DIFFUSION-BASED CHANNEL IMPAIRMENT MITIGATION AND RESOLUTION ENHANCEMENT FOR WIRELESS IMAGE TRANSMISSION

VAN HUNG LE

A Thesis Submitted to the Faculty of Graduate Studies

In Partial Fulfillment of the Requirements

for the Degree of Master of Science

Graduate Program in Electrical Engineering and Computer Science

York University

Toronto, Ontario

September 2026

© Van Hung Le, 2026

Abstract

Diffusion models have recently been explored to mitigate channel-induced degradation in wireless image transmission. However, many existing approaches rely on explicit channel state information (CSI) to guide the restoration process. Practically, acquiring accurate and timely CSI may introduce overhead and be unreliable in dynamic wireless environments due to estimation errors and outdated channel information. In bandwidth-constrained or hardware-impaired systems, transmitted images are often low-resolution with limited visual details. Therefore, the receiver must jointly mitigate channel distortion and enhance image resolution. In this thesis, a conditional latent diffusion framework is proposed that conditions exclusively on the degraded received image, without requiring explicit channel information. The proposed model jointly performs channel impairment mitigation and image resolution enhancement, enabling reconstruction of higher-quality images from low-resolution degraded input. The effectiveness of the proposed method is evaluated and validated via extensive simulations under additive white Gaussian noise and Rayleigh fading channels across different noise levels.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Ping Wang, for her continuous guidance, support, and encouragement throughout my research. Her insightful feedback helped me refine my ideas and guided me in the right direction, from the initial stage of a rough idea to the concrete implementation of this project. I would also like to sincerely thank Professor Aijun An and Professor Liang Xue for their help in reviewing my thesis and serving on my thesis examination committee.

Table of Contents

Abstract ii

Acknowledgements iii

Table of Contents iv

List of Tables vii

List of Figures viii

List of Abbreviations x

1 Introduction 1

1.1 Motivation 1

1.2 Contribution 2

1.3 Thesis Organization 3

2 Background 5

2.1 Wireless Image Transmission and Channel Impairments 5

2.1.1 Additive White Gaussian Noise (AWGN) Channel 6

2.1.2 Rayleigh Fading Channel 7

2.2 Image Super-Resolution and Image Restoration 7

2.3 Convolutional Neural Networks and U-Net 9

2.4 Swin Transformer 11
2.5 Diffusion Models (DMs) 11

3 Literature Review **14**
3.1 Channel Denoising and Robust Image Transmission 14
3.2 Generative Models for Semantic and Bandwidth Constrained Image Transmission 16

4 The Proposed Method **19**
4.1 Problem Statement 20
4.2 Overall System Design 20
4.3 Conditional Diffusion Model 21
4.3.1 Diffusion Forward Process 22
4.3.2 Conditional Reverse Process 23
4.4 Diffusion Model Preconditioning 24
4.5 Swin Transformer Autoencoder 26
4.6 Conditional Latent Diffusion Model 27
4.6.1 Inference Algorithm (Second-Order Heun Sampler) 28

5 Experiments **31**
5.1 Evaluation Metrics 31
5.2 Dataset 33
5.3 Implementation Details 33
5.4 Benchmark Methods 35
5.5 Experimental Results 35
5.5.1 Quantitative Results 36
5.5.2 Qualitative Results 41
5.6 Ablation study 44

6 Conclusion **47**

6.1 Summary 47

6.2 Future Work 48

Bibliography **49**

List of Tables

5.1 PSNR comparison between the proposed model and JSCC+CDDM under Rayleigh channel for different input resolutions and SNR values on DIV2K dataset. 35
5.2 SSIM comparison between the proposed model and JSCC+CDDM under Rayleigh channel for different input resolutions and SNR values on DIV2K dataset. 36
5.3 General reconstruction performance of our model on the FFHQ dataset under the Rayleigh fading channel with 16×16 transmitted images. 40
5.4 General reconstruction performance of our model on the FFHQ dataset under the Rayleigh fading channel with 64×64 transmitted images. 40
5.5 Cross-channel generalization performance of the proposed model. 45

List of Figures

2.1 Diagram of a wireless image transmission system, showing the transmission of a source image through a wireless channel and the reconstruction of the image at the receiver. 6

2.2 General architecture of a image super-resolution model. 8

2.3 General architecture of UNet. 10

2.4 General architecture of Swin Transformer. 11

2.5 Overview of the diffusion process. 12

4.1 Overall architecture of the proposed receiver-side conditional latent diffusion framework. A low-resolution image is transmitted through a wireless channel, encoded into a conditional latent representation, restored by the latent diffusion model, and decoded into a high-resolution reconstructed image. 19

4.2 Visualization of the proposed forward perturbation process in latent space. The clean high-resolution latent representation $z_0 = E_\phi(x_{\text{HR}})$ is corrupted by Gaussian noise at different noise levels σ_i to generate the noisy latent $z_{\sigma_i} = z_0 + \sigma_i \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. 25

5.1 PSNR comparison between the proposed diffusion model and the corresponding non-diffusion reconstruction baseline under AWGN and Rayleigh fading channels at different SNR levels. 37

5.2 SSIM comparison between the proposed diffusion model and the corresponding non-diffusion reconstruction baseline under AWGN and Rayleigh fading channels at different SNR levels. 38
5.3 PSNR comparison among the proposed diffusion model, the non-diffusion reconstruction baseline, and the JPEG2000-based baseline under AWGN and Rayleigh fading channels at different SNR levels. 39
5.4 Visual reconstruction results under the Rayleigh fading channel at different SNR levels on 16×16 image. The first row shows the degraded low-resolution received images, and the second row shows the corresponding high-resolution reconstructions produced by the proposed model. 41
5.5 Visual reconstruction results under the AWGN channel at different SNR levels on 16×16 image. The first row shows the degraded low-resolution received images, and the second row shows the corresponding high-resolution reconstructions produced by the proposed model. 41
5.6 Reconstruction results on the FFHQ dataset 16×16 under AWGN and Rayleigh fading channels. 42
5.7 Reconstruction results on the FFHQ dataset 64×64 under AWGN and Rayleigh fading channels. 43
5.8 Reconstruction results on the DIV2K dataset 64×64 under AWGN and Rayleigh fading channels. 44
5.9 Effect of conditional latent mismatch on PSNR under the Rayleigh fading channel. The conditioning latent is progressively corrupted by interpolating the correct conditional latent with a shuffled latent from another image, where larger η indicates stronger conditional mismatch. 46

List of Abbreviations

AE: Autoencoder

AWGN: Additive White Gaussian Noise

BCR: Bandwidth Compression Ratio

CDDM: Channel Denoising Diffusion Model

CNN: Convolutional Neural Network

CSI: Channel State Information

DDIM: Denoising Diffusion Implicit Models

DDPM: Denoising Diffusion Probabilistic Models

DIV2K: Diverse 2K Resolution High Quality Images

DM: Diffusion Model

EDM: Elucidated Diffusion Model

ELBO: Evidence Lower Bound

FFHQ: Flickr-Faces-HQ

GAN: Generative Adversarial Network

HR: High Resolution

IoT: Internet of Things

JSCC: Joint Source-Channel Coding

LDM: Latent Diffusion Model

LPIPS: Learned Perceptual Image Patch Similarity

LR: Low Resolution

MAE: Mean Absolute Error

MIMO: Multiple-Input Multiple-Output

MSE: Mean Squared Error

ODE: Ordinary Differential Equation

PSNR: Peak Signal-to-Noise Ratio

SDE: Stochastic Differential Equation

SNR: Signal-to-Noise Ratio

SR: Super-Resolution

SSIM: Structural Similarity Index Measure

ST: Swin Transformer

VAE: Variational Autoencoder

VGG: Visual Geometry Group

Chapter 1

Introduction

1.1 Motivation

Wireless image transmission has become an important research topic in modern communication systems, especially in scenarios where bandwidth is limited and the communication channel is severely corrupted by noise and fading. In some practical wireless systems, such as wireless edge networks and Internet of Things (IoT) applications, hardware limitations, including low-resolution image sensors, as well as bandwidth constraints, often necessitate the transmission of low-resolution images. This inherently limits the quality of reconstruction on the receiver side. As a result, the receiver is required to recover visually meaningful high-resolution images from observations that are degraded not only by channel noise and fading, but also by severe resolution loss.

The integration of deep neural networks into wireless communication systems has been extensively investigated in recent years. The surveys in [1, 2] review the applications of generative models, including variational autoencoders (VAEs) [3] and generative adversarial networks (GANs) [4], in tasks such as channel modelling, signal reconstruction, data compression, and joint source-channel coding (JSCC). However, GANs often suffer from unstable adversarial training and mode collapse, which may be limited in dynamic wireless channel conditions. VAEs offer

more stable and computationally efficient training, but their approximate posterior or posterior collapse may produce blurry or degraded outputs. In contrast, diffusion models provide stable training and quality sample generation. Furthermore, latent diffusion models reduce computational cost by performing the diffusion process in a compressed latent space. These advantages motivate the use of a conditional latent diffusion model in this work to reconstruct high quality images from low-resolution observations degraded by wireless channel impairments.

Recent works have incorporated diffusion models into wireless transmission [5, 6] and semantic communication [7, 8]. In particular, diffusion-based methods have been used to mitigate channel distortion and improve the reconstruction quality of received images. However, many existing approaches rely on channel state information (CSI), such as channel coefficients or equalization modules to guide the restoration process. In practice, obtaining accurate CSI requires additional channel estimation procedures, which consume bandwidth, time, and energy and therefore introducing significant communication overhead. In addition, estimated CSI may contain errors due to channel noise, interference, limited pilot resources, or imperfect estimation algorithms, and it may quickly become outdated in dynamic wireless environments. Consequently, there is a strong need for more practical wireless image transmission solutions that can operate effectively without requiring explicit CSI. Furthermore, existing diffusion-based methods for wireless image transmission primarily focus on channel denoising, while paying limited attention to the challenge of image resolution enhancement. Because of those reasons, we propose a conditional latent diffusion framework for wireless image transmission that explicitly address the joint problem of mitigating channel degradation while enhancing severely low-resolution inputs, without requiring explicit CSI.

1.2 Contribution

The main contributions of this work are summarized as follows:

1. We propose a receiver-side conditional latent diffusion framework that reconstructs high-

resolution images from degraded low-resolution received observations. The proposed framework jointly mitigates channel-induced degradation and enhances image resolution using only the received image as conditional guidance, without requiring explicit CSI.

2. We perform the diffusion-based restoration process in a compact latent space to reduce the computational cost of training and inference. Specifically, a Swin Transformer (ST) autoencoder is employed to extract latent representations from received observations. These latent representations are then used by an Elucidating Diffusion Model (EDM)-based conditional diffusion model to efficiently construct the clean high-resolution image.

3. We evaluate the proposed method a variety of channel environments including additive white Gaussian noise (AWGN) and Rayleigh fading channels across a range of signal-to-noise ratio (SNR) conditions using Peak Signal-to-Noise Ratio (PSNR) [1] and Structural Similarity Index Measure (SSIM) [2] The results show that the proposed method improves reconstruction quality over the baseline reconstruction model. We also compare the proposed method with a representative diffusion-based benchmark, highlighting that our framework outperforms the benchmark method and effectively performs joint channel impairment mitigation and resolution enhancement without requiring explicit CSI, while using fewer diffusion sampling steps.

1.3 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 presents the necessary background on wireless communication channels, image restoration and super-resolution, Swin Transformers, and diffusion models. Chapter 3 reviews existing studies on robust wireless image transmission, deep joint source-channel coding (DeepJSCC), generative models, and diffusion-based

[1] PSNR measures image reconstruction fidelity based on the pixel-wise error between the ground-truth image and reconstructed image. A higher PSNR indicates better reconstruction quality.

[2] SSIM is used to measure the structural similarity between the ground-truth image and reconstructed image, SSIM focuses on measuring luminance, contrast, and local structural information. A higher SSIM value indicates greater structural similarity.

communication systems. Chapter 4 introduces the proposed conditional latent diffusion framework, including the system model, Swin Transformer autoencoder, conditional EDM-based denoising network, training procedure, and inference process. Chapter 5 describes the experimental setup, datasets, channel configurations, evaluation metrics, implementation details, and baseline methods. This chapter also presents and discusses the quantitative and qualitative results under AWGN and Rayleigh fading channels. Finally, Chapter 6 concludes the thesis and outlines possible directions for future work.

Part of the work presented in this thesis has recently been accepted at the IEEE Global Communications Conference (GLOBECOM), a flagship conference of the IEEE Communications Society. An extended journal paper based on this research is currently in preparation.

Chapter 2

Background

In this chapter, we begin by introducing wireless image transmission and the key challenges that degrade image quality during wireless transmission. Then we introduce the fundamentals of image super-resolution. Next, we discuss convolutional neural networks, the U-Net architecture, and the Swin Transformer, which provide the architecture used for image feature extraction and reconstruction. Finally, we present diffusion models and their underlying principles, providing the necessary background to understand how they can effectively address the challenges in wireless image transmission.

2.1 Wireless Image Transmission and Channel Impairments

Wireless image transmission is an important topic in modern communication systems. It is the process of transmitting visual information from a source device to a receiver through a wireless communication channel. Wireless image transmission has been investigated in various practical applications. For example, in [9], DeepJSCC is adapted to address the challenges of noise and limited bandwidth in satellite-to-ground image transmission. In [10], deep neural networks (DNNs) are used for image recognition tasks in IoT systems with limited resources. In [11], a smart-city application is considered, where a vision-centered semantic communication framework is proposed

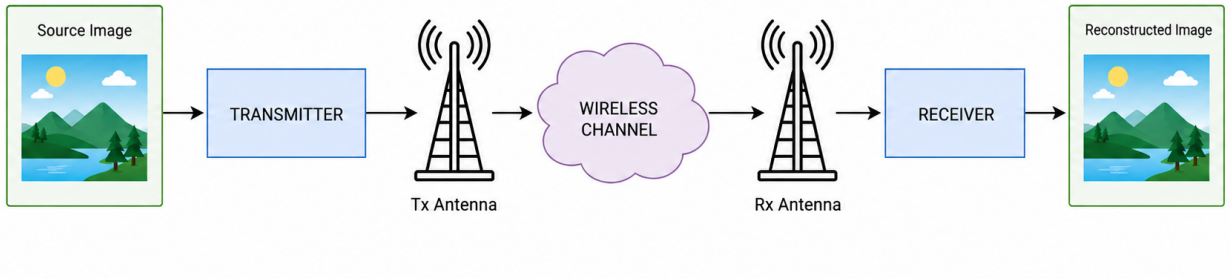


Figure 2.1: Diagram of a wireless image transmission system, showing the transmission of a source image through a wireless channel and the reconstruction of the image at the receiver.

to transmit images over noisy and limited wireless channels. Unlike wired transmission, wireless transmission is affected by channel uncertainties, including noise, fading, interference, and bandwidth limitations. These factors can distort the transmitted signal and degrade the quality of the received images.

