RESEARCH ARTICLE

DT-NeRF: A Diffusion and Transformer-Based Optimization Approach for Neural Radiance Fields in 3D Reconstruction

Bo Liu[1], **Runlong Li**[2,*], **Li Zhou**[3] **and Yan Zhou**[4]

[1] College of Computer Sciences, Northeastern University, Boston, MA 02115, United States
[2] Department of Electrical Engineering and Computer Science, University of California, Irvine, CA 92697, United States
[3] Desautels Faculty of Management, McGill University, Montréal, QC H3A 0G4, Canada
[4] Department of Mathematics, Northeastern University, Boston, MA 02115, United States

Abstract

This paper proposes a Diffusion Model-Optimized Neural Radiance Field (DT-NeRF) method, aimed at enhancing detail recovery and multi-view consistency in 3D scene reconstruction. By combining diffusion models with Transformers, DT-NeRF effectively restores details under sparse viewpoints and maintains high accuracy in complex geometric scenes. Experimental results demonstrate that DT-NeRF significantly outperforms traditional NeRF and other state-of-the-art methods on the Matterport3D and ShapeNet datasets, particularly in metrics such as PSNR, SSIM, Chamfer Distance, and Fidelity. Ablation experiments further confirm the critical role of the diffusion and Transformer modules in the model's performance, with the removal of either module leading to a decline in performance. The design of DT-NeRF showcases the synergistic effect between modules, providing an efficient and accurate solution for 3D scene reconstruction. Future research may focus on further optimizing the model, exploring more advanced generative models and network architectures to enhance its performance in large-scale dynamic scenes, with the ultimate goal of enabling intelligent systems to perceive and reconstruct complex 3D environments autonomously.

Keywords: diffusion model, NeRF, 3D reconstruction, detail recovery, transformer network, intelligent systems, scene understanding.

1 Introduction

Neural Radiance Fields (NeRF) have emerged as a novel 3D scene reconstruction technology, attracting widespread attention in the fields of computer vision and graphics in recent years. The core concept of NeRF is to utilize a multi-layer perceptron (MLP) to learn the propagation of light in a 3D scene from different viewpoints, enabling high-quality 3D reconstruction. NeRF is capable of generating intricate lighting and texture details at multiple scales, with recent extensions demonstrating significant

Academic Editor:
Seifedine Kadry

Submitted: 06 June 2025
Accepted: 05 July 2025
Published: 25 August 2025

Vol. 2, **No.** 3, 2025.
10.62762/TIS.2025.874668

***Corresponding author:**
Runlong Li
runlonl@uci.edu

Citation
Liu, B., Li, R., Zhou, L., & Zhou, Y. (2025). DT-NeRF: A Diffusion and Transformer-Based Optimization Approach for Neural Radiance Fields in 3D Reconstruction. *ICCK Transactions on Intelligent Systematics*, 2(3), 190–202.

© 2025 ICCK (Institute of Central Computation and Knowledge)

advancements in anti-aliasing and multi-scale scene representation for photorealistic rendering [1]. Despite its strong performance in 3D reconstruction, NeRF still faces challenges in generalizing across scenes and disentangling geometry from appearance, motivating geometry-aware generative extensions to the core radiance field framework [2]. Due to its strong reliance on viewpoint information, the reconstruction quality often suffers when the viewpoints are insufficient, leading to loss of details or poor consistency [3, 4]. These capabilities are of particular relevance to intelligent systems that rely on accurate environmental modeling for tasks such as autonomous navigation, robotic manipulation, and scene understanding.

As a result, traditional NeRF has some inherent limitations, especially in scenarios with sparse viewpoints and complex geometries, where reconstruction quality and efficiency are often constrained [5]. Specifically, NeRF struggles to effectively recover details when there are fewer viewpoints, and maintaining scene consistency becomes challenging. Moreover, NeRF's computational cost is high, particularly when handling complex geometries and large-scale scenes, where training time and inference speed become bottlenecks. While NeRF can generate high-quality images via volumetric rendering, yet its performance is sensitive to the accuracy of camera parameters, with reconstruction quality degrading substantially when precise camera poses are unavailable—a practical constraint that limits deployment in uncontrolled capture conditions [6].

To address these issues, recent studies have increasingly integrated other advanced deep learning techniques with NeRF to enhance its performance in complex scenes [7]. Some methods introduce generative models to gradually optimize the details and quality of images. These generative models have shown significant success in image restoration and enhancement, particularly in recovering details and improving generated image quality. For example, some denoising-based generative models can iteratively remove noise at each step, optimizing the final image's quality and consistency. Additionally, another class of methods introduces global feature aggregation strategies to effectively handle long-range dependencies in 3D point cloud data, enabling the model to better capture spatial relationships in complex geometric scenes, thereby improving the ability to reconstruct geometric details [8]. By combining these methods, it is expected that NeRF's detail recovery under sparse viewpoints, multi-view consistency, and the optimization of 3D scene geometries will be significantly enhanced-as demonstrated by depth-supervised approaches that achieve comparable reconstruction quality with substantially fewer input views [9]. The main contributions of this paper are as follows:

- We introduce a diffusion model into the training process of NeRF, which generates latent features that effectively compensate for the lack of viewpoints in sparse-view scenarios. This aids in detail recovery and improves image quality, addressing the limitations of traditional NeRF in these conditions.

- We embed a Transformer into the rendering process of NeRF, replacing the traditional MLP structure with self-attention mechanisms. This improvement enhances the model's ability to capture long-range dependencies in 3D scenes, which significantly boosts the accuracy of geometric modeling and detail reconstruction.

- By combining these two advanced techniques—diffusion models and Transformers—we propose an efficient joint optimization framework that results in significant improvements in image quality, geometric accuracy, and multi-view consistency, making the method particularly effective for complex 3D scene reconstruction tasks.

The structure of this paper is as follows: Section 2 reviews related works, including the basic principles and limitations of NeRF, applications of diffusion models in image generation, and relevant uses of Transformers in computer vision. Section 3 provides a detailed description of the proposed DT-NeRF model, including the overall architecture, specific designs of the diffusion model module and the Transformer module. Section 4 presents the experimental setup, datasets, evaluation metrics, and validation of our model's effectiveness through comparative and ablation experiments. Finally, Section 5 summarizes the contributions of this paper and outlines future research directions.

