

# RadioDiff: An Effective Generative Diffusion Model for Sampling-Free Dynamic Radio Map Construction

Xiucheng Wang<sup>✉</sup>, *Student Member, IEEE*, Keda Tao<sup>✉</sup>, Nan Cheng<sup>✉</sup>, *Senior Member, IEEE*,  
Zhisheng Yin<sup>✉</sup>, *Member, IEEE*, Zan Li<sup>✉</sup>, *Senior Member, IEEE*, Yuan Zhang<sup>✉</sup>, *Member, IEEE*,  
Xuemin (Sherman) Shen<sup>✉</sup>, *Fellow, IEEE*

**Abstract**—Radio map (RM) is a promising technology that can obtain pathloss based on only location, which is significant for 6G network applications to reduce the communication costs for pathloss estimation. However, the construction of RM in traditional is either computationally intensive or depends on costly sampling-based pathloss measurements. Although the neural network (NN)-based method can efficiently construct the RM without sampling, its performance is still suboptimal. This is primarily due to the misalignment between the generative characteristics of the RM construction problem and the discrimination modeling exploited by existing NN-based methods. Thus, to enhance RM construction performance, in this paper, the sampling-free RM construction is modeled as a conditional generative problem, where a denoised diffusion-based method, named RadioDiff, is proposed to achieve high-quality RM construction. In addition, to enhance the diffusion model's capability of extracting features from dynamic environments, an attention U-Net with an adaptive fast Fourier transform module is employed as the backbone network to improve the dynamic environmental features extracting capability. Meanwhile, the decoupled diffusion model is utilized to further enhance the construction performance of RMs. Moreover, a comprehensive theoretical analysis of why the RM construction is a generative problem is provided for the first time, from both perspectives of data features and NN training methods. Experimental results show that the proposed RadioDiff achieves state-of-the-art performance in all three metrics of accuracy, structural similarity, and peak signal-to-noise ratio. The code is available at <https://github.com/UNIC-Lab/RadioDiff>.

**Index Terms**—radio map, denoise diffusion model, generative problem, wireless network.

## I. INTRODUCTION

In wireless networks, pathloss quantifies the attenuation of signal strength between a pair of sender and receiver caused by free-space propagation loss and interactions of radio waves

with obstacles [1]–[3], which is critical to be measured for wireless resource management [4]–[6]. Traditionally, the pathloss measurement usually depends on pilot transmission and signal processing [7]. However, the dramatic increase of network nodes and antennas has led to the challenge of estimating high-dimensional channels [8], resulting in significant costs in training and feedback overhead, as well as signal processing complexity [9]. This issue further deteriorates in high-mobility scenarios with short channel coherence times for low-latency applications [10], [11]. Meanwhile, the upcoming 6G networks will introduce a large variety of node types, including passive equipment such as intelligent reflective surfaces (IRS) [12], which cannot actively transmit pilots or engage in digital signal processing to measure the pathloss.

The emergence of these new scenarios makes it necessary to efficiently obtain pathloss using easy-to-be-obtained information without pilot transmission and signal processing. Consequently, radio map (RM) technology has been developed, by which the pathloss can be acquired just through location information [13]. Traditional RM construction methods can be categorized into two types: (1) sampling-based approaches that rely on sampling position measurements (SPM) within the RM region, which are then used for interpolation or solving specific least squares problems to construct the RM [14]–[16], and (2) sampling-free methods which are achieved through environmental 3D modeling and electromagnetic ray tracing (ERT) [17]. However, both methods face their own inherent challenges. The sampling-based methods require SPM of the RM construction area, with too few or inaccurate measurements leading to poor construction quality, while a large number of high-precision measurements significantly increase RM construction costs. Moreover, such methods cannot be used to construct RM in never-to-reach regions, limiting its applicability, such as UAV trajectory plan. On the other hand, the ERT-based method, while avoiding measurement costs, is burdened with high computational complexity and struggles to achieve RM construction within acceptable timeframes. Furthermore, both methods are restricted to the construction of static RM (SRM) due to their construction principles, which do not account for the real-time transformation of elements affecting pathloss. Consequently, changes in pathloss resulting from factors such as vehicle motion or alterations in reflection diameters are not reflected in the RM. The sampling-based method requires multiple sampling points to be measured across a wide area, making it impractical to arrange sufficient measurement equipment for simultaneous pathloss collection.

This work was supported by the National Key Research and Development Program of China (2020YFB1807700), and the National Natural Science Foundation of China (NSFC) under Grant No. 62201432.

Xiucheng Wang, Keda Tao, Nan Cheng, Zhisheng Yin, and Zan Li are with the State Key Laboratory of ISN and School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: xcwang\_1@stu.xidian.edu.cn; KD.TAO.CC@outlook.com; dr.nan.cheng@ieee.org; {zsyin, zanli}@xidian.edu.cn). *Nan Cheng is the corresponding author.*

Yuan Zhang is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: zy\_loye@126.com).

Xuemin (Sherman) Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

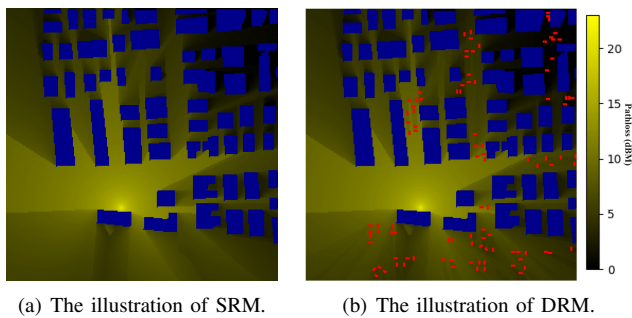


Fig. 1: Illustration of the RM, where the **yellow** elements represent the heatmap of pathloss; the brighter the yellow, the higher the pathloss. The **red** elements denote vehicles, and the **blue** elements are static buildings. Since, static buildings can completely block electromagnetic signals from entering their interiors, resulting in an internal pathloss of zero, to intuitively represent the impact of static buildings on the RM, we have colored them blue.

Additionally, the time-series measurement of different sampling points/sets leads to data collection at different times, rendering real-time RM construction unattainable [14], [15]. Similarly, the ERT-based method, which fundamentally relies on ray tracing of a static 3D scene, typically involves calculations that take several minutes or even longer, rendering it unsuitable for RM construction with dynamic environmental features [17].

To address these challenges, numerous researchers have exploited neural networks (NNs) with rapid inference capabilities to facilitate RM construction. The most notable pioneering work is the RadioUNet that leverages the U-Net, which is a classical architecture for image-to-image tasks, for RM construction [18]. Although some NN-based methods have demonstrated better performance than traditional SPM-based methods in RM construction, their performance, especially in details construction for the dynamic RM (DRM) with multiple dynamic obstacles<sup>1</sup>, is poor. Therefore, the majority of NN-based methods focus on RM construction in static environments, neglecting the exploration of RM construction in dynamic environments [18]–[21]. This is primarily attributed to two challenges: the heterogeneous propagation characteristics of electromagnetic waves and the complexity of RM texture features. The RM construction in static environments typically only concentrates on the influence of buildings on electromagnetic ray propagation, since it can be assumed that all objects affect electromagnetic rays in the same manner. In the context of DRM construction, the impact of moving objects, such as vehicles, on electromagnetic rays must be taken into account. As shown in Fig. 1, the influence of vehicles on electromagnetic ray propagation markedly differs from that of buildings. Different from static obstacles with large shapes and high heights, which can completely block direct electromagnetic rays from the base station (BS) to their surfaces, dynamic obstacles such as vehicles, due to their low height and small size, cannot entirely obstruct the electromagnetic

signal. This partial blockage results in a reduction of the pathloss rather than a complete obstruction, thereby increasing the complexity of RM construction and the texture features than SRM. Moreover, existing NN-based methods mainly train the NN through supervised learning in a discriminative style, whose objective is to minimize the mean square error (MSE) between the predicted RM and the ground truth. However, it has been demonstrated in [22], [23] that while the widely used MSE metric in discriminative training methods enhances the convergence speed of NN training, it degrades the NN's ability to capture the elaborately detailed features of the data, resulting in blurred edges. As shown in Fig. 1, RM (especially DRM) typically contains more detailed texture features. Neural networks trained with MSE-based supervised learning methods struggle to extract these detailed features.

