

School of Sciences | Department of Computer Science and Engineering

Replace this with the title of the thesis

Replace this with the subtitle (if applicable, if not delete the line)

Master of Science

in enter your discipline

Enter you name here

© Enter the month and year you are submitting

# Acknowledgements

# Abstract

*Write this last and remember what it is about = To provide a brief statement (no more than 1 page long) regarding the work performed. The statement should not go into too many specifics but should provide the reader with enough information to have a good idea of what the project is about.)*

# Table of Contents

[Acknowledgements 2](#_Toc496781209)

[Abstract 3](#_Toc496781210)

[Table of Contents 4](#_Toc496781211)

[1 Introduction 6](#_Toc496781212)

[1.1 Overview 6](#_Toc496781213)

[1.2 Aims and objectives 6](#_Toc496781214)

[1.3 Structure of thesis 7](#_Toc496781215)

[1.4 Summary 8](#_Toc496781216)

[2 Project Scope 9](#_Toc496781217)

[2.1 Introduction 9](#_Toc496781218)

[2.2 Title 1 **Error! Bookmark not defined.**](#_Toc496781219)

[2.3 Title 2 9](#_Toc496781220)

[2.4 Title 3 **Error! Bookmark not defined.**](#_Toc496781221)

[2.5 Summary 10](#_Toc496781222)

[3 Analysis and Design 12](#_Toc496781223)

[3.1 Introduction 12](#_Toc496781224)

[3.2 Title 1 **Error! Bookmark not defined.**](#_Toc496781225)

[3.3 Title 2 **Error! Bookmark not defined.**](#_Toc496781226)

[3.4 Title 3 **Error! Bookmark not defined.**](#_Toc496781227)

[3.5 Summary 20](#_Toc496781228)

[4 Implementation and Testing 21](#_Toc496781229)

[4.1 Introduction 21](#_Toc496781230)

[4.2 Title 1 **Error! Bookmark not defined.**](#_Toc496781231)

[4.3 Title 2 21](#_Toc496781232)

[4.4 Title 3 **Error! Bookmark not defined.**](#_Toc496781233)

[4.5 Summary **Error! Bookmark not defined.**](#_Toc496781234)

[5 Conclusions and Future Work 28](#_Toc496781235)

[5.1 Introduction 28](#_Toc496781236)

[5.2 Title 1 **Error! Bookmark not defined.**](#_Toc496781237)

[5.3 Title 2 28](#_Toc496781238)

[5.4 Title 3 31](#_Toc496781239)

[5.5 Summary 31](#_Toc496781240)

[Bibliography 33](#_Toc496781241)

[Appendices 34](#_Toc496781242)

[Appendix A 34](#_Toc496781243)

[Appendix B 35](#_Toc496781244)

[Appendix C 36](#_Toc496781245)

[Installation Manual 37](#_Toc496781246)

[Requirements 37](#_Toc496781247)

[Installation procedure 37](#_Toc496781248)

[Configuration 37](#_Toc496781249)

[User Manual 38](#_Toc496781250)

# Introduction

## Overview

Understanding the formation and evolution of galaxies at high redshifts is a fundamental objective in astrophysics. High-redshift galaxies, observed as they were billions of years ago, provide crucial insights into the early stages of the universe, including the initial burst of star formation, the buildup of galactic structures, and the growth of supermassive black holes. Observational data from space-based telescopes, such as the Hubble Space Telescope and the recently launched James Webb Space Telescope (JWST), have advanced our understanding of these early cosmic epochs. However, the inherent challenges of observing distant galaxies, including their faintness, dust obscuration, and the limited wavelength range of available instruments, mean that much of this epoch remains poorly understood.

To address these limitations, machine learning, and particularly deep generative models, have emerged as powerful tools for extrapolating and simulating the properties of galaxies based on available data. Generative Adversarial Networks (GANs) and Diffusion Models have proven effective in generating synthetic data that closely resembles real-world observations. GANs and diffusion models have been applied to bridge observational gaps and simulate high-redshift galaxy formations (Efstathiou et al., 2006; Stalevski et al., 2006; Siebenmorgen & Krugel, 2006). These models enable researchers to predict missing information, simulate various scenarios, and explore the conditions of the early universe in a controlled, computational environment.

Diffusion models, a class of probabilistic generative models, offer significant advantages over traditional GANs. By gradually transforming random noise into structured data, they have shown superior stability during training and better coverage of the underlying data distribution, as highlighted in recent studies in time series forecasting and image generation (Song et al., 2021; Ho et al., 2020)​. Their ability to generate high-fidelity outputs with more robust training dynamics makes them a promising tool for simulating high-redshift galaxy formations by generating synthetic spectra that can fill in missing observational data (Meijer et al., 2023; Yang et al., 2023 by filling gaps left by observational limitations, and improving our understanding of early galaxy evolution.

This thesis aims to explore the application of diffusion models in predicting and reconstructing the spectral energy distributions (SEDs) of high-redshift galaxies, comparing their performance with GANs. By leveraging advanced deep learning techniques, this study seeks to generate accurate synthetic spectra, enhancing our ability to study galaxies that are otherwise difficult to observe directly.

## Aims and objectives

The primary aim of this thesis is to develop and evaluate the effectiveness of diffusion models for generating synthetic spectral data of high-redshift galaxies. This involves creating a model capable of producing realistic and high-fidelity spectra that reflect the physical properties of galaxies in the early universe. A secondary aim is to conduct a comparative analysis between diffusion models and GANs, focusing on their ability to generate accurate and diverse galaxy spectra, as well as their training efficiency and stability.

The specific objectives of this thesis are as follows:

1. **Design and implement a diffusion model** tailored for generating synthetic galaxy spectra based on multi-wavelength observational data.
2. **Train and validate the diffusion model** using existing datasets, ensuring it can accurately replicate known spectral properties and predict missing information.
3. **Develop a GAN-based model** to serve as a benchmark for evaluating the performance of the diffusion model.
4. **Conduct a comparative analysis** of the two models, focusing on metrics such as fidelity, diversity, computational efficiency, and robustness.
5. **Interpret the results** to determine the strengths and limitations of each approach and provide insights into the most effective generative model for astrophysical applications.

