



Rugularizing generalizable neural radiance field with limited-view images

Wei Sun^{1,2} · Ruijia Cui¹ · Qianzhou Wang¹ · Xianguang Kong⁴ · Yanning Zhang^{2,3}

Received: 12 December 2023 / Accepted: 23 November 2024 / Published online: 19 December 2024
© The Author(s) 2024

Abstract

We present a novel learning model with attention and prior guidance for view synthesis. In contrast to previous works that focus on optimizing for specific scenes with densely captured views, our model explores a generic deep neural framework to reconstruct radiance fields from a limited number of input views. To address challenges arising from under-constrained conditions, our approach employs cost volumes for geometry-aware scene reasoning, and integrates relevant knowledge from the ray-cast space and the surrounding-view space using an attention model. Additionally, a denoising diffusion model learns a prior over scene color, facilitating regularization of the training process and enabling high-quality radiance field reconstruction. Experimental results on diverse benchmark datasets demonstrate that our approach can generalize across scenes and produce realistic view synthesis results using only three input images, surpassing the performance of previous state-of-the-art methods. Moreover, our reconstructed radiance field can be effectively optimized by fine-tuning the target scene to achieve higher quality results with reduced optimization time. The code will be released at <https://github.com/dsdefv/nerf>.

Keywords Neural radiance field · Limited views · Generalizable network · Diffusion model

Introduction

Synthesis of novel views is a fundamental and challenging task within the fields of computer vision and graphics [1]. The objective is to produce realistic images of a given scene from new viewpoints. This process necessitates the capacity to comprehend the three-dimensional layout and geometric interconnections of the scene, and to precisely predict and render an image from a previously unseen perspective [2, 3].

Early researches focus on modelling 3D shapes by discrete geometric 3D representations, such as mesh surface [4], point cloud [5] and voxel grid [6]. Although explicit 3D geometry-based representations are intuitive, they encounter challenges related to the need for explicit proxy geometry or dense input views. Under-constraint conditions is difficult to achieve estimation with high precision, resulting in rendering artifacts. A recent and promising approach, known as neural scene representations [7, 8], has demonstrated the potential to generate new views by representing scenes using neural network weights. This research area has made notable progress, particularly through the application of neural radiance field (NeRF) [9]. NeRF utilizes multi-layer perceptron (MLP) and positional encodings to learn a continuous 5D radiance field

✉ Wei Sun
sunwei052@gmail.com

Ruijia Cui
crj732575@163.com

Qianzhou Wang
wangqianzhou413@gmail.com

Xianguang Kong
kongxianguang@xidian.edu

Yanning Zhang
ynzhang@nwpu.edu.cn

¹ School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, China

² National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Xi'an, China

³ School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China

⁴ School of Mechano-Electronic Engineering, Xidian University, Xi'an, China

for a specific scene, enabling differentiable rendering from arbitrary viewpoints.

Neural scene representations, such as NeRF, have shown significant promise in accurately predicting and rendering novel views by leveraging image relationships within large datasets [10]. However, these models require the optimization of a scene-specific 3D representation for each individual scene, which is computationally expensive and fails to capture shared information across different scenes. Furthermore, NeRF often produces unrealistic geometries and low-quality surface appearances when relying on a limited number of input views, resulting in noticeable artifacts. To address these challenges, various hand-crafted regularizers and learned priors have been proposed. These include hand-engineered priors to constrain scene geometry [11], learned priors designed to ensure plausible renderings from arbitrary viewpoints [12], and approaches leveraging single-image depth and normal estimation to impose high-level constraints on scene geometry [13]. Additionally, approaches like ColNeRF [14] integrate collaborative compensation and constraints into NeRF, significantly improving its generalization capabilities. Despite these advancements, no existing methods explicitly learn a joint probability distribution of scene geometry and color. Moreover, inaccuracies persist due to inconsistencies in the mapping between 2D pixels in the source images and the corresponding queried 3D locations. As a result, these methodologies often encounter modeling inaccuracies, struggling to maintain geometric and visual coherence across diverse viewpoints. This underscores the critical importance of adequately accounting for inter-view correlations and collaborative interactions.

To address these issues, we present a practical approach to neural scene reconstruction and rendering that focuses on efficient radiation field estimation. The proposed network incorporates two core designs tailored for learning a general continuous 3D scene representation from a limited number of observed views. The first key element involves encoding scene information and utilizing key knowledge to aid density value and color estimation. Our designs build on the transformer-based framework and utilize the attention mechanism to integrate deep relationships between the rendering view and observed views. By leveraging the attention mechanism, we consider adjacent 3D points on the target ray and multiple views simultaneously, enhancing the local consistency of a 3D scene representation in both the ray-cast space and the surrounding view space. Consequently, we can learn a more comprehensive neural radiance field from limited-view images and effectively generalize to novel scenes without per-scene fine-tuning.

The second essential aspect under consideration is the generation of high-quality geometry and appearance under a limited number of input views. In contrast to prevailing methods in radiance field reconstruction [16, 17], which pri-

marily rely on constraining the estimation process with an $\mathcal{L}_2$ loss, our approach integrates a denoising diffusion model as a learned prior to promote the fitting of color and density fields by NeRFs. By incorporating the denoising diffusion model into the rendering pipeline, our proposed model effectively addresses the issue of low-quality geometry and appearance in scenarios with limited views, thereby improving rendering results with enhanced fidelity and visual appeal. Furthermore, our reconstructed radiance fields can be easily optimized through fine-tuning the estimation on the target scene, requiring less optimization time and leading to photo-realistic results that are comparable to or even better than those achieved through per-scene overfit NeRFs (see Fig. 1).

In brief, the main contributions of our work are summarized as follows:

- We present an innovative and feasible approach for reconstructing and rendering neural scenes by combining information from multiple viewpoints and leveraging prior knowledge. The use of cross-attention and relevant knowledge exploitation mechanisms enables the matching of point densities and colors across different views, thereby enhancing the precision of prediction.
- We integrate a denoising diffusion model as a regularization term to reduce noise and enhance the smoothness of the radiance field. This approach allows us to tackle the challenges associated with low-quality geometry and appearance in scenarios with limited-view images.
- We conduct experiments on various widely used benchmark datasets and show that the proposed method is comparable to existing methods both qualitatively and quantitatively, especially for realistic view synthesis when there are only limited input views.

Related work

In recent years, there has been considerable research and attention devoted to the complex issue of novel view synthesis in the field of computer vision. This challenge remains persistent due to the need to accommodate a wide range of observed views. Recent endeavours have focused on addressing novel view synthesis, with existing solutions categorized into image-based rendering and neural scene representation.

