Universität Zürich[UZH]

Geographisches Institut

Cloud-Aware Generative Reconstruction of Satellite Thermal Observations Using Synthetic Cloud Data

GEO 511 Master's Thesis

Author: Yaze Xue, 23-741-523

Supervised by: Dr. Jennifer Susan Adams, Franz Mauthner (f.mauthner@aee.at), Max Aragon (max@termatics.com), Carla Arellano (carla@termatics.com)

Faculty representative: Prof. Dr. Ross Purves

04.03.2026

University of Zurich[UZH]

Faculty of Geography

Cloud-Aware Generative Reconstruction of Satellite Thermal Observations Using Synthetic Cloud Data

GEO 511 Master’s Thesis

Supervisor:

Faculty Representative:

Dr Ross Stuart

University of Zurich (UZH):

Dr. Ross Stuart Purves; Dr. Jennifer Susan Adams

External (AEE INTEC, Termatics):

Franz Mauthner; Max Aragon; Carla Arellano

Submitted on March 4, 2026 by:

Yaze Xue

Matriculation number: 23-741-523

Acknowledgments

This thesis was conducted at the University of Zurich (UZH), Department of Geography, in close collaboration with the Austrian research institute AEE – Institute for Sustainable Technologies (AEE INTEC) and the industry partner Termatics, a company specializing in thermal remote sensing and advanced thermal analytics.

I am grateful to my university supervisors, Prof. Dr. Ross Stuart Purves and Dr. Jennifer Susan Adams, for their academic oversight and for the critical feedback provided throughout the development of this research. Their insights into geographic methodology were fundamental to the structure of this work.

I would also like to thank Franz Mauthner from AEE INTEC, and Max Aragon and Carla Arellano from Termatics, for their technical cooperation. Their assistance with the data infrastructure and their perspectives on the practical requirements of the sector were highly valued.

This work was conducted within the framework of the research project "TopView – Efficient Use of Remote Sensing Data for Climate Adaptation and Energy Planning", which is funded under the Austrian TIKS program for Energy and Environmental Technologies (FFG Project Number: FO999913467). From January 2025 to December 2026, the project supports the development of methodologies to integrate remote sensing data into urban climate adaptation strategies.

Abstract

Land surface temperature (LST) is a fundamental parameter in urban environmental science, as it directly governs the surface energy budget and the partitioning of sensible and latent heat fluxes. Therefore, it is a critical variable for urban climate modeling, energy planning, and the assessment of surface urban heat island (SUHI) intensity. However, thermal infrared satellite observations are frequently affected by cloud cover, resulting in spatially incomplete temperature fields and reduced temporal consistency. This limitation poses a significant challenge for applications that require spatially continuous temperature information.

This thesis investigates different reconstruction strategies for recovering cloud-contaminated thermal infrared satellite observations, with a particular focus on complex urban environments. A controlled evaluation framework is adopted using synthetically generated cloud masks, which enables reproducible comparison between reconstruction methods under varying degrees of information loss. Three categories of approaches are examined: classical geostatistical interpolation methods, a deterministic convolutional neural network based on the U-Net architecture, and a probabilistic diffusion-based generative model.

The results show that classical interpolation methods provide competitive local accuracy under limited cloud coverage but degrade when cloud-contamination becomes large and contiguous. The U-Net model improves global mean consistency but exhibits reduced robustness under dense cloud conditions. In contrast, the diffusion-based approach achieves the lowest cloud-only thermal Root Mean Square Error (RMSE) and the highest spatial correlation under thick cloud cover. In general, the findings highlight the advantages of generative models for reconstructing thermal infrared images under severe information loss.

Spatially continuous thermal fields provide a robust basis for climate-resilient urban energy planning within GIS-based decision frameworks. When integrated with multi-source geospatial datasets such as building inventories, land-use maps, demographic statistics, and urban green infrastructure layers, the reconstructed temperature fields enable the identification of neighbourhood-scale cooling demand peaks and thermally vulnerable areas, supporting more targeted and data-driven urban climate adaptation strategies. While the proposed framework demonstrates clear advantages under dense cloud conditions, its performance depends on the availability of representative training data and may be limited in regions with markedly different climatic or urban characteristics. Nevertheless, the results highlight the potential of generative reconstruction methods to improve the usability of thermal remote sensing data for urban climate analysis and energy planning.

Keywords: Cloud Removal, Deep Learning, U-Net, Diffusion Models, Urban Climate, Thermal remote sensing, GIS, Urban planning.

Table of Contents

Abbreviations

AdamW	Adaptive Moment Estimation with Weight Decay
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DDPM	Denoising Diffusion Probabilistic Model
DEM	Digital Elevation Model
EMA	Exponential Moving Average
Fmask	Function of Mask Cloud Detection Algorithm
GPU	Graphics Processing Unit
HLS	Harmonized Landsat and Sentinel-2
IDW	Inverse Distance Weighting
K	Kelvin (temperature unit)
LST	Land Surface Temperature
LULC	Land Use and Land Cover
MBE	Mean Bias Error
MSE	Mean Squared Error
MSI	MultiSpectral Instrument (Sentinel-2)
OK	Ordinary Kriging
OLI	Operational Land Imager (Landsat-8)
PSNR	Peak Signal-to-Noise Ratio
ReLU	Rectified Linear Unit
RGB	Red Green Blue
RMSE	Root Mean Squared Error
SCG	Satellite Cloud Generator
SSIM	Structural Similarity Index Measure
SUHI	Surface Urban Heat Island
SWIR	Short-Wave Infrared
TIR	Thermal Infrared
TIRS	Thermal Infrared Sensor
TOA	Top of Atmosphere
UHI	Urban Heat Island
USGS	United States Geological Survey
VNIR	Visible and Near-Infrared

List of Figures

Chapter 1

Introduction

1.1 Introduction

This research is conducted within the framework of the Austrian research project *TopView* (FFG-Project Number: FO999913467), which aims to support spatial energy planning and climate adaptation through the application of advanced remote sensing techniques (Austrian Research Promotion Agency (FFG), 2024). A central source of information in this context is thermal infrared satellite imagery, which provides spatially resolved insight into surface thermal conditions and radiative characteristics of the Earth's surface (Li et al., 2013). Thermal infrared observations have become an essential input for urban-scale analyses, including the investigation of surface urban heat island (SUHI) dynamics, the characterization of urban structure, and the assessment of heating demand and renewable energy potential (Voogt and Oke, 2003). High-resolution thermal imagery over cities such as Zurich, for example, reveals pronounced spatial temperature contrasts between densely built-up areas and vegetated zones, highlighting the importance of accurate and spatially consistent thermal information for urban environmental analysis.

Despite its relevance, the operational use of satellite-based thermal infrared observations is frequently constrained by cloud cover. Because thermal infrared radiation cannot penetrate clouds, cloud-contaminated imagery often contains large spatial gaps and exhibits reduced radiometric reliability, limiting the availability of spatially and temporally consistent thermal datasets (Li et al., 2013). This limitation poses a significant challenge for long-term environmental monitoring, urban climate research, and energy-related applications, all of which rely on continuous and reliable thermal surface information. Consequently, there is a clear need for robust reconstruction approaches capable of generating spatially continuous and radiometrically consistent thermal infrared datasets from incomplete observations.

1.2 Related Work

A broad range of strategies has been proposed to compensate for cloud-induced data loss in thermal infrared imagery. These strategies differ primarily in the type of information they exploit to estimate missing values and in the assumptions they make regarding data continuity and spatial structure. Some approaches rely on temporal continuity at the pixel level and infer missing observations from historical time series or complementary satellite overpasses (Crosson et al., 2012). Other methods focus on spatial relationships within individual scenes, estimating missing values based on neighboring pixels and auxiliary surface characteristics (Stathopoulou and Cartalis, 2009a). More advanced frameworks attempt to combine spatial and temporal information or to integrate auxiliary inputs derived from additional data sources.

Despite their methodological diversity, many existing approaches share common limitations. Their performance often deteriorates under persistent or dense cloud cover, they depend strongly on the availability and quality of auxiliary data, or they rely on assumptions that may not hold across different seasons and heterogeneous urban landscapes (Li et al., 2013). These constraints limit the applicability of traditional reconstruction methods in operational settings where data availability is highly variable.

Recent progress in machine learning, and in particular deep learning, has opened new possibilities for addressing these limitations. Deep learning models are capable of learning complex, non-linear relationships directly from data, thereby reducing the reliance on predefined statistical assumptions or handcrafted interpolation rules (LeCun et al., 2015). Within this context, probabilistic models such as Markov Random Fields have been used to explicitly model spatial and temporal dependencies, while recurrent neural network architectures, including Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks, have been applied to capture temporal dynamics in multi-temporal LST observations (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997). Convolutional Neural Networks (CNNs), which are particularly effective at extracting hierarchical spatial features, have also been widely adopted for cloud-related tasks in remote sensing imagery (Jeppesen et al., 2019; Zhang et al., 2020). Among CNN-based architectures, encoder–decoder models with skip connections have proven especially effective for pixel-level prediction tasks that require both contextual understanding and precise spatial localization. U-Net-based models, in particular, have been successfully applied to cloud and cloud-shadow detection in satellite imagery, demonstrating strong performance across different sensors and spectral configurations (Ronneberger et al., 2015). Their ability to preserve fine-scale spatial information while integrating broader contextual features makes them well suited for tasks involving spatially heterogeneous patterns, such as cloud-contaminated reconstruction.

Nevertheless, cloud detection and removal remain challenging due to the highly variable and stochastic nature of cloud distribution. In urban areas, this task is further complicated by the diverse reflectance of built environments, which can be spectrally indistinguishable from clouds. Such ambiguities pose significant risks to Urban Heat Island (UHI) quantification, as any misclassification directly propagates errors into LST retrieval. These persistent challenges in accurately separating and reconstructing urban surface signals necessitate more robust generative frameworks capable of capturing complex spatial dependencies.

More recently, diffusion-based generative models have emerged as an alternative framework for image restoration and reconstruction tasks. These models formulate data generation as an iterative denoising process, in which a neural network learns to reverse a gradual noise corruption procedure (Ho et al., 2020). Compared to earlier generative approaches, diffusion models have demonstrated stable training behavior and strong robustness in a variety of image synthesis and restoration applications (Song and Ermon, 2021). When applied to LST reconstruction, cloud-covered or unreliable pixels can be naturally interpreted as an image inpainting problem, where missing regions are reconstructed based on the surrounding spatial context (Bertalmio et al., 2000).

1.3 Research Objectives

Building on the methodological considerations outlined above, this thesis investigates the potential of U-Net–based models and diffusion models for the reconstruction of cloud-free thermal infrared satellite imagery, with a specific focus on complex urban environments. The overarching goal is to develop reconstruction approaches that remain robust under dense and persistent cloud cover, thereby enabling reliable assessment of urban heat patterns and supporting climate-resilient urban and energy infrastructure planning.

To this end, ten representative large European cities, including Munich, Vienna, and Paris, are selected as study areas. These cities cover a range of climatic conditions, urban morphologies, and land cover characteristics, enabling a comprehensive evaluation of model performance across diverse urban settings.

The specific research objectives are as follows:

1. To establish a controlled training and evaluation framework by generating synthetically cloud-contaminated thermal infrared datasets under varying cloud coverage levels and spatial configurations.

2. To benchmark deep learning–based reconstruction approaches against traditional methods, such as Inverse Distance Weighting (IDW) and Ordinary Kriging (OK), with a focus on reconstruction accuracy and spatial consistency.

3. To investigate the capability of U-Net architectures to remove cloud-induced information loss and to predict missing thermal infrared observations while preserving fine-scale urban thermal structures.

4. To evaluate the effectiveness of diffusion models for reconstructing thermal infrared imagery from cloud-contaminated inputs and to assess their robustness.

Addressing these objectives aims to strengthen the spatial data basis for integrated urban planning by generating spatially continuous thermal observations. The reconstructed temperature fields enable their integration into GIS-based analytical workflows and decision-support tools for urban energy planning and climate adaptation, and the implications and potential application fields of the proposed framework are discussed explicitly in this thesis.

Chapter 2

Methodology

This study follows a four-stage processing framework to reconstruct spatially continuous thermal imagery from satellite observations affected by cloud cover, as summarized in Figure 2.1. Stage 1 focuses on data preparation and cloud identification using harmonized Landsat–Sentinel (HLS) products, which provide radiometrically consistent multispectral and thermal observations across sensors(NASA Goddard Space Flight Center, 2026).

Cloud and cloud-shadow pixels are identified using the Fmask quality layer in combination with spectral characteristics and thermal anomaly information, after which all scenes are standardized into spatially consistent image patches together with corresponding cloud masks and near cloud-free reference samples. In Stage 2, a physically consistent synthetic cloud generation procedure is applied to address the limited availability of truly cloud-free observations; cloud fields with realistic spatial structure and controlled coverage are imposed consistently on optical reflectance and thermal radiance, resulting in paired clean and cloud-contaminated datasets suitable for supervised learning and controlled evaluation. Stage 3 uses these paired samples to train and compare different reconstruction strategies, including classical geostatistical interpolation methods, namely inverse distance weighting (IDW) and ordinary kriging (OK), as well as learning-based approaches based on a U-Net regression model and a diffusion-based generative model that formulates cloud removal as a conditional image inpainting problem. Finally, Stage 4 evaluates reconstruction performance under both synthetic and real cloud conditions using pixel-level and region-aggregated metrics, with emphasis on reconstruction accuracy, spatial consistency, and physical plausibility of the resulting cloud-free products for downstream urban climate applications.