## Structure of thesis

The thesis is structured as follows:

1. **Introduction**: Provides an overview of the research context, the importance of studying galaxy formation at high redshifts, and the role of generative models in addressing data limitations. It also states the thesis objectives and outlines the structure of the document.
2. **Literature Review**: Reviews current theories and findings related to galaxy formation and evolution at high redshifts. It also surveys existing generative models, such as GANs and Diffusion Models, and their applications in various domains, with a focus on astrophysical research. Previous works leveraging generative models to reconstruct and predict galaxy spectra are discussed in detail.
3. **Theoretical Background**: Introduces the mathematical foundations of diffusion models, covering key concepts like denoising diffusion probabilistic models (DDPMs) and Score-Based Generative Models (SGMs). It also explains GAN architectures and common challenges, along with an introduction to astrophysical concepts relevant to the study, such as spectral energy distributions and multi-wavelength data.
4. **Methodology**: Describes the datasets used in the study, including preprocessing techniques and data augmentation strategies. It details the design and implementation of the diffusion model, parameter tuning, and training procedures. The GAN architecture is presented as a baseline, and the evaluation metrics used for comparison are outlined.
5. **Experimental Results**: Presents the performance results of the diffusion model and GAN, including both quantitative metrics and qualitative visual comparisons. Specific case studies are highlighted to demonstrate the strengths of each model.
6. **Discussion**: Interprets the experimental findings, discussing the implications for future research in galaxy formation and generative modeling. It addresses the strengths and weaknesses of the models, identifies potential areas for improvement, and suggests future research directions.
7. **Conclusion and Future Work**: Summarizes the key contributions of the thesis, reaffirming the advantages of diffusion models for astrophysical applications. It also proposes future research directions, including model enhancements and broader applications in other areas of astrophysics.
8. **References**: Lists all academic literature, datasets, and software tools cited throughout the thesis.
9. **Appendices**: Provides additional figures, tables, technical details, and supplementary materials that support the main content of the thesis.

## Summary

This introduction has outlined the motivation, aims, and objectives of this thesis, which focuses on the application of diffusion models to simulate and predict the properties of high-redshift galaxies. By developing a robust generative model that can generate accurate synthetic spectra, this research seeks to fill gaps in observational data and enhance our understanding of galaxy formation during the early epochs of the universe. The thesis will also provide a detailed comparative analysis between diffusion models and GANs, highlighting the advantages and limitations of each approach, and ultimately contributing to the advancement of generative modelling techniques in astrophysics.

# Project Scope

## Introduction

The scope of this project is to investigate the use of diffusion models for generating synthetic spectral data of high-redshift galaxies and to perform a comparative analysis with Generative Adversarial Networks (GANs). Given the challenges of observing distant galaxies—such as faintness, dust obscuration, and limited wavelength coverage—the project aims to leverage machine learning techniques to address these limitations by generating accurate synthetic spectra. The study focuses on the design, development, and evaluation of these generative models, with the goal of improving our understanding of galaxy formation in the early universe. This chapter details the specific components of the project, the processes involved, and the expected outcomes.

## Data Production and Analysis

This project utilizes the SMART framework to generate spectral energy distributions (SEDs) of galaxies by combining flux contributions from four astrophysical components: starburst, AGN, spheroid, and polar dust. Each component is parameterized based on pre-computed radiative transfer simulations, as described in Efstathiou et al. (2000, 2022) and Rissakia et al. (2024)​(stae1141)​. These models enable accurate modeling of galaxy SEDs across diverse environments.

The total spectrum is calculated by summing the individual contributions of these components. Photometric data is derived by integrating high-resolution spectra over observational filter transmission curves, following methodologies from Efstathiou & Rowan-Robinson (1995) and other CYGNUS-based modeling approaches.Using the U5150 filter and a redshift of 1.89, photometry was generated across a wide range of wavelengths, from ultraviolet to submillimeter, producing data for 10,000 galaxies. This synthetic dataset facilitates detailed exploration of galaxy properties and spectral characteristics.

## Development and Implementation of Diffusion Models

A significant part of this project is the development and implementation of diffusion models specifically tailored to generate high-fidelity synthetic spectra for high-redshift galaxies. This work utilizes the TSDiff framework, an unconditional diffusion model adapted from recent advancements in time series forecasting by Kollovieh et al. (2023). The TSDiff model employs a self-guidance mechanism to condition the generative process during inference, enabling it to handle tasks such as forecasting, imputation, and synthetic data generation without requiring auxiliary networks or task-specific training.

The model development process includes preparing the data by collecting spectral energy distributions (SEDs) from various telescopic surveys and addressing astrophysical challenges such as missing data points. Advanced training and optimization techniques are employed to ensure the diffusion model can accurately replicate observed galaxy spectra and extrapolate missing information, leveraging the implicit probability density learned during training. The aim is to create a robust generative model that simulates the physical properties of galaxies, offering a powerful tool for analysing characteristics of high-redshift galaxies that are otherwise difficult to observe directly.

## Benchmarking with Generative Adversarial Networks (GANs)

In addition to developing a diffusion model, this project involves creating a benchmark using a GAN-based approach. This will provide a standard for comparing the performance of the diffusion model. The GAN architecture will be adapted from existing models proven effective in related fields, with modifications to accommodate the specific requirements of generating spectra for high-redshift galaxies.

This project investigates the use of diffusion models for generating synthetic spectral data of high-redshift galaxies and performing comparative analysis with Generative Adversarial Networks (GANs). GANs, widely used in image and signal processing, are known for capturing intricate details in synthetic data but can be prone to issues like mode collapse. Diffusion models, however, are inherently stable due to their structured noise injection mechanism and have shown higher fidelity in capturing complex spectral data across wavelengths (Dhariwal & Nichol, 2021; Ho et al., 2020)

Training the GAN model will involve similar processes as the diffusion model, ensuring that it is trained on the same datasets to maintain a consistent basis for comparison. The evaluation will focus on common issues associated with GANs, such as mode collapse and training instability, to highlight how these challenges impact performance. Through this comparison, the project will assess key performance indicators like fidelity, diversity, and computational efficiency, providing a clear understanding of each model's strengths and limitations.

## Evaluation Metrics and Experimental Analysis

A thorough evaluation of both the diffusion model and the GAN is crucial to this project. The evaluation process begins with the selection of appropriate metrics to assess the quality of the generated spectra. Metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Continuous Ranked Probability Score (CRPS) will be used to quantify the models' performance. Additionally, visual assessments will be conducted to compare the synthetic spectra against real observational data, allowing for a more nuanced understanding of each model's ability to replicate real-world patterns.

The project involves designing a comprehensive experimental setup to test the models under various conditions. Experiments will include training the models on different datasets, adjusting noise levels, and applying data augmentation techniques to evaluate the robustness and generalizability of the models. Specific case studies will be used to demonstrate the practical utility of the models, such as predicting missing data in known galaxy spectra or simulating the spectral energy distributions of hypothetical high-redshift galaxies. These scenarios will illustrate how the models can be applied to real-world problems in astrophysics, providing insights that can guide future research and observational strategies.

## Summary

This chapter has outlined the scope of the project, emphasizing the development, implementation, and evaluation of diffusion models for generating synthetic spectra of high-redshift galaxies. The project also involves creating a GAN-based benchmark to provide a comparative analysis. Through comprehensive testing and evaluation, the project aims to determine the most effective generative model for astrophysical applications, thereby enhancing our ability to study galaxies in the early universe. The outcomes of this research will contribute to the broader field of astrophysics by offering new methods to simulate and analyse galaxy properties that are otherwise challenging to observe directly.