2 Related Work

2.1 Application of Traditional Methods in Image Generation and 3D Scene Reconstruction

In recent years, with the rapid development of computer vision technologies, numerous methods have been proposed to address the challenges in image generation and 3D scene reconstruction [10]. Traditional image generation methods, such as multi-view stereo (MVS)-based techniques, restore the 3D geometric information of a scene by matching and fusing images from multiple viewpoints [11]. However, these methods rely on strong dependencies between viewpoints and often face performance bottlenecks in scenarios with sparse viewpoints or complex reconstruction details. Another class of methods is based on volumetric rendering, using ray tracing for scene reconstruction, such as employing the Marching Cubes algorithm and voxel grids for spatial partitioning and modeling [12]. While this approach has certain advantages in efficiently reconstructing scene structures, it incurs high computational overhead, a limitation well-documented across volumetric 3D scene representation methods including semantic scene completion [13]. Additionally, some methods use image-based reconstruction techniques, which utilize image segmentation and depth estimation technologies for rapid 3D scene reconstruction across diverse capture configurations, including omnidirectional and panoramic imaging setups. Although these methods are relatively straightforward, they still face limitations in detail representation and geometric complexity when applied to general indoor and outdoor environments [14]. Furthermore, SLAM (Simultaneous Localization and Mapping) technology, as a real-time 3D mapping method, enables dynamic scene reconstruction using information obtained from cameras or other sensors, but its performance in large-scale environments remains constrained by hardware limitations [15]. Recently, the deep learning-based NeRF method has achieved significant results in high-quality 3D reconstruction by modeling light propagation [16]. However, its dependence on numerous viewpoints and computational resources limits its application in scenarios with sparse viewpoints and complex scenes [17]. Recent advancements in diffusion models, such as Stable Diffusion and DDPM, have shown promise in overcoming these challenges, particularly for sparse-view reconstruction. Additionally, Vision Transformers (ViTs) have been advanced through focal attention mechanisms that enable fine-grained local interactions alongside coarse global context modeling [18], and have been successfully applied to multi-view 3D reconstruction tasks to improve long-range dependency capture in complex scenes [19].

Compared to these traditional methods, this paper introduces diffusion models and Transformers to optimize NeRF, aiming to address issues related to detail loss under sparse viewpoints and deficiencies in modeling complex geometries. Unlike traditional methods that rely on explicit geometric modeling or image matching, our approach enhances detail restoration through generative models and improves the capture of geometric and spatial information using self-attention mechanisms in deep learning, thereby improving both the effectiveness and efficiency of 3D reconstruction.

2.2 Innovative Application of Deep Learning in 3D Scene Reconstruction

In recent years, with the continuous development of deep learning technologies, many studies have gradually integrated deep neural networks with 3D scene reconstruction [20, 21], especially with significant advancements in the application of Neural Radiance Fields (NeRF) in this field. Many NeRF-based optimization methods have attempted to improve the model's performance in complex scenes. For example, NeRF in the Wild (NeRF-W) extends NeRF to handle unconstrained photo collections captured under varying illumination, transient occlusions, and inconsistent appearance conditions, significantly improving reconstruction robustness on in-the-wild image sets that lack controlled capture [22]. Additionally, FastNeRF optimizes the computational process of NeRF by utilizing hierarchical networks and acceleration techniques, improving training speed and inference efficiency, thus making real-time applications feasible [23]. However, while these methods enhance the model's efficiency and detail recovery capabilities, they still face certain limitations in handling complex lighting, long-range dependencies, and geometric structure modeling, particularly in large-scale and dynamic scenes, where detail recovery and consistency maintenance are not ideal. Meanwhile, another line of research has attempted to combine Generative Adversarial Networks (GANs) with NeRF to enhance the details and realism of reconstructed images. For instance, NeRF-GAN combines the generative capabilities of

GANs with NeRF's volumetric rendering, improving scene detail and texture representation [24]. However, it still faces challenges in maintaining multi-view consistency, particularly when processing dynamic scenes, where the generated images may exhibit inconsistencies. Other studies, such as Point-NeRF optimizes geometric modeling by incorporating point cloud representations into the radiance field [25], while the Scene Representation Transformer (SRT) leverages set-latent Transformer architectures for geometry-free novel view synthesis, capturing global scene context without explicit geometric primitives [26]. However, these methods still rely on a large amount of viewpoint information and involve high computational complexity. Thus, despite progress in some areas, existing methods continue to face challenges in detail recovery, geometric modeling, and computational efficiency [27]. The computational cost of NeRF-W and the instability of GAN-based NeRF methods further highlight the need for a more efficient and stable solution, which is addressed by the DT-NeRF model.

In contrast to these methods, the DT-NeRF model proposed in this paper combines diffusion models with Transformers to further enhance detail recovery and geometric modeling capabilities while retaining the advantages of traditional NeRF. We optimize the training process through latent features generated by the diffusion model and enhance global context modeling using Transformers. This approach addresses the shortcomings in sparse viewpoints and complex geometric modeling while also optimizing computational efficiency.

3 Methodology

3.1 Overall Model Architecture

The DT-NeRF model combines diffusion models and Transformers to optimize the performance of NeRF in complex 3D scene reconstruction. Figure 1 illustrates the architecture of DT-NeRF, where the diffusion model, Transformer, and NeRF decoder are closely integrated to form an innovative end-to-end 3D reconstruction framework. This framework first generates latent features through the diffusion model, which are then used as conditional inputs for the NeRF decoder. The Transformer module optimizes the input data through a self-attention mechanism, ultimately enhancing the accuracy of scene geometric modeling and detail recovery. The design of the overall architecture enables DT-NeRF to effectively address the challenges of sparse viewpoints and complex geometric reconstruction.

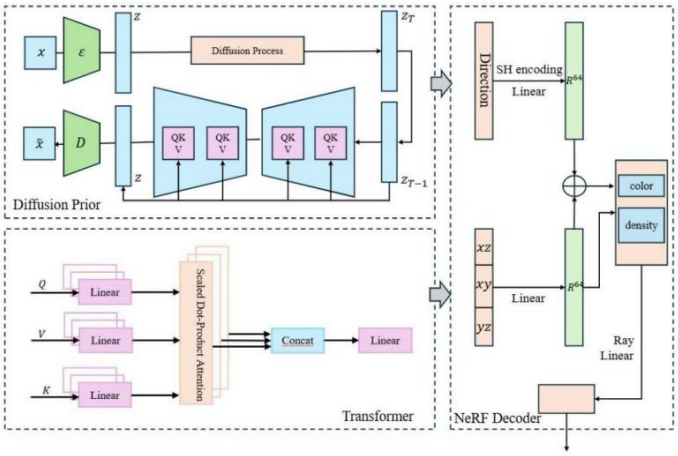


Figure 1. Architecture of the DT-NeRF model.