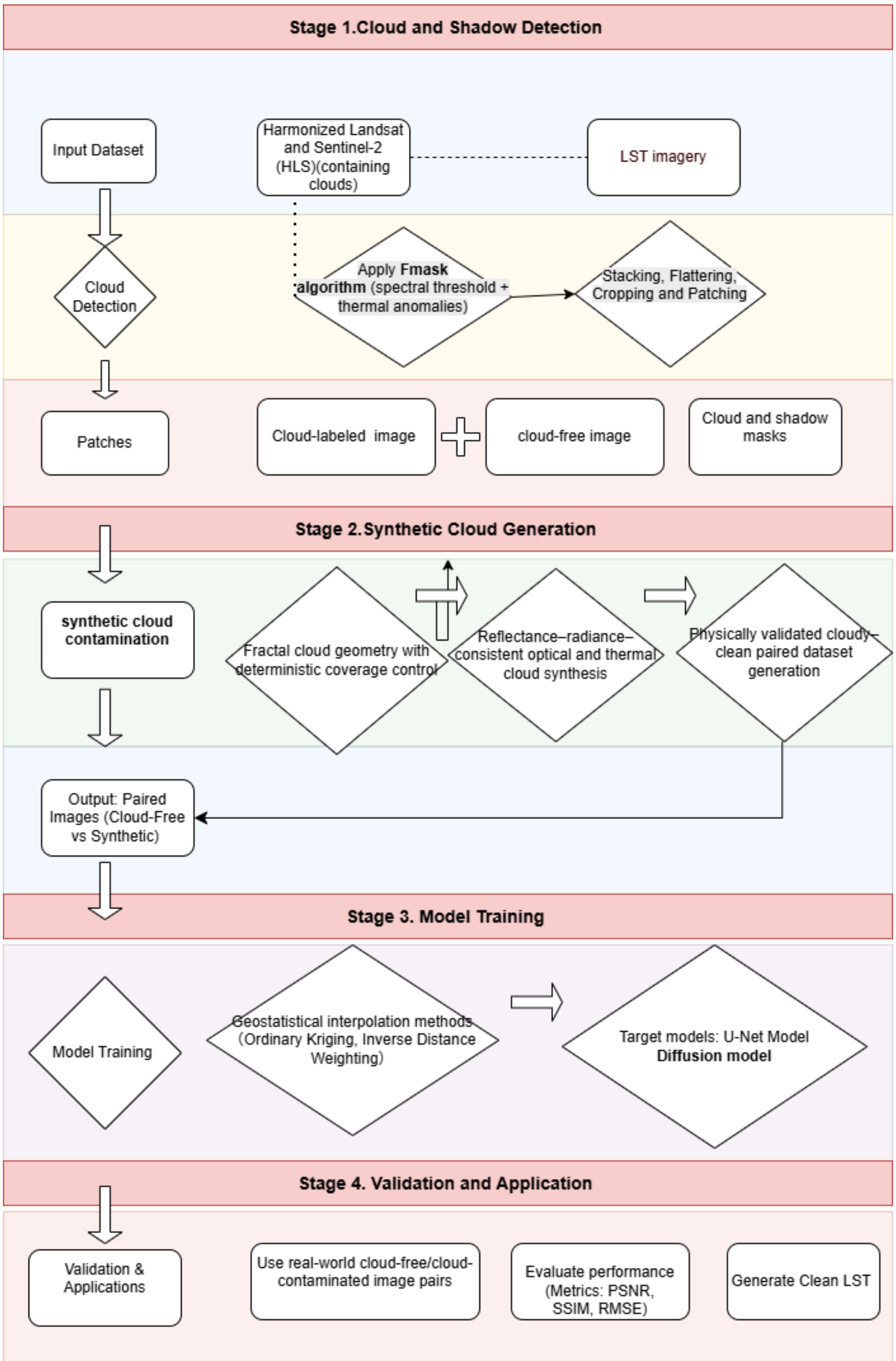


Figure 2.1: Overview of the methodological pipeline.

2.1 Dataset

2.1.1 Study area and data sources

The study area comprises a set of representative European cities, including Munich, Graz, Vienna, Prague, Warsaw, Milan, Paris and Brussels. The selection of these cities was guided by the objectives of the TopView project, with consideration given to their heterogeneous urban morphology and geographic contexts, which together offer a robust basis for assessing model generalization.

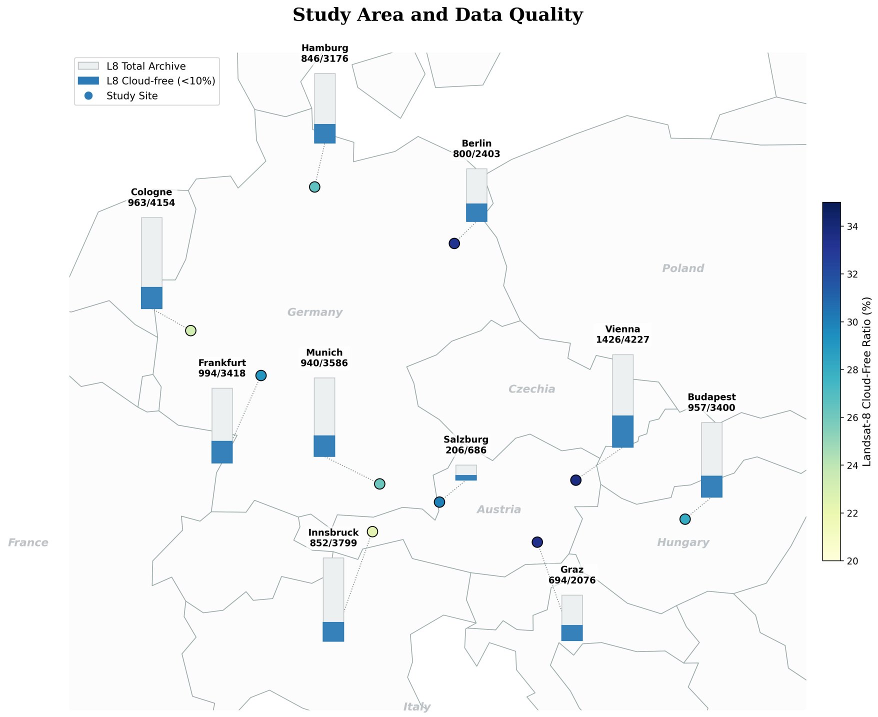


Figure 2.2: Study Area and Landsat-8 data statistics for the selected study areas. Blue bars indicate cloud-free patches (<10% cloud cover) relative to the total archive size.

Data were obtained from the Harmonized Landsat and Sentinel-2 (HLS) project jointly developed by NASA and USGS(NASA Goddard Space Flight Center, 2026). The Harmonized Landsat and Sentinel-2 (HLS) project is a NASA initiative aiming to produce a seamless surface reflectance record from the Operational Land Imager (OLI) and Multi-Spectral Instrument (MSI) aboard Landsat-8/9 and Sentinel-2A/B remote sensing satellites, respectively. Landsat-8 OLI reflectance bands B02–B07 (blue 0.45–0.51 µm, green 0.53–0.59 µm, red 0.64–0.67 µm, near-infrared 0.85–0.88 µm, SWIR1 1.57–1.65 µm, SWIR2 2.11–2.29 µm) and thermal infrared bands B10 and B11, corresponding to the TIRS1 (10.60–11.19 µm) and TIRS2 (11.50–12.51 µm) channels, were employed. Each image patch, therefore, consists of nine co-registered channels: six reflectance bands, two thermal bands, and one auxiliary Fmask channel indicating cloud and shadow classes. The analysis was restricted to Landsat-8 L30 products to focus on thermal remote sensing in urban environments. Thermal infrared bands B10 and B11 are uniquely available from Landsat-8 and, together with multispectral optical bands at 30 m spatial resolution, provide an appropriate balance between spatial detail and regional coverage for urban heat studies. All experiments are conducted on fixed-size image patches defined within the HLS tiling system, which serve as the basic spatial units for synthetic cloud generation, baseline interpolation, and learning-based reconstruction.

2.1.2 Cloud masking and cloud-free selection

The HLS Fmask/bitmask layer was used to identify clouds and cloud shadows. For each patch, an observation was considered cloud-free if the cloud fraction $p_{\text{cloud}} < 0.1$. The cloud-free ratio is computed as:

$$r_{\text{cf}} = \frac{N_{\text{cf}}}{N_{\text{total}}} \times 100\%,$$

where N_{cf} and N_{total} denote the number of cloud-free and total observations, respectively.

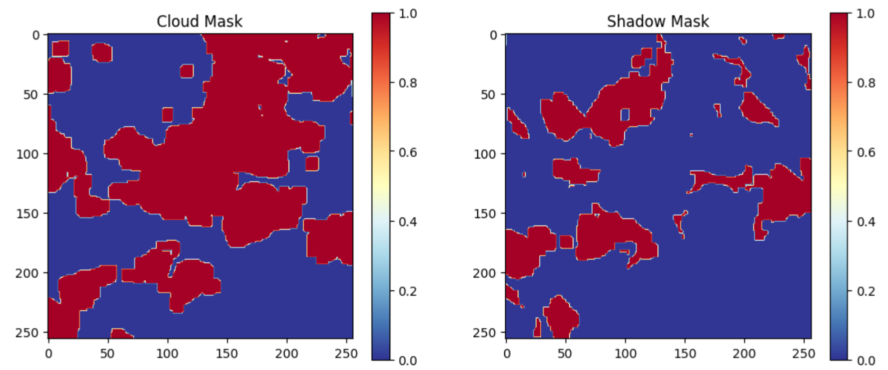


Figure 2.3: Example of cloud (left) and shadow (right) masks derived from the HLS Fmask data. Red areas indicate cloud/shadow pixels, while blue areas denote clear-sky regions.

Visible, near-infrared (VNIR), short-wave infrared (SWIR), and thermal infrared (TIRS B10/B11) bands from Landsat-8 were used. For quality inspection and visualization, the natural-color RGB composite (B04/B03/B02) is used. To ensure consistency, per-patch scaling (2–98 percentile linear stretching) is applied, and invalid pixels (clouds, clouds-shadows, saturation) are masked. A total of 106,736 patches were produced. For Landsat-8, 30,925 patches were available, of which 8,678 (28.1%) were classified as cloud-free.

2.2 Synthetic cloud generation

Building on the SatelliteCloudGenerator (SCG) framework for controllable cloud and cloud-shadow synthesis in multi-spectral imagery (Czerkawski et al., 2023), its effective rendering strategy for VIS, NIR, and SWIR bands is retained. Several modifications are introduced to improve reproducibility and physical consistency. Input data are cleaned, and channel definitions are standardized. Cloud coverage is controlled using deterministic seeding to ensure exact reproducibility. Optical-band compositing is performed in the reflectance domain. In addition, a physically consistent thermal-infrared module is introduced, operating in radiance space according to Planck's law, which is not explicitly modeled in the original SCG approach.

Given a cloud-free HLS/Landsat patch $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, representing a multi-band image tensor with C spectral channels and spatial dimensions $H \times W$, input values are preprocessed to ensure physical consistency. Invalid entries (NaN or infinite values) are removed, and remaining pixel values are clipped to physically plausible ranges. Following the design principles of the Harmonized Landsat and Sentinel-2 (HLS) surface reflectance product (Claverie et al., 2018), reflectance bands are treated in normalized reflectance units, while thermal infrared channels are consistently represented in absolute temperature units.

Thermal channels (B10/B11) are therefore harmonized to Kelvin irrespective of their original storage scale, allowing a uniform treatment of thermal information across samples. This harmonization step is intended to ensure internal consistency of the synthetic dataset rather than to replicate sensor-specific calibration procedures, which are documented elsewhere for Landsat 8 TIRS (Montanaro et al., 2014). The mapping to Kelvin is performed according to

$$T_{\mathrm{K}} = \begin{cases} T, & \text{(kelvin)}, \\ 200 + 130\,T, & \text{(scaled01)}, \\ 200 + 130\,\dfrac{T}{10000}, & \text{(scaled10k)}. \end{cases} \tag{2.1}$$

Cloud morphology is generated following the core concept of the SatelliteCloudGenerator (SCG) framework. A continuous opacity field $\alpha \in [0,1]^{H\times W}$ is sampled from multi-scale Perlin/Flex fractal noise blended across several spatial scales (e.g., $\sigma = 8/32/96$ px), and cloud boundaries are softened using Gaussian filtering. A prescribed cloud fraction $r^\star$ is obtained by determining a threshold τ such that

$$\text{cover} = \frac{1}{HW}\sum\nolimits_{i,j} \mathbf{1}\{\alpha_{ij} > \tau\} \approx r^\star, \tag{2.2}$$

with default values $r^\star \in [0.15, 0.25]$ and $\tau = 0.15$. All stochastic components are deterministically seeded using filename-based hashes to guarantee sample-level reproducibility.

Optical synthesis is performed in reflectance space by alpha compositing a channel-wise cloud reflectance template $\boldsymbol{\kappa} \in \mathbb{R}^C$ with a shadow-attenuated background. Cloud-shadows are approximated by spatially shifting regions of high opacity along the solar illumination direction, and background reflectance is attenuated accordingly. The resulting synthetic reflectance field is given by

$$\mathbf{Y}_{\text{refl}} = (1-\alpha) \odot (\gamma \mathbf{X}) + \alpha \odot \left(\boldsymbol{\kappa} \otimes \mathbf{1}_{H\times W}\right), \tag{2.3}$$

where γ controls the strength of shadow attenuation. For qualitative inspection only, clean and cloudy RGB previews are rendered using a joint 2–98% percentile stretch; an optional mild haze correction ($\gamma_{\text{haze}} \approx 1.03$–$1.15$) may be applied exclusively for visualization purposes.

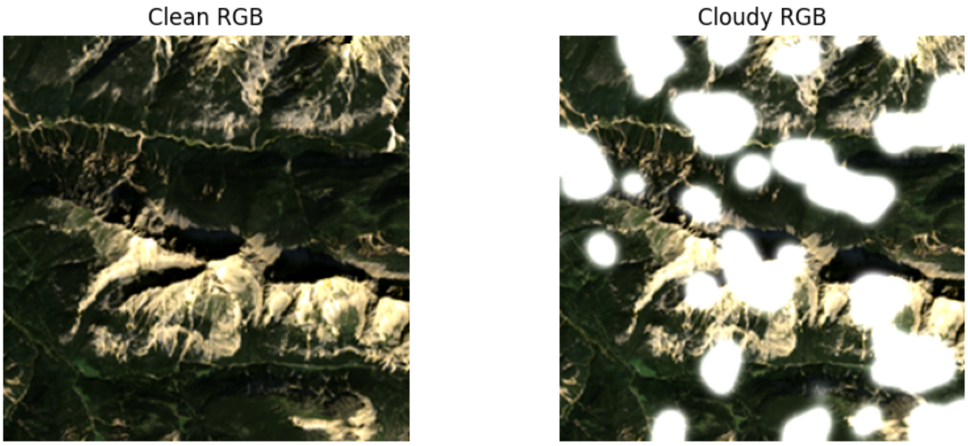


Figure 2.4: **Optical synthesis example (RGB).** Left: *Clean RGB* rendered with a joint 2–98% percentile stretch. Right: *Cloudy RGB* synthesized with the same stretch. Clouds appear bright with soft edges, and cloud shadows are cast consistently with the assumed sun direction. The clean and cloudy previews share the same contrast window to enable fair visual comparison.

To ensure physically plausible cloud-induced cooling in the thermal infrared (TIR) domain, linear mixing in temperature space is avoided. Instead, all thermal synthesis operations are carried out in radiance space, which is consistent with the physical interpretation of satellite-derived land surface temperature retrieval (Li et al., 2013). A smooth cooling field $\Delta T(x,y) \sim \mathcal{U}(5,15)\,\text{K}$ is sampled, Gaussian-smoothed, and modulated by the cloud opacity $\alpha(x,y)$. The selected range reflects typical cloud-induced cooling magnitudes observed in thermal infrared satellite imagery and is intended to represent a plausible variability rather than a sensor-specific physical parameter. The resulting cloud-affected brightness temperature proxy is defined as

$$T_{\text{cloud}}(x,y) = \text{clip}\Big(T_{\text{surf}}(x,y) - \Delta T(x,y)\,\alpha(x,y),\ 180,\ 330\Big), \tag{2.4}$$

where $\text{clip}(u,a,b) = \min(\max(u,a),b)$. Here, T_{cloud} serves as a synthetic brightness temperature proxy and does not represent a physically retrieved cloud-top temperature.

