figure 3: One sample of the LR, HR VV and all three conditionals *before* preprocessing.

Last, we faced the issue that many LR and HR images contained 'Not a Number' (NaN) values near the borders. These NaN values arise from variations in the satellite's viewing angle, which in SAR imaging can produce areas of missing data due to geometric distortions like layover, shadowing, or gaps at scene edges. To mitigate this, we cropped the images vertically and horizontally to remove the affected regions.

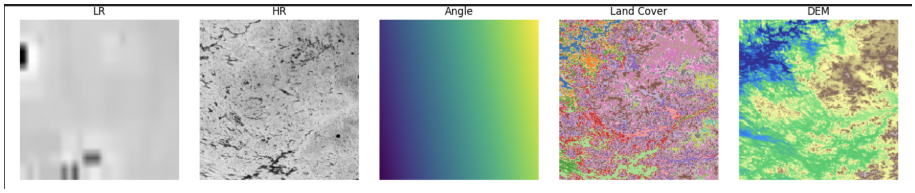

Figure 4: One sample of the LR, HR VV and all three conditionals *after* preprocessing.

Fig. 3 shows a sample from the raw dataset. In the top right corner of both the LR and HR images, we observe black pixels, which correspond to 'NaN' values. These regions need to be removed. As shown in Fig. 4, the 'NaN' values have been cropped and the figure shows the upsampled version of LR, HR and LC, to match the original 256x256 resolution.

3 Methodology

3.1 Vision Transformer Model

In this section, we present a regression model based on the ViT architecture as a baseline for the SR task. Originally introduced in natural language processing, transformers have demonstrated outstanding performance in computer vision, particularly in capturing long-range dependencies and modeling global context [17]. Unlike traditional convolutional networks that rely on local receptive fields, ViTs can aggregate information across large spatial regions, making them particularly suitable for image restoration tasks where spatial context is essential. ViTs have been shown to outperform convolutional networks in various image restoration tasks due to their ability to integrate information across large spatial regions [7].
For our SR problem, we adapt the ViT to reconstruct HR microwave observations from LR inputs and auxiliary conditional data. By modeling global relationships among these inputs, the ViT learns to directly predict the corresponding HR images from Sentinel-1.

3.1.1 Basic Model Architecture

This basic model is adapted from the standard ViT framework, tailored to process multi-channel satellite images for SR tasks. As shown in Fig. 5, the architecture consists of four main components:

- **Patch Embedding:** The input tensor $x \in \mathbb{R}^{B\times C\times H\times W}$ is partitioned into non-overlapping 8×8 patches with stride 8, resulting in a regular grid of patches across the image. Each patch is then flattened and projected into a 1024-dimensional embedding space by a linear layer while ensuring that patch features are layer normalized before entering the transformer. The number of patches is $(H/P) \times (W/P)$, where $P = 8$, denotes the side length of a patch.

- **Positional Encoding:** To retain spatial information among patches, a fixed two-dimensional sine-cosine positional encoding is added to the patch embeddings [17]. The encoding tensor $\mathrm{PE} \in \mathbb{R}^{1\times N\times D}$, where $N = H \cdot W/P^2$ is the number of patches and $D = 1024$ is the embedding dimension, is initialized externally and stored as a trainable parameter. The positional encoding is scaled by 0.01 and added to the patch embedding sequence, providing the transformer with explicit spatial priors. The final patch embedding is computed as

$$\mathrm{E} \leftarrow \mathrm{E} + 0.01 \cdot \mathrm{PE}. \tag{1}$$

- **Transformer Encoder:** A stack of 12 transformer layers processes the sequence of patch embeddings. Each layer contains multi-head self-attention (16 heads, $d_k = 64$ per head, $D = 1024$ total) and a feedforward MLP (2048 hidden size, GELU activation). Group normalization replaces layer normalization for greater stability in small-batch scenarios. The self-attention mechanism allows the model to aggregate global context and capture long-range dependencies essential for high-fidelity image reconstruction.

- **Reconstruction Head:** After processing the patch sequence, the tokens are reshaped into a 2D spatial grid and passed through a lightweight convolutional decoder to reconstruct the high-resolution image. The decoder applies a sequence of convolutional layers:

$$\text{Conv2D: } D \to D/2 \to D/4 \to C_{\text{out}} \cdot P^2, \tag{2}$$

where C_{out} denotes the number of output channels, and P is the patch size. The output is then rearranged from shape $[B, C_{\text{out}} \cdot P^2, H/P, W/P]$ back to the full-resolution image $[B, C_{\text{out}}, H, W]$.

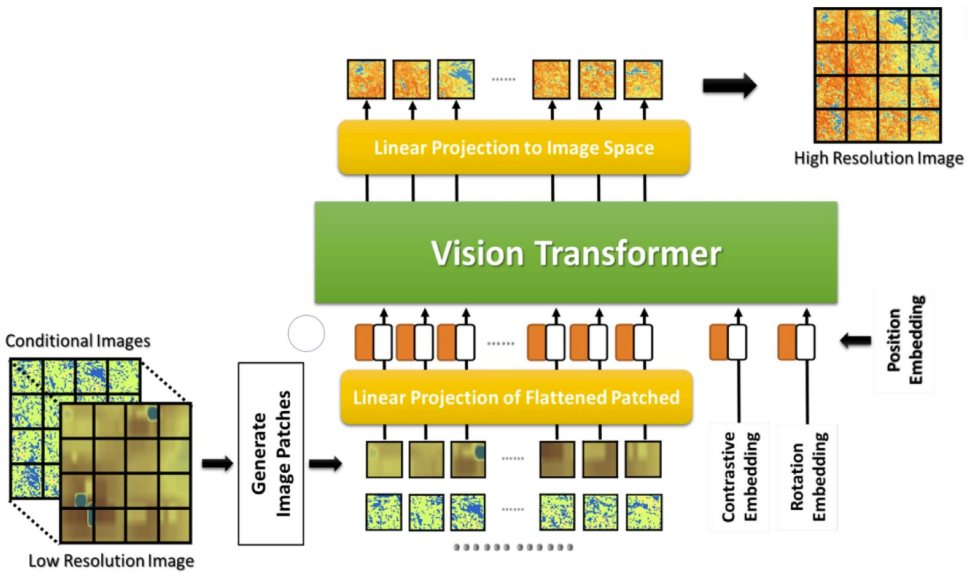

Figure 5: Basic ViT model architecture adapted for SR tasks

3.1.2 Model Improvements

While the basic ViT architecture provides a strong foundation for satellite image super-resolution, we introduce several enhancements to further boost model performance. These improvements target patch embedding and positional encoding, aiming to enhance both reconstruction accuracy and training stability.

- **Convolutional Stem for Patch Embedding:** Instead of flattening each patch directly, we apply a convolutional stem to project local regions into the embedding space [18]. Each patch is passed through a Conv2D layer with kernel size 8 and stride 8, followed by group normalization and GELU activation [19]:

$$E_{i,j} = \text{GELU}(\text{GroupNorm}(\text{Conv2D}(\text{patch}_{i,j}))). \tag{3}$$

This design captures local spatial features while avoiding explicit reshaping operations, and helps stabilize training. Group normalization is used in place of layer normalization to provide more stable normalization under small-batch settings [20].

Additionally, we implement a variant of patch embedding strategy: overlapping patches with stride P/2 and zero-padding. The overlapping version increases receptive field and reduces grid artifacts compared to the non-overlapping version. Grid artifacts refer to visible blocky patterns or discontinuities in the reconstructed image that arise from processing non-overlapping patches separately.

- **Learnable Positional Encoding:** In addition to fixed encoding, we implement a learnable positional embedding mechanism that treats spatial positional priors as trainable parameters. The positional embedding tensor $\mathrm{PE} \in \mathbb{R}^{1\times H\cdot W\times D}$ is initialized with truncated normal noise :

$$\mathrm{PE} \sim \mathcal{N}(0, 0.02^2), \tag{4}$$

so that only random values within the range of the mean plus or minus two standard deviations are kept, while values outside this range are resampled. This ensures that the parameters are not too large or too small at initialization, resulting in more stable training.

The PE is optimized jointly with the model weights during training. This approach allows the network to adaptively learn positional patterns directly from data, potentially capturing task-specific inductive biases beyond what fixed encoding can offer. Compared with the fixed sine-cosine scheme, learnable embeddings provide increased flexibility at the cost of additional parameters.

3.1.3 Loss Function

For training, the primary objective is the mean squared error (MSE) between the predicted high-resolution output $\hat{y}$ and the ground truth y:

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{BHW}\sum_{b=1}^{B}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(\hat{y}_{b,h,w} - y_{b,h,w}\right)^2, \tag{5}$$

where B is the batch size, and H, W denote the spatial dimensions of the image.
In addition, the total variation loss is introduced to improve local spatial consistency and suppress high-frequency artifacts:

$$\mathcal{L}_{\mathrm{TV}} = \frac{1}{B(H-1)(W-1)}\sum_{b=1}^{B}\sum_{h=1}^{H-1}\sum_{w=1}^{W-1}\left|\hat{y}_{b,h+1,w} - \hat{y}_{b,h,w}\right| + \left|\hat{y}_{b,h,w+1} - \hat{y}_{b,h,w}\right|. \tag{6}$$

This loss reduces local intensity oscillations and helps maintain smooth surfaces in low-texture areas [21]. Furthermore, to preserve edge sharpness, an edge-aware loss using Sobel operators S_x and S_y is introduced, which encourages alignment between the gradient maps of prediction and ground truth, thus preserving boundary and edge details. [22]:

$$\mathcal{L}_{\mathrm{edge}} = \frac{1}{BHW}\sum_{b=1}^{B}\left\|S_x * \hat{y}_b - S_x * y_b\right\|_1 + \left\|S_y * \hat{y}_b - S_y * y_b\right\|_1, \tag{7}$$

The overall loss used during training is a weighted combination of the three components:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{MSE}} + \lambda_{\mathrm{TV}}\mathcal{L}_{\mathrm{TV}} + \lambda_{\mathrm{edge}}\mathcal{L}_{\mathrm{edge}}, \tag{8}$$

where we empirically set $\lambda_{\mathrm{TV}} = 10^{-5}$ and $\lambda_{\mathrm{edge}} = 5\times 10^{-3}$.

3.2 Denoising Diffusion Model

In this section, we present a denoising diffusion model as an alternative to our existing ViT-based super-resolution network. We train a Denoising Diffusion Probabilistic Model (DDPM) [5] and use a deterministic Denoising Diffusion Implicit Model (DDIM) [6] for fast, high-fidelity inference.

DDPMs excel at modeling the complex, multi-modal distributions often encountered in remote sensing imagery. By learning to reverse a gradual, pixel-wise noising process, they inherently capture both low- and high-frequency components of natural scenes, e.g. terrain roughness, water boundaries and vegetation textures, without collapsing to an average solution. Unlike convolutional or transformer-based priors that may bias towards smooth reconstructions, diffusion models naturally represent sharp transitions (e.g. cliff edges or coastlines) through their learned score fields at each timestep. This capacity is further amplified when conditioning on LR images plus auxiliary channels (DEM, EWC, incidence angle): the denoiser learns to fuse geophysical context with spectral detail, yielding HR outputs that respect both sensor geometry and land-cover heterogeneity. Moreover, because the diffusion objective matches the KL divergence between the learned distribution and the data distribution [5], the model places extra emphasis on low-probability but materially important features such as narrow river channels or localized soil moisture anomalies that are often lost under MSE-based losses.

While DDPM sampling is stochastic and typically involves hundreds of steps, DDIM introduces a non-Markovian, deterministic sampling process that can significantly accelerate inference without degrading image quality [6]. We use DDIM during sampling, as prior work has shown it can provide sharper and more faithful reconstructions in image super-resolution tasks compared to standard DDPM sampling [23, 24]. This is particularly beneficial for HR outputs, where deterministic trajectories help preserve structural details.

3.2.1 Denoising Diffusion Probabilistic Model for Training

In our super-resolution task, where the ground truth HR image is defined as $x_0 \in \mathbb{R}^{1\times 256\times 256}$, the objective of DDPMs is to learn a parameterized distribution $p_\theta(x_0)$ that closely approximates the data distribution $q(x_0)$ and enables efficient sample generation. In general, DDPMs are latent variable models defined as $p_\theta(x_0) := \int p_\theta(x_{0:T})\,\mathrm{d}x_{1:T}$, where $x_1, \ldots, x_T$ are latent variables sharing the same dimensionality as the observed data x_0 [5, 6]. A Markov assumption is imposed on this *reverse process*, yielding the factorization $p_\theta(x_{0:T}) := p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t)$, where the prior $p(x_T)$ is set to a standard Gaussian. Following [5, 6], the transitions are modeled as isotropic Gaussians, $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I})$, where μ_θ is learned by a neural network and σ_t is a fixed variance schedule.