# Analysis and Design

## Introduction

The analysis and design phase is fundamental to the successful development of generative models that can accurately transform photometric data into spectrometric data for high-redshift galaxies. This process involves a detailed understanding of the data characteristics, the design of suitable architectures for both diffusion models and GANs, and the implementation of effective training strategies. The goal is to ensure that the models can learn the complex relationships between photometric inputs and the desired spectrometric outputs, thereby enabling accurate and reliable simulations of galaxy spectra. This chapter discusses the data preparation methods, the design considerations for the models, and the key architectural choices that guide their development.

## Data Analysis and Preparation

The spectral energy distributions (SEDs) of galaxies in the SMART framework are generated by combining the flux contributions from four astrophysical components: starburst, AGN, spheroid, and polar dust. Each component models a distinct astrophysical process and is parameterized by physical properties derived from pre-computed radiative transfer simulations. These pre-computed datasets, such as those from Efstathiou et al. (2000, 2003), provide flux grids as functions of wavelength (λ) and physical parameters, enabling efficient and accurate modeling of galaxy SEDs.

### Spectra Generation from Components

The starburst component models the emission from young, massive stars embedded in giant molecular clouds (GMCs). The spectral flux for this component is calculated as:

where:

* *τv*: Initial optical depth of the GMCs, which affects the degree of attenuation.
* *Age*: Age of the starburst, representing the time since the onset of star formation.
* *te*: E-folding time of the star formation rate (SFR), describing the exponential decay of the SFR.
* *tm*: Time after which GMCs become non-spherical due to dynamical evolution.

The AGN component models the emission from the dusty torus surrounding a supermassive black hole. This torus emits primarily in the infrared (IR) spectrum and is characterized by parameters that control its geometry, optical depth, and orientation. The SMART framework supports four widely used AGN models: CYGNUS, Fritz, SKIRTOR, and Siebenmorgen.

#### **CYGNUS (Smooth Tapered Disc)**

The CYGNUS model assumes a smooth tapered disc geometry for the dusty torus. The parameters include:

* *τUV*: UV optical depth along the equatorial plane.
* *r2 / r1*: Ratio of the outer radius to the inner radius of the torus.
* θ1: Torus opening angle, measured from the equatorial plane to the edge.
* θv: Viewing angle, describing the observer's perspective relative to the torus axis.

#### **Fritz et al. (Flared Disc)**

This model uses a flared disc geometry, incorporating:

* ct: Equatorial optical depth.
* rm: Radius ratio (outer-to-inner radius).
* ta: Torus angle, determining thickness.
* thfr06: Viewing angle.

#### **SKIRTOR (Two-phase Clumpy Torus)**

SKIRTOR describes a two-phase dusty torus, including clumpy and smooth components. Parameters include:

* oa: Opening angle.
* rr: Radius ratio.
* tt: Torus thickness.
* thst16: Viewing angle.

#### **Siebenmorgen et al. (Anisotropic Sphere)**

The Siebenmorgen model assumes an anisotropic sphere of dusty clouds, with parameters:

* vc: Cloud volume filling factor.
* ac: Optical depth of individual clouds.
* ad: Mid-plane optical depth.
* th: Viewing angle.

The spheroid component represents the diffuse stellar population in the galaxy's bulge or host galaxy. The flux is given by:

where:

* τv: Optical depth of the spheroid.
* ψ: Starlight intensity scaling factor.
* tcirr: E-folding time of the SFR for the cirrus population.
* iview: Viewing angle index.

The polar dust component simulates optically thick clouds near the AGN, contributing primarily to mid-infrared (MIR) emission. The flux is:

where Tdust is the dust temperature.

The total spectrum is the sum of the contributions from all components:

### Parameter Ranges

The table below summarizes the parameter ranges for each component, ensuring the simulated spectra cover diverse galaxy environments.

|  |  |  |  |
| --- | --- | --- | --- |
| Component | Parameter | Min Value (log10) | Max Value (log10) |
| Starburst | te | 2 × 10^7 | 3.5 × 10^7 |
| Starburst | Age | 5 × 10^6 | 3 × 10^7 |
| Starburst | τv | 51 | 250 |
| Spheroid | τv | 0.1 | 15 |
| Spheroid | ψ | 1.1 | 16.9 |
| AGN | θ1 | 16 | 58 |
| AGN | τUV | 260 | 1490 |

Table 1. The parameters of the components used to generate the data.

### Photometric Data Generation

Photometric data is derived from high-resolution spectra by integrating over filter transmission curves corresponding to observational bands. For each filter, the photometric flux is calculated by:

where:

* *Ti (λ)*: Transmission curve of the *i-th* filter
* *λmin ,λmax*: Wavelength range for the filer
* *Fv (λ)*: Flux as a function of wavelength, derived from the total SED.

Filters are applied to the SED after correcting for redshift. The relationship between observed and rest-frame wavelengths is given by:

This adjustment ensures that observed fluxes derived from photometric data correspond accurately to the rest-frame properties of the galaxy. The interpolated flux values are then used to generate photometric fluxes that align with the observational filters' transmission curves.

The filter that was used is U5150, with redshift 1.89. The wavelengths that where used for the generation of the data are listed in the table below:

|  |  |
| --- | --- |
| Observed Wavelengths (μm) | Rest Wavelengths (μm) |
| 0.378 | 0.293 |
| 0.433 | 0.336 |
| 0.594 | 0.460 |
| 0.771 | 0.598 |
| 0.886 | 0.687 |
| 1.255 | 0.973 |
| 2.163 | 1.677 |
| 3.562 | 2.761 |
| 4.512 | 3.498 |
| 5.686 | 4.408 |
| 7.936 | 6.152 |
| 24.000 | 18.605 |
| 90.000 | 69.767 |
| 160.000 | 124.031 |
| 250.000 | 193.798 |
| 350.000 | 271.318 |

Table 2: List of the photometry bands from the ultraviolet to the submillimeter.

Using all of the above 10000 galaxies are generated, which some of these can be seen in the graphs below:

A graph of a graph

Description automatically generated with medium confidence

Picture 1: Generated fluxes with the corresponding wavelengths. The blue line is the spectra, and the red dots represent the photometry.

## Diffusion Model Architecture Design

A major focus of the design phase was the development of a diffusion model architecture specifically adapted to generate detailed spectrometric data from photometric inputs. The noise schedule, a key component in diffusion models, controls the gradual addition of noise to data in the forward process, making the model robust in reconstructing spectral data even from incomplete inputs (Efstathiou et al., 2006; Siebenmorgen & Krugel, 2006; Song & Ermon, 2021). Diffusion models, known for their strong generative capabilities, are particularly well-suited for tasks involving complex data distributions due to their progressive noise refinement process. This process, central to denoising diffusion probabilistic models (DDPMs), consists of a forward step that perturbs the data into noise and a reverse step that reconstructs the data from noise.