Using band-center wavelengths $\lambda_{10} \approx 10.9\,\mu\text{m}$ and $\lambda_{11} \approx 12.0\,\mu\text{m}$, spectral radiance is computed according to Planck's law and inverted as

$$L_\lambda(T) = \frac{2hc^2}{\lambda^5} \frac{1}{\exp\left(\dfrac{hc}{\lambda k_B T}\right) - 1}, \tag{2.5}$$

$$T(L_\lambda) = \frac{hc}{\lambda k_B} \Big/ \ln\left(\frac{2hc^2}{\lambda^5 L_\lambda} + 1\right). \tag{2.6}$$

Radiances are then linearly mixed using a nonlinear opacity weighting α^p, with $p \approx 1.5$ to de-emphasize optically thin clouds,

$$L_{\text{out}} = (1 - \alpha^p)\, L_\lambda\big(T_{\text{surf}}\big) + \alpha^p\, L_\lambda\big(T_{\text{cloud}}\big), \tag{2.7}$$

$$T_{\text{out}} = T\big(L_{\text{out}}\big). \tag{2.8}$$

For persistence, the resulting temperature field is mapped back to the original storage scale according to

$$T_{\text{stored}} = \begin{cases} T_{\text{out}}, & \text{(kelvin)}, \\ \dfrac{T_{\text{out}} - 200}{130}, & \text{(scaled01)}, \\ 10000 \cdot \dfrac{T_{\text{out}} - 200}{130}, & \text{(scaled10k)}. \end{cases} \tag{2.9}$$

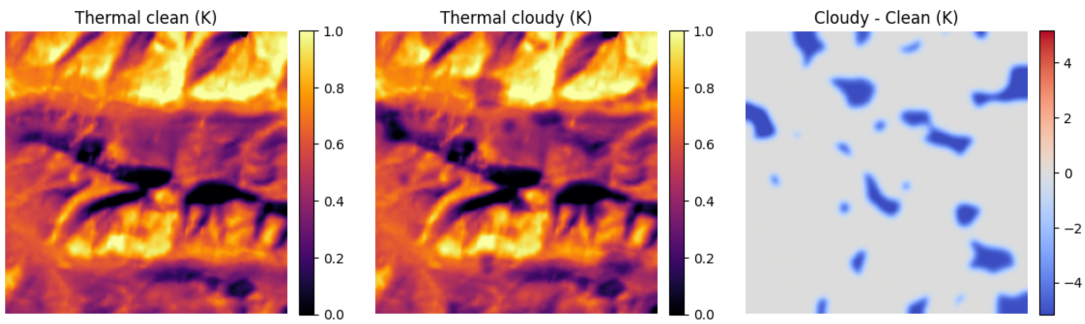


Figure 2.5: **Thermal QA in Kelvin (radiance-space synthesis).** Left: *Thermal clean (K)*. Middle: *Thermal cloudy (K)*. Right: *Cloudy − Clean brightness temperature difference (K)*. Cloudy regions show *negative* differences whose magnitude increases with opacity, exhibiting a smooth, thickness-dependent texture consistent with Planck-based radiance mixing.

Afterward, routine quality assurance procedures are applied. These include numerical sanitization and automatic regeneration of degenerate samples, verification of band alignment on cloud-free pixels using per-band correlation analysis with correction where required, and thermal quality assessment in Kelvin by visualizing *Clean (K)*, *Cloudy (K)* and *Cloudy − Clean (K)* fields. Differences are computed after radiance-space synthesis and subsequent inversion to brightness temperature. In cloudy regions, differences are negative and increase in magnitude with cloud opacity; for visualization only, optional Gaussian smoothing with $\sigma \approx 48$ px may be applied to the difference map. For diffusion-based inpainting, the synthetic cloudy RGB image and a binarized mask with the required polarity (known $=$ 255, unknown/cloud $= 0$) are exported.

Following the synthetic cloud generation procedure described above, a total of 8,678 data samples were produced. Each sample is defined on a single HLS tile, which serves as the basic spatial unit, and consists of a strictly co-registered pair of a cloud-contaminated image and its corresponding cloud-free reference, together with cloud mask information.

For each sample, the clean and cloudy images are stored as multi-channel tensors with a fixed spatial size of 256×256 pixels. Binary and continuous cloud masks are provided to indicate cloud-covered regions and their associated uncertainty. In addition, lightweight preview images and a metadata file are included to support quality inspection and to record essential information such as band ordering and key synthesis parameters. These paired cloudy/clean image tiles, along with their associated masks, constitute a unified data representation that is used consistently in subsequent baseline interpolation experiments and learning-based reconstruction models.

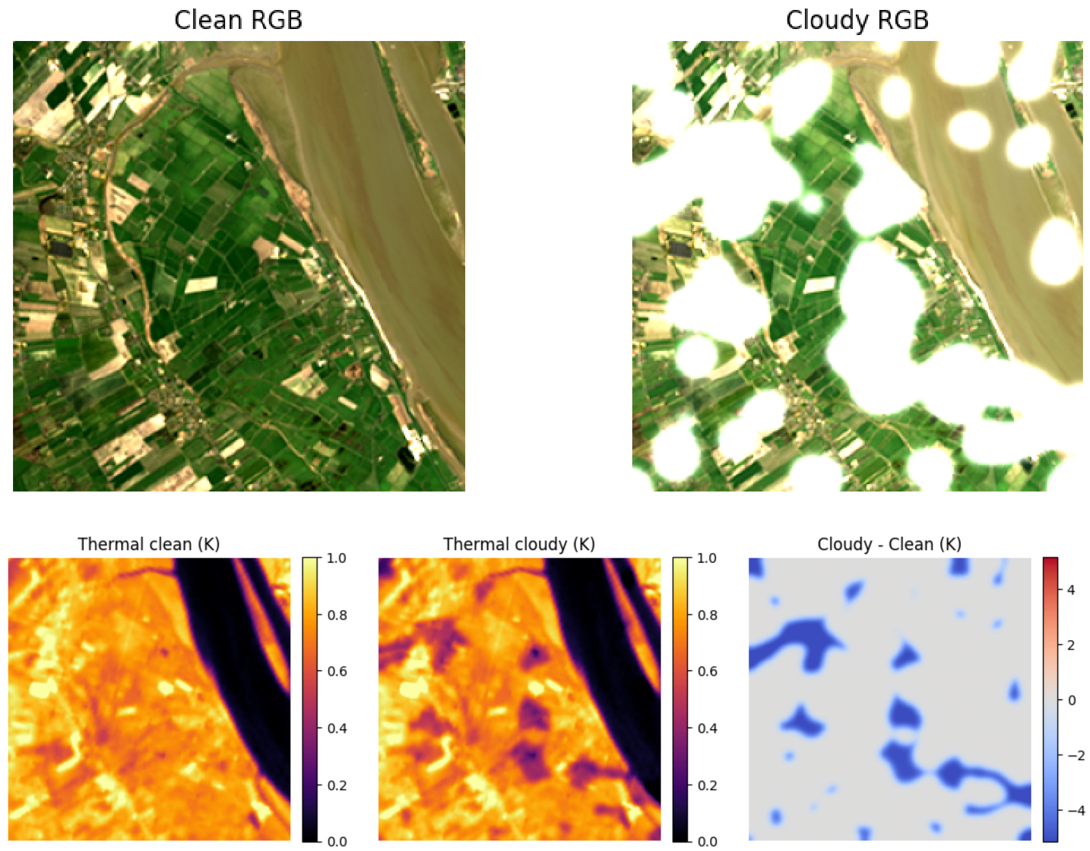


Figure 2.6: **Additional synthetic cloud generation examples in the optical and thermal domains.** Top: clean and cloud-contaminated RGB composites rendered using a joint 2–98% percentile stretch. Bottom: thermal infrared quality assessment in Kelvin, showing the clean reference, cloud-affected output, and the corresponding brightness temperature difference. Cloudy regions exhibit spatially smooth, opacity-dependent cooling patterns consistent with radiance-space synthesis.

2.3 Baseline Spatial Interpolation Methods

To evaluate the learning-based reconstruction models against transparent and reproducible reference approaches, two classical spatial interpolation techniques were adopted as benchmark methods: inverse distance weighting (IDW) and ordinary kriging (OK). These methods are widely used in remote sensing gap-filling studies from Landsat and MODIS imagery, because they can operate directly on spatially incomplete observations without requiring auxiliary predictors (Li et al., 2013; Stathopoulou and Cartalis, 2009b; Crosson et al., 2012; Neteler, 2010).

In operational processing chains, cloud- and shadow-contaminated pixels are typically identified using standardized algorithms such as Fmask for Landsat imagery (Zhu and Woodcock, 2012) or Sen2Cor for Sentinel-2 products (Main-Knorn et al., 2017). These procedures generate binary cloud masks that leave spatial gaps in thermal observations. Classical interpolation methods such as IDW and OK are commonly applied to such masked datasets to exploit local spatial redundancy while preserving scene-level radiometric consistency.

In this study, both methods operate strictly within the spatial domain of an individual HLS tile and utilize only cloud-free pixels as valid observations. No temporal context or auxiliary predictors are included. Under these constraints, the interpolation baselines provide a conservative and interpretable reference for reconstruction performance based solely on local spatial information.

2.3.1 Inverse Distance Weighting (IDW)

Inverse Distance Weighting (IDW) is a deterministic interpolation method in which the value at a target location is estimated as a weighted average of nearby observations, with weights decreasing as a function of spatial distance. For a cloud-covered pixel located at position p, the reconstructed value $\hat{X}(p)$ is defined as

$$\hat{X}(p) = \frac{\sum_{q \in \mathcal{N}(p)} w(p,q)\, X(q)}{\sum_{q \in \mathcal{N}(p)} w(p,q)}, \qquad w(p,q) = \frac{1}{d(p,q)^k}, \tag{2.10}$$

where $\mathcal{N}(p)$ denotes the set of cloud-free neighboring pixels within a predefined search radius, $d(p,q)$ is the Euclidean distance between pixel centers, and the distance decay exponent is fixed to $k = 2$. Only cloud-free pixels within the same HLS tile contribute to the interpolation.

2.3.2 Ordinary Kriging (OK)

Ordinary Kriging (OK) is a geostatistical interpolation method that explicitly accounts for spatial autocorrelation through a variogram model. Let $Z(p)$ denote a spatial random field corresponding to a spectral or thermal band within an HLS tile. Under the assumption of an unknown but locally constant mean, the value at a masked location p is estimated as

$$\hat{Z}(p) = \sum_{i=1}^{n} \lambda_i\, Z(p_i), \tag{2.11}$$

where $\{p_i\}_{i=1}^{n}$ are neighboring cloud-free pixels within the same tile and λ_i are interpolation weights obtained by solving the Kriging system. The weights are determined such that the estimator is unbiased and the estimation variance is minimized, given the spatial covariance structure described by the fitted variogram. Interpolation is performed within a local neighborhood to ensure numerical stability and to limit the influence of spatially distant samples.

2.4 Learning-based Reconstruction Models

The implementation of the learning-based reconstruction models follows the same data handling framework as the synthetic cloud generation procedure described above. In this study, each HLS tile patch is treated as an independent spatial processing unit, and supervised learning is conducted using paired cloud-contaminated and cloud-free samples. Cloud-affected regions are explicitly identified by a binary cloud mask and a continuous soft mask, which are consistently applied during both model training and evaluation. Thermal infrared channels are processed in a physically consistent domain to preserve the validity of radiometric quantities. All data are exported in standardized formats and accompanied by complete metadata to ensure experimental reproducibility.

Within this unified framework, two learning-based reconstruction strategies are investigated. First, a U-Net architecture is employed as a deterministic regression model that directly maps cloud-contaminated inputs to their corresponding cloud-free estimates. The model produces a single prediction at each pixel location and does not explicitly model uncertainty or data distributions. As such, it serves as a structurally clear and conceptually simple learning-based baseline. Second, a diffusion-based reconstruction model is introduced, in which cloud removal is formulated as a conditional generative problem. In this setting, the U-Net architecture is reused as the noise-prediction network within an iterative denoising process, while the overall model explicitly represents stochasticity in the reconstruction process through probabilistic sampling.

Although both approaches share the same U-Net architecture at the network level, their functional roles differ fundamentally. In the regression setting, the U-Net operates as a deterministic estimator of the conditional expectation. In contrast, within the diffusion-based model, it constitutes a core component of a probabilistic generative mechanism, enabling the synthesis of multiple plausible reconstructions under severe information loss.

2.4.1 Regression U-Net

As a first learning-based reconstruction approach, a U-Net architecture is adopted as a deterministic regression model for cloud-contaminated thermal reconstruction. The U-Net was originally proposed by Ronneberger et al. (Ronneberger et al., 2015) and has since become a widely used architecture for pixel-wise prediction tasks due to its symmetric encoder–decoder structure and skip connections, which effectively combine multi-scale contextual information with fine-grained spatial details.

In the context of remote sensing, U-Net-based models have demonstrated strong performance across a range of applications, including cloud detection in multi-spectral satellite imagery (Jeppesen et al., 2019) and lightweight adaptations designed for visible and thermal infrared data (Zhang et al., 2020). These studies indicate that the U-Net architecture is well-suited for spatially heterogeneous reconstruction problems, where both local continuity and broader spatial context play a critical role.

Within this study, the regression U-Net is formulated as a deterministic mapping from cloud-contaminated observations to their corresponding cloud-free estimates. For each pixel location, the model produces a single reconstructed value and does not explicitly represent uncertainty or alternative plausible solutions. As such, the model approximates the conditional expectation of the target given the input and serves as a conceptually simple yet effective learning-based baseline.

The U-Net architecture is a fully convolutional neural network designed for dense, pixel-wise prediction tasks. It adopts a symmetric encoder–decoder structure in which the contracting path progressively captures high-level contextual information through repeated convolution and downsampling operations, while the expanding path restores spatial resolution via upsampling layers. A defining characteristic of U-Net is the use of skip connections that directly transfer feature maps from the encoder to the corresponding decoder stages, enabling the fusion of coarse semantic information with fine-grained spatial details.

As illustrated in Figure 2.7, each resolution level consists of successive 3×3 convolutional layers followed by nonlinear activations, while spatial resolution is reduced or increased through pooling and upsampling operations, respectively. The skip connections concatenate multi-scale feature representations across the network depth, allowing the decoder to recover precise spatial structures that would otherwise be degraded by repeated downsampling. This architectural design has been shown to be particularly effective for pixel-wise prediction tasks in remote sensing, where both global context and local spatial continuity are essential.

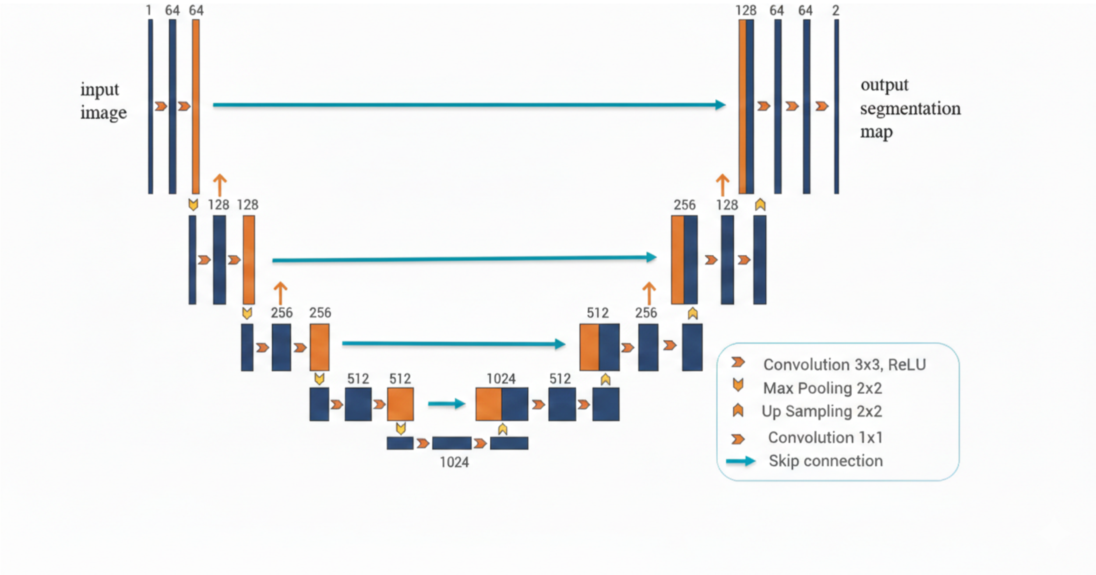


Figure 2.7: **U-Net architecture overview.** The network follows a symmetric encoder–decoder design composed of repeated 3×3 convolutional layers with nonlinear activation functions, 2×2 max-pooling for downsampling, and 2×2 upsampling for spatial resolution recovery. Skip connections transfer feature maps from the encoder to the corresponding decoder stages, enabling the fusion of multi-scale contextual and spatial information. *U-Net architecture diagram, adapted from the original study by Ronneberger et al.*(Ronneberger et al., 2015).