In DT-NeRF, the role of the diffusion model is critical. The diffusion model progressively denoises the initial noisy image to transform it into latent features. During this process, the diffusion model not only effectively enhances image details but also generates rich latent features during training, which are passed as conditional inputs to the NeRF decoder. In this way, the diffusion model provides more refined input data for NeRF while addressing the issue of information scarcity caused by sparse viewpoints. Particularly when the viewpoints are limited, the diffusion model compensates for the shortcomings of traditional NeRF methods in such scenarios by enhancing the latent features of the input, thus improving the detail and consistency of the reconstruction.

Complementing this is the introduction of the Transformer module. The primary role of the Transformer in DT-NeRF is to optimize NeRF's geometric modeling capabilities. By introducing the self-attention mechanism, the Transformer module enables global context modeling of 3D point cloud data, capturing long-range dependencies and details in complex geometric structures—capabilities that are critical for data-efficient 3D scene understanding, where learning rich contextual representations from limited annotations remains a key challenge [28]. This process significantly enhances the model's geometric modeling capabilities in complex scenes, particularly when dealing with scenes with highly intricate structures. The Transformer ensures that geometric details in the reconstruction process are more accurate by focusing on the relationships between different 3D points. Additionally, the Transformer-optimized feature inputs effectively improve the performance of the NeRF decoder, resulting in finer and more

consistent reconstructed images.

The design of the DT-NeRF model innovatively combines the advantages of both the diffusion model and the Transformer. The diffusion model generates high-quality latent features, enhancing NeRF's input and compensating for the lack of viewpoints in sparse-view scenarios. The Transformer improves geometric reconstruction and global context modeling by capturing long-range dependencies, boosting accuracy in complex scenes. Both the Diffusion and Transformer models are trained jointly within a unified optimization framework, sharing the same loss function. These modules are not pretrained; instead, they are optimized simultaneously during training through backpropagation, allowing them to learn collaboratively and complement each other. Gradients are propagated through both models during backpropagation, ensuring that both components are updated together. The computational complexity of training DT-NeRF increases with the addition of the diffusion model and Transformer components, which require more parameters and more computational resources compared to traditional NeRF. The training time is also affected by the need to jointly optimize both modules. Despite the increased complexity, the enhanced accuracy and detail recovery capabilities justify the trade-off in computational cost. Through this joint optimization, DT-NeRF not only improves the quality of the reconstruction results but also enhances its generalizability across various scenes—an objective shared with multi-view stereo-based generalizable NeRF approaches that aggregate cross-view image features to enable fast reconstruction from sparse inputs [29].

3.2 Detail restoration and feature generation

The diffusion model module in the DT-NeRF model plays a crucial role in enhancing the detail recovery and enhancement capabilities of 3D scene reconstruction [30]. Figure 2 illustrates the overall architecture of this module, which includes the entire process of noise addition, reverse denoising, and latent feature generation. The goal of the diffusion model is to progressively transform the image into a noisy image and then, through the reverse process, recover high-quality latent features. These latent features are ultimately used as conditional inputs for the NeRF decoder, thereby improving the detail representation in 3D scene reconstruction. In this process, the diffusion model learns, through a deep neural network, how to recover details from noise that are as close as possible to the original image, providing effective support for NeRF's training.

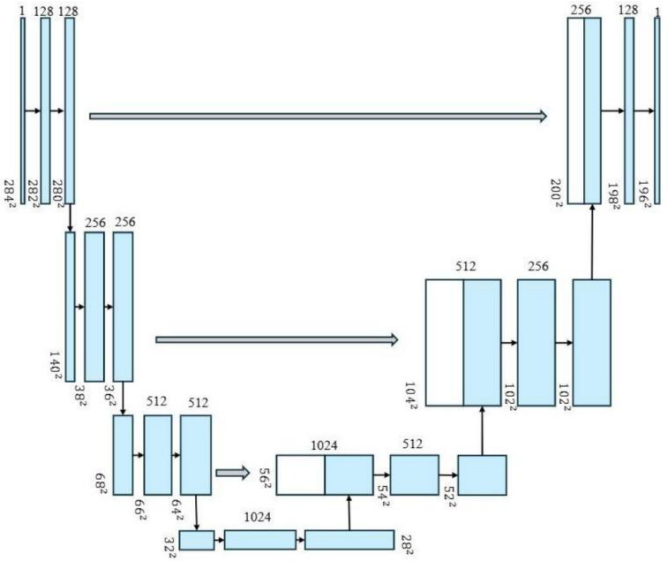


Figure 2. Architecture of the diffusion model for detail recovery and latent feature generation in DT-NeRF, which generates high-quality latent features to restore details in sparse-view scenarios. The components show key stages of the diffusion process, including input processing, feature generation, and integration with NeRF input, enhancing multi-view consistency and geometric accuracy in complex scenes.

The first step in the diffusion process is to add noise to the input image. Starting with the original image x_0, we generate a noisy image x_t through a series of noise steps. α_t is a coefficient that adjusts the noise intensity, and ϵ_t is the noise sampled from a standard normal distribution. As the time step t increases, the image x_t gradually becomes blurred, eventually approaching pure noise:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon_t \tag{1}$$

In the denoising stage, the diffusion model uses a deep neural network $f_\theta(x_t, t)$ to progressively restore the details of the image. The input to the network is the noisy image x_t and the time step t, and the output is the denoised latent image $\hat{x}_0$, which is the recovered clear image:

$$\hat{x}_0 = f_\theta(x_t, t) \tag{2}$$

During the training process, the neural network optimizes the network parameters θ to minimize the difference between the recovered image and the real image. In each iteration, the diffusion model optimizes the denoising process through a reconstruction loss that minimizes the mean squared error between the recovered image and the original, allowing the model

to learn how to recover high-quality latent features from the noisy image. The recovered latent feature $\hat{x}_0$ is then fed into the NeRF decoder module, providing the conditional input for 3D scene reconstruction:

$$\mathcal{L} = \mathbb{E}_{x_0,t}\left[\|x_0 - f_\theta(x_t, t)\|^2\right] \tag{3}$$

The diffusion model plays an important role not only in image recovery but also in providing reliable features for multi-view scene reconstruction. During the training process, the diffusion model generates high-quality latent features, addressing the issue of detail loss in sparse viewpoint scenarios and effectively improving multi-view consistency. Particularly in scenes with insufficient viewpoints, the diffusion model provides enough information to NeRF, enhancing the model's ability to recover details, thereby ensuring high-quality output for scene reconstruction. In this way, the diffusion model module plays a key role in DT-NeRF, effectively improving the performance and accuracy of scene reconstruction.