Image-based rendering aims to synthesize realistic rendered images from input views. Early research in this field focuses on synthesizing novel views through a weighted blending of reference images, with blending weights determined by approximate proxy geometry [18] or ray-space proximity [19]. In recent years, the advancement of deep learning has significantly contributed to image-based rendering. Researchers employ neural networks to estimate optical

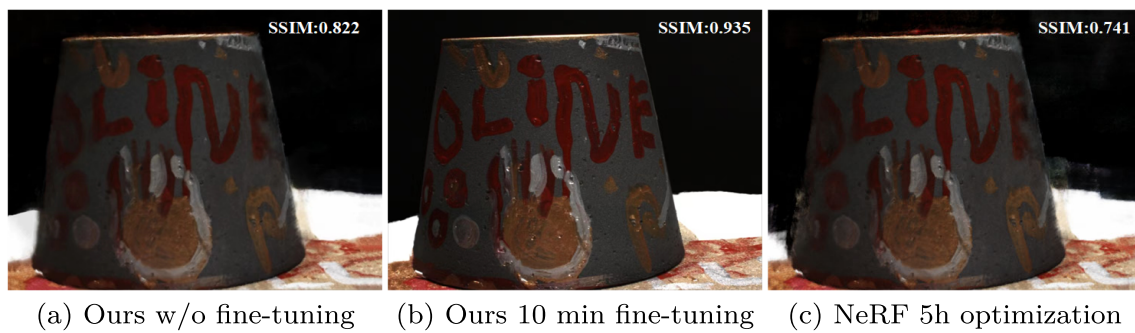


Fig. 1 Our network is trained using scenes from the DTU dataset [15] and demonstrates the ability to generalize effectively across a wide range of scenes with just three input view images (a). Furthermore, significant improvements in performance can be achieved by fine-tuning our

estimation on specific target scenes for a mere 10 min (b), surpassing the quality achieved by NeRF [12] with 5 h of per-scene optimization (c)

flow [20, 21], compute proxy geometry [22–25], and soft blending [26, 27]. While these methods can handle sparser views and have shown promising results in some cases, they are constrained by the performance of 3D reconstruction algorithms [28, 29]. They struggle in low-textured or reflective areas and cannot handle partially translucent surfaces. In contrast, our approach involves learning continuous volume densities and colors in an end-to-end manner and utilizes a denoising diffusion model to regulate neural radiance fields for improved rendering quality.

Neural scene representation focuses on utilizing neural network models to learn and depict the structural and visual information of 3D scenes. Recent developments in this area have made significant progress in representing the shape and appearance of scenes using neural networks. Previous approaches [10, 30, 31] to neural scene representation primarily utilize MLPs as implicit shape representations, with the weights of the MLPs establishing a mapping between continuous spatial coordinates and either signed distance or occupancy values. Advancements in rendering techniques [24, 32] have led to a shift in research towards learning the spatial structure and visual characteristics of scenes through the analysis of multiple viewpoints [33–35]. For example, the NeRF model [9] trains a deep neural network to encode the density and color information of a scene into a volume, yielding impressive results for novel view synthesis. However, NeRF focuses on optimizing the representation of each scene individually, without considering potential shared information among scenes, which can be time-consuming and inefficient. Other concurrent works [17, 36] attempt to address this issue and generalize NeRF, but the pooling-based fusion models employed in these methods have limited capability to effectively analyze the intricate interconnections among multiple perspectives, hindering comprehensive comprehension of 3D scenes. While our approach utilizes knowledge from multiple viewpoints and priors to regularize the rendering process, reducing artifacts and inconsistencies.

Our method

In this section, we first present the proposed architecture for novel view synthesis in “[Overview](#)”. Subsequently, we discuss the cost volume construction module in “[Cost volume construction](#)”. Following this, we provide a detailed explanation of the process for predicting density and color in “[Density and color prediction](#)”. Lastly, we focus on our design of rendering details in “[Rendering](#)”.

Overview

The proposed rendering network with attention and prior guidance design, depicted in Fig. 2, consists of three main modules: a cost volume construction module, a radiance field reconstruction module and a volume rendering module. Given N input views $I_i (i = 1, \dots, N)$ of a real scene and their corresponding camera parameters ϕ_i , the network initially projects the 2D image features from the input images captured from different viewpoints onto a planar scanning volume in the reference view. Subsequently, it constructs a cost volume at the reference view by warping the features from nearby input views onto sweeping planes in the reference view’s frustum, similar to [37], where each voxel contains information about feature variation at a specific depth.

In the module for reconstructing radiance fields, we employ a 3D Convolutional Neural Network (CNN) to generate a volumetric representation of a neural scene, containing neural features at each voxel that capture details about the local scene’s geometry and visual characteristics. We then employ cross-attention and relevant knowledge exploitation mechanisms to decode the volume density and radiance at arbitrary continuous locations. Unlike traditional methods that predict density and color independently, we utilize an attention mechanism to aggregate all sample information along the ray path and views, and allocate specific weights

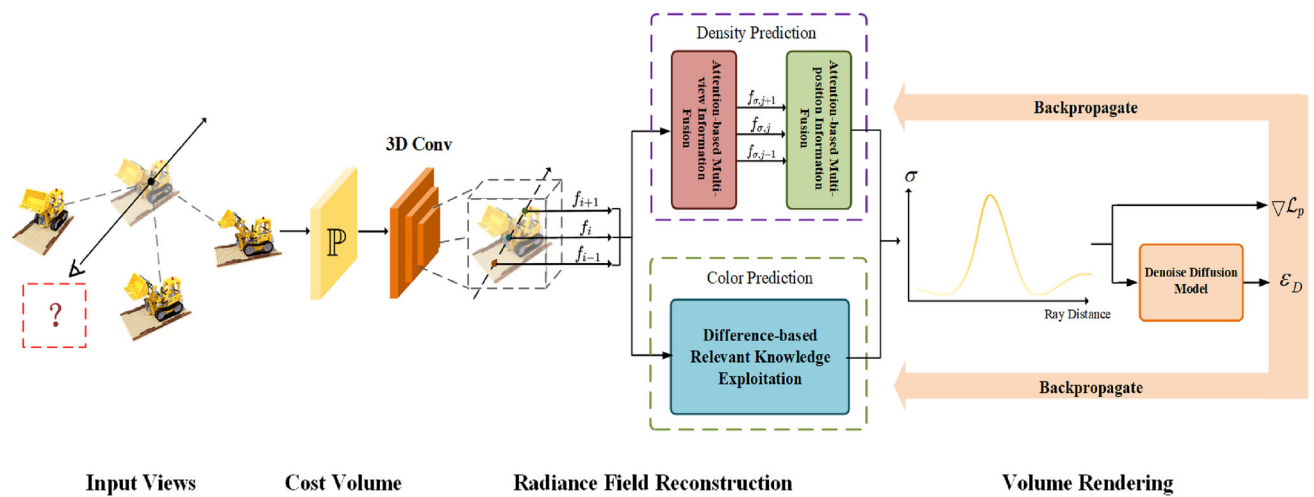


Fig. 2 The architecture of our whole network. Given input views, we first construct a cost volume. Subsequently, instead of processing each 3D point independently, we combine attention mechanism and view difference to learn the neural radiance field. Finally, volumetric rendering

predicts the expected directional color for the query 3D point on a target-viewing ray. Denoising diffusion model is introduced to regularize the reconstruction process and refine the results

to each sample point. This mechanism for assigning weights enables the network to prioritize points with higher weights, thereby disregarding inaccurate or inconsequential points, and consequently enhancing the accuracy of density and color prediction. Specifically, our approach can naturally decode observed views and spatial location information for a query 3D point in the rendering ray direction into its density. We combine the correlations between directional color representations and viewing angle differences to focus on key information from surrounding views for improving color prediction.