The regression U-Net is trained and evaluated using paired cloudy and cloud-free samples extracted from HLS tile patches. Each sample corresponds to a 256×256 pixel patch with nine spectral channels. Six reflectance bands (B02–B07) cover the visible, near-infrared, and shortwave infrared regions, while two thermal infrared bands (B10 and B11) provide temperature-related information. An additional Fmask channel is included as an auxiliary input to encode cloud- and shadow-related classification information derived from the operational cloud detection algorithm.

To further guide the reconstruction process, a continuous soft cloud mask with values in the range $[0, 1]$ is provided for each pixel, representing the estimated degree of cloud contamination. A corresponding binary cloud mask is available for strict delineation of cloudy regions when required. Cloud-free patches are used as regression targets, and the corresponding cloudy observations serve as model inputs. The dataset is divided into training, validation, and test subsets using an $80\%/10\%/10\%$ split.

By design, the regression U-Net isolates the representational capacity of a convolutional encoder–decoder architecture from the effects of probabilistic modeling. This allows its performance to be directly compared with both classical interpolation methods and more advanced generative approaches under identical spatial and data constraints.

2.4.2 Diffusion-based Reconstruction Model

To explicitly address the inherent uncertainty associated with cloud-induced information loss, a diffusion-based reconstruction model is further investigated. In this formulation, cloud removal is treated as a conditional generative problem, where the goal is to model a distribution of plausible cloud-free solutions conditioned on the observed cloudy input.

Diffusion models are based on a forward–reverse process. In the forward process, clean target samples are gradually corrupted by the addition of Gaussian noise over a sequence of timesteps, resulting in a near-isotropic noise distribution. The reverse process is learned as an iterative denoising operation that progressively removes noise and restores structure. Conditioning on the cloudy observation constrains this reverse process and guides the generation toward reconstructions that are consistent with the available information.

To address the spatial variability of thermal infrared observations and the limitations of deterministic reconstruction, this study adopts a generative framework based on Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020). Within this framework, cloud-gap filling is treated as a conditional image-to-image translation task. The model reconstructs the underlying thermal signals by reversing a progressive Gaussian diffusion process.

The theoretical foundation of the DDPM framework consists of two coupled processes:

1. **Forward Diffusion Process**: A clean image I_0 is incrementally transformed into a latent noise distribution across T timesteps. Controlled by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$, the noisy state I_t at any arbitrary step t can be sampled directly as:
$$I_t = \sqrt{\bar{\alpha}_t} I_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{2.12}$$
where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. As $t \to T$, the original thermal structure is systematically replaced by isotropic Gaussian noise.

2. **Reverse Denoising Process**: A U-Net-based neural network ϵ_θ is trained to invert the diffusion. The model predicts the specific noise component ϵ injected at each step, conditioned on the cloud-contaminated observation I^c and the corresponding cloud mask C. The learning objective is to minimize the Mean Squared Error (MSE) loss:
$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{I_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(I_t, t \mid I^c, C)\|^2\right] \tag{2.13}$$

During inference, a RePaint-style conditional sampling strategy is utilized to enforce fidelity to the observed data. At each reverse timestep $t \rightarrow t-1$, the known (cloud-free) regions are updated by sampling from the observed image with a noise level synchronized to the current step, while the unknown (cloud-covered) regions are generated by the denoising network. This stochastic blending is defined as:

$$I_{t-1} = C \odot I_{t-1}^{\text{pred}} + (1 - C) \odot \tilde{I}_{t-1}^{c} \tag{2.14}$$

where $\odot$ denotes element-wise multiplication and $\tilde{I}_{t-1}^{c}$ represents the noise-corrupted observation. To enhance the stability and generalization of the generative output, an Exponential Moving Average (EMA) of the model weights is maintained during training:

$$\theta_{\text{EMA}} \leftarrow \lambda \theta_{\text{EMA}} + (1 - \lambda)\theta \tag{2.15}$$

where λ is the decay rate. This ensures the reconstruction of complex, non-linear thermal structures that remain physically and statistically consistent with the surrounding clear-sky pixels.

Within this framework, the U-Net architecture is employed as a noise-prediction backbone rather than as a direct predictor of cloud-free outputs. At each diffusion step, the network estimates the noise component present in the current latent representation, given both the corrupted sample and the corresponding cloudy input. Through repeated application of this denoising operation, cloud-free reconstructions are gradually synthesized.

Unlike the regression U-Net, the diffusion-based model naturally supports stochastic sampling, allowing multiple reconstructions to be generated for the same cloudy input. This property is particularly relevant under dense or contiguous cloud cover, where large portions of the surface are unobserved and multiple physically plausible solutions may exist.

The diffusion-based reconstruction model was implemented as a conditional inpainting framework operating on multi-spectral Landsat-8 patches of size 256×256 pixels. Input channels consisted of VNIR/SWIR reflectance bands, thermal infrared bands B10 and B11 expressed in Kelvin brightness temperature, and a continuous soft cloud mask indicating cloud-affected regions. The noise-prediction network followed a U-Net backbone with two residual blocks per resolution level, 4 attention heads, and attention resolutions at 32, 16, and 8 pixels. The base feature width was set to 128 channels and doubled after each downsampling stage.

Training used a linear noise schedule with 1,000 diffusion timesteps, a batch size of 4, and a learning rate of 1×10^{-4} optimized with AdamW. The model was trained for 40,000 iterations with an exponential moving average (EMA) decay of 0.999.

Unless otherwise specified, remaining hyperparameters followed the default settings of the reference DDPM implementation. During inference, a RePaint-style conditional sampler was employed with 250 denoising timesteps, a jump length of 10, and 5 resampling cycles. Known cloud-free pixels were enforced at each iteration according to the cloud mask. All experiments were implemented in PyTorch 2.0.1 and executed on Ubuntu 22.04 within the Google Colab Pro environment, using an NVIDIA A100 40GB GPU (when available) together with standard Colab virtual machine CPUs. Random seeds were fixed to ensure reproducibility.

Chapter 3

Evaluation Methodology and Validation Protocol

This chapter describes the validation protocol and evaluation metrics used to assess reconstruction performance under cloud-induced information loss. Rather than introducing additional methodological components, the focus here is on establishing a unified and reproducible evaluation framework that enables a fair comparison between spatial interpolation baselines, deterministic regression models, and diffusion-based generative models. All quantitative results reported in Chapter 4 are derived exclusively from the metrics and aggregation procedures specified in this chapter.

3.1 Evaluation Scope and Priorities

The primary target variable of this study is thermal reconstruction. This focus is consistent with the critical role in monitoring surface energy balances and urban climate dynamics (Li et al., 2013). Accordingly, thermal accuracy metrics constitute the core of the evaluation framework. Optical reconstruction metrics are reported as complementary indicators of spatial structure and texture preservation within cloud-covered regions, but they are not treated as the principal objective of the study.

All pixel-level evaluation is restricted to cloud-covered regions to ensure that reported errors reflect reconstruction performance under genuine information loss rather than overall image similarity. Unless stated otherwise, the term *cloud-covered* refers to pixels identified by the binary cloud mask. In addition, a soft cloud probability map is used to stratify cloud-covered pixels into thin- and thick-cloud subsets, allowing reconstruction robustness to be assessed under varying degrees of cloud opacity.

3.2 Data Partitioning and Test Set Construction

Evaluation is performed on a held-out test set that is strictly separated from all data used for model training and hyperparameter tuning. Nearly cloud-free reference images are selected according to the criteria described in Chapter 2 and are never exposed to learning-based models during training.

Synthetic cloud masks are applied exclusively at the evaluation stage to generate paired cloud-contaminated inputs and cloud-free reference targets. Because the reference values are fully known, this setup allows reconstruction errors to be computed directly under controlled cloud configurations. All reconstruction methods are evaluated on identical test samples and cloud mask realizations, ensuring that observed performance differences can be attributed to methodological characteristics rather than data selection effects.

3.3 Cloud Masks and Evaluation Regions

Let $I \in \mathbb{R}^{C\times H\times W}$ denote a cloud-free reference image and I^c its cloud-contaminated counterpart. A binary cloud mask $M \in \{0,1\}^{H\times W}$ identifies pixels requiring reconstruction, where $M(p) = 1$ indicates cloud-covered pixels.

3.3.1 Cloud-only evaluation

Pixel-level errors are computed exclusively over the set

$$\Omega_{\text{cloud}} = \{p \mid M(p) = 1\}.$$

Cloud-free pixels are excluded from pixel-level evaluation, as their inclusion would artificially reduce error estimates by incorporating trivially known values. This restricted validation protocol is essential for identifying model artifacts that are otherwise masked by the high fidelity of clear-sky pixels, a common challenge in automated cloud removal verification (Zhu and Woodcock, 2012). This cloud-only protocol ensures that reported metrics reflect reconstruction performance under actual information loss.

3.3.2 Thin- and thick-cloud stratification

To further characterize reconstruction robustness under varying cloud opacity, cloud-covered pixels are stratified into thin- and thick-cloud subsets based on a soft cloud probability measure. The selection of stratification thresholds follows established practice in operational cloud masking for Sentinel-2 imagery.

Specifically, probability ranges consistent with the Sen2Cor cloud classification scheme are adopted (Main-Knorn et al., 2017). In Sen2Cor, cloud probability values between approximately 20% and 65% are commonly associated with optically thin clouds or cirrus, whereas probabilities exceeding 65% indicate optically thick clouds with strong attenuation of surface signal. These categories reflect different radiative impacts and degrees of information loss and are widely used in atmospheric correction and cloud screening workflows.

Let $S \in [0, 1]^{H \times W}$ denote the soft cloud probability map associated with a given test sample. Using fixed thresholds derived from the Sen2Cor convention, cloud-covered pixels are partitioned as

$$\Omega_{\text{thin}} = \{p \mid M(p) = 1 \ \wedge \ 0.20 \leq S(p) \leq 0.65\}, \tag{3.1}$$

$$\Omega_{\text{thick}} = \{p \mid M(p) = 1 \ \wedge \ S(p) > 0.65\}, \tag{3.2}$$

where M denotes the binary cloud mask. Pixels with cloud probability below 20% are not considered cloud-affected and are therefore excluded from reconstruction error evaluation.

This stratification is applied uniformly across all reconstruction methods. By reporting evaluation metrics separately for thin- and thick-cloud regions, the analysis explicitly distinguishes between moderate and severe information loss and provides insight into method robustness under increasingly challenging cloud conditions.

3.4 Evaluation Metrics

Reconstruction performance is quantified using a set of complementary metrics that capture different aspects of thermal reconstruction accuracy. Thermal metrics are expressed in physically meaningful units (Kelvin) and form the basis for method comparison. Additional statistical measures are reported to characterize systematic bias and spatial consistency.

3.4.1 Thermal accuracy metrics (primary)

Let $T(p)$ and $\hat{T}(p)$ denote the reference and reconstructed thermal infrared temperature proxy at pixel p, respectively.

Pixel-level RMSE. Pixel-level reconstruction accuracy is quantified using the root mean squared error (RMSE),

$$\mathrm{RMSE}_{\mathrm{pp}} = \sqrt{\frac{1}{|\Omega|} \sum_{p \in \Omega} \left(T(p) - \hat{T}(p) \right)^2}, \tag{3.3}$$

where Ω is set to Ω_{cloud} for overall cloud-only RMSE and to Ω_{thin} or Ω_{thick} for stratified thin- and thick-cloud RMSE. This metric reflects local reconstruction fidelity in regions affected by information loss.

Thermal bias (mean error). To assess systematic over- or underestimation of temperature, the mean error (bias) is additionally reported:

$$\mathrm{Bias} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left(\hat{T}(p) - T(p) \right), \tag{3.4}$$

with Ω defined analogously to the RMSE computation. While RMSE captures both variance and bias components, the bias metric isolates systematic offsets and provides complementary information regarding the physical plausibility of reconstructed temperature fields.

Global regional-mean RMSE. To evaluate large-scale consistency, each test sample is summarized by its global mean temperature. For a given sample, the reference and reconstructed global means are defined as

$$\bar{T} = \frac{1}{HW} \sum_{p \in \Omega_{\mathrm{all}}} T(p), \qquad \bar{\hat{T}} = \frac{1}{HW} \sum_{p \in \Omega_{\mathrm{all}}} \hat{T}(p), \tag{3.5}$$

where Ω_{all} denotes all pixels in the patch. The global regional-mean error is then computed across the full test set:

$$\mathrm{RMSE}_{\mathrm{mean,global}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\bar{T}_i - \bar{\hat{T}}_i \right)^2}. \tag{3.6}$$

Here, the term *region* is explicitly defined as the global patch mean, yielding a single global indicator of large-scale thermal consistency.

3.4.2 Spatial correlation (supplementary)

In addition to amplitude-based error metrics, spatial consistency between reconstructed and reference temperature fields is assessed using the Pearson correlation coefficient. While RMSE quantifies absolute deviation, it may fail to capture the preservation of structural information and spatial gradients, which are vital for interpreting urban thermal patterns (Wang et al., 2004). For each test sample, the correlation is computed at the patch level over cloud-covered pixels:

$$r = \mathrm{corr}\left(\{T(p)\}_{p\in\Omega_{\text{cloud}}}, \{\hat{T}(p)\}_{p\in\Omega_{\text{cloud}}}\right). \tag{3.7}$$

Patch-level correlation coefficients are subsequently averaged across the test set to obtain a global indicator. This metric quantifies the degree to which spatial temperature variations are preserved independently of absolute magnitude and is interpreted in conjunction with RMSE and bias.

3.5 Training Diagnostics versus Validation Metrics

For diffusion-based models, the noise-prediction mean squared error (noise MSE) is monitored during training as an optimization diagnostic. This quantity measures how accurately the network predicts injected Gaussian noise within the diffusion objective and is used exclusively to assess training stability and convergence behavior. Noise MSE is not a reconstruction accuracy metric and is therefore not used for cross-method performance ranking in Chapter 4.

3.6 Aggregation and Reporting

Pixel-level metrics (RMSE, bias, and spatial correlation) are computed independently for each test sample and subsequently averaged across the full test set to obtain global indicators. Global regional-mean RMSE is reported as a single value computed over all test samples. Where applicable, thin- and thick-cloud stratified results are reported alongside overall cloud-only metrics to provide a consistent view of robustness under varying cloud opacity.

This chapter establishes a unified evaluation framework centered on thermal reconstruction accuracy, with additional metrics capturing systematic bias and spatial consistency. By combining cloud-only RMSE, mean error, global regional-mean error, and patch-level spatial correlation, the framework provides a balanced characterization of reconstruction performance without introducing redundant or method-specific indicators. The comparative results presented in Chapter 4 can therefore be interpreted without ambiguity arising from inconsistent validation settings.