Figure 2.1 illustrates the general block diagram of a wireless image transmission system. The source image is first processed by the transmitter and then sent through the transmit antenna which propagates through the wireless channel. During propagation, the signal may be affected by noise, fading, interference, and bandwidth limitations. At the receiver side, the received signal is captured by the receive antenna and processed by the receiver to reconstruct the image.

2.1.1 Additive White Gaussian Noise (AWGN) Channel

The AWGN channel is one of the most commonly adopted channel models in wireless transmission. It models the received signal as the transmitted signal corrupted by additive Gaussian noise:

$$y = x + n, \quad n \sim \mathcal{N}(0, \sigma^2), \tag{2.1}$$

where y is the received signal, x is the transmitted signal, and n is a Gaussian noise component with zero mean and variance σ^2.

2.1.2 Rayleigh Fading Channel

In a Rayleigh fading channel, the transmitted signal is affected by a random fading coefficient in addition to additive noise. This model is commonly used to represent wireless environments where there is no dominant line-of-sight path between the transmitter and receiver. The received signal can be expressed as

$$y = hx + n, \quad n \sim \mathcal{N}(0, \sigma^2), \tag{2.2}$$

where y is the received signal, x is the transmitted signal, h is the Rayleigh fading channel coefficient, and n is the additive Gaussian noise. The magnitude of h, denoted by $r = |h|$, follows a Rayleigh distribution. Its probability density function is given by

$$f_R(r) = \frac{r}{\sigma_h^2} \exp\left(-\frac{r^2}{2\sigma_h^2}\right), \qquad r \geq 0,$$

where σ_h is the scale parameter of the Rayleigh distribution [12].

2.2 Image Super-Resolution and Image Restoration

Image super-resolution (SR) is the process of enhancing low-resolution (LR) images to achieve high-resolution (HR) images. It plays an important role in image restoration, as LR images often suffer from the loss of fine details, textures, and edge information. Image super-resolution is particularly useful in practical Internet of Things (IoT) scenarios, where captured images may have low quality due to limited bandwidth, low-resolution sensors, and hardware impairments.

In [13], a comprehensive overview of image super-resolution covers serveral major approaches including convolutional neural networks (CNNs) [14, 15], generative adversarial networks (GANs) [4, 16], and commonly used image quality assessment metrics.

In a typical supervised SR setting, an LR image is generated from a ground-truth HR image

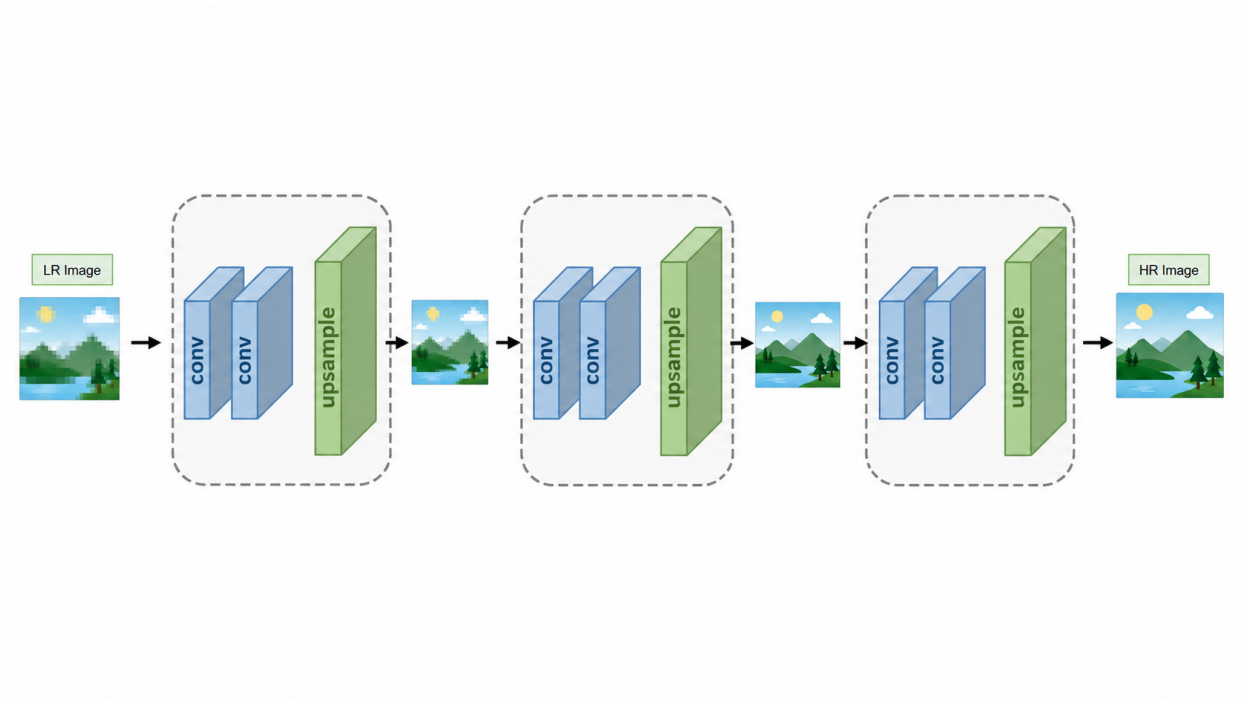


Figure 2.2: General architecture of a image super-resolution model.

through a degradation process. This can be formulated as

$$x = D(y),$$

where y represents the ground-truth HR image, x is the degraded LR image, and $D(\cdot)$ denotes the degradation function, such as down-sampling, blurring, or noise corruption.

The SR model then aims to reconstruct an estimated HR image from the LR input:

$$\hat{y} = F_\theta(x),$$

where $F_\theta(\cdot)$ represents the SR model and θ denotes the learnable network parameters. A common training objective is to minimize the mean squared error (MSE) between the reconstructed HR image $\hat{y}$ and the ground-truth HR image y:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \|y_i - \hat{y}_i\|_2^2$$

Figure 2.2 shows the overall structure of a convolutional neural network for image super-resolution. Given an LR input image, the network applies several convolutional layers to learn hierarchical feature representations. Upsampling layers are then used to increase the spatial resolution of the feature maps. Through repeated convolution and upsampling stages, the model gradually reconstructs an HR image that approximates the corresponding ground-truth image.

Image restoration aims to recover a clean image from a degraded observation. The study in [17] provides an overview of deep learning models used for image restoration. In general, a degraded image can be defined as

$$y = \mathcal{H}(x) + n,$$

where x is the clean image, y is the degraded observation, $\mathcal{H}(\cdot)$ represents the degradation process, and n denotes noise. The degradation process may include blur, downsampling, compression, or noise corruption. Traditional image restoration can be defined as

$$\hat{x} = \arg\min_{x} \|y - \mathcal{H}(x)\|^2 + \lambda\Phi(x),$$

where y is the degraded observation, $\mathcal{H}(\cdot)$ is the degradation operator, and $\lambda\Phi(x)$ represents an regularization term.

2.3 Convolutional Neural Networks and U-Net

Convolutional neural networks (CNNs) have been widely used in image processing because they are effective at extracting local image features, such as edges, textures, and spatial patterns. This makes them suitable for applications such as image denoising, image enhancement, or image reconstruction [18, 19, 15]. They are also widely adopted for image classification, such as AlexNet [20], Visual Geometry Group network (VGG) [21], and Residual Network (ResNet) [22], Densely Connected Convolutional Networks (DenseNet) and GoogleNet [23] demonstrated the effectiveness of deep CNNs for large-scale image recognition tasks. Traditional CNNs consist

of an input layer, convolutional layers, activation functions, and an output layer. In a convolutional layer, learnable filter kernels slide across the input image or feature map to extract meaningful features. The resulting feature map is usually passed through a nonlinear activation function, such as Rectified Linear Unit (ReLU) or sigmoid. The spatial size of the output feature map can be calculated as:

$$Output = \left\lfloor \frac{W - K + 2P}{S} \right\rfloor + 1$$

where (W) is the input size, (K) is the kernel size, (P) is the padding size, and (S) is the stride.

U-Net [24] is a CNN-based architecture that applies convolutional networks and reshapes them into a U-shaped encoder–decoder structure, as shown in Figure 2.3. The encoder path gradually contracts the input image through multiple convolutional layers and ReLU layers. The decoder path then expands it in reverse, and skip connections are used between corresponding layers to maintain important image features.

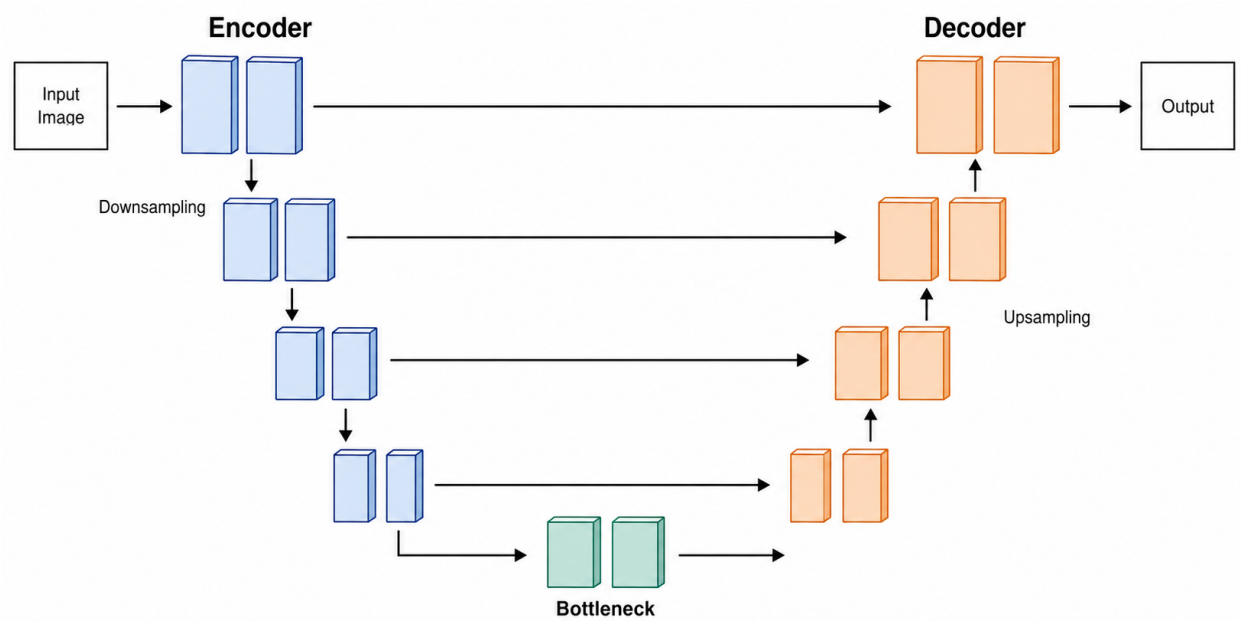


Figure 2.3: General architecture of UNet.

2.4 Swin Transformer

Swin Transformer (ST) [25] is a vision Transformer based architecture which is designed to work with computer vision tasks. Similarly to CNNs, it gradually extracts image features while reducing the spatial resolution to form hierarchical feature maps. Nevertheless, ST applies self-attention with local non-overlapping windows instead of traditional convolution layers. These windows are shifted between layers so as to capture both local visual patterns and global representation. Figure 2.4 illustrates the overall architecture of ST, where the input image is first divided into patches, which are treated as visual tokens. ST then applies self-attention within local windows, which reduces computational complexity. The window partition is shifted between consecutive layers.

2.5 Diffusion Models (DMs)

Diffusion probabilistic models were introduced by [26]. The main idea is to learn the data distribution by gradually adding noise to the data in forward process, and then train the deep neural network to remove this noise in reverse process.

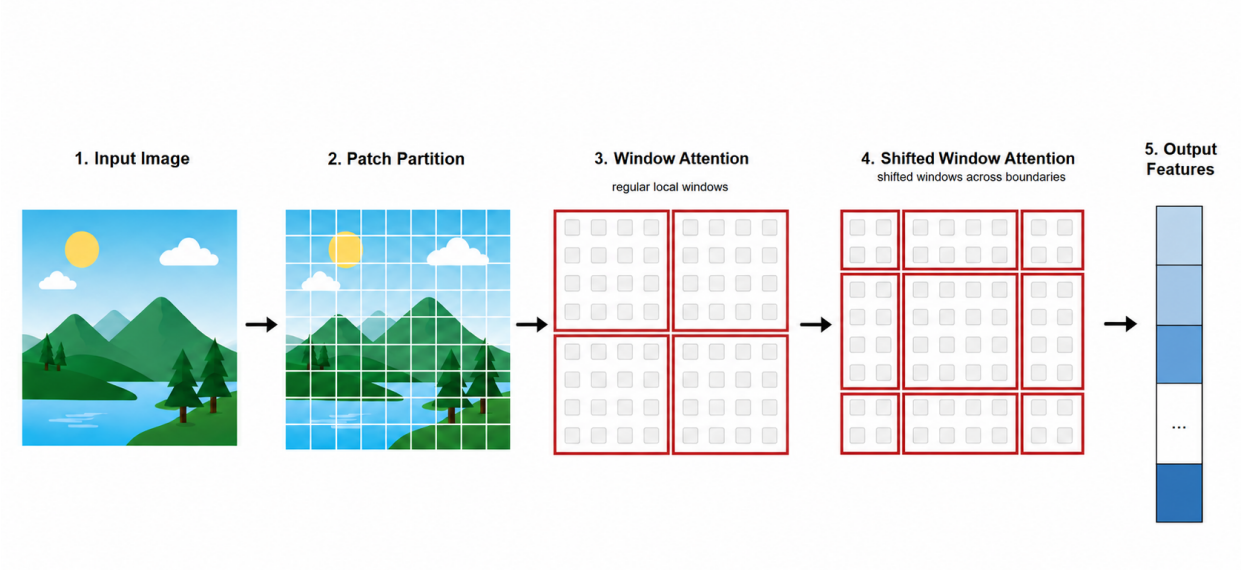


Figure 2.4: General architecture of Swin Transformer.