Finally, to achieve realistic rendering, we utilize volume rendering techniques to aggregate color and density values along the rays and produce the final rendering outcomes. However, when the input views are limited, the rendering procedure may give rise to numerous artifacts, commonly referred to as “ghosting”, which are erroneous projections not corresponding to the scene objects. To tackle this challenge, we introduce a denoising diffusion model as a learning prior for color and geometry. The denoising diffusion model continuously assesses the level of noise introduced to the image during the diffusion process and gradually learns to restore the noise-free image through denoising, thereby enabling the prediction of potential scene structure and details. This results in more realistic and accurate generated outcomes. The overall time complexity of the process is determined by the number of points in the scene and the sampling count per point. Generally, the time complexity of our model can be approximated as $O(NM)$, where N denotes the total number of points in the scene, and M represents the sampling count per point.

Cost volume construction

To capture the local visual characteristics, we first utilize a deep CNN to extract high-dimensional neural features based on individual input views. This sub-network consists of down-sampling convolutional layers that convert each input view $I_i \in \mathbb{R}^{H_i \times W_i \times 3}$ into a feature map $F_i \in \mathbb{R}^{H_i/4 \times W_i/4 \times C}$. Here, H and W represent the height and width of an image, respectively, while C denotes the desired number of feature channels.

Once the features are extracted, we perform coordinate transformations to map the features from their respective view to the reference view, similar to [37]. This process entails the adjustment of camera parameters and coordinates to ensure the alignment of all features within a unified coordinate system. Given the camera’s intrinsic and extrinsic parameters $\phi = [K, R, t]$, homographic deformation is formulated as

$$Q_i(g) = K_i \left(R_i R_1^T + \frac{(t_1 - t_i) n_1^T}{n_1} \right) K_1^{-1}. \quad (1)$$

Here, $Q_i(g)$ represents the matrix distortion from view i to the reference view at depth g . K denotes the intrinsic parameter matrix of the camera. R and t represent the rotation and translation matrices of the camera, respectively. Each feature map F_i can be warped to the reference view by

$$F_{i,g}(a, b) = F_i(Q_i(g)[a, b, 1]^T), \quad (2)$$

where $F_{i,g}$ represents the distorted feature map at depth g , and the coordinate (a, b) corresponds to the pixel position in the reference view.

Following the transformation of coordinates, the converted features are mapped to their corresponding positions within the planar scanning volume. Subsequently, the disparities of characteristics for each voxel are determined by comparing the distinctions between characteristics from the reference viewpoint and those from other input viewpoints. Feature disparity vectors depict the coherence of features across various depths and are preserved in the relevant depth layers of the scanning volume. The variance-based metric is employed to calculate the cost. For each voxel centered at (a, b, g) in $\mathbb{P}$, its cost feature vector is computed as follows:

$$\mathbb{P}(a, b, g) = \text{Var}(F_{i,g}(a, b)). \quad (3)$$

Here, $\text{Var}(\cdot)$ calculates the variance of the feature map at position (a, b) across N views. By following these steps, we construct a cost volume where each voxel encapsulates data pertaining to the depth of the scene and the coherence of features across diverse views.

Density and color prediction

Neural encoding volume. For rendering tasks, relying solely on geometric information is insufficient. To achieve high-quality rendering outcomes, a thorough comprehension of the scene's visual properties, including lighting, textures, and materials, is essential. Therefore, it is imperative to deduce appearance-related information from the cost volume and integrate it into the reconstruction process. Specifically, we employ a 3D CNN $\mathcal{B}$ with a UNet [38] structure to extract significant scene geometry and appearance characteristics from the constructed cost volume $\mathbb{P}$, and represent them as a new neural feature volume $\mathcal{S}$. This process can be formulated as

$$\mathcal{S} = \mathcal{B}(\mathbb{P}). \quad (4)$$

The scene encoding volume $\mathcal{S}$ can be seen as a high-dimensional feature space, where individual channels represent distinct attributes of the scene. Through the interpolation and transformation of these features, it becomes possible to infer characteristics such as volume density and brightness dependent on viewpoint. This process empowers us to leverage these features in subsequent tasks, thereby enhancing the precision of volume rendering results.

Volume density prediction. Firstly, density estimation is deduced through the examination of the coherence among the multi-view features $\{f_i\}_{i=1}^N$ pertaining to a specific point. The density feature encompasses a comprehensive representation of the view information relevant to that point. Specifically,

in our attention-based multi-views information fusion module (see Fig. 3), we employ a structure similar to PointNet [5] to utilize multiple view features and utilize variance as a global pooling operator. The mean μ and variance ν are calculated from view features $\{f_i\}_{i=1}^N$ to capture global information. Subsequently, each view feature f_i is concatenated with global features μ and ν , and fed into an MLP to integrate local and global information. This results in a multi-view aware feature f'_i and a weight vector $\omega_i \in [0, 1]$. Ultimately, the fused features are obtained through weighted averaging and mapped to the density feature f_σ by an MLP. Consequently, our approach, based on the weighted pooling design, can more accurately utilize information from different perspectives, eliminate interfering components, and enhance the precision of density estimation.

After consolidating data from multiple perspectives to derive the density feature f_σ , prior studies traditionally utilize a basic MLP to directly transform it into a singular density value σ . However, it has been observed that this simplistic approach is inadequate for accurately forecasting the density of scenes characterized by intricate geometric configurations. We attribute this limitation to the fact that relying solely on the observations of individual point features is insufficient for accurately predicting their density. Additional contextual information and global correlations are necessary for predicting density. Therefore, we introduce an attention-based multi-position information fusion module, which allows samples along a ray to attend to each other before predicting their densities (refer to Fig. 3). When dealing with a sequence of M samples along a ray, we apply positional encoding and a multi-head self-attention mechanism [39, 40] to the density feature sequence $\{f_{\sigma,j}\}_{j=1}^M$. By considering the correlations and significance of each sample in relation to other samples, we can predict the final density value σ for each sample.