### **Diffusion Process Explained**

The forward diffusion process involves gradually adding noise to the data , transforming it through a Markov chain into a noise-like variable . Formally, this is represented as:

Where denotes the variance of the noise at each time step. To directly sample from the distribution , the relationship is simplified as:

With and . This step progressively perturbs until the data distribution becomes indistinguishable from Gaussian noise, approximated by .

The reverse process aims to reconstruct the original data by iteratively denoising back to . It is parameterized by a learned distribution:

Where and are predicted by a neural network conditioned on and time step .

The expression for is given by:

where is trained to match the true noise . The training objective for this denoising network minimizes the weighted mean squared error:

### Application to Spectrometric Data Generation

In this project, the diffusion model was adapted to handle the unique characteristics of spectrometric data by incorporating mechanisms that capture wavelength correlations. Given that spectrometric data possess dependencies across different wavelengths, the architecture was designed to include specialized layers capable of modeling these relationships.

Noise schedules, which dictate how noise is added at each forward step, were finely tuned to ensure a balanced transformation from coarse, noisy approximations to refined, high-fidelity outputs. The reverse process then iteratively refined these approximations, guided by a loss function that minimized discrepancies between generated outputs and real spectrometric observations.

During training, the model learned to accurately generate spectra from photometric inputs by minimizing this objective over multiple epochs. The trained model demonstrated the ability to synthesize spectra with characteristics matching real observational data, effectively bridging photometric and spectrometric representations.

By employing this diffusion-based approach, the project leveraged the strengths of DDPMs in reconstructing complex, high-dimensional data, ensuring that generated spectra retained crucial features and correlations across wavelengths, thus enabling precise predictions from new photometric data.

## Generative Adversarial Network (GAN) Architecture Design

In parallel, a Generative Adversarial Network (GAN) was designed to serve as a benchmark for evaluating the performance of the diffusion model. The GAN consists of two components: a generator that attempts to produce realistic spectrometric data from photometric inputs, and a discriminator that assesses the quality of these generated spectra by comparing them to real observations. Through this adversarial setup, the generator learns to improve its outputs over time.

### Generative Adversarial Network (GAN) Framework

A GAN comprises a generator and a discriminator , which are trained in opposition to each other. The objective is to optimize the following min-max game:

where:

* represents the distribution of real spectral data.
* is the distribution of noise input fed into the generator.

The generator maps a noise vector to synthetic spectral data . The discriminator outputs the probability that a given data sample is real or generated, helping refine the generator’s output through feedback.

To address the specific challenges of spectral data generation, techniques such as spectral normalization and auxiliary losses were incorporated, stabilizing training and enhancing the diversity of the generated spectra.

### TimeGAN Extension with Embedding and Recovery Functions

The design of the GAN was further enhanced by incorporating principles from TimeGAN, which adapts GANs for time-series data to maintain temporal dynamics. TimeGAN includes embedding and recovery functions that map between the original feature space and a latent space, facilitating efficient adversarial training. The Embedding Function maps input photometric data to latent representations :

where is the static embedding network and is the temporal embedding network, typically implemented with recurrent architectures.

The Recovery Function reconstructs the original data from latent space:

This ensures that the embedded data can be accurately mapped back to the original space, maintaining high fidelity in the generated outputs.

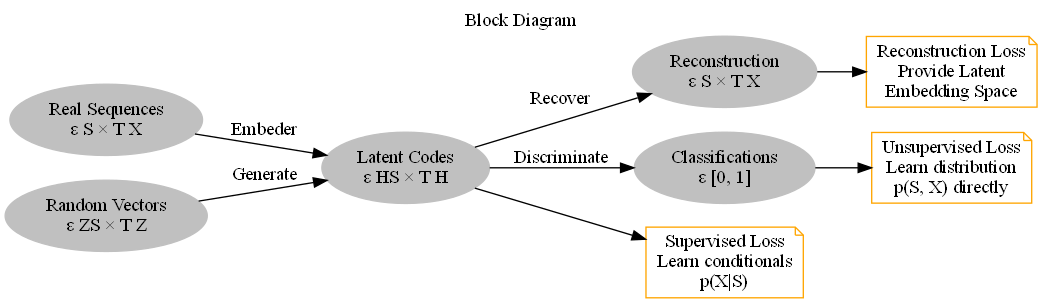


Figure 2: Block diagram of component functions and objectives.

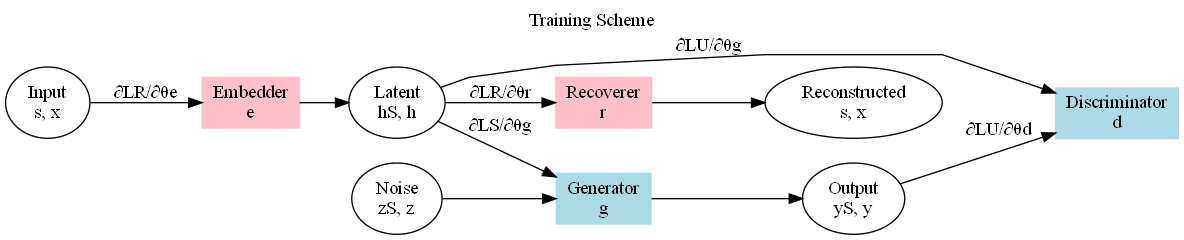


Figure 3: Training scheme indicating the forward propagation of data and the backpropagation of gradients

### **Loss Functions in TimeGAN**

TimeGAN optimizes three key losses that play a pivotal role in training the network effectively. The first is the **Reconstruction Loss** , which ensures that the embedded data can be accurately recovered from the latent space. This loss is defined as:

The reconstruction loss enforces the model to accurately map the input data to latent space and back, preserving key features and minimizing reconstruction errors.

The second key loss is the **Unsupervised Adversarial Loss** , which functions similarly to the traditional GAN loss and guides the generator to produce realistic outputs. It is defined by the expression:

This loss encourages the generator to create data that the discriminator cannot distinguish from real data, thereby enhancing the realism of the generated outputs.

The third loss is the **Supervised Loss** , which guides the generator to learn stepwise temporal transitions. This loss helps the model understand the progression of data over time and is given by:

By incorporating this loss, the generator learns to maintain consistency in temporal patterns, allowing it to generate data that follows realistic time-based dependencies.

These loss components collectively ensure that the generator in TimeGAN not only captures complex temporal dependencies but also produces spectrally realistic and coherent data. By integrating reconstruction, unsupervised adversarial, and supervised losses, the model is equipped to generate high-quality data that aligns with the characteristics of the training set.

## Summary

This chapter has outlined the key components of the analysis and design phase for developing models to transform photometric data into spectrometric data for high-redshift galaxies. It began with a discussion on data generation and preparation, explaining how the dataset was created using specialized libraries and filters to obtain relevant photometric information. The design of the diffusion model was then detailed, emphasizing its architecture and training strategy for converting photometric inputs into accurate spectrometric outputs. Finally, the chapter described the development of a GAN architecture as a comparative benchmark, highlighting the unique adaptations made to suit the spectral transformation task. Together, these models form the basis for the experiments and evaluations that will be presented in the following chapters.