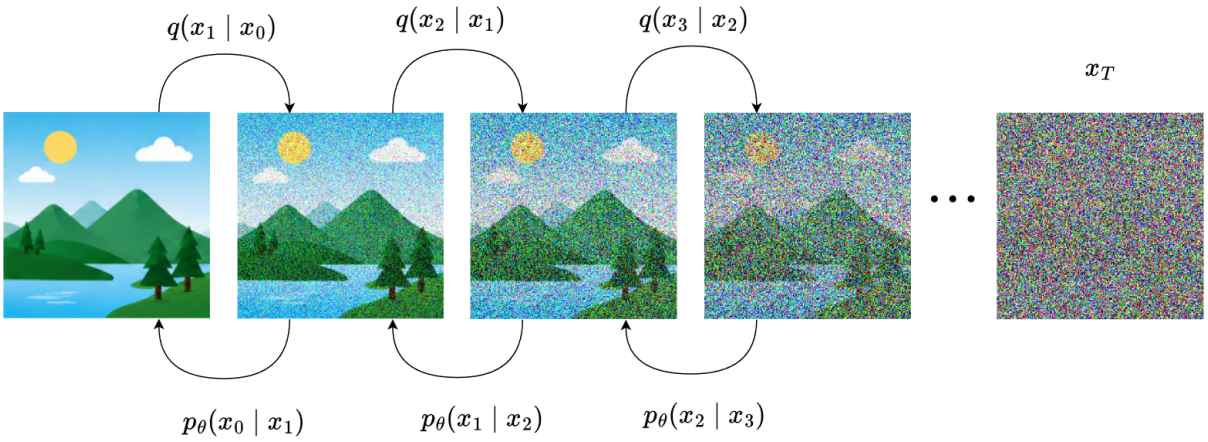


Figure 2.5: Overview of the diffusion process.

The forward process is defined as a Markov chain where a small amount of noise is added to the data at each timestep, and it can be defined as

$$q(x^{(0:T)}) = q(x^{(0)}) \prod_{t=1}^{T} q(x^{(t)} \mid x^{(t-1)}).$$

where $q(x^{(0)})$ is data distribution, $q(x^{(t)} \mid x^{(t-1)})$ is the forward process from one state to another.

The reverse process is defined in the opposite direction:

$$p(x^{(0:T)}) = p(x^{(T)}) \prod_{t=1}^{T} p(x^{(t-1)} \mid x^{(t)}).$$

where $p(x^{(T)})$ is pure noise, which can be Gaussian noise, $p(x^{(t-1)} \mid x^{(t)})$ is the backward process from one state to another.

Figure 2.5 illustrates the forward and reverse processes of a diffusion model. In the forward direction, the original image $\mathbf{x}_0$ is gradually corrupted by adding noise at each timestep according to $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$. Eventually, the image becomes pure Gaussian noise, denoted as $\mathbf{x}_T$. The reverse direction represents the learned denoising process. Starting from $\mathbf{x}_T$, the model $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ progressively remove noise and recover a clean sample.

Since the diffusion probabilistic model [26], many improved versions and different approaches have been proposed. The studies in [27, 28, 29] have further improved training and sampling ef-

ficiency, conditional modeling [30], restoration and enhancement capability [31], and computational efficiency [32]. Especially, DMs have higher generation quality and more stable training than GANs or VAEs [33, 34]. More recently, the work in [35] unified several design choices for diffusion training and sampling into a more effective and stable formulation. The role of DMs in wireless networks and communications is further investigated with recent surveys [5, 6] including channel modeling, signal detection, data reconstruction, and semantic communication. These studies suggest that diffusion models are particularly suitable for wireless reconstruction tasks due to their strong denoising capability and ability to model complex data distributions.

Chapter 3

Literature Review

In this chapter, we review the existing literature on wireless image communication, focusing on their applications in channel denoising and image quality improvement in bandwidth constrained transmission.

3.1 Channel Denoising and Robust Image Transmission

DeepJSCC was introduced in [36], which replaces the traditional separate source and channel coding pipeline with an end-to-end neural encoder-decoder framework. These encoder and decoder are built with CNNs architectures, as introduced in Section 2.3. Since then, several variants have been proposed to improve the robustness and adaptability of wireless image transmission under different channel conditions. For example, the attention-based JSCC method in [37] enables a single model to adapt image transmission across different SNR levels.

With the development of generative models, diffusion models have also been combined with DeepJSCC to further improve receiver-side reconstruction quality. In [38, 39], diffusion models are incorporated into DeepJSCC-based systems to mitigate channel distortion and improve the perceptual quality of received images. Similarly, the channel denoising diffusion model (CDDM) [8] introduces a diffusion-based framework for wireless semantic communication, where diffu-

sion models are used to mitigate AWGN and Rayleigh channel noise before image reconstruction. Moreover, the work in [7] combines DeepJSCC with semantic extraction and diffusion models for wireless image reconstruction. In this method, the image is encoded into latent channel symbols and transmitted over a wireless channel under a given bandwidth compression ratio (BCR). At the receiver, a diffusion model is employed as a channel denoiser before the latent symbols are decoded by the DeepJSCC decoder, which helps improve the reconstruction quality.

DiffCom [40] further uses a unified encoder and adopts diffusion models as probabilistic generative decoders, enabling more realistic reconstructions from received wireless signals. In addition, latent diffusion models are integrated with autoencoder-based semantic communication frameworks in [41, 42] to mitigate channel-induced degradation in wireless transmission and improve the perceptual quality of reconstructed images.

Besides diffusion-based approaches, other generative and iterative methods have also been explored to enhance the quality of received images. DeepJSCC is extended with GANs in [43], where InverseJSCC and GenerativeJSCC are proposed to improve the perceptual quality of received images under challenging conditions such as low BCR and low SNR. In a similar direction, the study in [44] applies an invertible neural network as guidance for the diffusion model to further enhance degraded DeepJSCC reconstructions. Moreover, the work in [45] improves the quality of DeepJSCC reconstruction with an iterative refinement approach that mitigates the noise of the impaired transmission channel. In [46] a masking vector quantized-variational autoencoder (VQ-VAE-based) semantic communication system is also proposed to improve robustness and reduce transmission overhead.

Overall, these works demonstrate that the channel noise problem can be mitigated and the received image quality can be improved by post-processing the reconstructed image at the receiver side. However, many existing works assume that a perfect channel state information (CSI) is available and integrate it into the denoising task. However, perfect CSI is an idealized assumption and cannot be achieved in practical systems. In addition, because these methods mainly focus on denoising, they do not explicitly enhance the received image resolution, such as upsampling it

under low-quality transmission caused by practical bandwidth limitations or hardware constraints.

3.2 Generative Models for Semantic and Bandwidth Constrained Image Transmission

In practical scenarios, hardware limitations can reduce the quality of transmission, resulting in received images with poor or low visual quality. With recent advances in generative models, a growing body of research has explored how these models can be applied to improve the quality of received images after transmission.

One important research direction is semantic communication, where compact semantic representations are transmitted instead of full pixel-level information to reduce bandwidth consumption. For example, stable diffusion is incorporated with JSCC in [47] for goal-oriented semantic communication. In this approach, compact semantic latent features are transmitted, and diffusion-based denoising is performed at the receiver for image reconstruction.

Similarly, the studies in [48, 49] improve semantic transmission efficiency by exploiting semantic priors and generative models. Instead of focusing on exact pixel recovery, these works aim to preserve and reconstruct semantic information while reducing the amount of transmitted data. In [48], the transmitter extracts and sends a compressed semantic map, such as a segmentation map, which is then used by a Denoising Diffusion Probabilistic Models (DDPM) at the receiver to reconstruct a semantically consistent image. In [49], the transmitter applies semantic change detection to transmit only important semantic updates, while the receiver uses a conditional DDPM with a semantic decoder to reconstruct the remote scene from the received data.

The work in [50] proposes a hybrid wireless image transmission scheme that integrates diffusion-based JSCC with digital transmission. In this work, a U-Net architecture, as described in Section 2.3, is applied to implement the diffusion-based reconstruction module rather than the JSCC encoder and decoder. This approach achieves bandwidth-efficient image delivery and improves the perceptual quality of the received image. The study in [51] also uses a low-resolution coarse image

combined with a compressed semantic segmentation map as a guide for image super-resolution on the receiver side. The receiver reconstructs a high-resolution image from these two inputs using a semantic-conditioned diffusion model with a U-Net backbone.

In [52], the noise from the wireless channel is treated as part of the diffusion forward process, and a compression module is used to reduce the size of the transmitted semantic features. This method improves reconstruction quality over various channels under bandwidth constraints. In a related but different direction, the study in [53] explores Swin Transformer-based methods for remote sensing image super-resolution, as introduced in Section 2.4. This method aims to enhance image resolution while reducing computational complexity through an efficient Swin Transformer architecture.

However, the studies discussed above mainly focus on reducing transmission bandwidth through semantic communication, where semantic representations are transmitted instead of raw pixel-level information. The study in [53] also addresses image quality enhancement via post processing and achieve image super resolution. Nevertheless, these works generally do not consider the impact of wireless channel degradation, such as noise or fading, during transmission. In [52], the authors consider both diffusion-based denoising and bandwidth-constrained transmission. However, their approach is similar to DeepJSCC-based methods, which extract meaningful semantic features and reduce the transmitted data size, as discussed in Section 3.1. Moreover, the method assumes that channel state information is available at the receiver and leverages this information in the denoising process.

Overall, existing studies address different aspects of wireless image transmission, including channel denoising, semantic feature compression, perceptual reconstruction, and image super-resolution. However, many diffusion-based wireless communication methods rely on explicit CSI or assume that the channel condition is known during the reconstruction process. Other methods mainly focus on reducing the transmission bandwidth through semantic compression, while channel degradation and resolution loss are often considered separately. In addition, most receiver-side denoising methods aim to recover the transmitted image at its original resolution and do not jointly

enhance a low-resolution image that is caused by bandwidth constraints or hardware limitations. These limitations motivate our work on developing a conditional latent diffusion framework that directly uses the degraded received image as the conditioning input, without requiring explicit CSI, and jointly performs channel impairment mitigation and image resolution enhancement.

Chapter 4

The Proposed Method

In this chapter, we present the proposed framework for wireless image transmission with joint channel impairment mitigation and image resolution enhancement.

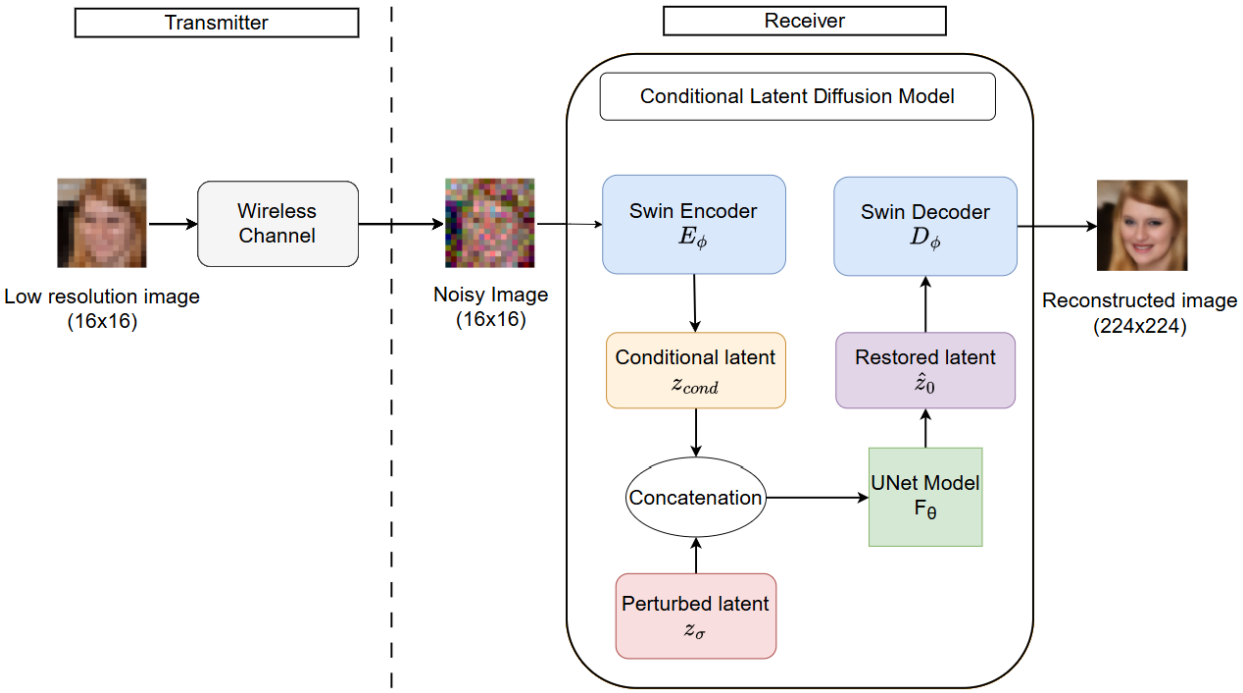


Figure 4.1: Overall architecture of the proposed receiver-side conditional latent diffusion framework. A low-resolution image is transmitted through a wireless channel, encoded into a conditional latent representation, restored by the latent diffusion model, and decoded into a high-resolution reconstructed image.

4.1 Problem Statement

Consider an IoT imaging system that transmits low-resolution images due to low-resolution image sensors or bandwidth constraints. Let $\mathbf{x}_{LR} \in \mathbb{R}^{C \times H_L \times W_L}$ denote the transmitted low-resolution image, where the dimensions correspond to the channel, height, and width respectively. During transmission over a wireless channel, the image is further degraded by channel fading and impairments, resulting in the received observation

$$\mathbf{y} = h\mathbf{x}_{LR} + n, \tag{4.1}$$

where h denotes the effective channel coefficient and n denotes the additive noise term. Given only distorted observation $\mathbf{y}$ without CSI estimation, the objective is to mitigate channel distortion and enhance low resolution image to construct a clean high-resolution image $\mathbf{x}_{HR} \in \mathbb{R}^{C \times H_H \times W_H}$, with $H_H > H_L$ and $W_H > W_L$.

4.2 Overall System Design

The proposed framework adopts a diffusion-based receiver-side image restoration strategy, as shown in Figure 4.1. The model does not require a forward diffusion process derived from an explicit CSI. Instead, it uses a standard latent diffusion formulation and restores the image directly from the degraded received observation. To make diffusion-based restoration more efficient, we operate in the latent space rather than directly on the raw image. The corrupted observation $\mathbf{y}$ is first mapped into a compact latent space by a ST encoder to produce the conditional latent representation

$$z_{cond} = E_\phi(\mathbf{y}) \in \mathbb{R}^{C_z \times H_z \times W_z}, \tag{4.2}$$

where C_z, H_z, and W_z denote the channel and spatial dimensions of the latent space. A conditional latent diffusion model then estimates the clean latent $\hat{z}_0$ from a noisy latent sample z_σ, guided by z_{cond}. Because the model relies only on the received observation, it does not require explicit CSI.

After restoration, the recovered latent is decoded by ST decoder $D_\phi(\hat{z}_0)$ to generate the final high-resolution image. This design enables joint channel impairment mitigation and image enhancement. The learned semantic encoder is not considered at the transmitter because the proposed framework is intended to perform restoration entirely at the receiver while keeping the transmitter computationally simple. Introducing a semantic encoder would require additional computation, memory, and energy. Furthermore, the transmitted image $\mathbf{x}_{\mathrm{LR}}$ has already been at low resolution, which contains substantially less spatial information than the ground truth high-resolution image. Introducing an additional learned compression bottleneck at the transmitter could discard further visual information that may be useful for reconstruction. For these reasons, the low-resolution image is transmitted without an additional semantic encoder, and the proposed method focuses on recovering missing details and mitigating channel distortion at the receiver.