In this way, the two attention-based modules are designed to focus on distinct objectives and are intricately combined, resulting in comprehensive localization of spatial position information and enhanced data propagation and processing. The integration of global and contextual key information is thoroughly and mutually taken into account, ultimately enhancing the accuracy of density prediction.

Color prediction. To predict the color of a specific point, we utilize a technique that involves forecasting the blending weights for the colors associated with the point in the source views, as shown in Fig. 4. In contrast to the conventional NeRF approach, which relies on absolute viewing directions, we take into account the relative viewing direction in relation to each input view i . A smaller difference indicates a closer similarity between the color of the target view and the color corresponding to view i . To forecast the blending weights for each input view's color c_i , we combine the feature f'_i with

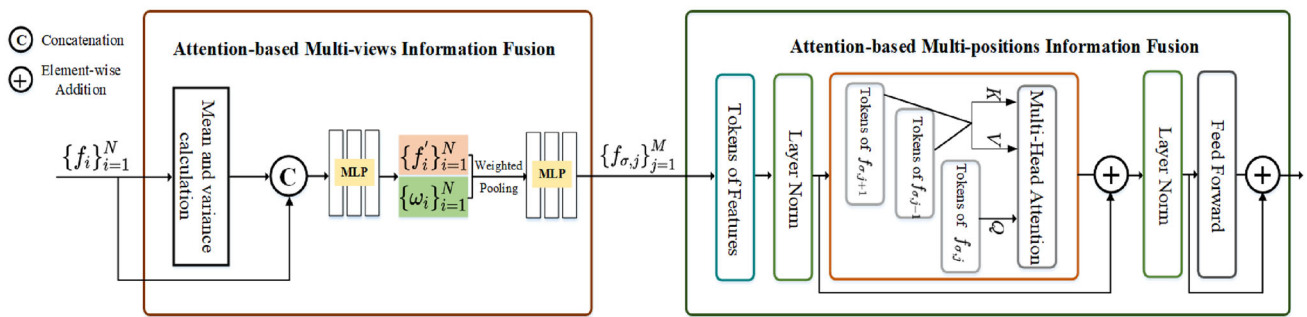


Fig. 3 The architecture of our density prediction network, which extracts correlated information from multi-views and ray positions to adapt density estimation

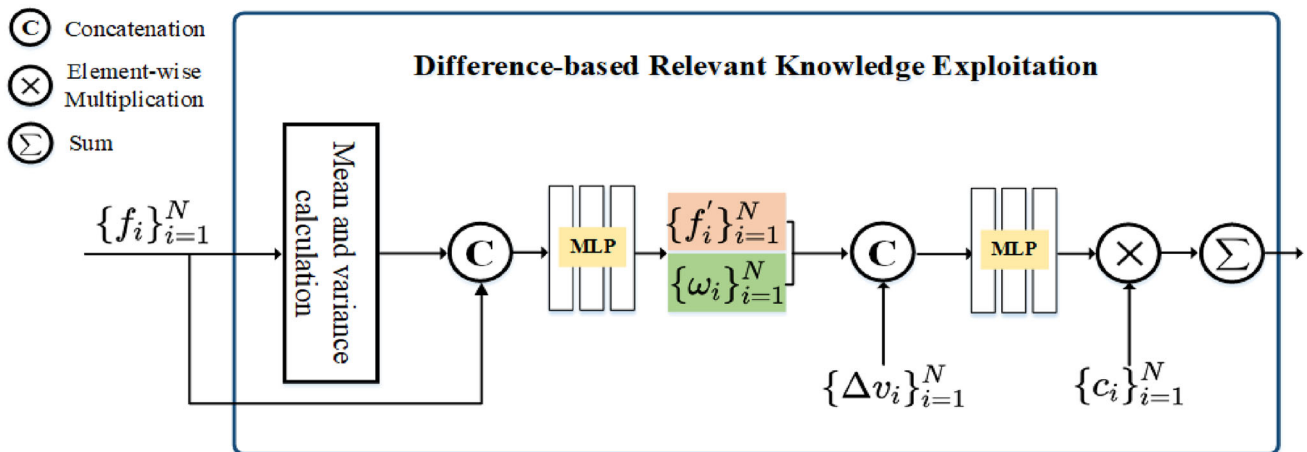


Fig. 4 The architecture of our color prediction network, which takes into account view disparity and integrates analogous colors from adjacent views to facilitate prediction

view disparity Δv_i and process the combined feature through an MLP to produce the blending weights ω_i^c . Ultimately, the color of the point is calculated as the weighted sum of the colors $c = \sum_{i=1}^N [c_i \exp(\omega_i^c) / \sum_{q=1}^N \exp(\omega_q^c)]$. By assigning different weights to the colors of the surrounding views, relevant knowledge is taken into account, thereby improving the accuracy of predictive capabilities.

Rendering

The processes described above enable us to calculate the color and density values at a specified location. In order to visualize the color of a ray r as it moves through the scene, we initially gather the color and density information from M samples along the ray. We then utilize volume rendering methods to reconstruct the image's color based on the color and density data obtained from these samples.

Volume rendering. Volume rendering involves the computation of pixel color through the accumulation of radiance along a ray. To be specific, a ray is traced from the camera position through each pixel, and sampling is conducted along the ray.

The ultimate pixel color is determined by aggregating the contribution from each sampling point, considering factors such as transmittance, radiance, and accumulated radiance at occlusion points. The rendering process along the ray r is given by:

$$\hat{c}(r) = \sum_{j=1}^M \lambda_j (1 - \exp(-\sigma_j)) c_j, \quad (5)$$

$$\text{and } \lambda_j = \exp\left(-\sum_{q=1}^{j-1} \sigma_q\right). \quad (6)$$

Here, $\hat{c}(r)$ is the final color of the point and λ denotes the volume transmittance. σ_j and c_j represent the density and color of the j -th sample on the ray r , respectively.

Denoise diffusion model. Exclusively restricting the color $\hat{c}(r)$ using an $\mathcal{L}_2$ loss may result in the production of low-quality geometry and appearance that deviate from the object surface when provided with restricted views. Therefore, we employ a denoising diffusion model [41] as a learned prior and backpropagate it to the preceding network as a form of regularization.

