ASSR-NeRF: Arbitrary-Scale Super-Resolution on Voxel Grid for High-Quality Radiance Fields Reconstruction

Ding-Jiun Huang
National Taiwan University
b08902028@ntu.edu.tw

Zi-Ting Chou
National Taiwan University
r11942101@ntu.edu.tw

Yu-Chiang Frank Wang
National Taiwan University
ycwang@ntu.edu.tw

Cheng Sun
NVIDIA Research
chengs@nvidia.com

Abstract

NeRF-based methods reconstruct 3D scenes by building a radiance field with implicit or explicit representations. While existing state-of-the-art methods for 3D scene reconstruction can capture geometry and appearance of a scene accurately, the quality of the rendered novel views is bounded by the training views. On the other hand, single-image super-resolution aims to restore low-resolution (LR) images to high-resolution (HR) counterparts and enhance details as well as textures simultaneously. To improve the rendering quality of a radiance field trained with LR training views, we propose Arbitrary-Scale Super-Resolution NeRF (ASSR-NeRF), a novel framework for super-resolution (SR) of neural radiance field. A voxel-based radiance field is first constructed with training views. Then, an attention-based VoxelGridSR module performs SR directly on the constructed radiance field, instead of the rendered 2D views, to generate finer details and textures in rendered views. Experiments with both quantitative and qualitative comparisons show that our proposed method significantly improves rendering quality.

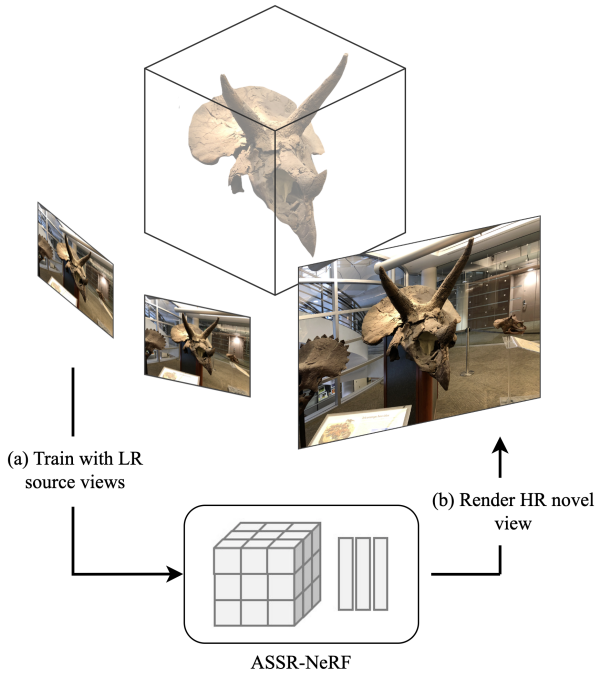

Figure 1. Given LR training views of a scene, our proposed *ASSR-NeRF* performs arbitrary-scale super-resolution to generate HR renderings with rich details.

1. Introduction

Novel view synthesis (NVS), or 3D scene reconstruction, aims to synthesize an image of a 3D scene from arbitrary viewing direction given multi-view images and camera poses. NeRF [22] first proposes to handle NVS with a multi-layer perceptron (MLP), which maps 3D positions and viewing directions to view-dependent colors and occupancy. Since NeRF learns a continuous volumetric representation, it is capable of synthesizing novel views at arbitrary resolution. Although NeRF can generate appealing results, it has lengthy training and rendering time. Many following works [3, 6, 23, 25, 28, 36, 37] then put emphasis on reducing training and rendering time. For example, Intant-ngp [23] takes only seconds to reconstruct a 3D scene. Another issue of NeRF is aliasing, happening when training views and rendering views are different in resolutions. Methods [2, 3, 10] then propose to mitigate this issue with mipmap techniques. While current state-of-the-art NVS methods can accurately synthesize geometry and appearance of a scene, they are limited by the quality of

training views. Specifically, these methods can "preserve" the observed details from training views in their radiance fields, but fail to "enrich" details in rendered novel views.

Single-image super-resolution (SISR) aims to synthesize a high-resolution (HR) image from its low-resolution counterpart. Different from mathematical up-sampling method, e.g., bicubic and bilinear interpolation, SISR methods [8, 13, 16, 17, 19, 31, 38] integrate deep learning models to enrich details and textures that are missed in LR images. Recently, generative-based methods [15, 30] shows exciting results with the advent of diffusion models.