4.3 Conditional Diffusion Model

In this part, we develop a receiver-side conditional latent diffusion framework for CSI-free wireless image restoration and super-resolution. The proposed framework leverages iterative conditioning [31] to progressively refine degraded observations while formulating the restoration process through a continuous-time stochastic differential equation (SDE) perspective [54, 55] and an efficient EDM-based implementation [35]. Specifically, the degraded low-resolution received image serves as the sole conditioning signal, enabling the model to simultaneously suppress channel distortions and reconstruct a high-resolution image without requiring explicit CSI or channel equalization modules. By integrating these components within a unified latent diffusion framework and tailoring them to the wireless image transmission scenario, the proposed approach achieves robust image recovery under various channel conditions. Additional architectural and training strategies inspired by diffusion-based restoration literature are further incorporated to enhance reconstruction performance.

4.3.1 Diffusion Forward Process

Let us first consider the DDPM forward process [27], where Gaussian noise is gradually added to the data over a sequence of discrete timesteps. The forward transition is defined as

$$\mathbf{x}_t = \sqrt{1-\beta_t}\,\mathbf{x}_{t-1} + \sqrt{\beta_t}\,\boldsymbol{\epsilon}_{t-1}, \qquad \boldsymbol{\epsilon}_{t-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{4.3}$$

where β_t denotes the noise variance at timestep $t = 1, \ldots, N$. As the step size $\Delta t = \frac{1}{N}$ approaches zero, the discrete noise schedule above can be represented by a continuous noise rate $\beta(t)$. In this continuous-time limit, the DDPM forward process converges to the variance-preserving SDE [54, 55]

$$d\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}\,dt + \sqrt{\beta(t)}\,d\mathbf{w}, \tag{4.4}$$

where $\mathbf{w}$ denotes a standard Wiener process. The SDE in equation (4.4) is a special case of the general diffusion process

$$d\mathbf{x} = f(\mathbf{x}, t)\,dt + g(t)\,d\mathbf{w}, \tag{4.5}$$

with drift coefficient $f(\mathbf{x}, t) = -\frac{1}{2}\beta(t)\mathbf{x}$ and diffusion coefficient $g(t) = \sqrt{\beta(t)}$. Following the EDM [35], this diffusion perturbation process equation (4.4), (4.5) can be described in a general form

$$q_{0t}(\mathbf{x}(t) \mid \mathbf{x}(0)) = \mathcal{N}\big(\mathbf{x}(t); s(t)\mathbf{x}(0), s(t)^2\sigma(t)^2\mathbf{I}\big)\,, \tag{4.6}$$

where $q_{0t}(\mathbf{x}(t) \mid \mathbf{x}(0))$ denotes the forward perturbation distribution from the clean sample $\mathbf{x}(0)$ to the noisy sample $\mathbf{x}(t)$ at time t and $\mathcal{N}(\cdot)$ represents a Gaussian distribution with mean $s(t)\mathbf{x}(0)$ and covariance $s(t)^2\sigma(t)^2\mathbf{I}$. Here, $s(t)$ determines how the clean signal is scaled over time, while $\sigma(t)$ controls the amount of Gaussian noise added to the data.

In the EDM setting, choosing $s(t) = 1$ removes the time-dependent signal scaling. As a result, the noisy sample can be obtained directly from the clean sample without explicitly simulating all intermediate diffusion steps, the distribution of forward process $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0, \sigma^2\mathbf{I})$ is:

$$\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4.7}$$

4.3.2 Conditional Reverse Process

Given the forward SDE in Equation (4.4), the corresponding probability flow ordinary differential equation (ODE) can be written as [54, 55]

$$d\mathbf{x} = \left[f(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right] dt. \tag{4.8}$$

This ODE defines a deterministic reverse-time trajectory whose marginal distributions match those of the corresponding reverse SDE. For image restoration in [31], the reverse process can be guided by an additional observation $\mathbf{x}_{\text{cond}}$, the DDPM forward process equation (4.3) with the distribution $q(x_t|x_{t-1}) = \mathcal{N}(x_t|\sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ has its corresponding conditional reverse process which is guided by an additional input x_{cond} defined as

$$p_\theta \left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_{\text{cond}} \right) = \mathcal{N} \left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta \left(\mathbf{x}_t, \mathbf{x}_{\text{cond}}, t \right), \sigma_t^2 \mathbf{I} \right),$$

where $\mathbf{x}_{\text{cond}}$ denotes the conditional observation, and $\boldsymbol{\mu}_\theta(\mathbf{x}_{\text{cond}}, \mathbf{x}_t, t)$ is the conditional mean of the inference process. Similarly, the probability flow ODE equation (4.8) can also be conditioned

$$d\mathbf{x} = \left[f(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t \left(\mathbf{x} \mid \mathbf{x}_{\text{cond}} \right) \right] dt. \tag{4.9}$$

Here, the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x} \mid \mathbf{x}_{\text{cond}})$ guides the noisy sample toward regions that are more consistent with the conditional observation. Following the EDM setting [35], the drift and diffusion coefficients from equations (4.5), and (4.6) can be parameterized using the signal scaling function $s(t)$ and the noise schedule $\sigma(t)$

$$f(\mathbf{x}, t) = \frac{\dot{s}(t)}{s(t)} \mathbf{x}, \qquad g(t) = s(t) \sqrt{2 \dot{\sigma}(t) \sigma(t)}, \tag{4.10}$$

where $\dot{s}(t)$ and $\dot{\sigma}(t)$ denote the derivatives of $s(t)$ and $\sigma(t)$, respectively. Substituting equation (4.10) into equation (4.9) and setting $s(t) = 1$ gives the conditional probability flow ODE:

$$d\mathbf{x} = -\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{x}} \log p\left(\mathbf{x} \mid \mathbf{x}_{\text{cond}}; \sigma(t)\right) dt, \tag{4.11}$$

where $\nabla_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{x}_{\text{cond}}; \sigma(t))$ is the score function that guides the sample to high-probability regions of the data distribution that are consistent with conditional observation. However, following EDM [35], we do not train the neural network to predict the score directly. Instead, the network is trained as a denoiser that predicts the clean sample from its noisy observation. This formulation is more stable in practice because the network output remains in the data space, while the corresponding score can still be recovered analytically from the denoiser output under Gaussian perturbation. Under the forward perturbation in equation (4.7), the score is approximated as

$$\nabla_{\mathbf{x}} \log p\left(\mathbf{x} \mid \mathbf{x}_{\text{cond}}; \sigma\right) \approx \frac{D_\theta\left(\mathbf{x}, \mathbf{x}_{\text{cond}}; \sigma\right) - \mathbf{x}}{\sigma^2}, \tag{4.12}$$

where D_θ denotes the conditional denoising network, and σ denotes the noise level. The corresponding training objective of our diffusion model is to minimize the following loss function

$$\mathcal{L}(\theta) = \mathbb{E}_{x, x_{\text{cond}}, \sigma, \epsilon}\left[\lambda(\sigma) \left\| D_\theta(\mathbf{x}_\sigma, \mathbf{x}_{\text{cond}}; \sigma) - x \right\|_2^2\right], \tag{4.13}$$

where $\mathbf{x}_\sigma$ is computed from equation (4.7), and $\lambda(\sigma)$ is the loss weighting function.

4.4 Diffusion Model Preconditioning

Directly parameterizing D_θ with a neural network can lead to unstable training, since the input and target magnitudes vary significantly with the noise level σ. Following EDM [35], we therefore express the denoiser using a preconditioned neural network F_θ as

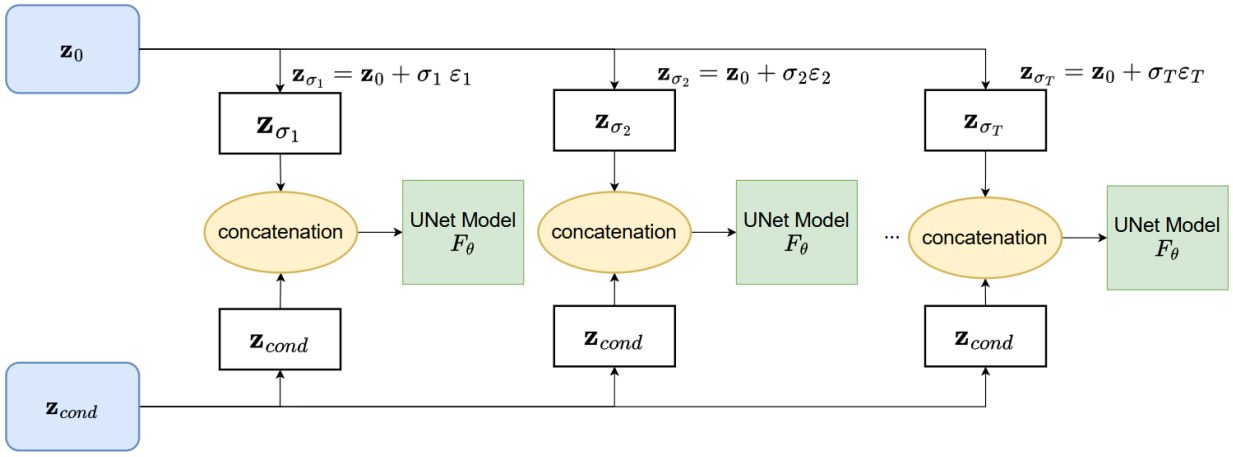


Figure 4.2: Visualization of the proposed forward perturbation process in latent space. The clean high-resolution latent representation $z_0 = E_\phi(x_{\text{HR}})$ is corrupted by Gaussian noise at different noise levels σ_i to generate the noisy latent $z_{\sigma_i} = z_0 + \sigma_i\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$.

$$D_\theta(\mathbf{x}, \mathbf{x}_{\text{cond}}; \sigma) = c_{\text{skip}}(\sigma)\,\mathbf{x}_\sigma + c_{\text{out}}(\sigma)\,F_\theta\Big(c_{\text{in}}(\sigma)\,[\mathbf{x}_\sigma, \mathbf{x}_{\text{cond}}], c_{\text{noise}}(\sigma)\Big) \tag{4.14}$$

where F_θ denotes the SongUNet architecture [55] and $[\mathbf{x}_\sigma, \mathbf{x}_{\text{cond}}]$ denotes channel-wise concatenation following Figure 4.2. For $\mathbf{x}_\sigma, \mathbf{x}_{\text{cond}} \in \mathbb{R}^{C_z \times H_z \times W_z}$, this operation keeps the spatial dimensions $H_z \times W_z$ unchanged and doubles the channel dimension C_z.

The coefficients are defined as

$$\begin{aligned} c_{\text{skip}}(\sigma) &= \frac{\sigma_{\text{data}}^2}{\sigma^2 + \sigma_{\text{data}}^2}, & c_{\text{out}}(\sigma) &= \frac{\sigma\,\sigma_{\text{data}}}{\sqrt{\sigma^2 + \sigma_{\text{data}}^2}}, \\ c_{\text{in}}(\sigma) &= \frac{1}{\sqrt{\sigma^2 + \sigma_{\text{data}}^2}}, & c_{\text{noise}}(\sigma) &= \frac{1}{4}\ln\sigma, \\ \lambda(\sigma) &= \frac{\sigma^2 + \sigma_{\text{data}}^2}{(\sigma\,\sigma_{\text{data}})^2}, & \ln\sigma &\sim \mathcal{N}\big(P_{\text{mean}},\, P_{\text{std}}^2\big)\,, \end{aligned} \tag{4.15}$$

where σ_{data} denotes the empirical standard deviation of the clean latent z_0 measured over the training set. The parameters P_{mean} and P_{std} control the log-normal inference distribution of the diffusion noise level σ, and their values are specified in Chapter 5. The coefficients $c_{\text{skip}}(\sigma)$, $c_{\text{in}}(\sigma)$, $c_{\text{out}}(\sigma)$, and $c_{\text{noise}}(\sigma)$ are used to precondition the denoising network. Specifically, $c_{\text{skip}}(\sigma)$ controls the skip connection from the noisy input, $c_{\text{in}}(\sigma)$ and $c_{\text{out}}(\sigma)$ scale the network input and output magnitudes,

respectively, and $c_{\text{noise}}(\sigma)$ maps the noise level into a conditioning embedding. This preconditioning keeps the network input and training target well scaled across different noise levels, improving training stability.

4.5 Swin Transformer Autoencoder

Directly performing diffusion in the image domain is computationally expensive, especially for high-resolution image reconstruction. Therefore, we employ a ST autoencoder to map images into a lower-dimensional latent representation before applying the diffusion model. Unlike autoencoder used in [32], our ST autoencoder is adopted as a deterministic representation module whose purpose is to preserve structural and visual information for image restoration, rather than to generate a probabilistic latent prior for image generation.

The ST autoencoder consists of an encoder $E_\phi(\cdot)$ and a decoder $D_\phi(\cdot)$. Given a clean high-resolution image $\mathbf{x}_{\text{HR}}$, the encoder maps it into a latent representation

$$\mathbf{z}_0 = E_\phi(\mathbf{x}_{\text{HR}}), \tag{4.16}$$

where $\mathbf{z}_0$ denotes the clean latent used as the target of the diffusion model. Similarly, the degraded received observation $\mathbf{y}$ is encoded as

$$\mathbf{z}_{\text{cond}} = E_\phi(\mathbf{y}), \tag{4.17}$$

where $\mathbf{z}_{\text{cond}}$ serves as the conditional guidance for the latent diffusion model.

The ST autoencoder and the conditional latent diffusion model are trained separately. In the first stage, the ST autoencoder is trained using clean images, with the objective of reconstructing the original high-resolution image from its latent representation. After convergence, the parameters of the autoencoder are frozen. In the second stage, the frozen encoder is used to extract both the clean latent $\mathbf{z}_0$ and the conditional latent $\mathbf{z}_{\text{cond}}$, and the diffusion model is trained to recover $\mathbf{z}_0$

Algorithm 1 Training the Swin Autoencoder

Input: Clean dataset $q(\mathbf{x})$, encoder E_ϕ, decoder D_ϕ, learning rate η
Output: Trained encoder E_ϕ and decoder D_ϕ
repeat
  $\mathbf{x}_{\text{HR}} \sim q(\mathbf{x})$
  $\hat{\mathbf{x}}_{\text{HR}} \leftarrow D_\phi(E_\phi(\mathbf{x}_{\text{HR}}))$
  $\mathcal{L}_{\text{AE}}(\phi) = \frac{1}{N} \|\hat{\mathbf{x}}_{\text{HR}} - \mathbf{x}_{\text{HR}}\|_2^2$
  $\phi \leftarrow \phi - \eta \nabla_\phi \mathcal{L}_{\text{AE}}(\phi)$
until convergence

from its Gaussian-perturbed version conditioned on $\mathbf{z}_{\text{cond}}$.