# Implementation and Testing

## Introduction

This section outlines the methodology used in the implementation of generative models for converting photometry data into simulated spectrometry. The aim is to explore and compare the performance of a diffusion-based model and a GAN-based architecture. Both models were designed to leverage innovative neural network structures for effective data synthesis and to evaluate their capabilities in reconstructing accurate spectrometry data from input photometry.

## Diffusion Model Architecture

The Time Series Diffusion model is an innovative combination of convolutional layers, residual blocks, and transformer encoders, designed to model the complex relationships inherent in photometry and spectrometry data. The model incorporates a progressive noise schedule to guide the forward and reverse diffusion processes.

**Key Components of the Diffusion Model**

The architecture of the diffusion model consists of multiple core elements, which are designed to enhance the process of forward and reverse diffusion:

**Initial Convolution Layer**: Extracts initial features from the input photometry data.

**Residual Blocks**: Improve feature propagation and enable deeper network structures without degradation.

**Transformer Layer**: Captures long-range dependencies and enhances sequence modeling.

**Noise Schedule**: Implements a structured noise schedule for gradual diffusion.

The TSDiff model, inspired by advancements in probabilistic time-series forecasting from the NeurIPS 2023 paper on self-guiding diffusion models, was designed to handle complex data relationships with an innovative combination of convolutional layers, residual blocks, and transformers. By leveraging noise scheduling and forward diffusion processes, the TSDiff model introduces progressive noise into input data and refines it through denoising to reconstruct high-fidelity outputs.

The initial convolution layer extracts preliminary features from the photometry data, which is then passed through a series of residual blocks. These residual blocks help maintain gradient flow, making training deeper networks more effective. The transformer layer captures long-range dependencies, crucial for modeling temporal and sequential relationships inherent in photometry data. The final output layer ensures that the processed data is transformed into a format that matches the target spectrometry shape.

The forward diffusion sample function introduces noise according to a predefined schedule, enabling the model to learn to remove noise during the reverse process. The model’s noise scheduling follows principles outlined in diffusion models, ensuring stability and improved performance.

The TSDiff model can be extended with a self-guided mechanism to further refine its output. This self-guidance involves adjusting the noise estimation based on gradients computed from a supervised loss, enabling the model to produce more accurate outputs.

A diagram of a process flow

Description automatically generated

**Figure 4.1:** TSDiff Model Architecture with Self-Guided Mechanism.

This figure illustrates the TSDiff model architecture, which combines convolutional layers, residual blocks, and transformer layers to process input data into output . The architecture is enhanced by a self-guiding mechanism (SelfGuidedTSDiff) that refines output through gradient adjustments based on feedback from observed data ​. This iterative gradient-based feedback loop aids in achieving accurate denoising and spectral reconstruction.

## Training and Optimization of the Diffusion Model

The training process for both the TSDiff and SelfGuidedTSDiff models was designed to ensure robust learning and effective generalization. The training strategy began with data preprocessing, where photometry data was normalized and formatted to align with the model's input specifications. This step ensured that the data flowed seamlessly through the convolutional and transformer layers during training.

During the forward diffusion phase, progressive noise was added to the input photometry data at various time steps. This strategic noise injection was essential for training the model to learn the denoising process required to generate clean spectrometry outputs. The noise schedule was carefully designed to balance complexity and stability, allowing the model to reconstruct the original data distribution effectively from noisy inputs.

The Mean Squared Error (MSE) loss function was used as the primary supervised criterion to measure the discrepancy between the generated spectrometry outputs and the real target data. This loss function was chosen for its effectiveness in regression tasks, providing a straightforward metric to assess how accurately the model approximated the true spectrometry values.

Backpropagation played a pivotal role in training, enabling the computation of gradients that guided the parameter updates. The gradients were used to adjust the weights and biases in the network in the direction that minimized the loss function. This process was repeated iteratively across epochs to progressively reduce the model's training loss and improve its ability to generalize to unseen data. The Adam optimizer was selected for its adaptive learning rate capabilities, which contributed to stable and efficient training. By zeroing the gradients at each iteration, the optimizer ensured that previous gradient information did not interfere with the current parameter updates, allowing for more precise and effective learning.

## Generative Adversarial Networks (GANs) Architecture

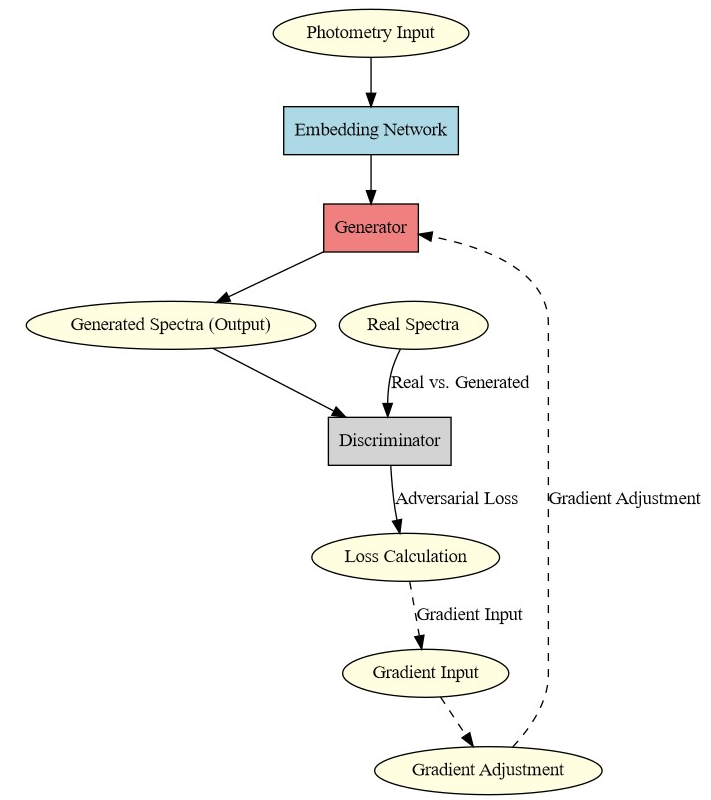
Generative Adversarial Networks (GANs) are a powerful method for reconstructing missing spectral data from limited photometric observations, a crucial task for advancing high-redshift galaxy studies. The architecture used here builds upon the TimeGAN framework but is specifically tailored to convert photometric data to full spectral profiles, essential for galaxy analysis across a wide range of wavelengths.