Chapter 4

Results

This chapter presents quantitative and qualitative reconstruction results under cloud-induced information loss. All evaluation results follow the unified validation protocol defined in Chapter 3, including (i) cloud-only pixel-level evaluation, (ii) thin/thick stratification based on Sen2Cor probability ranges (thin: 20%–65%; thick: $> 65\%$), and (iii) global regional-mean evaluation where the *region* is defined as the global patch mean. Thermal reconstruction is treated as the primary objective, while optical metrics are reported as auxiliary indicators of structural and textural fidelity within cloud-covered regions.

4.1 Dataset Preparation Outcomes and Quality Assessment

A major part of this study consisted of constructing a paired dataset of multi-spectral image patches suitable for controlled evaluation under cloud-induced information loss. The final dataset contains 7,695 co-registered cloudy/clean patch pairs, each associated with a soft cloud mask and metadata required for model training and evaluation. All samples were produced by the preprocessing workflow described in Chapter 2, and all performance metrics reported in this chapter follow the validation protocol defined in Chapter 3.

The synthetic cloud fields exhibit a controlled range of cloud coverage, with achieved cloud fractions between 0.145 and 0.254 and a mean value of 0.200. The 10th, 25th, 50th, 75th, and 90th percentiles are 0.160, 0.176, 0.201, 0.225, and 0.240, respectively. This distribution spans a representative range of cloud conditions while avoiding extreme cases that could dominate aggregated error statistics.

The distribution of samples across study areas is summarized in Fig. 4.1. Differences in patch counts reflect variations in scene availability, cloud-free reference coverage, and valid patch yield after quality filtering. The data set was constructed by pooling samples from all study areas rather than stratifying by city, so the figure provides an overview of the composition of the data rather than a training split by location.

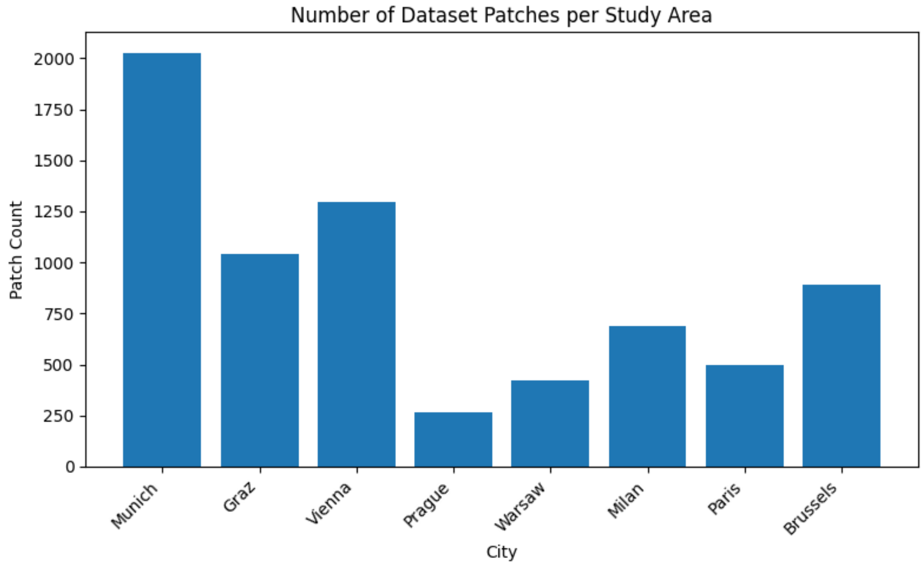


Figure 4.1: **Number of dataset patches per study area.** Bars show the number of paired cloudy/clean image patches derived from each city, based on HLS tile identifiers. Differences in counts reflect variations in data availability and valid patch yield across study areas.

Examples of synthetic cloud generation are shown in Fig. 4.2. The upper panels present optical RGB imagery before and after the introduction of synthetic clouds, while the lower panels show the corresponding thermal infrared reference, the synthetic cloudy image, and the brightness temperature difference. These examples illustrate the structure of the paired samples used throughout the experiments and document the result of the data preparation procedure. The resulting data set therefore provides consistent paired observations across a range of cloud-cover conditions and study areas, forming a controlled basis for comparing reconstruction methods under identical input conditions.

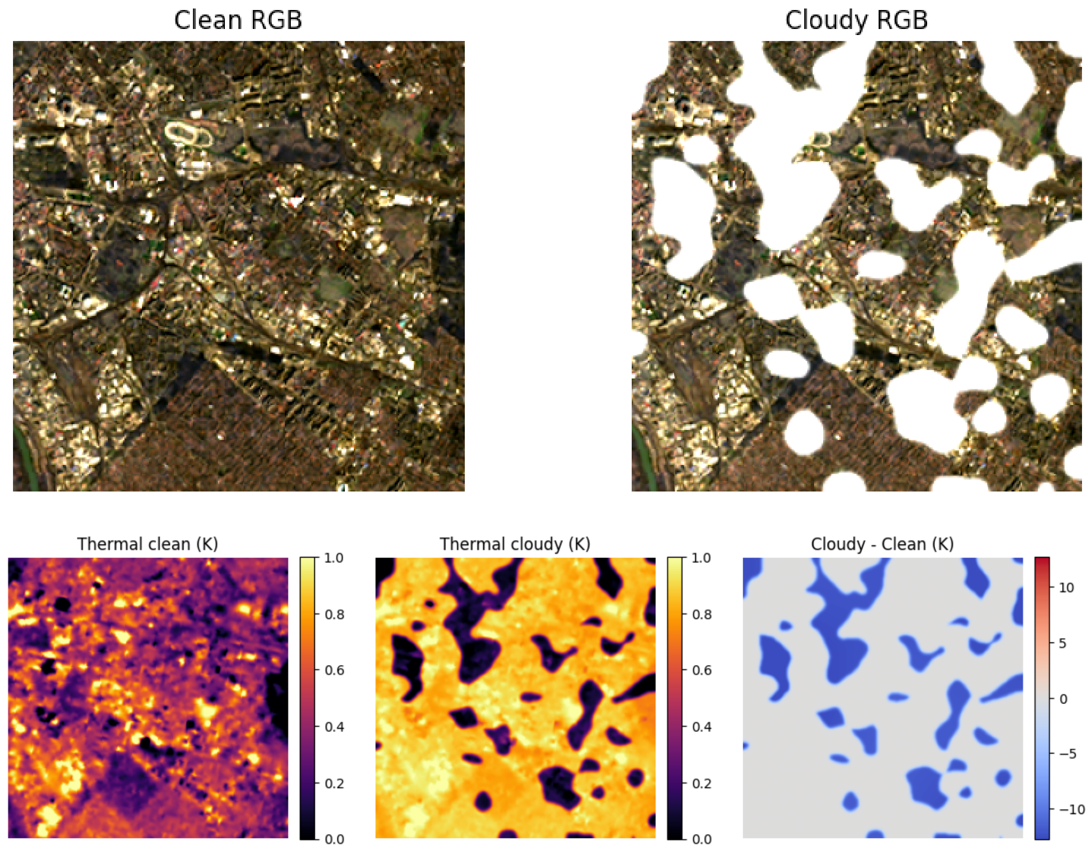
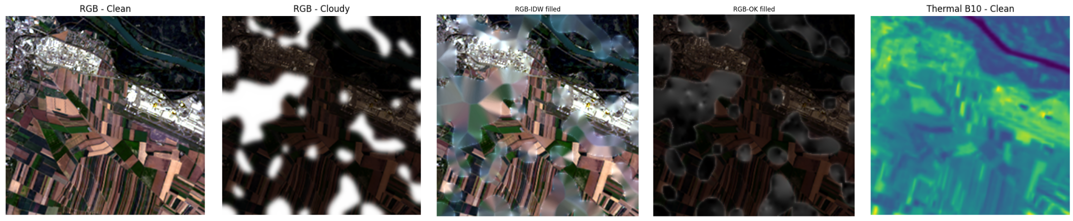


Figure 4.2: **Examples of synthetic cloud generation used for constructing paired training samples.** Top: clean and cloud-contaminated RGB imagery. Bottom: thermal infrared images showing the clean reference, synthetic cloudy image, and the corresponding brightness temperature difference.

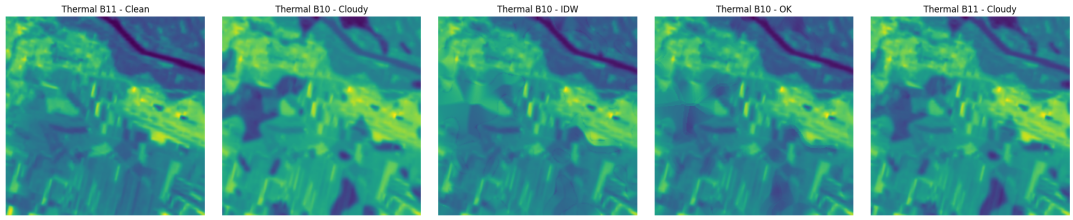
4.2 Geostatistical Interpolation Methods

To benchmark the learning-based reconstruction models under conditions consistent with standard remote sensing workflows, two widely used spatial interpolation methods, inverse distance weighting (IDW) and ordinary kriging (OK), were evaluated. In gap-filling studies, these methods are often applied after cloud masking using algorithms such as Fmask or Sen2Cor, and therefore provide a conservative and operationally relevant reference based solely on local spatial information.

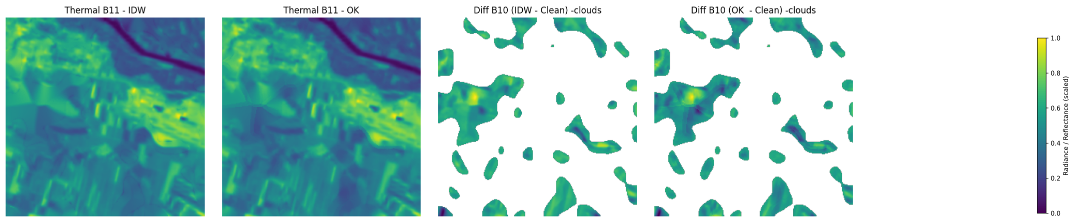
IDW reconstructs cloud-covered pixels as a weighted average of nearby valid observations with inverse-distance-squared weighting, whereas OK additionally accounts for spatial autocorrelation through an empirical variogram (here fitted with an exponential model). Interpolation was performed locally using the 50 nearest valid neighbours to balance computational efficiency and spatial fidelity. As expected, both methods yield spatially smooth reconstructions and may fail to recover fine-scale thermal structure when cloud gaps are large and contiguous. Because thermal



(a) RGB—Clean (b) RGB—Cloudy (c) RGB—IDW (d) RGB—OK (e) Thermal B10—Clean



(f) Thermal B11—Clean (g) Thermal B10—Cloudy (h) Thermal B10—IDW (i) Thermal B10—OK (j) Thermal B11—Cloudy



(k) Thermal B11—IDW (l) Thermal B11—OK (m) Diff B10 (IDW–Clean) (n) Diff B10 (OK–Clean) (o) Legend

Figure 4.3: Cloud-gap filling results for interpolation baselines. Row 1: RGB references and reconstructions plus B10 (Band 10 10.60–11.19 µm) clean. Row 2: B11 (Band 11, 11.50–12.51 µm) clean and B10 reconstructions. Row 3: B11 reconstructions and B10 difference maps. IDW = inverse distance weighting; OK = ordinary kriging. The rightmost panel shows the radiance/reflectance scale.

accuracy in this study is evaluated in Kelvin and aggregated over a held-out test set following the protocol in Chapter 3, the quantitative comparison between IDW, OK, and the learning-based models is reported in Section 4.4 (Tables 4.1 and 4.3).

Figure 4.3 provides a representative qualitative comparison between IDW and OK. Both interpolation methods produce spatially smooth reconstructions and may fail to recover fine-scale structure when cloud gaps are large and contiguous. Because thermal evaluation in this thesis is conducted in Kelvin and aggregated over the held-out test set following the protocol in Chapter 3, the quantitative comparison between IDW and OK is reported together with all other methods in Section 4.5 (Table 4.1 and Table **??**).

4.3 U-Net Model

This section reports the U-Net reconstruction performance under the evaluation protocol defined in Chapter 3. All pixel-level metrics are computed over cloud-covered pixels only. Thin/thick subsets follow the Sen2Cor-based cloud probability thresholds (thin: 20%–65%; thick: $> 65\%$).

4.3.1 Model Training and Validation

Figure 4.3 shows a representative example of cloud-gap filling for the same image patch. The first row presents the RGB reference image, the cloud-contaminated input, and the corresponding reconstructions using IDW and OK, followed by the clean thermal B10 reference. This row allows visual identification of cloud locations and assessment of spatial smoothing effects.

The second row shows the clean B11 reference together with B10 reconstructions, while the third row presents B11 reconstructions and B10 difference maps (reconstruction minus clean). The difference maps highlight where interpolation errors occur, which are typically largest within thick cloud regions and along sharp urban temperature gradients.

Both IDW and OK produce spatially smooth temperature fields, but fine-scale urban thermal structures are partially lost when cloud gaps are large and contiguous. OK generally preserves gradients slightly better than IDW, while both methods underestimate local temperature extremes. These qualitative observations are consistent with the quantitative evaluation reported in Section 4.5.

Figure 4.4 shows the evolution of training and validation loss during model calibration, where the loss represents the mean squared difference between reconstructed and reference land surface temperature values and thus provides a measure of average reconstruction error. The horizontal axis denotes the training epoch, defined as one complete pass through the training dataset, during which model parameters are iteratively adjusted to reduce this error. In a stable calibration process, both curves are expected to decrease and gradually stabilise, while a large or persistent separation between them would indicate overfitting and limited generalisation. In the present case, both losses decline rapidly during the first epochs, indicating that the dominant spatial temperature patterns are learned early, followed by a slower reduction that reflects refinement of smaller-scale thermal structures; the close agreement between the training and validation curves, with only minor fluctuations, suggests stable optimisation without pronounced overfitting. The eventual flattening of both curves indicates that convergence has largely been reached and that further training would yield only marginal improvement, implying that the calibrated model is likely to perform consistently on unseen image patches.

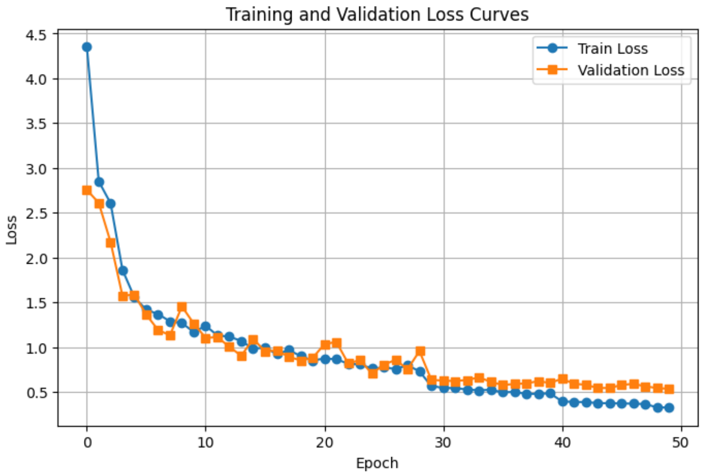


Figure 4.4: Training and validation loss curves of the regression U-Net across 50 epochs.

