Encoding SAR Ocean Signatures into Latent Space

Capturing Multi-Scale Ocean Phenomena in SAR Imagery with Variational Autoencoders

Master thesis

Isabel Slingerland

Delft University of Technology

TU Delft

Encoding SAR Ocean Signatures into Latent Space

Capturing Multi-Scale Ocean Phenomena in SAR Imagery with Variational Autoencoders

by

Isabel Slingerland

Student Name

Isabel Slingerland

Supervisor: Prof. dr. F. Lopez-Dekker
Supervisor: O. O'Driscoll
Supervisor: dr. I. Rocha
Faculty: Faculty of Geoscience and Civil Engineering, Delft

Cover: Sentinel-1A SAR ocean imagery (Wave Mode), Copernicus Programme, European Space Agency (ESA)

TU Delft

Preface

Starting this thesis, like most decisions in the past five years, began with an intuitive decision. A spark of interest, followed closely by fear of not knowing much about the subject, followed by the choice to do it anyway. Five years ago I left a career as a professional dancer and took a leap of faith to go back to university. I had a vague, but persistent feeling that I wanted to study something about Earth science. I could not have predicted that this would lead me to Synthetic Aperture Radar and machine learning. Looking back, what felt like one giant leap (or jeté in dance terms) was actually a sequence of small decisions, each one guided by the same instinct: if something interests me and frightens me in equal measure, it is probably worth pursuing. At the TU Delft open day, I chose a remote sensing workshop almost at random. It turned out to be led by my now-supervisor, Paco. Throughout my bachelor's I discovered an unexpected affinity for computational science which surprised me enormously given that I had arrived with no programming background. I chose to do a minor in Computational Science and Engineering precisely because it felt like a steep learning curve. It was. But it also, fittingly, led me into the field of computing uncertainty. Choosing a master's track brought its own anxieties. I worried about career prospects and about picking the wrong specialisation. I eventually recognised that this kind of thinking was paralysing me, and I returned to the same principle: follow what genuinely interests you, trust that and let go of worrying about the outcome. The Earth Observation track felt immediately right. When the opportunity arose to do an internship with the Sentinel-1 team at ESA (knowing it would delay my graduation) I took it. I am deeply grateful to my supervisor at ESA Tobias and the whole Sentinel-1 team for making a retired dancer feel so welcome in the space sector. Before I started this journey of going back to University in my 30ies, my mother warned me at the open day: "Just be aware that you might feel like an outsider, that you might not fit in." What I found instead was some of the strongest connections I have made were during these last 5 years: Elin, Hamra, Julian, Jiaqi, Mona, Gui and many others. I may not always have felt like I fit in, but I discovered that university life did fit me. This thesis itself was no exception to the pattern. I got stuck, felt overwhelmed, went through periods of genuine hating machine learning (and then liking it again) and with the challenge of keeping an overview of it all. I am sincerely grateful to my three supervisors, Paco, Owen, and Iuri for the conversations that always helped me take the next step when I was lost. I came out the other side not necessarily with all the solutions, but with findings, more knowledge and with new interests I am excited to explore. Finally, thank you to my family, for supporting me through the stressful final stretch of this thesis, and for the five years before it.

Isabel Slingerland
Delft, March 2026

Abstract

Synthetic Aperture Radar (SAR) satellites produce vast archives of high-dimensional ocean imagery, capturing complex multi-scale surface patterns induced by sea-air interaction processes. To estimate geophysical parameters such as turbulence fluxes, scientists traditionally apply domain knowledge to manually select physically meaningful features, an implicit form of dimensionality reduction. This thesis explores whether Variational Autoencoders (VAEs) can automate this process of dimensionality reduction by learning compressed latent representations directly from raw SAR ocean imagery. Using 220,000 Sentinel-1 Wave Mode images co-located with ERA5 reanalysis data, VAE architectures were trained across four latent dimensionalities (32, 64, 128 and 256). The multi-scale complexity of SAR scenes introduced a strong frequency bias: standard pixel-wise losses such as Mean Squared Error failed to capture fine-scale detail, and conventional metrics such as PSNR and SSIM proved insufficient to measure this. Frequency Focal Loss (FFL) was incorporated to address reconstruction in the spectral domain, alongside a dynamic weight matrix that refocuses the loss on difficult-to-learn features. Dynamically annealing loss term weights had a striking effect on reconstruction quality, and subsequent hyperparameter optimisation using Optuna further confirmed that loss function tuning dominates over architectural choices. This sensitivity to loss function design is a central finding of this work. VAEs successfully reconstructed large- and intermediate-scale ocean patterns at latent dimensions of 128 and 256. For air-sea flux estimation, three configurations were compared: a direct CNN regressor, a frozen VAE encoder with regression head, and a jointly trained VAE. All three underperform the physics-informed approach of O'Driscoll et al.[29], suggesting unsupervised deep learning alone cannot extract flux-relevant information, though task objectives incorporated into the loss function can redirect what the model learns. This points toward conditional or semi-supervised architectures as promising future directions.

Contents

1

Introduction

Ocean and atmosphere form a coupled system, whose interactions drive weather patterns, regulate the climate and distribute heat around the globe. Understanding these exchanges requires real-time, large-scale monitoring of air-sea parameters such as sea surface temperature, wind speed, humidity gradients, heat fluxes and turbulence. The complexity is that many parameters are difficult to observe directly, or are required to be observed in coordination with other parameters [11].

Synthetic Aperture Radar (SAR) satellites offer unique capabilities for observing the ocean and, indirectly, the atmosphere as well. SAR provides high-resolution, all-weather, day-and-night imagery of the ocean surface. SAR captures the backscatter from transmitted signals, which is highly sensitive to sea surface roughness composed of centimeter-scale waves. Interactions between the ocean and the atmosphere modulate these short waves, leaving distinct signatures of geophysical processes that SAR imagery can capture. Processes such as ocean waves, wind streaks and atmospheric and oceanic fronts can all be observed and classified [44]. Therefore, SAR data is inherently high-dimensional, capturing multiple processes that span many orders of magnitude in spatial scale. Over four decades of SAR data have accumulated since NASA's Seasat mission in 1978 [12], with current missions such as Sentinel-1 [40] and planned missions like Sentinel-1 Next Generation [42] and Harmony [25, 26] ensuring that the archive is ever growing.

This creates both an opportunity and a challenge: how can we systematically extract meaningful geophysical parameters from this high-dimensional, multi-scale, continually expanding dataset?

Current Approaches and Limitations

Existing methods for obtaining ocean-atmosphere characteristics from SAR rely on manual feature selection informed by physical intuition. Based on domain knowledge of how atmospheric turbulence appears in SAR ocean data, researchers extract hand-selected spectral features based on first and second order statistics. [50, 29].

Young et al. [50]developed theoretical correlations between wind field spectra and turbulence parameters and created analytical methods for inferring atmospheric boundary layer variables from SAR spectral characteristics. By combining domain-engineered features and machine learning, O'Driscoll et al. [29] improved turbulence parameter estimation. While these domain-knowledge approaches have achieved notable success, they rely on manual feature selection. This may not capture the full information content available in SAR imagery. Furthermore distinguishing between geophysical processes that operate at overlapping spatial scales can't be separated by spectral filtering alone. An automated, unsupervised, data-driven model might be useful in this situation.

Deep Learning for Automated Feature Extraction

Deep learning offers the potential for automated and data-driven feature extraction by convolutional neural networks (CNN) that are mainly used for processing imaging data. Unsupervised deep learning methods can discover the underlying structure of high-dimensional datasets that can be mapped to

a lower dimension latent space, without requiring manual specification of relevant features. The Variational Autoencoder (VAE) is a good example of a probabilistic generative model that works well for dimensionality reduction tasks [22, 31].

An image is mapped by the VAE encoder to a multivariate normal distribution around a point in the latent space. The VAE decoder can map this point back to an image. This way the VAE can generate new images by sampling different points from the latent space [22, 31]. Training is performed using a loss function comprising two terms that balance reconstruction accuracy against latent space regularisation, ensuring that the latent distributions remain smooth and structured. This enables both accurate reconstruction and the meaningful generation of diverse images.[1].

If successful, this data-driven approach could offer several advantages over manual feature engineering. An automated method could potentially capture information that domain experts might overlook, while also providing a scalable solution for processing the continually expanding SAR archive without requiring expert intervention for each new dataset or mission.

However, the ocean surface presents a unique challenge for automated feature extraction. SAR ocean imagery simultaneously captures geophysical processes spanning multiple spatial scales, from meter-scale capillary waves to kilometer-scale atmospheric fronts, all overlapping within a single scene. Moreover, the surface is in constant motion, meaning no two SAR scenes are ever identical. Yet despite this variability, the underlying geophysical processes generate recurring, spatially invariant signatures at their characteristic scales. For any dimensionality reduction approach to be effective, it must preserve information across this full range of spatial scales.

This multi-scale preservation is particularly critical for turbulence parameter estimation. Atmospheric boundary layer turbulence manifests across a spectrum of spatial scales in SAR imagery, from fine-scale turbulent eddies to larger organized structures. Manual feature engineering approaches [50, 29] address this by carefully selecting spectral features. However, if a VAE can learn to encode this multi-scale information into a compact latent representation, the resulting features may capture the full complexity of turbulent signatures more effectively than hand-picked spectral features.

Ocean SAR Imagery: A Distinct Challenge for VAEs

Previous VAE studies on SAR imagery have focused on land surface classification [46] and ship detection [47], applications where scenes are relatively static or where the target of interest (ships) is distinct from the background. These applications do not face the same multi-scale, complex air-sea signatures. The key question becomes: **Can a VAE learn a compact representation of multi-scale geophysical signatures from inherently variable ocean scenes, and then use this representation to both reconstruct scenes and generate new, physically plausible ones?**

Research Gap and Motivation

However, standard CNN-based VAEs have a draw back when applied to data containing small details and high frequencies. The F-principle[32] describes CNNs frequency bias, which causes them to learn low frequencies before high frequencies during training. For SAR ocean data, where geophysical processes span a large frequency range, this bias manifests as blurry reconstructions that lose high-frequency information. To address this limitation, we integrate Focal Frequency Loss (FFL) [19] into the VAE training objective. In order to preserve the different frequency information, FFL adaptively focusses the model on difficult-to-learn frequency components by explicitly optimising reconstruction in the frequency domain.

Research Approach

This study uses a VAE framework to accomplish two interconnected goals:

1. **Multi-scale reconstruction:** Compressing high-dimensional SAR ocean imagery to assess the capability of the reconstruction of multiple frequency scales while preserving multi-scale geophysical processes
2. **Unsupervised feature extraction:** Unsupervised learning can be used to discover latent features for downstream ocean parameter estimation, bypassing the need for human feature engineering.

The VAE framework naturally connects these goals through a multi-objective loss function that balances reconstruction accuracy, latent space regularization, frequency preservation (via Focal Frequency Loss). In extention to flux paramter estimation there will be an extended model that would use a loss function to predict turbulence fluxes. This work systematically optimizes this multi-objective function using hyperparameter optimization (Optuna [2]), demonstrating that careful tuning of architectural parameters, loss component weights, and training schedules is essential for achieving all goals simultaneously. This approach differs from earlier VAE applications to remote sensing SAR data, where loss functions are typically adopted with default settings. The complete mathematical formulation and optimization strategy are detailed in Chapter 3.

Research Objectives

Primary Research Question

Can a variational autoencoder effectively compress SAR ocean imagery into a low-dimensional latent representation that preserves turbulence-relevant geophysical information across multiple spatial scales?

Specific Research Questions

1. **Frequency bias mitigation:** To what extent does Focal Frequency Loss address the frequency bias inherent in CNNs when reconstructing SAR ocean imagery? What impact does this have on preserving geophysically relevant scales?
2. **Optimal latent dimensionality:** What is the minimum latent dimensionality required to preserve information sufficient for turbulence parameter estimation? How does compression ratio affect reconstruction quality and downstream parameter accuracy?
3. **Feature quality comparison:** How do VAE-derived latent features compare with manually picked features [29] for estimating turbulence parameters when validated against ERA5 reanalysis?
4. **Latent space structure:** How should the VAE latent space be regularized (through β-annealing and loss balancing) to produce physically meaningful, interpretable representations that enable controlled generation of varied ocean-atmosphere scenarios?

Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 establishes theoretical foundations in three areas: (1) SAR missions and SAR ocean remote sensing principles (2) atmospheric boundary layer turbulence and parameter estimation methods, and (3) variational autoencoders, including β-VAE formulation, frequency bias in CNNs, and Focal Frequency Loss.

Chapter 3 presents the complete methodology: dataset preparation from Sentinel-1 acquisitions, VAE architecture design and and training strategies (including β-annealing and FFL integration), hyperparameter optimization framework using Optuna.

Chapter 4 reports experimental results organized by research question: baseline VAE performance, impact of Focal Frequency Loss on reconstruction quality and frequency preservation, latent dimensionality analysis, turbulence parameter estimation validated against ERA5 data, comparison with O'Driscoll et al.'s approach [29], and latent space visualization and controlled generation experiments.

Chapter 5 Discusses the results in scientific context: VAE effectiveness as a dimensionality reduction tool for geophysical remote sensing, significance of frequency bias mitigation for SAR applications, physical interpretability of learned representations, comparison with related work, limitations, and generalizability to other SAR applications and outlines future research directions.

Chapter 6 synthesizes key contributions, summerises findings and implications in context of the research questions.

2

Background and Theoretical Foundation

2.1. Synthetic Aperture Radar for Ocean Monitoring

2.1.1. History and Evolution of SAR Missions

Since the launch of NASA's Seasat mission in 1978 vast amount of ocean-covering SAR data have been acquired, with archives now spanning over more than four decades. Seasat was the first Earth-orbiting satellite to carry a Synthetic Aperture Radar (SAR) sensor specifically designed for oceanographic applications[12]. This groundbreaking mission paved the way for ocean remote sensing using active microwave instruments and inspired many more satellite SAR missions. Some of the most successful missions were the ESA's ERS-1, ERS-2 and Envisat, which operated during the 1990s and 2000s. Another successful mission was Canada's RADARSAT, which was launched in 2002. All of these missions greatly improved ocean monitoring from space[16][3].

Building on this legacy, the European Commission and the European Space Agency (ESA) launched the Copernicus Programme, which uses space-based and in situ observations to monitor the Earth's environment and deliver operational data to users worldwide. The programme comprises a series of missions called Sentinels, with each mission consisting of two satellites that fulfil revisit and coverage requirements while providing open-access datasets for Copernicus services [37]. A cornerstone of this programme is the SAR Sentinel-1 mission, which provides continuous all-weather, day-and-night radar imagery for land and ocean monitoring[30][37].

Sentinel-1A was launched in 2014, followed by Sentinel-1B in 2016. Both satellites operate in C-band, enabling applications such as sea ice monitoring, ship detection and the retrieval of wave and surface current data. To ensure mission continuity, the European Space Agency (ESA) has developed Sentinel-1C and Sentinel-1D as follow up satellites, which launched in 2024 and 2025 respectively[41]. These satellites have now replaced the older versions A and B, and feature several enhancements. Looking ahead, ESA is planning the Sentinel-1 Next Generation (S1-NG) mission to maintain and enhance C-band SAR data continuity into the 2030s[42].

In addition to the Copernicus programme, the European Space Agency's (ESA) Earth Explorer Programme develops science-driven missions proposed by the scientific community. These missions are designed to address key gaps in our understanding of the Earth system[6]. Harmony has been selected as the tenth Earth Explorer mission and will consist of two bistatic SAR satellites flying in convoy with a Copernicus Sentinel-1 satellite, using its radar as an illuminator. The mission will improve the representation of small-scale ocean–atmosphere interactions in current Earth system models, including momentum flux and wave–current coupling, thereby addressing a critical gap. Understanding these processes is fundamental to comprehending air–sea energy exchange, extreme weather events, and climate feedback mechanisms.[25][26].

2.1.2. Sar Imaging theory

Synthetic aperture radar (SAR) is a side-looking imaging radar system mounted on a moving platform such as a satellite[16]. As the platform moves along its flight path, the radar system repeatedly transmits electromagnetic pulses towards the side and receives the echoes of backscattered signals through its antenna [27]. As the platform is in motion, each transmission and reception occurs from a different position, resulting in a series of observations of the same ground area from slightly different viewing angles.

SAR operates in the microwave spectrum, typically between 500 MHz and 100 GHz. At these frequencies, electromagnetic waves can penetrate clouds with minimal attenuation[9]. SAR can therefore take high-resolution surface images, regardless of cloud cover, weather conditions and solar illumination. These are capabilities unavailable to optical sensors, which require sunlight and clear skies[28]. These all-weather, day-night imaging capabilities are essential for operational oceanographic monitoring, where observations of wind fields and storm systems are often accompanied by clouds[9].

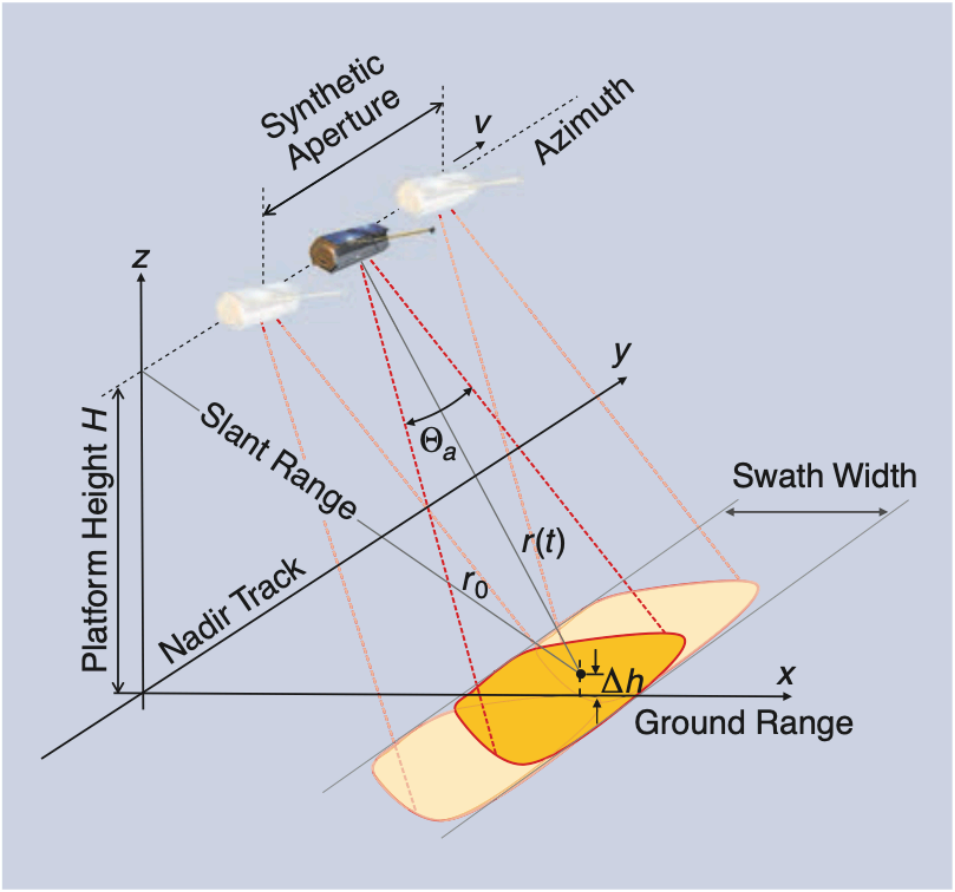


Figure 2.1: Illustration of SAR imaging geometry [27]. Θ_a is the azimuth beamwidth, v is the platform velocity, and r0 is the shortest approach distance.

SAR imaging geometry and resolution

SAR imaging geometry is defined by two orthogonal directions. The satellite's trajectory has an along-track direction (the direction of flight) and an across-track direction (perpendicular to the flight path). These map to radar-specific concepts: the along-track direction corresponds to azimuth angle, azimuth time, and azimuth position, while the across-track direction corresponds to slant range and range time. In the processed SAR image, these become the azimuth and range axes respectively [39]. The slant range direction follows the radar's line of sight from the antenna to any point on the ground surface. This geometry, shown schematically in Figure 2.1, determines how SAR measurements are organised and processed into two-dimensional images.

The radar transmits electromagnetic pulses towards the ground at an oblique incidence angle [9]. As these waves travel through space, they accumulate phase shifts proportional to the distance travelled. When the pulses encounter the Earth's surface, a fraction of the inbound radiation is reflected back to the antenna. The amplitude and phase characteristics of this reflection are determined by the distance of the target, surface properties, and geometric configuration [28]. The distance between the antenna and any fixed ground target changes continuously as the platform moves along the azimuth direction. It decreases as the platform approaches, then increases after it passes. This motion induces frequency shifts in the received signal, known as Doppler frequencies [39]. The magnitude and rate of the Doppler shift depend on the target's position within the antenna beam, providing a mechanism for discriminating between targets at different azimuth positions. Exploiting these Doppler frequency variations is essential for achieving fine azimuth resolution in SAR imagery.

SAR systems produce high-resolution, two-dimensional images of large areas through the use of range and azimuth information. Position along the range axis is determined by the time delay of the received pulse. Position along the azimuth axis is determined by the Doppler frequency of the return signal [16]. SAR achieves high resolution through two independent mechanisms. In the range direction, resolution is determined by the bandwidth of the transmitted pulse. Shorter pulses, or equivalently a wider bandwidth, enable targets at different distances to be more finely discriminated. The azimuth resolution of conventional radars is fundamentally limited by the physical length of their antennas—longer antennas produce narrower beams, enabling closely spaced targets to be discriminated in the azimuth direction [39].

SAR overcomes this constraint by exploiting platform motion. As the satellite travels along its orbit, the radar antenna repeatedly observes each ground target from a sequence of positions. By coherently combining these observations during processing, SAR synthesises an effective aperture length equal to the distance travelled while the target remains within the antenna beam [10][27]. This synthetic aperture can be orders of magnitude longer than any physically realisable antenna, enabling metre-scale azimuth resolution. However, creating a synthetic aperture requires maintaining phase coherence across all observations of a target. The radar system must therefore track the accumulated phase of each return signal precisely [39]. Signal processing, known as SAR focusing, is required to transform the recorded phase histories into interpretable two-dimensional images [27].

The final SAR image is constructed using the time delay (which determines the range position), the Doppler frequency (which determines the azimuth position), and the strength of the backscattered signal of each resolution cell. The intensity and phase of the backscattered returns primarily depend on the target's geometry, surface roughness, and dielectric properties [16]. SAR backscatter is usually converted to a normalized radar cross section (NRCS), representing the radar reflectivity of the surface per unit area, which allows comparisons to be made across different acquisition geometries. For ocean surfaces, these dependencies enable SAR to indirectly observe atmospheric and oceanic processes through their modulation of sea surface roughness. This forms the basis for SAR oceanography.

2.1.3. SAR Ocean monitoring

The imaging process along the azimuth direction can be understood as a correlation operation. The received signal is cross-correlated with a template signal as a function of time. Each azimuth position corresponds to a particular time. Image intensity is assigned according to the correlation strength at that time. Correct image formation depends on the signal phase behaving as expected based on the geometric relationship between the sensor and target[43]. This assumption holds well for static land surfaces. However, ocean surfaces present a fundamental challenge. The ocean surface is dynamic and moves both randomly and with organized motions such as ocean swell. These motions affect both the placement of radar echo energy on the SAR image and the focus of the image itself. Target motion relative to the imaging plane causes phase perturbations[43]. For ocean waves, the orbital motion introduces periodic variations in phase perturbation. This periodic variation corresponds to the periodic changes in wave orbital velocity projected along the radar line of sight. Consequently, image intensity is periodically misplaced in the azimuth direction. This creates periodic patterns in the image. Since these patterns follow the spatial period of the ocean wave, a wave image is generated despite the phase incoherency. The periodicity of the ocean waves is thus preserved in the SAR image, even though individual scatterers are misplaced. This principle enables SAR to detect ocean wave patterns

and extract information about their wavelength and direction[43].

Bragg Scattering and Ocean Surface Interaction/Ocean waves and internal waves
There are two main types of wave on the ocean surface: gravity waves and capillary waves. Gravity waves are formed when gravitational forces act on wind-disturbed bodies of water. These waves have longer wavelengths, and their height depends on wind speed, fetch length and wind duration. Capillary waves, on the other hand, arise from surface tension counteracting wind disturbances. They have shorter wavelengths, typically in the centimetre range[3]. Both wave types are illustrated in figure 2.2.

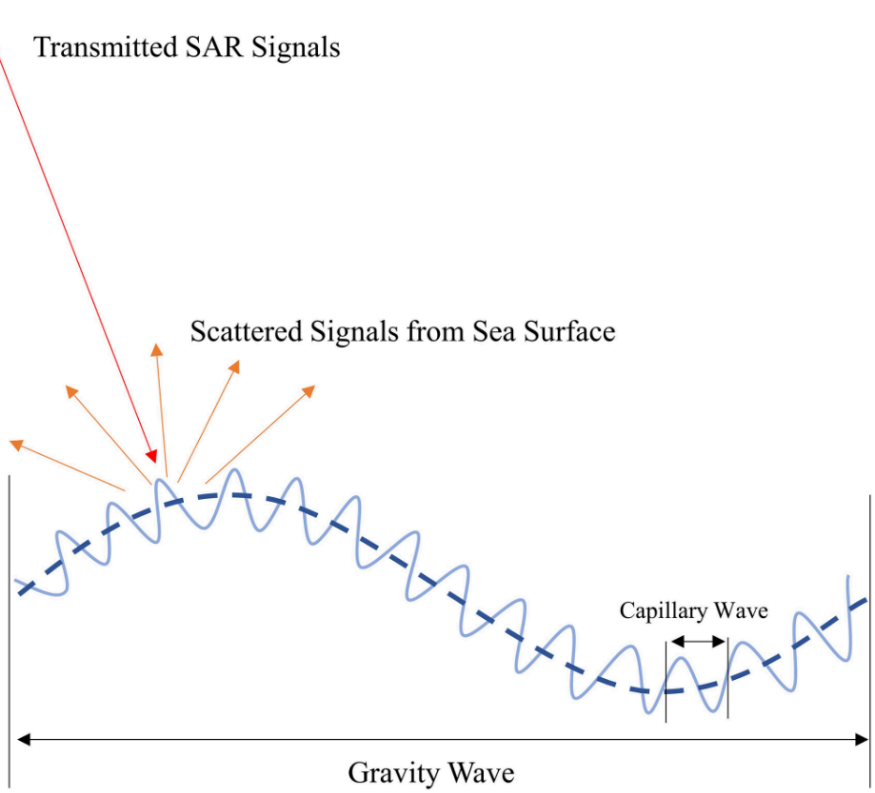


Figure 2.2: Capillary waves and gravity on the ocean's surface as ilustrated by [3]

The dominant SAR signal return at lower incidence angles on the ocean surface is Bragg scattering. This occurs when the radar wavelength and ocean wave wavelength satisfy a specific resonance condition. The backscattered signal is primarily influenced by Bragg interactions with capillary waves. However, non-Bragg mechanisms, such as whitecaps and breaking waves, also contribute to the signal[3]. When the ocean surface is perfectly flat, it acts as a specular reflector. Wind roughens the surface, generating centimeter-scale waves that scatter energy back toward the sensor. Wind speeds between 2 and 14 m/s produce enough backscatter signal above the noise floor, enabling detection of ocean surface features [3].

The NRCS depends on several factors including transmitted signal characteristics (frequency, polarization, imaging geometry) and ocean surface properties (roughness, multi-scale wave interactions, currents, and surface features such as oil slicks or ships). Incidence angle significantly affects the observed backscatter. At smaller incidence angles (around 20°), higher sea surface returns are typically observed. The returned energy decreases rapidly as the incidence angle increases. At larger incidence angles (around 40°), the sea surface appears darker, making surface objects more apparent. This incidence-angle dependency enables observation of gravity waves through their interaction with capillary waves. The front slopes of gravity waves face the transmitted signal at reduced incidence angles and appear brighter. Back slopes increase the effective incidence angle and appear darker [3].

SAR's sensitivity to centimeter-scale surface roughness enables indirect inference of atmospheric conditions. Wind stress at the ocean surface generates and modulates capillary and short gravity-capillary waves, which in turn modulate the radar backscatter [3]. Ocean surface roughness increases with wind speed and exhibits directional dependence, with the highest radar reflections occurring in the upwind direction and the lowest in the crosswind direction [39]. This relationship between atmospheric forcing and surface roughness forms the physical basis for retrieving wind speed and direction from SAR imagery.

Various approaches have been developed to exploit this relationship. Wind speed is commonly retrieved using geophysical model functions (GMFs) that relate NRCS to observation geometry and wind speed [3]. Wind direction can be estimated through spectral methods, such as Fast Fourier Transform (FFT) based approaches, that identify dominant wind streak orientations in the frequency domain[3]. Beyond wind parameters, SAR imagery provides information about wave height and direction (derived from Doppler shifts and backscatter modulations), surface currents and eddies (identified through wave modulation patterns), and internal waves and boundary layer turbulence (observed indirectly through their impact on surface roughness) [39, 3].

This indirect observational capability, whereby atmospheric properties are inferred from their impact on ocean surface roughness, is central to the motivation behind this thesis. If the processes of the atmospheric boundary layer leave detectable, recurring signatures in SAR imagery, then learning compact representations of these signatures could allow parameters to be retrieved systematically without the need for explicit, multi-step processes.

2.1.4. Atmospheric Boundary Layer Turbulence

As mentioned SAR images capture signatures of various geophysical phenomena that are associated with air-sea exchanges. Several of these phenomena factor significantly in the vertical transport of heat, moisture and momentum, and play key roles in the climate system[50]. Wind fields contain information on atmospheric stability. The key is that SAR backscatter can be converted to surface wind speed, by the assumption that backscatter is correlated to the amplitude of the surface wave spectrum and surface stress[50]. Microscale convection is caused by vertical temperatures and wind fluxes, while mesoscale convection is caused by larger scale horizontal fluxes. These fluxes are air-sea interactions, and climate models can better depict the characteristics of the marine atmospheric boundary layer (MABL) by using these parameters[11]. Young et al [50] developed a method to derive MABL turbulence and stability statistics from SAR backscatter imagery using Monin–Obukhov similarity theory, mixed layer similarity theory, and estimates of boundary layer depth. This method is based up on the assumption that wind-speed variability contains information on atmospheric stability. O'Driscoll et al [29] significantly improved turbulence parameter estimation by integrating machine learning with domain engineered features. They extracted multiple SAR-derived parameters including first and second-order wind speed statistics (mean, standard deviation). These features were then used in a machine learning regression model to estimate the Obukhov length L, a fundamental atmospheric stability parameter that characterizes the balance between shear- and buoyancy-driven turbulence. This study used ERA 5 data and point-wise buoy data as validation.