In our remote sensing super-resolution setup, we adapt this framework to model the conditional distribution $p_\theta(x_0 \mid c)$, where $c \in \mathbb{R}^{12\times 256\times 256}$ comprises the corresponding LR image and the auxiliary observations EWC, DEM and incidence angle map (see Section 2.2). These conditioning variables are provided to the model at each step of the reverse process by concatenating them channel-wise to the input of the denoising network. In particular, the reverse process becomes conditional, taking the form $p_\theta(x_{0:T} \mid c) := p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t, c)$, where each transition is modeled as $\mathcal{N}(x_{t-1}; \mu_\theta(x_t, t, c), \sigma_t^2\mathbf{I})$.

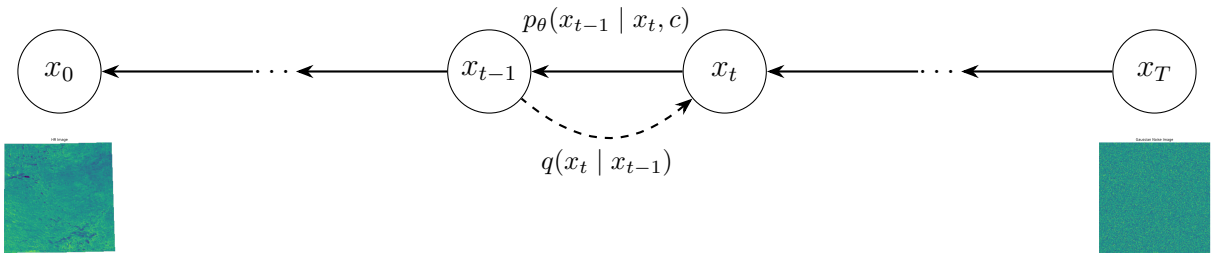

Figure 6: Denoising diffusion probabilistic process [5]. The model learns $p_\theta(x_{t-1} \mid x_t, c)$ to reverse the forward noising process $q(x_t \mid x_{t-1})$. Starting from pure noise x_T, we condition in every step on the noisy latent x_t and auxiliary obsevations c such as the LR image.

This formulation allows the model to exploit both the spatial structure encoded in the LR image and the physical priors encoded in the DEM, EWC, and incidence angle when learning the means $\mu_\theta(x_t, t, c)$.
Complementary to the learned reverse process, the forward process $q(x_{1:T} \mid x_0)$ is a fixed Markov chain that gradually adds Gaussian noise to the data according to a predefined variance schedule $\beta_1, \ldots, \beta_T$ [5, 6]:

$$q(x_{1:T} \mid x_0) := \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \quad \text{with} \quad q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}\right). \tag{9}$$

As illustrated in Fig. 6, this forward noising process gradually corrupts the original image x_0 until reaching nearly standard Gaussian noise at step T. Note that this process is independent of the conditionals c. Crucially, x_t can be sampled at any intermediate timestep t in closed form using the reparameterization:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{10}$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. As $\alpha_T \to 0$, the distribution $q(x_T \mid x_0)$ approaches a standard normal, and this is the reason for choosing the reverse process prior as $p(x_T) := \mathcal{N}(\mathbf{0}, \mathbf{I})$.
Training the model consists of maximizing the evidence lower bound on the conditional negative log-likelihood of the data. Following Ho et al. [5], instead of directly learning to predict x_0, our model is trained to predict the noise ϵ added at timestep t using the reparameterization in (10). This corresponds to learning a noise prediction network $\epsilon_\theta(x_t, t, c) \approx \epsilon$, conditioned on the noisy input x_t, timestep t, and the conditioning data c. This leads to a simplified and effective training objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t, x_0, \epsilon}\left[\|\epsilon - \epsilon_\theta\left(x_t, t, c\right)\|^2\right]. \tag{11}$$

In particular, this formulation leverages the auxiliary data c throughout the denoising process, guiding the model towards high-quality super-resolved images.

3.2.2 Denoising Diffusion Implicit Model for Sampling

While DDPMs define a generative model through a stochastic Markovian reverse process with Gaussian transitions, DDIMs offer a deterministic alternative that preserves the

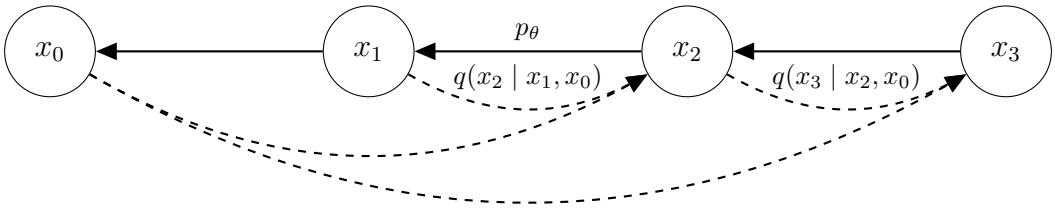

Figure 7: Non-Markovian DDIM (noising) process [6]. During the forward process, each x_t depends on x_{t-1} and x_0.

same marginal distributions as the DDPM forward process [6]. Based on this observation another non-Markovian forward process with the same marginals is introduced [6]:

$$q(x_{1:T} \mid x_0) := q(x_T \mid x_0) \prod_{t=2}^{T} q(x_{t-1} \mid x_t, x_0) \tag{12}$$

$$q(x_{t-1} \mid x_t, x_0) := \mathcal{N}\big(x_{t-1}\,;\ \sqrt{\alpha_{t-1}}\,x_0 \ +\ \sqrt{1-\alpha_{t-1}-\sigma_t^2}\ \frac{x_t - \sqrt{\alpha_t}\,x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\, I\big) \tag{13}$$

Specifically, we obtain $q(x_t \mid x_{t-1}, x_0) = \frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)}$. Note that every latent x_t variable does not only depend on x_{t-1} as in DDPM models but also on x_0. Fig. 7 illustrates this relation. Importantly, the forward diffusion process remains entirely unconditional, independent of our remote-sensing-specific conditionals c.
As the DDIM forward process is compatible with the DDPM training procedure, we can use our pretrained DDPM model for sampling with DDIM. For a decreasing sequence of timesteps $\tau_1 > \tau_2 > \cdots > \tau_K$, we sample as in [6]:

$$x_{\tau_{k-1}} = \sqrt{\bar{\alpha}_{\tau_{k-1}}}\,x_0 + \sqrt{1-\bar{\alpha}_{\tau_{k-1}}-\eta^2\sigma_{\tau_k}^2}\cdot\epsilon_\theta(x_{\tau_k},\tau_k,c) + \eta\,\sigma_{\tau_k}\cdot\epsilon, \quad \epsilon\sim\mathcal{N}(0,\mathbf{I}), \tag{14}$$

where $\sigma_{\tau_k}^2 = \frac{1-\bar{\alpha}_{\tau_{k-1}}}{1-\bar{\alpha}_{\tau_k}}\cdot\beta_{\tau_k}$. x_0 is estimated by rewriting (10):

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_{\tau_k}}}\left(x_{\tau_k} - \sqrt{1-\bar{\alpha}_{\tau_k}}\cdot\epsilon_\theta(x_{\tau_k},\tau_k,c)\right). \tag{15}$$

As we reuse the pretrained DDPM noise prediction network $\epsilon_\theta\,(x_t, t, c)$, the sampling process is conditioned on the conditionals c. The parameter $\eta \in [0, 1]$ controls the level of stochasticity in the reverse process [6]. We set $\eta = 0$ to eliminate the final noise term entirely and to obtain a fully deterministic sampling trajectory. Moreover, the non-Markovian structure enables accelerated sampling: since each reverse step depends only on the current latent x_{τ_k} and an estimate of x_0, we can evaluate the model on a coarse schedule of time steps (e.g., 20 or 50 instead of 1000) without significant degradation in sample quality. In our conditional super-resolution setting, this results in fast and stable generation of HR images.
Finally, training under the original DDPM framework remains reasonable even when sampling is performed with the deterministic DDIM sampler. First, the training process simplifies to a noise prediction task, where the model learns to estimate the Gaussian

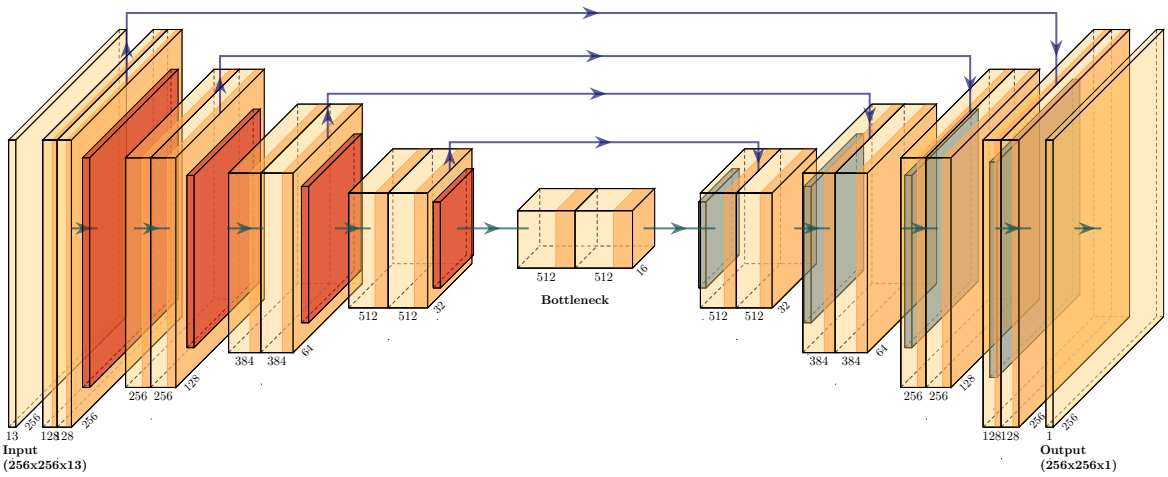

Figure 8: Encoder–Decoder U-Net architecture used in the diffusion pipeline.

noise added at each timestep, making it computationally efficient and easy to implement. Moreover, the Markovian forward process allows losses to be computed at random timesteps without needing to evaluate the full trajectory [5, 6]. Second, the stochastic DDPM training objective corresponds to learning the score function of the data distribution at all noise levels β_t, which has been shown to yield robust denoising networks that generalize well across both small and large perturbations of the input [5, 6]. Last, this makes us very flexible in adjusting the number of subsampled timesteps at inference.

3.2.3 U–Net

The diffusion model employs a U-Net encoder-decoder [25] whose overall layout is illustrated in Fig. 8. Following the work in [25, 26, 5], our architecture adopts four resolution stages, each with two residual blocks. The latter increases modeling capacity and stabilizes training. Further, the exact number of channels at each stage was determined through hyperparameter search during preliminary experiments. A concise description of each component is given below.

Encoder An initial 3×3, stride-1, padding-1 convolution projects the 13 input channels (noisy sample plus conditionals, see Section 2.3) to the 128 channels expected by the encoder of the U-Net. Then, the encoder comprises four resolution stages. In the k-th stage, a tensor of shape $H_k \times W_k \times C_k$ is passed through two consecutive residual units, each executing the sequence

$$\text{input} \xrightarrow{\text{GroupNorm}} \xrightarrow{\text{SiLU}} \xrightarrow[\substack{\text{stride } 1 \\ C_{\text{in}} \to C_{\text{out}}}]{3\times3 \text{ conv}} \xrightarrow{+t \text{ embed}} \xrightarrow{\text{GroupNorm}} \xrightarrow{\text{SiLU}} \xrightarrow[\substack{\text{stride } 1 \\ C_{\text{out}} \to C_{\text{out}}}]{3\times3 \text{ conv}} \xrightarrow{+ \text{ skip}} \text{out},$$

where in the first unit $C_{\text{in}} = C_k$ and $C_{\text{out}} = C_{k+1}$, and in the second unit both are C_{k+1}. Following Ho et al. [5], we add a learned linear embedding of the current timestep t of our diffusion process after the first convolution. After these two residual units, a learned 3×3 convolution with stride 2 downsamples the spatial dimensions $(H_k \times W_k) \to (\frac{H_k}{2} \times \frac{W_k}{2})$ while leaving the number of channels C_{k+1} invariant.

Bottleneck At the smallest scale, two more residual units - identical to those in the encoder - process the $16 \times 16 \times 512$ tensor. Each unit again injects the timestep embedding between its convolutions, but no spatial rescaling occurs, so the output remains $16 \times 16 \times 512$, carrying fully conditioned features into the decoder.

Decoder The decoder mirrors the encoder in reverse over four stages. Each stage begins by upsampling its input tensor of shape $H'_k \times W'_k \times C_{k+1}$ back to $H_k \times W_k$ using a transposed convolution with a 4×4 kernel, stride 2, and padding 1, maintaining the same number of channels C_{k+1}. The upsampled tensor is then concatenated along the channel dimension with a skip connection stored from the second residual block in the corresponding encoder stage, resulting in a tensor with $2C_{k+1}$ channels. The concatenated tensor is passed into the first residual unit, which follows

$$\text{input} \xrightarrow{\text{GroupNorm}} \xrightarrow{\text{SiLU}} \xrightarrow[\substack{\text{stride } 1 \\ 2C_{k+1} \to C_{k+1}}]{3\times3 \text{ conv}} \xrightarrow{+t \text{ embed}} \xrightarrow{\text{GroupNorm}} \xrightarrow{\text{SiLU}} \xrightarrow[\substack{\text{stride } 1 \\ C_{k+1} \to C_k}]{3\times3 \text{ conv}} \xrightarrow{+ \text{ skip}} \text{out}\,.$$

This block reduces the channels from $2C_{k+1}$ to C_k. The output then passes through a second residual unit that has identical structure but operates on C_k channels throughout, leaving the channel dimension unchanged. Both blocks preserve the spatial size ($H_k \times W_k$). Finally, a 3×3, stride-1, padding-1 convolution projects the 128 channels to a single map providing the noise estimate needed to be subtracted at time step t to restore your ground-truth.