The fundamental reason for the above challenges is the misalignment between the generative problem attributes of the RM construction and the discriminant methodology. RM construction shows the characteristics of generative problems in both data features and NN training methods, while the existing sample-free NN-based methods use discriminant methods to construct RM, which inevitably limits the construction performance of RM. Specifically, in terms of data features, neither the value nor location of the elements of the pathloss to be predicted exist in the input environmental data, thus NN is required to generate pathloss from the raw data. Moreover, since the elements in the RM are not discrete values, it is almost impossible to construct the RM by dividing a finite number of hyperplanes to classify the elements in the environmental data as a discriminative problem [24], [25]. Therefore, the use of discriminant training methods to generate multiple hyperplanes in the latent space using NN to predict pathloss will inevitably lead to poor performance. Furthermore, the self-supervised training method that uses data which is partially masked as input to train NN to predict masked information is mainly used in the training of generative models [26]. In the context of RM construction, the data representing environmental information, can be regarded as the RM whose pathloss is masked, and the NN-based RM construction methods use these masked RM data to recover the pathloss, which is a self-supervision training method. Based on the above analysis, RM construction is a generative problem, so from the perspective of the alignment of the problem and the methodology, a generative method should be used to effectively construct the RM. Moreover, the generative training methods enable NN training both using generative loss and MSE loss, where the NN's capacity to extract intricate texture features of RM can be enhanced. Generative adversarial networks (GANs) have been extensively studied in RM reconstruction; however, their potential is often hindered by extreme instability during training. In contrast, the diffusion model offers a novel solution to the RM reconstruction problem. Geographic maps and RM contain abundant high-frequency sharp edges. Directly feeding these into NN-based models can result in erratic predictions, noise, and artifacts. Diffusion models, however, demonstrate superior capability in capturing this edge information and predicting conditions. Therefore, to achieve high-performance RM construction by taking advantage of the generative NN training method, in this

<sup>1</sup>It is important to note that in this paper, we construct the RM based on a snapshot of the environment, without considering the Doppler effect caused by vehicle motion on the pathloss. Consequently, the main difference between DRM and SRM lies in the snapshot time interval and whether the influence of small-scale obstacles on the propagation of electromagnetic rays is considered.

paper, the RM construction problem is modeled as a conditional generative problem, where a generative diffusion-based method, named RadioDiff, is proposed to construct the RM effectively<sup>2</sup>. The main contributions of this paper are as follows.

- 1) For the first time, the sampling-free RM construction problem is modeled as a conditional generative problem, where the location of the BS, and the environment features are used as prompts for conditional generation. In addition, theoretically, a detailed analysis of why the RM construction is a generative problem from the perspective of data features and NN training methods is provided for the first time.
- 2) To the best of our knowledge, the diffusion-based generative model is first used for effective RM construction. Moreover, to enhance the performance and the inferencing efficiency of the diffusion model, a decoupled diffusion model is used in this paper.
- 3) To enhance the diffusion model's capability to extract dynamic environmental features, the prompts of the static and dynamic environmental features are represented by two matrices, respectively. Additionally, the adaptive fast Fourier transform module is employed to enhance the diffusion model's ability to extract high-frequency information resulting from dynamic environmental features in the data.
- 4) The experimental results demonstrate that the proposed RadioDiff achieves state-of-the-art (SOTA) RM construction performance in all three metrics of accuracy, structural similarity, and peak signal-to-noise ratio.

The remainder of this paper is organized as follows. We first overview the related works of RM construction and give preliminaries of the diffusion model in section-II, then the RM construction problem is formulated and analyzed in section-III. In section-IV the details of the proposed RadioDiff are introduced, while in section-V the experimental results are given. The section-VI concludes our work. The notations are shown in Table I.

## II. PRELIMINARIES AND RELATED WORK

### A. Radio Map Construction

The construction of RM can be categorized into two primary types: sampling-based and sampling-free. The sampling-based methods primarily utilize SPM to obtain pathloss in specific areas for interpolation. Although these methods are independent of knowing environmental details and BS location, they require pathloss measurement in the targeted regions for RM construction. Among these methods, the K-nearest neighbors technique obtains RM data for other locations by weighted averaging the pathloss values of the K nearest sparse measurement points [14]. Additionally, local multinomial regression is commonly employed in sampling-based RM construction. This approach determines the pathloss at a current point by solving a least squares problem that involves the pathloss values of nearby

<sup>2</sup>In this paper, the terms “diffusion model” and “denoise diffusion model” are used interchangeably, both referring to methods for generating the required data from noise input. The diffusion process involves diffusing the data into noise, while the denoising process reconstructs the data from this noisy input.

TABLE I: Notation Table

Variables	Definition
$\mathbf{x}_t$	Noise data after $t$ times of diffusion.
$\mathbf{z}_t$	The feature map of $\mathbf{x}_t$ .
$\hat{\mathbf{z}}_t$	The feature map predicted by NN.
$\mu_{\theta}(\cdot)$	The denoise NN with parameters $\theta$ .
$\mathbf{w}$	The parameters of AFT.
$\mathcal{E}$	VAE encoder.
$\mathcal{D}$	VAE decoder.
$q(\cdot)$	The probability of forward diffusion process.
$p(\cdot)$	The probability of inverse denoising process.
$\phi(\cdot)$	Flattened operator.
$\nu(\cdot)$	Trainable projection function.
$\mathcal{F}(\cdot)$	The FFT operation.
$\mathcal{F}^{-1}(\cdot)$	The inverse FFT operation.
$\mathbf{P}$	The ground truth of RM.
$\hat{\mathbf{P}}$	The RM predicted by NN.
$\mathbf{H}_s$	The static obstacle distribution matrix.
$\mathbf{H}_d$	The dynamic obstacle distribution matrix.
$\mathbf{R}$	The location of BS.
$\mathbf{C}$	The prompt of the diffusion model.
$\mathbf{I}$	The identity matrix.
$\mathcal{N}(\cdot, \cdot)$	The Gaussian stochastic distribution.
$\mathbb{R}^N$	$N$ -dimensional real space.
$\epsilon$	Gaussian stochastic variable as added noise.
$T$	Maximum number of adding noises.
$\alpha, \beta, \gamma, \delta$	Hyper-parameters.

sparse measurement points [15]. To further enhance the quality of sampling-based RM construction, Kriging interpolation has been proposed. This method treats RM construction as a stochastic process modeling and prediction problem based on a covariance function, thereby improving the accuracy of RM construction [16].

The aforementioned sampling-based methods face two significant challenges: dependence on SPM and low construction accuracy [18], [19]. Consequently, sampling-free RM construction methods, based on NN, have attracted the attention of researchers. These methods typically require knowledge of the environmental features, such as obstacle locations, heights, and the positions of BS, to construct RMs. A representative example is the RadioUNet, which is derived from the classic U-Net framework used in image-to-image tasks [18]. This approach leverages the U-Net framework, training NN using the MSE between the generated RM and the ground truth, yielding impressive results [18]. Inspired by RadioUNet and the success of the Vision Transformer (ViT) led to the development of Radionet, based on the transformer architecture for RM construction [19]. Additionally, more complex NN architectures, such as graph neural networks (GNN) [20], with stronger feature extraction capabilities, have been employed in RM construction. However, these approaches generally treat sampling-free RM construction as a discriminative supervised learning task. Although RME-GAN attempts to introduce generative adversarial networks (GAN) methods into RM construction, it is not sampling-free, as it relies on SPM within the construction area [21]. Different from existing NN-based RM construction methods, in this paper, the sampling-free RM

construction is modeled as a conditional generative problem, and a diffusion-based method is proposed which significantly improves the RM construction performance.

### B. Diffusion Model

The diffusion model is a category of generative models based on Markov chains, that progressively restore data through a learned denoising process. This model has emerged as a strong competitor to the GAN in various generative tasks, such as computer vision [27], and natural language processing [22]. Moreover, diffusion models exhibit significant potential in perception tasks, including image segmentation, object detection, and model-based reinforcement learning (RL) [28]. For example, in [29], the authors effectively utilized diffusion model-based soft actor-critic algorithms for optimal contract design. Besides, the authors in [30] innovatively leveraged diffusion model-based deep deterministic policy gradient algorithms for optimal Stackelberg game solutions. In the diffusion model, there are two procedures which are the forward diffusion procedure where the raw data is diffused into noise, and the reversed denoise procedure where an NN is used to remove the noise adding to the raw data, thus generating raw data from noise.

1) *Forward Diffusion Procedure*: From a probabilistic modeling standpoint, the essence of generative models lies in training them to produce data  $\hat{x} \sim p_\theta(\hat{x})$  that mirror the distribution of the training data  $x \sim p_t(x)$ . The denoising diffusion probability model (DDPM) employs two Markov chains: a forward chain that converts data into noise, and a backward chain that reverts the noise to data. In a formal context, given the data  $x_0$ , the progression of a forward Markov chain is realized by generating a series of stochastic variables  $x_1, x_2, \dots, x_T$ , which evolve following the transition kernel  $q(x_t|x_{t-1})$ . By employing the chain rule of probability in conjunction with the Markovian property, it is possible to deconstruct the joint probability distribution of  $x_1, x_2, \dots, x_T$  given  $x_0$ , expressed as  $q(x_1, \dots, x_T|x_0)$ , into an appropriate factorial form.

$$q(x_1, \dots, x_T | x_0) = \prod_{t=1}^T q(x_t | x_{t-1}). \quad (1)$$

Thus, a forward noising process which produces latent  $x_t$  through  $x_0$  by adding Gaussian noise at time  $t \in \{0, 1, \dots, T\}$  can be defined as follows.

$$q(x_t | x_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}), \quad (2)$$

where  $T$  and  $\beta_t \in (0, 1)$  are the total number of diffusion iterations and hyper-parameter of the variance scaling factor, respectively. By setting  $\alpha_t = 1 - \beta_t$  and  $\bar{\alpha}_t = \prod_{s=0}^t \alpha_s$ , the distribution of  $x_t$  condition on the  $x_0$  can be obtained as follow.

$$q(x_t | x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (3)$$

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad (4)$$

where  $\mathcal{N}(\cdot, \cdot)$  is Gaussian distribution,  $\mathbf{I}$  is identity matrix, and  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ . In addition, Eq. (4) describes how the noisy data  $x_t$  is generated by combining the original input  $x_0$  with a

Gaussian noise term  $\epsilon$ . The term  $\sqrt{\bar{\alpha}_t}x_0$  represents the scaled contribution of the original data, while  $\sqrt{1 - \bar{\alpha}_t}\epsilon$  represents the amount of noise added at each step. As  $t$  increases, the influence of the noise term grows, resulting in progressively noisier data.