2.2. Variational Autoencoders

2.2.1. Unsupervised learning and dimensionality reduction

In unsupervised learning, we typically assume that our data exhibits hidden patterns governed by unknown latent variables [4]. An important motivation for latent variable models is that many datasets exhibit the property that data points lie close to a manifold of much lower dimensionality than the original data space. If our goal is to compress data or reduce dimensionality, there can be significant benefits in exploiting this manifold structure [4].

In practice, data points will not be confined precisely to a smooth low-dimensional manifold, and departures from the manifold can be interpreted as noise. This leads naturally to a generative view of such models in which we first select a point within the manifold according to a latent variable distribution and then generate a data point by adding noise, drawn from a conditional distribution of the data variables given the latent variables [4].

The simplest continuous latent variable model assumes Gaussian distributions for both the latent and observed variables and makes use of a linear mapping between them. This leads to a probabilistic formulation of the well-known technique of principal component analysis (PCA), known as probabilistic PCA [4]. While linear methods such as PCA are effective for certain datasets, complex real-world data, such as SAR ocean imagery, often exhibit non-linear structures that cannot be adequately captured by linear dimensionality reduction. This motivates the use of nonlinear encoder-decoder architectures.

2.2.2. From autoencoders to variational autoencoders

An autoencoder is a neural network designed to learn a compressed, lower-dimensional representation of high-dimensional input data. The autoencoder consists of two parts [31][13]: an encoder that maps the input data $\mathbf{x}$ into a latent representation $\mathbf{z}$, and a decoder that reconstructs the data $\tilde{\mathbf{x}}$ from the latent space. The latent space acts as a bottleneck, forcing the network to capture only the most important features of the data while discarding noise and redundancies [31][13].

The objective of a standard autoencoder is to minimize the reconstruction error between the input $\mathbf{x}$ and its reconstruction $\tilde{\mathbf{x}}$, typically using a mean squared error (MSE) loss. When both the encoder and decoder consist of linear transformations, the autoencoder performs principal component analysis (PCA) [31]. However, standard autoencoders have important limitations: they can overfit to the training data, they produce deterministic point estimates in latent space, and they do not guarantee that the latent space is continuous or structured in a meaningful way [1].

Variational Autoencoders (VAEs) address these limitations through a probabilistic formulation (Figure 2.3). Rather than learning deterministic mappings, VAEs learn probability distributions over latent variables. The encoder, parametrised by θ, compresses high-dimensional input data x into a lower-dimensional latent representation by learning an approximate posterior distribution $q_\theta(z|x)$. The latent variable z is sampled from this distribution. This posterior is regularised to be close to a predefined prior distribution $p(z)$, which is typically a standard normal distribution. The decoder, parametrised by ϕ, constructs the output data $\tilde{x}$ from the latent variable z using a learned likelihood distribution $p_\phi(\tilde{x}|z)$ [1][31][13].

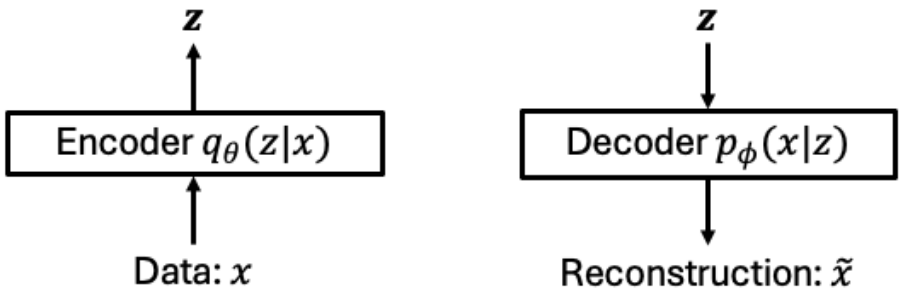


Figure 2.3: Schematic illustration of the encoder and decoder architecture of a variational autoencoder (VAE). The encoder maps input data x to latent variables z, which the decoder uses to generate the reconstruction $\tilde{x}$. Adapted from [22].

The training objective balances two competing goals: accurate reconstruction of the input and regularization of the latent space. This is achieved through a loss function that corresponds to the Evidence Lower Bound (ELBO) of the model:

$$ELBO(\theta, \phi) = \mathbb{E}_{q_\phi(z|x)}\left[\ln p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta\, D_{KL}\left(q_\phi(z|x)\|p(z)\right) \quad (2.1)$$

The first term encourages accurate reconstruction, while the second term, weighted by β, regularizes the latent distribution to match the prior. This regularisation prevents overfitting and enables the model to generalize beyond the training data, allowing it to represent not only observed patterns but also plausible variations within the learned distribution [1].

Mathematical theory of VAE

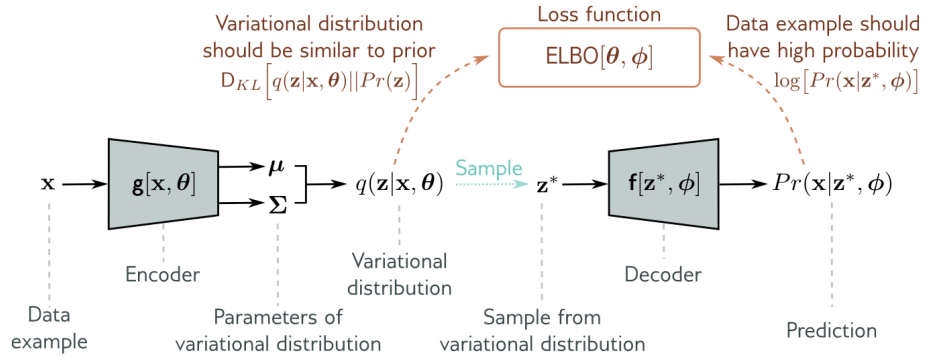


Figure 2.4: Scemetic ilustration of the VAE[31]. The encoder maps a data example x to a Variational distribution, the decoder samples from that distribution and makes a prediction. The loss function is the negative ELBO, which depends on how accurate this prediction is and how similar the variational distribution is to the prior.

Figure 2.4 illustrates the VAE model structure, consisting of an Encoder and Decoder and the ELBO which is a way of minimizing loss function. The model defines a joint distribution $p(\mathbf{x}, \mathbf{z})$ over observed data $\mathbf{x}$ and latent variables $\mathbf{z}$, where the marginal distribution over data is obtained by integrating over the latent space:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z}, \phi)p(\mathbf{z})\, d\mathbf{z} \quad (2.2)$$

The prior $p(\mathbf{z})$ and the likelihood $p(x|z, \phi)$ are multivariate Gaussian distributions:

$$\begin{cases} p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, \mathbf{I}) \\ p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|f_\phi(\mathbf{z}), \sigma^2\mathbf{I}) \end{cases} \quad (2.3)$$

The difficulty is that $f_\phi(\mathbf{z})$ is a non linear function and ϕ is the set of parameters that map z back to x (the decoder). [31]. This means the posterior $p(z|x)$ will not be a Gaussian distribution. The posterior could in principle be computed using Bayes law $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$, but it is intractable, because we can't evaluate $p(x)$ in the denominator [31]. To circumvent this, variational inference introduces an approximate posterior:

$$q_\theta(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu_q, \Sigma_q) \quad (2.4)$$

This approximated posterior is an isotropic Gaussian distribution with mean μ_q and a diagonal covariance matrix Σ_q, both dictated by a set of parameters θ. The encoder maps the input x to mean and variance vectors. It does not need to worry about the covariance between the dimensions because the isotropic quality means that all the distributions per dimension are independent.[13][1].

We want to obtain the parameters ϕ and θ we maximize the log-likelihood $\log p(x)$. We define an evidence lower bound (ELBO), which is always less than or equal to the log-likelihood for a given value of ϕ and θ. As seen in figure 2.4, the loss function $ELBO[\theta, \phi]$, will couple the encoder and decoder by training both sets of parameters θ and ϕ.

Evidence Lower Bound (ELBO) Derivation

The Evidence Lower Bound provides a tractable objective function for training the VAE. Starting from the log-likelihood of the data and introducing the approximated distribution $q_\theta(\mathbf{z}|\mathbf{x})$:

$$\log p(\mathbf{x}) = \log p(\mathbf{x}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{x}) + q_\theta(\mathbf{z}|\mathbf{x}) - q_\theta(\mathbf{z}|\mathbf{x}) \tag{2.5}$$

Rearrange these terms and take the expectation with respect to $q_\theta(\mathbf{z}|\mathbf{x})$:

$$\log p(\mathbf{x}) = \int q_\theta(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{x}, \mathbf{z})}{q_\theta(\mathbf{z}|\mathbf{x})} \, d\mathbf{z} - \int q_\theta(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{z}|\mathbf{x})}{q_\theta(\mathbf{z}|\mathbf{x})} \, d\mathbf{z} \tag{2.6}$$

The first term is the ELBO, and the second term is the KL divergence $D_{KL}(q_\theta(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}|\mathbf{x}))$, which is always non-negative. Therefore, the ELBO provides a lower bound on the log-likelihood. We can rewrite this term as [31]:

$$\log p(\mathbf{x}) \geq \text{ELBO}(\theta, \phi) = \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{x}, \mathbf{z}) - \log q_\theta(\mathbf{z}|\mathbf{x}) \right] \tag{2.7}$$

The joint can be decomposed into two terms $\log p(\mathbf{x}, \mathbf{z}) = \log p(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z})$. This means equation 2.7 can be rearranged:

$$\log p(\mathbf{x}) \geq \text{ELBO}(\theta, \phi) = \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{x}|\mathbf{z}, \phi) \right] + \underbrace{\mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{z}) - \log q_\theta(\mathbf{z}|\mathbf{x}) \right]}_{-D_{\text{KL}}(q_\theta(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))} \tag{2.8}$$

This formulation reveals again two terms that are competing objectives in VAE training. The first term, the expected log-likelihood or reconstruction term, measures how well samples from the approximate posterior can reconstruct the input data. The second term, the KL divergence $D_{KL}(q_\theta(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})$ between the approximate posterior and the prior acts as a regulariser that encourages the learned latent distribution to remain close to the prior [1][7].

The KL divergence term has a closed-form solution when both distributions are Gaussian. For a diagonal covariance structure with M latent dimensions, it can be computed as:

$$D_{KL}(q_\varphi \| p) = \frac{1}{2} \sum_{m=1}^{M} \left(1 + \ln(\sigma_m^2) - \mu_m^2 - \sigma_m^2 \right) \tag{2.9}$$

where μ_m and σ_m^2 are the mean and variance of the m-th dimension of the approximate posterior [31].

Reparametrization Trick The reconstruction term in the ELBO involves an expectation over the distribution $q_\theta(\mathbf{z}|\mathbf{x})$, which can be approximated using Monte Carlo sampling. However, direct sampling would create a problem in the backpropagation of the gradients. The reparametrization trick solves this problem by expressing the random variable $\mathbf{z}$ as a deterministic function of the parameters μ_q, σ_q and a noise variable $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ [31]:

$$\mathbf{z} = \mu_q + \sigma_q \odot \boldsymbol{\epsilon} \tag{2.10}$$

where $\odot$ denotes element-wise multiplication. This reparametrization moves the sampling outside the main computational path, by just sampling from $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, allowing gradients to flow through both the encoder and decoder networks during backpropagation [7][13][31].

Loss Function The VAE is trained by maximizing the ELBO, or equivalently, minimizing the negative ELBO[1]. The complete loss function combines reconstruction accuracy with latent space regularization:

$$\begin{aligned} \mathcal{L}_{\text{VAE}}(\theta, \phi) &= -\text{ELBO}(\theta, \phi) \\ &= -\mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})} \left[\log p_\phi(\mathbf{x}|\mathbf{z}) \right] + D_{\text{KL}} \left(q_\theta(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}) \right) \end{aligned} \tag{2.11}$$

For Gaussian likelihood, the reconstruction term simplifies to the mean squared error between the input and its reconstruction, scaled by the variance:

$$\log p_\phi(\mathbf{x}|\mathbf{z}) = -\frac{D}{2}\log(2\pi) - \frac{D}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left\|\mathbf{x} - f_\phi(\mathbf{z})\right\|^2 \tag{2.12}$$

where D is the dimensionality of the data space [31].

2.2.3. The Role of β

β-VAE Formulation

The VAE is trained by maximizing the ELBO and most of the time its not possible to maximise both terms in the ELBO. To control this trade-off a parameter β is introduced that controls the importance of the KL divergence term [31][1]:

$$\mathsf{ELBO}(\theta, \phi, \beta) = \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[\ln p(\mathbf{x}|\mathbf{z}, \phi)\right] - \beta\, D_{\mathsf{KL}}\left(q_\theta(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right) \tag{2.13}$$

Determining Optimal β

The β parameter scales the strength of the prior $p(\mathbf{z})$ and therefore governs a fundamental trade-off in VAE training. Higher values of β enforce stronger regularisation, pushing the approximate posterior $q_\theta(\mathbf{z}|\mathbf{x})$ closer to the prior $p(\mathbf{z})$. This penalises correlations between latent dimensions, encouraging disentanglement of latent features and promoting a smooth, continuous latent space. It also improves generalisation by preventing overfitting [31]. However, excessive regularisation compromises reconstruction quality, since the model prioritises matching the prior over accurately encoding the input data.

Conversely, lower values of β allow the model to focus on reconstruction accuracy, enabling the approximate posterior to deviate more from the prior to better capture the structure of the data. This is particularly important for complex, high-dimensional data where accurate reconstruction requires expressive latent representations [1].

The optimal β value is highly task-dependent and data-dependent. For dimensionality reduction tasks where the primary objective is to preserve as much information as possible from the original data, smaller β values have been found to be more optimal [1].

A critical challenge is selecting an appropriate value for β. A notorious problem is the posterior collapse phenomenon. This occurs when the approximated posterior $q_\theta(\mathbf{z}|\mathbf{x})$ aligns to closely to the prior $p(\mathbf{z})$, effectively causing the latent variables to carry no information about the input data. In this state, the KL divergence term approaches zero, and the decoder learns to generate outputs independently of the latent code, relying only on the bias terms or the mean of the training data[7][1].

To mitigate posterior collapse and improve the quality of learned latent representations, various β-annealing strategies have been proposed. These methods gradually increase β from a small initial value to a larger target value during training, allowing the model to first learn meaningful latent representations with minimal regularisation pressure before gradually imposing structural constraints [7][1].

The core principle behind β-annealing is to initially prioritize reconstruction over regularization. By starting with a small β, the model can freely learn to encode information in the latent space without the constraint of matching the prior. As training progresses and the encoder learns to capture meaningful features, β is gradually increased, introducing regularization that structures the latent space and encourages it to align with the prior distribution [1].

Several annealing schedules exist for gradually increasing β during training. This work employs cosine annealing, which follows a smooth cosine curve:

$$\beta(t) = \frac{1 - \cos(\pi t/T)}{2} \tag{2.14}$$

where t is the current training step and T is the total number of annealing steps. This schedule provides a gradual increase in regularisation strength and has been shown to yield high performance in similar generative modelling tasks [1].

A critical consideration in β-annealing is selecting the final target value of β. While prior work commonly anneals β to a fixed value of 1 as a conventional endpoint [7], recent work has proposed using metrics such as the Fraction of Variance Explained (FVE) to identify the optimal β value[1][5]. The FVE score measures how much variance in the data is explained by the latent representation, and the β value corresponding to the peak FVE score can be used as the annealing target [1]. For a dataset with n features, each characterized by an input data $\mathbf{x}$ and its reconstruction $\hat{\mathbf{y}}$, the FVE is defined as [5]:

$$\text{FVE} = 1 - \frac{\sum_{i=1}^{n} \|\mathbf{x}_i - \hat{\mathbf{y}}_i\|^2}{\sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2} \tag{2.15}$$

where $\bar{\mathbf{x}}$ is the mean input data. The FVE score ranges from 0 to 1. Higher values indicate that the latent representation preserves more of the original data variance. Computed over the validation set during training, the score acts as a regulariser.

The relationship between β and FVE follows a characteristic trajectory during VAE training. At the start of training with a small β, the model learns to reconstruct the data effectively, leading to an increase in the FVE score. As β increases during annealing, the model reaches an optimal balance between variance explanation and regularization, corresponding to the peak FVE score. Beyond this point, as β continues to increase, the weight of the KL divergence term grows, leading to stronger regularization of the latent space. This regularization forces the latent variables to conform more closely to the prior distribution, and create a posterior collapse. This will affect the ability of the model to explain the variance, resulting in a decline in the FVE score [1].

Eventually, as β becomes sufficiently large, the regularization fully dominates, and the model prioritizes aligning the latent space with the prior over capturing meaningful variance from the data. At this stage, the FVE score flattens out, indicating that the latent space is no longer effectively representing the data. The decoder essentially defaults to reconstructing the mean of the data, also known as posterior collapse [1][7].

By monitoring the FVE score as a function of β, the optimal value can be identified as the β corresponding to the peak FVE score [1]. Importantly, this FVE-based approach can be combined with annealing strategies: rather than arbitrarily annealing to $\beta = 1$, the model can be annealed from $\beta \approx 0$ to the optimal β value identified by the FVE curve. This strategy has been shown to prevent posterior collapse more effectively than fixed β values and to produce latent spaces with better clustering properties and more distinct structural features [1][5].

Despite its effectiveness, β-annealing is not a universal solution. The optimal annealing schedule and target β value are highly dependent on the data characteristics and the task at hand. Additionally, annealing requires careful tuning of hyperparameters such as the annealing rate and the number of epochs, which can be computationally expensive [1]. Nonetheless, when properly configured, β-annealing represents a powerful technique for training VAEs that balance reconstruction fidelity with structured, meaningful latent representations.

2.3. Frequency bias in CNNs and Mitigation Strategies

2.3.1. the F-priciple

Deep neural networks (DNNs) have a tendency to learn the low-frequency components of functions before the high-frequency components. This phenomenon has significant implications for image reconstruction tasks, including architectures based on convolutional neural networks (CNNs). Rahaman et al. [32] conducted systematic experiments demonstrating that deep ReLU networks favour low frequencies and thus exhibit a spectral bias towards smooth functions. Through controlled regression experiments involving combinations of sinusoids with different frequencies, they found that lower frequencies were consistently learned first, regardless of their amplitude in the target function. This learning bias is formalized by the F-principle [48] which describes the order of priority in which neural networks fit different frequency components, typically progressing from low to high frequencies [19]. A key insight from the work of Rahaman et al. [32] is that lower frequencies are more robust to parameter perturbations, whereas expressing higher frequencies requires finely tuned parameters to work in conjunction with each other. This observation has direct implications for training dynamics: while networks can eventually learn high-frequency components through extended training, this increases the risk of overfitting,

particularly since noise typically resides in higher frequency ranges. Their experiments on MNIST data set, involving the addition of frequency-specific noise, revealed that the validation performance decreased precisely when the higher-frequency components of the noise signals were being learned. Furthermore, Rahaman et al. [32] demonstrated that the geometry of the data manifold is crucial in determining the learnability of high frequencies. Specifically, they showed that low-frequency functions defined in the input space can have high-frequency components when restricted to lower-dimensional manifolds with complex shapes. In generative models, this frequency bias manifests as a gap in the frequency domain between the original and reconstructed images. Jiang et al. [19] demonstrated that standard reconstruction losses in VAE's lead to an under-representation of high-frequency details. As with regression networks, generative models prioritise frequencies according to the F principle. This results in reconstructions that are blurred and lack fine-scale features. The theoretical basis of this bias can be traced back to the spectral properties of neural networks themselves. As demonstrated in [15], the Fourier spectrum of ReLU networks exhibits polynomial decay rates that vary according to the network architecture. This implies that representing high-frequency components requires substantially more network capacity or training time than low-frequency components. This theoretical framework helps to explain why standard training procedures that minimise mean squared error (MSE) uniformly across all frequencies fail to capture high-frequency detail adequately.

2.3.2. Frequency Focal Loss (FFL)

To address the frequency domain gap between original and reconstructed images, Jiang et al. [19] developed the Focal Frequency Loss (FFL), which adaptively weights different frequency components based on reconstruction difficulty. The approach begins by transforming images into the frequency domain via the 2D discrete Fourier transform (DFT):

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \tag{2.16}$$

where $M \times N$ represents the image dimensions (with $M = N$ for square images), (x, y) denotes spatial coordinates, $f(x, y)$ is the pixel value at position (x,y). The coordinates in the frequency spectrum are represented by (u,v), and $F(u, v)$ is the complex-valued frequency coefficient. Following Euler's formula, the exponential term can be decomposed into real and imaginary components:

$$e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} = \cos\left(2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)\right) - i\sin\left(2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)\right) \tag{2.17}$$

This decomposition reveals that the image can be represented as a sum of complex exponential basis functions 2.16 which equivalently represents a superposition of orthogonal sinusoidal components. Following Euler's formula (Equation 2.17), each frequency component $F(u, v)$ consists of both a cosine term (real part, $R(u, v)$) and a sine term (imaginary part $I(u, v)$), together forming the complete frequency representation. The spectrum coordinate (u, v) encodes the directional orientation of each spatial frequency component, while $F(u, v)$ quantifies the image's response amplitude and phase at that frequency.

The Fourier coefficient $F(u, v)$ can be expressed in terms of its real and imaginary components:

$$F(u,v) = R(u,v) + iI(u,v) = a + bi \tag{2.18}$$

where $R(u, v) = a$ represents the real part and $I(u, v) = b$ represents the imaginary part. These components encode the two fundamental characteristics of each frequency, the amplitude and the phase:

1. Amplitude (Magnitude): The amplitude describes the strength of the image's response to a 2D sinusoidal wave at a specific frequency:

$$|F(u,v)| = \sqrt{R(u,v)^2 + I(u,v)^2} = \sqrt{a^2+b^2} \tag{2.19}$$

2. The phase represents the spatial shift of the 2D sinusoidal wave relative to a reference position (the origin of the cycle):

$$\angle F(u,v) = \arctan\left(\frac{I(u,v)}{R(u,v)}\right) = \arctan\left(\frac{b}{a}\right) \tag{2.20}$$

Both amplitude and phase information are essential for a full signal reconstruction, as demonstrated by Jiang et al. [19] through ablation studies showing that reconstruction quality degrades severely when either component is discarded.

Each complex frequency value can be mapped to a Euclidean distance vector in two-dimensional space, as can be show in figure 2.5 that shows a Cartesian representation of Equation 2.18. For an original image with spectrum $F_o(u,v)$ and a reconstructed image with spectrum $F_r(u,v)$, vectors $\vec{r}_o$ and $\vec{r}_r$ are constructed with magnitudes $|\vec{r}_o| = |F_o(u,v)|$ and $|\vec{r}_r| = |F_r(u,v)|$ (the amplitudes) and angles $\theta_o = \angle F_o(u,v)\, \theta_r = \angle F_r(u,v)$(the phases). As visualised in figure this geometric interpretation captures both amplitude and phase discrepancies in a single metric.

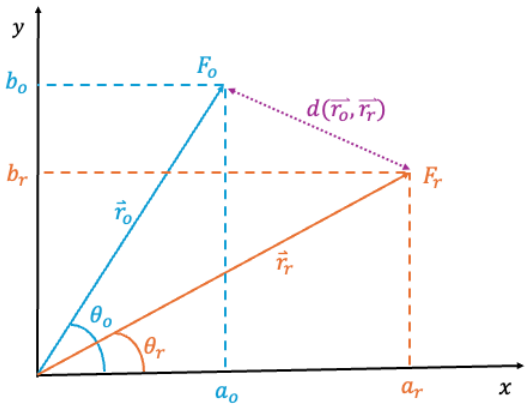


Figure 2.5: Frequency-domain distance between the original and reconstructed frequency representations at spectrum location (u,v). This figure is adapted from Jiang et al [19]. The vectors $\vec{r}_o$ and $\vec{r}_r$ are mapped from the corresponding frequency coefficients $F_o(u,v)$ and $F_r(u,v)$, respectively. The Euclidean distance $d(\vec{r}_o, \vec{r}_r)$ (purple dashed line) jointly accounts for both magnitude information ($|\vec{r}_o|$, $|\vec{r}_r|$) and phase information (angles θ_o and θ_r).

The frequency-wise distance is then computed as the squared Euclidean distance between these vectors:

$$d(\vec{r}_o, \vec{r}_r) = |\vec{r}_o - \vec{r}_r|_2^2 = |F_o(u,v) - F_r(u,v)|^2 \tag{2.21}$$

Doing this over all frequencies yields the total frequency distance:

$$d(F_o, F_r) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} |F_o(u,v) - F_r(u,v)|^2 \tag{2.22}$$

where the normalization by MN ensures scale invariance with respect to image size.

To adaptively focus training on harder-to-learn frequencies, FFL uses a dynamic weighting matrix, inspired by the focal loss mechanism of Lin et al. [24]. A spectrum weight matrix is constructed, where each element in this matrix $w(u,v)$ represents the weight for frequency component $F(u,v)$. This weight is determined by the reconstruction error at that frequency:

$$w(u,v) = |F_o(u,v) - F_r(u,v)|^\alpha \tag{2.23}$$

where $\alpha \geq 0$ is a scaling factor controlling the degree of focus (Jiang et al. [19] use $\alpha = 1$ in their experiments). The weight matrix is normalized to the range $[0,1]$, ensuring that easily-learned frequencies

(with small reconstruction errors) receive weights near zero and are down-weighted, while difficult frequencies receive weights closer to one. Crucially, gradients through the weight matrix are blocked, so it serves purely as a modulation factor rather than a learnable parameter. The complete Focal Frequency Loss is then formulated as:

$$\mathcal{L}\text{FFL} = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} w(u,v)|F_o(u,v) - F_r(u,v)|^2 \tag{2.24}$$

This loss can be interpreted as a weighted average of frequency-wise reconstruction errors, where the weights are dynamically adjusted during training to emphasize components that the model currently struggles to reproduce. As training progresses and the model becomes proficient at reconstructing certain frequencies, their weights automatically decrease, allowing the optimization to shift focus toward remaining problematic frequencies. This adaptive behaviour stands in contrast to static frequency-weighting schemes and enables progressive refinement of reconstruction quality across the entire spectrum as shown by Jiang et al. [19]. They demonstrated that FFL is complementary to spatial-domain losses such as MSE, and achieved best results when combining both objectives. For VAE-based reconstruction, their experiments showed that FFL significantly narrows the frequency domain gap, resulting in sharper images with better-preserved fine details while maintaining overall structural coherence.

3

Methodology

3.1. Data Acquisition and Preprocessing

This section describes the SAR imagery dataset and ground-truth turbulence flux estimates used to train and evaluate the proposed VAE-based models. Section 3.1.1 details the Sentinel-1 SAR data acquisition, preprocessing pipeline, and dataset composition. Section 3.1.2 describes the co-located buoyant heat flux estimates derived from ERA5 reanalysis and their use as training targets. Finally, Section 3.1.3 specifies the training, validation, and test partitioning strategy.

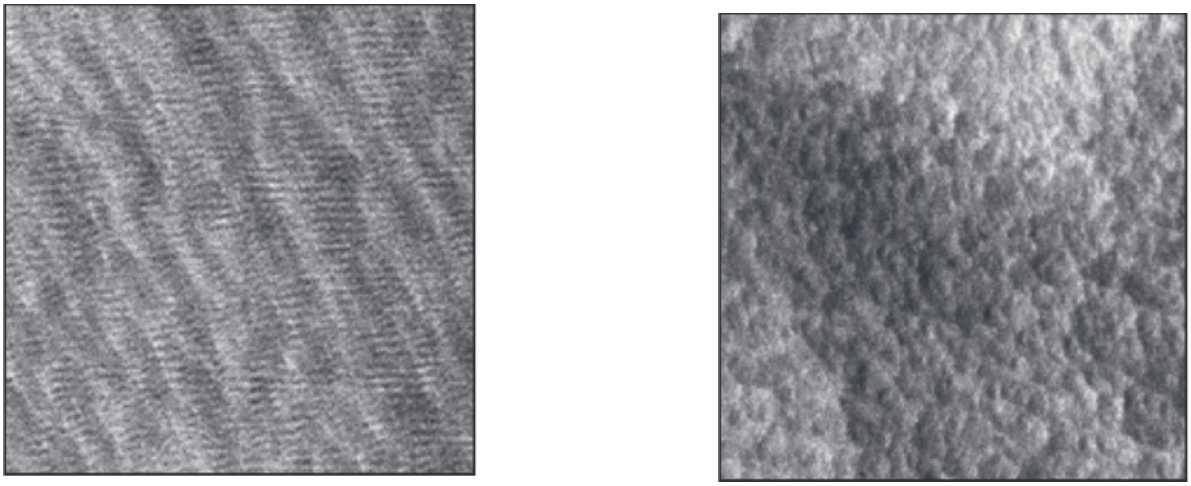
3.1.1. SAR Imagery

Sentinel-1 Mission and Data Acquisition

This study utilizes Synthetic Aperture Radar (SAR) imagery from the European Space Agency's Sentinel-1 (S-1) mission. The mission comprises two polar-orbiting satellites (Sentinel-1A and Sentinel-1B) operating in C-band with a frequency of 5.405 GHz, corresponding to a wavelength of 5.5 cm. Wave Mode acquisitions alternate between two incidence angles: WV1 at $23.8°$ and WV2 at $36.8°$, with an along-track sampling separation of 100 km. The S-1 WV images are collected in 20x20 km scenes [40]. All imagery analysed in this study was acquired in VV polarization, which accounts for more than 99% of Sentinel-1 WV acquisitions [14].