This design separates representation learning from diffusion-based restoration. The autoencoder acts as a deterministic dimensionality reduction module rather than a semantic transmission model. By operating in the latent space, the diffusion model performs channel impairment mitigation and resolution enhancement with lower computational cost than pixel-space diffusion.

The ST autoencoder is trained by minimizing the mean squared error (MSE) reconstruction loss:

$$\mathcal{L}_{\text{AE}}(\phi) = \mathbb{E}_{\mathbf{x}_{\text{HR}}} \left[\frac{1}{N} \|D_\phi(E_\phi(\mathbf{x}_{\text{HR}})) - \mathbf{x}_{\text{HR}}\|_2^2 \right], \tag{4.18}$$

where N denotes the total number of elements in each reconstructed image, and $E_\phi(\cdot)$ and $D_\phi(\cdot)$ denote the ST encoder and decoder, respectively. The ST autoencoder training procedure is summarized in Algorithm 1.

4.6 Conditional Latent Diffusion Model

Using the frozen ST encoder, the clean image and the degraded received observation are first mapped into the latent representations defined in equations (4.16) and (4.17). Therefore, the EDM forward perturbation in equation (4.7) is applied to the clean latent representation rather than the image-space sample. The conditional denoising network in equation (4.14) is then trained to recover the clean latent $\mathbf{z}_0$ from the noisy latent $\mathbf{z}_\sigma$, guided by the conditional latent $\mathbf{z}_{\text{cond}}$. Accordingly, the general training objective in equation (4.13) becomes the following latent diffusion

Algorithm 2 Conditional Latent Diffusion Model Training Procedure

Input: dataset $\mathcal{X}$, encoder $E_\phi(\cdot)$
for each mini-batch $\{\mathbf{x}_i\}$ **do**
  $\mathbf{z}_0 \leftarrow E_\phi(\mathbf{x_{HR}})$
  $\gamma \sim \mathcal{U}(0, 20)$ dB
  $\mathbf{x}_{\text{LR}} \leftarrow \text{Down}(\mathbf{x_{HR}})$
  $\mathbf{y} \leftarrow \mathcal{H}(\mathbf{x}_{\text{LR}}, \gamma)$
  $\mathbf{z}_{\text{cond}} \leftarrow E_\phi(\mathbf{y})$
  $\sigma \leftarrow \exp(\mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2))$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$
  $\mathbf{z}_\sigma \leftarrow \mathbf{z}_0 + \sigma\boldsymbol{\epsilon}$
  $L(\theta) \leftarrow \lambda(\sigma)\|D_\theta(\mathbf{z}_\sigma, \mathbf{z}_{\text{cond}}; \sigma) - \mathbf{z}_0\|_2^2$
  $\theta \leftarrow \theta - \eta\nabla_\theta L$
end for

objective:

$$\mathcal{L}_{\text{LDM}}(\theta) = \mathbb{E}_{\mathbf{z}_0, \mathbf{z}_{\text{cond}}, \sigma, \epsilon}\left[\lambda(\sigma)\left\|D_\theta(\mathbf{z}_\sigma, \mathbf{z}_{\text{cond}}; \sigma) - \mathbf{z}_0\right\|_2^2\right], \tag{4.19}$$

where $\mathbf{z}_0$ and $\mathbf{z}_{\text{cond}}$ are obtained from the frozen ST encoder as defined in equations (4.16) and (4.17), and $\mathbf{z}_\sigma$ is generated by the EDM forward perturbation in equation (4.7). The training procedure of the proposed DM is summarized in Algorithm 2. To improve robustness under varying channel conditions, the model is trained using degraded observations generated from a range of channel qualities. Specifically, during training, the received image is simulated as $\mathbf{y} = \mathcal{H}(\mathbf{x}_{\text{LR}}, \gamma)$, where $\mathcal{H}(\cdot, \gamma)$ denotes a certain wireless channel with SNR γ uniformly sampled from a broad range and $\text{Down}(\cdot)$ denotes a spatial resolution reduction operator used to obtain the low-resolution image from the dataset. As a result, the conditional latent $\mathbf{z}_{\text{cond}}$ captures channel distortions of different severities, enabling the diffusion model to learn a restoration mapping that generalizes across a broad range of noise and fading conditions.

4.6.1 Inference Algorithm (Second-Order Heun Sampler)

The inference procedure of the proposed conditional latent diffusion model is summarized in Algorithm 3. During inference, image reconstruction is performed by integrating the EDM probability flow ODE [35]. Starting from an initial Gaussian latent variable, the model progressively

Algorithm 3 Conditional Latent Diffusion Model Inference Procedure

Input: degraded latent z_{cond}, noise schedule $\{\sigma_i\}_{i=0}^{N}$
Sample initial latent $z_N \sim \mathcal{N}(0, \sigma_0^2 I)$
for $i = N, N-1, \ldots, 1$ **do**
  $\hat{\mathbf{D}}_i \leftarrow \mathbf{D}_\theta(z_i, \sigma_i, z_{\text{cond}})$
  $\mathbf{d}_i \leftarrow (z_i - \hat{\mathbf{D}}_i)/\sigma_i$
  $\tilde{z}_{i-1} \leftarrow z_i + (\sigma_{i-1} - \sigma_i)\mathbf{d}_i$
  if $i > 1$ **then**
    $\hat{\mathbf{D}}_{i-1} \leftarrow \mathbf{D}_\theta(\tilde{z}_{i-1}, \sigma_{i-1}, z_{\text{cond}})$
    $\mathbf{d}'_{i-1} \leftarrow (\tilde{z}_{i-1} - \hat{\mathbf{D}}_{i-1})/\sigma_{i-1}$
    $z_{i-1} \leftarrow z_i + (\sigma_{i-1} - \sigma_i)\left(\frac{1}{2}\mathbf{d}_i + \frac{1}{2}\mathbf{d}'_{i-1}\right)$
  else
    $z_{i-1} \leftarrow \tilde{z}_{i-1}$
  end if
end for
$\hat{x} \leftarrow \mathbf{D}_\phi(z_0)$

removes noise and refines the latent representation toward the clean latent estimate. Unlike training, where the noise level σ is randomly sampled from a log-normal distribution, the inference stage uses a fixed discrete noise schedule $\{\sigma_i\}_{i=0}^{N}$. The conditional latent representation $\mathbf{z}_{\text{cond}}$ is kept fixed during the reverse process and guides the denoising trajectory toward a reconstruction that is consistent with the degraded received observation. The noise schedule is defined as

$$\sigma_i = \left(\sigma_{\text{max}}^{1/\rho} + \frac{i}{N-1}\left(\sigma_{\text{min}}^{1/\rho} - \sigma_{\text{max}}^{1/\rho}\right)\right)^{\rho},$$

where $i = 0, \ldots, N-1$. The parameter ρ controls the distribution of discretization steps between large $\sigma_0 = \sigma_{max}$ and small noise levels $\sigma_{N-1} = \sigma_{min}$. This schedule allocates more steps near low-noise regions, where reconstruction details are more sensitive to numerical errors.

At each step, the denoiser first predicts a clean latent estimate $\hat{\mathbf{D}}_i$ from the current noisy latent $\mathbf{z}_i$. This prediction is used to compute the ODE direction $\mathbf{d}_i$, which points from the current latent toward the estimated clean latent.

Euler predictor

$$\hat{\mathbf{D}}_i = \mathbf{D}_\theta(z_i, z_{\text{cond}}; \sigma_i)\,, \qquad \mathbf{d}_i = \frac{z_i - \hat{\mathbf{D}}_i}{\sigma_i},$$

$$\tilde{z}_{i-1} = z_i + (\sigma_{i-1} - \sigma_i)\mathbf{d}_i.$$

The Euler predictor gives a first-order estimate $\tilde{\mathbf{z}}_{i-1}$ at the next noise level. To reduce discretization error, Heun's method re-evaluates the denoising direction at this predicted point and averages the two directions.

Heun corrector For $\sigma_{i-1} \neq 0$, the corrected update is given by

$$\hat{\mathbf{D}}_{i-1} = \mathbf{D}_\theta(\tilde{z}_{i-1}, z_{\text{cond}}; \sigma_{i-1})\,,$$

$$\mathbf{d}'_{i-1} = \frac{\tilde{z}_{i-1} - \hat{\mathbf{D}}_{i-1}}{\sigma_{i-1}},$$

$$z_{i-1} = z_i + (\sigma_{i-1} - \sigma_i)\left(\frac{1}{2}\mathbf{d}_i + \frac{1}{2}\mathbf{d}'_{i-1}\right).$$

When $\sigma_{i-1} = 0$, the correction step is omitted to avoid division by zero, and the Euler prediction is used directly. The final latent $\tilde{z}_0$ is decoded by the ST decoder to obtain the reconstructed image.

Chapter 5

Experiments

5.1 Evaluation Metrics

Five image quality metrics are used to quantitatively evaluate the reconstruction quality of the proposed method. They are peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), learned perceptual image patch similarity (LPIPS), mean squared error (MSE), and mean absolute error (MAE). These metrics evaluate different aspects of the reconstructed image, including pixel-level fidelity, structural preservation, perceptual similarity, and reconstruction error.

Mean Absolute Error (MAE) is a pixel-wise evaluation metric used to measure the average absolute difference between a reconstructed image and its reference image. Given a reference image I and a reconstructed image K, both of size $m \times n$, MAE is defined as

$$\text{MAE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |I(i,j) - K(i,j)| .$$

Mean squared error (MSE) is a basic pixel-wise evaluation metric used to measure the difference between a reconstructed image and its reference image. It calculates the average squared difference between corresponding pixel values. Given a reference image I and a reconstructed

image K, both of size $m \times n$, MSE is defined as

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 .$$

Peak signal-to-noise ratio (PSNR) is defined via MSE

$$\text{PSNR} = 10 \log_{10} \left(\frac{MAX^2}{\text{MSE}} \right),$$

where MAX is the maximum possible pixel value of the image. The higher the PSNR score, the closer the reconstructed image is to the ground-truth image.

Structural similarity index measure (SSIM) [56] is used to measure the structural similarity between the ground-truth image and reconstructed image. SSIM is better than PSNR in measuring luminance, contrast, and local structural information. It is defined as

$$\text{SSIM}(x, \hat{x}) = \frac{(2\mu_x \mu_{\hat{x}} + C_1)(2\sigma_{x\hat{x}} + C_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)}, \tag{5.1}$$

where x is the ground-truth image, $\hat{x}$ is the reconstructed image, μ_x and $\mu_{\hat{x}}$ are the mean sample, σ_x^2 and $\sigma_{\hat{x}}^2$ are the variances, and $\sigma_{x\hat{x}}$ is the covariance between the two images. The higher SSIM score, the better preservation of image structure.

Learned perceptual image patch similarity (LPIPS) [57] is used to measure the similarity between images using deep features which are extracted from a pretrained deep neural network. Unlike other measurements that compare pixel-wise differences, LPIPS compares images in a learned feature space that is similar to human visual perception. The lower the LPIPS score, the more perceptually similar the reconstructed image is to the ground-truth image.

5.2 Dataset

We evaluate the proposed framework using two image datasets: Flickr-Faces-HQ (FFHQ) [58] and Diverse 2K Resolution High Quality Images (DIV2K) [59]. DIV2K contains high quality images with diverse scenes, objects, textures, and structures. The dataset consists of 1,000 2K-resolution images, which are split into 800 training images, 100 validation images, and 100 test images. In our experiments, DIV2K images are randomly cropped and resized to (224×224) for training and testing.

FFHQ is used for human face image reconstruction experiments. The original FFHQ images have a resolution of (512×512) and contain high-quality face images with diverse poses, appearances, and lighting conditions. For training and evaluation, all images are resized to (224×224). The FFHQ dataset is divided into 41,600 training images, 5,200 validation images, and 5,200 test images.

To simulate bandwidth-constrained wireless image transmission, each HR image is first down sampled to a LR image using bicubic interpolation. The transmitted low resolution image has spatial resolution 16×16 and 64×64. The low resolution image is then transmitted through either an additive white Gaussian noise (AWGN) channel or a Rayleigh fading channel to generate the degraded received observation. All input images are converted to tensors and normalized to the range [-1, 1] using channel-wise mean and standard deviation values of (0.5,0.5,0.5).

5.3 Implementation Details

The proposed framework is trained in two stages. In the first stage, the ST autoencoder is trained on clean high-resolution images using the MSE reconstruction loss. The autoencoder follows the Swin-Unet design [60], but is simplified to two encoder-decoder stages with the long skip connections shown in Figure 2.3 removed. This prevents information from bypassing the latent bottleneck and forces the decoder to reconstruct the image solely from the compressed latent representation. This design is particularly suitable for diffusion-based inference, where only the

restored latent $\hat{z}_0$ is provided to the decoder, while the corresponding clean encoder features are unavailable.

The initial embedding dimension is set to 12. Both the encoder and decoder use two Swin Transformer stages with layer depths $[6, 2]$, corresponding to 6 and 2 ST blocks, respectively. Patch merging is applied between the encoder stages to reduce the spatial resolution and increase the feature dimension, while patch expansion is used in the decoder to recover the image resolution. Window-based self-attention with window size 7 is used in each Swin Transformer block. Between the encoder and decoder, a convolutional latent projection module is used to refine the latent feature. This module consists of convolutional layers with group normalization and Sigmoid Linear Unit (SiLU) activation. The final decoder output is produced by a $4\times$ patch expansion layer followed by a 1×1 convolution to reconstruct the RGB image.

After the autoencoder converges, both the encoder and decoder are frozen. In the second stage, the conditional latent diffusion model is trained in the latent space. Both the Swin Transformer autoencoder and the latent diffusion model are optimized using AdamW. For the latent diffusion model, the learning rate is initially set to 1×10^{-4} for the first 1000 epochs and then reduced to 5×10^{-5} for an additional 2000 epochs. The batch size is set to 128. During diffusion training, the channel SNR is uniformly sampled from the range of 0 dB to 20 dB for both AWGN and Rayleigh fading channels, i.e., $\gamma \sim \mathcal{U}(0, 1, \ldots, 20)$ dB. The diffusion noise level follows the log-normal distribution, i.e., $\ln \sigma \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$, where $P_{\text{mean}} = -4.6$ and $P_{\text{std}} = 1.2$, this P_{mean} and P_{std} may vary in different datasets. We choose P_{mean} so that the center of the noise distribution is consistent with the empirical scale of the clean latent representation $\mathbf{z}_0$. In our experiments, $\sigma_{\text{data}} \approx 0.01$. Following latent diffusion scaling in [32, 61], the output of the Swin Transformer encoder is scaled by a constant factor of 0.18215 to stabilize diffusion training.