2) *Reversed Denoise Procedure*: For the data generation, DDPM initially creates unstructured noise vectors from the prior distribution, subsequently removing the noise through a learnable Markov chain operated in reverse temporal order. To achieve this, the reverse process can be formulated as follows.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \beta_t \mathbf{I}) \quad (5)$$

where  $\beta_t$  is a hyper-parameter,  $\theta$  is the parameter of NN  $\mu_\theta$ . Applying a trained NN  $\mu_\theta$ , we can iteratively denoise  $x_t$  from  $t = T$  to  $t = 1$  as follows [28].

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \mu_\theta(x_t, t) \right) + \beta_t \mathbf{I}. \quad (6)$$

Remarkably, the second term in (6) serves the purpose of obtaining the same distribution of  $x_{t-1}$  obtained through denoising, particularly in terms of variance, as that is derived through forward diffusion. If the second term of (6) is removed, the  $x_{t-1}$  obtained by denoising is just equal to the mean of which obtained via forward diffusion procedure. Strictly, the scaling factor in the second term of (6) should be  $\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$  to ensure distribution consistency. However, as shown in [31], setting the scaling factor to  $\beta_t$  achieves the same performance as setting it to  $\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$ , while also reducing computational complexity. Therefore, the scaling factor in the second term of (6) is  $\beta_t$ , both for efficiency and effectiveness.

### III. PROBLEM FORMULATION OF RM CONSTRUCTION

We consider the scenario where the RM needs to be constructed within an area as a grid of size  $N \times N$ , since the grid is small enough, the pathloss in a grid is a constant, where the RM can be represented by a matrix  $\mathbf{P}$ . Within this region, there is a BS with a single antenna and multiple static and dynamic obstacles. The location of BS can be represented by a tuple  $\mathbf{R}$  as  $(h, d_x, d_y)$ , where  $h, d_x$  and  $d_y$  is the height and coordinates of the BS. The static obstacles have varying sizes and shapes but are composed of the same surface material, thus reflecting and diffracting electromagnetic waves in the same way, and similar to [18], [19], [21] the pathloss is zero within their interiors. The static obstacle information is represented by the matrix  $\mathbf{H}_s \in \mathbb{R}^{N \times N}$ , where a value of  $h_{i,j}^s = 0, \forall h_{i,j}^s \in \mathbf{H}_s$  indicates the absence of the static obstacle at the position of  $i$ -th row and  $j$ -th column. Moreover, dynamic obstacles, such as vehicles, impact electromagnetic wave propagation through shielding, reflecting, and diffracting effects, similar to fixed obstacles, however, due to their small size and short height, dynamic obstacles cannot completely block the propagation of electromagnetic rays in its direction. The information of dynamic obstacles is represented by the matrix  $\mathbf{H}_d \in \mathbb{R}^{N \times N}$ , where the  $h_{i,j}^d = 0, \forall h_{i,j}^d \in \mathbf{H}_d$  indicates the absence of dynamic obstacle at the position of  $i$ -th row and  $j$ -th column.

The objective of this paper is to train an NN  $\mu_\theta(\cdot)$  with parameters  $\theta$  to predict the pathloss matrix  $\hat{\mathbf{P}} \in \mathbb{R}^{N \times N}$



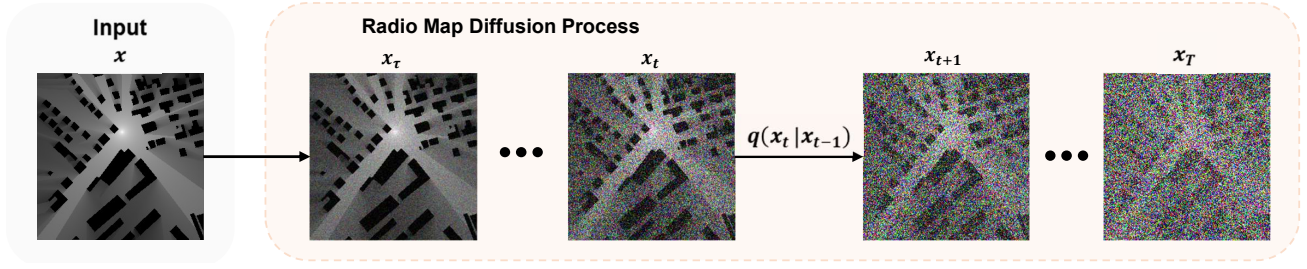


Fig. 2: The diffusion procedure of RM, where in diffusion procedure the RM is diffused into noise, and in the denoising procedure the RM is revealed from the noise and prompt.

based on the environmental features and location of the BS to minimize the difference between the predicted  $\hat{P}$  and the ground truth  $P$ . The difference can be measured by the criterion function  $\mathcal{L}(\hat{P}, P)$ . Therefore, the problem of RM construction can be formulated as

**Problem 1.**

$$\min_{\theta} \quad \mathcal{L}(\hat{P}, P), \quad (7)$$

$$s.t. \quad \hat{P} = \mu_{\theta}(H_s, H_d, R), \quad (7a)$$

The Problem 1 is a generative problem rather than a discriminative problem, which can be analyzed from both the data features and training method perspective. First, from the perspective of data features, the pathloss needs to be predicted from empty elements of the input data, as they do not exist in the environmental matrixes  $H_d$  and  $H_s$ , which should be generated by the NN  $\mu_{\theta}(\cdot)$ . Moreover, since the pathloss to be predicted is not a discrete value, it is almost impossible to construct the RM by dividing the finite hyperplane to classify the nodes. However, from the perspective of statistical learning, the fundamental of discriminative tasks, especially supervised learning, is to let NN learn from the latent space to the partitioning of hyperplanes to achieve data classification [24]. Thus, the discriminative method inevitably limits the performance of RM construction. Second, from the perspective of training methods, the self-supervised training method is mainly used by the generative model, where some elements of the raw data are masked and an NN is trained to recover the masked data through unmasked data in self-supervised learning [26]. In the RM construction, the environmental information  $\langle H_s, H_d \rangle$  can be regarded as  $P$  whose pathloss elements are masked, and the NN needs to predict the masked pathloss elements based on unmasked data that are  $\langle H_s, H_d \rangle$ . Therefore, the NN used for RM construction is trained in a self-supervised learning style. In addition, since the location of the BS  $R$  affects the distribution of pathloss, the BS can be regarded as a condition in self-supervision training, which means that the construction of RM is a condition generative problem.

#### IV. DIFFUSION-BASED RM CONSTRUCTION

As analyzed in section-III the RM construction is a generative problem, thus the SOTA generative diffusion model is used as a backbone to construct the RM effectively.

##### A. Initial Processing

To improve the convergence speed, the pathloss matrix  $P$  is encoded into a grayscale matrix through a process involving logarithmic scaling, normalization, and subsequent quantization [18]. Additionally,  $R$  is also represented as a grayscale matrix, with a pixel value of 1 denoting the location of the AP, while other pixels are set to 0. Subsequently, all the environment information is encoded into a three-channel tensor as a prompt tensor  $C = [H_s, H_d, R]$ . To further enhance the training efficiency of the denoise diffusion model, similar to [28], we initially train a variational autoencoder (VAE) to encode the raw data into the latent space for training and testing [32]. The encoder  $\mathcal{E}$  module of the VAE encodes  $P$  as a latent vector  $z_0$ , where the noise is added to the  $z_0$  according to (3) for training. After the noise is removed by the NN, the decoder module  $\mathcal{D}$  of the VAE is utilized to recover the RM from the diffusion prediction vector  $\hat{z}_0$ . Through the use of the VAE, the denoise diffusion model only needs to remove the noise vector added to  $z$  instead of the noise matrix originally added to  $P$ , thereby reducing the output space dimension of the diffusion model to enhance training efficiency. Thus, in the following of this paper, we use the latent vector  $z$  instead of  $x$  to denote the image to be generated. It is essential to highlight that the training of the VAE operates independently of the subsequent training of the denoise diffusion model, where the VAE only utilizes  $P$  for training the encoder  $\mathcal{E}$  and decoder  $\mathcal{D}$  trained in an autoencoder style. Then the VAE's parameters remain static and do not change in the subsequent training procedure of the diffusion model.