Dataset Evolution and Composition

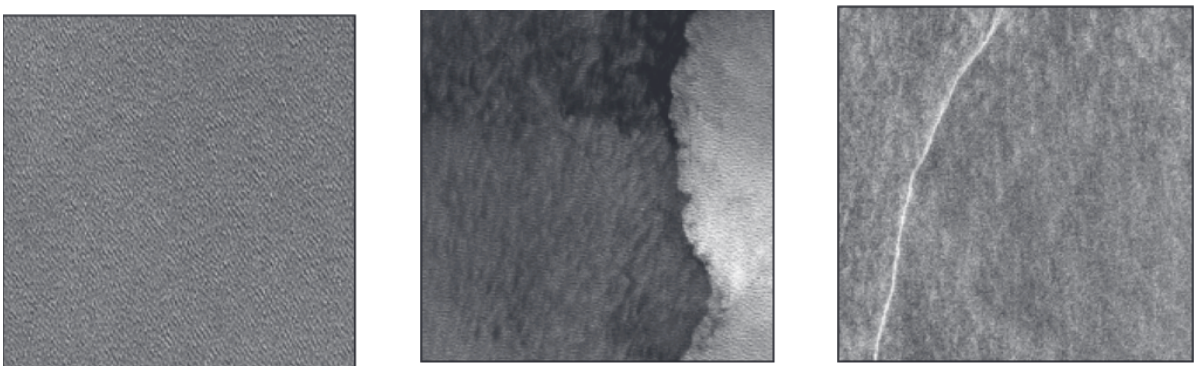
The development of the dataset proceeded in two phases, with the aim of progressively expanding the quantity and diversity of ocean scenes. Initially, a focused subset of approximately 23000 images was constructed following the classification methodology of Wang et al. [44]. This dataset was obtained from 2015 S-1A acquisitions and has exclusively wind streaks and convective cells as shown in figure 3.1 These are two ocean-atmosphere phenomena that produce distinct signatures in SAR ocean imagery. This initial dataset was partitioned into training (70%, approximately 16100 images) and validation (30%, approximately 6900 images) subsets to support initial model development and validation.



(a) Wind streaks **(b)** Convective cells

Figure 3.1: Classified SAR ocean scenes from Wang et al.[45] (a) wind streaks and (b) convective cells

The dataset was subsequently expanded to include a broader range of geophysical ocean and atmospheric phenomena by incorporating S1-A and S1-B WV data from 2021 to 2024. Following preprocessing to decompress the data, approximately 400,000 images representing diverse ocean-atmosphere phenomena were obtained. These phenomena can be classified into ten different geophysical categories, as described by [44]. To enable validation against turbulent flux targets, these images were co-located with ERA5 reanalysis products and buoy observations. During this stage, classes unsuitable for turbulence estimation, such as icebergs and sea ice, were filtered out. This resulted in a final dataset of around 220,000 images, which are stored as NetCDF (.nc) files. Each file contains the NRCS (σ_0), the incidence angle and the co-located heat flux estimates, as well as the heat flux estimates from O'Driscoll et al. [29]. As illustrated in the figure 3.2, it now also contains ocean waves, atmospheric and ocean fronts, as well as wind streaks and convective cells.



(a) Ocean waves **(b)** Atmospheric fronts **(c)** Oceanic fronts

Figure 3.2: Classified SAR ocean scene from Wang et al. [45], showing: (a) ocean waves, (b) atmospheric fronts, and (c) oceanic fronts

Image Preprocessing Pipeline

The preprocessing pipeline implements different steps adapted from established methodologies for analysing SAR ocean imagery, as described in [44, 14]. First, the incidence angle is normalised in the NRCS (σ_0) values to eliminate systematic trends of viewing geometry. Following the approach of Wang et al. [44], the NRCS is recalibrated using the CMOD5.N geophysical model function [17] to derive a measure of sea surface roughness (SSR) that is independent of incidence angle:

$$\text{SSR} = \frac{\sigma_0}{\text{CMOD5.N}(U_{10} = 10 \text{ m/s}, \theta, \phi = 45, \text{VV})} \tag{3.1}$$

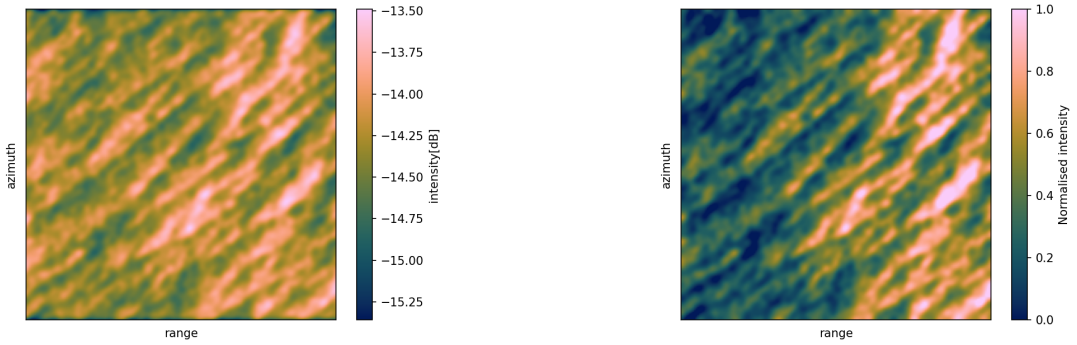
where U_{10} is the $10m/s$ neutral wind speed, θ is the radar incidence angle, and ϕ is the relative wind-platform angle.

The Sentinel-1 WV mode scenes are at $5m$ resolution but are effectively coarsened by processing artifacts such as smearing, azimuthal cut-off, and velocity bunching [8, 21, 33] and are resampled to a $100m$ resolution grid [29]. A Gaussian filter ($\sigma = 2$) is subsequently applied to reduce speckle noise.

Intensity normalization is then performed following the approach of Glaser et al. [14], mapping the SSR values to an 8-bit unsigned integer range [0, 255] using the 1st and 99th percentiles as reference points:

$$\text{SSR}_{\text{norm}} = 255 \times \frac{\text{SSR} - P_1}{P_{99} - P_1} \tag{3.2}$$

where P_{01} and P_{99} denote the 1st and 99th percentiles of the SSR distribution, respectively. Values below P_{01} are clipped to 0, and values above P_{99} are clipped to 255. The normalized images are saved as grayscale Portable Network Graphics (PNG) files for efficient storage and subsequent processing. For machine learning applications, these grayscale images are converted to floating-point arrays by dividing pixel values by 255, yielding normalized intensities in the range [0, 1](see Figure 3.3).



(a) S1 WV mode image before intensity normalization

(b) S1 WV mode image after intensity normalization

Figure 3.3: Data example of an S1 image before and after intensity normalisation

3.1.2. Ground-Truth Turbulence Parameters

For each SAR scene in the 2021-2024 dataset, two types of buoyant heat flux estimates were obtained from O'Driscoll et al. [29]. Since ERA5 reanalysis does not directly provide the buoyant heat flux, reference values (`hbb_coare`) were computed by inputting ERA5 atmospheric state variables (sea surface temperature, air temperature, humidity, wind speed, and pressure) into the COARE3.6 bulk flux algorithm. Secondly, O'Driscoll et al. provided their machine learning-derived flux predictions (`y_pred`). These predictions serve as a benchmark for comparison with the VAE-derived flux estimates developed in this study. These COARE-derived heat fluxes (`hbb_coare`) serve as the primary training targets and validation references for the VAE regression models.

3.1.3. Splitting the data

The complete dataset of approximately 220000 SAR-ERA5 co-located samples is partitioned into training (70%), validation(15%), and test(15%) subsets to enable robust model development, hyperparameter tuning, and unbiased performance evaluation. This meant approximately 154000 samples are used for VAE training, 33000 samples for validation and 33000 samples for testing.

3.2. Model Architecture

3.2.1. Base VAE Design

The baseline VAE architecture was adapted from recent applications of VAEs to SAR imagery [47]. This previous study primarily focused on ship detection [47] and differs fundamentally from large-scale ocean surface imagery in terms of scene complexity and spatial scale. Nevertheless, CNN-based encoder-decoder frameworks provide a suitable starting point for this application [13].

Development Strategy of architecture

The model development proceeded in two phases, each addressing distinct research objectives:

Phase 1: Proof of Concept (Dataset 2015) Initial experiments utilized a dataset consisting of approximately 23000 SAR images from 2015, containing two ocean-atmosphere phenomena classes as described by Wang et al[44], namely convective cells and wind streaks. This phase addressed three fundamental questions:

1. *Dimensionality Reduction:* Can a VAE effectively compress high-dimensional SAR ocean imagery into a low-dimensional latent representation while maintaining adequate reconstruction quality?
2. *Reconstruction Optimization:* How can reconstruction quality be improved through modifications to the standard VAE loss function and capture multiple frequency scales?

Phase 2: Generalization and Application (Dataset 2021–2024)

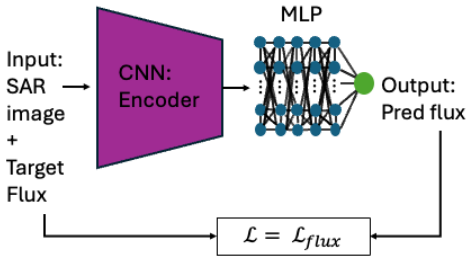
Following the successful completion of the proof of concept, the model is evaluated on a substantially larger and more diverse dataset comprising 220,000 SAR images spanning the period from 2021 to 2024. This dataset encompasses a wider variety of atmospheric and oceanic phenomena and is used to address three key questions:

1. *Generalization:* Does the VAE maintain reconstruction quality across more varied SAR scenes, and can it faithfully reconstruct all phenomenon classes?
2. *Latent Space Structure:* Does the latent space preserve discriminative information about different phenomenon classes, and does sampling from the latent space reproduce this diversity?
3. *Turbulence Parameter Regression:* Can turbulence parameters be accurately estimated from the latent representations?

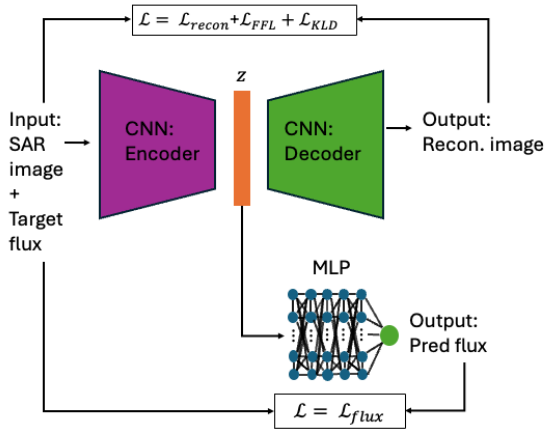
To address the third question, the ability to estimate turbulence parameters from learned representations is investigated by comparing three model configurations, illustrated schematically in Figure 3.4:

- **Model 0 — Direct CNN + MLP Regression (Baseline):** End-to-end CNN trained with a regression head to directly regress turbulence parameters from SAR images.
- **Model 1 — Two-Stage Training:** A VAE is first trained, after which its encoder is frozen and a regression head is trained on the fixed latent representations.
- **Model 2 — Joint Training:** The VAE and regression head are trained simultaneously end-to-end, allowing regression gradients to propagate through the encoder.

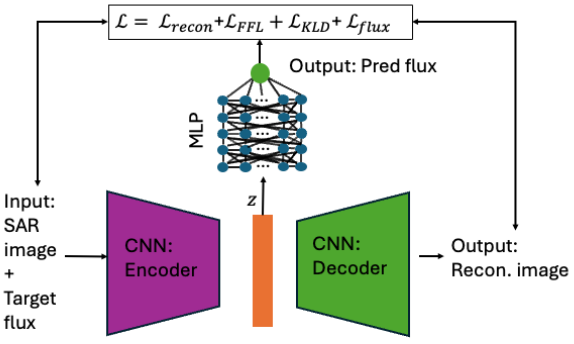
Each model uses a Multi-Layer Perceptron (MLP) as the regression head to map from a feature representation to the target heat flux estimates. The MLP is a feed-forward neural network comprising multiple layers of linear transformations followed by nonlinear activations [4]. In Models 1 and 2, the input to the MLP is the latent vector $\mathbf{z}$, sampled from the posterior distribution $q_\theta(\mathbf{z}|\mathbf{x})$ of the VAE encoder (see Section 2.2.2). In Model 0, the MLP operates directly on the feature embedding produced by the CNN encoder. The loss function components governing each model are shown in Figure 3.4 and detailed in Section 3.3.



(a) Model 0: Direct CNN + MLP Regression (Baseline).



(b) Model 1: VAE trained first, encoder frozen, regression head trained on fixed latent representations.



(c) Model 2: VAE and regression head trained jointly, with gradients from the flux loss propagating through the encoder.

Figure 3.4: Schematic overview of the three model configurations used to estimate turbulent heat fluxes from SAR imagery. Each model includes an MLP regression head and its associated loss function components. Model 0 regresses directly from CNN features; Model 1 uses a frozen VAE latent vector $\mathbf{z}$; Model 2 jointly optimises the VAE and regression objectives.

Latent Dimensionality Selection

A critical hyperparameter in VAE design is the latent dimensionality z_{dim}, which governs the information bottleneck between encoder and decoder. In the absence of an established optimization criterion for determining optimal minimum dimensionality, a categorical evaluation approach was adopted. Four discrete latent dimensions were systematically investigated: $z_{\text{dim}} \in \{32, 64, 128, 256\}$.

Encoder-Decoder Architecture

The VAE employs a symmetric encoder-decoder structure built from convolutional blocks inspired by residual network architectures [31, 47, 13]. Organising convolutional layers into blocks with batch normalisation and activation functions has proven effective for training deep networks on image tasks [31]. Each block consists of the following sequence:

1. Convolutional layer (kernel size $k \times k$, where $k \in \{3, 5, 7\}$)
2. Batch normalisation
3. ReLU activation
4. Convolutional layer (kernel size $k \times k$, where $k \in \{3, 5, 7\}$)
5. Batch normalisation
6. ReLU activation

Encoder Architecture The encoder progressively reduces spatial dimensions while increasing the number of feature channels, enabling hierarchical feature extraction from coarse to fine scales[31]. The encoder consists of:

- N_{blocks}: where N_{blocks} varies with latent dimension (as demonstrated above)
- Spatial downsampling by a factor of 2 after each block via average or max pooling [47][13]
- Zero-padding applied to maintain dimensional consistency across operations

The final encoder layer produces two outputs which define the approximate posterior distribution $q(z|x)$ [22]:

- Mean vector: $\boldsymbol{\mu} \in \mathbb{R}^{z_{\text{dim}}}$
- Log-variance vector: $\log \boldsymbol{\sigma}^2 \in \mathbb{R}^{z_{\text{dim}}}$

Decoder The decoder mirrors the encoder, progressively upsampling the latent representation via transposed convolutions to reconstruct the original image dimensions. The final layer produces a single-channel output passed through a sigmoid activation, constraining reconstructed values to $[0, 1]$.

Architectural Hyperparameters

The architectural hyperparameters, N_{blocks}, convolutional kernel sizes, feature channel progression, and pooling strategy, are optimised separately for each latent dimension z_{dim} using the Optuna framework [2], as detailed in Section 3.5. Initial choices were informed by Xu et al. [47]. During the first stage of optimisation in Optuna, it was found that the number of residual blocks, kernel sizes and feature channels were of low importance. For simplification, these can be fixed in the second stage of optimisation, after a consensus can be reached on the basic architecture for all four z_{dim}. The complete architecture is illustrated in Figure 3.5

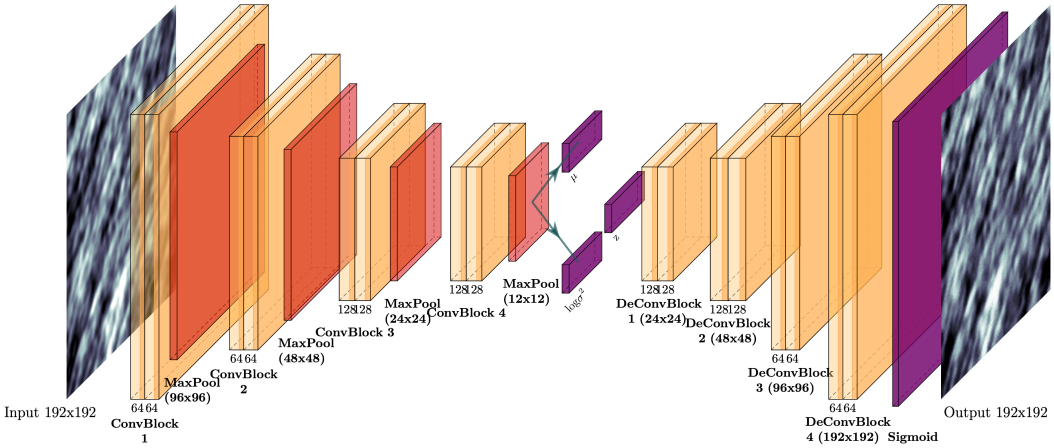


Figure 3.5: Complete VAE architecture

3.3. Loss Function Design

The choice of loss function is fundamental to VAE training, because it governs both the quality of image reconstructions and the structure of the learned latent space. This section presents the baseline VAE loss function, followed by the integration of Frequency Focal Loss. The final subsection details the loss scheduling strategy employed during training. The loss function design is first tested on the Dataset 2015 to asses reconstruction quality. The findings will be taken into phase 2 on the larger dataset (2021 – 2024).

3.3.1. Baseline VAE Loss

Loss Function Formulation

The choice of loss function is fundamental to VAE training, this governs both reconstruction quality and the regularisation of the learned latent space. This section presents the complete training objective, followed by the evaluation metrics used to assess performance, and the scheduling strategy employed during training. All configurations are first tested on the 2015 dataset before being applied to the larger 2021–2024 dataset.

3.3.2. Training Objective

Baseline VAE Loss

The standard VAE loss minimises the negative Evidence Lower Bound (ELBO) [22, 31, 13] (see Section 2.2.2):

$$\mathcal{L}_{\text{VAE}}(\theta, \phi) = \underbrace{-\mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x})}\left[\log p_\phi(\mathbf{x}|\mathbf{z})\right]}_{\text{Reconstruction Term}} + \underbrace{\beta\, D_{\text{KL}}\left(q_\theta(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\right)}_{\text{KL Divergence Term}} \tag{3.3}$$

Where θ and ϕ denote the encoder and decoder parameters respectively, and β controls the relative importance of latent space regularisation. For simplicity, this is written as follows:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\text{KL}} \tag{3.4}$$

Frequency Focal Loss Integration

To address the frequency bias, the Focal Frequency Loss (FFL) of Jiang et al. [19] is incorporated (see Section 2.3.2 for the full derivation). FFL adaptively weights frequency components based on reconstruction difficulty, concentrating optimisation on poorly-reconstructed frequencies. The complete training objective becomes:

$$\mathcal{L} = \gamma_{\text{recon}}\,\mathcal{L}_{\text{recon}} + \beta\,\mathcal{L}_{\text{KL}} + \lambda\,\mathcal{L}_{\text{FFL}} \tag{3.5}$$

where γ_{recon} is a weight on the reconstruction term (detailed in Section 3.3.3), β controls KL regularisation, and λ controls the contribution of FFL.

3.3.3. Loss Scheduling Strategy

The relative weighting of loss components is allowed to dynamically change during training. This gradually changing of the weight of a loss term is referred to as annealing.

Reconstruction Term Annealing

The weight γ_{recon} in Equation 3.5 is linearly decreased over training according to an annealing schedule, gradually shifting emphasis from the spatial reconstruction term toward FFL:

$$\gamma_{\text{recon}}(t) = 1.0 - (t/T) \tag{3.6}$$

where t is the current epoch and T is the total number of annealing epochs. The weight starts at 1 and decreases toward 0, ensuring rapid learning of overall image structure early in training while allowing FFL to dominate in later epochs when fine frequency detail becomes the focus.

Experimental Configurations

The following configurations are systematically evaluated on the 2015 dataset:

1. **Baseline**: $\gamma_{\text{recon}} = 1.0$, $\lambda = 0$ (no FFL)
2. **Fixed FFL**: $\gamma_{\text{recon}} = 1.0$, $\lambda = 1.0$, for $\alpha \in \{1, 2, 4, 6\}$
3. **Annealing + FFL**: Equation 3.6 with $T_{\text{anneal}} \in \{10, 20, 30, 40\}$ epochs, $\lambda = 1.0$

3.4. β-Annealing Strategy

This study follows the methodology of Adhikari et al. [1] as a way to optimize β, the weight of the KL divergence term, by gradually increasing it during training. The primary objective is to identify the optimal β value that maximizes variance retention in the latent representation while avoiding posterior collapse. The β-Annealing schedule is first evaluated on the 2015 dataset to assess its effect on latent space organisation, before being applied to the larger 2021–2024 dataset.

3.4.1. Determination of Optimal β using FVE

The Fraction of Variance Explained (FVE, Equation 2.15) is used as the primary metric to determine the optimal β value. Computed on the validation set after each training epoch, FVE quantifies the proportion of input variance retained in the latent representation. The VAE is trained while slowly annealing β, and the FVE score is monitored at each epoch. The β value corresponding to the maximum FVE is identified as the optimal value β_{opt}. Annealing follows the cosine schedule given in Equation 2.14.

3.4.2. Training Protocol

The training protocol consists of two phases:

Phase 1: Optimal β identification

The VAE is trained with β annealed from 0 to a specified maximum over a fixed number of epochs. Following Adhikari et al. [1] a slow annealing over 500 epochs is used for the 2015 dataset. On the larger 2021–2024 dataset, the greater computational cost per epoch makes it necessary to reduce the schedule to 100 epochs. Furthermore, experimentation showed that the effective range of β for this dataset is substantially smaller than in Adhikari et al.; to achieve a maximum FVE the upper annealing limit is set to $\beta_{\text{max}} = 1 \times 10^{-3}$ rather than 1.

To verify that FVE competently captures latent space regularisation, the distribution of each latent dimension is examined at multiple training stages by plotting every 4 epochs. Three behaviours are expected whilst slowly increasing β:

1. Early training (low β): Distributions are expected to be non-Gaussian as the model prioritizes reconstruction
2. In the range of β_{opt}: Distributions should approximate a standard normal distribution $\mathcal{N}(0, 1)$, indicating proper regularization without collapse
3. Beyond β_{opt} we should see that distributions of different latent dimension start to collapse towards the mean, indicating posterior collapse

Phase 2: Training with optimal β

The VAE is retrained using cosine annealing from 0 to β_{opt} over the same number of epochs as Phase 1. Reconstructed and sampled images are then compared against those from a standard VAE trained with a fixed β to evaluate the benefit of the annealing approach.

3.4.3. Evaluation Metrics

To assess VAE performance, both spatial- and frequency-domain metrics are employed.

Spatial-Domain Metrics

- **PSNR** (higher is better): Measures pixel-wise reconstruction accuracy in decibels:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}_I^2}{\text{MSE}}\right) \tag{3.7}$$

  where MAX_I is the maximum pixel value [18].
- **SSIM** (higher is better): Evaluates perceptual similarity considering luminance, contrast, and structure [45].

Frequency-Domain Metrics

- **Log Frequency Distance (LFD)** (lower is better): Introduced by Jiang et al. [19] to quantify frequency-domain reconstruction error:

$$\text{LFD} = \log\left[\frac{1}{MN}\sum_{u,v}|F_o(u,v) - F_r(u,v)|^2 + 1\right] \tag{3.8}$$

  where F_o and F_r are the Fourier transforms of the original and reconstructed images respectively.

In addition to these quantitative metrics, 2D power spectra of original and reconstructed images are visualised to directly observe any frequency-domain gap:

$$P(u,v) = |F(u,v)|^2 \tag{3.9}$$

For each configuration, PSNR, SSIM, and LFD are reported on the validation set, alongside training curves and frequency spectrum visualisations. The optimal configuration is then carried forward to the 2021–2024 dataset.

3.5. Optimization with Optuna

3.5.1. Overview

Hyperparameter optimisation is conducted using the Optuna framework [2] to systematically tune the VAE models across four latent dimensions $z_{\text{dim}} \in \{32, 64, 128, 256\}$. Given the computational cost of training deep CNN-based VAEs on large datasets, a two-stage strategy is employed: the first stage identifies which hyperparameters most influence performance, and the second stage concentrates resources on fine-tuning only those hyperparameters.

3.5.2. Staged Optimisation strategy

Stage 1: Importance screening

Thirty trials are run on 50% of the training data, sampling across the full hyperparameter space. Optuna's parameter importance analysis quantifies each hyperparameter's contribution to the objective

function. Parameters with low importance are subsequently fixed to the values from the best-performing trial, reducing the search space for Stage 2.

Stage 2: Final Optimization

The remaining high-importance hyperparameters are optimised on the full training dataset (100%). An initial 50 trials are performed, after which convergence is assessed by inspecting the progression of the best trial objective value across trials. If the best objective has stabilised, meaning that subsequent trials no longer improve upon it, the search is considered converged. Since Optuna stores old trials, additional trials can be appended seamlessly if convergence has not yet been reached at 50 trials, without discarding prior evaluations.

Reproducibility and Pruning

To ensure reproducibility, a fixed random seed was employed throughout all optimization experiments. The Tree-structured Parzen Estimator (TPE) algorithm was used as the sampling strategy for hyperparameter selection, providing an efficient balance between exploration and exploitation. Optuna's pruning functionality was employed to terminate underperforming trials early. This mechanism, which implements a variant of the Successive Halving algorithm[2], monitors intermediate validation loss values and prunes trials that show no promise of outperforming the current best configuration. This approach significantly reduced total computation time while maintaining search quality.

3.5.3. Search Space Definition

The hyperparameters being optimised span the architecture, training configuration, and loss function weights for each model. Tables 3.1–3.3 summarise the full search space per model.

Table 3.1: Model 0 (Direct CNN Regression) hyperparameters.

Parameter	**Type**	**Search Space**
CNN Architecture		
CNN blocks	Categorical	$\{3, 4, 5, 6\}$
Kernel size	Categorical	$\{3, 5, 7\}$
Pooling type	Categorical	{Max, Avg}
Batch normalisation	Categorical	{True, False}
Base channels	Categorical	$\{32, 64\}$
Embedding dim	Categorical	$\{1, 8, 16\}$
MLP Architecture		
Number of layers	Categorical	$\{4, 5, 6, 7, 8, 9\}$
Hidden dimension	Categorical	$\{64, 128, 256\}$
Dropout	Float	$[0.0,\ 0.5]$
Training		
Learning rate	Float	$[10^{-5},\ 10^{-3}]$
Weight decay	Float	$[10^{-6},\ 10^{-3}]$
Batch size	Categorical	$\{16, 32, 64\}$

Table 3.2: VAE architecture and training hyperparameters (Model 1 and Model 2).

Parameter	Type	Search Space
CNN Architecture		
CNN blocks	Categorical	$\{3, 4, 5, 6\}$
Batch normalisation	Categorical	{True, False}
Feature channels	Categorical	$\{32, 64, 128\}$
Pooling type	Categorical	{Max, Avg}
MLP Regression Head		
Number of layers	Categorical	$\{3, 4, 5, 6\}$
Hidden dimensions	Categorical	$\{32, 64, 128, 256\}$
Weight decay	Float	$[10^{-6}, 10^{-3}]$
Dropout	Float	$[0.0, 0.5]$
Training		
Learning rate	Float	$[10^{-5}, 10^{-3}]$
Optimiser	Categorical	{Adam, AdamW, SGD}
Batch size	Categorical	$\{16, 32, 64\}$

Table 3.3: Loss function weight hyperparameters (Model 1 and Model 2).

Parameter	Type	Search Space
β (KL weight)	Float	$[10^{-8}, 10^{-4}]$
λ (FFL weight)	Float	$[1, 3]$
α (flux weight, Model 2 only)	Float	$[10^{-6}, 10^{-3}]$
γ_{max} (cool-down start)	Float	$[0.5, 1.5]$
γ_{min} (cool-down end)	Float	$[0.01, 0.1]$

The loss function weights β, λ, and α are treated as hyperparameters directly. For the reconstruction cool-down schedule (Section 3.3.3), the decreasing slope of γ_{recon} was found to influence frequency reconstruction quality. Rather than fixing the start and end values as in the initial exploration (where γ_{recon} decayed from 1.0 to 0.1), Optuna is used to determine the optimal start value γ_{max} and end value γ_{min}, generalising Equation 3.6 to:

$$\gamma_{\text{recon}}(t) = \gamma_{\text{max}} - (\gamma_{\text{max}} - \gamma_{\text{min}}) \cdot (t/T) \tag{3.10}$$

3.5.4. Model-Specific Loss Functions and Objectives

Models 1 and 2 differ in their training objectives. Model 1 trains the VAE and regression head sequentially, whereas Model 2 trains them jointly. Their respective loss functions are:

$$\mathcal{L}_1 = \gamma_{\text{recon}}\,\mathcal{L}_{\text{recon}} + \beta\,\mathcal{L}_{\text{KL}} + \lambda\,\mathcal{L}_{\text{FFL}} \tag{3.11}$$

$$\mathcal{L}_2 = \gamma_{\text{recon}}\,\mathcal{L}_{\text{recon}} + \beta\,\mathcal{L}_{\text{KL}} + \lambda\,\mathcal{L}_{\text{FFL}} + \alpha\,\mathcal{L}_{\text{flux}} \tag{3.12}$$

Objective Functions

During β-annealing (see Section 3.4), the FVE score proved to be an effective indicator of the optimal β,and is therefore incorporated into the Optuna optimisation scheme. FVE is an R^2-type metric, bounded above by 1 and requiring maximisation. In order to maintain a consistent objective

that can be easily combined for Model 2, where reconstruction quality, latent space regularisation and flux prediction must be optimised jointly, all terms are formulated as R^2-based metrics. Since $R^2 = 1 - \text{MSE}/\text{Var}(y)$, maximising R^2 is equivalent to minimising MSE normalised by the target variance, ensuring that each term contributes on the same scale. Using just the MSE for the flux term would introduce inconsistent scaling in the composite objective. To maintain consistency, R^2_{flux} , is also adopted as the objective for Model 0 and Model 1 Stage 2 (the regression models), enabling direct comparison of flux prediction performance across all three models.

Model 0 — Direct CNN Regression (Baseline)

Maximise the coefficient of determination for flux prediction directly from SAR images:

$$\text{Objective} = \max\left(R^2_{\text{flux}}\right) \tag{3.13}$$

As Model 0 has no generative objective, only regression performance is optimised. This model serves as an upper-bound reference for flux prediction quality.

Model 1, Stage 1 — VAE Training

Maximise the combined reconstruction quality in both spatial and frequency domains:

$$\text{Objective} = \max\left(\text{FVE} + R^2_{\text{freq}}\right) \tag{3.14}$$

where R^2_{freq} measures the proportion of explained variance in the frequency domain:

$$R^2_{\text{freq}} = 1 - \frac{\sum(F_o - F_r)^2}{\sum(F_o - \bar{F}_o)^2} \tag{3.15}$$

Model 1, Stage 2 — Flux Regression

Maximise the coefficient of determination for flux prediction:

$$\text{Objective} = \max\left(R^2_{\text{flux}}\right) \tag{3.16}$$

R^2_{flux} is used rather than minimising MSE because its normalised, dimensionless nature allows consistent comparison across different flux variables and trial configurations, and a value of 1 provides an interpretable upper bound on prediction quality.