A straightforward way of solving the above-mentioned quality issue of NVS methods is to directly apply SISR methods on the rendered views. However, applying SISR on each view independently will cause multi-view inconsistency, i.e., the geometry or appearance of objects isn't consistent among the results. [29] first proposes NeRF-SR for super-resolution (SR) of neural radiance field. Given a set of LR training views and 1 HR reference view of the same scene, NeRF-SR refines the rendered view through supersampling and a patch-based refinement module. Although it shows satisfying results, requiring an HR reference view for every scene is not practical. With similar ideas, Super-NeRF [9] and CROP [35] propose to employ an SR module to guide the HR renderings of NeRF, and rendered novel views can be recycled to guide the SR module, making its SR outputs view-consistent. However, an lengthy optimization is required for every scene, and the super-resolution is only optimized to a specific scale, limiting the flexibility. A pre-print [1] proposes to decompose a neural radiance field to tri-plane [5], and applies a pre-trained SISR model to the 2D feature planes. While this design improves the generalizability of SR module, i.e., a trained or fine-tuned SR module in a scene can be directly applied to another scene, applying SR on tri-plane independently causes inconsistency between the planes.

In this work, we propose arbitray-scale super-resolution NeRF (ASSR-NeRF) without the above-mentioned issues. Inspired by [28, 36, 37], we first construct the radiance field of a scene with explicit voxel grids. Similar to [1] that apply SR on volumetric representation, we propose an attention-based VoxelGridSR module that will directly super-resolve on the voxel grids. Since our super-resolution is performed in 3D space, there won't be multi-view inconsistency. The design of VoxelGridSR is inspired by LIIF [7], which treats SR as a mapping problem between pixel coordinates on HR images and colors. Given a coordinate of queried point in the 3D voxel grid, VoxelGridSR performs *density-distance-aware attention* on its nearest neighbors and outputs a refined voxel feautre. Since the coordinates in 3D space is continuous, VoxelGridSR is capable of performing ASSR. To make VoxelGridSR generalizable across scenes, we propose a cross-scene RGBNet to regularize the latent distribution of voxel features. Finally, a two-stage multi-scene training is elaborately designed.

In summary, our key contributions of our work are as follows:

- We propose a novel framework, ASSR-NeRF, for arbitrary-scale super-resolution on neural radiance fields.
- We propose VoxelGridSR as the core of ASSR-NeRF. It benefits multi-view consistency by performing super-resolution on 3D voxel grids.
- We propose a cross-scene RGBNet in ASSR-NeRF framework to enable a two-stage multi-scene training, improving the generalizability of VoxelGridSR.
- Experiments of both quantitative and qualitative comparisons show that our method can effectively performs super-resolution without multi-view inconsistency.

2. Related Work

2.1. Single-Image Super-Resolution

Single-image super-resolution (SISR) aims to restore an HR image from its LR counterpart. Early SISR methods [8, 13, 17, 38] adopt a deep convolutional neural network (CNN) to improve performance. After the advent of the attention mechanism, methods such as SwinIR [16] and ESRT [19] achieve competitive performance using a transformer-based architecture. To further enrich details in SR results, GAN-based and diffusion based methods [15, 30, 31] generate finer details as well as rich textures through adversarial training and powerful diffusion models respectively. Although excelling in SISR tasks, most methods can only perform SR on one fixed scale, failing to fit in real-world scenarios where display devices come in different resolutions.

2.2. Arbitrary-Scale Super-Resolution

To perform arbitrary-scale super-resolution (ASSR), one could first properly upscale the input image, then apply existing SISR methods. However, this approach is time-consuming and would lead to unsatisfied results with large scales. Recently, several methods [4, 7, 11, 14, 33, 34] are proposed to tackle ASSR with a single model. LIIF [7] maps arbitrary coordinates to RGB colors with an MLP, taking encoded image latent as input. With the same idea as LIIF, CiaoSR [4] further applies attention mechanisms for an enlarged receptive field and ensemble of local predictions.

2.3. Neural Radiance Fields

NeRF [22] has emerged as a prominent method for novel view synthesis (NVS), showcasing remarkable results with several input views and known camera poses. Specifically, NeRF encodes appearance and geometry of a 3D scene into a multi-layer perceptron (MLP), which takes 3D positions

and viewing directions as input and predicts corresponding colors and densities. Volume rendering techniques then accumulate the queried properties along a camera ray to formulate the color of a pixel. Many follow-ups extend this idea to different settings and scenarios. Some methods dramatically improve training or rendering efficiency with explicit structures. DVGO and Plenoxel [28, 36] employ voxel grids as explicit scene representations, leading to fast convergence. TensoRF [6] represents a scene with a tri-plane structure, greatly reducing both training time and memory usage. On the other hand, some methods focus on rendering quality. Mip-NeRF [2] leverages mipmapping to achieve anti-aliasing when rendering at different resolutions, and Tri-MipRF [10] further integrates hash encoding, inspired by Instang-ngp [23], to enable both instant reconstruction and anti-aliased rendering.