3.3 Model Configurations

We conducted a series of experiments using varied data and architectural configurations to assess the performance and generalization of our models. These were specifically designed to evaluate the models' robustness across different conditions.

3.3.1 Experimental Setups

We define four settings that introduce variation across time, space, and polarization. All experiments are performed on the Nunavik dataset presented in Section 2.2. Further, unless otherwise stated, we super-resolve only the VV polarization.

- **Full-Data**: Models are trained on all available data across all time periods and all six Nunavik regions. Specifically, the training and validation split was conducted by randomly shuffling the dataset and assigning 90% to training and 10% to validation. This setup provides an upper-bound baseline by leveraging full temporal and spatial coverage.
- **Generalization across Time**: Models are trained and validated on samples from two temporal splits across all regions. First, we train on data from the year 2017 to 2020 and test on 2021 samples. Second, we repeat the same experiment by training on 2017, 2019-2021 and test on 2018 to rule out outliers in the climate cycle. This setting assesses the model's ability to generalize on unseen time periods, a crucial requirement for long-term Earth observation forecasting.

- **Generalization across Space**: Models are trained and validated on five source regions and then tested on the unseen sixth region to assess spatial transferability. For the ViT model, we further perform fine-tuning on the unseen data to enable domain adaptation.
- **Dual-Channel Reconstruction**: As all other experiments focus on super-resolving the VV polarization, we extend our models in this experiment to jointly predict both VV and VH polarizations. This multi-output setting supports richer structural modeling and encourages the network to learn complementary spatial features across channels. In particular, we evaluate the extent to which the performance achieved on VV polarization generalizes to VH polarization. For this experiment, we use the same regional splits as in the the full-data setup.

3.3.2 Evaluation Metrics

To compare the performance of our models quantitatively, we adopt four evaluation metrics: Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Spearman Correlation Coefficient (SCC). These metrics capture different aspects of image quality, including pixel-wise error, perceptual fidelity, and spatial structure preservation.

- **MSE**: Measures the average squared difference between the predicted image $\hat{I}$ and the ground truth I:

$$\text{MSE} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(I_{ij} - \hat{I}_{ij} \right)^2 \tag{16}$$

  Lower MSE indicates higher pixel-wise accuracy.

- **PSNR**: Evaluates the reconstruction fidelity in decibels. It is computed from the MSE as:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{(R_{\max} - R_{\min})^2}{\text{MSE}} \right) \tag{17}$$

  where $R_{\max}$ and $R_{\min}$ are the global maximum and minimum values in the reference image.

- **SSIM**: Captures perceptual quality by comparing luminance, contrast, and structure between the predicted and ground truth images:

$$\text{SSIM}(I, \hat{I}) = \frac{(2\mu_I \mu_{\hat{I}} + C_1)(2\sigma_{I\hat{I}} + C_2)}{(\mu_I^2 + \mu_{\hat{I}}^2 + C_1)(\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2)} \tag{18}$$

  where μ, σ^2, and $\sigma_{I\hat{I}}$ are local statistics, and C_1, C_2 are small constants. SSIM ranges from 0 to 1.

- **SCC**: Measures spatial correlation by computing the Spearman rank-order correlation between the predicted and ground truth image intensities:

$$\text{SCC} = \rho_s = \text{corr}(\text{rank}(I), \text{rank}(\hat{I})) \tag{19}$$

  Higher SCC indicates better preservation of structural and regional information.

3.3.3 Training and Hyperparameter Setup

Next, we introduce our training configurations. For both models and in all experiments, z-score normalization is applied across the input, including the conditionals. Specifically, given an input value x, the z-score normalized value z is centered around zero with unit variance: $z = (x - \mu)/\sigma$ where μ and σ denote the mean and standard deviation of the data per channel. This provides more balanced gradient signals, which helps stabilize and accelerate training. As both the LR and HR images are working with a dynamic range, i.e. 32-bit floating points, this method is particularly well-suited, as it does not constrain the data to a fixed interval.
For training the ViT model, we use the Adam optimizer [27] with an initial learning rate of 10^{-5} during pretraining and reduce it to 2×10^{-7} during fine-tuning for spatial transfer. The StepLR scheduler from the PyTorch library [28] is applied with a step size of 300 and decay factor $\gamma = 0.1$. The model is trained for 500 epochs. Input channels are obtained from the preprocessed data mentioned in Section 2.3.
During training of the denoising diffusion model, we use $T = 1000$ diffusion steps and a linearly increasing DDPM variance schedule $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ as in [5]. For sampling with our DDIM noise scheduler we used $N = 50$ uniformly spaced timesteps, i.e. $t_i = \lfloor \frac{i \cdot T}{N} \rfloor$ where $i \in [N]$. For optimization, we use Adam optimizer [27] with decoupled weight decay regularization [29] and a cosine annealing learning rate schedule [30]. Last, dropout is applied to the ResNet blocks in the U-Net after the second activation function with a probability of 10%.
Last, in certain experiments, we work with data augmentation. Specifically, we apply horizontal and vertical flipping during training. However, more complex augmentations such as rotation are avoided due to potential structural distortion in repetitive remote sensing patterns.

3.3.4 Runtime Environment

All experiments are conducted on a server running Ubuntu 20.04.4 LTS with Linux kernel version 5.4.0-216-generic, equipped with dual AMD EPYC 7413 24-core CPUs (96 threads total), 503GB of RAM, and a single NVIDIA GeForce RTX 3090 GPU with 24.GB of memory. The software stack comprises CUDA 11.4, NVIDIA driver version 470.256.02, and PyTorch 2.3.0. All training and inference procedures are executed in mixed-precision where applicable.

4 Experiments and Results

In this section, we present the results and performance of our ViT and denoising diffusion model obtained on the Nunavik dataset presented in Section 2.2.

4.1 Preliminary Experiments

To find an optimal training procedure, we started our experiments with a hyperparameter search and tested the models performance with different setups.

4.1.1 ViT Model

Table 1: Comparision of different training settings of the ViT Models trained on all conditions for 500 epochs on all regions

Setup	MSE ↓	PSNR (dB) ↑	SSIM ↑	SCC ↑
Baseline	4.48	27.06	0.63	0.50
+ No Incidence Angle	4.42	27.19	0.63	0.52
+ Data Argumentation	4.34	27.23	0.64	0.52
+ No Angle + Aug	**4.26**	**27.27**	**0.64**	**0.55**

As shown in Table 1, we systematically examined the influence of input conditions and training strategies on the performance of our improved ViT models introduced in Section 3.1.2. Removing the incidence angle channel from the input resulted in a modest, yet consistent, improvement across all quantitative metrics. This suggests that the model is capable of maintaining robust performance even when certain auxiliary information is omitted, likely because the incidence angles are provided as full input images, introduce a large amount of redundant or weakly relevant information into the model. While these values occupy significant capacity during training, their actual contribution to super-resolution performance appears minimal in our specific setting.
Furthermore, we assessed the effect of data augmentation through random vertical and horizontal flipping of the input images. For the ViT model, data augmentation led to additional gains in both PSNR and SSIM, indicating improved reconstruction fidelity and structural preservation. The best overall results were achieved when both the incidence angle channel was excluded and data augmentation was applied, with the model attaining its highest PSNR, SSIM, and spatial correlation scores. In all subsequent experiments, we therefore adopt this configuration as our default setting and train for 500 epochs.

4.1.2 Denoising Diffusion Model

To determine an appropriate optimizer step size for the diffusion training process, we conducted a brief hyperparameter search over several learning rates. An initial learning rate of 10^{-4} was found to be optimal. Although lower rates also allowed convergence, they resulted in a markedly slower decrease of the training loss. Conversely, higher rates led to degraded performance-manifesting as unstable loss fluctuations and poorer final accuracy-most likely because the larger step size caused the optimizer to overshoot minima and hinder effective gradient-based refinement.
Next, we tested the impact of data augmentation, i.e. random flipping. Contrary to our expectations, applying these augmentations led to a substantial drop in both training and validation performance, with no measurable improvement on the validation data. This behavior is most likely a consequence of our highly specific remote-sensing scenario: although we use around 2.500 images, they originate from only six regions whose spatial patterns are extremely repetitive. By distorting these patterns, augmentation prevents the model from learning the region-specific textures and structures that are critical for accurate reconstruction.

Table 2: Different training setups for denoising diffusion model. The baseline model is trained including all three conditionals (DEM, EWC and incidence angle). In all setups, we trained over 500 epochs on all 6 Nunavik regions.

Setup	**MSE ↓**	**PSNR (dB) ↑**	**SSIM ↑**	**SCC ↑**
Baseline	4.52	27.48	0.78	0.65
+ No Incidence Angle	4.84	27.75	0.80	0.69
+ Data Augmentation	5.92	25.67	0.57	0.32
+ Dropout (p=0.1)	**3.88**	**28.43**	**0.86**	**0.79**
+ No Angle + Aug	5.79	26.15	0.71	0.54

Subsequently, we investigated the impact of dropout regularization. Incorporating dropout into the U-Net backbone of our DDPM led to a notable improvement in both training stability and generalization. We attribute this gain to dropout's capacity to disrupt co-adaptation among filters, thereby encouraging the network to learn more robust, redundant feature representations. In the context of our highly repetitive, region-specific remote sensing imagery, these regularized features better capture subtle spatial variations without overfitting to idiosyncratic patterns.
Finally, we evaluated the effect of omitting the incidence angle conditioning from our model. Since the incidence angle does not directly encode soil or subsurface properties, its utility was uncertain a priori. In our experiments, the incidence angle did not prove beneficial. However, given the limited size and geographic scope of our data, the potential advantages of incorporating the incidence angle as an additional conditioning variable may become more evident when applied to broader, globally distributed collections. All of our preliminary analysis is summarized in Table 2. Based on these observations, we trained the denoising diffusion model without applying data augmentation or incorporating incidence angle conditionals, while including dropout regularization. Compared to the setup in Table 2, we further extended the training duration to 1000 epochs, as this led to a slight improvement in performance.

4.2 Quantitative Evaluation on Full Data

To evaluate region-level performance, we first assess the full-data ViT model trained on all six Nunavik regions on all time periods. Notably, Nu911 achieves the best MSE and SCC among all regions, likely due to its simpler terrain and smoother water bodies. These results highlight the ViT model's capacity to model complex spatial patterns using purely attention-based architectures. However, performance in regions with more fragmented water structures (e.g., Nu6XX) is comparatively lower, suggesting room for improvement in modeling fine-grained local textures. Fig. 9 shows one super-resolved image from each region predicted by the ViT Model.
In terms of the denoising diffusion model, we obtain superior results in the Nu9XX regions, especially in Nu911, similar to the results of the ViT. Again, this performance can be attributed to the lower spatial complexity of these areas, which feature prominent rivers and large lakes that are easily identifiable and learnable. In contrast, the Nu6XX regions exhibit numerous small water bodies and highly varied spatial structures, posing greater

Table 3: Per-model performance comparison across regions.

Model	Region	MSE ↓	PSNR (dB) ↑	SSIM ↑	SCC ↑
ViT	Nu621	4.36	26.83	0.6259	0.4621
	Nu622	4.55	27.28	0.6369	0.4983
	Nu623	5.66	26.41	0.5985	0.4882
	Nu911	2.78	28.49	0.6991	0.6312
	Nu912	4.15	27.04	0.6509	0.5655
	Nu913	4.33	27.32	0.6293	0.6179
Diffusion	Nu621	3.86	28.00	0.8614	0.7329
	Nu622	3.16	29.48	0.8491	0.7721
	Nu623	4.41	28.36	0.8086	0.7239
	Nu911	2.79	30.14	0.9099	0.8639
	Nu912	3.81	28.13	0.8746	0.7923
	Nu913	4.25	28.15	0.8630	0.8159
Overall	ViT	**4.26**	**27.27**	**0.6423**	**0.5537**
	Diffusion	**3.74**	**28.68**	**0.8618**	**0.7890**

challenges for the model. Fig. 10 shows one super-resolved image from each region.
To facilitate a more direct comparison between the two approaches, we summarize the performance of both ViT and denoising diffusion models across all six regions in Table 3. The diffusion-based model consistently outperforms the ViT baseline, particularly in perceptual and structural metrics such as SSIM and SCC, further confirming the effectiveness of probabilistic generative modeling for complex spatial reconstruction.