Model 2 — Joint VAE + Regression

Combine all objectives:

$$\text{Objective} = \max\left(\text{FVE} + R^2_{\text{freq}} + R^2_{\text{flux}}\right) \tag{3.17}$$

4

Results

This chapter presents experimental results in two phases. Phase 1 (Section 4.1) establishes proof-of-concept using the 2015 dataset, investigating loss function design and reconstruction quality across four latent dimensions: $z_{\text{dim}} \in \{32, 64, 128, 256\}$. Phase 2 (Section 4.2) extends to the larger 2021–2024 dataset, hyperparameter search outcomes, and final model performance for both Model 1 (two-stage training) and Model 2 (joint training).

4.1. Phase 1: Proof of Concept (Dataset 2015)

Phase 1 experiments utilized approximately 23,000 SAR images from 2015 containing convective cells and wind streaks. This phase systematically evaluated baseline VAE performance, the effect of Frequency Focal Loss (FFL) and the different α values for the focus weight matrix, and reconstruction term cooling-down strategies.Throughout the experiments in Phase 1, the FFL loss weight was held fixed at $\lambda_{\text{FFL}} = 1$, following the default setting of Jiang et al. [19]. The parameter varied across experiments was α, the scaling factor of the dynamic spectrum weight matrix $w(u, v) = |F_r(u, v) - F_f(u, v)|^{\alpha}$, which controls the degree to which the loss focuses on hard-to-reconstruct frequency components

4.1.1. Baseline Performance and Frequency Focal Loss Effects

The baseline VAE with standard ELBO loss (MSE + KL divergence) achieved PSNR of 28.46, SSIM of 0.45, and LFD of 16.95 for $z_{\text{dim}} = 256$ (Table 4.1). Three values of the focal parameter $\alpha \in \{1, 2, 6\}$ were evaluated to assess adaptive focusing on difficult-to-reconstruct frequencies. Spatial-domain metrics (PSNR, SSIM) showed minimal variation across α values, with PSNR ranging from 28.42 to 28.44 and SSIM from 0.45 to 0.46.

Figure 4.2 displays validation loss components during training. The FFL term decreased with increasing α, confirming stronger weighting of poorly-reconstructed frequencies, though total validation loss remained elevated due to the additional loss term. Training stability analysis revealed KLD spikes at $\alpha = 4$, leading to the selection of $\alpha = 2$ for subsequent experiments as this appeared more stable during training.

4.1.2. Annealing of the Reconstruction Term

The reconstruction weight γ_{recon} was decreased linearly from 1.0 to 0.1 over annealing periods of 10, 20, 30, and 40 epochs, while the FFL weight λ_{FFL} was held constant. Table 4.1 presents evaluation metrics at the epoch of minimum total validation loss. PSNR increased modestly from 28.32 (10 epochs) to 28.35 (40 epochs), while SSIM increased from 0.42 to 0.45 with longer annealing duration.

Figure 4.3 shows the validation loss trajectories for all annealing configurations alongside the fixed-weight baseline. The total validation loss was consistently lower for the annealed configurations than for the baseline. The KL divergence validation loss decreased across all configurations over the course of training.

Figure 4.1 shows representative reconstructions for $z_{\text{dim}} = 256$ across the loss configurations evaluated in this section. Progressing from the baseline to the addition of the FFL term ($\alpha = 2$) and then to the 10- and 40-epoch annealing schedules demonstrates an improvement in the preservation of fine detail, which is consistent with the frequency-domain metrics reported in Section 4.1.3.

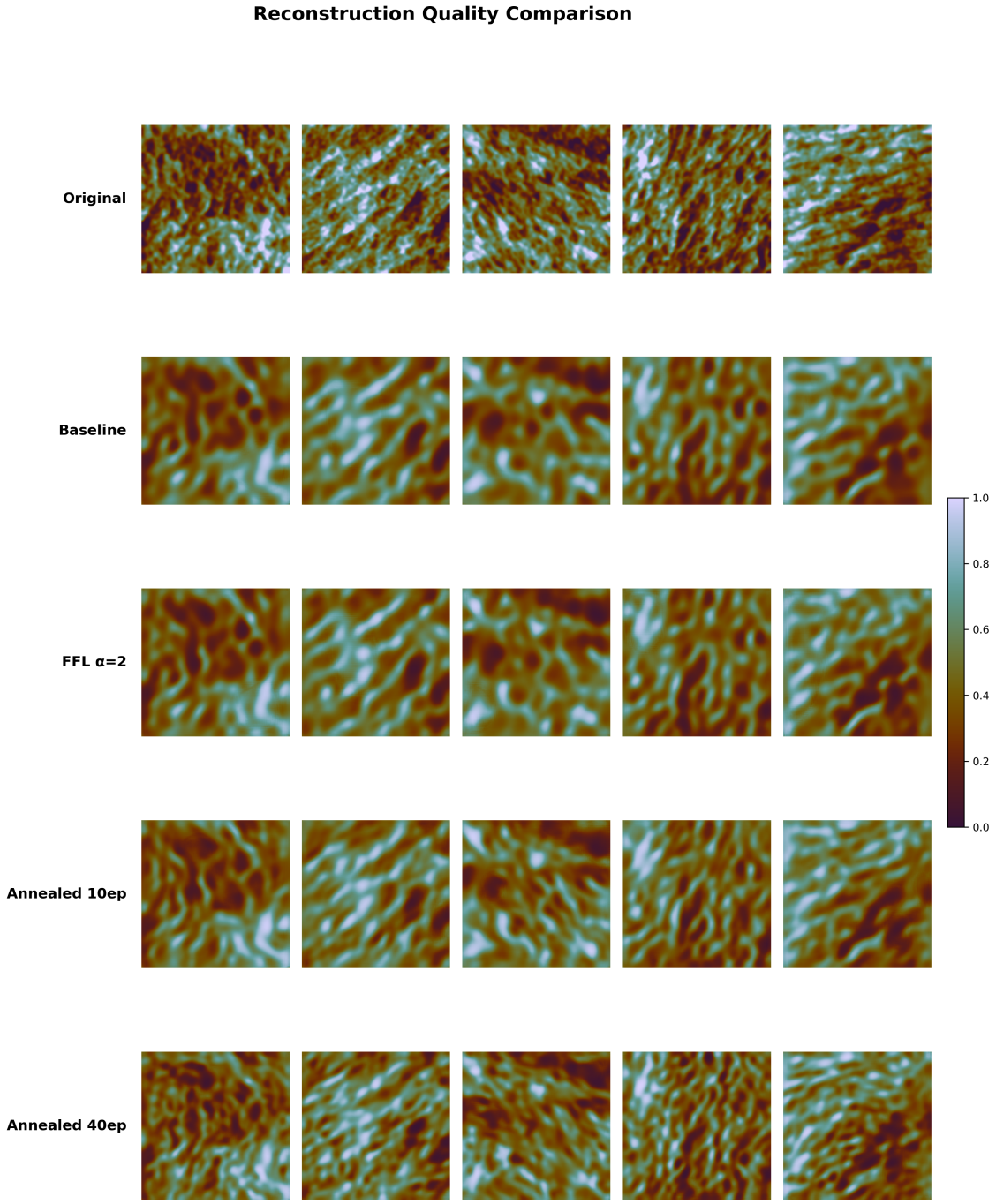
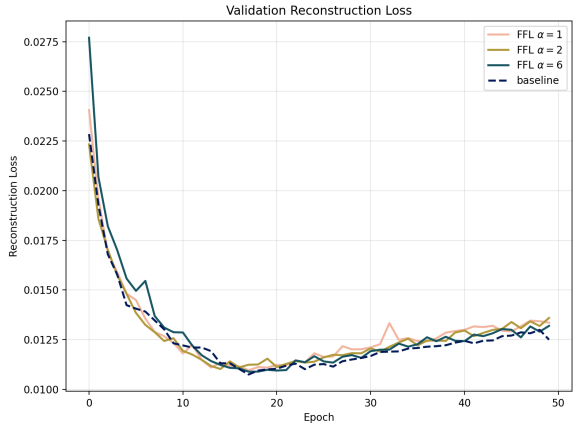
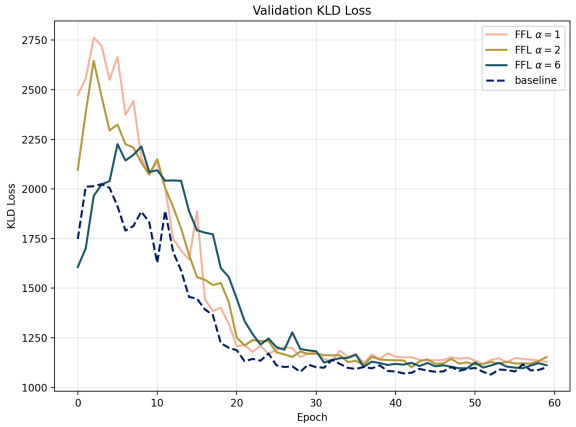


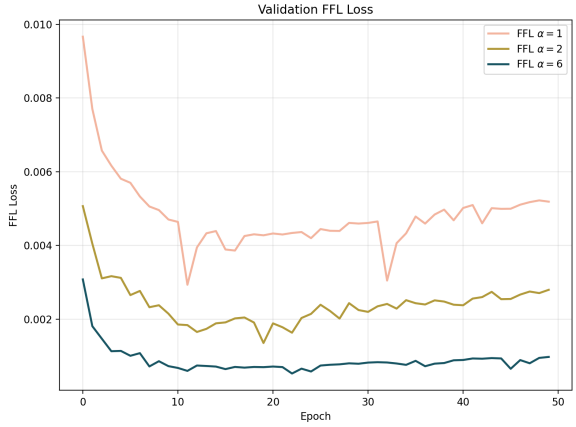
Figure 4.1: Representative reconstruction for $z_{\text{dim}} = 256$ across loss configurations. Row 1: Original. Row 2: Baseline. Row 3: FFL ($\alpha = 2$). Row 4: FFL + 10-epoch annealing. Row 5: FFL + 40-epoch annealing. Visual improvement in fine detail preservation correlates with frequency-domain metrics.



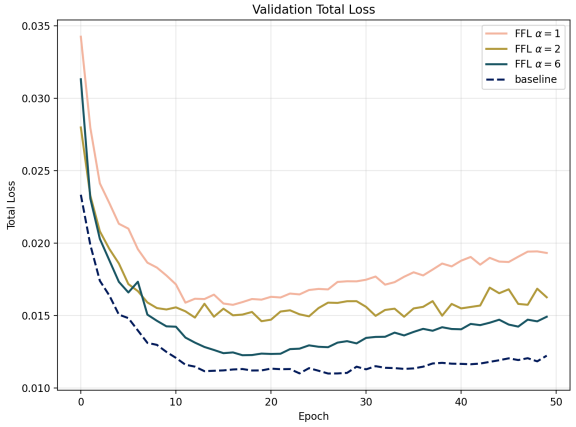
(a) reconstruction loss (MSE)



(b) KL divergence loss

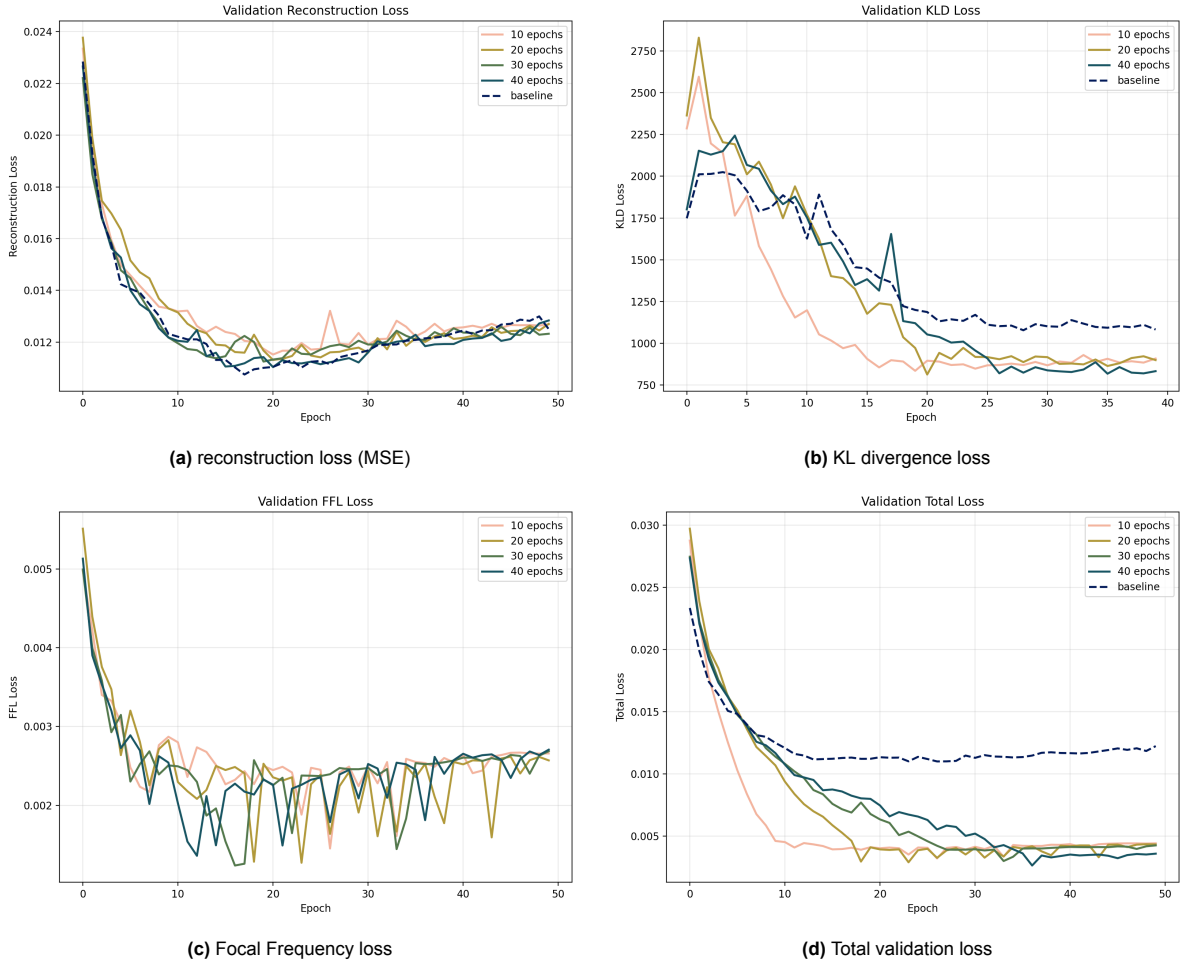


(c) Focal Frequency loss



(d) Total validation loss

Figure 4.2: Validation losses for varying α in the FFL term (experiments of $z_{dim} = 256$)



(a) reconstruction loss (MSE)

(b) KL divergence loss

(c) Focal Frequency loss

(d) Total validation loss

Figure 4.3: Validation losses for reconstruction term annealing over varying durations (experiments of $z_{dim} = 256$)

Table 4.1: Reconstruction metrics achieved at minimum validation loss for $z_{\text{dim}} = 256$. Despite minimal PSNR/SSIM variation, frequency-domain analysis reveals substantial differences (Section 4.1.3).

Experiment	PSNR (dB)	SSIM	LFD
Baseline	**28.46**	0.45	16.96
10 epochs	28.32	0.42	17.01
20 epochs	28.33	0.41	17.01
30 epochs	28.33	0.44	16.98
40 epochs	28.35	0.45	16.97
FFL $\alpha = 1$	28.44	**0.46**	16.97
FFL $\alpha = 2$	28.44	0.46	**16.94**
FFL $\alpha = 6$	28.43	0.45	16.98

4.1.3. Frequency-Domain Analysis

Figure 4.4 directly visualizes the frequency gap between original and reconstructed images through 2D power spectra. The baseline (leftmost) shows strong attenuation of high frequencies (radially outward from center), progressively improving through FFL integration, 10-epoch annealing, and finally 40-epoch annealing (rightmost). The corresponding spatial reconstructions (top row) confirm that spectral improvements translate to visible preservation of fine details and overlapping patterns.

The high-frequency preservation is assessed through energy ratio and correlation metrics (Table 4.2). The 40-epoch annealing schedule achieves optimal balance with energy ratio 0.82 and the highest

correlation of all experiments 0.91.

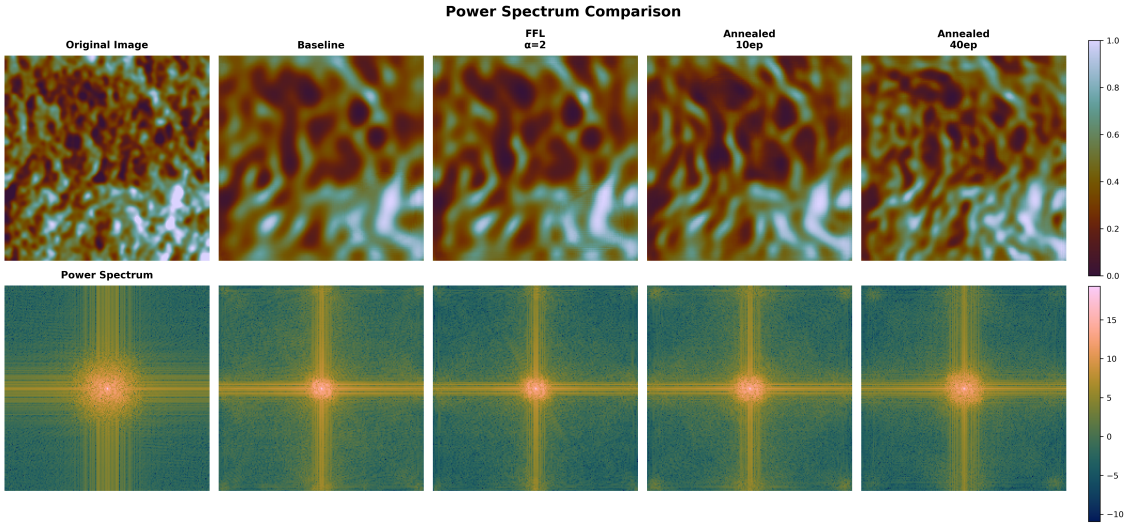


Figure 4.4: Power spectra comparison for $z_{\text{dim}} = 256$. Top row: reconstructed images. Bottom row: corresponding power spectra. From left to right: Baseline, FFL ($\alpha = 2$), FFL + 10-epoch annealing, FFL + 40-epoch annealing. The frequency gap decreases progressively with FFL integration and extended annealing.

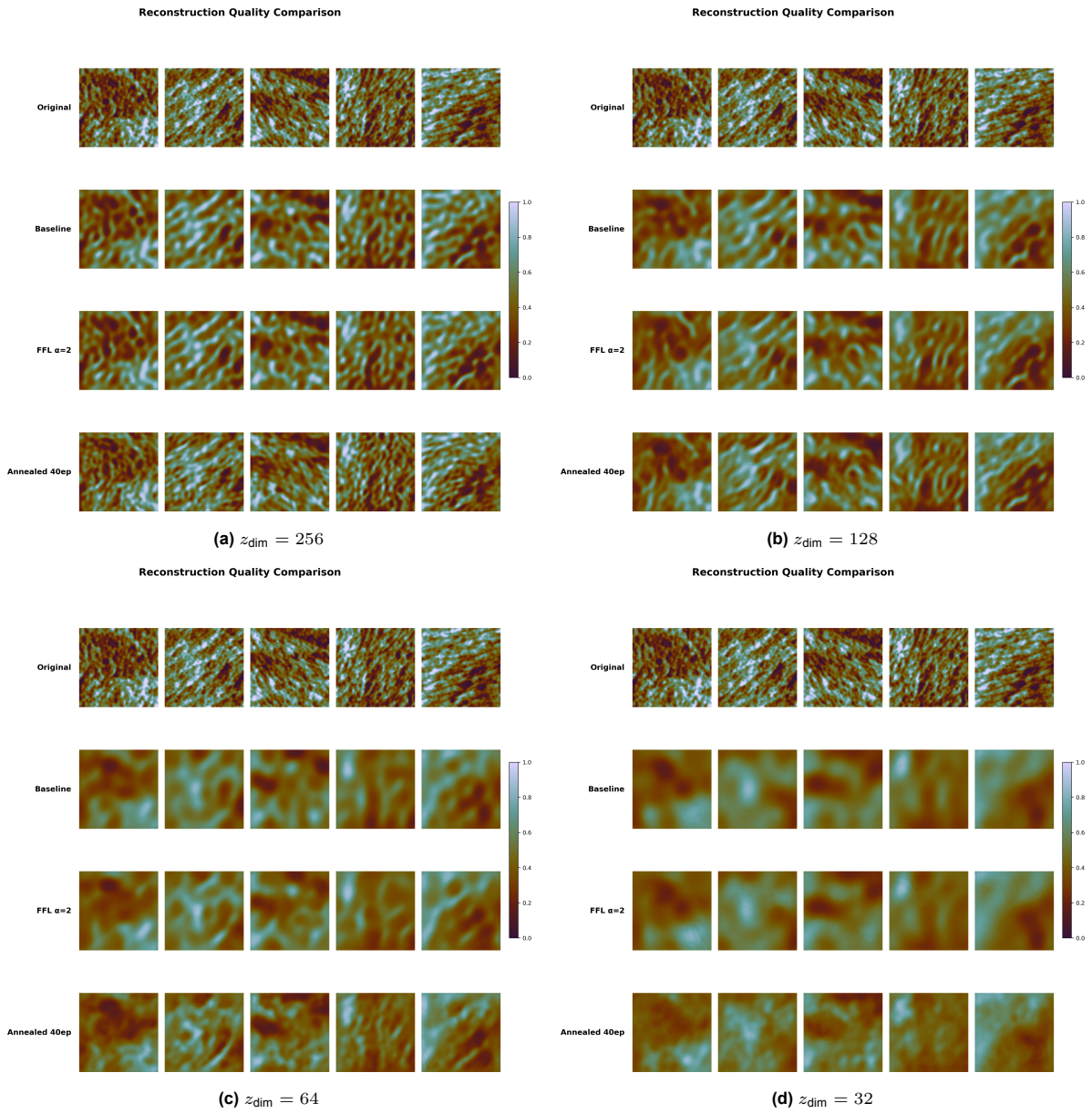
Table 4.2: High-frequency band metrics for $z_{\text{dim}} = 256$. Energy ratio measures proportion of original high-frequency energy preserved; correlation assesses pattern similarity. Bold values indicate optimal performance.

Experiment	Energy Ratio	Correlation
Baseline	0.71	0.81
$\alpha = 1$	0.70	0.88
$\alpha = 2$	0.73	0.89
$\alpha = 6$	0.67	0.90
Annealed (10 epochs)	1.09	0.86
Annealed (20 epochs)	0.70	0.89
Annealed (40 epochs)	**0.82**	**0.91**

4.1.4. Effects Across Latent Dimensions

Figure 4.5 presents reconstruction comparisons across all four latent dimensions for a representative test image. Higher dimensions ($z_{\text{dim}} \in \{128, 256\}$) produce sharper reconstructions with better fine-scale structure preservation. Lower dimensions capture large-scale patterns but lose detail, with $z_{\text{dim}} = 32$ producing notably blurry outputs. Importantly, the progression from baseline to adding the FFL term to annealing shows consistent improvement across all dimensions.

Complete reconstruction galleries showing 50 test images per dimension are provided in Appendix A.1.1, demonstrating consistent performance patterns across diverse scene types.



(a) $z_{\text{dim}} = 256$ **(b)** $z_{\text{dim}} = 128$

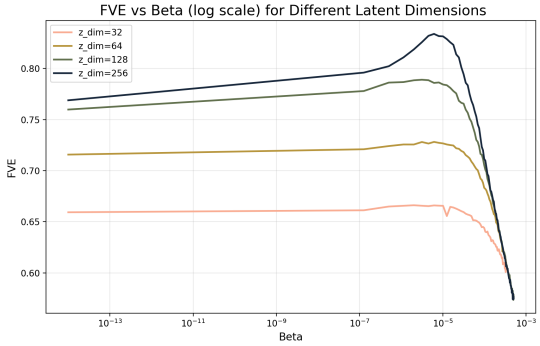
(c) $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 32$

Figure 4.5: Reconstruction comparison across latent dimensions for a representative test image. Each panel shows original (row 1), baseline (row 2), FFL (row 3), and annealed configurations. Reconstruction quality degrades with lower dimensions, but FFL + annealing consistently improves detail preservation across all z_{dim} values.

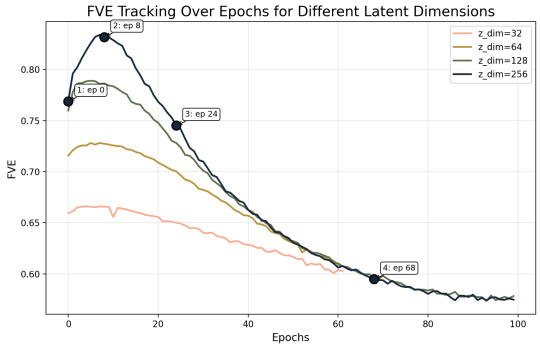
4.1.5. β-Annealing Optimization

Figure 4.6 presents the Fraction of Variance Explained (FVE) as a function of β during cosine annealing from 1×10^{-14} to 1×10^{-3} over 100 epochs. While $z_{\text{dim}} = 256$ exhibits a pronounced peak with a clearly identifiable optimal β, lower dimensions display flatter curves over a broader range of β values, making the optimal β less precisely defined. The optimal β values identified via FVE maximization are reported in Table 4.3; for $z_{\text{dim}} \in \{32, 64, 128\}$, a range is given rather than a single value, reflecting the broader plateau in FVE.

The right panel of Figure 4.6 marks four training epochs (1, 8, 24, 60) for $z_{\text{dim}} = 256$, corresponding to latent space distributions that are visualised in Figure 4.7. These distributions evolve from non-Gaussian at epoch 1 to approximate standard normal near the optimal β, and collapse progressively at higher β values, consistent with the FVE trajectory.



(a) FVE vs. β (log scale)

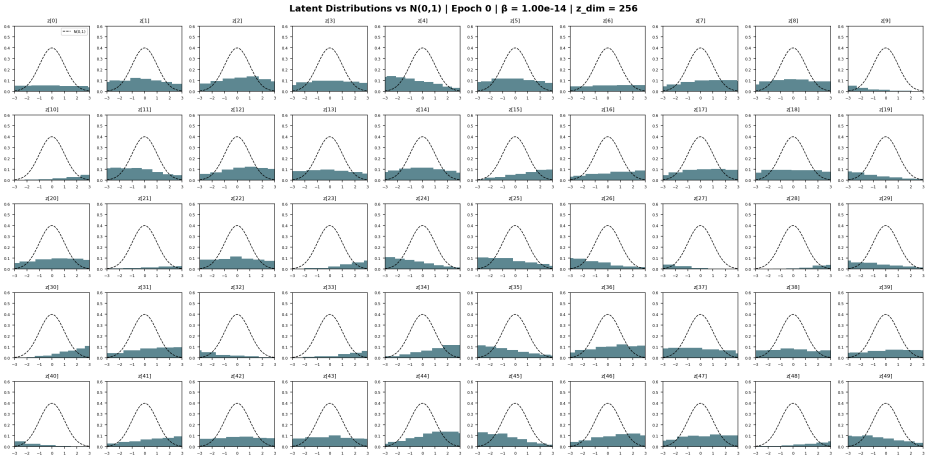


(b) FVE vs. training epoch

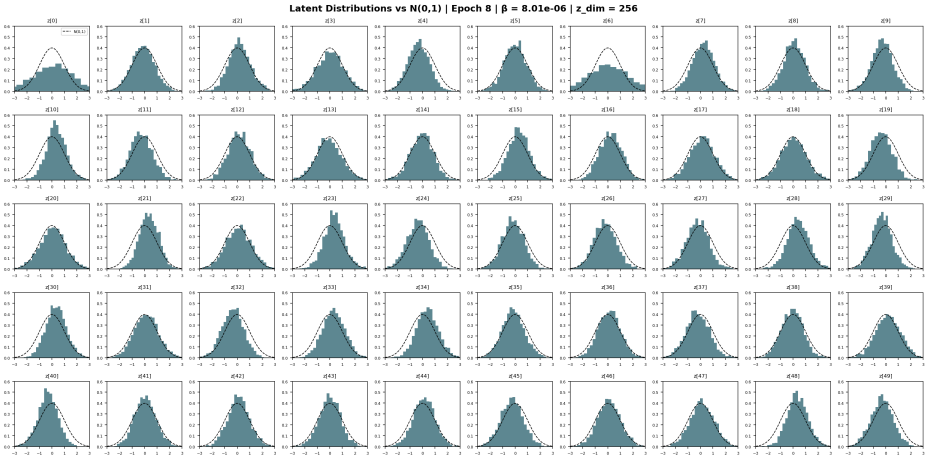
Figure 4.6: Fraction of Variance Explained during β-annealing. (a) FVE vs. β reveals distinct optima for each latent dimension. (b) FVE vs. epoch shows peak values occurring early in training. Marked epochs for $z_{\text{dim}} = 256$ correlate with distribution visualizations in Appendix 4.7

Table 4.3: Optimal β values identified via FVE maximization during cosine annealing. Higher-dimensional latent spaces achieve higher maximum FVE at lower optimal β values.