2.4. Super-Resolution of Neural Radiance Field

Since NeRF [22] learns a continuous volumetric representation for NVS, it can directly render novel views in arbitrary resolutions. However, the rendering procedure adopted by NeRF samples a scene with a single ray per pixel, therefore producing renderings with aliasing, blurs or artifacts when training and rendering views vary in resolutions. Supersampling, which samples multiple rays per pixel, is an effective solution, but it leads to heavy computational burden for MLP queries. Applying existent SISR methods to rendered novel views is another straightforward approach. Nevertheless, super-resolving each view independently would cause multi-view inconsistency, i.e., geometry of an object in different views varies. Several methods [2, 3, 10] are proposed to mitigate this quality issue, but they only "preserve" details, failing to "enrich" details in an HR renderings. For example, given LR training views of an antique vase, Mip-NeRF [2] can generate anti-aliased HR novel views but fail to restore finer patterns on the vase. NeRF-SR [29] first proposes a module to refine details for rendered HR novel views with one HR reference view of the same scene. Following the same idea, RefSR-NeRF [12] performs reference-based SR and reaches massive speedup. [35] further weakens the assumption that there's always an HR reference image for each scene, proposing to super-resolve novel views with only LR training views. While these methods show impressive results, the SR modules are all trained with a fixed scale and a per-scene optimization is required.

3. Preliminaries

NeRF [22] performs 3D scene reconstruction by encoding the geometry and occupancy of a scene into a multi-layer perceptron (MLP). The MLP maps a 3D position x and a viewing-direction d to the corresponding view-dependent color c and density σ. NeRF marches ray to render the color $\hat{\mathbf{C}}(r)$ of each pixel, where r represents the ray marched from camera center through the pixel. Along each ray, a total of K points are sampled, and the corresponding color and density are queried by the MLP. $\hat{\mathbf{C}}(r)$ can then be obtained by the following equations:

$$\hat{\mathbf{C}}(r) = \sum_{i=1}^{K} T_i \, \alpha_i \, c_i \, , \tag{1a}$$

$$\alpha_i = 1 - \exp(-\sigma_i \, \delta_i) \, , \tag{1b}$$

$$T_i = \prod_{j=1}^{i-1} (1 - \alpha_j) \, , \tag{1c}$$

where T_i is the accumulated transmittance from the starting point to i along the ray; α_i is the opacity; $(T_i \alpha_i)$ represents the probability of termination at point i; δ_i is the distance to adjacent points. NeRF model can then be trained with a photometric loss:

$$L_{\text{photo}} = \frac{1}{|R|} \sum_{r \in R} \| \mathbf{C}(r) - \hat{\mathbf{C}}(r) \|_2^2 \, . \tag{2}$$

While NeRF shows appealing performance on novel view synthesis, it struggles with lengthy training and rendering time. Subsequent works [6, 23, 28, 36] improve training efficiency by replacing the MLP with grid-based representations. We adopt DVGO [28] as our base model where modalities of interest, e.g., density, color, of a 3D position are explicitly stored as voxel features and can be queried via trilinear interpolation:

$$\text{interp}(x, V) : (R^3, R^{C \times N_x \times N_y \times N_z}) \rightarrow R^C \tag{3}$$

where V represents the voxel grid, x is the 3D position, C is the dimension of the modality, and N_x, N_y, N_z represents the 3 dimensions of the grid respectively. A shallow RGBNet is additionally employed to map queried voxel feature and viewing-direction to view-dependent color.

4. Method

4.1. Overview

In this section we describe our method for neural radiance field super-resolution. As shown in Fig. 2, we propose a novel framework of arbitrary-scale super-resolution NeRF (**ASSR-NeRF**). We propose a voxel-based radiance field that explicitly represents a 3D scene (Sec. 4.2). To enrich details in rendered HR novel views, VoxelGridSR module is proposed to directly super-resolve on the volumetric representation (Sec. 4.3). Finally, to make VoxelGridSR generalizable across different scenes, a cross-scene RGBNet is proposed to enable a multi-scene training (Sec. 4.4).