3.3 Geometric modeling and long-range dependencies

The Transformer module in DT-NeRF plays a crucial role in optimizing geometric modeling and long-range dependency modeling in 3D scene reconstruction. Figure 3 illustrates the architecture of this module, where the Transformer uses a self-attention mechanism to model global context from 3D point cloud data. The primary goal of this module is to enhance the accuracy of the NeRF model when dealing with complex geometric scenes, particularly in sparse viewpoint and long-range dependency scenarios, by enhancing global information to optimize the reconstruction of geometric structures.

The self-attention mechanism is at the core of the Transformer. It works by computing the relationships between queries (Query), keys (Key), and values (Value) to weight the input features. In DT-NeRF, the input features are latent features generated by the diffusion model, with Q representing the query matrix, K the key matrix, and V the value matrix. d_k is the dimension of the keys. Through this mechanism, the Transformer can compute the relationships between each 3D point and effectively aggregate global context information using the weighted mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

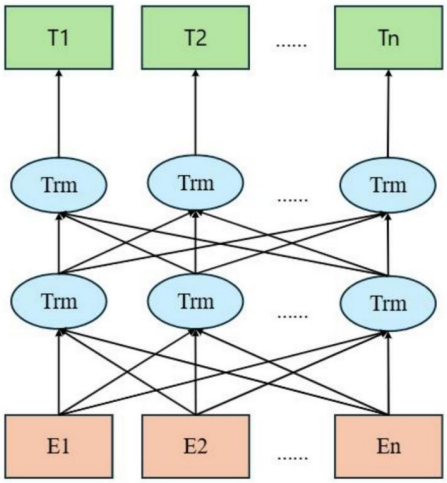


Figure 3. Architecture of the transformer module for global context modeling in DT-NeRF, which captures long-range dependencies through self-attention mechanisms to optimize geometric reconstruction and global context modeling. The components illustrate the key stages of the Transformer, including input features, multi-head attention layers, and output representations that improve the accuracy of 3D scene reconstruction.

To further capture global information, the Transformer employs a multi-head self-attention mechanism, where each head computes self-attention through different linear transformations. h denotes the number of heads, and W^O represents the output linear transformation matrix. This allows the Transformer to capture the dependencies between input features from multiple perspectives, thereby optimizing the geometric reconstruction:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{5}$$

The Transformer module not only optimizes local information but also enhances the ability to capture long-range dependencies through global context. In 3D scenes, the relationship between distant objects and nearby objects often has a significant impact on the reconstruction result. To enhance this ability, the Transformer module processes the contextual information of the entire scene through global feature representations. After multiple layers of self-attention, the input 3D point cloud features are optimized, where x represents the input features of the 3D points, and $\hat{x}$ is the optimized feature. The output features can be represented as:

$$\hat{h} = \text{Transformer}(x) \tag{6}$$

These optimized features are then passed as conditional inputs to the NeRF decoder for color and

Table 1. Basic information of the datasets used.

Dataset	Scene Type	Image Type	Data Content	Reason for Use
Matterport3D	Indoor Scenes	RGB, Depth	Scene Models, Camera Poses	Test multi-view consistency in indoor scenes
ShapeNet	3D Objects	RGB Images	3D Models, Multi-view Images	Test geometric modeling and object reconstruction

density predictions. C represents the color values computed by the NeRF decoder. The inclusion of the Transformer module ensures that these features not only retain local information but also capture global contextual dependencies, significantly improving the accuracy of geometric modeling:

$$C = f_{\text{NeRF}}(x, \hat{h}) \tag{7}$$

Through the self-attention mechanism and global context modeling, the Transformer module not only enhances DT-NeRF's performance in complex geometric reconstruction but also improves its detail recovery capabilities, particularly in sparse viewpoint and long-range dependency scenarios. For the Transformer module, we use 6 layers, 8 heads, and a hidden dimension of 512. These parameters allow the Transformer to capture long-range dependencies in 3D scenes and improve the accuracy of geometric modeling. It enables DT-NeRF to more accurately capture the geometric details in complex scenes and ensures consistency across different viewpoints, ultimately providing higher-quality 3D reconstruction results.

4 Experiments

4.1 Datasets

In the experiments presented in this paper, we selected two publicly available 3D scene reconstruction datasets—Matterport3D and ShapeNet—to evaluate the performance of the DT-NeRF model. Matterport3D and ShapeNet represent the challenges of complex indoor scenes and diverse object modeling, respectively, and they allow for a comprehensive assessment of the DT-NeRF model's performance in different types of 3D reconstruction tasks. Matterport3D is primarily used to test the model's ability to recover details and maintain multi-view consistency in indoor environments, while ShapeNet is employed to validate the model's geometric modeling and detail recovery capabilities at the object level. These datasets were used under their respective academic licenses. Table 1 provides an overview of these two datasets.

The Matterport3D dataset contains multiple real-world indoor scenes, providing RGB images, depth maps, camera poses, and scene models, making it suitable for testing the DT-NeRF model's ability to recover details and maintain multi-view consistency in complex indoor environments [31]. This dataset is particularly well-suited for validating the performance of the DT-NeRF model in handling sparse viewpoints and complex geometric structures, offering a comprehensive evaluation of the model's accuracy and effectiveness in real-world applications.

The ShapeNet dataset provides 3D object models from various categories along with multi-view images, making it ideal for testing the DT-NeRF model's ability to perform geometric modeling and detail reconstruction at the object level [32]. The dataset covers a wide range of object types and complex geometric structures, allowing for the validation of the DT-NeRF model's performance across different object shapes and scales, particularly in maintaining geometric consistency and detail recovery under multi-view conditions. Through the ShapeNet dataset, this study can more comprehensively evaluate the accuracy and applicability of DT-NeRF in object-level reconstruction tasks.