During inference, we use the EDM second-order Heun sampler with $\sigma_{\text{min}} = 0.0005$, $\sigma_{\text{max}} = 0.9$, $\rho = 7$, and $N = 17$ sampling steps. All models are trained on an NVIDIA H100 GPU.

5.4 Benchmark Methods

To evaluate the effectiveness of the proposed method, we compare it with two baseline schemes. The first baseline is the corresponding reconstruction model without diffusion restoration. The second baseline is a conventional JPEG2000 based scheme, where its compression and coding system is applied to the low-resolution source image at the transmitter. Our proposed method and the baseline are both evaluated under the same channel conditions. In addition to the comparison with conventional reconstruction baselines, we also compare the proposed method with JSCC+CDDM [8], which is a diffusion-based wireless semantic communication method designed for channel denoising. JSCC+CDDM is selected as a representative diffusion-based baseline because it also applies a diffusion model at the receiver to mitigate channel distortion. However, the comparison should be interpreted as a reference comparison rather than a strictly identical experimental setting, since JSCC+CDDM uses explicit channel state information (CSI), including the estimated Rayleigh fading coefficient, during receiver-side denoising. In contrast, our proposed model performs restoration using only the degraded received image as the conditional input and does not require explicit CSI.

5.5 Experimental Results

We evaluate the proposed conditional latent diffusion model under both Rayleigh fading and AWGN channels using PSNR and SSIM as reconstruction metrics.

Method	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**20 dB**
Our Model, 64×64	19.453	20.331	20.805	20.955	21.044
JSCC+CDDM, 64×64	20.646	23.766	25.810	26.867	27.320

Table 5.1: PSNR comparison between the proposed model and JSCC+CDDM under Rayleigh channel for different input resolutions and SNR values on DIV2K dataset.

5.5.1 Quantitative Results

Tables 5.1 and 5.2 compare the proposed model with JSCC+CDDM under the Rayleigh fading channel at different SNR values and input resolutions on the DIV2K dataset. For the (64×64) input setting, JSCC+CDDM [8] achieves higher PSNR and SSIM than the proposed method across all SNR values. However, our proposed model still obtains comparable reconstruction performance while using significantly fewer diffusion sampling steps, requiring only 17 steps compared with 100 steps in JSCC+CDDM. In addition, JSCC+CDDM uses explicit channel state information (CSI), including the Rayleigh fading coefficient, to equalize the received signal before diffusion-based denoising. In contrast, the proposed model does not require explicit CSI and performs restoration using only the degraded received image as conditional guidance. This highlights the efficiency and practicality of the proposed framework under channel-impaired wireless image transmission.

Figure 5.1 and Figure 5.2 present the quantitative performance under different SNR conditions. In both channel settings, the proposed diffusion model consistently outperforms the corresponding baseline without diffusion restoration. The gain is particularly evident at low and moderate SNR values, where the received observations suffer from more severe distortion. As the SNR increases, the reconstruction quality of all methods improves, while the proposed model maintains a clear advantage in both PSNR and SSIM. These results indicate that the latent diffusion restoration process effectively mitigates channel degradation while enhancing the quality of reconstructed images.

Figure 5.3 compares the proposed method with a JPEG2000 baseline under the AWGN and Rayleigh fading channels. The proposed model achieves higher PSNR across the evaluated SNR

Method	**0 dB**	**5 dB**	**10 dB**	**15 dB**	**20 dB**
Our Model, 64×64	0.5916	0.6161	0.6305	0.6375	0.6388
JSCC+CDDM, 64×64	0.6584	0.6584	0.7150	0.7355	0.7408

Table 5.2: SSIM comparison between the proposed model and JSCC+CDDM under Rayleigh channel for different input resolutions and SNR values on DIV2K dataset.

range.

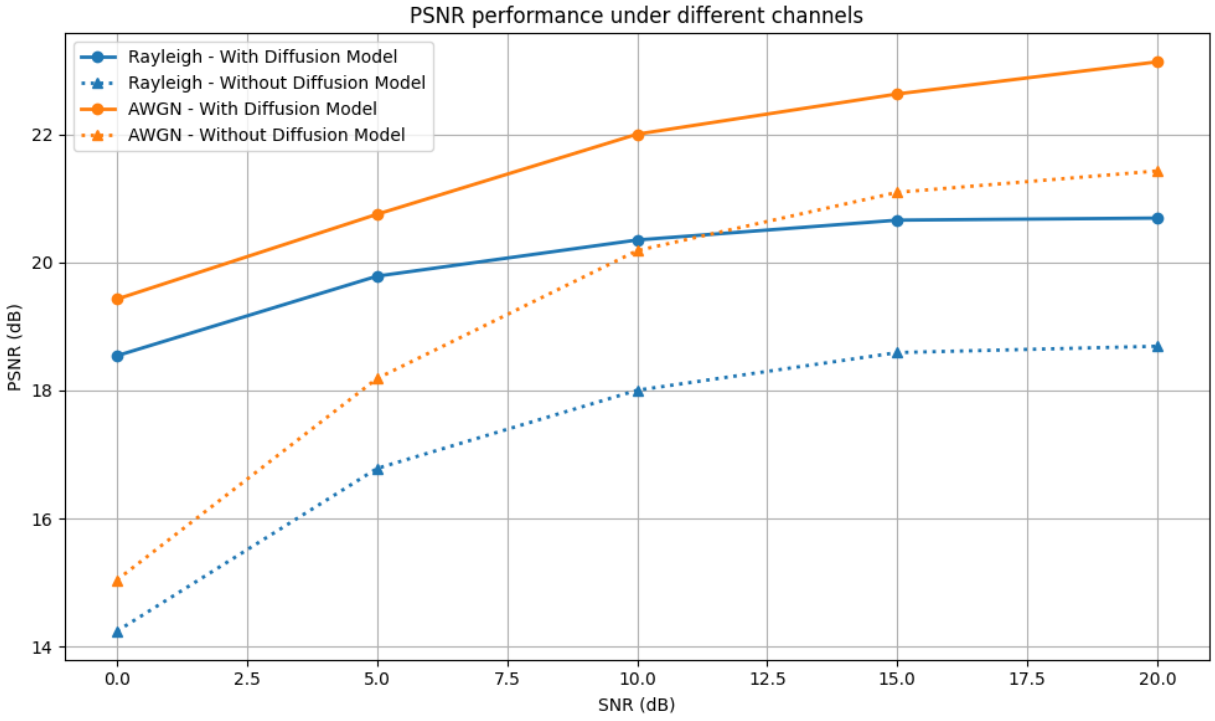


Figure 5.1: PSNR comparison between the proposed diffusion model and the corresponding non-diffusion reconstruction baseline under AWGN and Rayleigh fading channels at different SNR levels.

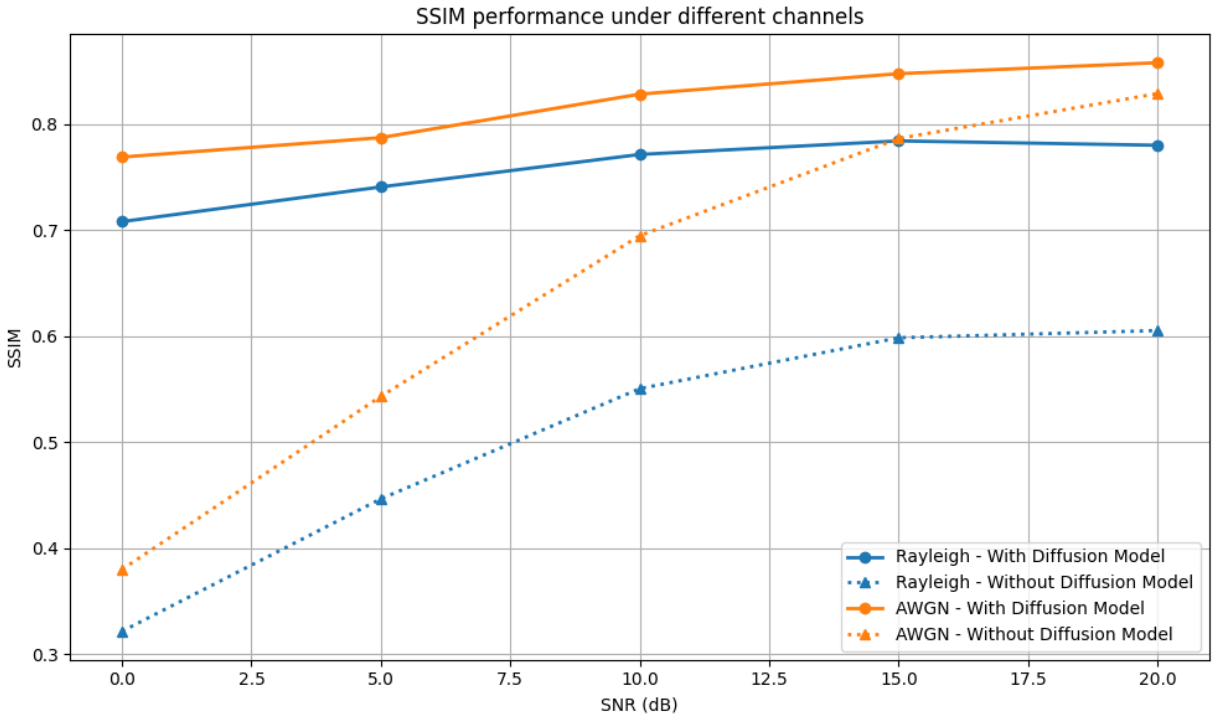


Figure 5.2: SSIM comparison between the proposed diffusion model and the corresponding non-diffusion reconstruction baseline under AWGN and Rayleigh fading channels at different SNR levels.

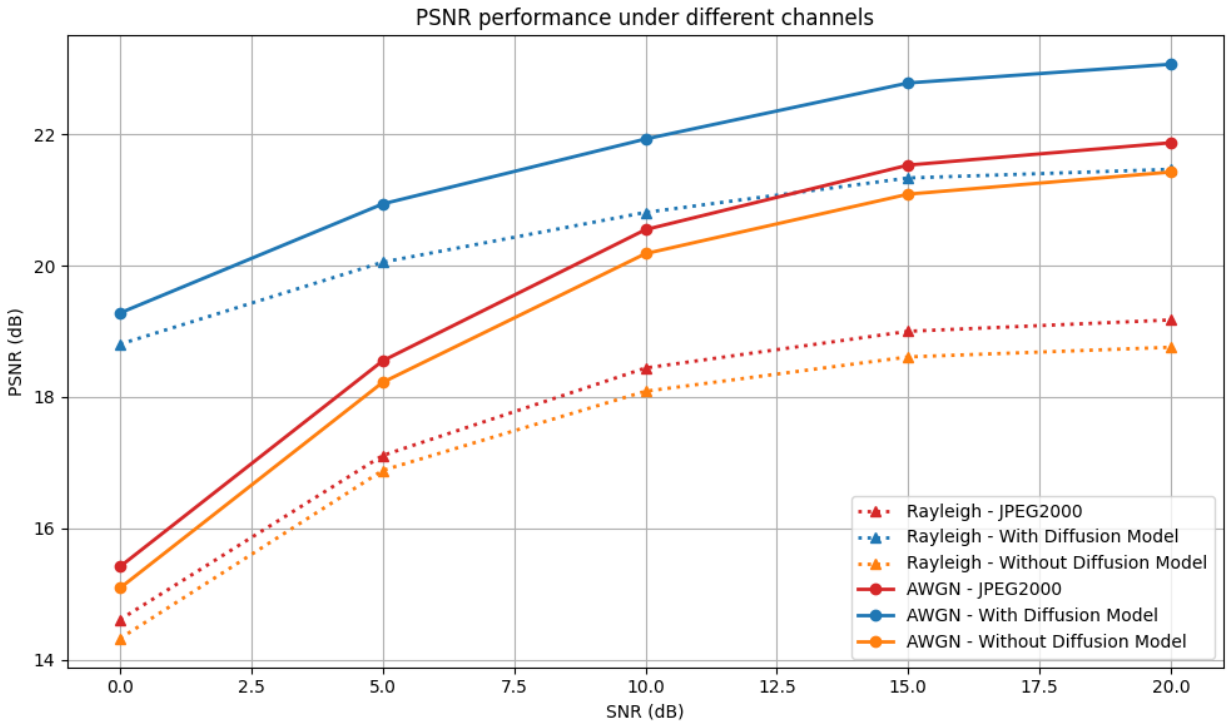


Figure 5.3: PSNR comparison among the proposed diffusion model, the non-diffusion reconstruction baseline, and the JPEG2000-based baseline under AWGN and Rayleigh fading channels at different SNR levels.

Table 5.3 and Table 5.4 report the general reconstruction performance of our model on the FFHQ dataset under the Rayleigh fading channel with 16×16 and 64×64 transmitted images. As the SNR increases, the proposed model achieves higher PSNR and SSIM, while the reconstruction errors measured by MSE and MAE decrease. The LPIPS values also remain relatively low across the evaluated SNR range, indicating that the reconstructed images preserve perceptual similarity to the ground truth. These results show that the proposed latent diffusion framework can improve both pixel-level fidelity and perceptual reconstruction quality under severe low-resolution wireless transmission.

SNR (dB)	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**MSE↓**	**MAE↓**
0	17.073	0.658	0.396	0.0809	0.2061
5	18.279	0.708	0.365	0.0621	0.1722
10	19.473	0.748	0.336	0.0475	0.1504
15	20.009	0.764	0.322	0.0424	0.1418
20	20.261	0.779	0.328	0.0404	0.1378

Table 5.3: General reconstruction performance of our model on the FFHQ dataset under the Rayleigh fading channel with 16×16 transmitted images.

SNR (dB)	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**MSE↓**	**MAE↓**
0	22.409	0.819	0.255	0.024	0.106
5	24.142	0.847	0.221	0.054	0.087
10	25.120	0.865	0.203	0.036	0.077
15	25.744	0.868	0.196	0.031	0.071
20	25.841	0.869	0.193	0.029	0.070

Table 5.4: General reconstruction performance of our model on the FFHQ dataset under the Rayleigh fading channel with 64×64 transmitted images.