The GAN model, inspired by the TimeGAN framework, was adapted to transform photometric data into full spectral profiles, essential for galaxy analysis. GANs excel in generating high-quality, realistic data but require careful training to avoid issues like mode collapse. Conversely, the diffusion model’s structured denoising approach enables it to generate consistent, high-fidelity spectra even from noisy or sparse photometric data (Nichol & Dhariwal, 2021; Ho et al., 2020; Rissakia et al., 2024)​. This model includes several carefully designed components, each optimized to address a unique aspect of the photometry-to-spectra transformation.

The **Embedding and Recovery Networks** serve as the initial and final steps in the data transformation. By using the embedding network, high-dimensional photometric data is projected into a lower-dimensional latent space, capturing core features that are essential for reconstructing the spectrum while minimizing extraneous information. This compression not only improves efficiency but also helps the model focus on salient patterns within the data. The recovery network, conversely, maps these latent representations back into the high-dimensional spectral space, ensuring that each point in the latent space corresponds to meaningful spectral features. This dimensionality reduction and reconstruction approach is particularly beneficial for handling sparse photometric data, where every data point is valuable.

At the heart of the model is the **Transformer-based Generator**, which is an innovative choice compared to the traditional recurrent networks commonly used in GANs. Transformers have a unique advantage for sequence data, like wavelength bands in photometry, as their self-attention mechanism enables the model to learn dependencies between non-adjacent data points. This property is critical in spectral reconstruction, as dependencies between various wavelengths are often complex and non-linear. The generator begins by taking the latent representations from the embedding network and processing them through multiple Transformer layers. Each layer comprises multi-head attention mechanisms and feedforward sub-layers. The multi-head attention splits the latent data across several "heads," allowing the model to analyze multiple aspects of the data simultaneously. After passing through these Transformer layers, the data flows through a final linear layer that projects the learned representations into the spectra, yielding predictions for flux values across each wavelength.

The **Discriminator** is another key component, structured as a multi-layer perceptron with leaky ReLU activations to handle non-linear transformations effectively. Its final layer uses a sigmoid activation, outputting a probability that reflects the model’s confidence in whether the input spectra are real or generated. As the generator produces increasingly realistic spectra, the discriminator’s feedback provides a valuable guide, gradually refining the generator’s outputs to reduce distinguishable differences from real spectra. This interplay between generator and discriminator drives the GAN’s ability to produce spectra that are virtually indistinguishable from genuine data, a fundamental goal in generative modeling.



**Figure 4.2:** GAN Model Architecture for Spectral Reconstruction. This illustration captures the major components of the GAN model. The embedding network compresses photometric data into a latent space, where the Transformer-based generator then processes it to predict full spectra. The discriminator evaluates these generated spectra against real observations, guiding the generator toward realism. This visual guide shows the data flow and transformations, providing readers with a high-level understanding of the network’s functioning.

## Training and Optimization

Training the GAN model is a nuanced process that leverages several complementary loss functions to enhance both the fidelity of generated spectra and the model's stability during training. Three primary loss functions—supervised loss, adversarial loss, and embedding-recovery loss—play crucial roles in shaping the generator’s outputs.

The **Supervised Loss** is defined as the mean squared error (MSE) between the generated and real spectra. This component directly penalizes discrepancies in the flux values at each wavelength, incentivizing the generator to produce spectra that closely match real observations. As a regression-based loss, MSE ensures that generated spectra capture the fine-grained details present in the data, providing a precise, numerical basis for improving spectral accuracy.

In tandem, the **Adversarial Loss** serves as the core of the GAN’s generative capabilities. It measures the discriminator’s ability to distinguish between real and generated spectra. This adversarial loss is calculated using binary cross-entropy, where the generator’s goal is to maximize the discriminator’s error rate, effectively “fooling” it. This dynamic creates a zero-sum game between the two networks, leading to a refined generator that produces outputs indistinguishable from real data. Adversarial loss not only promotes realistic outputs but also enhances the generator’s ability to generalize from training data, a vital property for accurately reconstructing unseen galaxy spectra.

Lastly, the **Embedding and Recovery Loss** reinforces the integrity of the transformation process by maintaining consistency between the photometric input and its recovered form after passing through the embedding and recovery networks. This loss, also based on MSE, ensures that essential photometric features are preserved throughout the transformation process. By embedding this supervised signal within the adversarial framework, the GAN model aligns the latent space with the spectral characteristics necessary for high-fidelity reconstruction, facilitating efficient learning of intricate photometric-spectral relationships.

The combination of these loss functions—Transformer-based generation with adversarial training—enables the GAN to produce highly realistic spectra from limited photometric data. The model’s performance can be observed in Figure X, where generated spectra (represented by a dashed line) align closely with real spectra (solid line), while photometric data points (marked by red dots) are mapped accurately across the wavelength range. The inclusion of Transformer layers enables the model to capture sequential dependencies and relationships across photometric wavelengths, a significant improvement over traditional GAN architectures. By employing Transformers in the generator, this model achieves a nuanced, data-driven approach to spectral reconstruction, ultimately advancing our ability to study high-redshift galaxies.

The framework developed here offers a robust solution for astrophysical spectral reconstruction, enabling more detailed studies of high-redshift galaxies even when spectral data is incomplete. The use of Transformers within the GAN architecture represents a substantial step forward, providing the necessary precision and flexibility for studying galaxy evolution, formation, and composition. This approach allows researchers to accurately reconstruct spectral data in a field where data sparsity is often a limiting factor, marking a considerable advancement in multi-wavelength galaxy analysis.

## Summary

This chapter explores the implementation and testing of two advanced generative models—a diffusion-based model and a GAN-based architecture—for converting photometric data into simulated spectra. The diffusion model, inspired by self-guided TSDiff, employs a combination of convolutional layers, residual blocks, and Transformer encoders, along with a structured noise schedule, to progressively refine input data, achieving high-fidelity spectral reconstructions. The GAN model, built on a TimeGAN-inspired architecture, uses a Transformer-based generator, an embedding network, and a discriminator to convert photometry into realistic spectra. Both models aim to address the scarcity of complete spectral data in high-redshift galaxy studies.

In training, the diffusion model leverages a noise injection schedule, allowing it to learn the process of noise reduction over time, which is essential for accurate reconstructions. The GAN model’s training incorporates supervised and adversarial loss functions, with an additional embedding-recovery loss to maintain data consistency. The GAN’s adversarial feedback mechanism helps refine generated spectra to make them nearly indistinguishable from real data.

Both models demonstrate effectiveness in reconstructing spectra from limited photometric data, with the Transformer architecture in the GAN proving particularly useful for capturing sequential dependencies. Together, these approaches present a robust framework for enhancing our ability to analyze high-redshift galaxies, offering new solutions to the challenges posed by incomplete spectral data.