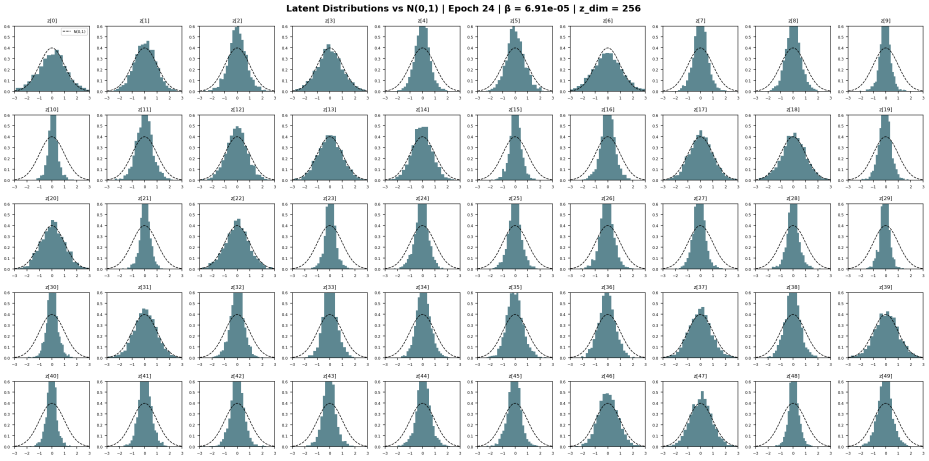
z_{dim}	β_{opt}	Max FVE
32	$3 \times 10^{-7} - 2 \times 10^{-5}$	0.67
64	$9 \times 10^{-7} - 1 \times 10^{-5}$	0.73
128	$5 \times 10^{-7} - 9 \times 10^{-6}$	0.79
256	2.0×10^{-6}	0.83



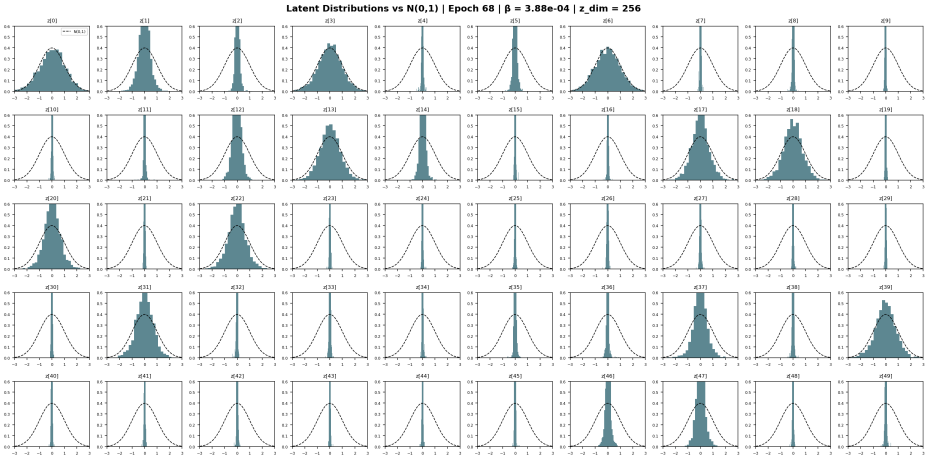
(a) 1: Epoch 0



(b) 2: Epoch 8



(c) 3:Epoch 24



(d) 4: Epoch 68

Figure 4.7: Evolution of latent space distributions for a model with $z_{\text{dim}} = 256$ during β annealing. The latent distributions correspond to the epochs highlighted in the plot 4.6b

Following β_{opt} identification, models were retrained with cosine annealing from 1×10^{-14} to β_{opt} over 80 epochs. Reconstructions , latent space samples, and final distributions are presented in the section of the Appendix A.1.2.

4.2. Phase 2: Generalization and Application (Dataset 2021--2024)

Phase 2 experiments utilised the expanded 2021–2024 dataset comprising approximately 220,000 SAR images with diverse geophysical phenomena. This phase presents systematic hyper parameter search results and final performance for all three model configurations: Model 0 (direct CNN regression), Model 1 (two-stage VAE), and Model 2 (joint training).

4.2.1. Hyper parameter Optimization Results

Model 0: Direct CNN Regression

Hyper parameter optimization for Model 0 employed a single-stage Optuna search over 60 trials, covering CNN architectural parameters (number of convolutional blocks, kernel size, pooling type, batch normalisation) and MLP regressor parameters (number of layers, hidden dimension, dropout, weight decay), alongside training parameters (learning rate, batch size). The best hyperparamters found in the searcg are reported in Table 4.4.

Table 4.4: Optimised hyper parameters for Model 0 (direct CNN regression) identified via Optuna search.

Category	Parameter	Optimised Value
CNN Architecture	`n_blocks`	4
	`pooling_type`	max
MLP Regressor	`n_layers`	5
	`hidden_dim`	128
	`dropout`	0.001
	`weight_decay`	1.31×10^{-6}
Training	`lr`	1×10^{-4}
	`batch_size`	16

Model 1 and Model 2: VAE-Based Configurations

For Models 1 and 2, hyper parameter optimization employed a two-stage Optuna strategy. Stage 1 (50% data, 30 trials) explored the full search space (Tables A.1–A.3) and identified low-importance parameters via importance analysis. These parameters were fixed for Stage 2 (100% data, $\pm$60 trials), focusing computational resources on the most influential hyper parameters.

4.2.2. Hyper parameter Importance Analysis

Across both models, training and loss function parameters have the highest importance, while architectural parameters (number of blocks, pooling type, kernel size) consistently ranked low. Consequently, architecture was fixed for Stage 2, with all trials converging to 4 residual blocks with max pooling. The best hyper parameters found are documented in Tables 4.5–4.7

For the Stage 2 MLP regression (Model 1), the most influential parameters were number of layers, optimizer, learning rate, and hidden dimension. Weight decay, dropout, and batch size were fixed based on Stage 1 findings. The best found hyper paramters can be found in Table 4.6

Table 4.5: Optimised hyper parameters for Model 1 (VAE, Stage 1) per latent dimension, identified via Optuna search. Fixed parameters are shared across all z_{dim} values.

Parameter	$z_{\text{dim}} = 32$	$z_{\text{dim}} = 64$	$z_{\text{dim}} = 128$	$z_{\text{dim}} = 256$
Optimised parameters				
β	1.62×10^{-6}	3.70×10^{-7}	2.18×10^{-7}	6.34×10^{-7}
λ_{FFL}	1.055	1.241	1.261	2.206
$\gamma_{\text{recon,max}}$	1.290	1.828	1.522	1.567
$\gamma_{\text{recon,min}}$	0.075	0.097	0.057	0.045
`lr`	4.42×10^{-4}	1.81×10^{-4}	2.35×10^{-4}	1.16×10^{-4}
`batch_size`	32	32	32	32
`optimizer`	AdamW	Adam	Adam	AdamW
Fixed parameters (all z_{dim})				
`n_blocks`	4			
`channels`	[64, 64, 128, 128]			
`kernel_size`	5			
`pooling_type`	max			
`use_batchnorm`	True			

Table 4.6: Optimised hyperparameters for Model 1 regression head (Stage 2) per latent dimension. The VAE encoder was frozen and only regression-specific parameters were tuned.

Parameter	$z_{\text{dim}} = 32$	$z_{\text{dim}} = 64$	$z_{\text{dim}} = 128$	$z_{\text{dim}} = 256$
Optimised parameters				
`n_reg_layers`	5	5	5	5
`reg_hidden_dim`	96	128	96	96
`reg_dropout`	0.100	0.139	0.119	0.118
`lr`	3.60×10^{-4}	2.28×10^{-4}	1.36×10^{-4}	2.30×10^{-4}
`optimizer`	Adam	Adam	Adam	Adam
`weight_decay`	6.80×10^{-6}	6.99×10^{-5}	1.48×10^{-4}	2.61×10^{-5}
`batch_size`	64	64	64	64

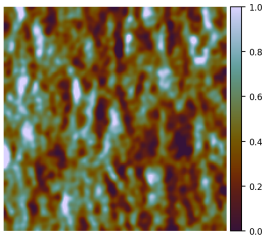
Table 4.7: Optimised hyperparameters for Model 2 (joint training, Stage 1) per latent dimension, identified via Optuna search. Fixed parameters are shared across all z_{dim} values.

Parameter	$z_{\text{dim}} = 32$	$z_{\text{dim}} = 64$	$z_{\text{dim}} = 128$	$z_{\text{dim}} = 256$
Optimised parameters				
β	1.57×10^{-5}	7.17×10^{-6}	2.17×10^{-6}	1.60×10^{-7}
λ_{FFL}	2.814	2.223	1.233	1.434
$\gamma_{\text{recon,max}}$	0.787	0.753	0.881	0.728
$\gamma_{\text{recon,min}}$	0.099	0.080	0.024	0.054
`lr`	1.53×10^{-4}	8.48×10^{-5}	1.43×10^{-4}	2.22×10^{-4}
`reg_dropout`	0.112	0.100	0.127	0.168
α_{flux}	1.041	1.868	1.115	1.315
`reg_hidden_dim`	128	128	128	128
Fixed parameters (all z_{dim})				
`n_blocks`	4			
`channels`	[64, 64, 128, 128]			
`kernel_size`	5			
`pooling_type`	max			
`n_reg_layers`	7			
`batch_size`	32			

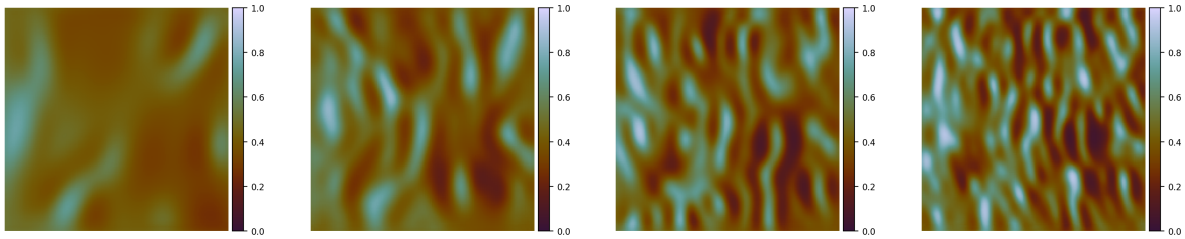
4.2.3. Final Model Performance: Reconstruction Quality

Model 1: Two-Stage Training

Model 1 achieves the best reconstruction metrics across all latent dimensions (Table 4.8) compared to the reconstructions of the 2015 dataset. Figure A.8 presents representative test set reconstructions for $z_{\text{dim}} \in \{128, 256\}$ to give a broader overview. Figures 4.8 –4.10 highlight some individual reconstructions for closer comparison. Windstreaks4.8 and fronts4.10 are well reconstructed, smaller cells are harder to reconstruct4.9. Lower dimensions ($z_{\text{dim}} \in \{32, 64\}$, produce noticeably blurrier outputs. A broader overview of different reconstructions can be found in the figures in the Appendix A.9 and A.8

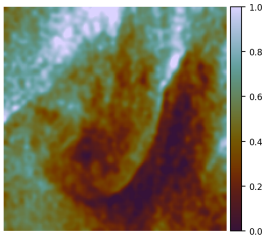


(a) Original SAR image

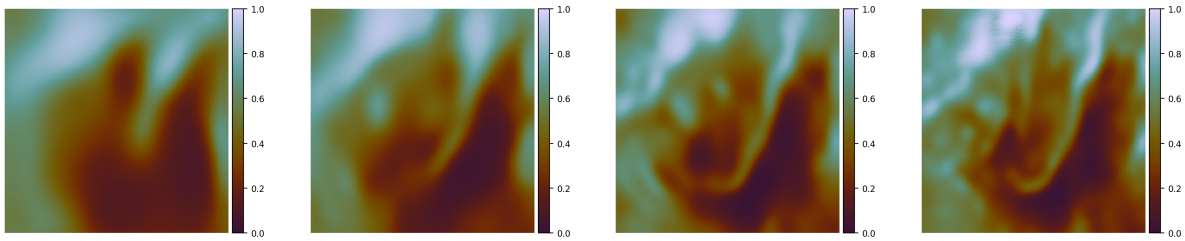


(b) $z_{\text{dim}} = 32$ **(c)** $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 128$ **(e)** $z_{\text{dim}} = 256$

Figure 4.8: Model 1 (SimpleCVAE) reconstruction quality across latent dimensions. Top: original SAR image. Bottom row: reconstructions for $z_{\text{dim}} \in \{32,\ 64,\ 128,\ 256\}$.

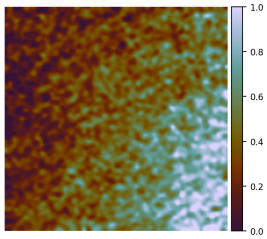


(a) Original SAR image

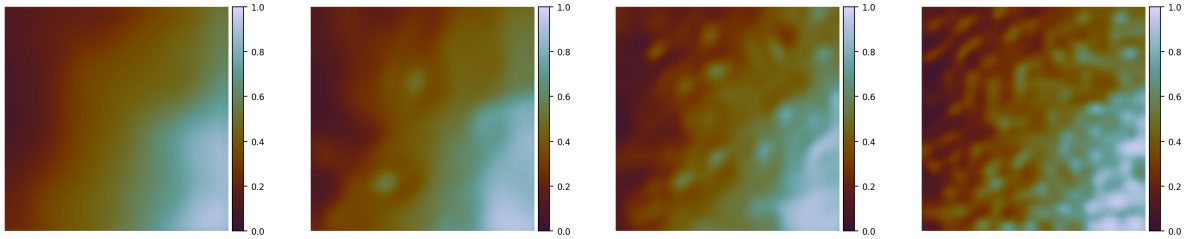


(b) $z_{\text{dim}} = 32$ **(c)** $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 128$ **(e)** $z_{\text{dim}} = 256$

Figure 4.9: Model 1 (SimpleCVAE) reconstruction quality across latent dimensions. Top: original SAR image. Bottom row: reconstructions for $z_{\text{dim}} \in \{32,\ 64,\ 128,\ 256\}$.

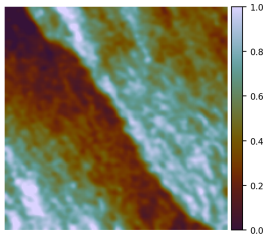


(a) Original SAR image

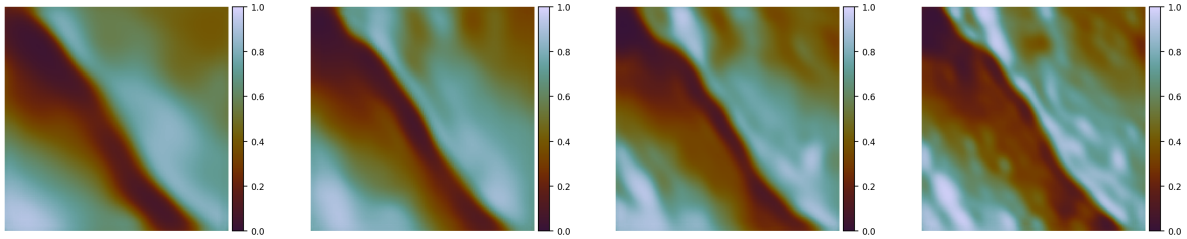


(b) $z_{\text{dim}} = 32$ **(c)** $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 128$ **(e)** $z_{\text{dim}} = 256$

Figure 4.10: Model 1 (SimpleCVAE) reconstruction quality across latent dimensions. Top: original SAR image. Bottom row: reconstructions for $z_{\text{dim}} \in \{32, 64, 128, 256\}$.



(a) Original SAR image



(b) $z_{\text{dim}} = 32$ **(c)** $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 128$ **(e)** $z_{\text{dim}} = 256$

Figure 4.11: Model 1 (SimpleCVAE) reconstruction quality across latent dimensions. Top: original SAR image. Bottom row: reconstructions for $z_{\text{dim}} \in \{32, 64, 128, 256\}$.

Table 4.8: Model 1 reconstruction metrics across latent dimensions. Metrics improve monotonically with increasing dimension. Bold values indicate best performance.

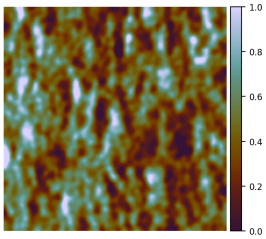
Metric	$z = 32$	$z = 64$	$z = 128$	$z = 256$
PSNR (dB)	28.31	28.47	28.67	**28.90**
SSIM	0.36	0.39	0.46	**0.54**
LFD	17.41	17.13	16.83	**16.47**

Model 2: Joint Training

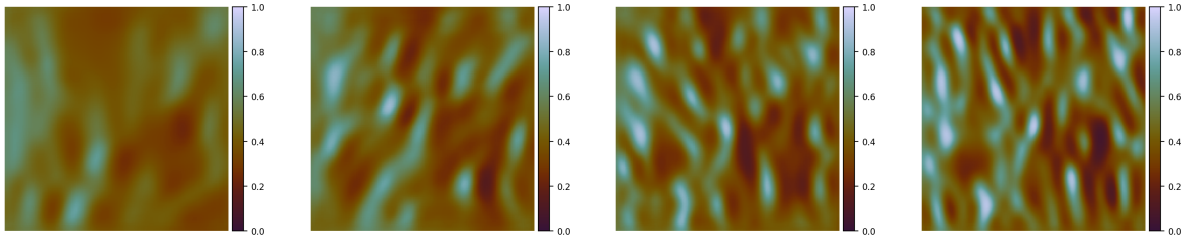
Model 2 exhibits slightly degraded reconstruction quality compared to Model 1 (Table 4.9 vs. Table 4.8), despite the additional flux prediction objective.

Figures 4.12 –4.14 highlight some individual reconstructions for closer comparison. Reconstructions remain generally faithful but lose some fine detail present in Model 1 outputs. A broader set of reconstructions across all latent dimensions is provided in Appendix A.11 and Appendix A.10.

This degradation suggests that adding the flux loss introduces competing objectives that compromise reconstruction fidelity, rather than guiding the network toward better frequency representation.

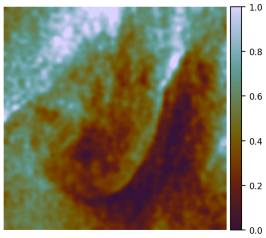


(a) Original SAR image

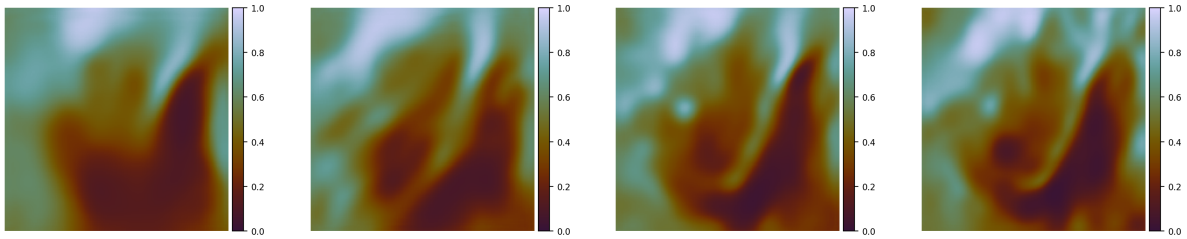


(b) $z_{\text{dim}} = 32$ **(c)** $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 128$ **(e)** $z_{\text{dim}} = 256$

Figure 4.12: Model 1 (SimpleCVAE) reconstruction quality across latent dimensions. Top: original SAR image. Bottom row: reconstructions for $z_{\text{dim}} \in \{32, 64, 128, 256\}$.

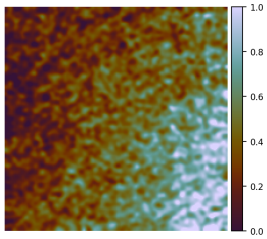


(a) Original SAR image

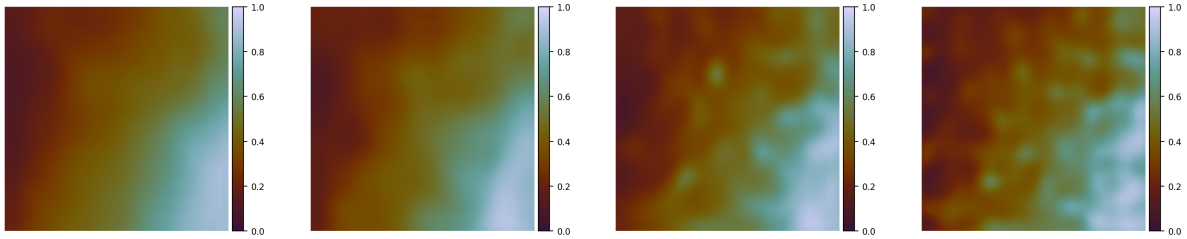


(b) $z_{\text{dim}} = 32$ **(c)** $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 128$ **(e)** $z_{\text{dim}} = 256$

Figure 4.13: Model 1 (SimpleCVAE) reconstruction quality across latent dimensions. Top: original SAR image. Bottom row: reconstructions for $z_{\text{dim}} \in \{32, 64, 128, 256\}$.

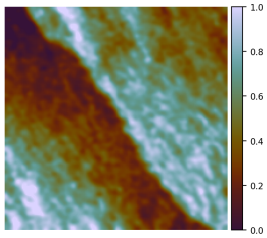


(a) Original SAR image

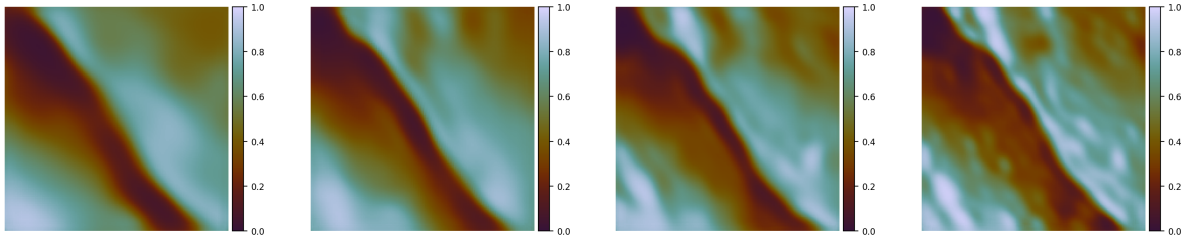


(b) $z_{\text{dim}} = 32$ **(c)** $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 128$ **(e)** $z_{\text{dim}} = 256$

Figure 4.14: Model 1 (SimpleCVAE) reconstruction quality across latent dimensions. Top: original SAR image. Bottom row: reconstructions for $z_{\text{dim}} \in \{32, 64, 128, 256\}$.



(a) Original SAR image



(b) $z_{\text{dim}} = 32$ **(c)** $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 128$ **(e)** $z_{\text{dim}} = 256$

Figure 4.15: Model 1 (SimpleCVAE) reconstruction quality across latent dimensions. Top: original SAR image. Bottom row: reconstructions for $z_{\text{dim}} \in \{32, 64, 128, 256\}$.

Table 4.9: Model 2 reconstruction metrics across latent dimensions. All metrics are slightly lower than Model 1 counterparts (Table 4.8).

Metric	$z = 32$	$z = 64$	$z = 128$	$z = 256$
PSNR (dB)	28.29	28.40	28.56	**28.70**
SSIM	0.34	0.37	0.42	**0.46**
LFD	17.47	17.26	16.99	**16.78**

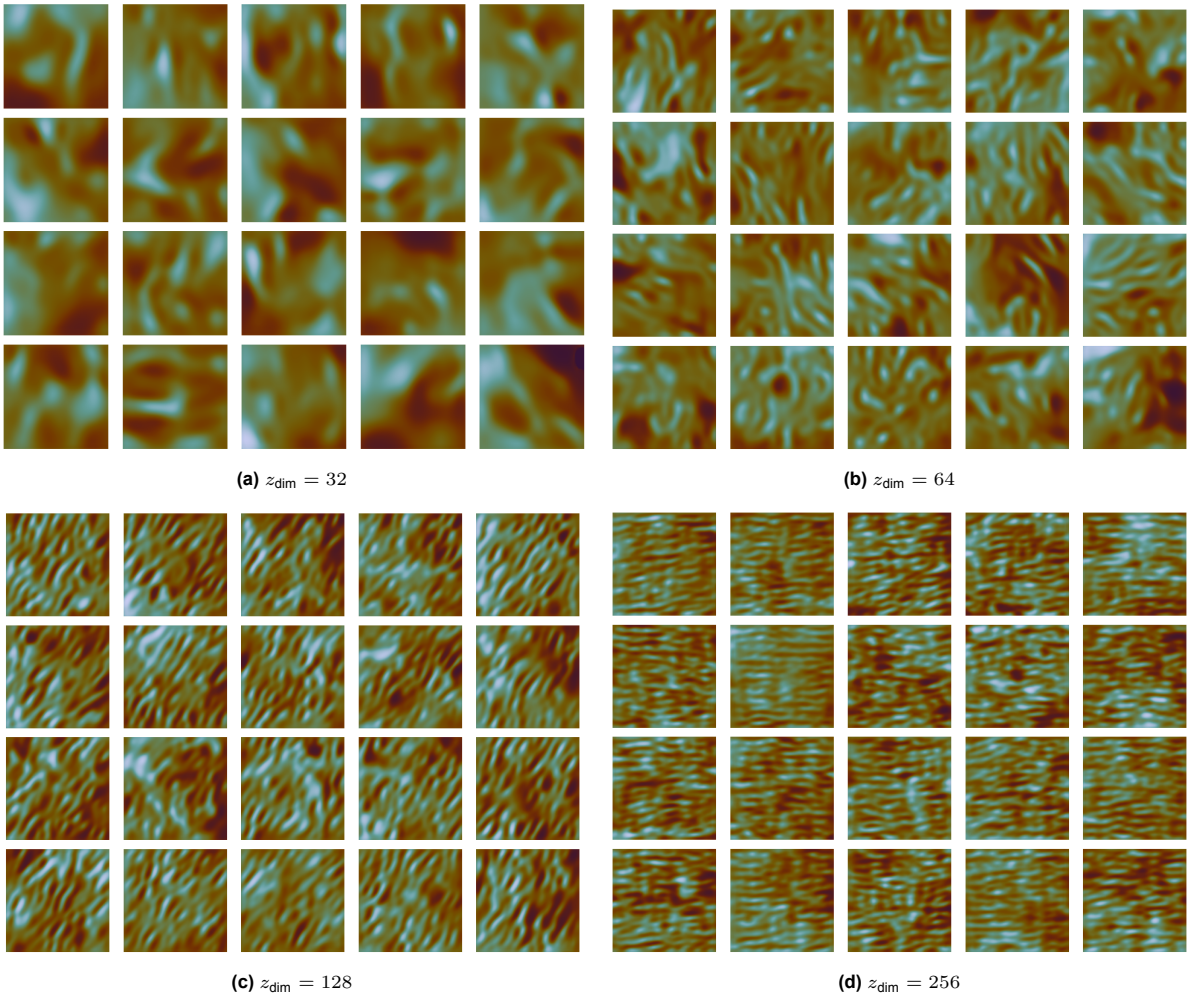
4.2.4. Latent Space Structure and Generative Capacity

Random samples drawn from the latent space distribution reveal characteristic patterns in the learned representations. Figure 4.16 presents Model 1 samples across all dimensions. Despite balanced training data containing multiple phenomenon classes, generated samples predominantly show windstreaks with uniform east-west alignment. This selective representation confirms lack of disentanglement discussed in Section 5.3.3: the latent space has not learned to independently vary phenomenon class, orientation, and spatial scale.

Model 2 samples (Figure 4.17) show qualitatively different behaviour. Dimensions 64 and 128 produce notably vague, less-structured samples compared to Model 1, suggesting degraded latent organization from competing training objectives. However, $z_{\text{dim}} = 256$ exhibits more spatial variation than Model 1, with windstreak patterns showing greater diversity in location and intensity.

Both models were independently optimised, therefore loss function hyperparameters will be different between configurations. When the optimised β values are compared (Tables 4.5 and 4.7), Model 2 converged to β values that are approximately one order of magnitude higher than those of Model 1 across most dimensions. The exception is $z_{\text{dim}} = 256$, for which Model 2 yields a marginally lower β.

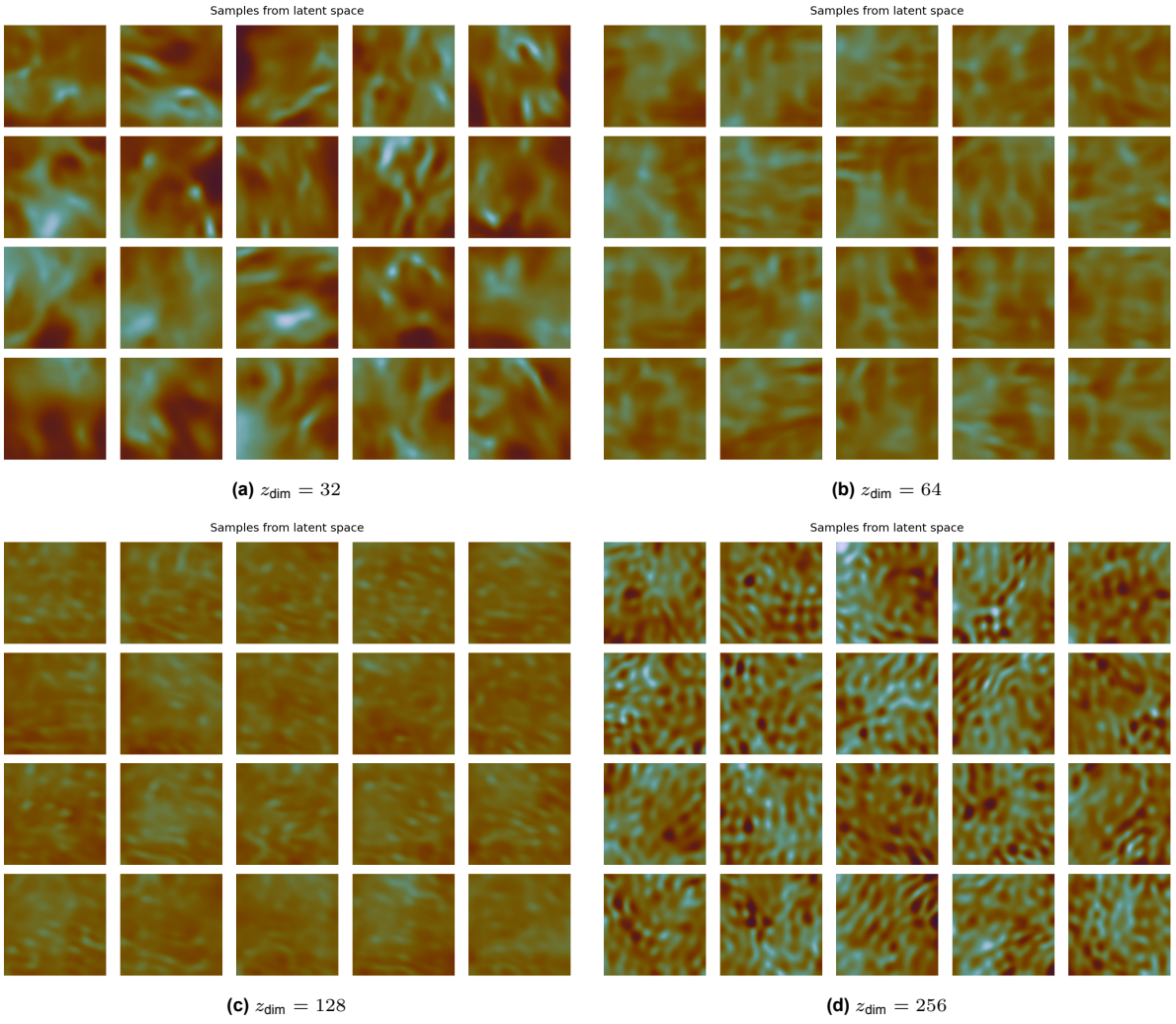
Complete latent distribution plots for both models are provided in Appendix, confirming that all dimensions achieve good marginal regularization (individual dimensions approximate standard normal) despite lack of disentanglement.