##### B. RadioDiff via Decoupled Diffusion Model

Although DDPM and latent diffusion models (LDM) have shown considerable potential across various fields, their extensive inference times and prolonged training durations motivate the use of a decoupled diffusion model (DDM) [33] for further enhancement. In DDM, the procedure of diffusing  $z_0$  to  $z_t$  is modeled as a two-stage continuous Markov process. First,  $z_0$  gradually diffuses into a  $0$  vector, then the noise  $\epsilon$  is added to this  $0$  vector to form  $z_t$ . The distribution of  $z_t$  can be formulated as follows.

$$q(z_t | z_0) = \mathcal{N}(\gamma_t z_0, \delta_t^2 I), \quad (8)$$

where  $\gamma_t$  and  $\delta_t$  are hyper-parameters, and  $\delta_t$  is designed to increase gradually over time while  $\gamma_t$  decreases. Compared

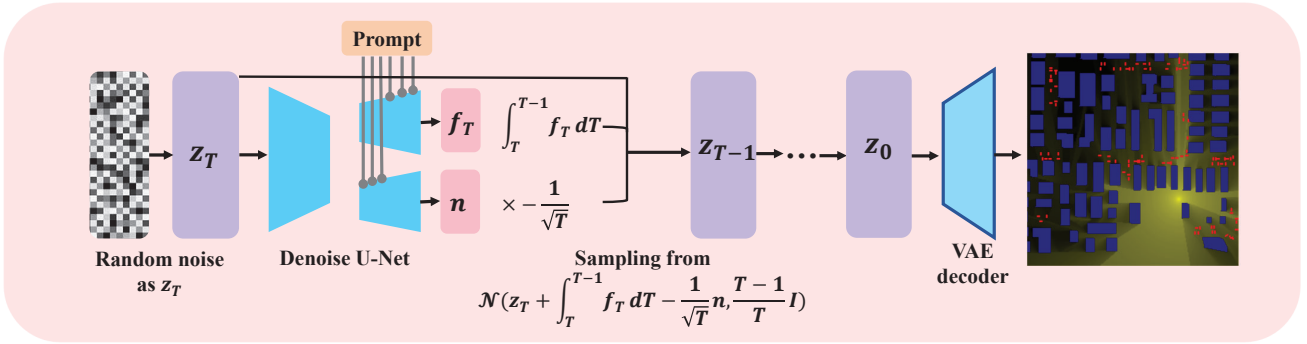
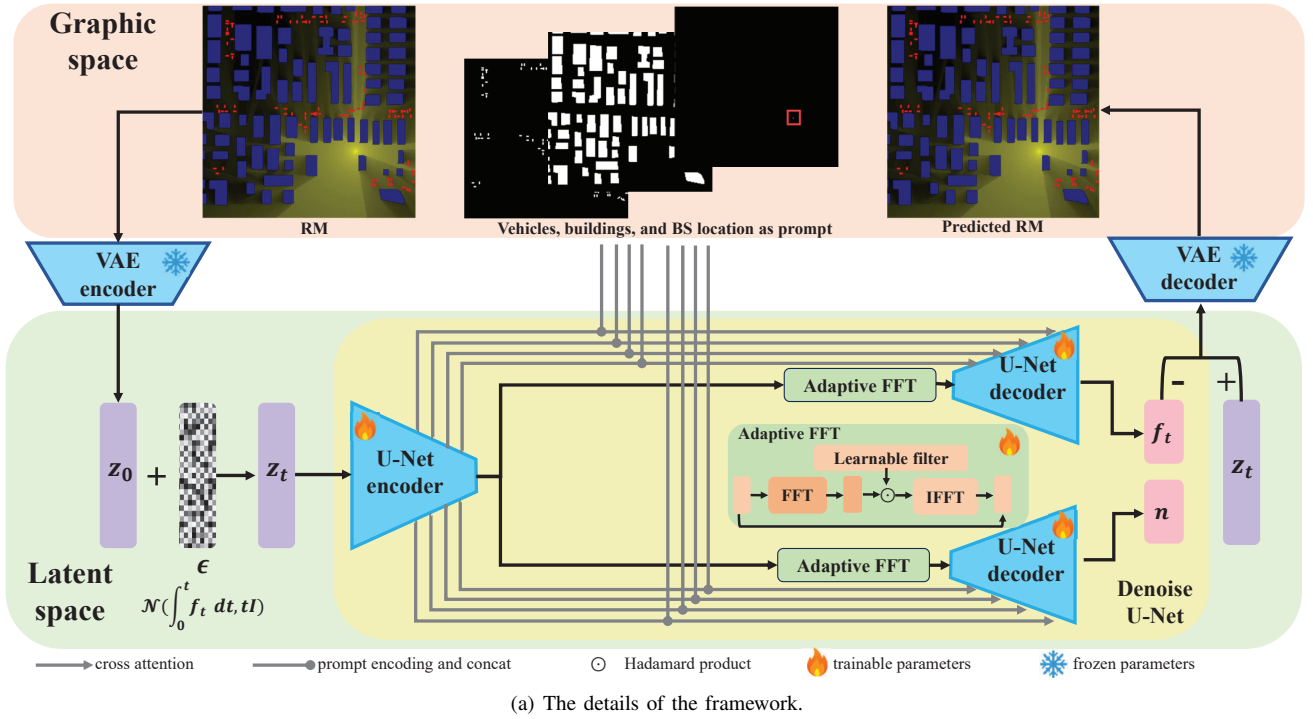


Fig. 3: The illustration of the proposed RadioDiff framework. The VAE is employed to encode the RM into a latent vector, thereby reducing the dimension of the input/output space for the denoise diffusion model. The framework incorporates a U-Net architecture, consisting of an encoder and decoder, to facilitate the denoising process. The prompt is represented as a grayscale diagram with three channels, each channel depicting the features of buildings, vehicles, and AP. After encoding the prompt, it is concatenated into the U-Net network, enabling the model to generate RMs under environmental conditions.

to traditional diffusion models where the diffusion process directly adds noise to the original input  $x_0$  or  $z_0$ , the decoupled diffusion model used in RadioDiff splits this process into two distinct stages. In DDM,  $z_0$  first diffuses into a  $\mathbf{0}$  vector, which effectively decouples the contribution of the original input from the added noise. This allows for a more controlled diffusion process, as the noise  $\epsilon$  is only introduced after  $z_0$  has been reduced to a  $\mathbf{0}$  vector. As a result, DDM reduces the variance in early diffusion steps and improves stability during both training and inference. Furthermore, this decoupled structure helps mitigate the prolonged inference time seen in traditional diffusion models, enabling more efficient generation in RadioDiff. Thus, according to [34], (8) can also be expressed as the following differential equation.

$$dz_t = f_t z_t dt + g_t d\epsilon_t, \quad (9)$$

$$f_t = \frac{d \log \gamma_t}{dt}, \quad (10)$$

$$g_t^2 = \frac{d\delta_t^2}{dt} - 2f_t \delta_t^2, \quad (11)$$

Based on the above equations, inverting  $z_t$  to  $z_0$  can be derived as follows.

$$dz_t = [f_t z_t - g_t^2 \nabla_x \log q(z_t)] dt + g_t d\bar{\epsilon}_t, \quad (12)$$

where  $\bar{\epsilon}_t$  is the Gaussian random variable in the reversed diffusion, playing a similar role as the second term of (6). By applying the decoupled diffusion strategy, the forward diffusion process can be redefined as follows.

$$z_t = z_0 + \int_0^t f_t dt + \int_0^t d\epsilon_t, \quad (13)$$

$$z_0 + \int_0^t f_t dt = \mathbf{0}, \quad (14)$$

where  $z_0 + \int_0^t f_t dt$  describes the data attenuation process and  $\int_0^t d\epsilon_t$  denotes the noise addition process. The function  $f_t$  is a differentiable function of  $t$ , and  $\epsilon_t$  is the standard Wiener process. Corresponding to equation (8), the conditional distribution can be simplified as follows.

$$q(z_t|z_0) = \mathcal{N}\left(z_0 + \int_0^t f_t dt, t\mathbf{I}\right). \quad (15)$$

By using the conditional probability formula the reverse sampling process,  $z_{t-\Delta t}$  can be obtained as follows.

$$q(z_{t-\Delta t} | z_t, z_0) = \mathcal{N}\left(z_t + \int_t^{t-\Delta t} f_t dt - \frac{\Delta t}{\sqrt{t}}\epsilon, \frac{\Delta t(t-\Delta t)}{t}\mathbf{I}\right). \quad (16)$$

According to (16), to reverse the noisy  $z_t$  into  $z_0$ , which corresponds to the feature map of RM  $\mathbf{P}$  in this paper, the NN only needs to predict two terms:  $\int_t^{t-\Delta t} f_t dt$  and  $\epsilon$ , since  $z_t$  is a known vector to denoise. As shown in Fig. 3, two U-Net decoder networks are employed to predict  $\int_t^{t-\Delta t} f_t dt$  and  $\epsilon$ , respectively. according to (13), the label for  $\epsilon$  is straightforward to obtain since the noise is added manually. However, the label for  $\int_t^{t-\Delta t} f_t dt$  must be obtained by solving the differential equation (14). In the training procedure, the ground truth is obtained by solving  $z_0 + \int_0^1 f_t dt = \mathbf{0}$ . For a simple example where  $f_t$  is equal to a constant  $c$ , the  $f_t$  can be easily determined as  $f_t = c = -z_0$ . For another scenario where the value  $f_t$  is in linear dependence with  $t$  that is  $f_t = at + b$ , solving the two parameters  $a$  and  $b$  using one equation is infeasible. To address this, we should sample one parameter from  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  and substitute it into  $z_0 + \int_0^1 f_t dt = \mathbf{0}$  to solve the other parameter. In this way, we concatenate  $a$  and  $b$  to obtain the ground truth. The ground truths for other functions can be determined similarly. Since different  $f_t$  calculation methods have different effects on the data quality generated by diffusion, the coefficients of the equations for calculating ground truth of  $f_t$  can also be considered as hyper-parameters.