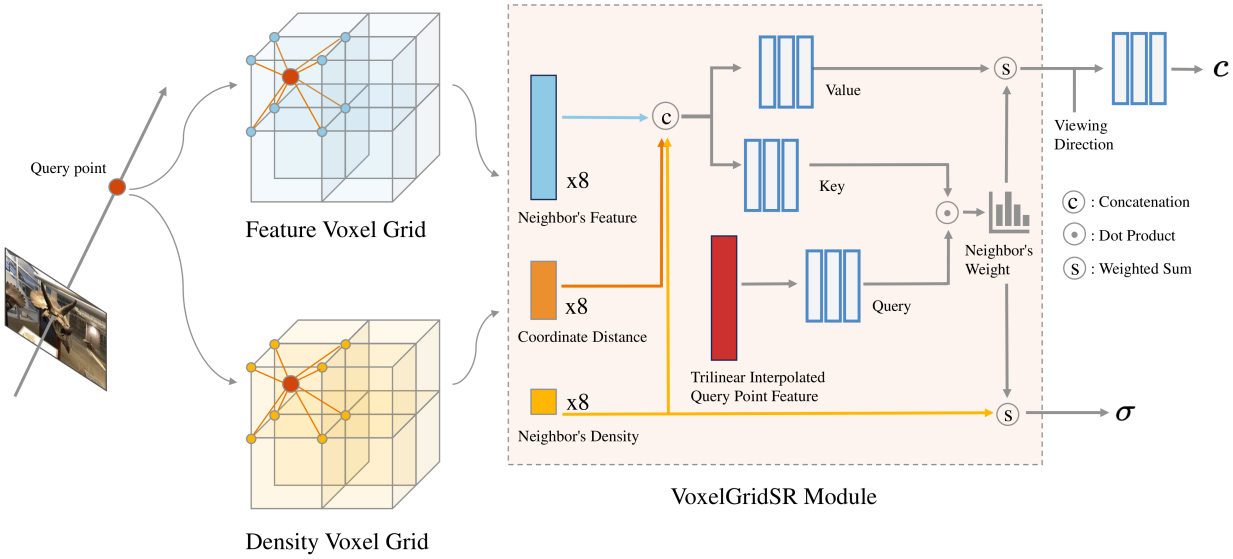

Figure 2. **Overview of ASSR-NeRF**: Given a query point x^q along a ray, features and densities of the nearest 8 neighbors are first sampled from voxel grids. Considering all the above modalities, VoxelGridSR then performs scaled dot-product attention for refined feature and density. Finally, view-dependent color is obtained through a cross-scene RGBNet.

4.2. Voxel-Based Radiance Field

Following [28, 36], we employ two voxel grids: a density voxel grid $V_d \in R^{1\times N_x\times N_y\times N_z}$ and a feature voxel grid $V_f \in R^{C\times N_x\times N_y\times N_z}$, to explicitly represent geometry and appearance of a 3D scene respectively, where C is the dimension of feature-space. Given a 3D position x^q, we first sample densities σ_i^n, features f_i^n of the 8 nearest points x_i^n where $i \in [1, 8]$, as shown in Fig. 2. Relative distances from x^q to x_i^n can then be calculated by:

$$s_i^n = x_i^n - x^q \tag{4}$$

We also obtain the trilinearly interpolated feature f^{tri} at x^q:

$$f^{tri} = \text{interp}(x^q, V_f) \tag{5}$$

To improve both performance and efficiency, we adopt tricks including free space skipping, progressively voxel grid upscaling, which are proposed by [28].

4.3. VoxelGridSR

In previous voxel-based methods [28, 36], a MLP query directly map f^{tri} and a viewing direction d to a view-dependent color. However, when trained on LR views, V_f only learns low-quality features, and trilinear interpolation can't help refine the features, resulting in lack of details in HR renderings.

Inspired by ASSR methods [4, 7], we propose VoxelGridSR module to learn a mapping between low-quality and high-quality features. Instead of applying SR on 2D feature space like previous methods, VoxelGridSR apply SR directly on 3D volumetric representation, strongly guarantees multi-view consistency, which is the main challenge of super-resolution of neural radiance field. VoxelGridSR consists of two parts: *Density-Distance-Aware Attention* and *Weighted Density Aggregation*.

4.3.1 Density-Distance-Aware Attention

While current state-of-the-art SISR methods achieve impressive performances, directly applying them to 3D voxel grids is impractical due to the booming computational cost. Inspired by [20, 26, 27], which introduce point-wise attention in point cloud object detection task, we propose voxel-wise attention considering both queried densities and features.

For a given 3D position x^q, the query, key and value are defined as:

$$\begin{cases} \mathbf{Q} : \text{MLP}_q(f^{tri}) \\ \mathbf{K} : \text{Stack}(\text{MLP}_k([f_i^n; s_i^n; \sigma_i^n]), i \in [1, 8]) \\ \mathbf{V} : \text{Stack}(\text{MLP}_v([f_i^n; s_i^n; \sigma_i^n]), i \in [1, 8]) \end{cases} \tag{6}$$