4.3 Generalization across Time

One of the main applications of our super-resolution models is time series forecasting. As described in Section 2.2, our dataset spans from 2017 to 2021, making it suitable for evaluating temporal generalization. In this setting, we assess the ability of both ViT model and denoising diffusion model to generalize across time. Following Section 3.3.1, both are trained exclusively on data from 2017 to 2020 and tested on samples from 2021. Additionally, we repeat the same experiment for training on 2017 as well as 2019-2021 and validating on 2018. Results and comparison are illustrated in Table 4.
For both models, we observe a substantial drop in performance compared to the randomly shuffled validation results reported in Section 4.2. For the 2018 validation split, the ViT and diffusion model achieve relatively strong results, with the latter still reaching an average PSNR of 27.22. However, when evaluating on 2021, performance deteriorates significantly, with both models dropping to a PSNR of below 25.
These results can be attributed to two primary factors. First, the comparatively stronger performance in 2018 is anticipated, as this year lies near the center of the temporal range, with training data available both before and after the corresponding timestamps. This facilitates interpolation, which is generally less challenging than the extrapolation required for the 2021 split at the end of the timeline. Second, the performance loss compared to

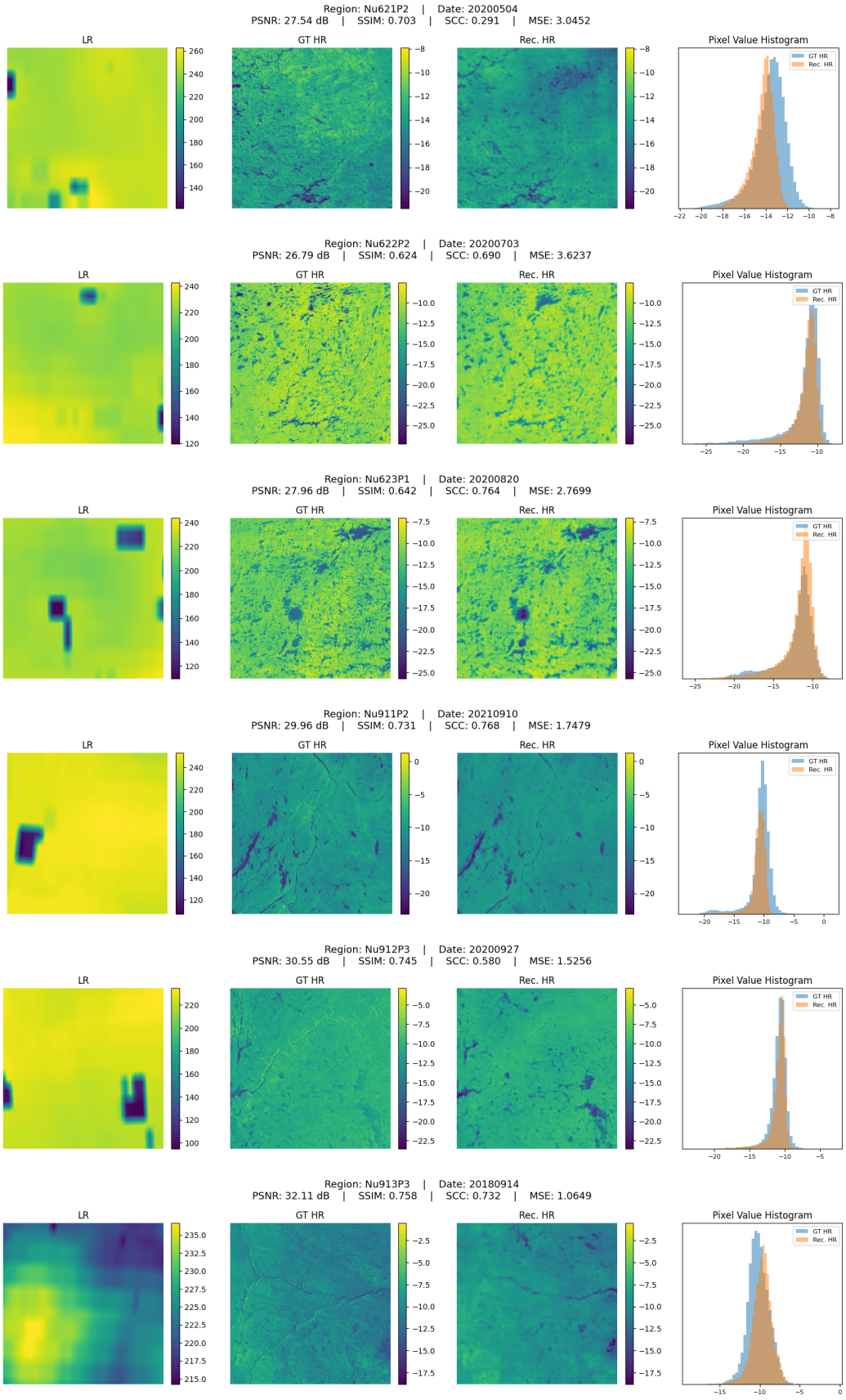

Figure 9: Samples generated by the ViT Model with data argumentation enabled and no incidence angle included from a randomly shuffled validation set. One sample per region is shown.

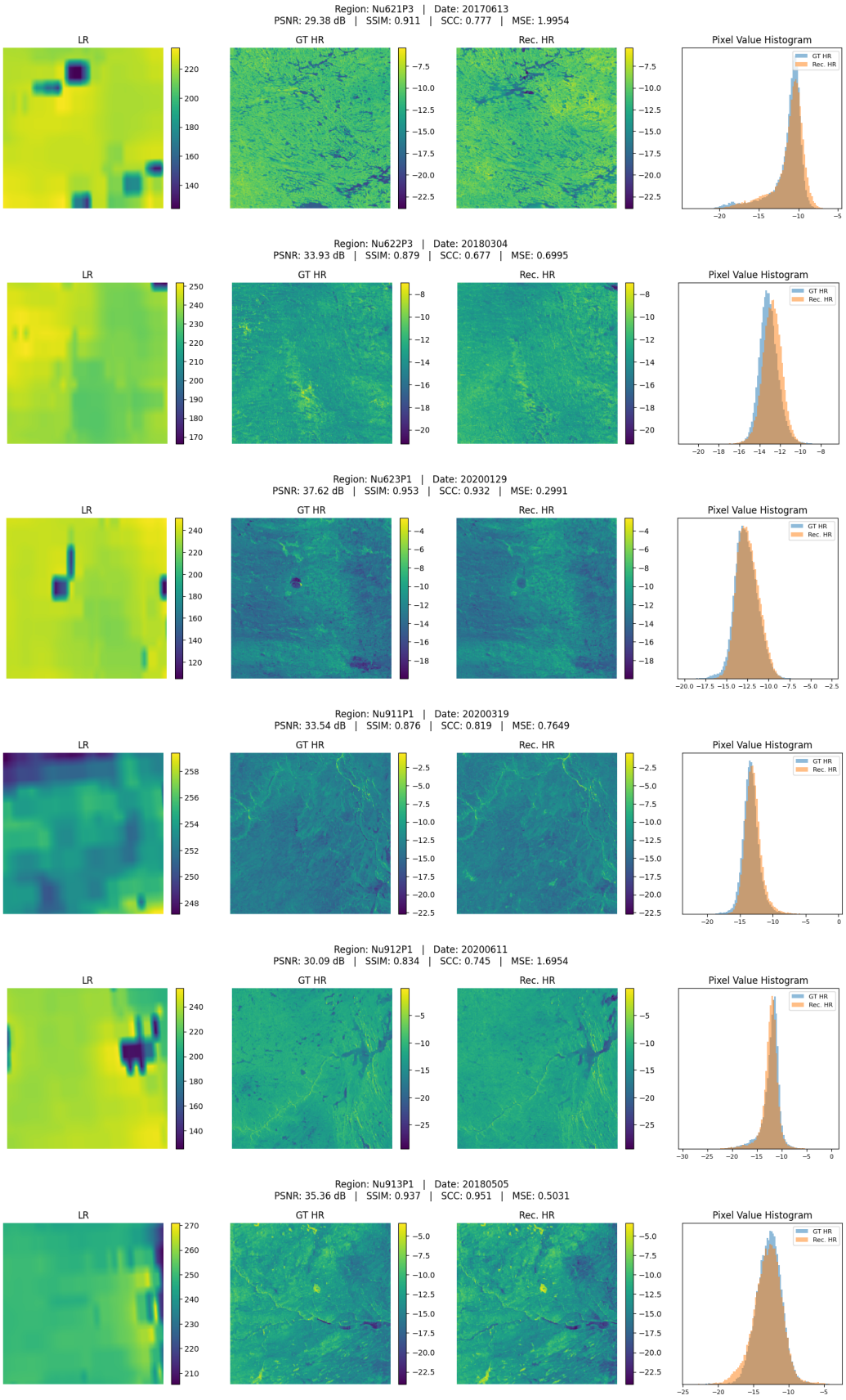

Figure 10: Samples generated by the denoising diffusion model with dropout enabled and no incidence angle included from a randomly shuffled validation set. One sample per region is shown.

Table 4: Temporal transfer capabilities of ViT and denoising diffusion models. Metrics are reported for validation years 2018 and 2021.

Model	**Region**	**MSE ↓**		**PSNR (dB) ↑**		**SSIM ↑**		**SCC ↑**	
		2018	2021	2018	2021	2018	2021	2018	2021
ViT	Nu621	5.00	10.32	26.64	22.29	0.66	0.54	0.45	0.24
	Nu622	5.66	10.24	26.74	22.30	0.65	0.55	0.42	0.31
	Nu623	6.65	11.16	25.69	21.74	0.56	0.51	0.34	0.25
	Nu911	4.79	6.71	27.21	24.60	0.73	0.65	0.57	0.43
	Nu912	5.00	9.79	26.84	22.56	0.67	0.59	0.50	0.41
	Nu913	7.18	9.79	25.90	22.61	0.65	0.60	0.57	0.56
Diffusion	Nu621	5.15	5.39	27.30	25.32	0.78	0.59	0.60	0.27
	Nu622	5.92	5.50	26.97	25.13	0.73	0.55	0.58	0.24
	Nu623	5.84	8.15	27.36	24.30	0.75	0.47	0.60	0.19
	Nu911	4.51	5.00	28.02	25.86	0.85	0.71	0.72	0.47
	Nu912	5.35	7.15	27.21	24.29	0.81	0.63	0.68	0.45
	Nu913	7.16	9.14	26.49	23.67	0.79	0.65	0.67	0.52
ViT	**Overall**	**5.76**	**9.65**	**26.49**	**22.69**	**0.65**	**0.57**	**0.48**	**0.37**
Diffusion	**Overall**	**5.69**	**6.72**	**27.22**	**24.76**	**0.78**	**0.60**	**0.64**	**0.36**

the shuffled data can be explained by seasonal variations in land cover (e.g., freeze-thaw cycles, flooding), which is difficult to capture without temporal diversity in the training set.

Although both models, in particular the denoising diffusion model, demonstrate a certain degree of temporal transferability, the results of this experiment indicate that they fail to capture broader climatic cycles. This limitation arises primarily from the restricted temporal scope of our dataset, which spans only five years. A potential future line of research to address this shortcoming would be to incorporate additional conditioning variables, e.g. surface temperature, that may help to better account for environmental anomalies. More results on the temporal transferability are found in Appendix A.4 and Appendix B.2.

4.4 Generalization across Space

Having seen the temporal transfer capabilities of our models, we now assess the spatial transfer abilities on completely unseen regions. First, we construct a dataset by excluding all samples from the Nu621 region (Nu621P1, Nu621P2, Nu621P3) during the training phase. These samples are used exclusively for evaluation and in case of the ViT model for a post-processing fine-tuning stage.

The fine-tuning stage is conducted as follows: The model is initialized from a pretrained checkpoint and trained for 70 additional epochs on the unseen region. The learning rate is reduced to 2×10^{-7} to prevent large parameter updates. Further, to preserve global priors while adapting to local characteristics, we freeze the first 70% of Transformer encoder blocks and allow only the remaining 30% and the convolutional decoder head to be up-

Table 5: Spatial transferability performance comparison of ViT and Diffusion models on Nu621 (target region) and Nu912 (seen region). Fine-tuning improves ViT performance on the target without harming generalization. Diffusion results for comparison.

Region	Method	MSE ↓	PSNR (dB) ↑	SSIM ↑	SCC ↑
Nu621	ViT Baseline	8.26	21.25	0.51	0.40
	ViT Fine-Tuned	6.05	25.72	0.56	0.19
	Diffusion	7.74	22.75	0.39	0.08
Nu912	ViT Baseline	4.84	26.74	0.69	0.62
	ViT Fine-Tuned	5.65	26.49	0.70	0.62

dated. All LayerNorm layers are kept trainable to support domain-specific normalization. This protocol enables effective domain adaptation while avoiding catastrophic forgetting. A detailed overview of the transfer learning procedure is given in Algorithm 1.