It should be emphasized that the above diffusion and denoising processes are based on constraint-free RM construction. However, according to Problem 1, we need to achieve conditional RM construction based on environmental and BS location information, ensuring that the generated RM is correlated with these features. To enable the NN to generate the required RM based on the prompt  $\mathbf{C}$ , we employ a conditional generative model using an attention-based architecture. This approach leverages the attention mechanism, allowing the output of the NN to correlate with the given attention key-value vectors  $\mathbf{K}$  and  $\mathbf{V}$ . Since the attention architecture cannot directly handle two-dimensional data, an extractor NN  $\nu(\cdot)$  projects the prompt  $\mathbf{C}$  into an embedding space. This embedding is then projected to the intermediate layers of the U-Net through a cross-attention layer, implementing the attention mechanism  $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right) \cdot \mathbf{V}$  as follows.

$$\mathbf{Q} = \mathbf{W}_Q \cdot \phi(z_t), \quad (17)$$

$$\mathbf{K} = \mathbf{W}_K \cdot \nu(\mathbf{C}), \quad (18)$$

$$\mathbf{V} = \mathbf{W}_V \cdot \nu(\mathbf{C}), \quad (19)$$

where  $\phi(z_t)$  is the flattened operator applied on the output from the U-Net, which executes the function  $\theta$  and applies the transformation  $\mathbf{W}_V$ . Additionally,  $\mathbf{W}_Q$  and  $\mathbf{W}_K$  represent trainable projection matrices [35]. The training loss function can be represented as follows.

$$\mathcal{L} = \mathbb{E}_{z,c,t,\epsilon} \left[ \|\epsilon - \epsilon_\theta\|^2 + \|f - f_\theta\|_2^2 \right], \quad (20)$$

$$\epsilon_\theta, f_\theta = \mu_\theta(z_t; \mathbf{C}, t). \quad (21)$$

### C. Adaptive FFT Filter for DRM Enhancement

As shown in Fig. 1, RM exhibits numerous edge texture features, especially in DRM, which generate substantial high-frequency information in the frequency domain [36]. Although conventional convolutional layers within the denoising U-Net effectively extract features, they struggle to precisely capture high-frequency components. As a result, neural networks (NNs) based solely on traditional convolutional layers often produce overly smooth RM outputs, leading to blurred representations and suboptimal performance in dynamic environments.

To address this issue, as depicted in Fig. 3, an adaptive Fast Fourier Transform (FFT) filter (AFT) module is introduced, specifically to enhance the model's capacity for extracting high-frequency features. The AFT module operates by transforming the 2D feature maps  $\mathbf{z}$ , generated by the encoder with spatial dimensions  $H \times W$  and channel count  $C$ , from the spatial domain to the frequency domain via the FFT operation, represented as  $\mathbf{z}_c = \mathcal{F}(\mathbf{z})$ , where  $\mathcal{F}$  denotes the FFT. To improve the model's ability to focus on relevant frequency components, AFT incorporates a learnable weight matrix  $\mathbf{w} \in \mathbb{C}^{H \times W \times C}$ . This matrix is applied to the frequency-domain features  $\mathbf{z}_c$  via the Hadamard product,  $\mathbf{w} \odot \mathbf{z}_c$ . The learnable weight matrix dynamically adjusts the model's response to different frequency distributions in the target data, emphasizing important frequencies while attenuating irrelevant ones. This adaptive spectral filtering allows the model to perform global frequency adjustments based on data-driven relevance, ensuring that high-frequency features are enhanced while less critical frequencies are suppressed.

Following this adaptive filtering process, the modified frequency-domain features are transformed back into the spatial domain using the Inverse Fast Fourier Transform (IFFT). To preserve key information and mitigate potential losses during filtering, a residual connection is introduced between the original feature map  $\mathbf{z}$  and the output feature map. The complete operation of the AFT module can be described as follows:

$$\mathbf{z} = \mathbf{z} + \mathcal{F}^{-1}(\mathbf{w} \odot \mathbf{z}_c), \quad (22)$$

$$\mathbf{z}_c = \mathcal{F}(\mathbf{z}). \quad (23)$$

## V. EXPERIMENTS

### A. Datasets

In this study, we evaluate the performance of the proposed method using the RadioMapSeer dataset provided by the

pathloss RM construction challenge [37]. The dataset consists of 700 maps, each with unique geographic information (e.g., building data), with each map containing 80 transmitter locations and their corresponding ground truth data. Each map contains between 50 and 150 buildings. We selected 500 maps for the training dataset and the remaining 200 maps for the test dataset. There is no overlapping terrain information between the training and test datasets.

The city maps include data from cities such as Ankara, Berlin, Glasgow, Ljubljana, London, and Tel Aviv, sourced from OpenStreetMap. In the dataset, both transmitter and receiver heights are set at 1.5 meters, while building heights are set at 25 meters. Each map is converted into a  $256 \times 256$  pixel morphological 2D image with binary pixel values (0/1), where each pixel represents one square meter: '1' for areas inside buildings and '0' for areas outside. Transmitter positions are stored in a two-dimensional numerical format and depicted in morphological images, with the transmitter's pixel set to '1' and all others to '0'. The transmitter power is set to 23 dBm, and the carrier frequency is 5.9 GHz. To obtain a sufficiently accurate RM as ground truth for training, the ground truth RM in the dataset is constructed using Maxwell's equations, where the pathloss is calculated by the reflection and diffraction of electromagnetic rays. The RMs that only consider the impact of the static buildings on the electromagnetic rays are used as the ground truth of SRM. Additionally, the RMs in the dataset that consider the impact both of the static buildings and the vehicles, which are randomly generated along the roads, are used as the ground truth of the DRM, as shown in Fig. 1.

### B. Metrics

To comprehensively evaluate the quality of RM construction, we begin by adopting the parameters commonly used in previous studies [18], namely NMSE and RMSE. Additionally, we observe that the accurate generation of structural information and details is a key objective in RM reconstruction tasks, whereas the MSE index focuses on overall error and does not directly address these specific requirements. Therefore, we propose introducing structural similarity index measurement (SSIM) and peak signal-to-noise ratio (PSNR) as additional metrics in this paper. SSIM evaluates the preservation of structural information to emphasize the accuracy of structural detail reconstruction, while PSNR measures the signal-to-noise ratio to assess the fidelity of RM construction, particularly with respect to edge signal reconstruction.

1) *MSE*: MSE is calculated by averaging the squared differences between the pixel intensities of the original and final images as follows.

$$MSE = \frac{1}{NM} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e(m, n)^2, \quad (24)$$

where  $e(m, n)$  is the error difference between the ground truth and the predicted RM, and  $M, N$  is the length and width of the image, respectively. The normalized MSE (NMSE) is a scaled version of MSE employed to assess the predictive accuracy of

TABLE II: **Quantitative Comparison.** Results in bold red and underlined blue highlight the highest and second highest, respectively. The  $\uparrow$  indicates metrics whereby higher values constitute improved outcomes, with higher values preferred for all other metrics.