where f^{tri}, s_i^n and σ_i^n are concatenated before the MLP,

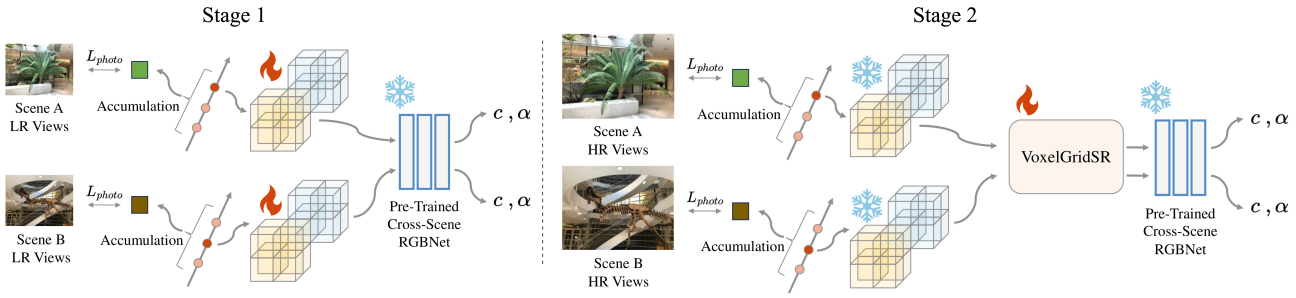

Figure 3. **Two-Stage Multi-Scene Training**: In stage 1, volumetric representation from every scene is first constructed with a shared pre-trained *cross-scene RGBNet*. In stage 2, all volumetric representations are fixed, and only the VoxelGridSR module is trained. Note that training views in stage 1 are in LR and those in stage 2 are in HR.

and will be stacked to $\mathbf{K}$ and $\mathbf{V}$ matrix. Attention can then be performed by scaled dot-product attention:

$$f^{\text{refine}} = \text{softmax}(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{D_K}})\mathbf{V} \tag{7}$$

where D_K is the dimension of voxel feature.

Compared with trilinear interpolation, VoxelGridSR considers not only the distance from x^q to its neighbors but also the relevance of features. Density information is as well beneficial because it helps differentiate interfaces between objects and air.

4.3.2 Weighted Density Aggregation

To generate novel views with finer geometry, we aggregate σ_i^n with the attention weights $w_i \in [0, 1], i \in [1, 8]$ obtained from the scaled dot-product attention:

$$\sigma = \sum_{i=1}^{8} \sigma_i^n w_i \tag{8}$$

4.4. Multi-Scene Training and Cross-Scene RGBNet

Before we describe the training procedure of ASSR-NeRF in this section, we first define our objective and the training settings. In SISR, there are HR/LR 2D training image pairs and LR testing images. In our case, which is super-resolution of neural radiance field, we have training scenes S_{train} of HR/LR image pairs $I^{hr}_{train}/I^{lr}_{train}$ and testing scenes S_{test} of only LR images I^{lr}_{test}. Our main objective is to train the proposed VoxelGridSR with S_{train}, and then the trained VoxelGridSR can be directly applied to any volumetric representations, i.e., voxel grids, trained with S_{test}.

Fig. 3 shows our two-stage training procedure. In the first stage, a volumetric representation for every scene $s_{train} \in S_{train}$ is initialized with I^{lr}_{train}. Instead of maintaining per-scene RGBNet, we employ a pre-trained cross-scene RGBNet, shared by every scene. RGBNet represents a mapping function between voxel feature space and color space. Without this unified mapping function, the latent space of voxel features can greatly vary between scenes, hindering the training of VoxelGridSR. In the second stage, we fix voxel grids and train only VoxelGridSR with I^{hr}_{train}. In this way, VoxelGridSR learns the mapping between low-quality voxel features to high-quality. After this two-stage training procedure, VoxelGridSR can directly be applied to volumetric representaion for $s_{test} \in S_{test}$, generating HR renderings with rich details.

5. Experiments

5.1. Implementation Details and Dataset

We impelment ASSR-NeRF with PyTorch [24]. To sample all the densities, features and relative distances of nearest neighbors efficiently given a 3D position, we design custom CUDA extensions. We set an expected number of voxels 256^3 for both density and feature voxel grids. The cross-scene RGBNet consists of 3 MLP layers with dimensions of 64, and is pre-trained with a carefully picked 3D scene, *fern* in LLFF [21] in our experiments, that generalizes well. Each volumetric representation in the first stage is trained for $30k$ iterations, and VoxelGridSR is trained for $240k$ iterations in the second stage. In the second stage of multi-scene training described in 4.4, VoxelGridSR is trained with voxel grids, V_f, V_d, of a scene for 2 iterations in-turn. In our experiments, we use LLFF [21] dataset, as it contains scenes varying from large-scale hall room to zoomed-in object, which is very close to real-life scenarios.

scene: *Fern*	x2			x4			x5			x8		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Plenoxel [36]	24.49	**0.783**	0.275	22.48	**0.663**	0.430	21.62	**0.626**	0.489	21.38	**0.628**	0.545
DVGO [28]	24.71	0.765	0.276	22.65	0.647	0.414	21.77	0.609	0.464	21.52	0.609	0.508
ASSR-NeRF (ours)	**24.92**	0.773	**0.267**	**22.81**	0.657	**0.406**	**21.91**	0.620	**0.458**	**21.64**	0.620	**0.507**

scene: *Trex*	x2			x4			x5			x8		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Plenoxel [36]	24.70	0.841	0.224	22.79	0.735	0.402	22.44	0.707	0.471	22.19	0.689	0.531
DVGO [28]	25.28	0.867	0.201	23.16	0.756	0.358	22.78	0.724	0.430	22.51	0.699	**0.500**
ASSR-NeRF (ours)	**25.33**	**0.868**	**0.198**	**23.21**	**0.760**	**0.357**	**22.83**	**0.729**	**0.429**	**22.55**	**0.703**	**0.500**