(a) $z_{\text{dim}} = 32$ **(b)** $z_{\text{dim}} = 64$

(c) $z_{\text{dim}} = 128$ **(d)** $z_{\text{dim}} = 256$

Figure 4.16: Model 1 random samples from latent prior. Despite balanced training data, generated images predominantly show windstreaks with uniform east-west alignment, indicating lack of disentanglement.



(a) $z_{\text{dim}} = 32$ **(b)** $z_{\text{dim}} = 64$

(c) $z_{\text{dim}} = 128$ **(d)** $z_{\text{dim}} = 256$

Figure 4.17: Model 2 random samples from latent prior. Dimensions 64 and 128 show degraded structure compared to Model 1. Dimension 256 exhibits more spatial variation but maintains windstreak dominance.

4.2.5. Turbulence Flux Estimation Performance

Individual Model Performance

Flux prediction scatter plots for Model 0, Model 1, and Model 2 across all latent dimensions are presented in Figures 4.18, 4.19, and 4.20. Model 1 exhibits severe underprediction, with predictions clustered near the mean regardless of true flux values. Model 2 shows improved spread and reduced bias relative to Model 1. Model 0 achieves the best performance of the three configurations.

Comparison with O'Driscoll et al.'s method

Table 4.10 compares flux prediction performance across all three model configurations and O'Driscoll et al.'s hand-crafted spectral feature approach on the validation set. O'Driscoll et al. achieves $R^2 = 0.675$, explaining 67.5% of flux variance.

Model 0 achieves the highest performance of the three models evaluated here, with $R^2 = 0.359$, though this remains substantially below O'Driscoll et al. Model 2 outperforms Model 1 across all latent dimensions ($R^2 \approx 0.22$–0.27 vs. $R^2 \approx 0.03$–0.05), with $z_{\text{dim}} = 64$ achieving the highest R^2 among VAE-based configurations ($R^2 = 0.272$). Model 1 explains only 3–5% of flux variance across all dimensions.

Table 4.10: Validation flux prediction performance comparison. O'Driscoll et al.'s methods vs Model 0, Model 1 and Model 2.

Method	R^2	**RMSE (W/m^2)**
O'Driscoll et al.	**0.675**	**16.44**
Model 0	0.359	21.0
Model 2 ($z = 32$)	0.236	25.33
Model 2 ($z = 64$)	0.272	24.59
Model 2 ($z = 128$)	0.258	24.84
Model 2 ($z = 256$)	0.216	25.52
Model 1 ($z = 32$)	0.045	28.60
Model 1 ($z = 64$)	0.051	28.53
Model 1 ($z = 128$)	0.047	28.40
Model 1 ($z = 256$)	0.048	28.39

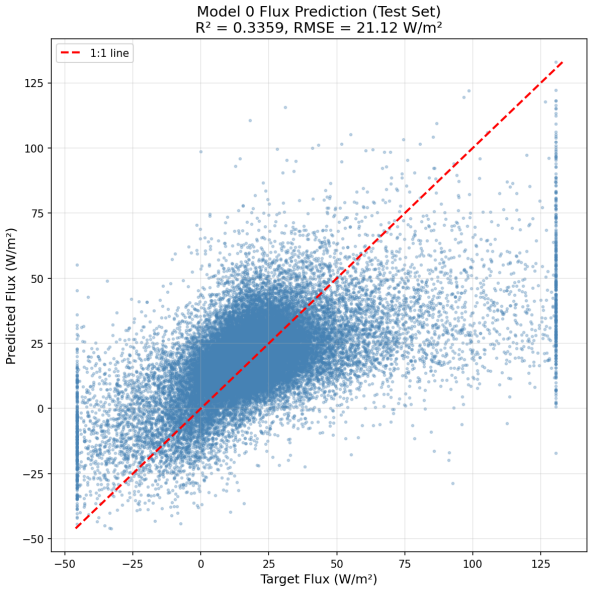
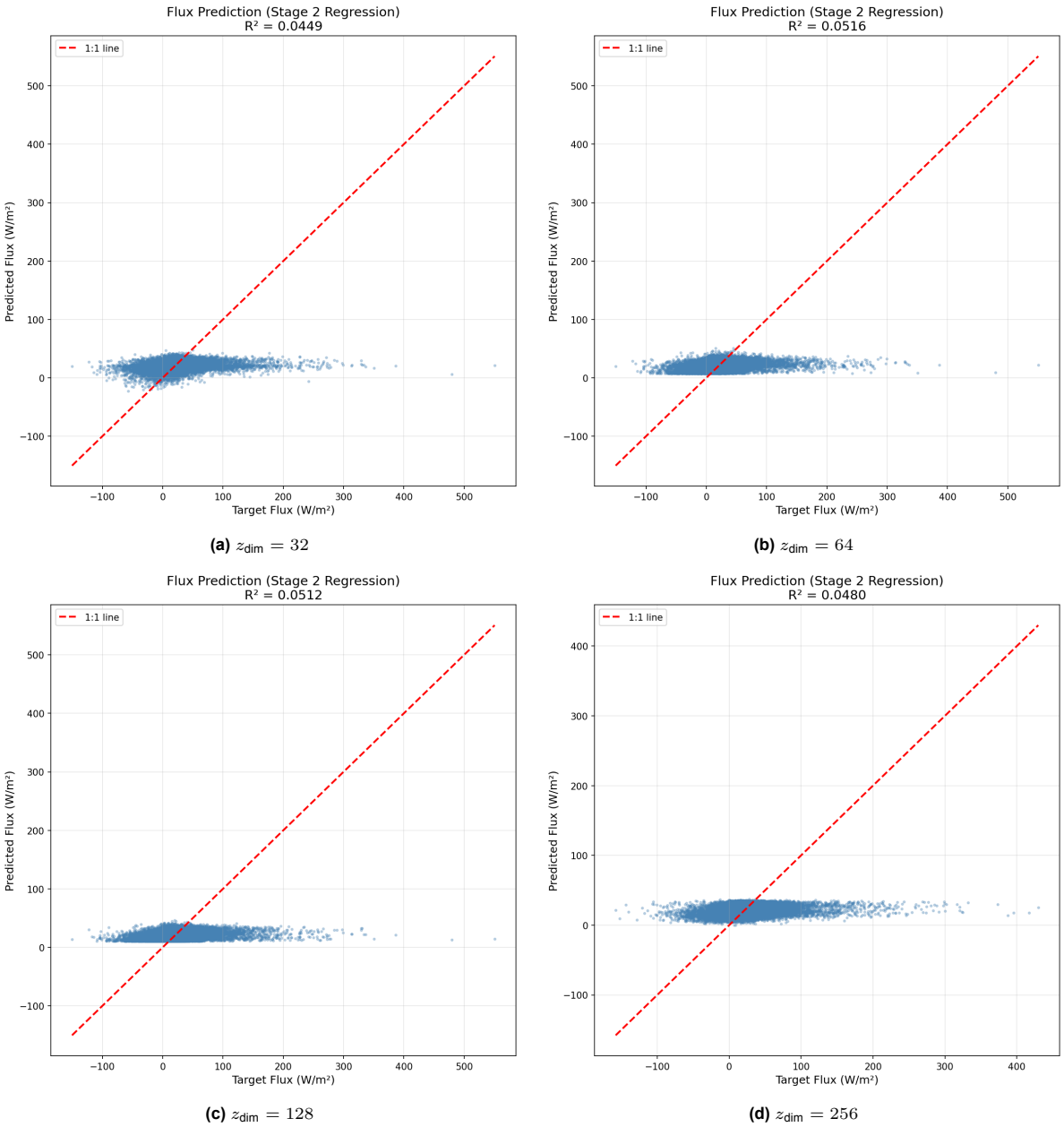


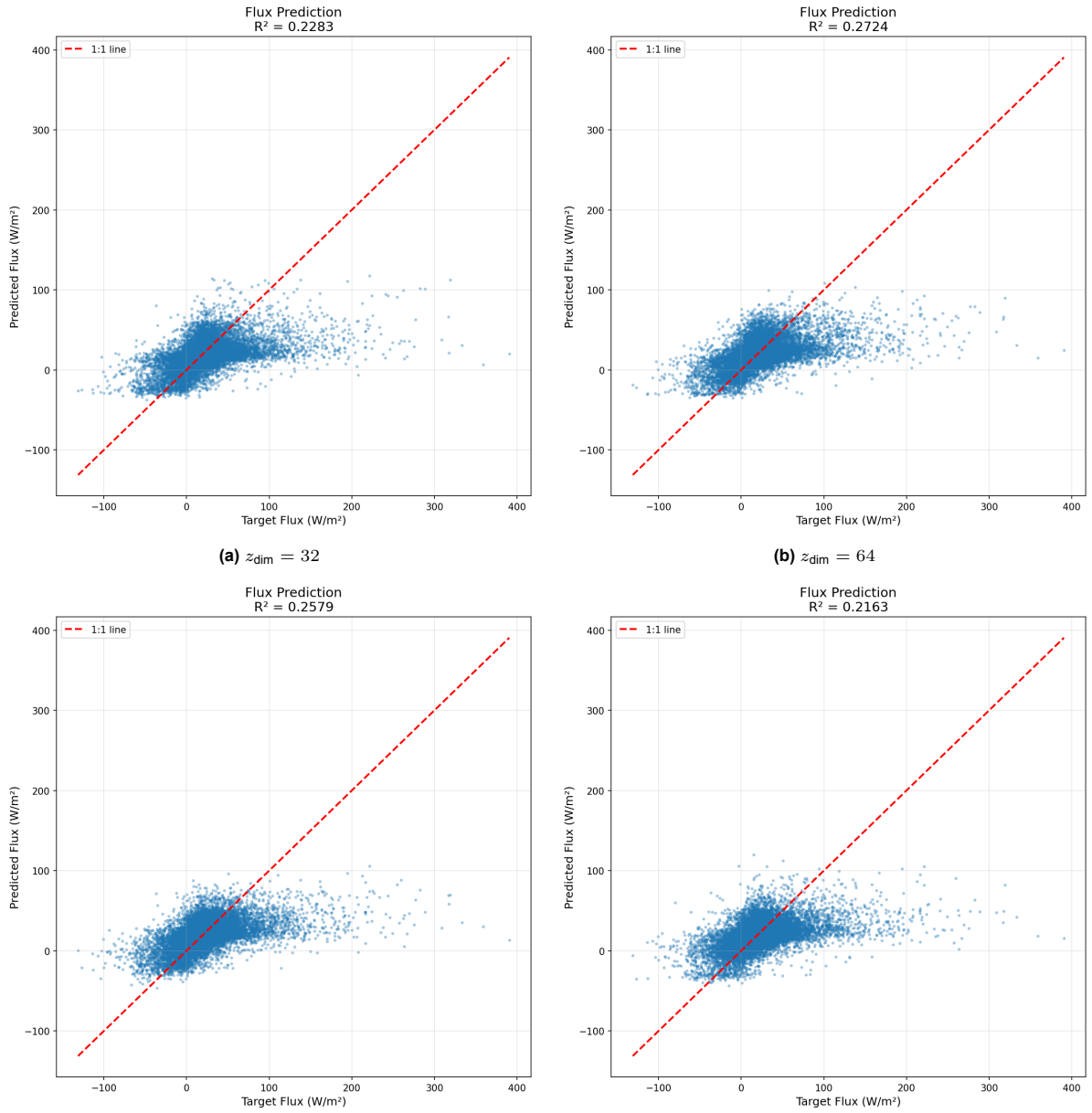
Figure 4.18: Model 0: flux predictions on test set. Target fluxes are plotted against predicted fluxes



(a) $z_{\text{dim}} = 32$ **(b)** $z_{\text{dim}} = 64$

(c) $z_{\text{dim}} = 128$ **(d)** $z_{\text{dim}} = 256$

Figure 4.19: Model 1 flux predictions on test set. Predictions cluster near the mean across all dimensions, indicating frozen encoder latents lack flux-relevant information.



(a) $z_{\text{dim}} = 32$ **(b)** $z_{\text{dim}} = 64$

(c) $z_{\text{dim}} = 128$ **(d)** $z_{\text{dim}} = 256$

Figure 4.20: Model 2 flux predictions on test set. Predictions show improved spread compared to Model 1 but remain inadequate for operational use.

5

Discussion

This chapter provides an interpretation of the findings presented in Chapter 4. The discussion is structured around three core themes: (1) the role of loss function design in capturing high-frequency detail (Section 5.1), (2) the challenge of latent space regularisation through β-annealing and hyperparameter optimisation (Section 5.2), and (3) the comparison between two-stage (Model 1) and joint training (Model 2) for turbulent flux estimation, evaluated against the direct CNN regression baseline (Model 0) (Section 5.3). This structure emphasises that VAE performance critically depends on hyperparameter tuning, particularly of loss function components. Section 5.4 discusses practical implications for operational turbulence parameter estimation, and Section 5.5 addresses limitations and future research directions.

5.1. Phase 1: Loss Function Design and High-Frequency Reconstruction

5.1.1. The Inadequacy of Spatial-Domain Metrics for Evaluating Frequency Content

The minimal variation in PSNR and SSIM across different loss configurations (Table 4.1) contrasts sharply with the visual improvements observed in the frequency domain and reconstruction images (Figures 4.1 and 4.4). This discrepancy reveals a fundamental limitation: pixel-wise metrics fail to capture the preservation of high-frequency content [18]. Similarly, while SSIM incorporates structural information through local windows, it remains dominated by low-frequency components that carry the majority of image energy. Therefore, these metrics cannot detect the improved high-frequency reconstruction visible in spectral analysis. This finding is consistent with research on perceptual loss functions[20], which shows that human perception of image quality differs from pixel-wise error metrics.

Consequently, the objective of these experiments is not to achieve the best possible PSNR or SSIM scores, but rather to demonstrate the qualitative differences in reconstruction across loss components, as well as differences that are best captured through frequency-domain analysis or visual inspection.

5.1.2. Frequency Focal Loss and the Frequency Gap

Figure 4.4 directly visualises the 'frequency gap' identified by Jiang et al. [19]: standard VAE training preferentially captures low frequencies (visible as the bright centre and main crossing bands), while high frequencies are poorly reconstructed. Adding the FFL term alone (middle panel) produces a modest improvement. However, combining FFL with MSE annealing over 40 epochs substantially reduces the frequency gap, progressively recovering detail from the centre outwards.

The corresponding spatial reconstructions (Figure 4.1) confirm this spectral improvement: the FFL + annealing combination captures finer details and complex overlapping patterns more effectively than either component alone. In quantitative terms, the configuration annealed over 40 epochs achieves the highest energy ratio (0.82) and correlation (0.91) in the high-frequency band (Table 4.2).

This differs from the findings of Jiang et al. [19], where FFL alone was good enough for face images. This discrepancy is likely due to the greater complexity of the patterns in SAR ocean imagery: faces primarily contain localised features (e.g. eyes and mouths) that benefit from simple sharpening, whereas SAR scenes exhibit multi-scale overlapping phenomena and patterns. Capturing these diverse scales simultaneously requires dynamic loss weighting that goes beyond what FFL alone provides.

5.1.3. The Role of Reconstruction term cooling down in Establishing Coarse-to-Fine Reconstruction

The cooling-down schedule for the reconstruction weight, $\gamma_{\text{reconstruction}}(t) = 1.0 - (t/T_{\text{anneal}})$ (Equation 3.6), implements a coarse-to-fine learning strategy: the initially high reconstruction weight encourages the model to establish large-scale spatial structure in early training, while its gradual reduction progressively transfers emphasis toward the Frequency Focal Loss (FFL), enabling finer detail to be captured as training matures.

Experimental comparison of annealing durations suggests that the rate of this reduction is a sensitive hyperparameter. A gradual reduction over 40 epochs yielded lower final validation loss than a short reduction over 10 epochs (Figure 4.2), with the longer schedule producing a steady, monotonic decline in validation loss throughout training. By contrast, the 10-epoch schedule reached a minimum loss relatively early and subsequently exhibited signs of overfitting, with validation loss increasing after the reconstruction weight had already been fully reduced. Visual reconstructions from the 10-epoch schedule were also found to miss fine-scale detail, and the corresponding energy ratio exceeded 1.0, indicating that premature emphasis on high-frequency components led to artefact amplification rather than improved reconstruction fidelity.

This sensitivity to the rate of reconstruction-weight reduction mirrors findings from the β-annealing literature, where the pace at which regularisation pressure is introduced has been shown to be critical for stable latent space formation. Bowman et al. [7] [1]. Yan et al. [49] explicitly note that the reconstruction loss weight γ_{recon} and KL weight β are inversely related. The gradual annealing of the reconstruction weight allows competing loss terms to come into balance progressively, rather than imposing an abrupt shift in the optimisation landscape that the model cannot adequately accommodate. Given that the slope of the cooling-down schedule appears to be an important determinant of training stability and final validation performance, this parameter was carried forward into Phase 2, where a starting value γ_{max} and a final value γ_{min} are included in the hyperparameter optimisation over the full training run.

5.1.4. FFL Parameter Sensitivity and Training Stability

The FFL focusing parameter α controls the degree of down-weighting for well-reconstructed frequencies through $w(u, v) = |F_o(u, v) - F_r(u, v)|^{\alpha}$. As α increases, poorly reconstructed frequencies receive progressively stronger emphasis. Experiments confirmed this mechanism, as can be seen in Figure 4.2, the FFL term decreased with increasing α. This demonstrates a successful identification and weighting of problematic frequencies.

However, training stability constrained viable α values. At $\alpha = 4$, training exhibited KLD spikes, while lower values maintained stable reconstruction and KLD trajectories (Jiang et al [19] kept α at 1). Given the trade-off between reconstruction metrics (Table 4.1) and stability, $\alpha = 2$ was selected for Phase 2.

5.1.5. Latent Dimensionality and Information Bottleneck Effects

As the latent dimension decreases, reconstruction quality degrades (Figure **??**), with $z_{\text{dim}} = 32$ capturing only large-scale patterns. This aligns with information bottleneck theory [38]: lower dimensions force the encoder to prioritize low-frequency, high-variance features. This why it seems to still capture the larger main patterns (such as atmospheric fronts). Even though the reconstructions lack detail and are blurry, you can see that the combination of FFL and reconstruction annealing still has an effect even for $z_{\text{dim}} = 32$ (Figure A.1).

Detailed comparisons across dimensions (Figures A.1–A.4 can be found in the appendix, showing 50 images per dimension. These reveal a consistent pattern: all configurations struggle with the smallest-scale phenomena (convective cells, fine turbulence), but the reconstruction of intermediate-scale features such as the windstreaks and the athmospheric front are well captured.

5.2. Beta-Annealing

Building on the findings of the reconstruction term cooling-down experiments, which established that the dynamic tuning of loss weight terms has a critical effect on reconstruction quality, this section investigates the role of the KL divergence weight β in regularising the latent space.

5.2.1. Fraction of Variance Explained as an Indicator of Optimal β

The clear peak in FVE at intermediate β values (Figure 4.6a) supports the use of FVE as a practical indicator of the optimal regularisation range, consistent with the approach proposed by Adhikari et al.[1].

The evolution of the latent distribution during annealing (Appendix 4.7) provides an insight into the behaviour of latent space regularisation. At low β, distributions are non-Gaussian and widely dispersed. At the FVE-optimal β, distributions approximate a standard normal, representing the balance between regularisation and expressiveness. Beyond this point, distributions progressively collapse toward the prior mean, confirming posterior collapse [7]. This establishes a direct link between the FVE metric and interpretable latent space behaviour.

5.2.2. Sensitivity of Dynamic Beta-Annealing

As discussed in Section 5.1.3, dynamically changing one loss weight term implicitly influences the others. This coupling proved equally consequential for β-annealing. In practice, the annealing schedule requires training over a sufficient range to allow the FVE peak to emerge, yet model selection based on best validation loss sometimes interrupted this process before a well-regularised latent distribution could fully establish itself, particularly for lower latent dimensions. A further practical constraint is the number of epochs required to anneal β sufficiently slowly. Adhikari et al. [1] employ 500 training epochs to achieve this, which would be computationally expensive and impractical for the dataset sizes used in this work.

For $z_{\text{dim}} = 256$, convergence toward a standard normal distribution was achieved (Figure A.7a), and ancestral samples (Figure A.6a) display recognisable windstreak and cellular patterns. However, sampled windstreaks appear preferentially aligned in the range direction, suggesting that the latent space has not generalised sufficiently. For lower dimensions, regularisation was less complete (Figure A.7d), with $z_{\text{dim}} = 32$ producing highly vague ancestral samples (Figure A.6d). This is reflected in the FVE curves (Figure 4.6), where lower-dimensional latent spaces exhibit a less pronounced peak, suggesting that the optimal regularisation range is both narrower and more difficult to identify reliably. In all cases, a clear upper bound on β is apparent: beyond the FVE peak, FVE declines sharply, consistent with the onset of posterior collapse [7].

Given the complexity of the interactions between all three dynamically changing loss terms, and the absence of clear improvement in latent space sample quality and variety under the annealing scheme, a pragmatic decision was made to fix β as a static hyperparameter. The FVE score, having demonstrated its value as a reliable indicator of the optimal regularisation range, is used within the Optuna optimisation framework to identify a well-calibrated fixed β value, reducing sensitivity to schedule design while retaining the diagnostic insight provided by the metric.

5.2.3. Hyperparameter Optimization: Architecture vs. Loss Function Dynamics

Divergent Importance Patterns Between Model 1 and Model 2

Hyperparameter importance rankings can be found in the Appendix (Tables A.6 and A.7) and reveal the following:

Model 1 (VAE-only training): Optimizer, batch size, learning rate, and the loss function weights γ, λ β show to have more importance in the optimiastion scheme. Architectural hyperparameters (number of blocks, pooling type, kernel size) show low importance.

Model 2 (joint training): Regression dropout, reconstruction weight γ_{recon}, flux-specific learning rate α_{flux}, and β rank highest. Architectural hyperparameters again show low importance.

The consistently low importance of architectural parameters contrasts with domains like image classification where architecture engineering significantly impacts performance [36]. This outcome aligns with limitations of deep CNNs on SAR imagery noted by Xu et al. [47]. The convergence of best trials

across all four z_{dim} values toward 4 residual CNN blocks with max pooling aligns with findings from other VAE-SAR studies [47, 34].

Consequently, the next optimisation stage could focus more on the loss function and training hyperparameters. This allowed us to simplify the optimisation process by fixing the architectural hyperparameters.

Multi-Objective Optimization in Joint Training

The elevated importance of loss function weigths in both Model 1 and Model 2 highlights the difficulty of multi-objective optimisation. During joint training, gradients from flux prediction directly impact latent representations, making the relative weighting of objectives critical for achieving both accurate reconstruction and flux estimation simultaneously. This finding reinforces a central observation of this work: VAE performance critically depends on the tuning of loss function hyperparameters, with architectural choices playing a secondary role for SAR ocean imagery.

A broader limitation of the joint optimisation approach is that the loss function weights ($\gamma_{\text{MSE, max}}$, $\gamma_{\text{MSE, min}}$, α_{FFL}, α_{flux}, and β) were optimised independently for each model and latent dimension. While this approach produced improvements in reconstruction quality, as reflected in the metrics reported in Table 4.8, it makes cross-configuration comparison difficult: differences in reconstructed images between models or latent dimensions cannot be attributed directly to a single cause, as the underlying hyperparameter configurations differ. This is perhaps the fundamental challenge of joint objective optimisation, where latent space regularisation, broadband reconstruction fidelity, frequency bias correction, and flux estimation are simultaneously optimised. These objectives have shown themselves to be competing, and combining them into a single optimisation objective function does not necessarily yield the best outcome for any individual objective. This is particularly evident in the quality of ancestral samples, which lack both fidelity and variety, suggesting that latent space regularisation was not achieved to a satisfactory degree under the joint optimisation scheme.

This raises the question of whether the selected β values represent the best possible regularisation or merely the best possible balance given the other objectives that are active at the same time. A more principled approach would be to optimise β categorically. This could involve testing a discrete range of fixed β values, as was done for $z_(dim)$, and optimising the remaining hyperparameters independently for each candidate β before evaluating latent space quality through ancestral sampling. However, given that this work already evaluates two models across four latent dimensions, extending this to a systematic sweep over β values represents a practical limitation of time and computational resources, and is therefore identified as a direction for future work.

5.3. Model 1 vs. Model 2: Two-Stage vs. Joint Training

Having established the importance of loss function design and hyperparameter optimisation, we now compare the two training paradigms across reconstruction quality, latent space structure, generative capacity, and flux estimation performance.

5.3.1. Reconstruction Quality: The Trade-off Between Specialization and Generalization

Model 1 achieves superior reconstruction metrics (Table 4.8) across all z_{dim} values, and reconstructions of the 2024 dataset represent a clear improvement over images from the 2015 dataset in terms of metric scores. Visual inspection (Figure A.8) confirms that higher dimensions ($z_{\text{dim}} \in \{128, 256\}$) capture diverse geophysical classes effectively. At $z_{\text{dim}} = 256$, reconstructions exhibit sharp features including windstreaks and atmospheric fronts. Also $z_{\text{dim}} = 128$ maintains a good differentiation between windstreaks and atmospheric fronts, with the primary difference being feature sharpness rather than class representation.

Lower dimensions ($z_{\text{dim}} \in \{32, 64\}$) successfully reconstruct atmospheric fronts and broad windstreak patterns, though $z_{\text{dim}} = 32$ produces noticeably blurry outputs. Across all dimensions, both models struggle with small-scale phenomena such as convective cells and fine turbulence structures. This is a limitation with direct implications for flux estimation, which requires information across all spatial scales.

Model 2 shows degraded reconstruction quality relative to Model 1, despite the additional flux predic-

tion objective. This degradation suggests that the flux loss term does not guide the network toward improved frequency representation as initially hypothesised. Instead, it introduces competing gradient signals that compromise reconstruction fidelity. For tasks requiring high-quality reconstruction, the results indicate that sequential training (Model 1) outperforms joint training (Model 2).

5.3.2. Flux Prediction Performance

Table 4.10 summarises flux prediction performance across all three model configurations compared to the method of O'Driscoll et al. [29]. The results follow the expected ordering: Model 0 performs best ($R^2 = 0.359$), followed by Model 2 ($R^2 = 0.216$–0.272), and Model 1 performing worst ($R^2 = 0.034$–0.048). This ordering is informative in itself. Model 0, a direct CNN with a regression head trained end-to-end on flux prediction, represents the upper bound of what a convolutional architecture can achieve on this task without the constraints imposed by the VAE objective. Model 2, which incorporates flux prediction into joint VAE training, outperforms Model 1 by a substantial margin, confirming that latent representations can be partially guided toward flux-relevant information when the flux loss is included during training. Model 1, where a regression head is applied to frozen latent features learned without any flux supervision, explains only 3–5$\%$ of flux variance, indicating that unsupervised VAE representations do not spontaneously encode the information necessary for flux estimation.

Despite this expected ordering, all three models substantially underperform O'Driscoll's method, who achieves $R^2 = 0.675$ using dedicated machine learning regressors trained on manually picked features, specifically designed for flux estimation [50]. Even Model 0, which is trained to just focus on fluxes, falls nearly 32 percentage points short.

The comparison with O'Driscoll's method is instructive. Their approach explicitly encodes domain knowledge through the selection of physically meaningful features, providing pre-knowledge to the regression models. Whereas all three models presented are unsupervised learning models and learn representations from raw SAR imagery without such inductive bias. The performance gap therefore does not necessarily indicate a failure of the deep learning approach, but rather reflects the advantage of prior physical knowledge in a setting where labelled training data are scarce. This interpretation is consistent with the observation that Model 0, still substantially underperforms [29], suggesting that the limiting factor is not the training paradigm but the difficulty of extracting flux-relevant information from raw SAR backscatter without explicit spectral feature engineering.

For the VAE-based models, an additional constraint is the tension between reconstruction quality and flux estimation within the joint objective. As discussed in Section 5.3.1, the flux loss term in Model 2 introduces competing gradient signals that compromise reconstruction fidelity, and the resulting flux prediction performance remains below that of Model 0. This suggests that combining both objectives into a single training scheme imposes a cost on each: Model 0 outperforms Model 2 on flux prediction precisely because its objective is focused exclusively on that task, while Model 1 achieves better reconstruction than Model 2 for the same reason in the opposite direction. Rather than complementing one another, the two objectives appear to compete, and the joint optimisation scheme does not resolve this tension satisfactorily for either. Achieving a better balance would likely require an architectural reformulation that decouples the two objectives more explicitly. Or using pre knowledge as away of conditioning the VAE, such as a conditional VAE [35] or a semi-supervised approach in which flux supervision is applied selectively to labelled samples [14], rather than being imposed as a competing term within a single joint loss.

5.3.3. Latent Space Structure

Latent dimension distributions across all z_{dim} values show good adherence to the priors standard normal distribution, maintaining spread and avoiding posterior collapse. This pattern is described by Foster et al.'s observations [13]. However, unlike Foster's CelebA results where individual latent dimensions corresponded to interpretable factors (hair color, smile), our latent space does not exhibit clear disentanglement.

This divergence likely reflects the fundamental differences between face images and Synthetic Aperture Radar (SAR) ocean imagery. Faces contain relatively independent generative factors (for example, one can vary smile intensity without affecting eye colour), whereas ocean phenomena exhibit strong physical correlations between wind, turbulence, ocean waves, and convective cells. Therefore, infor-

mation about these phenomena can be spread over many different latent variables. A useful analogy here can be found in genetics, where many genes collectively contribute to single traits.

Despite the fact that each dimension is individually Gaussian, the joint latent distribution requires multiple dimensions to represent a single physical factor. This pattern emerges in random samples from the latent space, where windstreaks are predominantly generated despite a balanced class representation in the training data. The uniformity of the windstreaks in range direction(figure 4.16) and the absence of atmospheric fronts in the samples further confirms entanglement: the latent space has not learnt to vary phenomenon class, orientation and scale independently.