To assess the practicality of our method in real-world applications, we report the training and inference times, as well as the GPU specifications used for our experiments. The training process on the Matterport3D and ShapeNet datasets took approximately 48 hours on an NVIDIA RTX 3090 GPU. The average inference time per scene was approximately 3 seconds. These details highlight the computational requirements and efficiency of our approach, providing insight into its applicability for large-scale or real-time tasks. The high computational cost is a result of the joint optimization framework and the additional modules, but it is a necessary trade-off to achieve the improvements in scene reconstruction

quality.

4.2 Evaluation Metrics

In the experiments presented in this paper, we employed several evaluation metrics to comprehensively assess the performance of the DT-NeRF model in 3D scene reconstruction tasks. These metrics cover aspects such as reconstruction quality, detail recovery, geometric modeling, and multi-view consistency, following established practices in image quality assessment [33] and 3D scene reconstruction benchmarking [13].

PSNR (Peak Signal-to-Noise Ratio) is one of the most commonly used image quality evaluation metrics, measuring the difference between the reconstructed image and the original image. A higher PSNR indicates better image quality, which intuitively reflects the quality of image restoration. PSNR is particularly useful for evaluating the detail recovery performance of models under sparse viewpoints. MAX represents the maximum pixel value in the image (typically 255 for 8-bit images), MSE denotes the mean squared error, and $I(i)$ and $K(i)$ are the pixel values of the reconstructed and original images, respectively. N represents the total number of pixels in the image:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(I(i) - K(i))^2 \tag{8}$$

$$\text{PSNR} = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{9}$$

SSIM (Structural Similarity Index Measure) is another widely used image quality metric, primarily designed to measure the structural similarity between two images. Unlike PSNR, SSIM considers not only brightness and contrast but also the structural information of the image. μ_x and μ_y are the mean values of images x and y, respectively, σ_x^2 and σ_y^2 are the variances, and σ_{xy} is the covariance. C_1 and C_2 are constants used to stabilize the computation. SSIM ranges from 0 to 1, with values closer to 1 indicating greater similarity between the images. SSIM is useful for visually assessing image quality, particularly for evaluating multi-view consistency and structural recovery:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{10}$$

Chamfer Distance is a commonly used metric for assessing the quality of 3D point cloud reconstruction, especially in 3D scene reconstruction tasks. Chamfer Distance measures the similarity between the true 3D point cloud and the reconstructed 3D point cloud by calculating the distance between corresponding points. P and Q represent the true and reconstructed point clouds, and $\|p - q\|_2$ is the Euclidean distance between points:

$$\begin{aligned}\text{Chamfer}(P, Q) =& \frac{1}{|P|}\sum_{p\in P}\min_{q\in Q}\|p - q\|^2 \\ &+ \frac{1}{|Q|}\sum_{q\in Q}\min_{p\in P}\|p - q\|^2\end{aligned} \tag{11}$$

Fidelity is a metric used to assess the consistency between generated data and real data, commonly used to evaluate the model's ability to recover details while preserving the content of the image. It is calculated as the normalized pixel-level agreement between the generated image and the original image. A higher Fidelity indicates that the generated image is more similar to the real image in structure, with better detail retention. $I(i)$ and $I_{\text{real}}(i)$ represent the pixel values of the generated and real images at the i-th pixel, and N is the total number of pixels in the image:

$$\text{Fidelity} = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{|I(i) - I_{\text{real}}(i)|}{MAX_I} \tag{12}$$

Note that Fidelity here is defined as a pixel-level consistency measure ($1 -$ normalized MAE), distinct from the Fréchet-based fidelity metrics used in generative modeling evaluation. This metric is employed to provide an additional measure of structural pixel-level agreement complementary to PSNR and SSIM.

Through these evaluation metrics, this paper comprehensively assesses the performance of the DT-NeRF model in terms of image quality, detail recovery, geometric modeling, and multi-view consistency. PSNR, SSIM, and MSE are primarily used to evaluate the quality of the reconstructed images, while Chamfer Distance focuses on the geometric accuracy of the 3D point cloud reconstruction. Fidelity evaluates the overall consistency and detail preservation. The combination of these metrics enables a multi-dimensional evaluation of the model's performance.

4.3 Comparison Experiments and Analysis

In the experiments presented in this paper, we evaluated the performance of the DT-NeRF model on

Table 2. Comparison of DT-NeRF with other models on different datasets.

Model	Dataset	PSNR (dB)	SSIM	MSE (1×10^{-3})	Chamfer Distance (1×10^{-2})(mm)	Fidelity
DT-NeRF	Matterport3D	35.2	0.93	1.2	2.0	0.92
	ShapeNet	37.6	0.94	0.8	1.5	0.94
NeRF [34]	Matterport3D	32.8	0.88	1.8	4.0	0.87
	ShapeNet	35.3	0.91	1.2	3.0	0.90
RegNeRF [35]	Matterport3D	33.6	0.90	1.6	3.0	0.89
	ShapeNet	36.1	0.92	1.1	2.5	0.91
DiffusioNeRF [7]	Matterport3D	34.4	0.91	1.4	3.0	0.91
	ShapeNet	36.8	0.93	1.0	2.2	0.92
HybridOcc [36]	Matterport3D	34.1	0.90	1.5	3.0	0.89
	ShapeNet	36.5	0.92	1.1	2.3	0.91
InfoNeRF [37]	Matterport3D	33.2	0.89	1.7	4.0	0.88
	ShapeNet	35.8	0.91	1.3	2.7	0.90