Methods	RME-GAN	RadioUNet	UVM-Net	RadioDiff (Ours)
SRM	NMSE	0.0115	<u>0.0074</u>	0.0085
	RMSE	0.0303	<u>0.0244</u>	0.0304
	SSIM $\uparrow$	0.9323	<u>0.9592</u>	0.9320
	PSNR $\uparrow$	30.54	<u>32.01</u>	30.34
DRM	NMSE	0.0118	0.0089	<u>0.0088</u>
	RMSE	0.0307	<u>0.0258</u>	0.0301
	SSIM $\uparrow$	0.9219	<u>0.9410</u>	0.9326
	PSNR $\uparrow$	30.40	<u>31.75</u>	30.42

the RM construction, where rooted MSE (RMSE) is the rooted MSE, which are defined as follows.

$$NMSE = \frac{\sum_{m=1}^M \sum_{n=1}^N (I_b(m, n) - I(m, n))^2}{\sum_{m=1}^M \sum_{n=1}^N I^2(m, n)}, \quad (25)$$

$$RMSE = \sqrt{NMSE} \quad (26)$$

2) *SSIM*: SSIM is a quality assessment metric inspired by the human visual system. Since SSIM focuses on measuring texture differences, and there are lots of high-frequency details in RM, SSIM is suitable for evaluating the quality of the generated results. We also believe that greater attention should be given to the brightness of the signal radiation, the contrast between the signal radiation and the surrounding area, and the accuracy of geographic map in RM reconstruction. This aligns with the SSIM metric, which evaluates three key components: brightness, contrast, and structural information, which can be calculated as follows.

$$l(x, y) = \frac{2\mu_X(x, y)\mu_Y(x, y) + C_1}{\mu_X^2(x, y) + \mu_Y^2(x, y) + C_1} \quad (27)$$

$$c(x, y) = \frac{2\sigma_X(x, y)\sigma_Y(x, y) + C_2}{\sigma_X^2(x, y) + \sigma_Y^2(x, y) + C_2} \quad (28)$$

$$s(x, y) = \frac{\sigma_{XY}(x, y) + C_3}{\sigma_X(x, y)\sigma_Y(x, y) + C_3} \quad (29)$$

where  $x, y$  correspond to two different input images and  $\mu_x, \sigma_x^2, \sigma_{xy}$  denote the mean and variance of  $x$  and the covariance of  $x$  and  $y$  respectively. In addition,  $C_1, C_2$ , and  $C_3$  are constants which are defined as follows.

$$C_1 = (K_1 L)^2, \quad C_2 = (K_2 L)^2, \quad C_3 = \frac{C_2}{2},$$

where  $L$  represents the dynamic range of the data. Based on these parameters, the structural similarity can be computed as described as follows.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (30)$$

3) *PSNR*: The PSNR is defined as the ratio between the maximum possible power of a signal and the power of interfering noise that affects the fidelity of its representation. PSNR is typically expressed in decibels (dB) and provides an approximate measure of the perceived quality of reconstruction. In image evaluation, a higher PSNR generally indicates better image quality. For RMs, an accurate edge signal is crucial; therefore, PSNR is used not only to assess overall image quality



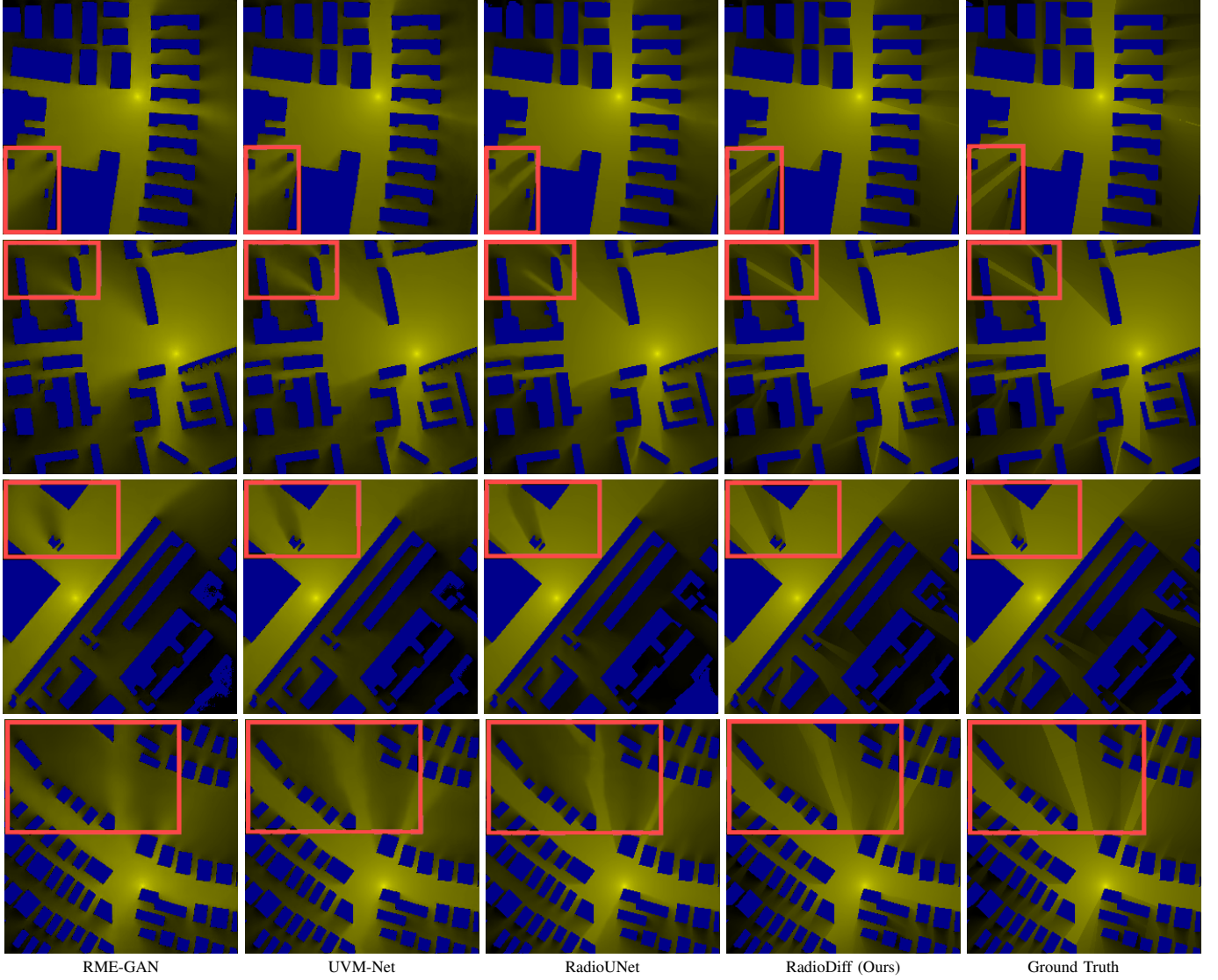


Fig. 4: The comparisons of constructed SRM on different methods.

but also to determine the quality of edge detail in the generated RMs. PSNR can be calculated as follows.

$$PSNR = 10 \log_{10} \left( \frac{r^2}{MSE} \right) \quad (31)$$

where  $r$  is the maximal variation in the input image data.

### C. Implementation

We implemented our RadioDiff framework using the PyTorch framework. The training process is divided into two phases, both utilizing an AdamW optimizer with a decaying learning rate, starting from  $5 \times 10^{-5}$  and reducing to  $5 \times 10^{-6}$ . In the initial phase, the autoencoder is trained using RM images from the entire dataset's training set as ground truth. This phase, which trains the VAE with  $z$ -channels set to 3 and embedding dimension of 128, takes approximately 120 hours on  $4 \times$  NVIDIA A100 SXM GPUs with a batch size of 2. The subsequent phase involves training the denoise diffusion U-net model, which takes around 360 hours using  $4 \times$  NVIDIA A100 SXM GPUs with a batch size of 64. In this implementation,

the input image size is  $256 \times 256$ . The diffusion process uses  $T = 1000$  in training, and the loss function is  $l_2$ -based with an objective of predicting KC. The start distribution is set to normal, and the perceptual weight is set to 0. Although a larger batch size can achieve better training results and higher training speed [26], the batch size in the first stage is significantly smaller than that in the second stage. In the first stage, both the input and output are raw image data, which consumes substantial video memory. In contrast, during the second stage, when training the diffusion model, the data is processed in the latent space encoded by the VAE, thereby reducing video memory consumption. Consequently, a larger batch size can be employed to enhance training speed and accuracy. This highlights the importance of using the VAE to encode data into the latent space. The hyper-parameter for the diffusion timesteps  $T$  is set to 500 for inference, and the diffusion process is decoupled with  $f_t$  defined as  $-z_0$ .