scene: *Orchids*	x2			x4			x5			x8		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Plenoxel [36]	20.46	0.712	0.250	19.59	0.587	0.369	19.20	**0.548**	0.423	19.10	**0.549**	0.496
DVGO [28]	20.54	0.715	0.251	19.69	0.580	0.372	19.30	0.533	0.421	19.20	0.524	0.478
ASSR-NeRF (ours)	**20.64**	**0.721**	**0.241**	**19.78**	**0.591**	**0.365**	**19.38**	0.546	**0.416**	**19.28**	0.539	**0.475**

Table 1. Quantitative results on LLFF [21]: We randomly split LLFF dataset to 7 training scenes s_{train} and 1 testing scene s_{test}. ASSR-NeRF along with VoxelGridSR is trained with $I^{hr}_{train}/I^{lr}_{train}$. All methods then render HR novel views of s_{test} given I^{lr}_{test}. ASSR-NeRF achieves the best performance in all metrics.

Figure 4. Qualitative results on LLFF [21]: We train ASSR-NeRF with $I^{lr}_{test}/I^{llr}_{test}$ of resolutions $[756 \times 1008]/[378 \times 504]$, and train other methods with I^{lr}_{test} of resolution $[756 \times 1008]$. Then, we generate renderings of resolution $[3024 \times 4032]$ for comparisons.

5.2. Comparisons and Discussions

We compare ASSR-NeRF and other state-of-the-art methods with few different settings. We first show the generalizability of VoxelGridSR in Sec. 5.2.1, where VoxelGridSR is trained with $s_{train} \in S_{train}$ and tested with $s_{test} \in S_{test}$. In Sec. 5.2.2, we prove VoxelGridSR's ability to generate. finer details and textures.

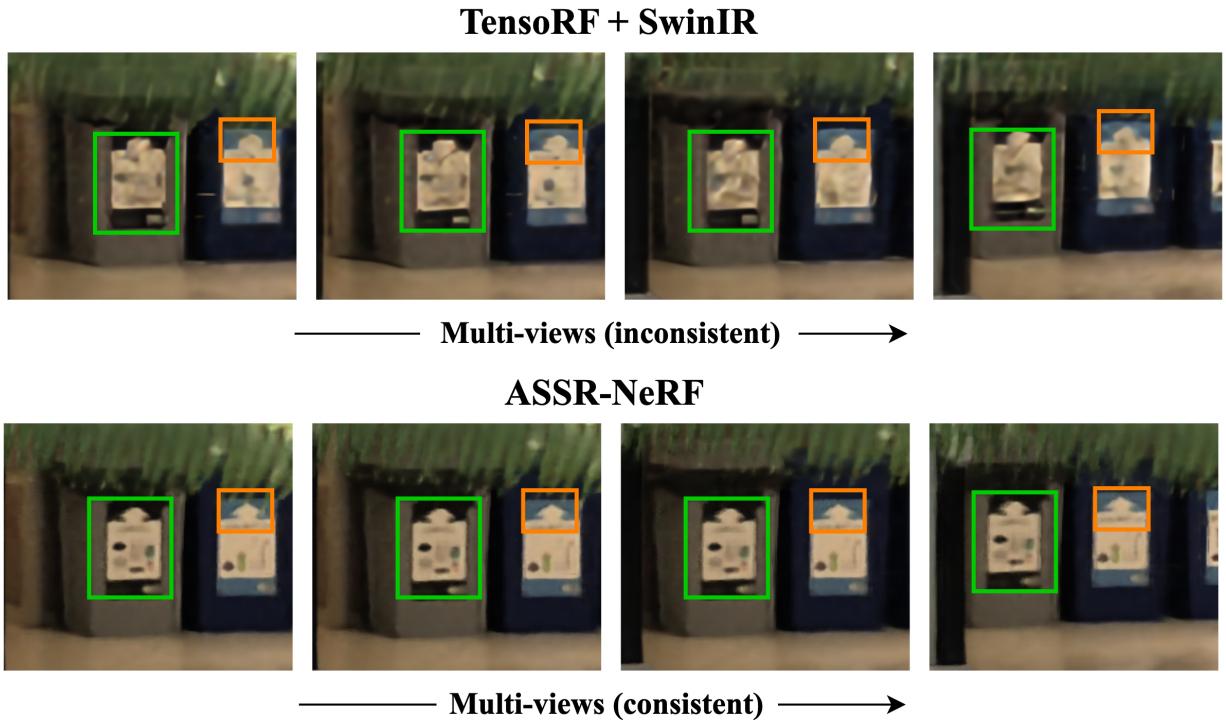

Figure 5. Qualitative results of multi-view consistency. A hybrid approach first renders novel views in LR then super-resolve them to HR, leading to multi-view inconsistency. ASSR-NeRF, on the other hand, generates results with consistent textures and shapes.