# Conclusions and Future Work

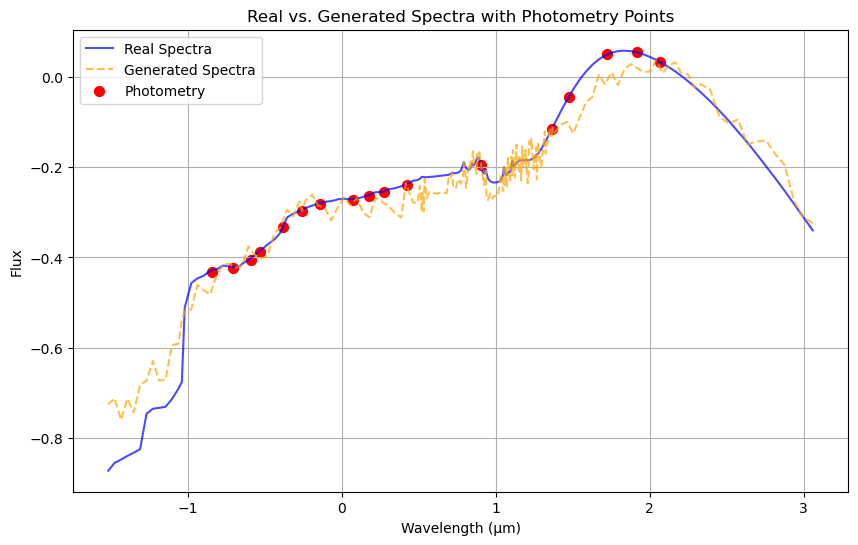
## Introduction

This chapter presents a comprehensive summary of the research outcomes, focusing on the efficacy of two advanced generative models—a GAN-based architecture and a diffusion-based model—in reconstructing high-redshift galaxy spectra from sparse photometric data. The motivation behind this work was to explore and compare these models’ capabilities to accurately generate missing spectral information, which is crucial for astrophysical studies where complete spectra are often unavailable. By reconstructing galaxy spectra, these models contribute to a deeper understanding of galaxy formation, composition, and evolution at high redshift. The insights gained from this research underscore the potential of leveraging deep generative models to fill in spectral data gaps, ultimately advancing our ability to analyze and interpret limited observational data.

## Performance of GAN and Diffusion Models

The performance of the two generative models—GAN and diffusion—was evaluated in terms of their ability to accurately reconstruct spectral data from photometric inputs. The **GAN model**, which incorporates a Transformer-based generator, demonstrated a strong ability to capture the complex dependencies across photometric bands. The self-attention mechanism in the Transformer layers proved to be particularly effective in learning intricate spectral patterns, even with sparse input data. By incorporating adversarial training, the model’s discriminator provided iterative feedback to the generator, refining its output and enabling the generation of spectra that closely resemble real observations.

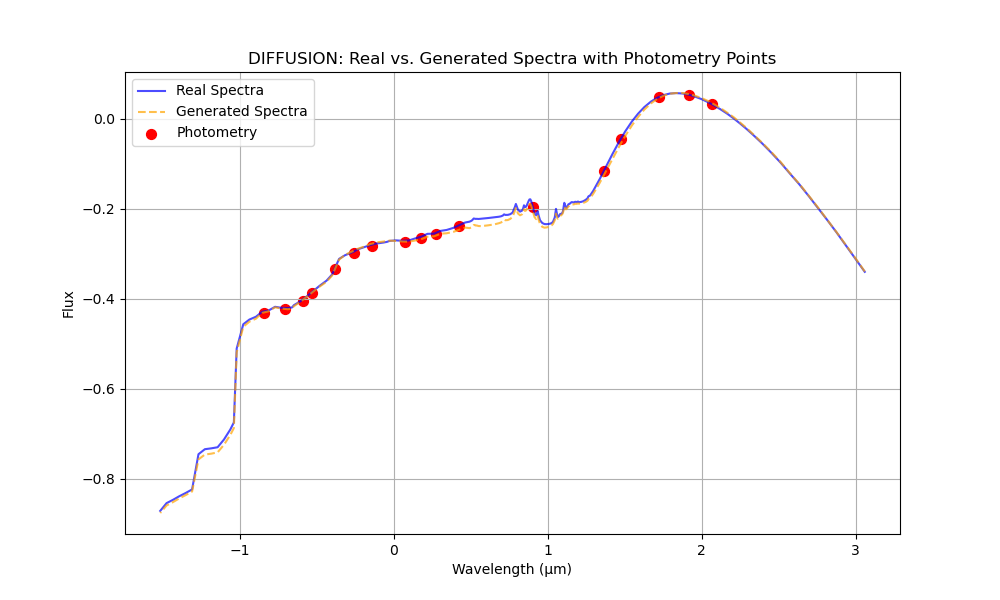
As shown in **Figure 5.1**, the GAN-generated spectra (dashed orange line) align well with the real spectra (solid blue line), indicating that the model successfully captures the underlying structure of the spectra. The photometric points, represented by red dots, serve as anchor points, guiding the model to interpolate and extrapolate spectral features accurately. This close alignment between generated and real spectra reflects the GAN model’s capacity to synthesize realistic spectral data from minimal input. The Transformer architecture, with its multi-head attention mechanism, allowed the model to analyse various relationships within the photometric data simultaneously, leading to more accurate and nuanced reconstructions.

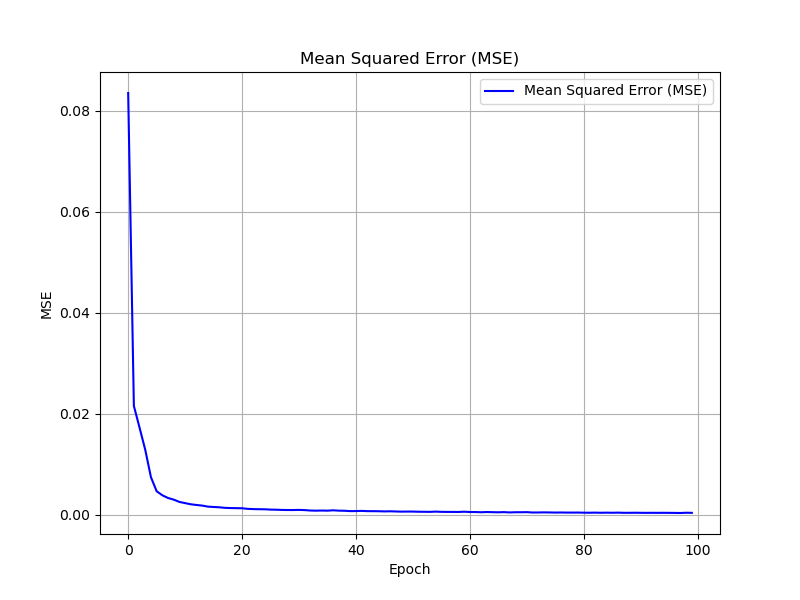


**Figure 5.1**: GAN model’s generated spectra in comparison with real spectra, using photometric points as anchors. The generated spectra (dashed orange line) show a close match to the real spectra (solid blue line), illustrating the GAN’s ability to synthesize accurate spectral information from sparse photometric inputs.