The forward diffusion process involves the incremental introduction of small Gaussian noise to a data sample $x_0 \sim g(x_0)$, resulting in the generation of progressively more noised versions, so

$$x_k = \sqrt{\eta_k}x_{k-1} + \sqrt{\gamma_k}\epsilon_{k-1}, \quad (7)$$

where x_k represents the data sample with added noise at time step k , while ϵ_{k-1} denotes the noise component conforming to a Gaussian distribution with mean 0 and identity covariance matrix $\mathbf{I}$. $\eta_k = 1 - \gamma_k$ and $\{\gamma_k\}_{k=1}^K$ controls the level of noise. Employing the reparameterization technique enables us to derive the following result

$$g(x_k|x_0) = \mathcal{N}(x_k; \sqrt{\eta_k}x_0, (1 - \bar{\eta}_k)\mathbf{I}). \quad (8)$$

Here, $\bar{\eta}_k = \prod_{a=0}^k \eta_a$ allows for efficient generation of noisy samples for arbitrary k . As the value of k tends towards infinity, the distribution of noisy samples x_k approaches convergence to an isotropic unit Gaussian distribution.

The objective of the denoise diffusion model is to acquire knowledge of the inverse diffusion process, which entails estimating the distribution of the preceding noisy sample x_{k-1} based on the current noisy sample x_k :

$$p(x_{k-1}|x_k) = \mathcal{N}(x_{k-1}; \mu(x_k, k), \hat{\gamma}_k\mathbf{I}), \quad (9)$$

where $p(x_{k-1}|x_k)$ is a Gaussian distribution and $\hat{\gamma}_k = (1 - \bar{\eta}_{k-1})\gamma_k / (1 - \bar{\eta}_k)$.

When provided with the input x_k , the mean $\mu(x_k, K)$ can be computed by estimating the noise term ϵ_{k-1} using the noised input:

$$\mu(x_k, k) = \frac{1}{\sqrt{\eta_k}} \left(x_k - \frac{\gamma_k}{\sqrt{1 - \bar{\eta}_k}} \varepsilon_D(x_k, k) \right), \quad (10)$$

where $\varepsilon_D(x_k, k)$ is a neural network.

Hence, the disturbance component can be approximated through the training of a neural network $\varepsilon_D(x_k, K)$ using a noised input and noise level. It is noteworthy that the noise estimator integrated with the denoise diffusion model is linked to the concept of score matching [42] and is directly related to the score function $\varepsilon_D(x_k, K) \propto \nabla_x \log p(x)$.

In accordance with the Bayes theorem, the posterior probability of density and color fields given training views $\{I_i\}_{i=1}^N$ is expressed as $p(\sigma, c) \propto p(\{I_i\}_{i=1}^N | \sigma, c) p(\sigma, c)$. The loss function is derived from the gradient of the log-posterior and formulated as

$$\nabla \log p(\sigma, c | \{I_i\}_{i=1}^N) \propto \nabla \log p(\sigma, c) + \nabla \log p(\{I_i\}_{i=1}^N | \sigma, c). \quad (11)$$

By substituting $\varepsilon_D(x_k, K) \propto \nabla_x \log p(x)$ into the aforementioned equation, we can employ a diffusion model as a prior for (σ, c) . In the context of $p(\{I_i\}_{i=1}^N | \sigma, c)$, we utilize a $\mathcal{L}_2$ loss function $\mathcal{L}_p = \|\hat{c}(r) - c_g\|_2^2$, where c_g represents the true color. In this way, the gradient for our loss function is

$$\nabla \mathcal{L} = \nabla \mathcal{L}_p + \lambda_D \varepsilon_D. \quad (12)$$

Here λ_D controls the weight of the regularize and is set as 0.01 when training our model.

Experiments

In this section, we will commence by outlining the experimental parameters in “[Experimental settings](#)”. Following this, we will assess the efficacy of our method through qualitative and quantitative analyses in “[Evaluations and results](#)”.

Experimental settings

Datasets

To evaluate the effectiveness, a series of comprehensive experiments are carried out using both synthetic and real datasets. These datasets include the NeRF synthetic dataset [43], DTU dataset [15] and LLFF dataset [44]. The NeRF synthetic dataset comprises eight scenes and eight distinct objects, each with 100 input views and 200 test views. The DTU dataset consists of images of objects placed on a table against a black background, captured by a 6-axis industrial robot equipped with a camera and a structured light scanner. The LLFF dataset contains eight real-world scenes captured by a handheld camera, with camera poses, intrinsic parameters, and 3D scene bounds estimated using COLMAP [29]. The resolution for the synthetic, DTU and LLFF datasets are 800×800 , 800×600 and 1008×756 , respectively.

Implementation details

In our experiment, we train our network using the DTU dataset with the aim of acquiring a generalizable framework. The DTU dataset is divided into 88 training scenes and 16 testing scenes, following a similar approach to PixelNeRF [33]. We also evaluate the performance of our model (merely trained on the DTU dataset) on the NeRF synthetic dataset and LLFF dataset, which possess distinct scene and view distributions compared to our training set. For each testing scene, we select 20 nearby views, from which 3 center views are chosen as input, and 4 images are designated as testing views. In cases where fine-tuning is necessary, the remaining 13 images are utilized as additional input for per-scene fine-tuning. Our feature extraction employs 32 channels, which

Table 1 Quantitative evaluations on our network architecture

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
MVSNeRF [17]	20.49	0.822	0.197
No AMIF	24.85	0.938	0.179
No DRKE	25.96	0.943	0.214
No DDM	25.74	0.893	0.181
Ours full model	26.47	0.952	0.172

Since the proposed model profits greatly from cross-attention-based density prediction, difference-based color prediction and diffusion regularization, it obtains better performances and indicates the effectiveness of the architecture

Bold indicates the best results

also corresponds to the number of channels in our cost volume and neural encoding volume. In the attention-based multi-position information fusion module, we sample 128 points on each marching ray. Our multi-head attention design utilizes 4 heads, and the MLP decoder is implemented with 6 layers. We train our network with one RTX 3090Ti GPU and use Adam optimizer with an initial learning rate of $5e-4$.

Evaluations and results

Evaluations on our network. We first conduct studies on our network architecture to validate the importance of different modules in our whole model, including cross-attention-based density prediction, difference-based color prediction and diffusion regularizer. To ensure a fair comparison, all models are retrained, and additional layers are incorporated into the various variants to ensure a similar number of parameters as our final model. For “No AMIF”, we remove the design of our cross-attention-based density prediction and replace it with MLPs. In this way, the density prediction for samples on each ray is independent and local. As shown in Table 1 and Fig. 5, when removing this module, the network cannot accurately predict the object density and local artifacts appear. For “No DRKE”, we remove the difference-based relevant knowledge exploitation for color prediction and employ MLPs to estimate color. Removing view differences reduces our network’s ability to combine relevant information from surrounding views to perform predictions and harmful factors affect the accuracy of color predictions. For “No DDM”, we directly train our model with $\mathcal{L}_p$ loss, resulting in blurry reconstruction results and disappearance of detailed information.