Algorithm 1 Transfer Learning for ViT

Input: M_{pre} (pretrained ViT), D_{target} (dataset for target region)
Output: M_{ft} (fine-tuned ViT)
$(D_{\text{train}}, D_{\text{val}}) \leftarrow \text{split}(D_{\text{target}}, 0.9)$ ▷ Split into train and val
$model, opt, sched \leftarrow \text{load\_model}(M_{\text{pre}})$ ▷ Load pretrained
$\text{freeze}(model, :[:70\%])$; $\text{unfreeze}(model, [70\% :] \cup \{\text{norm}, \text{head}\})$ ▷ Prepare Head
$M_{\text{ft}} \leftarrow \text{fine\_tune}(model, D_{\text{train}}, opt, sched, E)$ ▷ Fine-tune on target

Table 5 reports the performance of the denoising diffusion, the ViT and the fine-tuned ViT model on the unseen region Nu621. For both the denoising diffusion and the ViT model, we can observe a very low performance across all metrics. In particular, the spatial similarity (SSIM and SCC) of the samples from the denoising diffusion model drops to a minimum. Accordingly, both models are not usable in a real world application on unseen regions.

With the help of the fine-tuning stage of the ViT, we are able to partially address this issue. After fine-tuning, the model exhibits a marked improvement on the target region Nu621, with a reduction of MSE from 8.26 to 6.05, and an increase in PSNR from 21.25 dB to 25.72 dB. This indicates the effectiveness of lightweight domain adaptation.

To ensure that fine-tuning does not compromise generalization to other regions, we also report performance on Nu912 which was part of the initial training. As seen in the table, the metrics on Nu912 remain stable or slightly improved after fine-tuning, suggesting that the model preserves its performance on the original training distribution while gaining better locality for the new target. Finally, we remark that it is an open question to which extend one can adapt the fine-tuning technique of the ViT model to the denoising diffusion model. More results on the spatial transferability of the diffusion model, in particular, are found in Appendix B.3.

4.5 Dual-Channel Reconstruction

All previous experiments focused exclusively on super-resolving the VV polarization. To extend this setup, we now perform an additional experiment aimed at jointly predicting both VV and VH polarizations within a single forward pass. This joint modeling approach enables both the ViT and denoising diffusion models to exploit shared spatial information while allowing polarization-specific differences to emerge in the two output channels. It also enables more compact architectures and reduces inference cost compared to training two separate models.

For the ViT Model, we introduce the following key changes to support dual-channel prediction: The number of output channels is set to 2, corresponding to VV and VH. The final convolutional layer in the head outputs $2 \times \text{patch\_size}^2$ channels, which are reshaped to $[B, 2, H, W]$. The input tensor format remains unchanged (11 channels excluding the incidence angle). The ground truth is provided as two separate HR images for VV and VH, which are concatenated along the channel dimension. Last, during training, the loss is computed separately for the two channels using MSE and then averaged:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{2BHW} \sum_{b=1}^{B} \sum_{c=1}^{2} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(\hat{I}_{b,c,h,w} - I_{b,c,h,w}\right)^2, \tag{20}$$

Additional regularization terms, such as total variation loss and edge loss (based on Sobel gradients), are applied to each channel independently and added to the total loss.

Regarding the denoising diffusion model, the final layer of the U-Net introduced in Section 3.2.3 needs to be adjusted to output two noise estimates instead of one. Specifically, at each diffusion step t the model receives a noisy 2-channel input $\mathbf{x}_t \in \mathbb{R}^{2\times256\times256}$ representing VV and VH polarizations with added Gaussian noise, and conditional inputs, such as the LR image (see Section 3.2.3), resulting in a total of 14 input channels to the U-Net. The model outputs an estimate of the noise $\hat{\boldsymbol{\epsilon}}_t \in \mathbb{R}^{2\times256\times256}$, jointly for both polarization channels, which is used to denoise the noisy input $\mathbf{x}_t$. Similar to the ViT, the MSE objective is averaged over both channels during training:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{2BHW} \sum_{b=1}^{B} \sum_{c=1}^{2} \sum_{h=1}^{H} \sum_{w=1}^{W} \left(\hat{\epsilon}_{b,c,h,w} - \epsilon_{b,c,h,w}\right)^2, \tag{21}$$

where B is the batch size, $H = W = 256$, and $c = 1, 2$ index the VV and VH channels.

Finally, Table 6 summarizes the performance for both models in dual-channel reconstruction. Both polarizations yield similar results within each model. However, the denoising diffusion model clearly outperforms the ViT, achieving performance close to single-channel VV prediction (Table 3). This improvement arises because the denoising diffusion model leverages two independent noise channels, effectively introducing additional input diversity and enabling the network to learn distinct representations for each polarization. In contrast, the ViT processes both channels deterministically and cannot easily adapt to separate stochastic inputs for different predictions. This represents a fundamental limitation of the ViT architecture.

Table 6: Overall performance comparison of ViT and denoising diffusion model in dual-channel reconstruction.

Model	Channel	MSE ↓	PSNR (dB) ↑	SSIM ↑	SCC ↑
ViT	VV	4.15	20.79	0.43	0.57
	VH	5.84	19.27	0.45	0.68
Diffusion	VV	4.32	27.54	0.77	0.62
	VH	3.99	26.31	0.79	0.75

In Appendix B.4, we present additional visualizations illustrating the strong performance of the denoising diffusion model for dual-channel reconstruction. Fig. 21 shows a super-resolved sample for both polarizations, complemented by two diagnostic plots per sample in Fig. 22. The sixth image in this figure displays the difference between the VV and VH polarizations in the ground truth, while the seventh shows the corresponding difference in the model prediction. These polarization difference maps offer insight into the relative backscattering behavior and enable an assessment of whether the model captures not only individual polarizations but also the inter-polarization structure characteristic of SAR imagery.
Figure 23 in Appendix B.4 presents two additional difference maps per sample. The sixth image depicts the discrepancy between the ground truth VV and the super-resolved VV, while the seventh shows the corresponding difference for the VH polarization. These visualizations enable localized inspection of reconstruction errors for each polarization and provide further insight into the distinct characteristics of the VV and VH polarization in the original SAR data. Combined with the VV-VH difference maps discussed previously, they contribute to a more comprehensive assessment of the model's polarimetric fidelity.

4.6 Runtime Analysis

Finally, we compare the runtime characteristics of the ViT and the denoising diffusion model. The latter is computationally intensive, particularly during inference, where a complete U-Net pass is required for each sampling step. In contrast, our ViT performs inference with only a single forward pass. Although we substantially accelerated inference in the denoising diffusion model using the DDIM sampling procedure, it remains significantly slower than our ViT, as illustrated in Fig. 11.

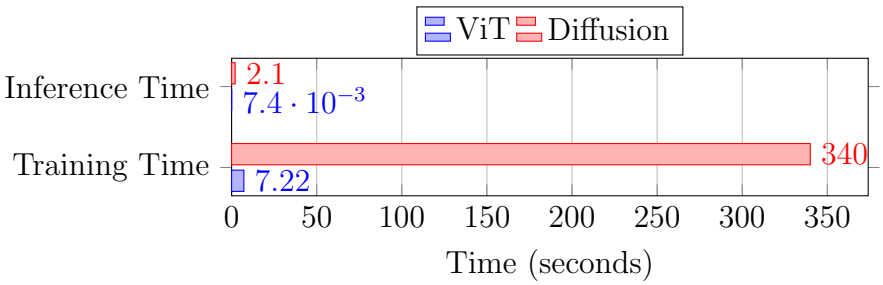

Figure 11: Comparison of training times per epoch (2250 images) and inference times for one super-resolved sample for ViT and denoising diffusion model.

A similar pattern is observed during training: we must optimize our denoising diffusion model across all timesteps of the noise schedule, resulting in higher computational cost per epoch and generally requiring more epochs to achieve performance comparable to that of our ViT. Nevertheless, in our application, predictive accuracy is prioritized over runtime efficiency, since inference is conducted offline and real-time performance is not a constraint.

5 Discussion and Future Work

Despite the strong performance of our models across multiple experimental settings, several challenges remain: Firstly, the denoising diffusion model entails considerable computational demands, both during training and inference. Although DDIM sampling significantly accelerates inference, identifying more efficient training strategies warrants further research. In particular, latent diffusion models offer a promising direction for reducing complexity and preliminary small-scale experiments have shown promising results. However, comprehensive evaluation lies beyond the scope of this work.
Secondly, our assessment of temporal transfer revealed limitations in capturing long-term climate variability, constrained by the relatively short five-year span of the available remote sensing data. It remains to be determined whether incorporating additional conditional variables, such as surface temperature, could mitigate this issue and enhance the models' temporal generalization capabilities.
Thirdly, while transfer learning has proven effective in enabling the ViT to adapt to previously unseen spatial regions, it is yet to be established whether similar benefits extend to the denoising diffusion model.
Lastly, our study has so far focused exclusively on the Nunavik region in Canada. The scalability of the proposed methods to larger datasets encompassing diverse global regions remains unexplored. Broader geographic coverage could potentially improve spatial generalization and further validate the practical applicability of these approaches.

6 Conclusion

In this work, we developed a ViT architecture and a denoising diffusion model for super-resolution in microwave remote sensing. Our ViT implementation serves as a solid, lightweight baseline, offering high computational efficiency during both training and inference. In contrast, our denoising diffusion model, while computationally more demanding, consistently achieves superior super-resolved image quality.
Both models demonstrate the ability to perform temporal transfer, successfully super-resolving images from previously unseen time periods. However, although temporal transfer can be achieved, the performance drops compared to seen time periods, indicating that temporal generalization remains a challenge. In terms of spatial transfer to entirely new regions, neither model achieves performance levels sufficient for immediate real-world deployment. To address this limitation, we introduced a post-processing fine-tuning procedure for the ViT that yields strong results on unseen regions with only minimal additional training effort.

Finally, we investigated the simultaneous prediction of VV and VH polarizations. While our ViT model struggles with this dual-channel prediction task, the denoising diffusion model maintains performance comparable to single-channel VV prediction. To our knowledge, this is the first system capable of jointly predicting both polarizations within a super-resolution framework, representing a significant contribution that enables more compact model architectures and substantially reduces computational costs for multi-polarization applications.

References

[1] Edwin T Engman. Applications of microwave remote sensing of soil moisture for water resources and agriculture. *Remote sensing of environment*, 35(2-3):213–226, 1991.

[2] Kamal Nasrollahi and Thomas B Moeslund. Super-resolution: a comprehensive survey. *Machine vision and applications*, 25:1423–1468, 2014.

[3] Yueli Chen and Ralf Ludwig. Exploring the merging potential of high temporal resolution and high spatial resolution microwave remote sensing data. EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023, 2023. EGU23-8999.

[4] Fawwaz T Ulaby, Richard K Moore, and Adrian K Fung. *Microwave Remote Sensing: Active and Passive. Volume I: Microwave Remote Sensing Fundamentals and Radiometry*. Artech House, reprint edition, 2014.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[6] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[8] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017.

[9] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[10] Charis Lanaras. *Enhancing the Spectral and Spatial Resolution of Remote Sensing Images*, volume 122. ETH Zurich, 2018.

[11] Yunliang Qi, Meng Lou, Yimin Liu, Lu Li, Zhen Yang, and Wen Nie. Advancing image super-resolution techniques in remote sensing: A comprehensive survey. *arXiv preprint arXiv:2505.23248*, 2025.

[12] Michel Allard, Michel Lemay, et al. Nunavik: A regional perspective of permafrost grounded in field observations and climate change projections. *Geological Survey of Canada, Ottawa*, 5480(2007):1–20, 2007.

[13] S. K. Chan, P. E. O'Neill, T. J. Jackson, E. G. Njoku, and R. Bindlish. Development and assessment of the smap enhanced passive soil moisture product. *Remote Sensing of Environment*, 204:243–256, 2018.

[14] Li Zhu, Bingfang Wu, and Shang Huang. Polarization decomposition of sentinel-1 sar data for monitoring vegetation and soil moisture dynamics in cold regions. *Remote Sensing of Environment*, 259:112418, 2021.

[15] R. Touzi. Sar polarization signatures: Interpretation of polarimetric sar measurements over natural targets. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12):3807–3821, 2007.

[16] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[18] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross B. Girshick. Early convolutions help transformers see better. *CoRR*, abs/2106.14881, 2021.

[19] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016.

[20] Yuxin Wu and Kaiming He. Group normalization. *CoRR*, abs/1803.08494, 2018.

[21] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[22] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[23] Axi Niu, Kang Zhang, Trung X. Pham, Jinqiu Sun, Yu Zhu, In So Kweon, and Yanning Zhang. Cdpmsr: Conditional diffusion probabilistic models for single image super-resolution, 2023.

[24] Yiyang Ma, Huan Yang, Wenhan Yang, Jianlong Fu, and Jiaying Liu. Solving diffusion ODEs with optimal boundary conditions for better image super-resolution. In *The Twelfth International Conference on Learning Representations*, 2024.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[26] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

[27] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep learning library.* Curran Associates Inc., Red Hook, NY, USA, 2019.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[30] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Appendix

A Vision Transformer

In this section, we supplement detailed information about model structure based on Section 3.1 and include additional experiment results obtained.