5.3.4. Optimal Configuration and Generalization Across Datasets

The current VAE state cannot yet support operational turbulence flux estimation, this is a limitation consistent across all four latent dimensions. For reconstruction-focused applications, Model 1 with $z_{\text{dim}} \in \{128, 256\}$ achieves to capture diverse phenomena (atmospheric fronts, windstreaks, broader cell patterns).

Importantly, the model generalizes well across datasets (2015 vs. 2021–2024), with similar reconstruction quality, latent distributions, and sample characteristics despite different scene diversity. This robustness suggests that the optimized hyperparameters can capture the fundamental properties of SAR ocean imagery and is capable to provide good quality reconstructions.

5.3.5. Performance Variation Across Phenomenon Classes

Despite balanced training data, the dominance of windstreaks in latent space samples reveals class-dependent learning difficulty. Wang et al.'s classification [44] provides insight: wind streaks exhibit periodic linear features (with wavelengths ranging from 0.8 to 5 km) that dominate entire images, whereas ocean waves and convective cells consist of smaller, more localised features. Atmospheric fronts, which are characterised by large-scale intensity gradients and occluded boundaries, theoretically should appear in the samples given their spatial extent, but yet they do not.

This selective representation likely reflects the CNN's frequency bias and the VAE's tendency towards 'average' reconstructions. Windstreak periodicity aligns with the mid-frequency components that CNNs naturally emphasise, whereas the fine structure of cells falls into the high frequencies that the frequency gap missed to capture. Atmospheric front gradients also require precise phase information, which MSE-based reconstruction smooths away. While the reconstruction quality at higher dimensions does capture all classes (including fronts), this indicates that the encoder and decoder structure successfully represents them. However, the sampling behaviour from the latent space seems to mainly generate windstreak patterns.

5.4. Practical Implications and Limitations

5.4.1. Current Limitations for Operational Deployment

The most significant limitation is incomplete capture of small-scale phenomena. This maintains a challenge for CNNs given frequency bias [32] and VAEs given their tendency toward blurry reconstructions [31]. This limitation precludes direct operational use for turbulence flux estimation, which requires information across all scales including the finest resolvable features.

The lack of disentanglement further limits the interpretability and controllability of the model. Without clear latent factors corresponding to physical variables such as wind direction, convective cells, fronts and turbulence signatures, the model cannot support downstream tasks such as ocean parameter estimation.

5.4.2. Insights for Future VAE Development on SAR Imagery

Despite its limitations, this work offers valuable methodological insights.

Loss function design and optimisation improves performance. The systematic exploration of FFL integration, Reconstruction term annealing, and β optimization demonstrates that training hyperparameters and in particularly loss component weights and dynamic training schedules, exert far greater influence than architectural choices for SAR ocean imagery. Optimization frameworks like Optuna prove essen-

tial for navigating this complex, sensitive hyperparameter landscape.

Optimization metrics must align with task goals. Standard metrics (PSNR, SSIM) fail to capture high-frequency content preservation, while FVE successfully identifies optimal regularization and R^2_{FFL} (adapted from the FFL term) managed to decrease the frequency gap with reconstructions. They can be successfully used as a objective by optimization frameworks such as Optuna.

Dynamic loss weighting requires careful implementation. While dynamic tuning of loss weights such as β-annealing and reconstruction term annealing demonstrably affects both reconstruction quality and latent space regularisation, it also introduces complexity through the implicit coupling between terms. Adjusting one weight modulates the effective influence of the others, as discussed in Section 5.1.3, making it difficult to isolate the contribution of any individual scheduling decision. Practical challenges such as model selection timing and dataset sensitivity can further undermine the effectiveness of annealing schemes. Fixed hyperparameter optimisation via Optuna proved more robust across both datasets in this regard. Future work should investigate the interactions between β and reconstruction term annealing more systematically, isolating the effect of each schedule on reconstruction quality and latent space regularisation independently before combining them, in order to better understand the dynamics of joint loss scheduling in VAEs trained on SAR imagery.

Latent space regularisation benefits from structured β selection. A more principled approach to comparing latent space regularisation across models and latent dimensions would be to define a discrete set of fixed β values, optimise all remaining hyperparameters independently for each combination of β and z_{dim}, and then evaluate the resulting latent space quality through ancestral sampling and FVE scores. This structured grid over β and z_{dim}, would allow a more direct and interpretable comparison of latent space regularisation outcomes across configurations, and provide a clearer basis for selecting an optimal β without the confounding influence of jointly optimised dynamic schedules.

5.5. Future Directions

5.5.1. Conditional VAEs and Semi-Supervised Learning

The lack of disentanglement and class-balanced generation of latent space samples, suggests that VAEs may be insufficient for SAR ocean imagery. Conditional VAEs (cVAEs) [35], which pass class labels or important information (incidence angle, phenomena class) to both encoder and decoder, could address this limitation. The result is that the latent space does not need to encode the class information, this might prove better for downstream tasks.

This semi-supervised approach could further incorporate classification losses during training. Glaser et al. [14] demonstrate promising results using semantic embeddings for downstream tasks.

5.5.2. Hierarchical Priors and Diffusion Models

As Prince [31] notes, high-quality VAE sampling requires latent priors that are more sophisticated than simple Gaussians. Hierarchical priors are one option. These connect naturally to diffusion models, which can be viewed as VAEs with hierarchical priors and have demonstrated superior sample quality for complex natural images. Given the multi-scale and overlapping phenomena in SAR ocean images, hierarchical representations may be particularly well suited to capturing different spatial scales at different levels.

5.5.3. Hybrid Approaches for Flux Estimation

Rather than expecting VAE latents to directly support flux estimation, hybrid architectures might prove more effective:

Multi-task architectures with auxiliary branches: Extend the VAE encoder with separate pathways optimized for different tasks (reconstruction, flux estimation, classification), allowing task-specific feature extraction while maintaining shared low-level representations.

5.5.4. Alternative Generative Models

Generative Adversarial Networks (GANs) avoid the blurry reconstruction problem through adversarial training but lack the structured latent space and training stability of VAEs. Recent hybrid approaches

(VAE-GAN [23]) might offer advantages, combining VAE interpretability with GAN reconstruction quality.

6

Conclusion

This thesis set out to explore whether VAEs can perform dimensionality reduction by learning compressed latent representations directly from SAR ocean imagery. The answer is nuanced: VAEs can successfully compress SAR ocean imagery to substantially reduced latent dimensions (128–256) while reconstructing diverse large- and intermediate-scale geophysical phenomena, but preserving the full multi-scale information required for turbulence parameter estimation remains beyond the capabilities of current models.

The main finding is that careful tuning of loss function components and training strategies has a significant impact on both reconstruction quality. A systematic investigation of the integration of frequency focal loss, MSE annealing schedules and β-optimisation reveals the complex dynamics between reconstruction, regularisation and task-specific objectives, which are highly sensitive to hyperparameter configuration. Computational optimisation frameworks such as Optuna are essential for navigating this landscape when paired with objective functions that are aligned with the requirements of the task in question.

Although current models cannot yet support the operational estimation of turbulence fluxes, they demonstrate that unsupervised learning can extract meaningful compressed representations of complex, multi-scale physical phenomena. Moving forward, we need conditional architectures that leverage domain knowledge and hierarchical priors that can naturally accommodate multi-scale structure. We also need hybrid approaches that combine VAE representation learning with physics-informed components.

Successful model development in geophysical remote sensing requires more than architectural choices alone. It demands joint optimisation of loss function design, training procedures and effective evaluation metrics. This is reinforced by a central finding of this work: VAE performance is critically sensitive to loss function hyperparameters, with architectural choices playing a secondary role for SAR ocean imagery. Notably, the dynamic annealing of loss function terms was found to have a striking effect on reconstruction quality, demonstrating that not only the choice of loss components but also how they are scheduled during training is critical. Although generic deep learning approaches are a valuable starting point, they must be systematically fine tuned to capture the specific spatial scales, spectral characteristics, and physical relationships inherent in SAR ocean data in order to encode ocean signatures into latent space.

References

[1] Subinoy Adhikari and Jagannath Mondal. “Elucidating Protein Dynamics through the Optimal Annealing of Variational Autoencoders”. In: *Journal of Chemical Theory and Computation* (2025).

[2] Takuya Akiba et al. “Optuna: A next-generation hyperparameter optimization framework”. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2623–2631.

[3] Reza Mohammadi Asiyabi et al. “Synthetic aperture radar (SAR) for ocean: A review”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2023).

[4] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[5] Emanuele Boattini et al. “Autonomously revealing hidden local structures in supercooled liquids”. In: *Nature communications* 11.1 (2020), p. 5479.

[6] Maurice Borgeaud et al. “Status of the ESA earth explorer missions and the new ESA earth observation science strategy”. In: *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2015, pp. 4189–4192.

[7] Samuel Bowman et al. “Generating sentences from a continuous space”. In: *Proceedings of the 20th SIGNLL conference on computational natural language learning*. 2016, pp. 10–21.

[8] Claus Brüning, Werner Alpers, and Klaus Hasselmann. “Monte-Carlo simulation studies of the nonlinear imaging of a two dimensional surface wave field by a synthetic aperture radar”. In: *Remote Sensing* 11.10 (1990), pp. 1695–1727.

[9] Keith R Carver, Charles Elachi, and Fawwaz T Ulaby. “Microwave remote sensing from space”. In: *Proceedings of the IEEE* 73.6 (2005), pp. 970–996.

[10] Fabrice Collard, Fabrice Ardhuin, and Bertrand Chapron. “Extraction of coastal ocean wave fields from SAR images”. In: *IEEE Journal of Oceanic Engineering* 30.3 (2005), pp. 526–533.

[11] Meghan F Cronin et al. “Air-sea fluxes with a focus on heat and momentum”. In: *Frontiers in Marine Science* 6 (2019), p. 430.

[12] Diane L Evans et al. “Seasat—A 25-year legacy of success”. In: *Remote Sensing of Environment* 94.3 (2005), pp. 384–404.

[13] David Foster. *Generative deep learning*. ” O’Reilly Media, Inc.”, 2022.

[14] Yannik Glaser et al. “WV-Net: A foundation model for SAR WV-mode satellite imagery trained using contrastive self-supervised learning on 10 million images”. In: *arXiv preprint arXiv:2406.18765* (2024).

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014).

[16] Ramon F Hanssen. *Radar interferometry: data interpretation and error analysis*. Springer, 2001.

[17] Hans Hersbach. “Comparison of C-band scatterometer CMOD5. N equivalent neutral winds with ECMWF”. In: *Journal of Atmospheric and Oceanic Technology* 27.4 (2010), pp. 721–736.

[18] Alain Hore and Djemel Ziou. “Image quality metrics: PSNR vs. SSIM”. In: *2010 20th international conference on pattern recognition*. IEEE. 2010, pp. 2366–2369.

[19] Liming Jiang et al. “Focal frequency loss for image reconstruction and synthesis”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 13919–13929.

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. “Perceptual losses for real-time style transfer and super-resolution”. In: *European conference on computer vision*. Springer. 2016, pp. 694–711.

[21] Vincent Kerbaol, Bertrand Chapron, and Paris W Vachon. “Analysis of ERS-1/2 synthetic aperture radar wave mode imagettes”. In: *Journal of Geophysical Research: Oceans* 103.C4 (1998), pp. 7833–7846.

[22] Diederik P Kingma. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[23] Anders Boesen Lindbo Larsen et al. “Autoencoding beyond pixels using a learned similarity metric”. In: *International conference on machine learning*. PMLR. 2016, pp. 1558–1566.

[24] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[25] Paco López-Dekker et al. “Harmony: An Earth explorer 10 mission candidate to observe land, ice, and ocean surface dynamics”. In: *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2019, pp. 8381–8384.

[26] Paco López-Dekker et al. “The Harmony mission: End of phase-0 science overview”. In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE. 2021, pp. 7752–7755.

[27] Alberto Moreira et al. “A tutorial on synthetic aperture radar”. In: *IEEE Geoscience and remote sensing magazine* 1.1 (2013), pp. 6–43.

[28] ZN Musa, I Popescu, and A Mynett. “A review of applications of satellite SAR, optical, altimetry and DEM data for surface water modelling, mapping and parameter estimation”. In: *Hydrology and Earth System Sciences* 19.9 (2015), pp. 3755–3769.

[29] Owen O’driscoll et al. “Obukhov length estimation from spaceborne radars”. In: *Geophysical Research Letters* 50.15 (2023), e2023GL104228.

[30] Pierre Potin et al. “Copernicus Sentinel-1 constellation mission operations status”. In: *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*. IEEE. 2019, pp. 5385–5388.

[31] Simon JD Prince. *Understanding deep learning*. MIT press, 2023.

[32] Nasim Rahaman et al. “On the spectral bias of neural networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 5301–5310.

[33] Igor G Rizaev et al. “Modeling and SAR imaging of the sea surface: A review of the state-of-the-art with simulations”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 187 (2022), pp. 120–140.

[34] Saumya Sinha et al. “Variational autoencoder anomaly-detection of avalanche deposits in satellite SAR imagery”. In: *Proceedings of the 10th International Conference on Climate Informatics*. 2020, pp. 113–119.

[35] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. “Learning structured output representation using deep conditional generative models”. In: *Advances in neural information processing systems* 28 (2015).

[36] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[37] Jean-Noël Thépaut et al. “The Copernicus programme and its climate change service”. In: *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2018, pp. 1591–1593.

[38] Naftali Tishby and Noga Zaslavsky. “Deep learning and the information bottleneck principle”. In: *2015 ieee information theory workshop (itw)*. Ieee. 2015, pp. 1–5.

[39] Kiyo Tomiyasu. “Tutorial review of synthetic-aperture radar (SAR) with applications to imaging of the ocean surface”. In: *Proceedings of the IEEE* 66.5 (1978), pp. 563–583.

[40] Ramon Torres et al. “GMES Sentinel-1 mission”. In: *Remote sensing of environment* 120 (2012), pp. 9–24.

[41] Ramon Torres et al. “Sentinel 1 evolution: Sentinel-1C and-1D models”. In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. 2017, pp. 5549–5550.

[42] Ramon Torres et al. “Sentinel-1 next generation: Enhanced C-band data continuity”. In: *EUSAR 2024; 15th European Conference on Synthetic Aperture Radar*. VDE. 2024, pp. 1–4.

[43] John F Vesecky and Robert H Stewart. “The observation of ocean surface phenomena using imagery from the SEASAT synthetic aperture radar: An assessment”. In: *Journal of Geophysical Research: Oceans* 87.C5 (1982), pp. 3397–3430.

[44] Chen Wang et al. “A labelled ocean SAR imagery dataset of ten geophysical phenomena from Sentinel-1 wave mode”. In: *Geoscience Data Journal* 6.2 (2019), pp. 105–115.

[45] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[46] Qihan Xu et al. “Synthetic aperture radar image compression based on a variational autoencoder”. In: *IEEE Geoscience and Remote Sensing Letters* 19 (2021), pp. 1–5.

[47] Yanbing Xu et al. “SAR target recognition based on variational autoencoder”. In: *2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*. Vol. 1. IEEE. 2019, pp. 1–4.

[48] Zhi-Qin John Xu et al. “Frequency principle: Fourier analysis sheds light on deep neural networks”. In: *arXiv preprint arXiv:1901.06523* (2019).

[49] Chaochao Yan et al. “Re-balancing variational autoencoder loss for molecule sequence generation”. In: *Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics*. 2020, pp. 1–7.

[50] George S Young, Todd D Sikora, and Nathaniel S Winstead. “Inferring marine atmospheric boundary layer properties from spectral characteristics of satellite-borne SAR imagery”. In: *Monthly weather review* 128.5 (2000), pp. 1506–1520.

A

Appendix

A.1. tables and figures

A.1.1. Phase 1: Dataset 2015

Reconstructions for different z dimensions

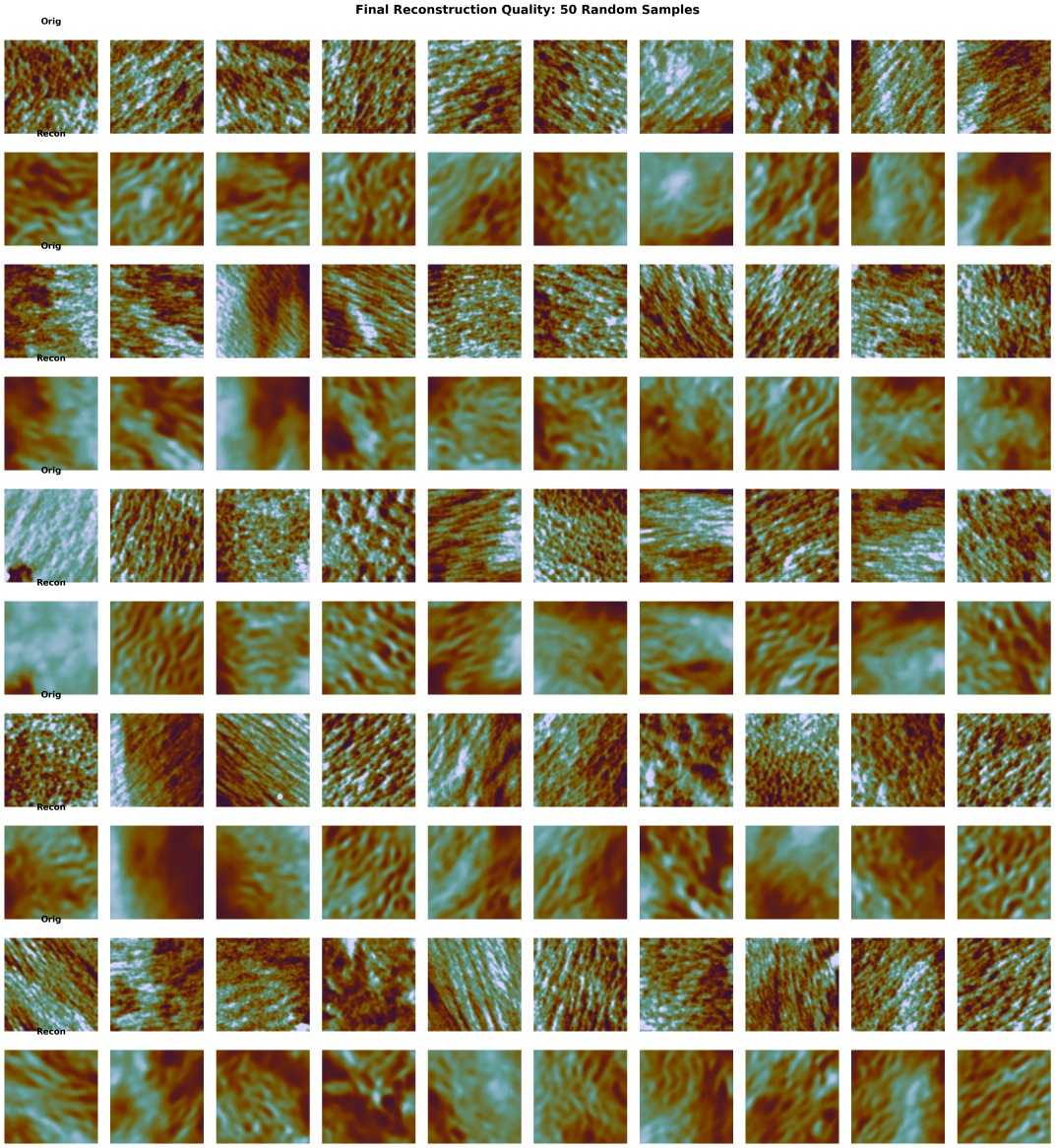


Figure A.1: Reconstructions of test set Dataset 2015, squeezed to 32 dimensions

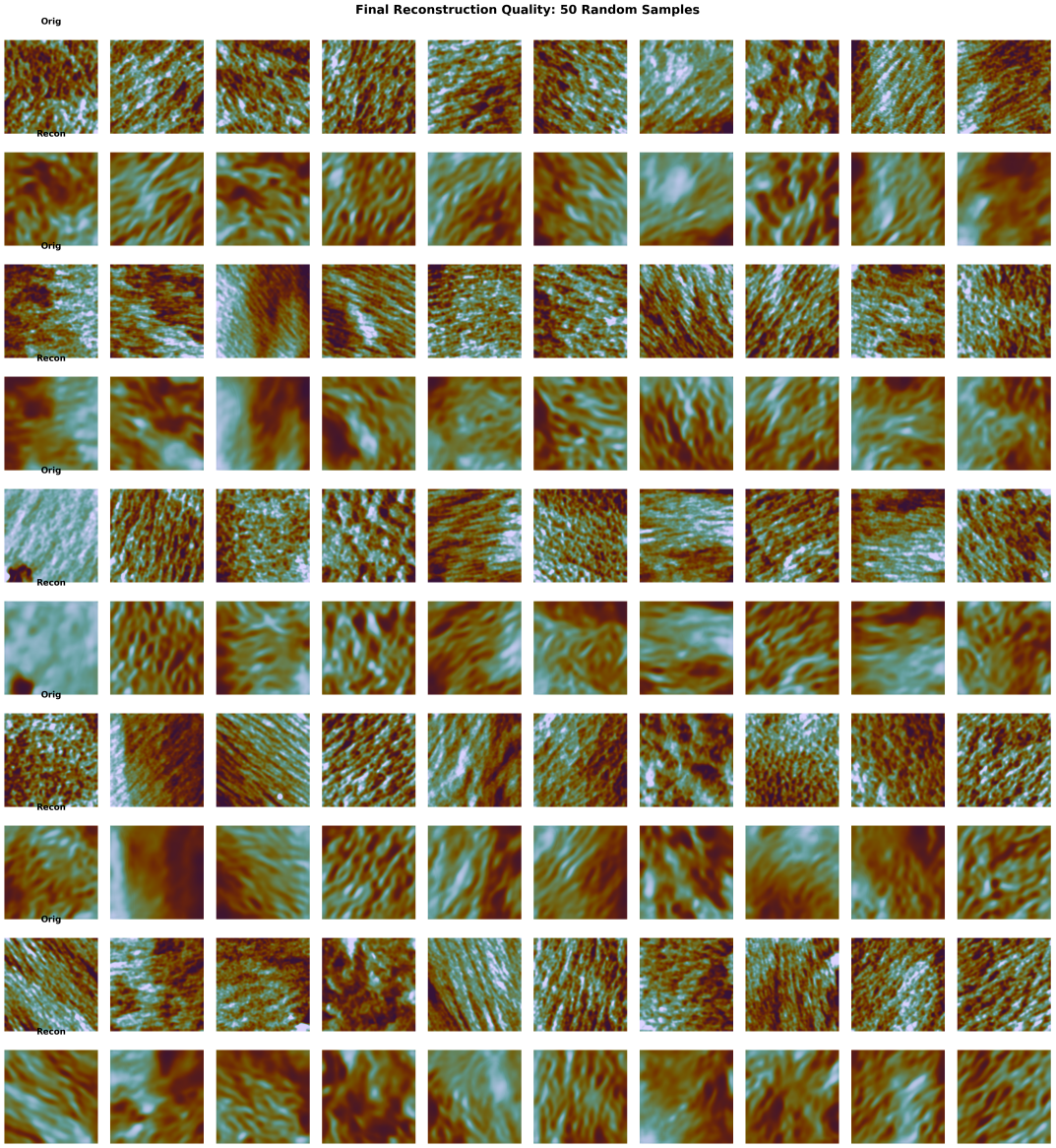


Figure A.2: Reconstructions of test set Dataset 2015, squeezed to 64 dimensions

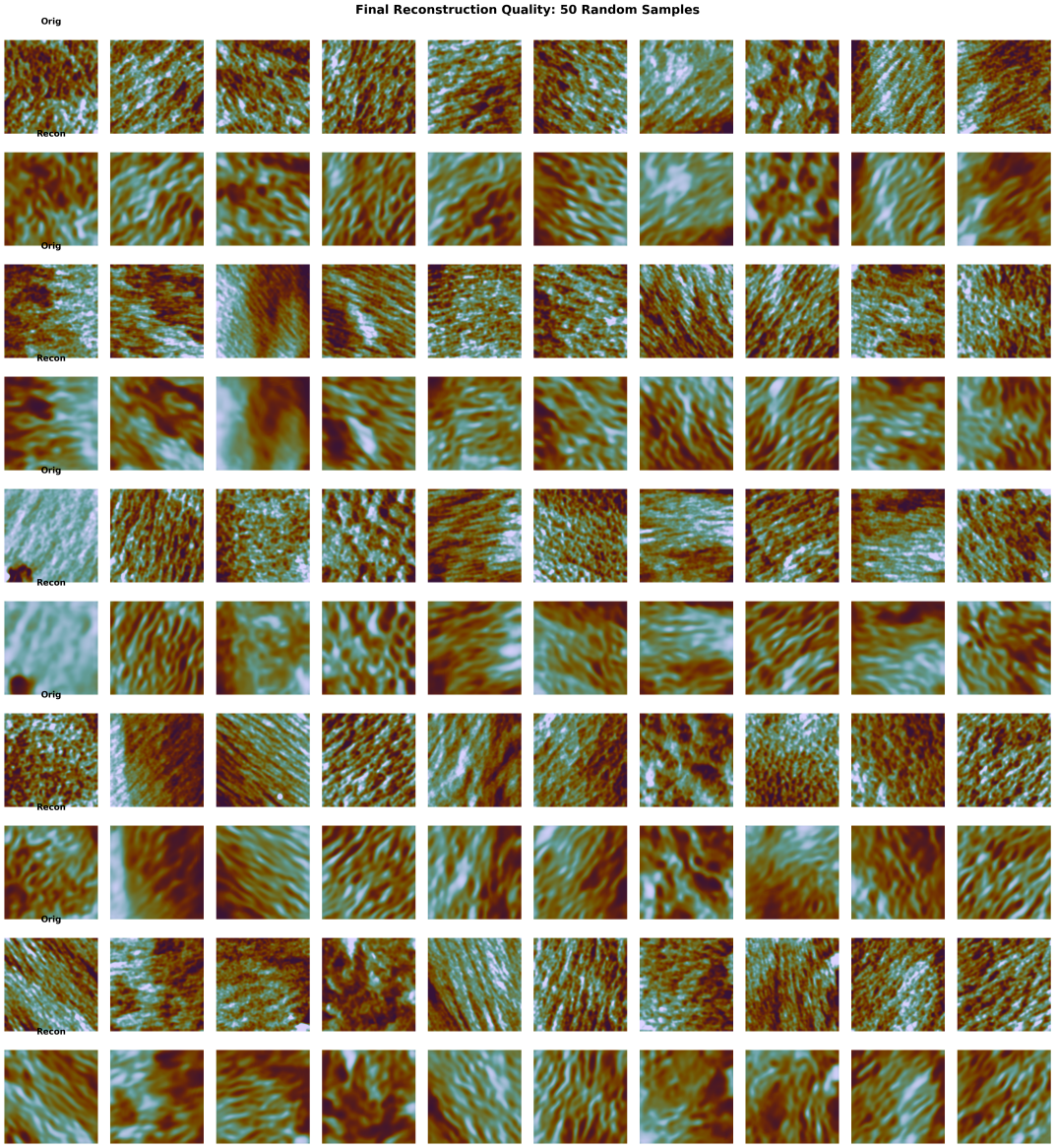


Figure A.3: Reconstructions of test set Dataset 2015, squeezed to 128 dimensions

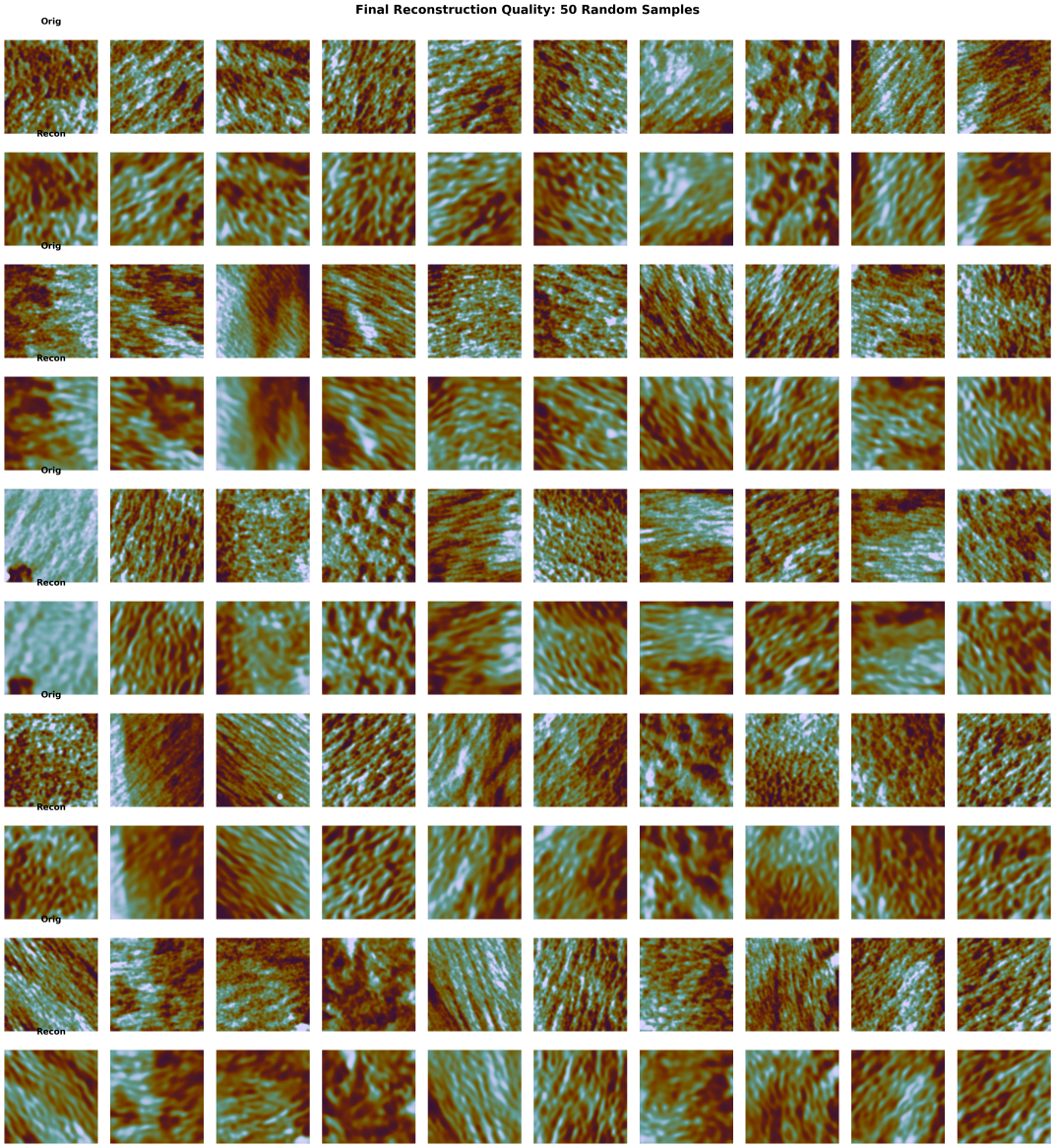
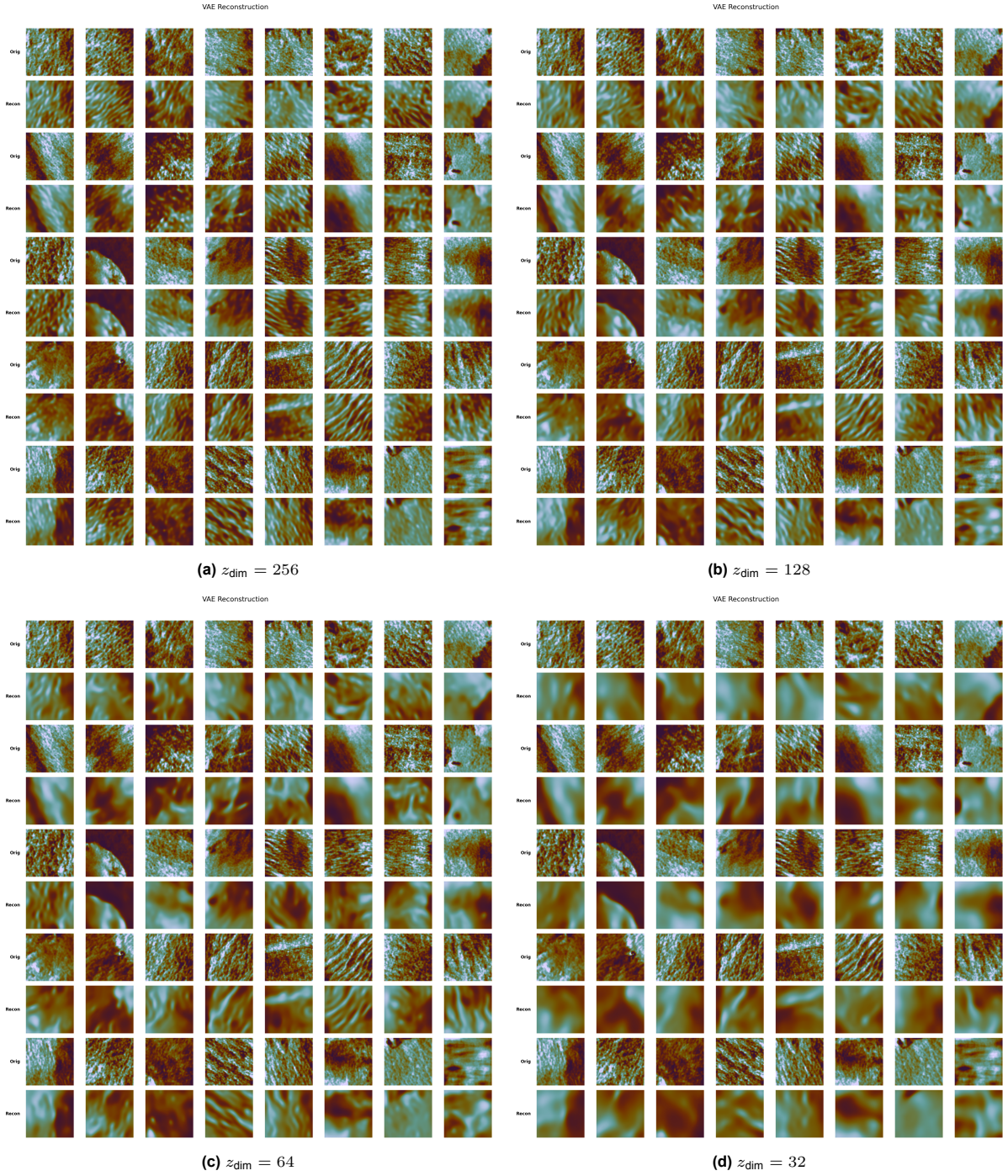