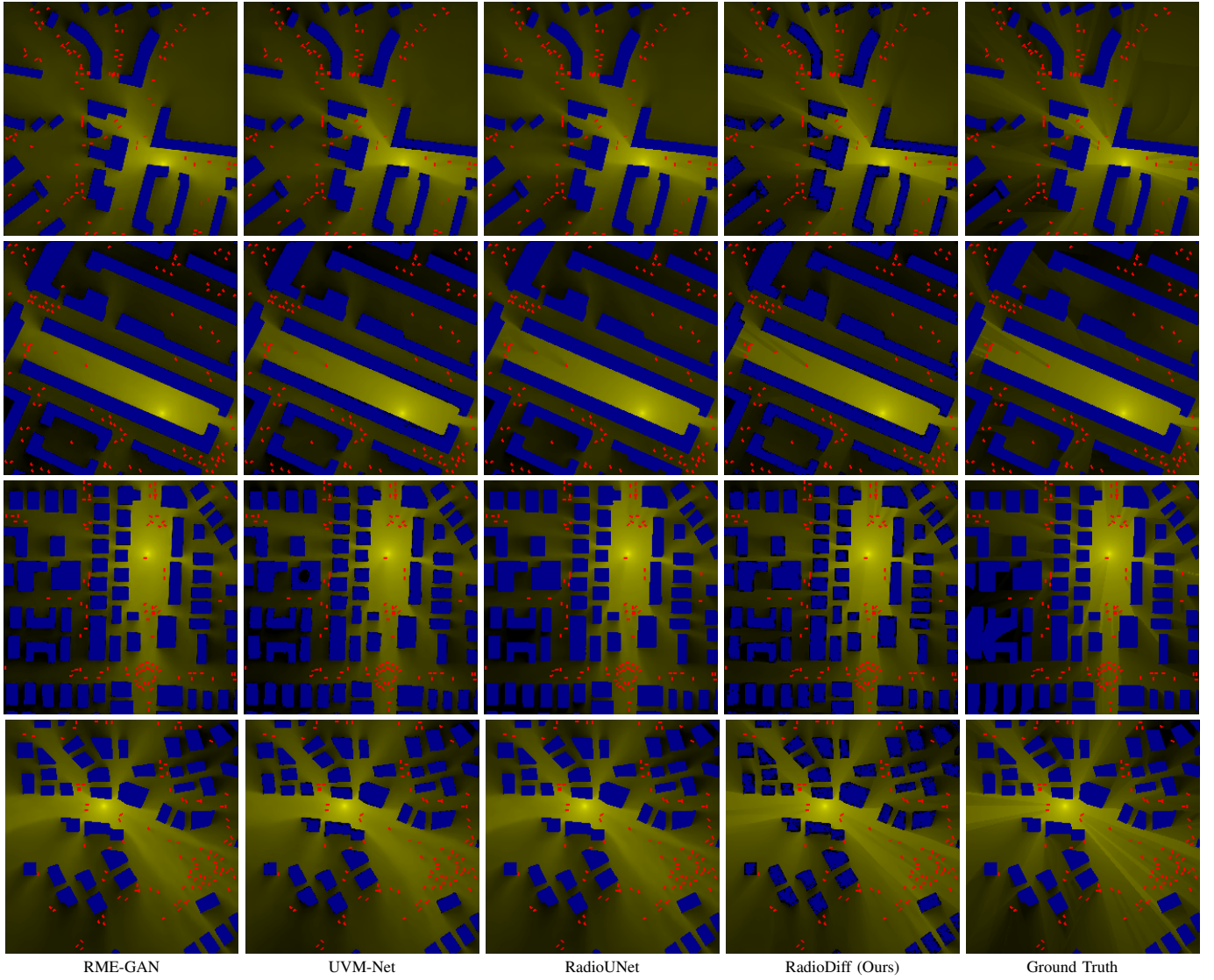


Fig. 5: The comparisons of constructed DRM on different methods.

#### D. Comparisons with SOTA Methods

To evaluate the proposed RadioDiff model, we compare it with other SOTA methods. To ensure a comprehensive comparison of the experiments, we compare the CNN-based, GAN-based, and Mamba-based methods separately, which represent the primary architectures used in the current RM reconstruction task based on deep learning. For the detailed parameter settings of the comparison model, we adhere to the description provided in the article [18], [21], [38]. Additionally, for fairness, the training and test data will be aligned with RadioDiff. The following method is used to compare.

- RadioUNet [18]: RadioUNet is one of the most effective sampling-free NN-based RM construction method, where a convolutional U-Net is used as the backbone NN, and the supervised learning is used to train the RadioUNet. RadioUNet reconstructs the wireless propagation graph using a simple yet effective network architecture that learns environmental characteristics, making it one of the most representative reconfiguration algorithms based on

reel machines.

- UVM-Net [38]: The training method and settings of UVM-Net are same as RadioUNet, but the backbone network is replaced by a convolutional layer with the latest SSMs (State Space Sequence Models). SSMs are specifically designed to handle long-sequence data and are particularly well-suited for modeling long-range dependencies. SSMs map the input sequence to a hidden state through a state space model and predict the output based on this hidden state. As a result, SSMs exhibit enhanced local feature capture and efficient remote modeling capabilities. We selected UVM-Net as the baseline to evaluate the performance differences between this SSM-driven architecture and traditional convolutional networks in the wireless propagation graph reconstruction task.
- RME-GAN [21]: The SOTA NN-based RM construction method, which uses a generative model cGAN to construct RM. However, RME-GAN is sampling-based, it not only uses environmental features but also uses the sampling pathloss as the input to construct the RM. For a fair

TABLE III: Ablation Study about AFT.

RadioDiff	NMSE	RMSE	PSNR	SSIM
w/o AFT	0.0067	0.0259	31.62	0.9465
w/ AFT	<b>0.0049</b>	<b>0.0190</b>	<b>35.13</b>	<b>0.9691</b>

comparison, the RME-GAN in this paper only uses environmental features as input. As GAN models have been at the forefront of generative model research in recent years, they effectively showcase the unique capabilities and challenges of generative adversarial networks in handling RM reconstruction tasks.

1) *Comparisons for SRM*: For the quantitative comparison on the RadioMapSeer-Test dataset for SRM scenarios is given in the first part of Table II and Fig. 4, our model outperforms others in error metrics, i.e., NMSE, RMSE, and structural metrics, i.e., SSIM, PSNR, indicating that our predictions and generated RM are more accurate. Notably, RadioDiff excels in the PSNR metric, indicating that the RMs it generates have clearer and sharper structural edges compared to other methods. Furthermore, the qualitative comparison presented in Fig. 4 demonstrates that the RMs constructed by RadioDiff closely resemble the ground truth, with well-defined edge features. This precision stems from the diffusion model's heightened sensitivity to the high-frequency signals of edge information, while the AFTs effectively enhance and isolate these signals. In contrast, structures such as CNNs and Mamba are less responsive to edge information, as is shown in RadioUNet and UVM-Net since they show inaccuracies in radiation signal positioning and edge blurring. As for the generative model RME-GAN, which relies on adversarial strategies, it tends to yield ambiguous and less precise results since it relies on the sampling position measurements, so in the sampling-free RM construction scenario it performs poorly.

2) *Comparisons for DRM*: As shown in the second part of Table II and Fig. 5, the quantitative comparison on the RadioMapSeer-Test dataset for DRM scenarios is given. In DRM scenarios, the models must account for additional dynamic environmental factors. Despite a general decline in performance, the second part of Table II shows that RadioDiff consistently delivers the best results across all indicators. As shown in Fig. 5, the RadioDiff model exhibits enhanced sensitivity to dynamic environmental factors such as vehicles, whereas the RME-GAN, RadioUNet, and UVM-Net models struggle with these elements, often resulting in significant blurring and distortion. This further highlights that RadioDiff has stable high performance under more challenging conditions, particularly in scenarios characterized by complex environments and overlapping signals.

### E. Ablation Study

In this section, we analyze the impact of the AFT on the performance of the RadioDiff model. The qualitative results in Fig. 6 vividly illustrate the visual disparity between radio patterns generated with and without employing the AFT. By incorporating the AFT, our model exhibits an enhanced capability to accurately detect and represent edge signals. This improvement is particularly noticeable when dealing

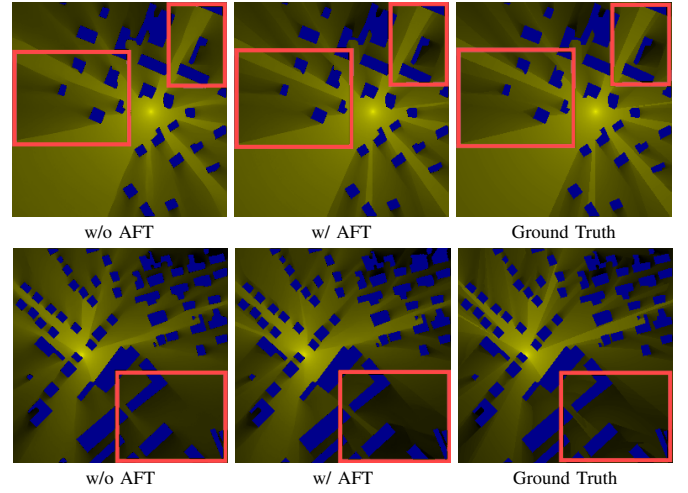


Fig. 6: Ablation study about AFT. The qualitative results demonstrate that incorporating the AFT further enhances the model's sensitivity to edge signals. This leads to RM images with more accurate edges and more robust results when multiple signals overlap.

with scenarios involving signal superposition, as evidenced by images produced by our model equipped with AFTs that possess increasingly sharper edges. Furthermore, Table III presents quantitative comparison results, demonstrating that a better performance can be achieved through the utilization of the AFT.

### F. Limitations and Discussion

Although this paper proposes a pioneer exploration of utilizing diffusion models for RM construction, and the proposed RadioDiff achieves SOTA performance, the issue of efficiency remains a key consideration for such a large generative model-based method. Table IV presents a comparison of the inference time and memory usage between the RadioDiff model and alternative models, showing that the diffusion model requires more resources and has a longer inference time than other methods. However, it should be emphasized that although the proposed RadioDiff is more time-consuming than other NN-based methods, the inference delay is still less than one second, which remains acceptable for dynamically constructing RM. For training data, RadioDiff uses the same dataset as other NN-based methods, and no limitations regarding its access to training data were identified in the experiment. In addition, although VAE needs to be trained separately before RadioDiff's formal training, this is due to LDM being used for the RM reconstruction task for the first time. In subsequent research, the relevant pre-trained weights of VAE can be directly used without re-training, which will significantly streamline future research efforts. Moreover, techniques such as NN compression and efficient inference methods like denoising diffusion implicit models (DDIM) [39] can be leveraged to notably enhance efficiency, raising a trade-off between performance and efficiency, which stands as a promising direction for future research.