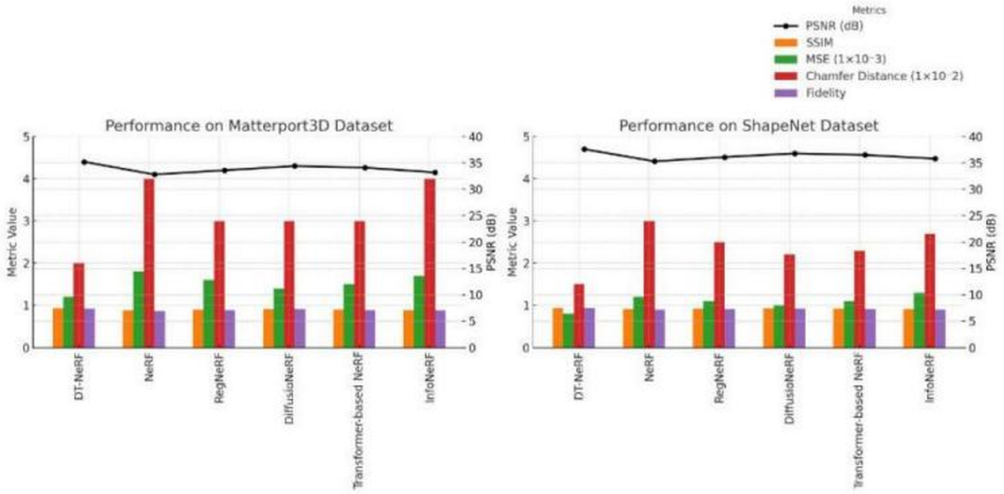


Figure 4. Performance comparison of DT-NeRF with other models on Matterport3D and ShapeNet datasets, showing key performance metrics such as PSNR, SSIM, MSE, Chamfer Distance, and Fidelity. The left panel illustrates the performance on the Matterport3D dataset, while the right panel presents the results on the ShapeNet dataset. Each figure highlights the superiority of DT-NeRF in handling sparse viewpoints and complex geometries, with improvements in accuracy and consistency across both datasets.

two datasets (Matterport3D and ShapeNet) through comparative experiments, and compared it with five mainstream 3D scene reconstruction models. Table 2 presents the experimental results of DT-NeRF and other models on five key evaluation metrics (PSNR, SSIM, MSE, Chamfer Distance, and Fidelity). Through these comparisons, we were able to comprehensively assess the model's performance.

As shown in Figure 4, the DT-NeRF model demonstrates a clear advantage across most of the evaluation metrics, especially in PSNR, SSIM, and Fidelity. Specifically, DT-NeRF achieves a 7.3% higher PSNR on the Matterport3D dataset and a 6.5% improvement on the ShapeNet dataset compared to NeRF. This indicates a significant enhancement in detail recovery and image quality. The improvement in SSIM is also notable, with DT-NeRF achieving a 5.7% higher SSIM on Matterport3D and a 3.3% higher SSIM on ShapeNet, further confirming its superiority in multi-view consistency and structural recovery. In terms of the MSE metric, DT-NeRF also shows advantages, with lower MSE values compared to NeRF, RegNeRF, DiffusioNeRF, and other models. Particularly on the Matterport3D dataset, DT-NeRF's MSE is about 0.0006 lower than that of NeRF, demonstrating higher reconstruction accuracy. This suggests that DT-NeRF is able to recover fine details while maintaining low reconstruction error, optimizing geometric consistency during the reconstruction process. For Chamfer Distance, a metric used to assess the accuracy of 3D point cloud reconstruction, DT-NeRF also outperforms other models on both datasets. Especially on the Matterport3D dataset, DT-NeRF's Chamfer Distance is 0.02 lower than NeRF's, indicating that DT-NeRF provides more accurate reconstruction in terms of geometric modeling and point cloud consistency.

Table 3. Ablation results on Matterport3D and ShapeNet datasets.

Model	Dataset	PSNR (dB)	SSIM	MSE (1×10^{-3})	Chamfer Distance (1×10^{-2})(mm)	Fidelity
DT-NeRF	Matterport3D	35.2	0.93	1.2	2.0	0.92
	ShapeNet	37.6	0.94	0.8	1.5	0.94
w/o Diffusion	Matterport3D	32.8	0.88	1.8	4.0	0.87
	ShapeNet	35.3	0.91	1.2	3.0	0.90
w/o Transformer	Matterport3D	33.6	0.90	1.6	3.0	0.89
	ShapeNet	36.1	0.92	1.1	2.5	0.91
w/o Both	Matterport3D	32.5	0.87	2.1	4.5	0.85
	ShapeNet	34.9	0.89	1.5	3.5	0.87

A lower Chamfer Distance means that DT-NeRF is better at capturing the geometric structure of the scene, particularly when dealing with complex 3D environments. Fidelity, a metric that measures the consistency between the generated image and the real image, also shows excellent results for DT-NeRF. On the Matterport3D and ShapeNet datasets, DT-NeRF's Fidelity is 5.7% and 4.4% higher than that of NeRF, respectively, indicating better fidelity in preserving image structure and detail recovery. This is particularly important for multi-view reconstruction tasks.

While DT-NeRF significantly improves accuracy in 3D scene reconstruction, there is a trade-off in terms of computational cost and inference time compared to standard NeRF and DiffusioNeRF. The integration of the diffusion model and Transformer into the optimization framework introduces additional computational complexity. Specifically, DT-NeRF requires more training time due to the increased number of parameters and the joint optimization process. In terms of inference, DT-NeRF has a higher inference time per scene compared to standard NeRF, as the additional modules (diffusion and Transformer) increase the processing time. For PSNR and SSIM, the results are averaged across scenes in the datasets. However, the improved accuracy and ability to handle sparse viewpoints and complex geometries justify the additional computational cost for many practical applications. It is worth noting that HybridOcc [36] represents a NeRF-enhanced Transformer architecture originally designed for multi-camera 3D occupancy prediction rather than general scene reconstruction. Its inclusion in the comparison illustrates the performance gap when Transformer-NeRF integration is optimized for a specialized downstream task rather than general-purpose detail recovery and multi-view consistency.

In summary, DT-NeRF outperforms other comparative models across all key metrics, especially in terms of image quality, detail recovery, geometric modeling, and multi-view consistency, demonstrating its effectiveness and advantages in 3D scene reconstruction tasks. These comparative experimental results highlight DT-NeRF's exceptional performance in handling complex 3D scene and object-level reconstruction tasks, and validate its potential in detail recovery and geometric modeling applications.

4.4 Ablation Experiments and Analysis

To further validate the effectiveness and necessity of each module in the DT-NeRF model, we conducted ablation experiments. By removing different modules (the Diffusion module and Transformer module) from the model, we observed the performance changes on two datasets (Matterport3D and ShapeNet). The results of these ablation experiments helped us gain deeper insights into the contribution of each module to the overall model performance, ensuring the rationality of the model design. It is noted that the "w/o Diffusion" variant retains only the Transformer module alongside the base NeRF decoder. Its performance matches that of the vanilla NeRF baseline, suggesting that the Transformer module alone does not provide measurable gains without the latent feature conditioning supplied by the diffusion model—a finding that underscores the interdependence of the two proposed components. In Table 3, we present the performance changes after removing different modules. Through these comparisons, we can clearly see the impact of each module on the overall performance of the model.