Figure A.4: Reconstructions of test set Dataset 2015, squeezed to 256 dimensions

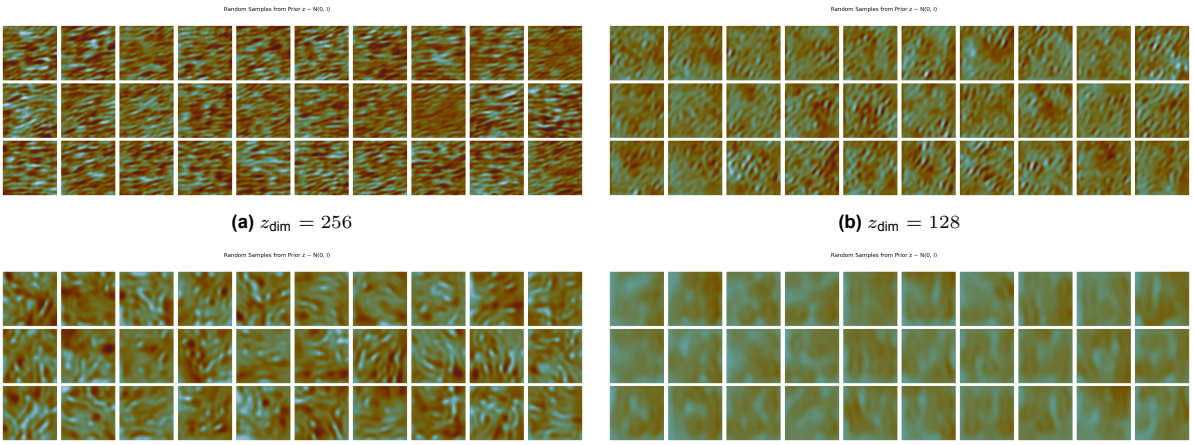
A.1.2. β annealing



(a) $z_{\text{dim}} = 256$ **(b)** $z_{\text{dim}} = 128$

(c) $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 32$

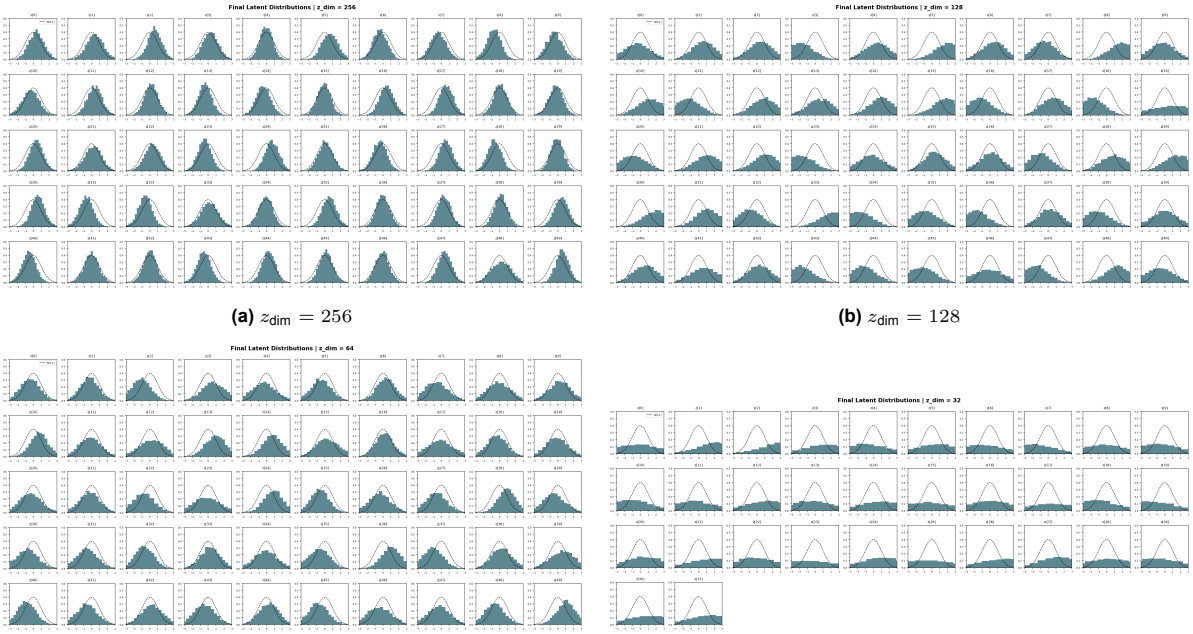
Figure A.5: reconstructions for different latent dimensions z_{dim} on test set after performing β annealing up to $\beta_{optimal}$



(a) $z_{\text{dim}} = 256$ **(b)** $z_{\text{dim}} = 128$

(c) $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 32$

Figure A.6: Sampling from the latent space for different latent dimensions z_{dim} on test set after performing β annealing up to $\beta_{optimal}$



(a) $z_{\text{dim}} = 256$ **(b)** $z_{\text{dim}} = 128$

(c) $z_{\text{dim}} = 64$ **(d)** $z_{\text{dim}} = 32$

Figure A.7: latent dimension distribution for different z_{dim} on test set after training with β annealing up to $\beta_{optimal}$

A.1.3. Optuna Optimisation

Table A.1: Phase 1 hyperparameter search space for Model 1 VAE optimization.

Parameter	Search Range / Options
Architecture	
n blocks	$\{3, 4, 5, 6\}$
base channels	$\{32, 64, 96, 128\}$
kernel size	$\{3, 5, 7\}$
pooling type	$\{\text{max}, \text{avg}\}$
Training	
learning rate	$[10^{-5}, 10^{-3}]$
batch size	$\{16, 32, 64\}$
optimizer	$\{\text{Adam}, \text{AdamW}, \text{SGD}\}$
Loss Function	
β	$[10^{-7}, 10^{-5}]$
λ_{freq}	$[1.0, 3.0]$
$\gamma_{\text{MSE, max}}$	$[0.5, 1.5]$
$\gamma_{\text{MSE, min}}$	$[0.01, 0.1]$
anneal epochs	$\{10, 20, 40\}$

Table A.2: Hyperparameter search space for regression training (Model 1).

Parameter	Search Range
number of regression layers (n_{reg})	$\{5, 10\}$
regression hidden dimension	$\{64, 96, 128\}$
regression dropout	$[0.1, 0.4]$
batch size	$\{32, 64\}$
learning rate	$[8 \times 10^{-5}, 4 \times 10^{-4}]$ (log scale)
optimizer	{Adam, AdamW, SGD}
weight decay	$[10^{-6}, 10^{-3}]$

Table A.3: Phase 1 hyperparameter search space for Model 2 joint training optimization.

Parameter	Search Range / Options
VAE Architecture	
n_blocks	$\{3, 4, 5, 6\}$
base_channels	$\{32, 64, 96, 128\}$
kernel_size	$\{3, 5, 7\}$
pooling_type	$\{\text{max}, \text{avg}\}$
Regression Architecture	
n_reg_layers	$\{5, 6, 7, 8, 9\}$
reg_hidden_dim	$\{32, 64, 96, 128\}$
reg_dropout	$[0.15, 0.35]$
Training	
learning_rate	$[10^{-5}, 10^{-3}]$
batch_size	$\{16, 32, 64\}$
optimizer	$\{\text{Adam}, \text{AdamW}, \text{SGD}\}$
Loss Function	
β	$[10^{-7}, 10^{-5}]$
λ_{freq}	$[1.0, 3.0]$
α_{flux}	$[1.0, 2.0]$
$\gamma_{\text{MSE, max}}$	$[0.5, 1.5]$
$\gamma_{\text{MSE, min}}$	$[0.01, 0.1]$
anneal_epochs	$\{10, 20, 40, 80\}$

Table A.4: Phase 1, Model 1: Best Hyperparameter Configuration for each latend dimension

Hyperparameter	$z = 32$	$z = 64$	$z = 128$	$z = 256$
Loss Hyperparameters				
β	1.78×10^{-7}	1.86×10^{-7}	1.00×10^{-7}	1.38×10^{-7}
λ_{freq}	1.89	1.28	1.88	1.70
γ_{max}	0.743	1.285	0.843	0.857
γ_{min}	0.072	0.046	0.094	0.058
anneal_epochs	40	20	10	40
Architecture Hyperparameters				
n_blocks	5	4	3	4
base_channels	128	96	128	32
kernel_size	3	7	7	7
pooling_type	max	max	max	max
Training Hyperparameters				
learning rate (lr)	3.86×10^{-4}	3.33×10^{-5}	8.70×10^{-5}	1.18×10^{-4}
batch_size	64	32	64	64
optimizer	Adam	Adam	Adam	AdamW

Table A.5: Phase 1, Model 2: Best Hyperparameter Configuration for Each Latent Dimension

Hyperparameter	$z = 32$	$z = 64$	$z = 128$	$z = 256$
β	3.0×10^{-6}	2.0×10^{-6}	0.0	2.0×10^{-6}
λ_{freq}	2.4512	2.4705	2.4753	1.5915
α_{flux}	1.9872	1.3927	1.6260	1.1911
γ_{max}	0.8187	0.9779	0.7688	0.9702
γ_{min}	0.0973	0.0717	0.0445	0.0546
anneal epochs	10	10	10	20
n blocks	4	4	4	3
channels	[128,64,32]	[64,64,128]	[64,64,128]	[64,64]
kernel size	5	7	5	5
pooling type	max	max	avg	max
n reg layers	5	7	7	6
reg hidden dim	96	64	128	96
reg dropout	0.2590	0.1619	0.2159	0.1936
learning rate	1.64×10^{-4}	1.85×10^{-4}	1.06×10^{-4}	1.37×10^{-4}
batch size	32	32	32	32

Table A.6: Average hyperparameter importance for Model 1 Phase 2 optimization, computed across all latent dimensions.

Hyperparameter	Importance
optimizer	0.229
batch_size	0.119
β	0.109
learning_rate	0.099
$\gamma_{\text{MSE, max}}$	0.098
$\gamma_{\text{MSE, min}}$	0.092
λ_{freq}	0.081
base_channels	0.032
kernel_size	0.021
use_batchnorm	0.020
n_blocks	0.011
pooling_type	0.010

Table A.7: Top 15 most important hyperparameters for Model 2 Phase 1 optimization, averaged across all latent dimensions.

Hyperparameter	Importance
reg_dropout	0.193
$\gamma_{\text{MSE, max}}$	0.125
learning_rate	0.110
α_{flux}	0.106
β	0.099
$\gamma_{\text{MSE, min}}$	0.095
reg_hidden_dim	0.088
λ_{freq}	0.085
kernel_size	0.027
ch_block_2	0.025
n_reg_layers	0.018
n_blocks	0.011
pooling_type	0.007

A.1.4. Stage 2

Table A.8: Final optimized hyperparameters for Model 1 VAE (Stage 1) across different latent dimensions z_{dim}. All other architectural parameters were fixed

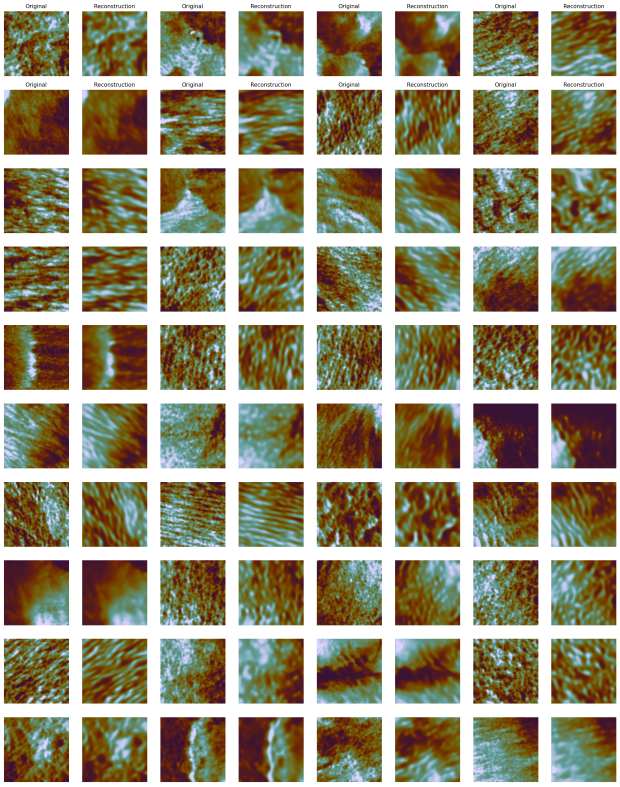
Hyperparameter	$z_{\text{dim}} = 32$	$z_{\text{dim}} = 64$	$z_{\text{dim}} = 128$	$z_{\text{dim}} = 256$
β	1.62×10^{-6}	3.70×10^{-7}	2.18×10^{-7}	6.34×10^{-7}
λ_{freq}	1.05	1.24	1.26	2.21
recon weight min	0.075	0.097	0.057	0.045
recon weight max	1.29	1.83	1.52	1.57
learning rate	4.42×10^{-4}	1.81×10^{-4}	2.35×10^{-4}	1.16×10^{-4}
batch size	32	32	32	32
optimizer	AdamW	Adam	Adam	AdamW

Table A.9: Final optimized hyperparameters for Model 2 (final stage) across latent dimensions z_{dim}. Fixed across all runs were: anneal epochs = 10, kernel size = 5, channels = [64, 64, 128], batch size = 32, number of regression layers = 7, number of blocks = 4, and max pooling.

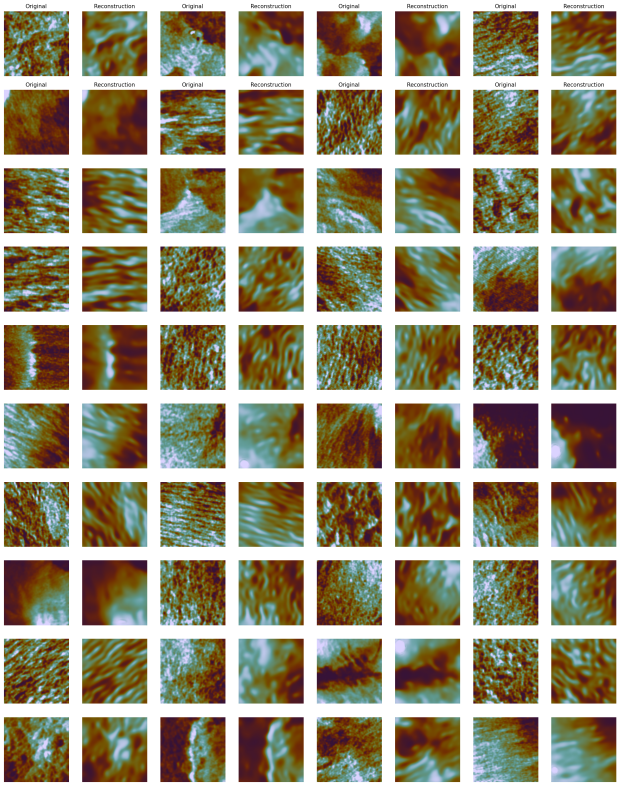
Hyperparameter	$z_{\text{dim}} = 32$	$z_{\text{dim}} = 64$	$z_{\text{dim}} = 128$	$z_{\text{dim}} = 256$
reg hidden dim	96	128	96	96
reg dropout	0.112	0.100	0.127	0.168
recon weight min	0.099	0.080	0.024	0.054
recon weight max	0.787	0.753	0.881	0.728
λ_{freq}	2.81	2.22	1.23	1.43
α_{flux}	1.04	1.87	1.11	1.32
β	1.57×10^{-5}	7.17×10^{-6}	2.17×10^{-6}	1.60×10^{-7}
learning rate	1.53×10^{-4}	8.48×10^{-5}	1.43×10^{-4}	2.22×10^{-4}

A.2. Best Model performance

A.2.1. Model 1

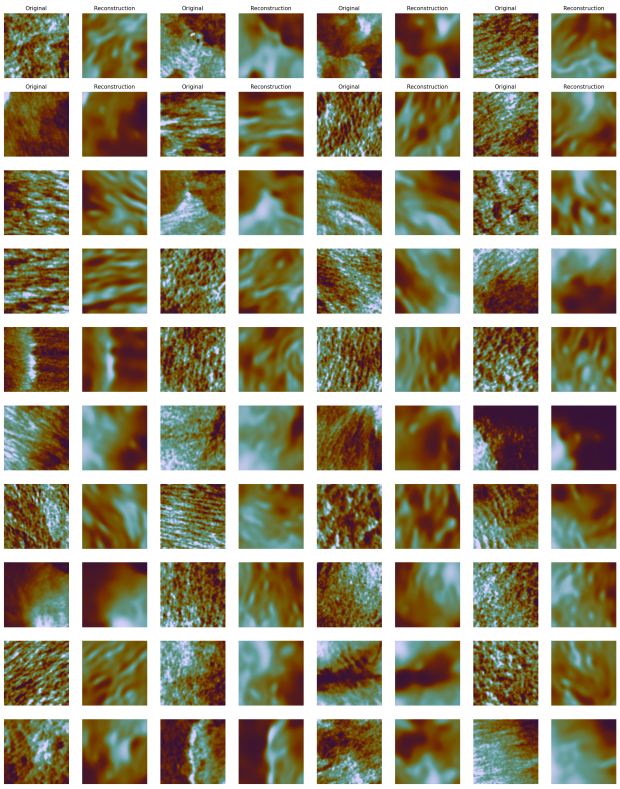


(a) $z_{\text{dim}} = 256$

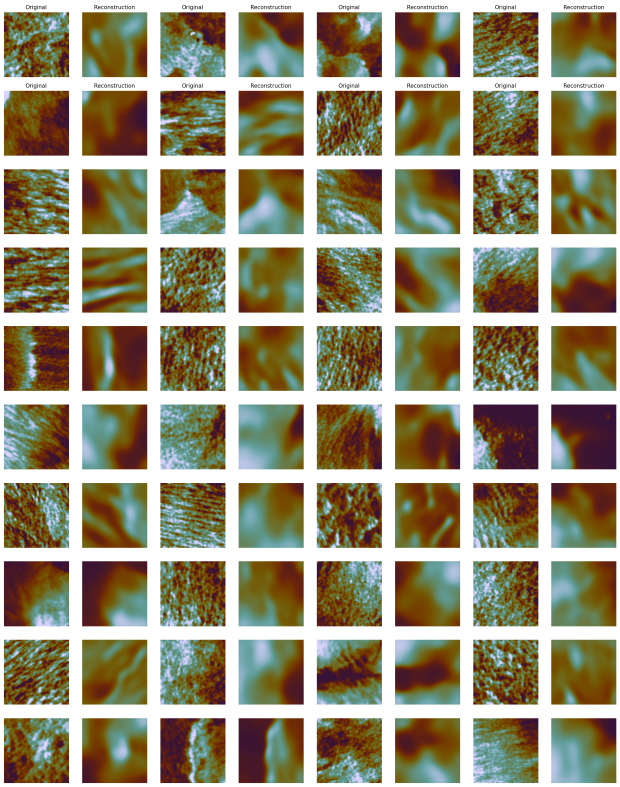


(b) $z_{\text{dim}} = 128$

Figure A.8: Model 1 test set reconstructions for higher dimensions. Columns alternate between original and reconstructed images. Both dimensions capture diverse phenomena including windstreaks, atmospheric fronts, and ocean patterns with high fidelity.



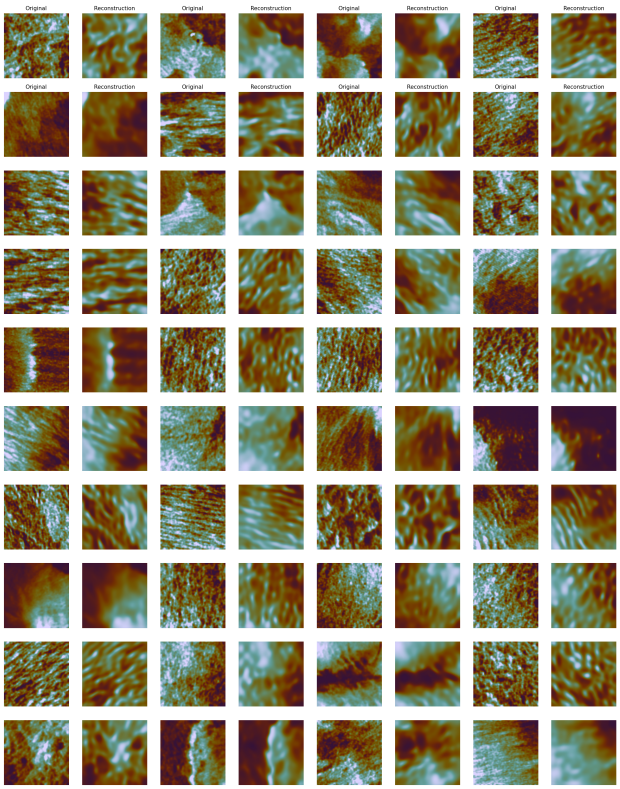
(a) $z_{\text{dim}} = 64$



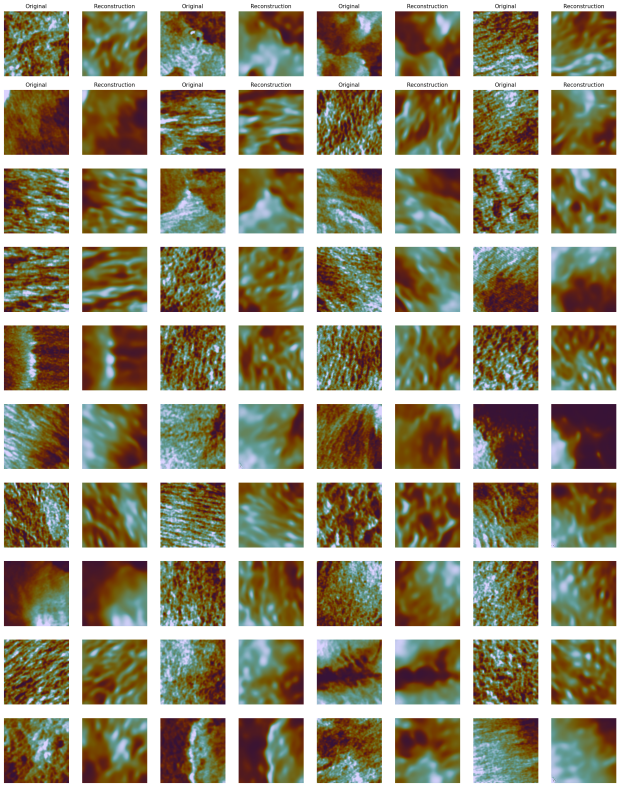
(b) $z_{\text{dim}} = 32$

Figure A.9: Model 1 test set reconstructions for lower dimensions. Large-scale patterns are preserved, but fine details are lost, particularly at $z_{\text{dim}} = 32$.

A.2.2. Model 2

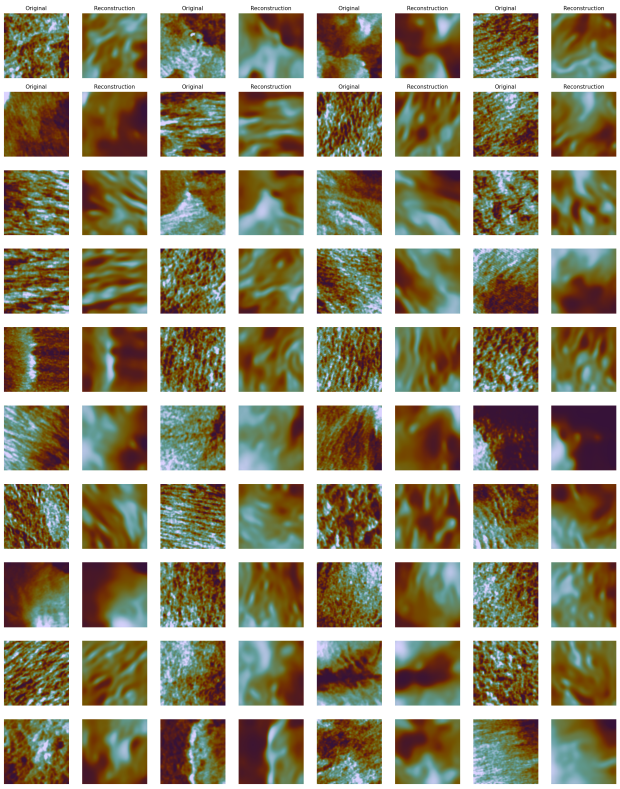


(a) $z_{\text{dim}} = 256$

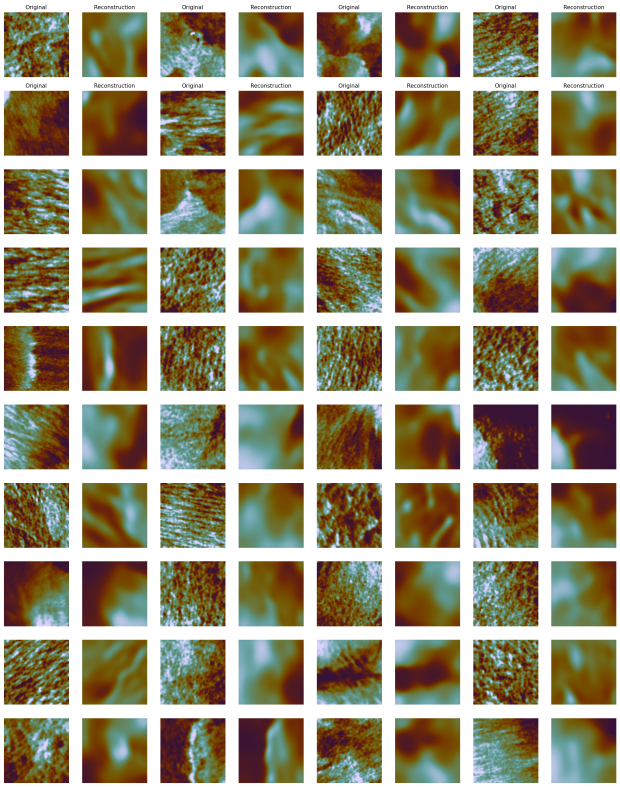


(b) $z_{\text{dim}} = 128$

Figure A.10: Model 2 test set reconstructions for higher dimensions. Compared to Model 1 (Figure A.8), reconstructions show slightly reduced sharpness despite flux prediction integration.



(a) $z_{\text{dim}} = 64$



(b) $z_{\text{dim}} = 32$

Figure A.11: Model 2 test set reconstructions for lower dimensions.