dataset: LLFF [21]	x2			x4		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
TensoRF [6]+ Real-ESRGAN [32]	24.44	0.782	0.249	23.49	0.721	**0.353**
TensoRF [6]+ SwinIR [16]	25.40	0.811	0.258	24.35	0.741	0.417
ASSR-NeRF (ours)	**26.09**	**0.827**	**0.220**	**25.03**	**0.769**	0.376

Table 2. Comparisons with 2D image SR methods on LLFF [21]: ASSR-NeRF directly renders HR novel views. For comparison, TensorRF [6] first renders LR novel views, then SOTA image SR methods super-resolve the views to HR. Our approach achieves the best performances in all the metrics.

5.2.1 ASSR-NeRF with Multi-Scene Training

Following the two-stage multi-scene training described in Sec. 4.4 and Fig. 3, we train ASSR-NeRF and VoxelGridSR with $I^{hr}_{train}/I^{lr}_{train}$ from $s_{train} \in S_{train}$.Then all the methods render HR novel views of s_{test} given I^{lr}_{test}. In this experiment, $I^{hr}_{train}/I^{lr}_{train}$ of resolutions $[378 \times 504]/[189 \times 252]$ is down-sampled from the $4k$ LLFF [21] dataset. For fair comparisons, we train DVGO [28] and Plenoxel [36] with the same grid resolution 256^3. From Tab. 1, we can see that ASSR-NeRF reaches the best performance in all testing scenes, indicating that VoxelGridSR is generalizable across scenes, and is effective in super-resolution of neural radiance field. Note that while Plenoxel [36] has the highest SSIM scores in certain scales, ASSR-NeRF always achieves the highest LPIPS score, indicating the competence of maintaining high perceptual quality.

5.2.2 ASSR-NeRF with Single-Scene Training

In this experiment, we want to especially show VoxelGridSR's ability to enrich details in HR renderings with qualitative comparisons. Specifically, given only $I^{lr}_{test} \in s_{test}$, we can first down-sample I^{lr}_{test} to I^{llr}_{test} with an arbitrary scale, and take this self-made HR/LR image pairs to train ASSR-NeRF along with VoxelGridSR. With this strategy, ASSR-NeRF can generate HR renderings with finer details and textures. We train ASSR-NeRF with $I^{lr}_{test}/I^{llr}_{test}$ of resolutions $[756 \times 1008]/[378 \times 504]$, and train other methods [6, 28] with I^{lr}_{test} of resolution $[756 \times 1008]$. Then, we

generate renderings of resolution $[3024 \times 4032]$ for comparison. As shown in Fig. 4, ASSR-NeRF achieves better perceptual quality when rendering at HR. Shown in the first row, novel view from DVGO [28] and TensoRF [6] contain artifacts while ASSR-NeRF render a clearer view. In addition, ASSR-NeRF renders sharper lines, words and edges, as shown in the second and the third rows. This experiment indicates that VoxelGridSR provides an effective aggregation to refine voxel features.

5.2.3 Multi-View Consistency

We also compare our method with hybrid approaches using state-of-the-art image SR methods [16, 32]. Real-ESRGAN [32] adopts a GAN-based architecture and proposes a high-order degradation modeling process to simulate real-world degradations of images. SwinIR [16] integrates Swin Transformer [18] into its architecture, reaching SOTA SR performance. In a hybrid approach, a NeRF-like model first renders LR novel views. Then, an image SR model super-resolves the novel views to HR. We first train ASSR-NeRF with $I^{lr}_{test}/I^{llr}_{test}$ of resolutions $[756 \times 1008]/[378 \times 504]$, using the same strategy in Sec. 5.2.2. For hybrid approaches, we train TensoRF [6] with I^{llr}_{test} of resolution $[378 \times 504]$ and fine-tune pre-trained image SR models with $I^{lr}_{test}/I^{llr}_{test}$ of resolutions $[756 \times 1008]/[378 \times 504]$. During rendering, ASSR-NeRF directly renders novel views in different scales, while hybrid approaches first render novel views of resolution $[378 \times 504]$ then super-resolve them. We show a qualitative comparison in Fig. 5, where ASSR-NeRF and a hybrid approach render multiple views of the scene *Fern* from LLFF [21] dataset. We can see that ASSR-NeRF not only achieves better perceptual quality but also maintains multi-view consistency, i.e., the shapes and textures remain the same across different views. On the other hand, the hybrid approach generates blurry and inconsistent results. This multi-view inconsistency is expected, since SR is applied to 2D image features independently. Tab. 2, we provide quantitative comparisons with 2 upsampling scales. We can see that ASSR-NeRF achieves the best performances in almost all metrics, suggesting that applying SR on volumetric representations is a better strategy than applying SR on 2D features.