4.3.2 Qualitative assessment

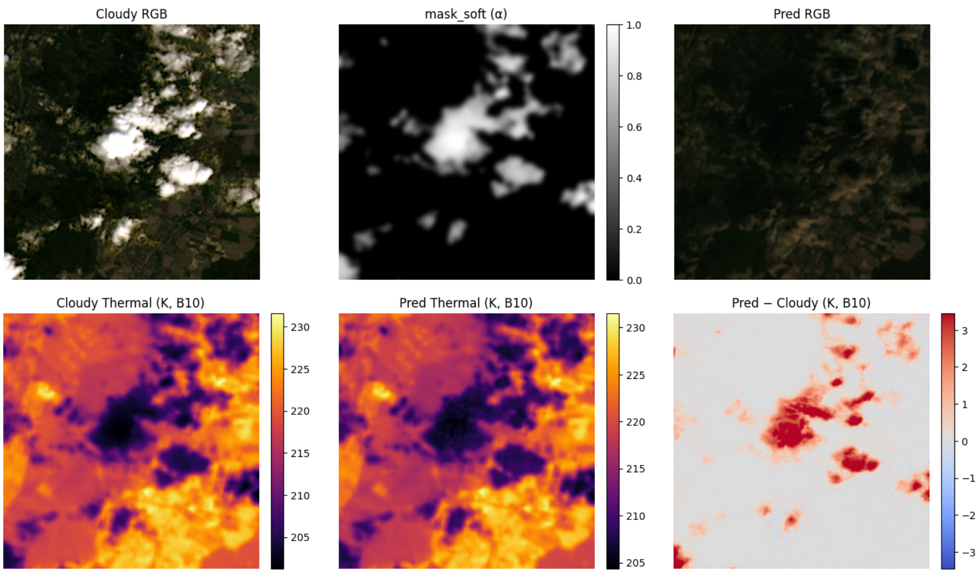


Figure 4.5: **Qualitative reconstruction results for cloud-contaminated observations.** The upper panel shows the input RGB composite, soft cloud mask, and reconstructed RGB output. The lower panel illustrates the cloudy thermal infrared band (B10), predicted thermal signal, and the corresponding difference map (residual analysis in Kelvin).

Figures 4.5 and 4.6 present reconstruction results for samples randomly drawn from the test set, following common evaluation practice in recent diffusion-based reconstruction studies. In Fig. 4.5, the model recovers continuous surface texture in the RGB channels and coherent thermal structures in band B10 over thin-cloud and

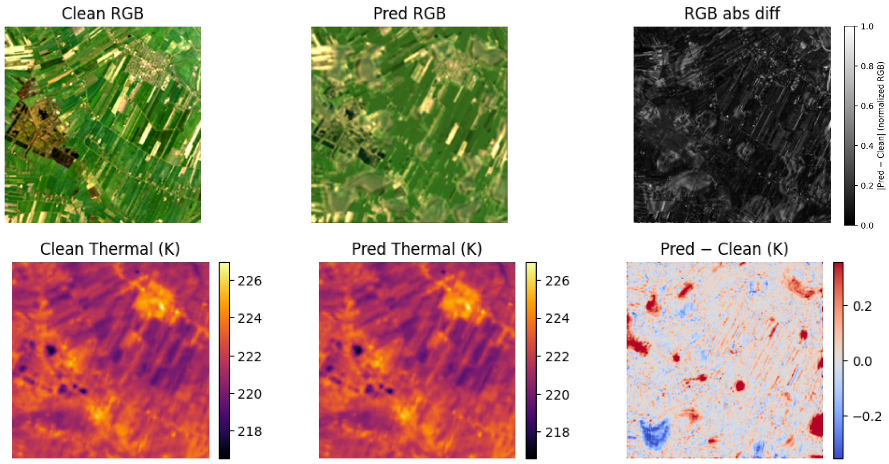


Figure 4.6: **Qualitative comparison with cloud-free reference data.** The upper panel presents the reference RGB composite, the reconstructed RGB output, and the absolute RGB difference. The lower panel displays the reference thermal band (B10, K), the predicted thermal signal (B10, K), and the residual map (prediction − reference, K).

partially occluded areas, with residual maps indicating temperature differences close to 0 K for most pixels; larger errors are spatially concentrated in regions fully covered by optically thick clouds, where no reliable observations are available. In Fig. 4.6, comparison with cloud-free reference data shows that reconstructed RGB images preserve major spatial structures such as field boundaries, rivers, and roads, while thermal residuals are predominantly within ± 0.5 K and locally reach about ± 0.75 K without systematic bias. These error magnitudes are substantially smaller than the imposed synthetic cloud cooling range (5–15 K), indicating that the method reconstructs both spatial patterns and temperature magnitudes with high fidelity, and that the remaining discrepancies are primarily attributable to information loss in densely cloud-covered regions.

4.4 Diffusion Model

A generative cloud-removal approach based on diffusion modelling was also examined. In this framework, cloud-gap filling is formulated as a conditional inpainting problem, in which cloud-contaminated pixels are reconstructed through a gradual denoising process while cloud-free observations are constrained to remain unchanged. This formulation allows the spatial context of surrounding valid observations to guide the reconstruction of missing thermal information. Figure 4.7 presents the evolution of noise-prediction mean squared error (noise MSE) during diffusion model training across 50 epochs. The loss decreases rapidly in the initial epochs, from approximately 0.22 to below 0.03, demonstrating that the network efficiently captures the dominant structure of the data distribution. Following this phase, the curve transitions into a gradual convergence regime and stabilizes near 0.007, exhibiting only minor fluctuations. This trend indicates that the optimization process remains numerically stable and that the selected learning rate, noise schedule, and exponential moving average (EMA) weighting effectively prevent divergence or oscillatory behavior.

The inset focusing on epochs 10 to 50 reveals a smooth downward trend with reduced variance, which typically occurs as the model shifts from learning coarse global structure to refining higher-frequency details. Within denoising diffusion probabilistic models, minimizing noise MSE enhances the estimation of the conditional distribution $p_\theta(x_{t-1} \mid x_t)$. As the model improves its prediction of the injected Gaussian noise at each timestep, the reverse diffusion process reconstructs samples that are increasingly consistent with the observed cloudy input and the learned spatial statistics of cloud-free fields. This loss serves as a training diagnostic rather than a direct reconstruction metric. A low, stable MSE for noise indicates that the model has learned a reliable approximation of the denoising process. However, final reconstruction quality should be assessed using thermal RMSE, bias, and spatial correlation in cloud-covered regions. The observed convergence behavior nonetheless provides indirect evidence that the diffusion framework is well-suited to the synthetic cloud dataset and can learn meaningful temperature priors without overfitting.

Stable convergence is particularly relevant for reconstruction under dense cloud cover, where extensive missing information necessitates reliance on contextual inference. The smooth training dynamics indicate that the model can integrate spatial context and spectral cues in a controlled manner, consistent with the improved robustness observed in subsequent quantitative comparisons. Collectively, the training curve supports the conclusion that the diffusion model can be trained reliably on the proposed dataset and establishes a solid foundation for subsequent conditional inpainting of cloud-contaminated thermal imagery.

4.4.1 Qualitative results

Figure 4.8 presents representative reconstructions in both reflective and thermal domains. Compared with the interpolation baselines and deterministic regression, the diffusion-based approach tends to produce reconstructions with coherent spatial patterns across larger missing regions, consistent with its iterative refinement procedure.

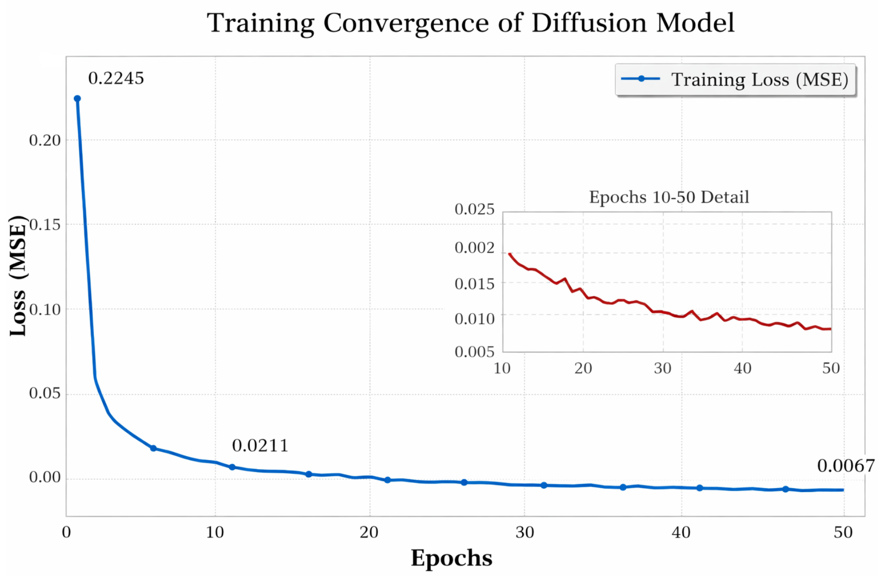


Figure 4.7: Training convergence of the diffusion model. The curve shows the mean squared error (MSE) of noise prediction over 50 epochs.

4.5 Unified Cross-Method Comparison

To enable a consistent comparison across all methods, this section summarizes quantitative results using the evaluation protocol defined in Chapter 3. Thermal reconstruction is evaluated using cloud-only RMSE as the primary accuracy metric, complemented by mean error (bias) and patch-level spatial correlation (Pearson r). Robustness to cloud opacity is assessed via thin/thick stratified cloud-only RMSE. Global consistency is evaluated using the global regional-mean RMSE where the region is defined as the global patch mean. Optical reconstruction quality is evaluated using the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM). PSNR quantifies radiometric fidelity by measuring pixel-wise intensity differences between reconstructed and reference images, while SSIM assesses perceptual image quality by jointly considering luminance, contrast, and structural consistency. Compared with simple error-based metrics, SSIM is more sensitive to structural distortions that are visually and physically relevant in remote sensing imagery. Both metrics are widely adopted for image quality assessment in reconstruction and restoration tasks (Wang and Bovik, 2006).

4.5.1 Pixel-level thermal accuracy (cloud-only)

Table 4.1 shows that the diffusion-based model attains the lowest cloud-only RMSE among all evaluated methods. Its bias is close to zero (+0.05 K), suggesting limited systematic over- or underestimation, whereas the interpolation baselines exhibit a modest cold bias and the U-Net shows a small warm bias. The Pearson

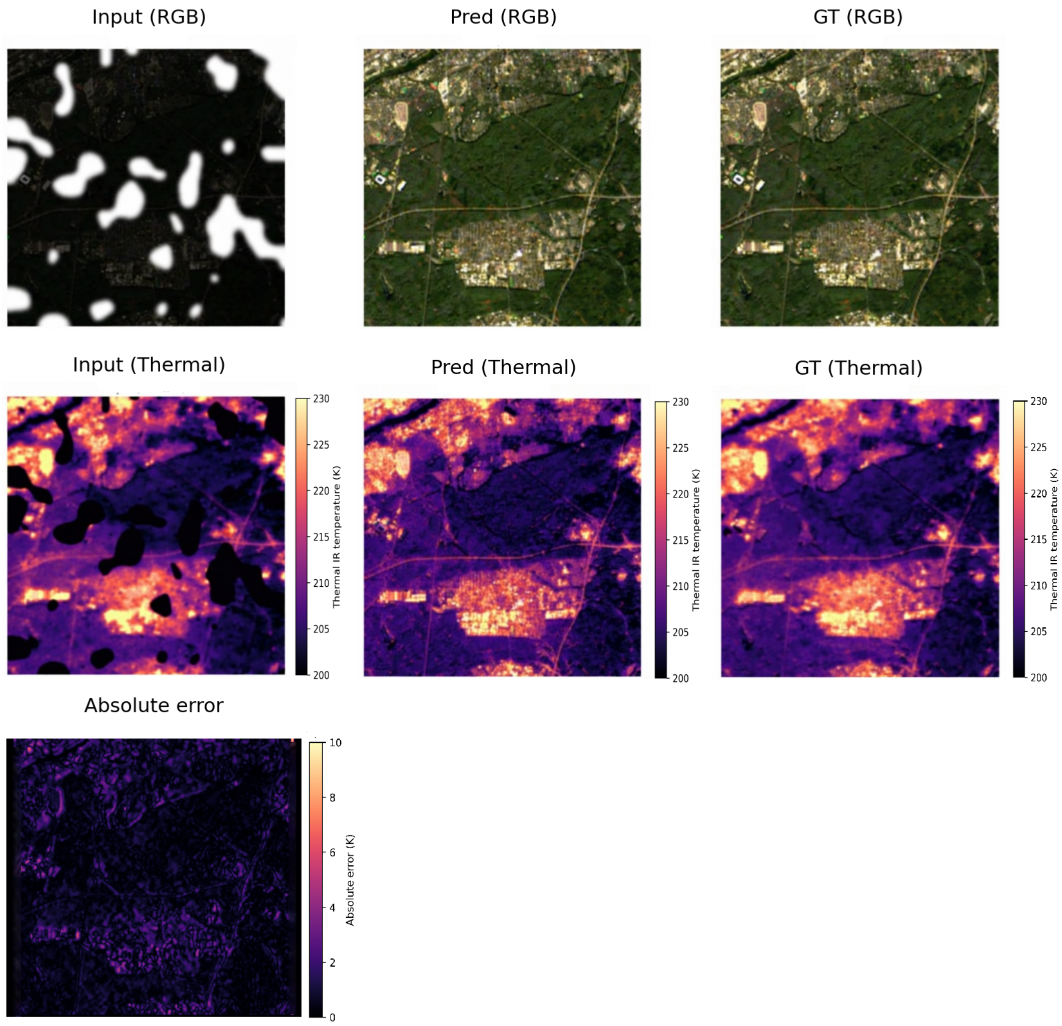


Figure 4.8: **Visual reconstruction results of the proposed framework.** The top row shows the optical domain, including the cloud-contaminated RGB input, the reconstructed (predicted) RGB image, and the corresponding ground-truth (GT) reference. The middle row presents the thermal infrared observations (in Kelvin), including the masked input thermal image, the reconstructed thermal prediction, and the ground-truth thermal reference. The bottom panel illustrates the spatial distribution of the absolute reconstruction error in the thermal domain (K).

Table 4.1: Cloud-only pixel-level thermal reconstruction accuracy (primary). Bias and Pearson r follow Chapter 3.

Method	RMSE (K) ↓	Bias (K)	Pearson r ↑
IDW	1.68	-0.32	0.71
OK	1.54	-0.21	0.74
U-Net	1.75	$+0.18$	0.69
Diffusion	1.48	$+0.05$	0.79

correlation coefficient further indicates that the diffusion-based approach better preserves spatial temperature variations within cloud-covered regions ($r = 0.79$), while the U-Net exhibits a lower correlation ($r = 0.69$), consistent with smoother reconstructions under severe information loss.

4.5.2 Thin- and thick-cloud thermal robustness

Table 4.2: Thermal RMSE (K) stratified by thin and thick cloud (cloud-only). Thin cloud: 20%–65%; thick cloud: $> 65\%$ cloud probability.