Compared to the previous methods, our variants exhibit distinct advantages and roles. While both can enhance accuracy, they serve different roles in handling radiation field reconstruction tasks. The cross-attention design in density prediction excels in capturing long-range dependencies and complex relationships within the radiation field. It effectively handles pixel interactions and global information, thereby

Table 2 Depth reconstruction

	Abs err $\downarrow$	Acc (0.01) $\uparrow$	Acc (0.05) $\uparrow$
MVSNeRF [17]	0.057/0.065	0.551/0.493	0.715/0.685
No AMIF	0.036/0.047	0.635/0.593	0.737/0.681
No DRKE	0.025/0.049	0.629/0.624	0.752/0.747
No DDM	0.032/0.038	0.612/0.584	0.741/0.656
Ours full model	0.023/0.035	0.663/0.651	0.814/0.792

We evaluate our depth reconstruction on the DTU testing dataset [15] with our variants and other neural rendering methods. The two numbers of each item denote the depth at reference/novel views. Our method is superior to other rendering methods and achieves high accuracy comparable to other structure variants

Bold indicates the best results

improving the accuracy and detail preservation of the reconstruction results. Our difference-based relevant knowledge exploitation establishes the connection between the difference in viewing direction and color, and we focus on the relevant color information under different views to jointly predict the color of a certain point. Our denoise diffusion model constrains the reconstruction process as a prior to mitigate interfering factors and achieve more stable and reliable results. By combining advantages, our final model leverages multi-level knowledge and obtains superior performances.

Depth reconstruction. In order to verify our ability of geometry reconstruction, we evaluate various variants by comparing depth reconstruction results. These results are generated from the volume density using a weighted sum of the depth values of sampled points on marched rays, as detailed in [9]. The quantitative performance of different methods is summarized in Table 2. Abs err represents the disparity between the calculated depth estimation results and the actual depth map for each scene and view. Meanwhile, Acc measures the proportion of pixels in the depth estimation map with a relative error from the true depth map that falls below a specified threshold. Note that, although our network is trained with only rendering supervision and no depth supervision, we obtain advantages with the help of our designed modules. Combined with visual results in Fig. 5, it is clear that we can generate high-quality of geometry reconstruction, which is one critical factor that leads to our realistic rendering.

Comparisons. To validate the effectiveness of the proposed model, a comparative analysis is conducted between our approach and existing state-of-the-art neural rendering algorithms. We use the released code and retrain the generalizable model, including MVSNeRF [17], IBRNet [16], SRF [36], PixelNeRF [33], ColNeRF [14] and ours, on the DTU training dataset [15], and test these existing models using three input views as used in our paper. This evaluation is performed

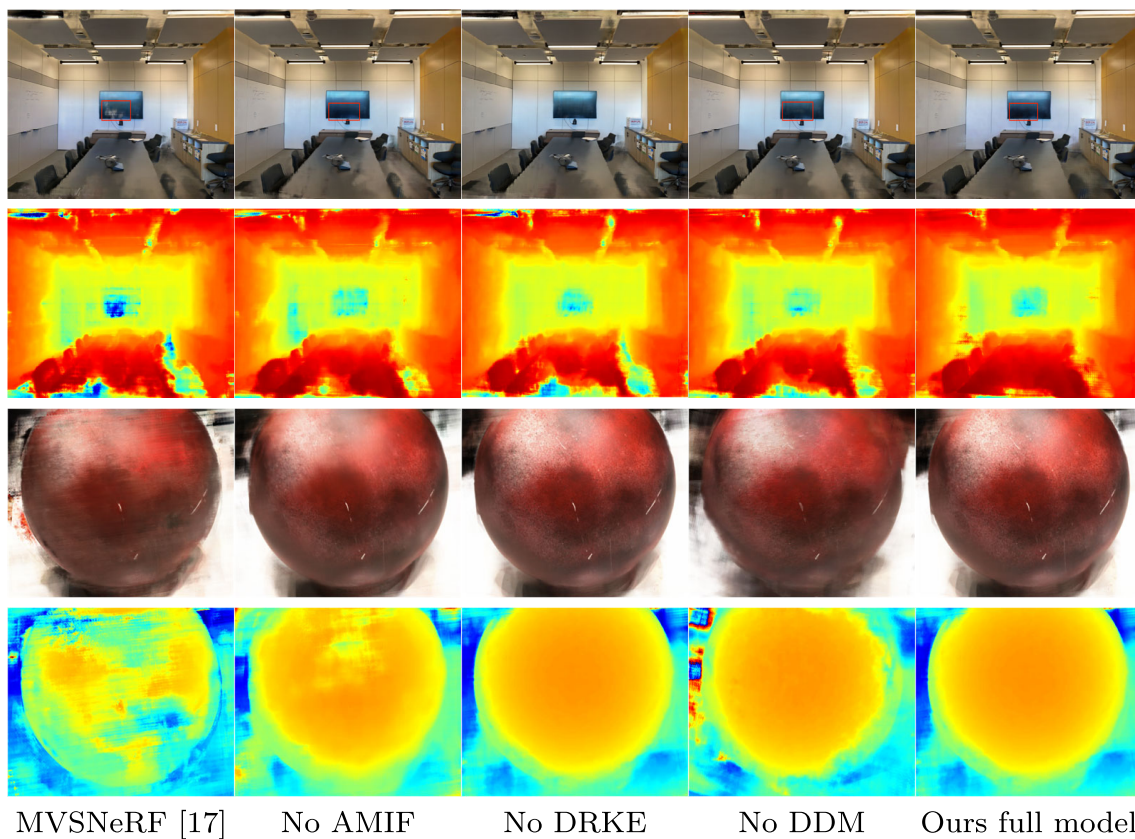


Fig. 5 Results of rendering and depth reconstruction based on different variants

Table 3 Quantitative evaluation of novel view synthesis

Method	Setting	Synthetic data [43]			Real data (DTU [15] / LLFF [44])		
		PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
MVSNerF [17]	No per-scene optimization	23.41	0.897	0.176	20.49/19.98	0.822/0.776	0.197/0.311
IBRNet [16]		22.44	0.897	0.195	24.04/19.79	0.891/0.782	0.190/0.379
SRF [36]		19.21	0.584	0.302	15.32/12.34	0.671/0.250	0.304/0.591
PixelNeRF [33]		22.65	0.808	0.202	20.56/16.17	0.753/0.438	0.223/0.512
ColNeRF [14]		22.50	0.843	0.198	19.55/18.96	0.716/0.587	0.362/0.447
Ours	Per-scene optimization	23.59	0.905	0.172	26.47/20.03	0.952/0.784	0.172/0.306
MipNeRF [46]		25.47	0.934	0.163	16.54/14.62	0.741/0.351	0.198/0.495
DifNeRF [47]		21.37	0.784	0.241	20.34/19.79	0.818/0.568	0.160/0.209
RegNeRF [12]		22.89	0.822	0.196	22.20/19.08	0.741/0.587	0.190/0.336
DSNeRF [35]		21.42	0.792	0.261	21.23/18.90	0.730/0.649	0.264/0.350
FreeNeRF [48]		22.78	0.832	0.217	19.92/19.63	0.787/0.612	0.182/0.308
Ours $f_{t-10min}$		26.62	0.939	0.161	27.07/21.79	0.956/0.877	0.168/ 0.236

We compare our approach with state-of-the-art rendering methods on various datasets with complex scenes. Our approach achieves optimal results in most cases of novel view synthesis
 Bold indicates the best results

on three widely-used benchmark datasets with the same input views and uses four additional images to test each scene.