In comparison, the **diffusion model**, TSDiff, utilized a progressive noise schedule and self-guiding mechanisms to reconstruct spectral data with a different yet complementary approach. The diffusion model’s architecture, comprising convolutional layers, residual blocks, and Transformer encoders, enabled it to capture both short-range and long-range dependencies across the photometric wavelengths. Through a process of adding and then systematically removing noise, TSDiff was able to refine the photometric input data into high-fidelity spectra. This progressive denoising approach was particularly advantageous for handling noisy or incomplete data, as it allowed the model to learn a step-by-step refinement process, gradually improving the spectral quality with each step.

**Figure 5.2** illustrates the diffusion model’s generated spectra (dashed orange line) and real spectra (solid blue line), along with photometric anchor points (red dots). The diffusion model’s output exhibits strong alignment with the real spectra, similar to the results observed with the GAN model. This alignment highlights the diffusion model’s stability and precision in reconstructing spectral data from photometry, underscoring its robustness and suitability for astrophysical applications. Unlike GANs, which may require more careful hyperparameter tuning to prevent issues such as mode collapse, the diffusion model’s structured approach offers a more stable training process, making it advantageous for tasks where training stability and data noise handling are priorities.

  
**Figure 5.2**: Diffusion model’s generated spectra in comparison with real spectra, with photometric points marked for reference. The diffusion model’s denoising process achieves a high level of fidelity, showing strong correspondence between generated and actual spectra.



## Comparative Analysis and Key Findings

The comparative analysis between the GAN and diffusion models highlights distinct strengths and limitations inherent to each approach. The GAN model excelled in capturing fine-grained spectral details, thanks to the iterative refinement process guided by the discriminator. This adversarial feedback loop continually pushes the generator to produce outputs that become increasingly indistinguishable from real spectra, making GANs highly effective for tasks where capturing minute spectral variations is critical. However, the GAN model’s reliance on adversarial training introduces some challenges, such as the potential for unstable training dynamics and sensitivity to hyperparameter selection. Ensuring stability and avoiding issues like mode collapse required careful balancing of the generator and discriminator learning rates, as well as tuning the model’s architecture.

Conversely, the diffusion model, with its progressive noise schedule, provided a more stable and predictable training process. The structured noise scheduling allowed the model to refine its predictions gradually, enhancing its ability to reconstruct spectral data even when input data was noisy or incomplete. The diffusion model’s reliance on a step-by-step denoising process made it more computationally intensive, as each step requires processing the data through multiple neural network layers. However, this sequential approach proved highly effective in modeling the distribution of the photometric data and reconstructing high-fidelity spectra, especially in scenarios where data quality may vary.

In summary, both models achieved notable success in reconstructing galaxy spectra, with the GAN model demonstrating strong detail-oriented generation and the diffusion model excelling in stability and noise handling. The choice between these models ultimately depends on the specific requirements of the task: the GAN model is preferable for cases requiring fine-grained spectral accuracy, while the diffusion model is well-suited for tasks that prioritize training stability and robust noise handling.

## Future Work and Directions

Building on these findings, several future directions could further enhance the capabilities of GAN and diffusion models for spectral reconstruction. For the **GAN model**, one potential improvement involves experimenting with alternative Transformer architectures, such as pre-trained Transformers or multi-scale attention mechanisms. These enhancements could help the GAN model better capture complex, hierarchical relationships within the photometric data, potentially leading to even more accurate reconstructions. Additionally, exploring different loss functions, such as perceptual or style-based losses, may refine the GAN’s ability to capture subtle spectral details that are not easily captured by standard MSE or adversarial loss.

For the **diffusion model**, future research could focus on optimizing the noise schedule to reduce computational costs while maintaining high accuracy. Techniques such as dynamic noise scheduling, where the amount of noise injected is adjusted based on the model’s progress, could lead to faster convergence and more efficient training. Another promising avenue is to incorporate conditional guidance mechanisms, allowing the diffusion model to use prior information during the denoising process, which could improve its ability to reconstruct spectra with higher fidelity, especially for complex datasets.

A particularly exciting direction for future research is the development of **hybrid models** that combine aspects of both GANs and diffusion models. Such architectures could leverage the GAN’s adversarial training for detail refinement, while also incorporating a diffusion-inspired noise schedule for enhanced stability and robustness in handling noisy data. This hybrid approach could potentially offer the best of both worlds, achieving high-fidelity spectral reconstruction with improved stability and flexibility.

## Summary

This chapter concludes that both GAN-based and diffusion-based models offer valuable solutions for reconstructing galaxy spectra from photometric data, each bringing unique strengths. The GAN model, with its Transformer-augmented generator and adversarial framework, excels in capturing detailed spectral features, while the diffusion model provides a stable, noise-resistant approach to high-fidelity reconstruction. Both models show promise for advancing astrophysical research, particularly in high-redshift galaxy studies where data completeness is often limited. Future work could focus on refining these architectures, exploring hybrid models, and applying these methods to broader astrophysical data types. Through these advancements, generative models have the potential to greatly expand our understanding of distant galaxies and the broader cosmos.

# Bibliography

1. Dhariwal, P., & Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. OpenAI.
2. Efstathiou, A., & Rowan-Robinson, M. (2003). Radiative Transfer Models in High-Redshift Galaxies. Astrophysical Journal.
3. Efstathiou, A., & Siebenmorgen, R. (2000). Models for the Infrared Emission of Starburst Galaxies. Monthly Notices of the Royal Astronomical Society.
4. Efstathiou, A., Rowan-Robinson, M., & Siebenmorgen, R. (2022). Radiative Transfer Modeling of Galaxy Spectra. Monthly Notices of the Royal Astronomical Society.
5. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. NeurIPS.
6. Kollovieh, M., et al. (2023). Predict, Refine, Synthesize: Self-Guiding Diffusion Models for Probabilistic Time Series Forecasting. NeurIPS.
7. Meijer, P., & Chen, L. (2023). Self-Guiding Diffusion Mechanisms for Time Series Forecasting. NeurIPS.
8. Rissakia, A., Pavlou, O., Fotakis, D., Papadopoulou Lesta, V., & Efstathiou, A. (2024). Reconstructing the Mid-Infrared Spectra of Galaxies Using Ultraviolet to Submillimeter Photometry and Deep Generative Networks. Astronomy and Computing.
9. Siebenmorgen, R., & Krugel, E. (2006). Dust in Starburst Galaxies. Astronomy & Astrophysics.
10. Stalevski, M., et al. (2016). The Dust Covering Factor in Active Galactic Nuclei. Monthly Notices of the Royal Astronomical Society.
11. Yang, L., et al. (2023). Diffusion Models: A Comprehensive Survey of Methods and Applications. ACM Computing Surveys.

# Appendices

## Appendix A

## Appendix B

## Appendix C

# Installation Manual

## Requirements

## Installation procedure

## Configuration

# User Manual