Furthermore, it is noteworthy that almost all existing NN-based RM construction methods, including this paper, concentrate on predicting pathloss from a BS to a specific



TABLE IV: Inference Time and Memory Consuming. Results in bold red and underlined blue highlight the highest and second highest, respectively.

Method	Average time (s)	Memory consuming (MB)
RME-GAN	0.042	2923
UVM-Net	<u>0.095</u>	<b>8738</b>
RadioUNet	0.056	3927
RadioDiff (Ours)	<b>0.6</b>	<u>6062</u>

location, known as one-to-any (O2X) RM. However, there is also another type of RM that is any-to-any (X2X) RM scenario, which aims at obtaining pathloss between any two points through their positions. X2X RM construction poses a challenge with sampling-based approaches, as these methods typically require a fixed BS location to estimate pathloss of any other position to the fixed BS. In contrast, by adopting a sampling-free RM construction approach in this paper and incorporating the location of BS as input for the diffusion model as part of the prompt, it is feasible to predict pathloss between any two points by adjusting the prompt of the BS position, thereby enabling the construction of X2X RM.

## VI. CONCLUSION

In this paper, we have proposed RadioDiff, a diffusion-based RM generative model to effectively construct the RM. By incorporating various techniques, including AFTs and the decoupled diffusion model, RadioDiff can construct accurate and sharp RM effectively. Extensive experiments demonstrate the qualitative and quantitative superiority of the proposed RadioDiff. As the first application of diffusion models to RM construction tasks, RadioDiff sets a new benchmark for future technological advancements. In future work, we will focus on how to leverage the diffusion model to generate the environment features based on the sparse RM information.

## REFERENCES

- [1] R. Liu, B. P. L. Lau, K. Ismail, A. Chaturanga, C. Yuen, S. X. Yang, Y. L. Guan, S. Mao, and U.-X. Tan, "Exploiting radio fingerprints for simultaneous localization and mapping," *IEEE Pervasive Comput.*, 2023.
- [2] Z. Jiwei, C. Jiacheng, S. Zeyu, S. Yuhang, Z. Haibo, and X. Shen, "Channel-feedback-free transmission for downlink FD-RAN: A radio map based complex-valued precoding network approach," *China Commun.*, vol. 21, no. 4, pp. 10–22, 2024.
- [3] M. Xu, D. Niyato, J. Kang, Z. Xiong, S. Mao, Z. Han, D. I. Kim, and K. B. Letaief, "When large language model agents meet 6G networks: Perception, grounding, and alignment," *IEEE Wireless Communications*, pp. 1–9, 2024.
- [4] N. Cheng, F. Lyu, W. Quan, C. Zhou, H. He, W. Shi, and X. Shen, "Space/aerial-assisted computing offloading for IoT applications: A learning-based approach," *IEEE J. Select. Areas Commun.*, vol. 37, no. 5, pp. 1117–1129, 2019.
- [5] X. Wang, L. Fu, N. Cheng, R. Sun, T. Luan, W. Quan, and K. Aldubaikhy, "Joint flying relay location and routing optimization for 6G UAV-IoT networks: A graph neural network-based approach," *Remote Sensing*, vol. 14, no. 17, p. 4377, 2022.
- [6] X. Wang, N. Cheng, L. Fu, W. Quan, R. Sun, Y. Hui, T. Luan, and X. Shen, "Scalable resource management for dynamic mec: An unsupervised link-output graph neural network approach," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2023, pp. 1–6.
- [7] X. Wang, X. Wang, S. Mao, J. Zhang, S. C. Periaswamy, and J. Patton, "Indoor radio map construction and localization with deep gaussian processes," *IEEE Internet Things J.*, vol. 7, no. 11, pp. 11 238–11 249, 2020.

- [8] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?" *Nature Electronics*, vol. 3, no. 1, pp. 20–29, 2020.
- [9] X. Shen, J. Gao, M. Li, C. Zhou, S. Hu, M. He, and W. Zhuang, "Toward immersive communications in 6G," *Frontiers in Computer Science*, vol. 4, p. 1068478, 2023.
- [10] N. Cheng, F. Chen, W. Chen, Z. Cheng, Q. Yang, C. Li, and X. Shen, "6G omni-scenario on-demand services provisioning: vision, technology and prospect(in chinese)," *Sci Sin Inform.*, vol. 54, pp. 1025–1054., 2024.
- [11] Y. Liu, S. Zhang, F. Gao, J. Ma, and X. Wang, "Uplink-aided high mobility downlink channel estimation over massive MIMO-OTFS system," *IEEE J. Select. Areas Commun.*, vol. 38, no. 9, pp. 1994–2009, 2020.
- [12] C. You, B. Zheng, and R. Zhang, "Wireless communication via double irls: Channel estimation and passive beamforming designs," *IEEE Wireless Commun. Lett.*, vol. 10, no. 2, pp. 431–435, 2020.
- [13] Y. Zeng, J. Chen, J. Xu, D. Wu, X. Xu, S. Jin, X. Gao, D. Gesbert, S. Cui, and R. Zhang, "A tutorial on environment-aware communications via channel knowledge map for 6G," *IEEE Commun. Surveys Tuts.*, 2024.
- [14] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [15] F. J. Breidt and J. D. Opsomer, "Local polynomial regression estimators in survey sampling," *Annals of statistics*, pp. 1026–1053, 2000.
- [16] Y. Qiu, X. Chen, K. Mao, X. Ye, H. Li, F. Ali, Y. Huang, and Q. Zhu, "Channel knowledge map construction based on a UAV-assisted channel measurement system," *Drones*, vol. 8, no. 5, p. 191, 2024.
- [17] S. Oh and N.-H. Myung, "MIMO channel estimation method using ray-tracing propagation model," *Electronics letters*, vol. 40, no. 21, p. 1, 2004.
- [18] R. Levie, Ç. Yapar, G. Kutyniok, and G. Caire, "RadioUNet: Fast radio map estimation with convolutional neural networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 4001–4015, 2021.
- [19] H. Li, K. Gupta, C. Wang, N. Ghose, and B. Wang, "RadioNet: Robust deep-learning based radio fingerprinting," in *2022 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2022, pp. 190–198.
- [20] G. Chen, Y. Liu, T. Zhang, J. Zhang, X. Guo, and J. Yang, "A graph neural network based radio map construction method for urban environment," *IEEE Commun. Lett.*, 2023.
- [21] S. Zhang, A. Wijesinghe, and Z. Ding, "RME-GAN: a learning framework for radio map estimation based on conditional generative adversarial network," *IEEE Internet Things J.*, 2023.
- [22] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg, "Structured denoising diffusion models in discrete state-spaces," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 981–17 993, 2021.
- [23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [24] V. N. Vapnik, V. Vapnik *et al.*, *Statistical learning theory*. wiley New York, 1998.
- [25] X. Wang, Q. Qiu, and N. Cheng, "Reliable projection based unsupervised learning for semi-definite QCQP with application of beamforming optimization," *arXiv preprint arXiv:2407.03668*, 2024.
- [26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [27] X. Lin, J. He, Z. Chen, Z. Lyu, B. Fei, B. Dai, W. Ouyang, Y. Qiao, and C. Dong, "Diffbir: Towards blind image restoration with generative diffusion prior," *arXiv preprint arXiv:2308.15070*, 2023.
- [28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.
- [29] J. Wen, J. Nie, Y. Zhong, C. Yi, X. Li, J. Jin, Y. Zhang, and D. Niyato, "Diffusion-model-based incentive mechanism with prospect theory for edge aige services in 6G IoT," *IEEE Internet of Things Journal*, vol. 11, no. 21, pp. 34 187–34 201, 2024.
- [30] B. Lai, J. Wen, J. Kang, H. Du, J. Nie, C. Yi, D. I. Kim, and S. Xie, "Resource-efficient generative mobile edge networks in 6G era: Fundamentals, framework and case study," *IEEE Wireless Communications*, vol. 31, no. 4, pp. 66–74, 2024.
- [31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [32] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [33] Y. Huang, Z. Qin, X. Liu, and K. Xu, "Decoupled diffusion models: Simultaneous image to zero and zero to noise," 2024.
- [34] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.



- [35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [36] L. Chen, L. Gu, D. Zheng, and Y. Fu, "Frequency-adaptive dilated convolution for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3414–3425.
- [37] Ç. Yapar, F. Jaensch, R. Levie, G. Kutyniok, and G. Caire, "The first pathloss radio map prediction challenge," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–2.
- [38] Z. Zheng and C. Wu, "U-shaped vision mamba for single image dehazing," *arXiv preprint arXiv:2402.04139*, 2024.
- [39] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.