Method	$\text{RMSE}_{\text{thin}}$ (K)	$\text{RMSE}_{\text{thick}}$ (K)
IDW	1.12	2.43
OK	1.05	2.21
U-Net	0.96	2.68
Diffusion	0.88	1.94

Thin/thick stratification provides a diagnostic view of robustness under varying cloud opacity. Band-specific thermal behavior can differ between B10 and B11, particularly under thin-cloud conditions, and is therefore summarized using stratified RMSE values. All methods show increased error in thick-cloud regions, reflecting stronger attenuation and reduced information content. The diffusion-based model exhibits the smallest degradation from thin to thick cloud and achieves the lowest RMSE in both subsets (0.88 K thin; 1.94 K thick). In contrast, the U-Net performs competitively under thin clouds (0.96 K) but degrades more strongly under thick clouds (2.68 K), suggesting reduced robustness when the missing region is dominated by severe information loss. For cross-method comparison, thin/thick RMSE values are aggregated across thermal bands (B10 and B11) following the protocol in Chapter 3. Explanations are discussed in the Discussion.

4.5.3 Global thermal consistency and auxiliary optical reconstruction

Global-scale thermal consistency and optical reconstruction metrics are employed as complementary indicators. These indicators are used to evaluate the stability of regional statistics and the preservation of spatial structure. Global thermal accuracy is quantified using the regional mean root-mean-square error (RMSE). Each image patch is represented by its spatial mean temperature, as defined in Chapter 2. Optical metrics, including peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), are computed only for cloud-covered pixels. These metrics serve as auxiliary evidence of structural fidelity.

Table 4.3: Global regional-mean thermal consistency and auxiliary optical reconstruction quality.

Method	$RMSE_{mean,global}$ (K) ↓	PSNR (dB) ↑	SSIM ↑
IDW	0.92	21.4	0.71
OK	0.85	22.1	0.74
U-Net	0.74	28.6	0.89
Diffusion	0.62	29.3	0.91

The regional-mean RMSE reflects the deviation between the reconstructed and reference patch-average temperatures, expressed in Kelvin. All methods exhibit lower errors at the global mean scale compared with pixel-level evaluation, because spatial averaging suppresses local noise and high-frequency residuals. Nevertheless, systematic differences remain. Interpolation baselines show the largest errors, with IDW and OK yielding 0.92 K and 0.85 K, respectively. These results indicate that approaches relying solely on local spatial continuity tend to smooth temperature gradients and underestimate regional contrasts under conditions of extensive cloud gaps, which results in biased patch-level mean values.

Learning-based models provide improved regional consistency. The U-Net reduces the global-mean RMSE to 0.74 K, corresponding to reductions of 0.18 K relative to IDW and 0.11 K relative to OK. The diffusion-based model achieves the lowest error, 0.62 K, representing reductions of 0.30 K (33%) compared with IDW, 0.23 K (27%) compared with OK, and 0.12 K (16%) compared with the U-Net. These results demonstrate that the generative framework more effectively preserves large-scale thermal structure and maintains regional temperature balance. Such behaviour is consistent with the pixel-level accuracy reported in previous sections and suggests improved stability when regional mean thermal values or aggregated thermal indicators represent the primary quantities of interest.

PSNR and SSIM quantify the fidelity of reconstructed spatial texture within cloud-covered areas. Interpolation methods yield relatively low PSNR values (21–22 dB) and SSIM below 0.75, reflecting loss of local detail and edge information. Learning-based approaches achieve substantially higher values, with the U-Net reaching 28.6 dB and 0.89. The diffusion model further improves these metrics to 29.3 dB and 0.91, which indicates more consistent preservation of structural information. Although the numerical differences between U-Net and diffusion are modest, An increase in SSIM generally corresponds to improved edge continuity and spatial coherence in reconstructed regions. Across all metrics, a consistent ranking is observed: interpolation baselines < U-Net < diffusion model. The most substantial improvements are observed in the primary thermal metric, where the diffusion-based approach achieves sub-kelvin regional-mean error, providing reliable estimates for applications based on aggregated LST statistics, such as urban heat island analysis or regional climate monitoring. Optical metrics provide supporting evidence of improved structural reconstruction, but play a secondary role in the evaluation framework.

4.5.4 City-wise Thermal Reconstruction Accuracy

In addition to the aggregated evaluation over the full test set, city-wise Reconstruction errors are reported to assess model generalization across different urban environments. For each study area listed in Section 2.1.1, cloud-only pixel-level RMSE was computed independently following the evaluation protocol defined in Chapter 3. Mean and standard deviation across cities are provided to quantify variability in performance.

Table 4.4: City-wise cloud-only thermal RMSE (Kelvin) for the diffusion-based model.

City	RMSE (K)
Munich	1.42
Graz	1.37
Vienna	1.45
Prague	1.53
Warsaw	1.61
Milan	1.48
Paris	1.39
Brussels	1.36
Mean	1.45
Std. Dev.	0.09

City-wise RMSE values range between 1.36 and 1.61 K, indicating limited variation in reconstruction accuracy across study areas. These results are consistent with the aggregated performance reported in Table 4.1 and do not indicate pronounced city-specific bias.

4.6 Summary of Results

The experimental results reveal distinct trade-offs inherent in each reconstruction paradigm. Interpolation-based methods (IDW and OK) provide computationally efficient baselines but are fundamentally constrained by local spatial continuity assumptions. As illustrated in the qualitative comparisons (Figure 4.3), these methods fail to recover fine-scale thermal features within large, contiguous cloud gaps, thereby reducing contrast and structural fidelity.

The U-Net architecture improves global mean thermal consistency and achieves higher scores in the auxiliary optical metrics (Table 4.3). However, it exhibits a notable lack of robustness under thick-cloud conditions (Table 4.2). Specifically, the U-Net's cloud-only pixel-level RMSE is higher than that of the geostatistical baselines (Table 4.1). This phenomenon is intrinsic to deterministic regression architectures optimized via ℓ_2-type loss functions. Such objectives tend to favor the conditional mean of the data distribution, resulting in spatial smoothing. While this smoothing suppresses noise and improves regional summary statistics, it inevitably filters out local high-frequency thermal variations. In scenarios dominated by severe information loss (thick clouds), the absence of strong observational constraints causes the model to revert to a smooth average, thereby degrading point-wise accuracy and spatial correlation compared to local interpolation.

In contrast, the diffusion-based approach attains the most favorable balance across all primary and auxiliary thermal metrics. It achieves the lowest cloud-only RMSE, maintains the highest spatial correlation ($r = 0.79$), and exhibits minimal systematic bias ($+0.05$ K). The stratified results confirm that its advantage is most pronounced under thick clouds, where it outperforms the U-Net by a significant margin (1.94 K vs. 2.68 K). This suggests that iterative refinement and the use of global latent priors enable the diffusion model to synthesize coherent spatial patterns even when local information is entirely obscured. The elevated root mean square error (RMSE) observed in thick-cloud regions results from several interacting factors, including information loss, data generation processes, and evaluation methodology—thick clouds significantly attenuate surface thermal radiation within the synthetic cloud framework. Opacity-dependent radiance mixing leads to a cooling of several Kelvin. As cloud opacity increases, the observed signal is dominated by cloud-top emission, which limits the availability of surface information for reconstruction. In contrast, thin clouds retain some underlying radiometric structure, enabling the model to utilize actual temperature gradients and spatial texture during inference. Thick clouds also tend to form larger contiguous occlusions within fixed-size patches, thereby reducing the number of valid neighboring pixels and increasing inpainting uncertainty. Thin clouds, however, often appear near cloud edges, where contextual information remains accessible. The synthetic cloud generation process introduces larger temperature perturbations under high-opacity conditions, thereby increasing the potential magnitude of error even when spatial patterns are reconstructed accurately. Additionally, the thin/thick stratification is determined by cloud probability rather than physical optical thickness. As a result, the “thin-cloud” subset may include partially transparent or boundary cases that are inherently easier to reconstruct, while the “thick-cloud” subset may contain complex multilayer structures. The RMSE metric is also sensitive to outliers, so occasional large deviations in heavily occluded regions can disproportionately increase the reported error. Collectively, these factors account for the systematically lower reconstruction accuracy in thick-cloud regions, despite the common perception that thin clouds are more challenging to detect or segment.

Despite these gains, the diffusion model’s superior performance comes at the expense of increased inference latency due to its iterative sampling nature. Nevertheless, the substantial improvements in thermal accuracy and structural stability position diffusion-based models as a highly promising direction for high-fidelity cloud-gap filling in urban thermal applications.

Chapter 5

Discussion

The results show that generative reconstruction can improve the spatial completeness and radiometric consistency of thermal infrared observations affected by cloud cover. The broader aim of this thesis, however, is not only methodological. It is to improve the spatial data basis available for integrated urban planning. Continuous thermal fields are essential for analysing urban surface energy balance patterns, identifying neighbourhood-scale heat exposure, and supporting climate-resilient energy planning in GIS-based workflows.

In practice, reconstructed temperature fields can be combined with building inventories, land-use data, demographic information, and urban green infrastructure maps. This allows planners to identify thermally vulnerable neighbourhoods, cooling-demand hotspots, and areas where greening or building retrofits may be most effective. In this way, reconstructed thermal data can move from exploratory remote sensing analysis toward practical decision-support in urban energy planning and climate adaptation.

These benefits are especially relevant in cities with frequent cloud cover, where consistent thermal observations are difficult to obtain. Reducing spatial gaps improves the continuity of urban temperature monitoring, supports more reliable SUHI indicators, and strengthens GIS-based modelling workflows used in integrated urban planning.

5.1 Implications for GIS Applications and Urban Energy Planning

The principal aim of reconstructing spatially continuous thermal infrared temperature fields is to generate gap-free datasets that can be integrated with additional spatial information in GIS-based analyses.

In urban energy planning scenarios, reconstructed LST fields can be combined with city level GIS layers, such as building footprints, cadastral parcels, land use maps, population distribution, and green infrastructure inventories, to delineate vulnerable neighbourhoods or to identify spatially resolved peak cooling demands that may affect local energy infrastructure needs. This spatially explicit perspective enables the targeted prioritization of mitigation and adaptation measures, such as urban greening, surface material modification, or the deployment of appropriate technical interventions (e.g., passive and active cooling strategies, blue–green infrastructure). At neighbourhood scale, such analyses may support the protection of vulnerable population groups and contribute to maintaining thermal comfort and urban liveability. More broadly, gap free LST time series enable central analyses to urban remote sensing studies, including the investigation of relationships between urban morphology, land cover composition, and surface temperature dynamics. With regard to the Topview research project, integrating reconstructed LST fields with multi-source spatial datasets provides a reliable, gap-free thermal data basis for high-resolution urban climate assessment, long-term monitoring, and evidence-based decision making in integrated urban energy planning. This approach strengthens the evidence base for climate-resilient and spatially differentiated urban energy planning.

5.2 Limitations

While the proposed framework demonstrates consistent performance under the defined evaluation protocol, several limitations related to the experimental design should be acknowledged.

A primary limitation arises from the reliance on synthetically generated cloud-contaminated data. Synthetic cloud masking enables controlled and reproducible assessment of reconstruction performance under cloud-induced information loss. However, such an approach cannot fully capture the complexity of real atmospheric conditions. In particular, effects associated with heterogeneous cloud structures, multi-layer cloud coverage, and spatially varying thermal attenuation are only approximated. As a result, the reported metrics should be interpreted as reflecting relative algorithmic performance under controlled degradation rather than guaranteed accuracy across all real-world meteorological scenarios.

In addition, the current evaluation does not explicitly quantify statistical uncertainty. All learning-based models are trained and evaluated using a single fixed data partition, which ensures methodological consistency across baselines but limits insight into variability arising from random initialization or data sampling. This consideration is especially relevant for diffusion-based models, where stochastic sampling during inference may introduce variability in fine-scale texture reconstruction. Accordingly, the reported performance differences should be regarded as indicative rather than statistically conclusive.

Another limitation concerns the data partitioning strategy. The dataset was randomly divided at the patch level, without explicitly accounting for potential spatial autocorrelation between samples originating from the same city or satellite overpass. Consequently, training and test subsets may contain spatially adjacent or temporally correlated patches, which could introduce a degree of data leakage and lead to optimistic performance estimates. In remote sensing applications, spatial dependence is often non-negligible, and evaluation protocols that separate training and testing data by geographic region—such as training on one city and evaluating on another—provide a more stringent assessment of cross-regional generalization. The current experimental design therefore primarily reflects within-distribution performance and does not fully quantify the model's ability to generalize across distinct urban or climatic contexts.

The accuracy of reconstruction is fundamentally dependent on the quality of cloud detection. Cloud masking in thermal infrared imagery presents significant challenges. This difficulty arises because cloud-top temperatures can be similar to those of cold land surfaces or water bodies. Additionally, thin or warm clouds may not be detected in thermal bands alone. Misclassification of thin clouds, cloud edges, or bright urban surfaces in existing Landsat cloud masks can introduce uncertainty into both reconstruction targets and evaluation metrics. Surfaces in existing Landsat cloud masks can propagate uncertainty into both reconstruction targets and evaluation metrics.

Finally, the improved spatial coherence achieved by generative reconstruction comes at the cost of increased computational complexity. Compared with interpolation methods and deterministic regression, the diffusion-based approach requires substantially higher inference time. While this overhead is acceptable in offline analysis and retrospective studies, it may constrain applicability in time-sensitive or resource-limited operational settings.

5.3 Conclusion

The research objectives formulated in Chapter 1 have been effectively addressed through both methodological innovation and applied validation. By establishing a controlled evaluation framework centered on synthetically cloud-contaminated thermal infrared imagery, this research enabled a systematic and reproducible benchmarking of geostatistical interpolation, deterministic convolutional architectures, and diffusion-based generative models across varying gradients of information loss. Such rigorous experimentation extends earlier paradigms in satellite LST reconstruction, which have traditionally pivoted on more constrained spatial or temporal interpolation strategies (Crosson et al., 2012; Stathopoulou and Cartalis, 2009a).

Comparative performance metrics indicate that while classical interpolation remains competitive under sparse cloud cover, its inherent reliance on local spatial autocorrelation fails to resolve fine-scale urban thermal heterogeneities. Although Regression U-Net models—representative of modern convolutional benchmarks (Ronneberger et al., 2015)—improve global structural consistency, they remain prone to spatial smoothing as information loss intensifies. Conversely, diffusion-based reconstruction, leveraging recent paradigms in probabilistic generative modelling (Ho et al., 2020), demonstrates a superior capacity for reconciling point-wise thermal fidelity with spatial coherence, particularly under opaque cloud conditions. These findings directly validate the robustness of generative approaches within the complex context of fragmented urban environments.

Beyond algorithmic performance, this study contextualizes diffusion-based reconstruction within the broader trajectory of remote sensing applications. Despite the recognized necessity of thermal infrared observations for urban climate analysis and energy planning (Voogt and Oke, 2003; Li et al., 2013), their operational utility is frequently compromised by persistent atmospheric interference. By achieving sub-kelvin regional accuracy at Landsat resolution, this research suggests that generative reconstruction can effectively bridge the gap between theoretical modelling and operational urban monitoring.

Consequently, the core contribution of this thesis lies in the synthesis of generative deep learning with long-standing challenges in urban remote sensing. The results underscore that diffusion-based models, when integrated with physically consistent LST retrieval and GIS frameworks, provide a viable pathway for generating spatially continuous thermal datasets. Such high-fidelity products are essential for advancing climate-resilient urban energy planning and facilitating evidence-based environmental decision making.

5.4 Future Work

Several directions for future research naturally follow from the limitations identified in this study.