Table 3 summarizes the quantitative performances achieved by different methods. The evaluated metrics of

PSNR, SSIM and LPIPS [45] are exhibited on different datasets. Overall, our rendering method demonstrates superior performance in the majority of cases and effectively addresses view synthesis using only three images as input.



Fig. 6 Rendering quality comparisons with different methods on the selected challenging samples. Our 10-min fine-tuning results are also exhibited

In contrast, generalizable PixelNeRF [33] and ColNeRF [14] tends to overfit the training setting on DTU and exhibits failure on the other two datasets. Despite being trained solely on the DTU dataset, our network exhibits strong generalization capabilities on two other datasets that have highly different view distributions and scenes. The designs of our cross-attention and relevant knowledge exploration focus on key information and improve the accuracy of prediction.

Furthermore, we also compare the results of the proposed generalizable framework with the state-of-the-art methods that need optimization on each testing scene, including Mip-

NeRF [46], DifNeRF [47], RegNeRF [12], DSNeRF [35] and FreeNeRF [48]. Due to the the strong initialization achieved from our network, we only fine-tune our neural reconstruction for a short period of 10 min, which already yields photorealistic results. In contrast to the results generated with substantially longer optimization time (up to 2h), our initial rendering results can be notably enhanced with just 10 min of fine-tuning. Our approach reconstructs localized radiance field representation as a neural encoding volume with explicit 3D structures. This leads to superior generalizability and rendering quality across diverse test-

ing scenes. Moreover, we integrate a denoising diffusion model as a regularization term to address the challenges of low-quality geometry and appearance in limited-view scenarios. The consistent performance improvement over other state-of-the-art methods is consistent with the visual results depicted in Fig. 6. By leveraging our generic network with the designed modules, we can capture more intricate details and achieve superior visual performance in fine structures. Some failed examples can also be found when the captured images contain degrading interference information, such as mirror-like reflections. This can complicate the reconstruction process. Therefore, further image pre-processing and joint optimization with the reconstruction process are future research directions.

Conclusion

In this paper, we propose a learning model with attention and prior guidance design, which presents a novel approach for optimizing NeRF in data-limited scenarios. The model operates under the assumption that scenes remain static across different views and that the multi-view data is precise and reliable. Our key insight is that NeRF's performance significantly deteriorates in sparse input scenes due to the difficult optimization of scene geometry. To overcome this limitation, we leverage a cross-attention mechanism to integrate information from surrounding views and neighbouring points along query rays, facilitating more effective learning of scene geometry. Additionally, we analyze and fuse view differences to highlight critical information for improving color prediction. Recognizing the impact of limited input views on geometry and appearance quality, we further incorporate a denoising diffusion model to learn scene priors, enhancing regularization during training. Consequently, our approach learns consistent 3D representations, enabling the synthesis of high-quality novel views. Comparative evaluations across diverse test sets and scenes show that our method outperforms existing approaches, especially under conditions with a limited number of input views. However, for highly challenging scenes featuring high glossiness, mirror-like reflections, or motion blur, directly recovering pronounced view-dependent effects through network inference remains difficult. Addressing these complexities may require extended fine-tuning and additional view pre-processing to fully enhance these effects. Consequently, further improvements in this area are necessary and will be actively pursued. In future research, we aim to develop more efficient methods to improve the learning process under extreme imaging conditions. We will focus on several strategies, including integrating multi-scale feature volume representations, enhancing the utilization of sampled points, and applying frequency regularization constraints.

Acknowledgements The work of W. Sun and Y. Zhang was partially supported by grants NSF of China (62201458, 61901384, 61231016), Natural Science Foundation of Shaanxi Province (no. 2022JQ-577, 2024JC-YBQN-0651), Shaanxi Provincial Education Department (no. 23KJ0669) and National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology. The work of R. Cui was supported by Graduate Innovation Fund of Xi'an University of Posts and Telecommunications (CXJJYL2023036).

Data Availability The data will be released at <https://github.com/dsdefv/nerf>.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Li J, Feng Z, She Q, Ding H, Wang C, Lee GH (2021) Mine: towards continuous depth mpi with nerf for novel view synthesis. In: IEEE international conference on computer vision, pp 12578–12588
2. Rahaman DM, Paul M (2017) Virtual view synthesis for free viewpoint video and multiview video compression using gaussian mixture modelling. *IEEE Trans Image Process* 27(3):1190–1201
3. Ham B, Min D, Oh C, Do MN, Sohn K (2013) Probability-based rendering for view synthesis. *IEEE Trans Image Process* 23(2):870–884
4. Gkioxari G, Malik J, Johnson J (2019) Mesh r-cnn. In: IEEE international conference on computer vision, pp 9785–9795
5. Qi CR, Su H, Mo K, Guibas LJ (2017) Pointnet: deep learning on point sets for 3d classification and segmentation. In: IEEE conference on computer vision and pattern recognition, pp 652–660
6. Yang B, Rosa S, Markham A, Trigoni N, Wen H (2018) Dense 3d object reconstruction from a single depth view. *IEEE Trans Pattern Anal Mach Intell* 41(12):2820–2834
7. Chen Z, Zhang H (2019) Learning implicit fields for generative shape modeling. In: IEEE conference on computer vision and pattern recognition, pp 5939–5948
8. Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A (2019) Occupancy networks: learning 3d reconstruction in function space. In: IEEE conference on computer vision and pattern recognition, pp 4460–4470
9. Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R (2021) Nerf: representing scenes as neural radiance fields for view synthesis. *Commun ACM* 65(1):99–106