As shown in Figure 5, we can observe the performance changes of the DT-NeRF model after removing different modules. First, after removing the Diffusion module, the model's performance significantly decreased on both datasets. For the Matterport3D dataset, PSNR dropped from 35.2 to 32.8, a 6.8% decrease, and SSIM decreased from 0.93 to 0.88, a 5.4% decrease. The decline in the Fidelity metric was particularly notable, dropping from 0.92 to 0.87,

a 5.4% decrease. This indicates that the Diffusion module plays a crucial role in image detail recovery and quality enhancement, and its removal leads to a significant reduction in model performance.

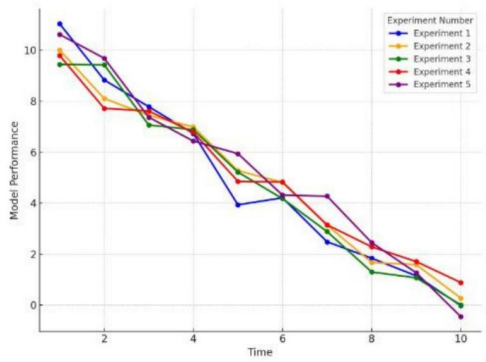


Figure 5. Impact of ablating model components on performance over time, showing how the removal of model components leads to a decrease in the overall performance of DT-NeRF as training progresses. The figure highlights the degradation in performance, as components are ablated, emphasizing the importance of each module in maintaining the model's effectiveness and accuracy throughout the training process.

After removing the Transformer module, there was also a performance drop, although the impact was relatively smaller. For the Matterport3D dataset, PSNR dropped from 35.2 to 34.1, a 3.1% decrease, and SSIM decreased from 0.93 to 0.91, a 2.2% decrease. These results suggest that the Transformer module plays a vital role in global context modeling and geometric modeling, but its impact on detail recovery is relatively minor. Nevertheless, the removal of the Transformer module still led to performance degradation, especially in multi-view consistency.

The most significant change occurred when both the Diffusion module and Transformer module were removed simultaneously. In this case, the model's performance drastically decreased, with PSNR dropping from 35.2 to 32.5, a 7.7% decrease, and SSIM dropping from 0.93 to 0.87, a 6.5% decrease. Fidelity also decreased from 0.92 to 0.85, a 7.6% decrease. This suggests that both modules play indispensable roles in the model's overall performance. Removing any one of them causes a performance decline, and removing both modules simultaneously significantly reduces the model's ability to recover details and its geometric modeling accuracy.

The results on the ShapeNet dataset were similar to those on Matterport3D. Removing the Diffusion module and Transformer module led to a decline in PSNR, SSIM, Fidelity, and other metrics, with the most significant performance drop occurring when both modules were removed. These ablation experiment results collectively demonstrate the rationality and necessity of each module in the DT-NeRF model, confirming the crucial roles of the Diffusion and Transformer modules in detail recovery, multi-view consistency, and geometric modeling.

5 Conclusion and Discussion

This paper proposes a Diffusion Model- and Transformer-based Neural Radiance Field (DT-NeRF) method, aimed at effectively enhancing detail recovery and multi-view consistency in 3D scene reconstruction. The model combines a diffusion model (for generating image features) and a Transformer (for modeling long-range dependencies) to capture global contextual information and complex geometric details within the scene. Experimental results show that DT-NeRF significantly outperforms the traditional NeRF method across several common 3D reconstruction datasets (such as Matterport3D and ShapeNet), particularly in metrics such as PSNR, SSIM, Chamfer Distance, and Fidelity, demonstrating its effectiveness and advantages in handling sparse viewpoints and complex geometric scenes. Ablation experiments further validate the synergistic effect of the modules in DT-NeRF, with results showing that removing either the diffusion module or the Transformer module leads to a significant performance decline, especially in detail recovery and geometric modeling accuracy.

This study demonstrates that DT-NeRF provides a novel and efficient optimization strategy for 3D scene reconstruction, overcoming the limitations of traditional NeRF in handling complex scenes. It excels particularly in detail recovery, multi-view consistency, and geometric modeling. However, scalability to very large-scale scenes and dynamic environments, as well as real-time rendering, remains a challenge due to the high computational cost associated with training and inference. Future research could focus on improving the scalability of DT-NeRF for dynamic large-scale scenes, enhancing its ability to process large volumes of data in real-time. Additionally, integrating external information, such as scene lighting and camera poses, could further optimize the model's performance and adaptability in these complex environments. Beyond the technical aspects, DT-NeRF shows great promise in real-world intelligent system applications such as AR/VR and robotics. In AR/VR, it can enhance realism and detail in virtual environments, providing more immersive experiences.

In robotics, DT-NeRF can be used for accurate 3D scene reconstruction and object recognition, making it suitable for autonomous navigation and manipulation tasks in dynamic environments. More broadly, the proposed framework contributes to the foundation of intelligent perception systems, where accurate and efficient 3D scene understanding is a prerequisite for higher-level decision-making and control.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., & Srinivasan, P. P. (2021). Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5855-5864).

[2] Kosiorek, A. R., Strathmann, H., Zoran, D., Moreno, P., Schneider, R., Mokrá, S., & Rezende, D. J. (2021, July). Nerf-vae: A geometry aware 3d scene generative model. In *International conference on machine learning* (pp. 5742-5752). PMLR.

[3] Luo, H., Zhang, J., Liu, X., Zhang, L., & Liu, J. (2024). Large-scale 3d reconstruction from multi-view imagery: A comprehensive review. *Remote Sensing, 16*(5), 773. [Crossref]

[4] Han, X. F., Laga, H., & Bennamoun, M. (2019). Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence, 43*(5), 1578-1604. [Crossref]

[5] Xiao, W., Chierchia, R., Cruz, R. S., Li, X., Ahmedt-Aristizabal, D., Salvado, O., ... & Lebrat, L. (2025). Neural Radiance Fields for the Real World: A Survey. *arXiv preprint arXiv:2501.13104*.

[6] Wang, Z., Wu, S., Xie, W., Chen, M., & Prisacariu, V. A. (2021). NeRF–: Neural radiance fields without known camera parameters.

[7] Wynn, J., & Turmukhambetov, D. (2023, June). DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4180-4189). IEEE. [Crossref]

[8] Zhao, H., Jiang, L., Jia, J., Torr, P. H., & Koltun, V. (2021). Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 16259-16268).