A.1 Model Architecture

A.1.1 Self-Attention and Transformer Blocks

The core of our ViT model is a stack of transformer encoder blocks that operate on patch embeddings. Each encoder block employs multi-head self-attention to model spatial dependencies between patches, allowing the network to aggregate global context across the image. Given an input sequence of patch embeddings, each attention head computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{22}$$

where $Q, K, V \in \mathbb{R}^{N\times d_k}$ are the query, key, and value matrices, N is the number of patches, and d_k is the dimension of each head [17]. In our implementation, the model uses 16 attention heads and embedding dimension $D = 1024$, resulting in $d_k = 64$ per head.

Each self-attention layer is followed by a feedforward MLP with a hidden size of 2048 and GELU activations [19]. Unlike standard ViT models that employ LayerNorm, we adopt GroupNorm in both the patch embedding and transformer blocks. This improves stability during training with small batch sizes, which is typical in high-resolution remote sensing data. The encoder is composed of 12 such transformer layers.

By capturing non-local dependencies across the entire spatial domain, the encoder learns global structural patterns essential for accurate super-resolution.

A.1.2 Positional Encoding

To retain spatial structure in the transformer architecture, we incorporate two-dimensional positional encodings $\text{PE} \in \mathbb{R}^{1\times N\times D}$ into the patch embeddings, where $N = H \cdot W$ is the total number of patches and D is the embedding dimension. We support both fixed sinusoidal encodings and learnable encodings.

For the fixed encoding, we implement a 2D extension of the sinusoidal encoding proposed in [17], with a temperature scaling factor $\tau = 0.5$. Given a patch grid of size $H \times W$, we first generate spatial coordinate matrices $x \in [0, W-1]$ and $y \in [0, H-1]$ using meshgrid indexing. A frequency vector $\omega \in \mathbb{R}^{D/4}$ is constructed as:

$$\omega_i = \frac{1}{\tau^{i/(D/4-1)}}, \quad i = 0, 1, \ldots, \frac{D}{4} - 1. \tag{23}$$

The encoding is then computed by applying sine and cosine functions to both x and y:

$$\text{PE}_x = [\sin(\omega \cdot x), \cos(\omega \cdot x)], \tag{24}$$

$$\text{PE}_y = [\sin(\omega \cdot y), \cos(\omega \cdot y)], \tag{25}$$

and concatenated to form the final embedding for each position:

$$\mathrm{PE}_{(x,y)} = \mathrm{concat}(\mathrm{PE}_x, \mathrm{PE}_y) \in \mathbb{R}^D. \tag{26}$$

A.2 Generalization across Time

Here, we include two super-resolved samples from our temporal transfer setup. Both visually and in terms of quantitative metrics, the reconstructed images from 2018 consistently outperform those from 2021. The 2018 samples exhibit higher PSNR, SSIM, and SCC values, and show closer visual resemblance to the ground truth, with better preservation of texture and structural details. This observation aligns well with the analysis in Section 4.4.

A.3 Generalization across Space

In Fig. 14 and Fig. 15, the upper row shows the reconstruction results before fine-tuning, while the lower row presents the results after fine-tuning. The Nu621 region serves as an unfeasible dataset that was not seen during the initial training, whereas the Nu912 region is included as a control group representing regions already involved in training. By comparing these results, we observe that before fine-tuning, the ViT model is able to reconstruct only very limited spatial features in the unseen Nu621 region. After fine-tuning, however, the model can recover much finer spatial details in this region. Both quantitatively and visually, fine-tuning substantially improves the reconstruction quality for the previously unfeasible region, while maintaining high image quality in regions already used for training.

A.4 Dual-Channel Reconstruction

Fig. 16 presents results from two different regions using the dual-channel output ViT model. Both quantitatively and visually, the reconstructed images show a lower degree of fidelity to the ground truth compared to those produced by the one-channel model.

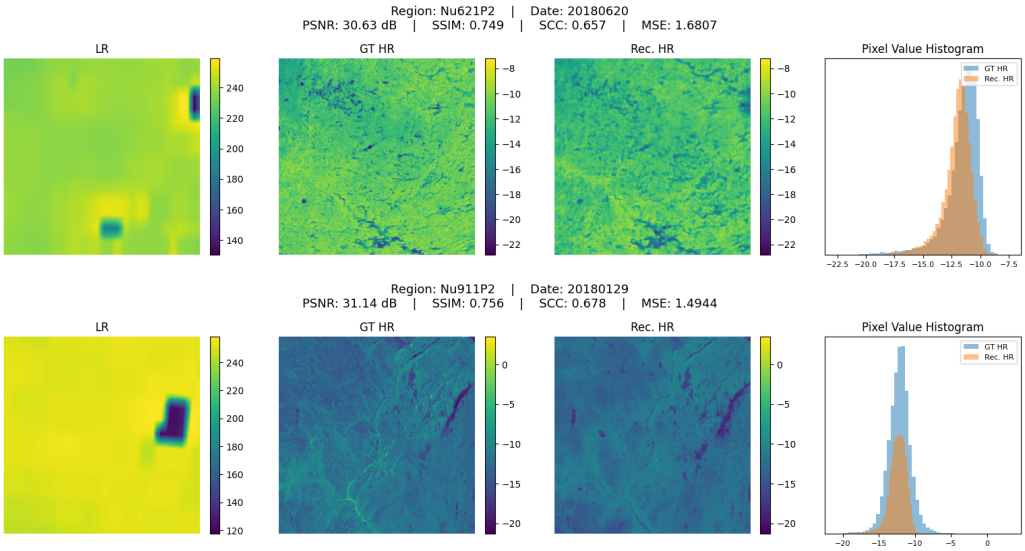

Figure 12: One sample from temporal transfer inference from 2018 from the ViT model.

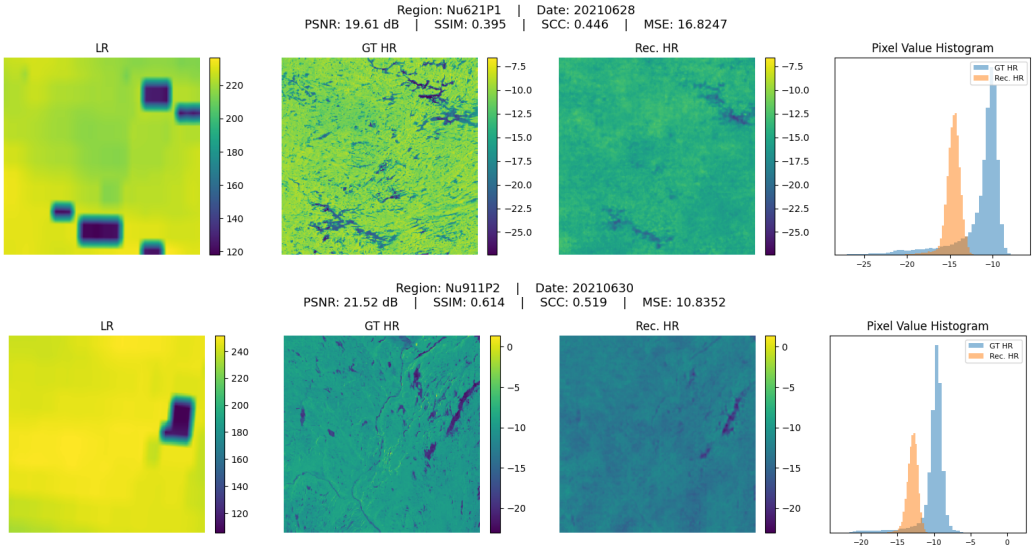

Figure 13: One sample from temporal transfer inference from 2021 from the ViT model.

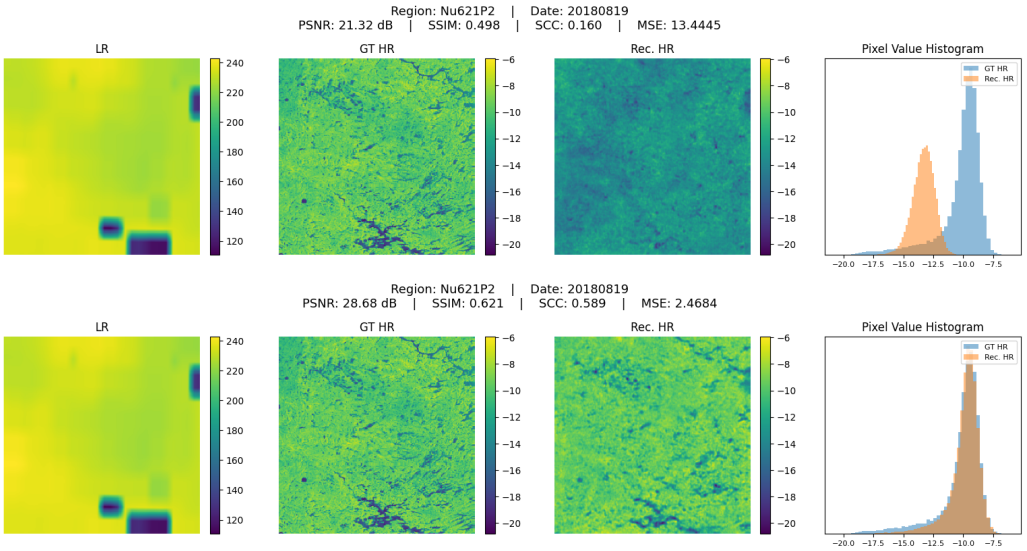

Figure 14: One sample from spatial transfer inference from the region Nu621 before and after fine-tuning.

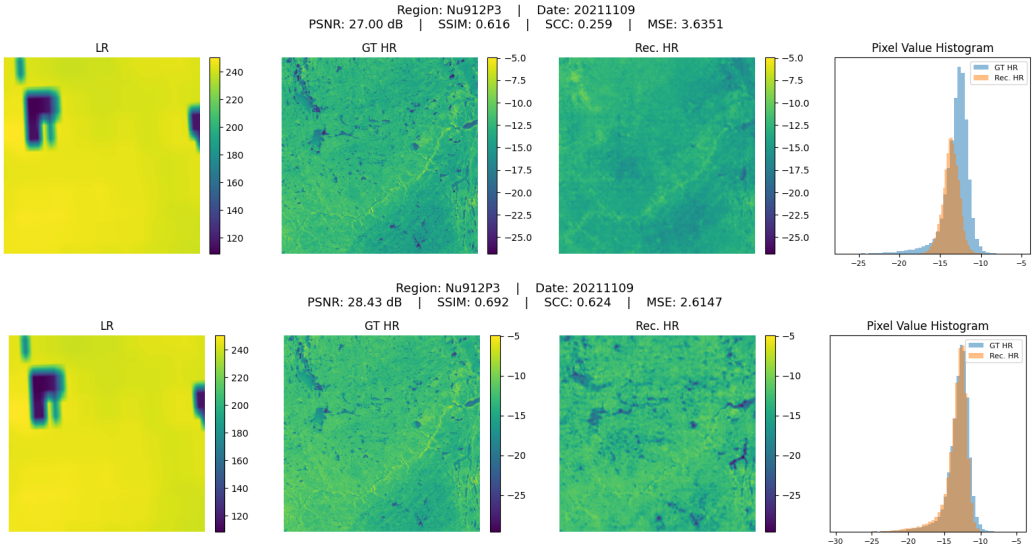

Figure 15: One sample from spatial transfer inference from the region Nu912 before and after fine-tuning.

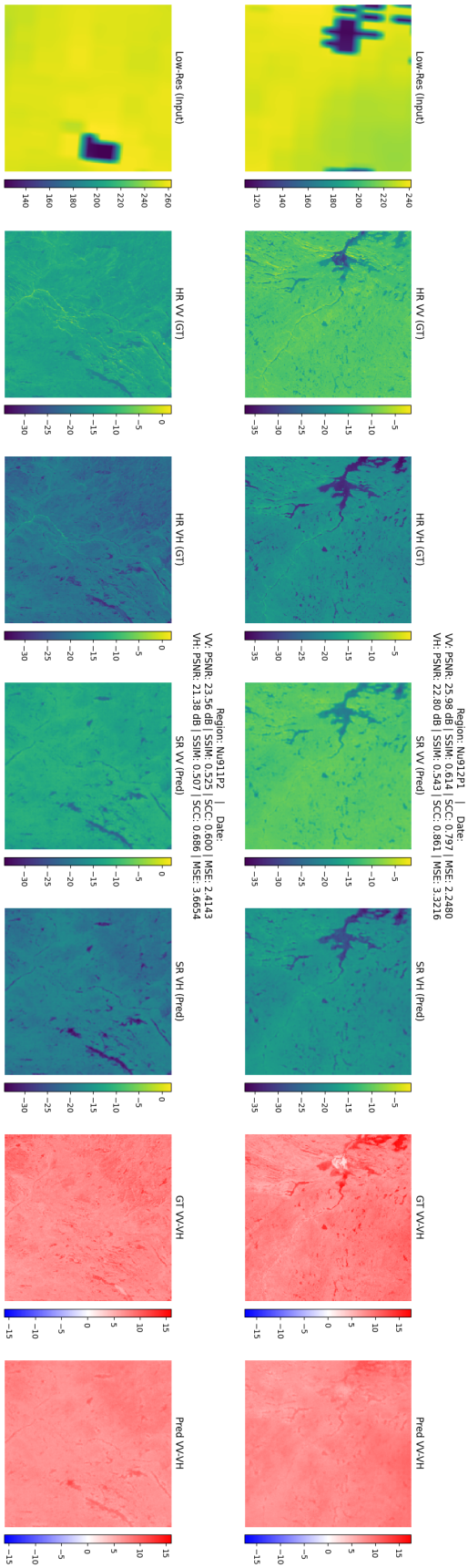

Figure 16: Samples generated by the dual-channel reconstruction ViT model from a randomly shuffled validation set. Two samples including VV-VH differences for ground truth and prediction.