10. Peng S, Niemeyer M, Mescheder L, Pollefeys M, Geiger A (2020) Convolutional occupancy networks. In: European conference computer vision, pp 523–540
11. Barron JT, Mildenhall B, Verbin D, Srinivasan PP, Hedman P (2022) Mip-nerf 360: unbounded anti-aliased neural radiance fields. In: IEEE conference on computer vision and pattern recognition, pp 5470–5479
12. Niemeyer M, Barron JT, Mildenhall B, Sajjadi MS, Geiger A, Radwan N (2022) Regnerf: regularizing neural radiance fields for view synthesis from sparse inputs. In: IEEE conference on computer vision and pattern recognition, pp 5480–5490
13. Wang J, Wang P, Long X, Theobalt C, Komura T, Liu L, Wang W (2022) Neuris: neural reconstruction of indoor scenes using normal priors. In: European conference on computer vision, pp 139–155
14. Ni Z, Yang P, Yang W, Wang H, Ma L, Kwong S (2024) Colnerf: collaboration for generalizable sparse input neural radiance field. AAAI Conf Artif Intell 38:4325–4333
15. Jensen R, Dahl A, Vogiatzis G, Tola E, Aanæs H (2014) Large scale multi-view stereopsis evaluation. In: IEEE conference on computer vision and pattern recognition, pp 406–413
16. Wang Q, Wang Z, Genova K, Srinivasan PP, Zhou H, Barron JT, Martin-Brualla R, Snavely N, Funkhouser T (2021) Ibrnet: learning multi-view image-based rendering. In: IEEE conference on computer vision and pattern recognition, pp 4690–4699
17. 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: IEEE international conference on computer vision, pp 14124–14133
18. Buehler C, Bosse M, McMillan L, Gortler S, Cohen M (2023) Unstructured lumigraph rendering. In: Seminal graphics papers: pushing the boundaries, pp 497–504
19. Heigl B, Koch R, Pollefeys M, Denzler J, Van Gool L (1999) Plenoptic modeling and rendering from image sequences taken by a hand-held camera. In: Mustererkennung, pp 94–101
20. Kalantari NK, Wang T-C, Ramamoorthi R (2016) Learning-based view synthesis for light field cameras. ACM Trans Gr 35(6):1–10
21. Hedman P, Philip J, Price T, Frahm J-M, Drettakis G, Brostow G (2018) Deep blending for free-viewpoint image-based rendering. ACM Trans Gr 37(6):1–15
22. Kellnhofer P, Jebe LC, Jones A, Spicer R, Pulli K, Wetzstein G (2021) Neural lumigraph rendering. In: IEEE conference on computer vision and pattern recognition, pp 4287–4297
23. Liu L, Gu J, Zaw Lin K, Chua T-S, Theobalt C (2020) Neural sparse voxel fields. Adv Neural Inf Process Syst 33:15651–15663
24. Liu S, Zhang Y, Peng S, Shi B, Pollefeys M, Cui Z (2020) Dist: rendering deep implicit signed distance function with differentiable sphere tracing. In: IEEE conference on computer vision and pattern recognition, pp 2019–2028
25. Hedman P, Ritschel T, Drettakis G, Brostow G (2016) Scalable inside-out image-based rendering. ACM Trans Gr 35(6):1–11
26. Riegler G, Koltun V (2020) Free view synthesis. In: European conference computer vision, pp 623–640
27. Penner E, Zhang L (2017) Soft 3d reconstruction for view synthesis. ACM Trans Gr 36(6):1–11
28. Jancosek M, Pajdla T (2011) Multi-view reconstruction preserving weakly-supported surfaces. In: IEEE conference on computer vision and pattern recognition, pp 3121–3128
29. Schonberger JL, Frahm J-M (2016) Structure-from-motion revisited. In: IEEE conference on computer vision and pattern recognition, pp 4104–4113
30. Choy CB, Xu D, Gwak J, Chen K, Savarese S (2016) 3d-r2n2: a unified approach for single and multi-view 3d object reconstruction. In: European conference computer vision, pp 628–644
31. Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S (2019) DeepSDF: learning continuous signed distance functions for shape representation. In: IEEE conference on computer vision and pattern recognition, pp 165–174
32. Jiang Y, Ji D, Han Z, Zwicker M (2020) Sdfdiff: differentiable rendering of signed distance fields for 3d shape optimization. In: IEEE conference on computer vision and pattern recognition, pp 1251–1261
33. Yu A, Ye V, Tancik M, Kanazawa A (2021) Pixelnerf: neural radiance fields from one or few images. In: IEEE conference on computer vision and pattern recognition, pp 4578–4587
34. Jain A, Tancik M, Abbeel P (2021) Putting nerf on a diet: semantically consistent few-shot view synthesis. In: IEEE international conference on computer vision, pp 5885–5894
35. Deng K, Liu A, Zhu J-Y, Ramanan D (2022) Depth-supervised nerf: fewer views and faster training for free. In: IEEE conference on computer vision and pattern recognition, pp 12882–12891
36. Chibane J, Bansal A, Lazova V, Pons-Moll G (2021) Stereo radiance fields (srf): learning view synthesis for sparse views of novel scenes. In: IEEE conference on computer vision and pattern recognition, pp 7911–7920
37. Yao Y, Luo Z, Li S, Fang T, Quan L (2018) Mvsnet: depth inference for unstructured multi-view stereo. In: European conference on computer vision, pp 767–783
38. Li X, Chen H, Qi X, Dou Q, Fu C-W, Heng P-A (2018) H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE Trans Med Imaging 37(12):2663–2674
39. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)
40. Sun W, Kong X, Zhang Y (2023) Attention-guided video super-resolution with recurrent multi-scale spatial-temporal transformer. Complex Intell Syst 9(4):3989–4002
41. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 33:6840–6851
42. Song Y, Garg S, Shi J, Ermon S (2020) Sliced score matching: a scalable approach to density and score estimation. In: Uncertainty in artificial intelligence, pp 574–584
43. Niemeyer M, Mescheder L, Oechsle M, Geiger A (2020) Differentiable volumetric rendering: learning implicit 3d representations without 3d supervision. In: IEEE conference on computer vision and pattern recognition, pp 3504–3515
44. Mildenhall B, Srinivasan PP, Ortiz-Cayon R, Kalantari NK, Ramamoorthi R, Ng R, Kar A (2019) Local light field fusion: practical view synthesis with prescriptive sampling guidelines. ACM Trans Gr 38(4):1–14
45. Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE conference on computer vision and pattern recognition, pp 586–595
46. Barron JT, Mildenhall B, Tancik M, Hedman P, Martin-Brualla R, Srinivasan PP (2021) Mip-nerf: a multiscale representation for anti-aliasing neural radiance fields. In: IEEE international conference on computer vision, pp 5855–5864
47. Wynn J, Turmukhambetov D (2023) Diffusionnerf: regularizing neural radiance fields with denoising diffusion models. In: IEEE conference on computer vision and pattern recognition, pp 4180–4189
48. Yang J, Pavone M, Wang Y (2023) Freenerf: improving few-shot neural rendering with free frequency regularization. In: IEEE conference on computer vision and pattern recognition, pp 8254–8263