An important next step is to evaluate the performance of the reconstruction under real cloud conditions. Validating the proposed methods using observed cloud-contaminated imagery and, where available, in-situ reference measurements would provide a more comprehensive assessment of generalization beyond synthetic benchmarks. Combining controlled synthetic experiments with real-world observations may further strengthen confidence in practical deployment scenarios.

Future work could also focus on more detailed spectral analysis of thermal reconstruction. In particular, explicitly examining band-specific behavior for Landsat-8 thermal bands B10 and B11 may help clarify how differences in atmospheric absorption and noise characteristics influence reconstruction accuracy under thin and thick cloud conditions. Such analysis could provide additional insight into the physical consistency of reconstructed land surface temperature fields.

Another important direction is the incorporation of a physically consistent Land Surface Temperature (LST) retrieval framework into the reconstruction pipeline. The thermal bands are currently reconstructed at the radiometric level without explicitly accounting for surface emissivity or atmospheric effects. Although this formulation is appropriate for relative performance comparison, it does not fully represent the physical processes that govern surface thermal emission. Future work should therefore integrate established LST retrieval procedures, including emissivity estimation and atmospheric correction, either as a dedicated post-processing step or within a physics-informed modeling framework. Coupling the current data-driven reconstruction approach with radiative transfer–based methods would improve physical consistency and enhance the suitability of the reconstructed thermal fields for quantitative applications, particularly in studies requiring cross-regional comparability and surface energy balance analysis.

Future research should extend the integrated analysis of reconstructed LST fields with detailed land use and urban morphology datasets within a GIS-based urban energy planning framework. Relating spatial temperature patterns to land-cover composition, building density, and vegetation fraction would improve the attribution of urban heat to specific structural and functional characteristics of the built environment and clarify their implications for cooling demand. Such analyses are directly supported by the Topview project, which provides access to consistent multi-layer spatial datasets. Coupling reconstructed LST with land-use information would strengthen evidence-based urban heat mitigation planning and enable more reliable spatial modeling of risk assessments at the neighborhood level for vulnerable populations and infrastructure measures.

Another direction for future work is the development of practical GIS-based indicators derived from the reconstructed LST fields. While this study demonstrates that spatially continuous thermal observations can be obtained under cloud contamination, additional work is needed to translate these data into indicators that are directly usable in urban planning. Examples include neighbourhood-scale heat exposure indices, proxies for cooling demand, and measures of thermal vulnerability that combine reconstructed LST with land-use, building density, and demographic information. Establishing transparent GIS workflows for indicator calculation, spatial aggregation, and uncertainty reporting would improve comparability across cities and make the reconstruction results more accessible for decision-support applications in urban climate adaptation and energy planning. Finally, improving the computational efficiency of diffusion-based reconstruction remains a relevant research direction. Approaches such as accelerated sampling schemes, model distillation, or hybrid frameworks that combine deterministic regression with generative refinement could help reduce inference time while preserving the spatial detail required for urban thermal analysis.

Bibliography

Austrian Research Promotion Agency (FFG) (2024). *TopView: Deep learning and remote sensing for urban climate and energy planning*. Accessed: 2026-01-27. url: `https://projekte.ffg.at/projekt/5124572`.

Bertalmio, M., G. Sapiro, V. Caselles, and C. Ballester (2000). „Image Inpainting“. In: *Proceedings of SIGGRAPH*, pp. 417–424.

Claverie, M., J. Ju, J. G. Masek, J. L. Dungan, E. F. Vermote, J.-C. Roger, S. Skakun, and C. Justice (2018). „The Harmonized Landsat and Sentinel-2 (HLS) surface reflectance data set“. In: *Remote Sensing of Environment* 219, pp. 145–161. doi: `10.1016/j.rse.2018.09.002`.

Crosson, W. L., A. S. Limaye, C. A. Laymon, and D. A. Quattrochi (2012). „A daily merged MODIS Aqua–Terra land surface temperature dataset“. In: *Remote Sensing of Environment* 119, pp. 315–324.

Czerkawski, M., R. Atkinson, C. Michie, and C. Tachtatzis (2023). „SatelliteCloudGenerator: Controllable Cloud and Shadow Synthesis for Multi-Spectral Optical Satellite Images“. In: *Remote Sensing* 15.17, p. 4138. doi: `10.3390/rs15174138`.

Ho, J., A. Jain, and P. Abbeel (2020). „Denoising Diffusion Probabilistic Models“. In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 6840–6851.

Hochreiter, S. and J. Schmidhuber (1997). „Long Short-Term Memory“. In: *Neural Computation* 9.8, pp. 1735–1780. doi: `10.1162/neco.1997.9.8.1735`.

Jeppesen, J. H., R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard (2019). „A cloud detection algorithm for satellite imagery based on deep learning“. In: *Remote Sensing of Environment* 229, pp. 247–258. doi: `10.1016/j.rse.2019.04.033`.

LeCun, Y., Y. Bengio, and G. Hinton (2015). „Deep learning“. In: *Nature* 521.7553, pp. 436–444. doi: `10.1038/nature14539`.

Li, Z.-L., B.-H. Tang, H. Wu, H. Ren, G. Yan, Z. Wan, I. F. Trigo, and J. A. Sobrino (2013). „Satellite-derived land surface temperature: Current status and perspectives“. In: *Remote Sensing of Environment* 131, pp. 14–37. doi: `10.1016/j.rse.2012.12.008`.

Main-Knorn, M., B. Pflug, J. Louis, V. Debaecker, U. Müller-Wilm, and F. Gascon (2017). „Sen2Cor for Sentinel-2“. In: *Image and Signal Processing for Remote Sensing XXIII*. Vol. 10427. SPIE, pp. 37–48. doi: `10.1117/12.2278218`.

Montanaro, M., A. Gerace, A. Lunsford, and D. Reuter (2014). „Radiometric Calibration of the Landsat 8 Thermal Infrared Sensor“. In: *Remote Sensing* 6.9, pp. 8803–8821. doi: `10.3390/rs6098803`.

NASA Goddard Space Flight Center (2026). *Harmonized Landsat–Sentinel (HLS) Product*. Accessed January 2026. url: https://hls.gsfc.nasa.gov/.

Neteler, M. (2010). „Estimating daily land surface temperatures in mountainous environments by reconstructed MODIS LST data“. In: *Remote sensing* 2.1, pp. 333–351.

Ronneberger, O., P. Fischer, and T. Brox (2015). „U-Net: Convolutional Networks for Biomedical Image Segmentation“. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

Schuster, M. and K. K. Paliwal (1997). „Bidirectional Recurrent Neural Networks“. In: *IEEE Transactions on Signal Processing* 45.11, pp. 2673–2781. doi: 10.1109/78.650093.

Song, Y. and S. Ermon (2021). „Score-Based Generative Modeling through Stochastic Differential Equations“. In: *International Conference on Learning Representations (ICLR)*.

Stathopoulou, M. and C. Cartalis (2009a). „Downscaling AVHRR land surface temperatures for improved surface urban heat island intensity estimation“. In: *Remote Sensing of Environment* 113.12, pp. 2592–2605.

Stathopoulou, M. and C. Cartalis (2009b). „Downscaling AVHRR land surface temperatures for improved surface urban heat island intensity estimation“. In: *Remote Sensing of Environment* 113.12, pp. 2592–2605.

Voogt, J. A. and T. R. Oke (2003). „Thermal remote sensing of urban climates“. In: *Remote Sensing of Environment* 86.3, pp. 370–384.

Wang, Z. and A. C. Bovik (2006). *Modern Image Quality Assessment: From Error Visibility to Structural Similarity*. Morgan & Claypool.

Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli (2004). „Image Quality Assessment: From Error Visibility to Structural Similarity“. In: *IEEE Transactions on Image Processing* 13.4, pp. 600–612. doi: 10.1109/TIP.2003.819861.

Zhang, J., X. Li, L. Li, P. Sun, and X. Su (2020). „Lightweight U-Net for cloud detection of visible and thermal infrared remote sensing images“. In: *Optical and Quantum Electronics* 52.11, pp. 1–17. doi: 10.1007/s11082-020-02500-8.

Zhu, Z. and C. E. Woodcock (2012). „Object-based cloud and cloud shadow detection in Landsat imagery“. In: *Remote Sensing of Environment* 118, pp. 83–94. doi: 10.1016/j.rse.2011.10.028.

.5 Appendix A: Reproducibility and Implementation Details

This appendix documents the implementation details required to reproduce the experiments presented in this thesis. The procedures described below correspond to the experimental setup used to generate all results reported in Chapters 3 and 4.

.5.1 Software and Hardware Environment

All experiments were implemented in Python 3.10 using PyTorch 2.0.1. Training was conducted in the Google Colab Pro environment with GPU acceleration, typically using NVIDIA A100 hardware or equivalent configurations. Additional libraries included NumPy, SciPy, rasterio, scikit-learn, and GDAL.

Random seeds were fixed across Python, NumPy, and PyTorch wherever possible to ensure reproducibility.

.5.2 Data Sources and Preprocessing

The dataset consists of multi-spectral Landsat-8 patches derived from the Harmonized Landsat and Sentinel-2 (HLS) L30 products. Each patch has a spatial size of 256×256 pixels and includes:

- Six reflectance bands (B02–B07; VNIR/SWIR),
- Two thermal infrared bands (B10 and B11),
- One auxiliary Fmask cloud classification band,
- A continuous soft cloud mask.

Thermal bands were converted to Kelvin brightness temperature following the harmonization procedure described in Chapter 2. Invalid values were removed and pixel values were clipped to physically plausible ranges.

Cloud-free patches were defined as observations with cloud fraction $p_{\text{cloud}} < 0.1$. Synthetic cloud masks were generated using the modified SatelliteCloudGenerator procedure described in Section 2.2, producing paired clean and cloud-contaminated samples.

The dataset was divided into training, validation, and test subsets using an 80/10/10 split.

.5.3 Model Architectures

Regression U-Net. The U-Net followed a symmetric encoder–decoder structure with four resolution levels. Each level contained two 3×3 convolutional layers with nonlinear activation. The model was trained using mean squared error loss and the AdamW optimizer.

Diffusion-Based Model. The diffusion reconstruction model was implemented as a conditional DDPM framework using a U-Net noise-prediction backbone. Training used a linear noise schedule with 1,000 diffusion steps. Inference employed a RePaint-style conditional sampling procedure.

.5.4 Training Procedure

All models were trained on paired cloudy/clean patches with cloud masks provided as auxiliary inputs. Data augmentation included random rotations and flips.

Models were trained until validation performance stabilized, typically after 30–50k iterations. Early stopping was applied based on validation loss.

Training data were shuffled each epoch. No temporal information was used.

.5.5 Evaluation Metrics

Evaluation followed the protocol defined in Chapter 3. Pixel-level metrics were computed only on cloud-covered pixels.

Reported metrics include:

- Cloud-only RMSE (Kelvin),
- Mean Bias Error,
- Global regional-mean RMSE,
- Patch-level spatial correlation.

Thin- and thick-cloud subsets were defined using cloud probability thresholds of 20% and 65%, consistent with Sen2Cor conventions.

.5.6 Code and Data Availability

Due to ongoing project collaboration and data licensing constraints, the source code and processed datasets cannot be released publicly at the time of thesis submission. The repositories will be made publicly available after thesis evaluation.

Additional implementation details can be provided upon reasonable academic request.

Appendix A

Appendix B: Synthetic Cloud Generation

The synthetic cloud generation procedure described in Section 2.2 is implemented using multi-scale random opacity fields combined with physically consistent thermal cooling.

```python
import numpy as np
from scipy.ndimage import gaussian_filter

def generate_cloud_opacity(shape=(256,256), seed=0):
    rng = np.random.default_rng(seed)
    noise = rng.normal(size=shape)

    # smooth multi-scale cloud structure
    cloud = gaussian_filter(noise, sigma=12)
    cloud = (cloud - cloud.min()) / (cloud.max() - cloud.min())
    return cloud
```

Listing A.1: Cloud opacity generation

Thermal synthesis follows the radiance-space cooling approximation outlined in Chapter 2:

```python
def synthesize_thermal(T_surface, alpha):
    delta_T = np.random.uniform(5.0, 15.0)
    T_cloud = T_surface - delta_T * alpha
    alpha_p = alpha ** 1.5
    T_out = (1 - alpha_p) * T_surface + alpha_p * T_cloud
    return T_out
```

Listing A.2: Thermal cloud synthesis

Appendix B

Appendix C: Diffusion Model Configuration

The diffusion-based reconstruction model is implemented using a conditional DDPM framework with a U-Net backbone.

```
from dataclasses import dataclass

@dataclass
class DiffConfig:
    image_size = 256
    batch_size = 4
    learning_rate = 1e-4
    train_steps = 40000
    diffusion_steps = 1000
    beta_schedule = "linear"
```

Listing B.1: Diffusion configuration

Example initialization of the noise-prediction network and scheduler:

```
from diffusers import UNet2DModel, DDPMScheduler

model = UNet2DModel(
    sample_size=256,
    in_channels=8, # VNIR/SWIR + B10/B11 + mask
    out_channels=8,
    layers_per_block=2,
    block_out_channels=(128, 256, 256),
)

scheduler = DDPMScheduler(
    num_train_timesteps=1000,
    beta_schedule="linear"
)
```

Listing B.2: UNet backbone initialization

Appendix C

Appendix D: Conditional RePaint Sampling

Cloud removal is formulated as a conditional inpainting problem. Known pixels are enforced at each reverse diffusion step.

```python
def repaint_step(pred, observed, mask):
    return mask * pred + (1 - mask) * observed

for t in reversed(range(T)):
    noise_pred = model(x_t, t)
    x_t = scheduler.step(noise_pred, t, x_t).prev_sample
    x_t = repaint_step(x_t, observed_image, mask)
```

Listing C.1: RePaint-style inpainting

Appendix D

Appendix E: Cloud-only RMSE Evaluation

All accuracy metrics in Chapter 4 are computed exclusively over cloud-covered pixels, as defined in Chapter 3.

```python
import numpy as np

def cloud_only_rmse(pred, ref, mask):
    diff = pred[mask == 1] - ref[mask == 1]
    return np.sqrt(np.mean(diff**2))
```

Listing D.1: Cloud-only RMSE

Appendix E

Appendix F: Patch Dataset Loader

The following example shows the dataset loading pipeline used during training.

```python
import torch
from torch.utils.data import Dataset

class HLSPatchDataset(Dataset):
    def __init__(self, cloudy, clean, mask):
        self.cloudy = cloudy
        self.clean = clean
        self.mask = mask

    def __len__(self):
        return len(self.cloudy)

    def __getitem__(self, idx):
        x = torch.tensor(self.cloudy[idx]).float()
        y = torch.tensor(self.clean[idx]).float()
        m = torch.tensor(self.mask[idx]).float()
        return x, y, m
```

Listing E.1: HLS patch dataset loader

Declaration

“ Personal declaration: I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis. We hereby declare that we completed this paper on my own using only the sources listed in the indexes or in the comments. Subject to other restrictive requirements by the responsible supervisor of this work, the following applies to the use of technical instruments that at least partially autonomously generate text, data, code or image material: The essentially unchanged adoption of such content must be marked. The labeling requirement is to be fulfilled on the one hand by clearly marking all affected parts of the work graphically and on the other hand by listing all the instruments specifically used in the directories. We also confirm that this paper has not already been used for any other assessment and that we will not use it in this way in the future. ”

Zurich, March 4, 2026

Zurich, March 4, 2026
Yaze Xue

Matriculation number: 23-741-523

Zurich, March 4, 2026

Zurich, March 4, 2026