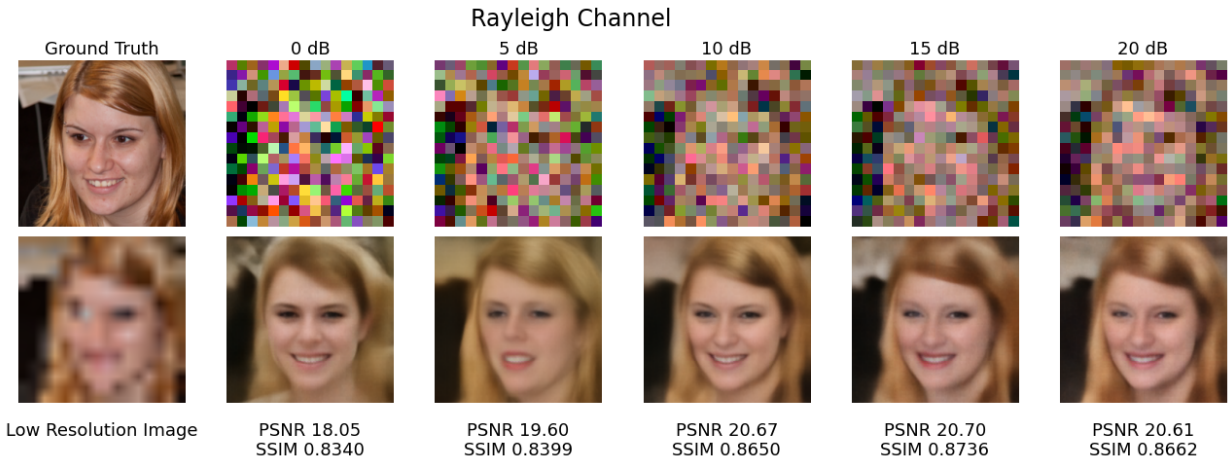


Figure 5.4: Visual reconstruction results under the Rayleigh fading channel at different SNR levels on 16×16 image. The first row shows the degraded low-resolution received images, and the second row shows the corresponding high-resolution reconstructions produced by the proposed model.

5.5.2 Qualitative Results

Representative examples of joint channel distortion mitigation and image resolution enhancement under Rayleigh fading and AWGN channels with different SNR levels are illustrated in Figure 5.4 and Figure 5.5, respectively. The first row shows the degraded LR observations after transmission over the channel under different SNR levels on 16×16 image, while the second row

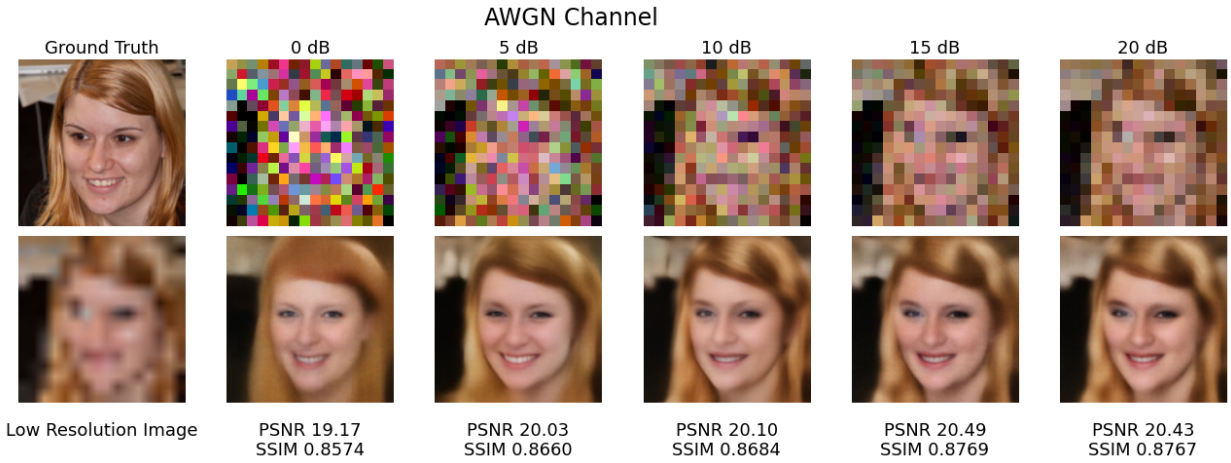


Figure 5.5: Visual reconstruction results under the AWGN channel at different SNR levels on 16×16 image. The first row shows the degraded low-resolution received images, and the second row shows the corresponding high-resolution reconstructions produced by the proposed model.

shows the corresponding reconstructed images. The ground truth and the LR image (without channel impairments) are included in the first column for reference. The results clearly demonstrate that the proposed method simultaneously improves image resolution and effectively mitigates channel impairments. As the SNR increases, facial structure and texture become progressively clearer, and even under severe channel degradation, the model is able to reconstruct a recognizable high-resolution image.

Figure 5.6, Figure 5.7 and Figure 5.8 show additional qualitative reconstruction results on the FFHQ and DIV2K datasets, respectively. Under both AWGN and Rayleigh fading channels, the received observations are severely degraded due to the combined effects of low-resolution trans-

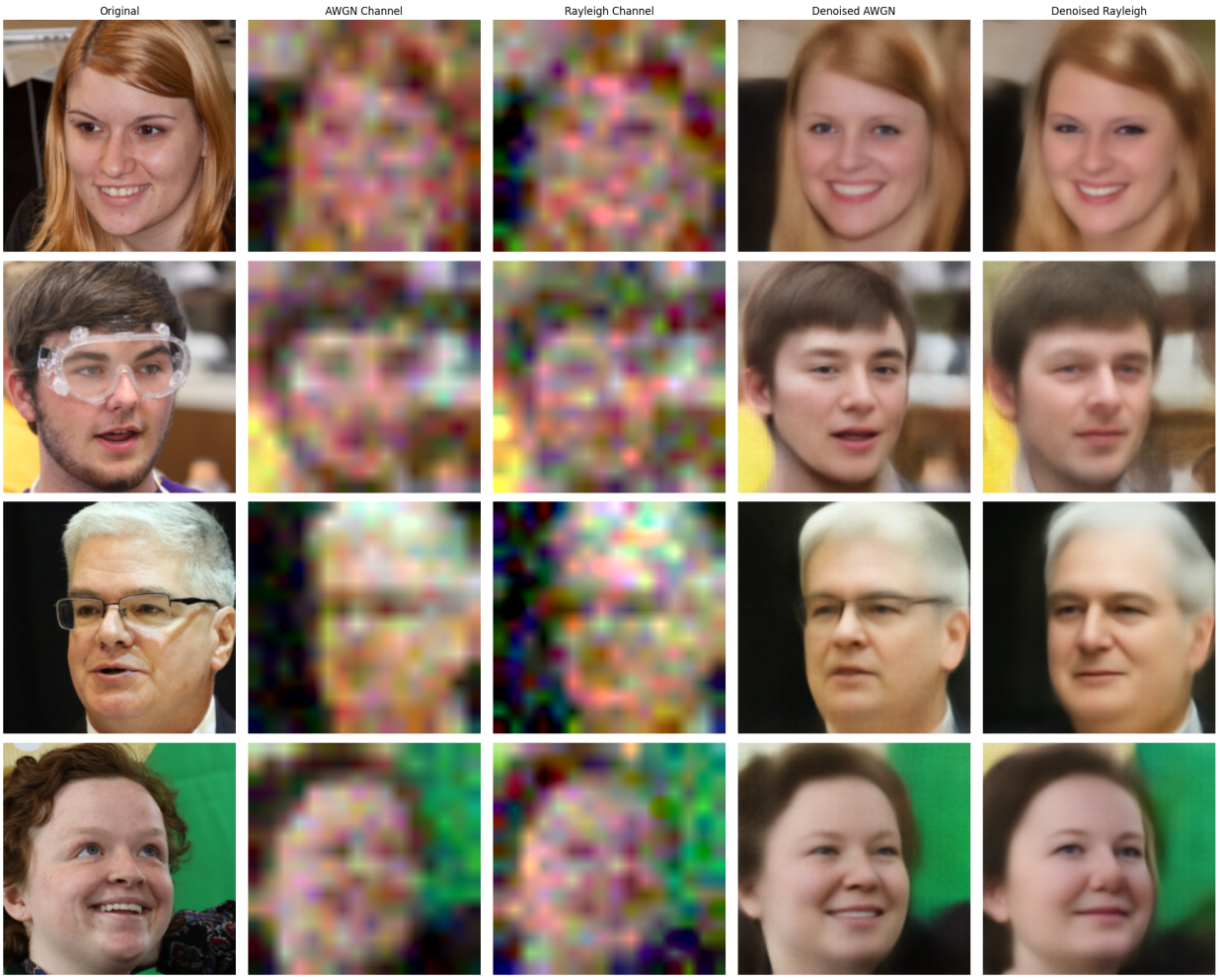


Figure 5.6: Reconstruction results on the FFHQ dataset 16×16 under AWGN and Rayleigh fading channels.

mission and channel distortion. In all three figures, the columns from left to right show the original high-resolution image, degraded AWGN observation, degraded Rayleigh observation, reconstruction from the AWGN-trained diffusion model, and reconstruction from the Rayleigh-trained diffusion model. On FFHQ, the proposed latent diffusion model reconstructs clearer facial structures and produces images with more coherent global appearance compared with the degraded observations. On DIV2K, the reconstruction task is more challenging because the dataset contains more diverse natural scenes with complex textures, edges, and object structures. Although some high-frequency details are not fully recovered, the proposed model reduces visible channel distortion and preserves the main image structures. These qualitative results demonstrate the ability of the



Figure 5.7: Reconstruction results on the FFHQ dataset 64×64 under AWGN and Rayleigh fading channels.

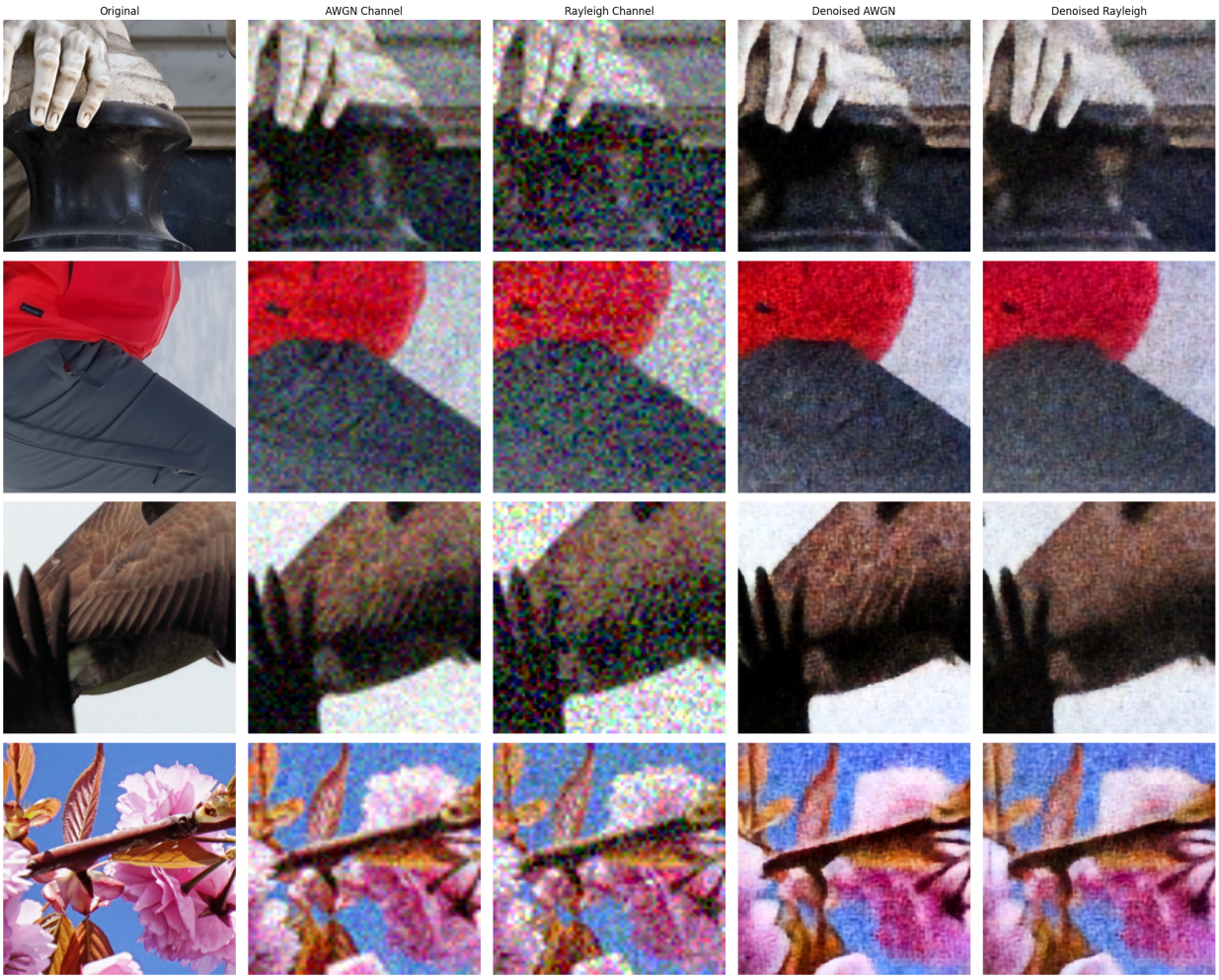


Figure 5.8: Reconstruction results on the DIV2K dataset 64×64 under AWGN and Rayleigh fading channels.

proposed framework to jointly mitigate channel distortion and enhance image resolution, while also indicating that more diverse training data and stronger latent representations can further improve general-image reconstruction quality.

5.6 Ablation study

To examine the robustness of the proposed framework under channel mismatch, Table 5.5 reports the cross-channel generalization performance when the model is trained and tested on different channel types. When trained and evaluated on the same channel, the model achieves its

Train Channel	PSNR (dB)			SSIM		
	AWGN	Rayleigh	Rician	AWGN	Rayleigh	Rician
AWGN	**22.0478**	19.6737	19.2402	**0.8031**	0.7245	0.7118
Rayleigh	22.0971	**21.515**	20.8573	0.8087	**0.7790**	0.7608

Table 5.5: Cross-channel generalization performance of the proposed model.

best performance for the AWGN setting. However, when the AWGN-trained model is tested on the Rayleigh fading channel, the performance drops noticeably. This result indicates that a model trained only under AWGN does not generalize well to the more severe and structurally different distortions introduced by Rayleigh fading. By contrast, the Rayleigh-trained model shows better cross-channel robustness. When evaluated on the AWGN channel, it achieves performance that is slightly higher than that obtained under its matched-channel condition. This observation suggests that a model trained under Rayleigh fading may learn a more robust restoration behavior that partially transfers to the AWGN case, and the AWGN channel is easier to handle in terms of channel impairment mitigation than Rayleigh fading channel. Nevertheless, the overall results confirm that channel mismatch still affects reconstruction quality and that a single model may not be sufficient to handle all channel conditions effectively.