B Denoising Diffusion Model

In this section, we include additional experiments and results obtained while developing our denoising diffusion model.

B.1 Baseline

To highlight the effectiveness of our proposed training configuration, we present samples generated by the model trained under the baseline conditions, without data augmentation, dropout, or selective conditioning, as shown in Fig. 17. Compared to the results in Fig. 10, which reflect our best-performing setup, the baseline outputs exhibit substantially lower quality, appearing notably blurrier and lacking spatial consistency.

B.2 Generalization across Time

As mentioned when presenting our results, we evaluated our denoising diffusion models capabilities to super-resolve LR images from unseen time periods. In Fig. 18, we show generated samples from 2018 from our validation dataset. In Fig. 19, we repeat the experiment for validation samples from 2021. Overall, the performance aligns with our expectations and is adequate for super-resolving data from periods such as 2022, during which HR images are unavailable due to satellite failure.

B.3 Generalization across Space

Analogous to the temporal transfer evaluation, we assessed the model's performance on entirely unseen regions. Representative samples generated for the Nu621 region, which serves as the validation set in this context, are presented in Fig. 20. Unlike in the temporal transfer scenario, these samples are unsuitable for any practical real-world application. Although the value range appears to be correctly estimated, which we attribute to the conditional information, the outputs exhibit pronounced over-smoothing and lack discernible spatial structure.

B.4 Dual-Channel Reconstruction

As mentioned in our results, we evaluated our denoising diffusion model's capabilities to super-resolve LR images and reconstruct the VV and VH channels. Fig. 21, Fig. 22, and Fig. 23 show that our model was able to effectively reconstruct spatial and structural patterns present in the ground truth. The generated outputs demonstrate high visual fidelity, recovering fine textures and preserving important features, even in challenging areas, and that shows the model's strong generalization ability.

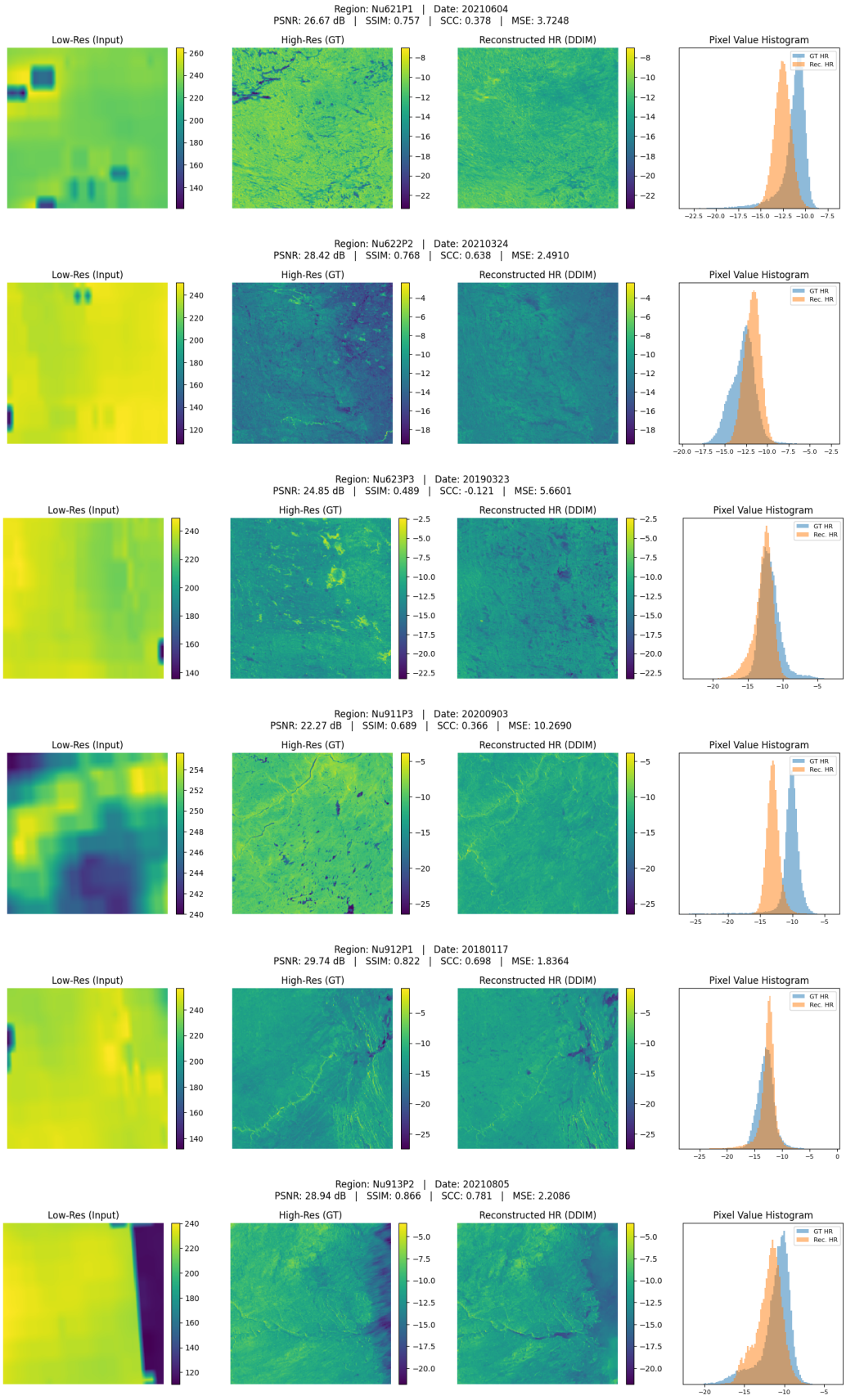

Figure 17: Samples generated by the Denoising Diffusion Model from a randomly shuffled validation set learned with our baseline training setup. One sample per region is shown.

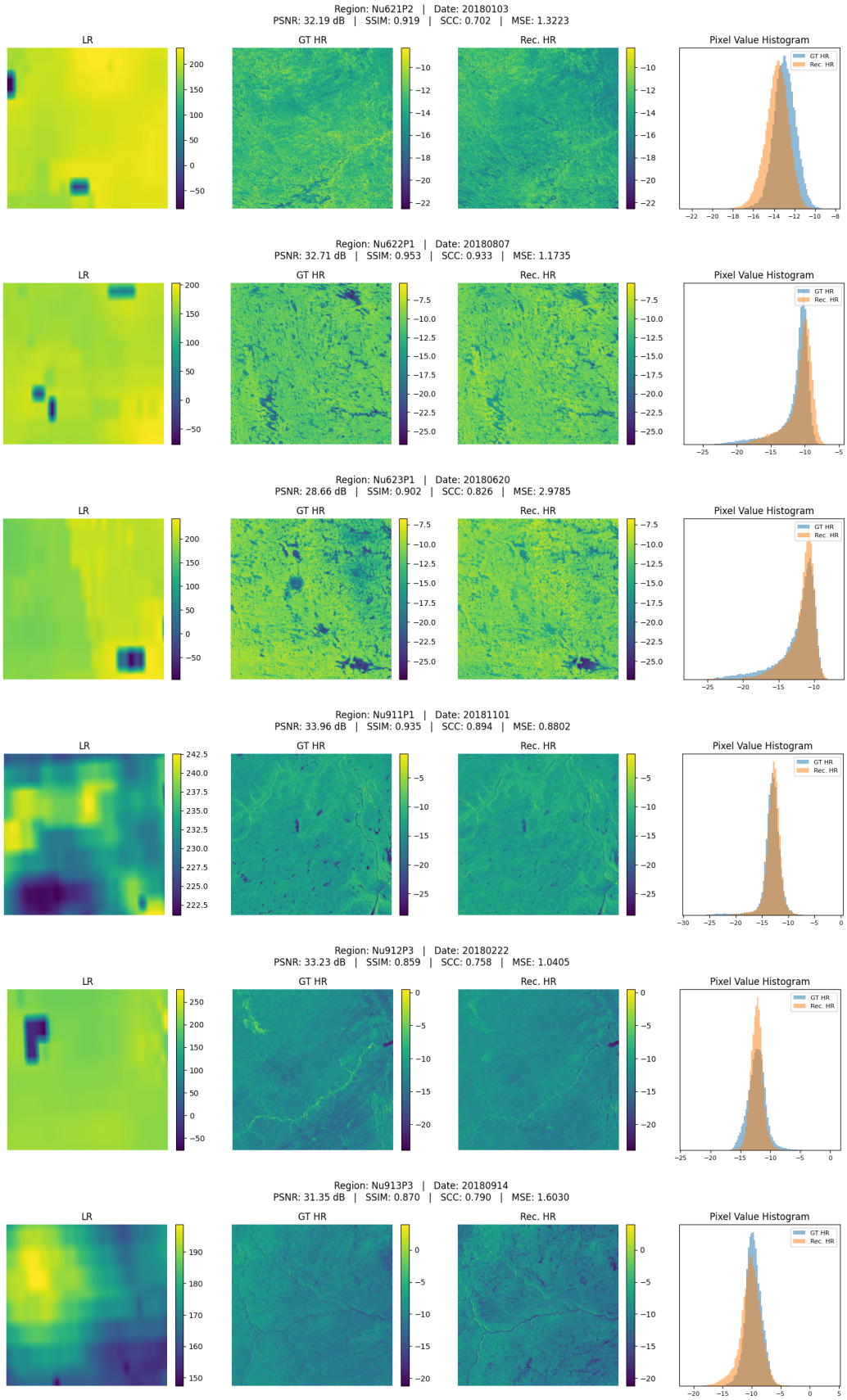

Figure 18: Samples generated by the denoising diffusion model for temporal transfer validated on data from 2018. One sample per region is shown.

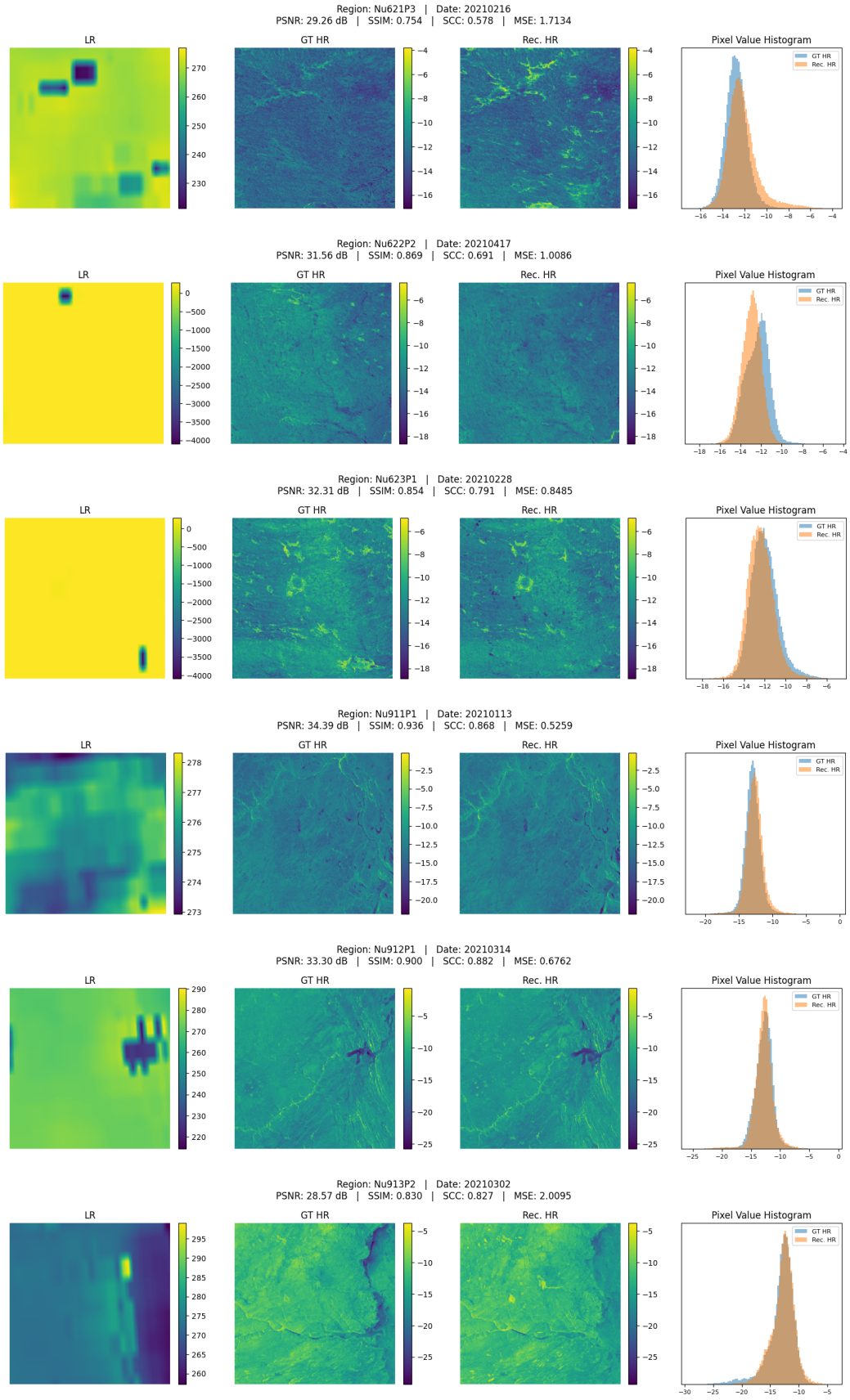

Figure 19: Samples generated by the denoising diffusion model for temporal transfer validated on data from 2021. One sample per region is shown.