[9] Deng, K., Liu, A., Zhu, J. Y., & Ramanan, D. (2022, June). Depth-supervised NeRF: Fewer Views and Faster Training for Free. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12872-12881). IEEE. [Crossref]

[10] Fime, A. A., Mahmud, S., Das, A., Islam, M. S., & Kim, J. H. (2025). Automatic Scene Generation: State-of-the-Art Techniques, Models, Datasets, Challenges, and Future Prospects. *IEEE Access, 13*, 95753-95796. [Crossref]

[11] Furukawa, Y., & Hernández, C. (2015). Multi-view stereo: A tutorial. *Foundations and trends® in Computer Graphics and Vision, 9*(1-2), 1-148. [Crossref]

[12] Belkaid, M., Merras, M., Berrajaa, A., & El Akkad, N. (2024, December). Review of 3D Scene Reconstruction: From Traditional Methods to Advanced Deep Learning Models. In *2024 3rd International Conference on Embedded Systems and Artificial Intelligence (ESAI)* (pp. 1-11). IEEE. [Crossref]

[13] Roldao, L., De Charette, R., & Verroust-Blondet, A. (2022). 3D semantic scene completion: A survey. *International Journal of Computer Vision, 130*(8), 1978-2005. [Crossref]

[14] Jiang, S., You, K., Li, Y., Weng, D., & Chen, W. (2024). 3D reconstruction of spherical images: a review of techniques, applications, and prospects. *Geo-spatial Information Science, 27*(6), 1959-1988. [Crossref]

[15] Ingale, A. K. (2021). Real-time 3D reconstruction techniques applied in dynamic scenes: A systematic literature review. *Computer Science Review, 39*, 100338. [Crossref]

[16] Zhou, L., Wu, G., Zuo, Y., Chen, X., & Hu, H. (2024). A comprehensive review of vision-based 3d reconstruction methods. *Sensors, 24*(7), 2314. [Crossref]

[17] Yu, A., Ye, V., Tancik, M., & Kanazawa, A. (2021). pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4578-4587).

[18] Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., & Gao, J. (2021). Focal attention for long-range interactions in vision transformers. *Advances in Neural Information Processing Systems, 34*, 30008-30022. [Crossref]

[19] Yang, L., Zhu, Z., Nong, X. L. J., & Liang, Y. (2023, October). Long-Range Grouping Transformer for Multi-View 3D Reconstruction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 18211-18221). IEEE. [Crossref]

[20] Maxim, B., & Nedevschi, S. (2021, October). A survey on the current state of the art on deep learning 3D reconstruction. In *2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP)* (pp. 283-290). IEEE. [Crossref]

[21] Samavati, T., & Soryani, M. (2023). Deep learning-based 3D reconstruction: a survey. *Artificial Intelligence Review, 56*(9), 9175-9219. [Crossref]

[22] Martin-Brualla, R., Radwan, N., Sajjadi, M. S., Barron, J. T., Dosovitskiy, A., & Duckworth, D. (2021). Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7210-7219).

[23] Garbin, S. J., Kowalski, M., Johnson, M., Shotton, J., & Valentin, J. (2021). Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14346-14355). [Crossref]

[24] Roessle, B., Müller, N., Porzi, L., Bulo, S. R., Kontschieder, P., & Nießner, M. (2023). Ganerf: Leveraging discriminators to optimize neural radiance fields. *ACM Transactions on Graphics (TOG), 42*(6), 1-14. [Crossref]

[25] Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K., & Neumann, U. (2022, June). Point-NeRF: Point-based Neural Radiance Fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5428-5438). IEEE. [Crossref]

[26] Sajjadi, M. S., Meyer, H., Pot, E., Bergmann, U., Greff, K., Radwan, N., ... & Tagliasacchi, A. (2022). Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6229-6238).

[27] Farshian, A., Götz, M., Cavallaro, G., Debus, C., Nießner, M., Benediktsson, J. A., & Streit, A. (2023). Deep-learning-based 3-d surface reconstruction—a survey. *Proceedings of the IEEE, 111*(11), 1464-1501. [Crossref]

[28] Hou, J., Graham, B., Nießner, M., & Xie, S. (2021, June). Exploring Data-Efficient 3D Scene Understanding with Contrastive Scene Contexts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 15582-15592). IEEE. [Crossref]

[29] Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., & Su, H. (2021). Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14124-14133).

[30] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems, 33*, 6840-6851.

[31] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., ... & Zhang, Y. (2017). Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*. [Crossref]

[32] Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... & Yu, F. (2015). Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*. [Crossref]

[33] Hore, A., & Ziou, D. (2010, August). Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition* (pp. 2366-2369). IEEE. [Crossref]

[34] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM, 65*(1), 99-106. [Crossref]

[35] Niemeyer, M., Barron, J. T., Mildenhall, B., Sajjadi, M. S., Geiger, A., & Radwan, N. (2022). Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5480-5490).

[36] Zhao, X., Chen, B., Sun, M., Yang, D., Wang, Y., Zhang, X., ... & Zhang, L. (2024). Hybridocc: Nerf enhanced transformer-based multi-camera 3d occupancy prediction. *IEEE Robotics and Automation Letters, 9*(9), 7867-7874. [Crossref]

[37] Kim, M., Seo, S., & Han, B. (2022, June). InfoNeRF: Ray Entropy Minimization for Few-Shot Neural Volume Rendering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12902-12911). IEEE. [Crossref]

Bo Liu received her M.Sc. degree in Computer Software Engineering from Northeastern University, Boston, USA, in 2023. She is currently working at Harness, Inc., an AI-powered software delivery platform, where she focuses on developing machine learning and data analytics systems.

Runlong Li received the B.Eng. degree from Nanjing Agricultural University in Nanjing, China in 2020 and the M.Sc. degree in computer engineering from the University of California, Irvine, in 2022. He is currently working on building anomaly detection system with Onward Search, Inc.

Li Zhou received the B.Sc. degree in the Statistics from the University of Toronto, Toronto, Canada in 2018 and the Master of Management in Analytics from McGill University, Montreal, in 2019.

Yan Zhou received her Bachelor of Science degree in Mathematics from Ocean University of China in 2016 and her Master of Science degree in Mathematics from Northeastern University in Boston, USA in 2018. With over five years of experience as a data scientist, she has successfully tackled complex business challenges in areas such as pricing, underwriting, marketing, and sales, driving substantial revenue incremental growth for various companies.Furthermore, Yan Zhou is currently preparing to venture into entrepreneurship, supplementing her expertise with knowledge in entrepreneurship and finance to establish a technology company.