To evaluate the role of the conditional latent $\mathbf{z}_{\text{cond}}$, we progressively corrupt the conditioning signal by interpolating the true conditional latent with a shuffled latent from another image $\tilde{\mathbf{z}}_{\text{cond}} = (1-\eta)\,\mathbf{z}_{\text{cond}} + \eta\, z_{\text{cond}}^{\text{shuffle}}$, where $\eta \in [0,1]$ controls the mismatch level, and demonstrate the results in Figure 5.9. When $\eta = 0$, the correct conditional latent is used, whereas larger values of η correspond to increasingly mismatched conditional guidance. As the mismatch level η increases, the reconstruction performance decreases, showing that the reverse diffusion process is sensitive to the consistency of the conditional guidance. In particular, severe mismatch with $\eta = 0.7$ results in a large performance drop because the model is guided by latent features from a completely mismatched image. At low SNR, severe channel distortion results in low PSNR for the received-image-only baseline. This severe distortion also makes $\mathbf{z}_{\text{cond}}$ less informative and less reliable as conditional guidance. Therefore, the incorrect image features introduced by a mildly mismatched latent may have a weaker influence on the reconstruction, allowing the conditional

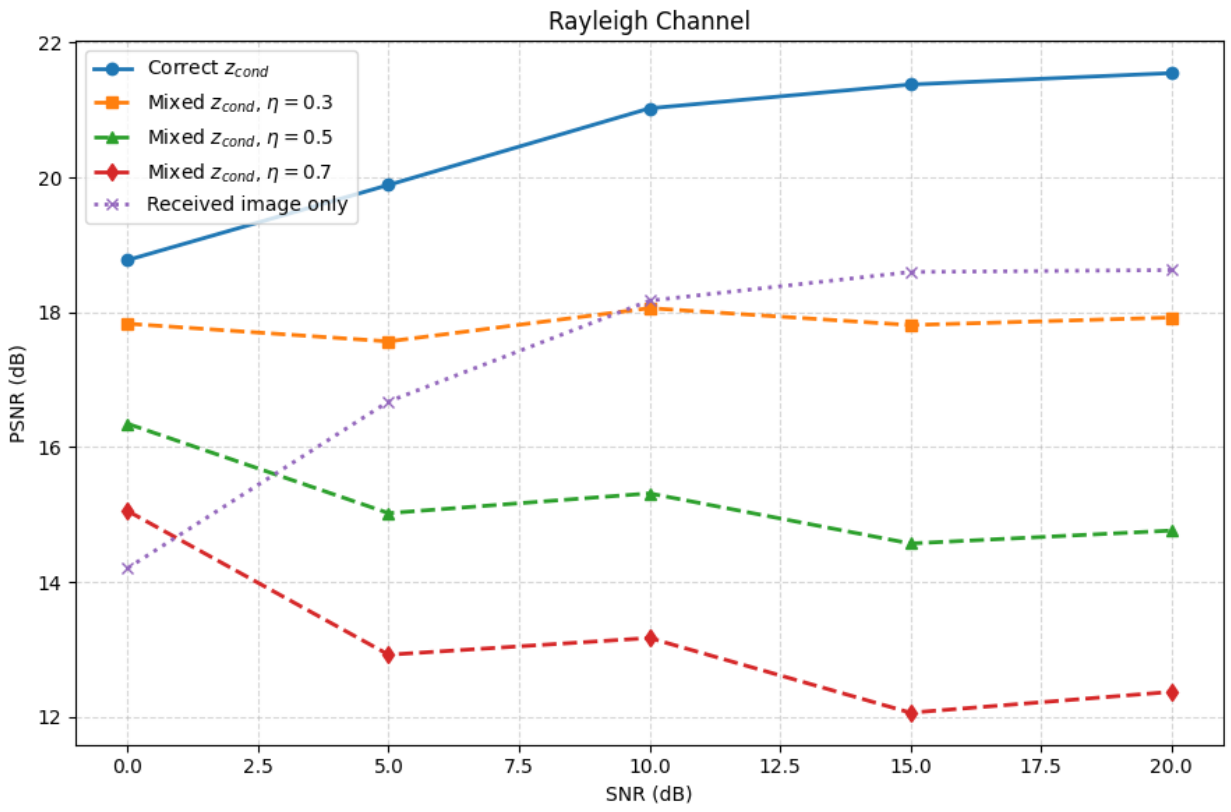


Figure 5.9: Effect of conditional latent mismatch on PSNR under the Rayleigh fading channel. The conditioning latent is progressively corrupted by interpolating the correct conditional latent with a shuffled latent from another image, where larger η indicates stronger conditional mismatch.

latent diffusion model to rely more on its learned image prior. For that reason, some mildly mismatched cases can still outperform the received-image-only baseline at low SNR. As the SNR increases, however, the shuffled latent components contain more reliable image-specific features. These features provide stronger but incorrect semantic guidance, directing the diffusion process toward a reconstruction that is inconsistent with the transmitted image. Consequently, strongly mismatched conditions can perform worse than the received-image-only baseline. This result indicates that using the received image without diffusion-based restoration may be preferable when the available conditional guidance is substantially mismatched.

Chapter 6

Conclusion

6.1 Summary

In this thesis, we proposed a conditional latent diffusion framework for wireless image transmission that jointly performs channel impairment mitigation and image resolution enhancement from distorted low-resolution received images. The model does not require explicit CSI to guide the denoising process. Instead, the model learns to recover the clean high-resolution image by conditioning on the degraded received observation in latent space. To achieve this, we integrated a ST autoencoder with an EDM-based diffusion model, where the autoencoder provides a compact latent representation and the diffusion model restores the clean latent from its noisy counterpart under conditional guidance. This design enables efficient restoration while avoiding the need to define a channel-specific forward diffusion process. Experimental results under both AWGN and Rayleigh fading channels demonstrate that the proposed method improves reconstruction quality over CSI-free reconstruction baselines across a broad range of SNR conditions. The performance is evaluated using multiple reconstruction metrics, including PSNR, SSIM, LPIPS, MSE, and MAE, showing improvements in both pixel-level fidelity and perceptual quality. Additional qualitative results on FFHQ and DIV2K further show that the proposed framework can recover meaningful high-resolution structures from severely degraded low-resolution observations, while

also highlighting that reconstruction on diverse natural images remains more challenging than on face images.

6.2 Future Work

As shown in the ablation study, the performance of the proposed model may degrade when deployed in practical channel environments that differ from those encountered during training. In real wireless transmission scenarios, the channel may not belong to only one fixed type, but can be affected by different kinds of noise, fading, and interference at the same time. Therefore, a model trained on a single channel condition may not always generalize well when the testing channel is changed or mixed with other types of degradation. Future work should further investigate this issue and improve the robustness of the model under more realistic and mixed channel environments.

Another important direction is the role of the conditional latent representation $\mathbf{z}_{\text{cond}}$. From the condition mismatch experiment, the quality of $\mathbf{z}_{\text{cond}}$ has a strong effect on the final reconstruction performance. When the conditional latent is mismatched or corrupted, the diffusion model may receive misleading guidance, which can reduce the reconstruction quality. In future work, channel characteristics such as SNR, fading type, or estimated channel information could be incorporated into the conditional guidance. Another possible direction is to integrate channel-related information into the forward diffusion process, so that the diffusion model can better match the actual degradation caused by the wireless channel. This may help the model achieve more stable performance under practical transmission conditions.

Bibliography

[1] N. Van Huynh, J. Wang, H. Du, D. T. Hoang, D. Niyato, D. N. Nguyen, D. I. Kim, and K. B. Letaief, "Generative ai for physical layer communications: A survey," *IEEE Transactions on Cognitive Communications and Networking*, vol. 10, no. 3, pp. 706–728, 2024.

[2] D. Fan, R. Meng, X. Xu, Y. Liu, G. Nan, C. Feng, S. Han, S. Gao, B. Xu, D. Niyato, T. Q. S. Quek, and P. Zhang, "Generative diffusion models for wireless networks: Fundamental, architecture, and state-of-the-art," *IEEE Communications Surveys Tutorials*, vol. 28, pp. 5632–5677, 2026.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[5] N. C. Luong, N. D. Hai, D. V. Le, H. T. Nguyen, T.-H. Vu, T. Huynh-The, R. Zhang, N. D. D. Anh, D. Niyato, M. D. Renzo, D. I. Kim, and Q.-V. Pham, "Diffusion models for future networks and communications: A comprehensive survey," *arXiv preprint arXiv:2508.01586*, 2025.

[6] D. Fan, R. Meng, X. Xu, Y. Liu, G. Nan, C. Feng, S. Han, S. Gao, B. Xu, D. Niyato, T. Q. S. Quek, and P. Zhang, "Generative diffusion models for wireless networks: Fundamental, architecture, and state-of-the-art," *IEEE Communications Surveys & Tutorials*, vol. 28, pp. 5632–5677, 2026.

[7] M. Zhang, H. Wu, G. Zhu, R. Jin, X. Chen, and D. Gündüz, "Semantics-guided diffusion for deep joint source-channel coding in wireless image transmission," *IEEE Transactions on Wireless Communications*, vol. 25, pp. 1547–1564, 2025.

[8] T. Wu, Z. Chen, D. He, L. Qian, Y. Xu, and M. Tao, "Cddm: Channel denoising diffusion models for wireless semantic communications," *IEEE Transactions on Wireless Communications*, vol. 23, no. 9, pp. 11168–11183, 2024.

[9] Z. Tan, C. Liu, K. Xie, Y. Ouyang, J. Wen, G. Zhang, D. Chen, G. Xie, and K. Li, "Adaptive semantic communication system for high-quality remote sensing image transmission in unstable wireless environments," *IEEE Transactions on Wireless Communications*, vol. 25, pp. 3001–3015, 2026.

[10] C.-H. Lee, J.-W. Lin, P.-H. Chen, and Y.-C. Chang, "Deep learning-constructed joint transmission-recognition for internet of things," *IEEE Access*, vol. 7, pp. 76547–76561, 2019.

[11] Y. Gong, Z. Chu, Z. Zhu, P. Xiao, M. Zeng, Y. Wang, H. Pandey, and J. Hou, "Wireless vision-centered semantic communication for smart city environment: Pretrained network and quantization," *IEEE Transactions on Consumer Electronics*, vol. 72, no. 1, pp. 2383–2398, 2026.

[12] B. Sklar, "Rayleigh fading channels in mobile digital communication systems. i. characterization," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 136–146, 1997.

[13] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2021.

[14] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), vol. 8692 of *Lecture Notes in Computer Science*, (Cham), pp. 184–199, Springer, 2014.

[15] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017.

[17] J. Su, B. Xu, and H. Yin, "A survey of deep learning approaches to image restoration," *Neurocomputing*, vol. 487, pp. 46–65, May 2022.

[18] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[19] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic mr image reconstruction," in *IEEE Transactions on Medical Imaging*, vol. 37, pp. 491–503, IEEE, 2018.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), pp. 234–241, Springer International Publishing, 2015.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, October 2021.

[26] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 2256–2265, PMLR, 07–09 Jul 2015.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020.

[28] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[29] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 21696–21707, Curran Associates, Inc., 2021.

[30] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2022.

[31] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2023.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[33] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 8780–8794, Curran Associates, Inc., 2021.

[34] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171, PMLR, 18–24 Jul 2021.

[35] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Advances in Neural Information Processing Systems*, 2022.

[36] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[37] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2315–2328, 2022.

[38] S. F. Yilmaz, X. Niu, B. Bai, W. Han, L. Deng, and D. Gündüz, "High perceptual quality wireless image delivery with denoising diffusion models," in *IEEE Conference on Computer Communications Workshops*, pp. 1–5, 2024.

[39] P. Yang, G. Zhang, and Y. Cai, "Rate-adaptive generative semantic communication using conditional diffusion models," *IEEE Wireless Communications Letters*, vol. 14, no. 2, pp. 539–543, 2025.

[40] S. Wang, J. Dai, K. Tan, X. Qin, K. Niu, and P. Zhang, "Diffcom: Channel received signal is a natural condition to guide diffusion posterior sampling," *IEEE Journal on Selected Areas in Communications*, vol. 43, no. 7, pp. 2651–2666, 2025.

[41] B. Xu, R. Meng, Y. Chen, X. Xu, C. Dong, and H. Sun, "Latent semantic diffusion-based channel adaptive de-noising semcom for future 6g systems," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, pp. 1229–1234, 2023.

[42] B. Xu, S. Han, X. Xu, W. Li, R. Meng, C. Dong, and P. Zhang, "Semantic prior aided channel-adaptive equalizing and de-noising semantic communication system with latent diffusion model," *IEEE Transactions on Wireless Communications*, vol. 24, no. 6, pp. 4614–4630, 2025.

[43] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 8, pp. 2645–2657, 2023.

[44] J. Chen, D. You, D. Gündüz, and P. L. Dragotti, "CommIN: Semantic image communications as an inverse problem with INN-guided diffusion models," in *Proceedings of the IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 13436–13440, 2024.

[45] C. Lee, X. Hu, and H.-S. Kim, "Deep joint source-channel coding with iterative source error correction," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* (F. Ruiz, J. Dy, and J.-W. van de Meent, eds.), vol. 206 of *Proceedings of Machine Learning Research*, pp. 3879–3902, PMLR, 25–27 Apr 2023.

[46] Q. Hu, G. Zhang, Z. Qin, Y. Cai, G. Yu, and G. Y. Li, "Robust semantic communications with masked vq-vae enabled codebook," *IEEE Transactions on Wireless Communications*, vol. 22, no. 12, pp. 8707–8722, 2023.

[47] N. Li and Y. Deng, "Goal-oriented semantic communication for wireless image transmission via stable diffusion," in *ICC 2025 – IEEE International Conference on Communications*, pp. 3748–3753, 2025.

[48] E. Grassucci, S. Barbarossa, and D. Comminiello, "Generative semantic communication: Diffusion models beyond bit recovery," *IEEE Transactions on Cognitive Communications and Networking*, vol. 12, pp. 8171–8185, 2026.

[49] W. Yang, Z. Xiong, H. Du, Y. Yuan, and T. Q. Quek, "Semantic change driven generative semantic communication framework," in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 01–06, 2024.

[50] X. Niu, X. Wang, D. Gündüz, B. Bai, W. Chen, and G. Zhou, "A hybrid wireless image transmission scheme with diffusion," in *2023 IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 86–90, 2023.

[51] F. Pezone, O. Musa, G. Caire, and S. Barbarossa, "Semantic-preserving image coding based on conditional diffusion models," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13501–13505, 2024.

[52] L. Guo, W. Chen, Y. Sun, B. Ai, N. Pappas, and T. Q. S. Quek, "Diffusion-driven semantic communication for generative models with bandwidth constraints," *IEEE Transactions on Wireless Communications*, vol. 24, no. 8, pp. 6490–6503, 2025.

[53] X. Kang, P. Duan, J. Li, and S. Li, "Efficient swin transformer for remote sensing image super-resolution," *IEEE Transactions on Image Processing*, vol. 33, pp. 6367–6379, 2024.

[54] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[55] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations (ICLR)*, 2021.

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

[57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

[58] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

[59] R. Timofte, S. Gu, J. Wu, L. Van Gool, L. Zhang, M.-H. Yang, M. Haris, *et al.*, "Ntire 2018 challenge on single image super-resolution: Methods and results," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[60] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, “Swin-unet: Unet-like pure transformer for medical image segmentation,” in *ECCV 2022 Workshops*, pp. 205–218, 2023.

[61] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, vol. 1, pp. 4172–4182, 2023.