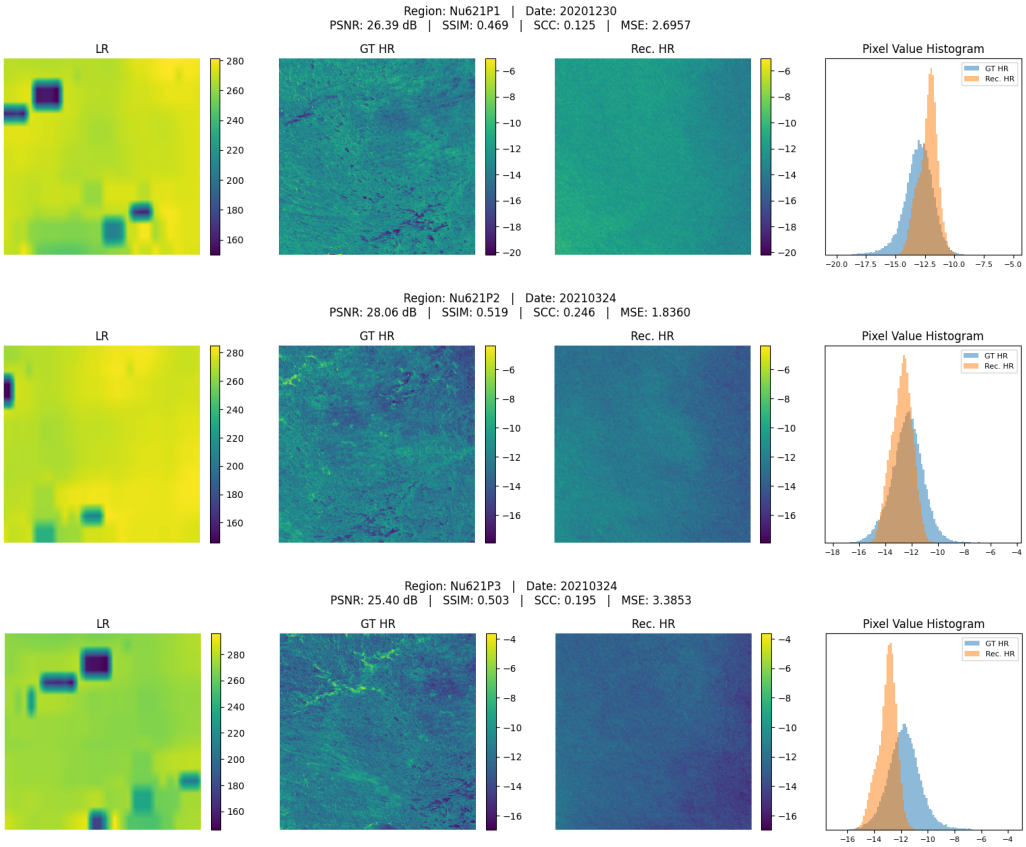

Figure 20: Samples generated by the denoising diffusion model for spatial transfer validated on region Nu621. One sample per subregion of Nu621 is shown.

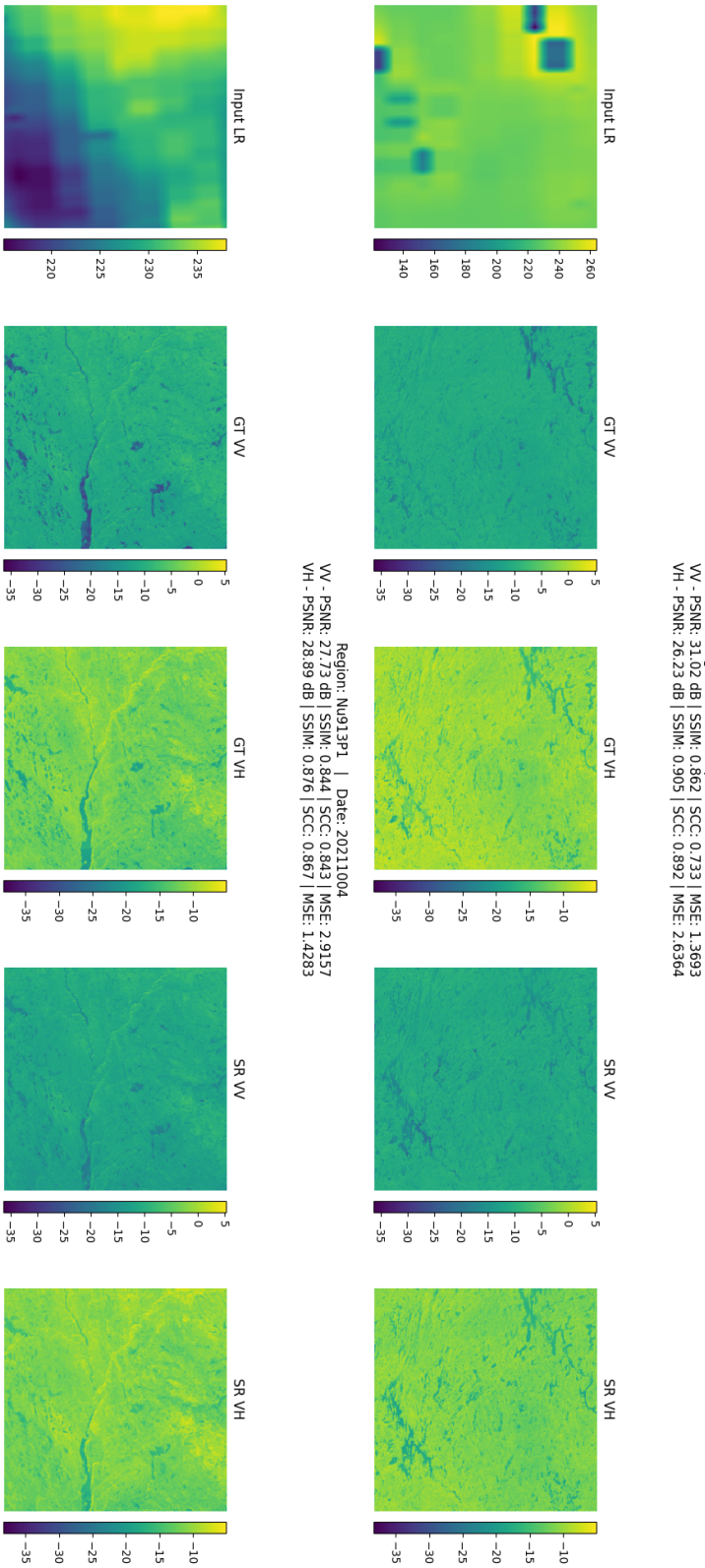

Figure 21: Samples generated by the dual-channel reconstruction denoising diffusion model from a randomly shuffled validation set. Two samples including ground truth and predictions of both polarizations.

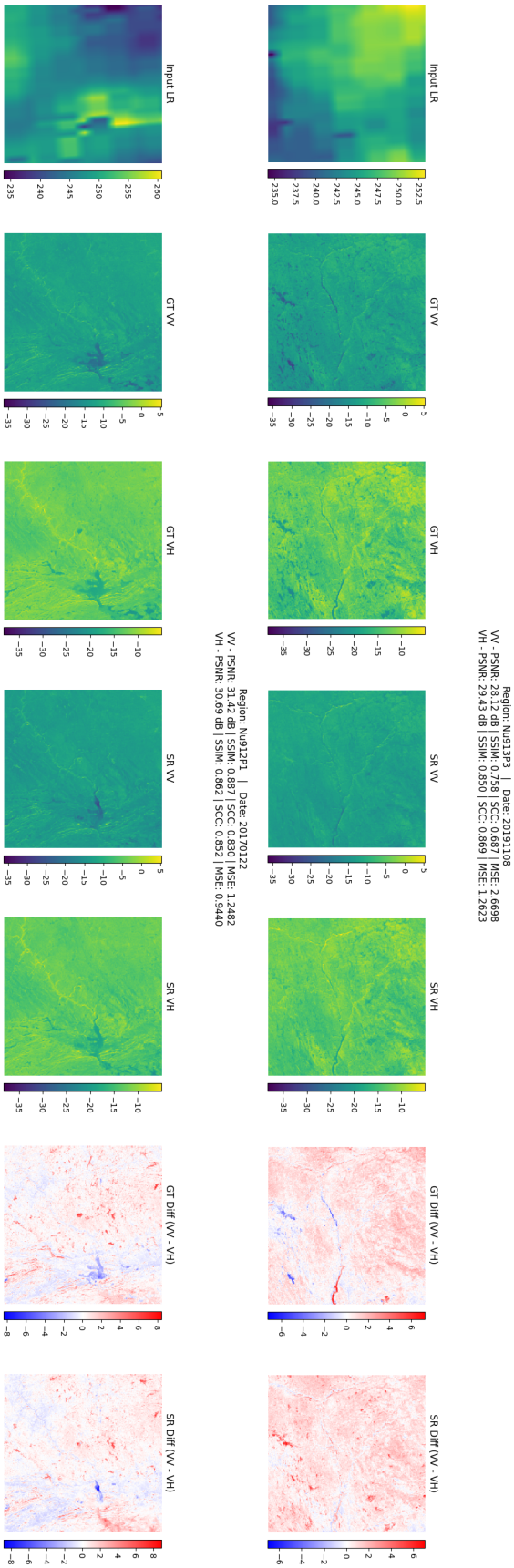

Figure 22: Samples generated by the dual-channel reconstruction denoising diffusion model from a randomly shuffled validation set. Two samples including VV-VH differences for ground truth and prediction.

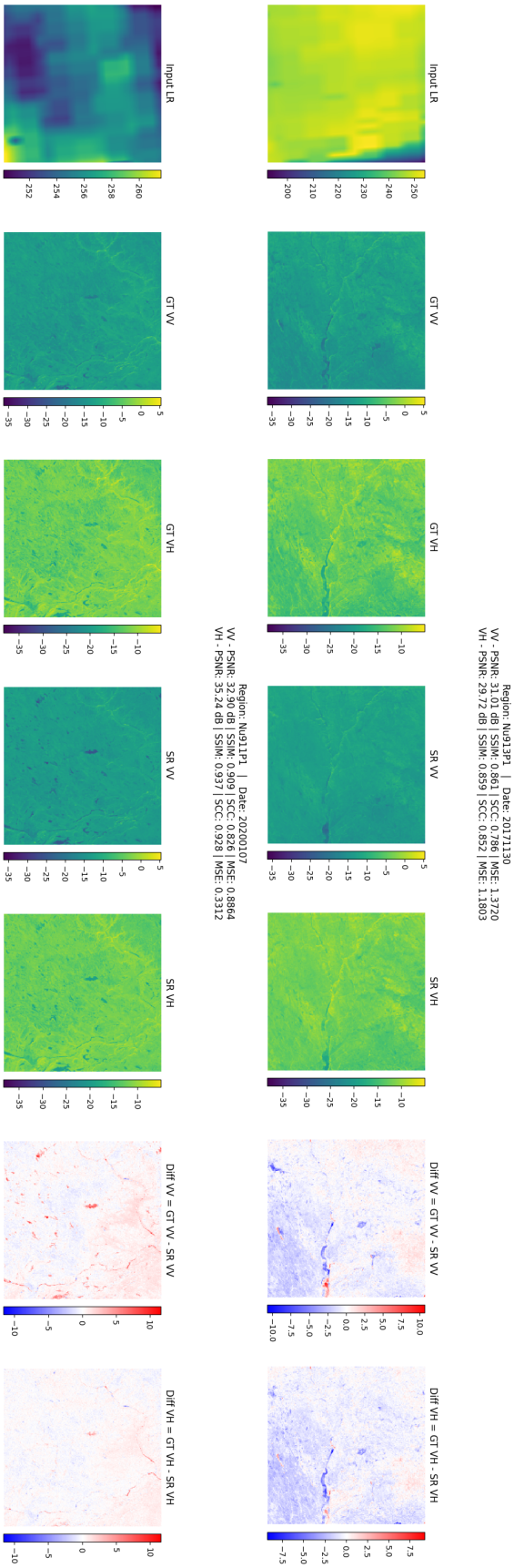

Figure 23: Samples generated by the dual-channel reconstruction denoising diffusion model from a randomly shuffled validation set. Two samples including polarization-ground truth differences.