5.3. Ablation Studies

In this section, we present an ablation study to analyze our proposed cross-scene RGBNet. Described in Sec. 4.4, the purpose of cross-scene RGBNet is to unify the latent distribution of voxel features, enabling multi-scene training for VoxelGridSR. As shown in Fig. 6, ASSR-NeRF along with VoxelGridSR trained with a unified cross-scene RGBNet could render satisfied HR results with fine details. However, if the volumetric representation of every scene initializes its own RGBNet, or per-scene RGBNet, VoxelGridSR won't be trained properly, leading to blurry results with wrong color tone and artifacts.

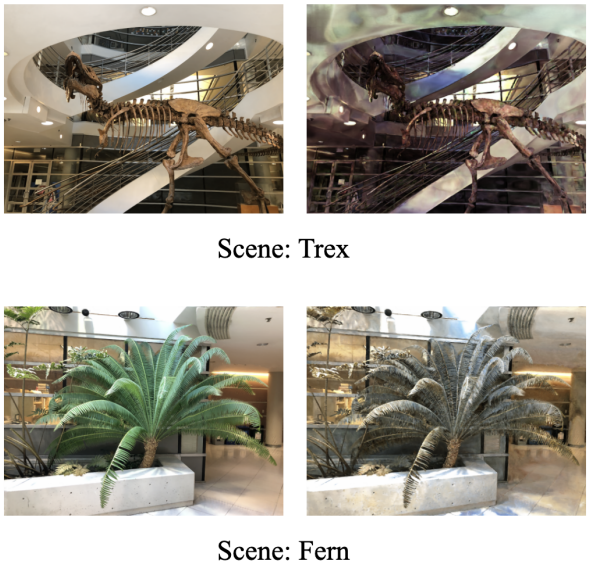

Figure 6. Ablation study of applying VoxelGridSR to volumetric representations with a cross-scene RGBNet (left) and with a per-scene RGBNet (right).

5.4. Discussions

The experiments show that our framework achieves competitive performance. However, one of limitations is the increased training time because swapping voxel grids in GPU memory causes lots of overhead. It takes about $10hr$ to train VoxelGridSR with 7 scenes, described in Sec. 5.1. Another limitation is the insufficient 3D training scenes compared to SISR methods which have extremely huge amount of 2D training data. Making use of 2D images to compensate the relative deficient 3D data is a potential research direction.

6. Conclusions

In this work, we propose a novel framework for super-resolution of neural radiance field. We propose to apply super-resolution directly on volumetric representation to eliminate multi-view consistency. To improve flexibility for real-world applications, we design an arbitrary-scale super-resolution module, VoxelGridSR, dedicated to refining the aggregated voxel features. We also propose a cross-scene RGBNet in our framework to regularize and unify the latent distributions of voxel features. Experiments on various benchmarks as well as qualitative comparisons show that our framework is strongly effective in improving rendering quality.

References

[1] Yuval Bahat, Yuxuan Zhang, Hendrik Sommerhoff, Andreas Kolb, and Felix Heide. Neural volume super-resolution, 2023. 2

[2] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields, 2021. 1, 3

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields, 2023. 1, 3

[4] Jiezhang Cao, Qin Wang, Yongqin Xian, Yawei Li, Bingbing Ni, Zhiming Pi, Kai Zhang, Yulun Zhang, Radu Timofte, and Luc Van Gool. Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution, 2023. 2, 4

[5] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks, 2022. 2

[6] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields, 2022. 1, 3, 7, 8

[7] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function, 2021. 2, 4

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:295–307, 2014. 2

[9] Yuqi Han, Tao Yu, Xiaohang Yu, Yuwang Wang, and Qionghai Dai. Super-nerf: View-consistent detail generation for nerf super-resolution, 2023. 2

[10] Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, and Yuewen Ma. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields, 2023. 1, 3

[11] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution, 2019. 2

[12] X. Huang, W. Li, J. Hu, H. Chen, and Y. Wang. Refsr-nerf: Towards high fidelity and super resolution view synthesis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8244–8253, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 3

[13] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2016. 2

[14] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function, 2022. 2

[15] Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhi hai Xu, Qi Li, and Yue ting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2021. 2

[16] Jingyun Liang, Jie Cao, Guolei Sun, K. Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021. 2, 7, 8

[17] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017. 2

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 8

[19] Zhisheng Lu, Juncheng Li, Hong Liu, Chao Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 456–465, 2021. 2

[20] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection, 2021. 4

[21] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines, 2019. 5, 6, 7, 8

[22] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. 1, 2, 3

[23] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41 (4):1–15, 2022. 1, 3

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. 5

[25] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps, 2021. 1

[26] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection, 2021. 4

[27] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection, 2022. 4

[28] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction, 2022. 1, 2, 3, 4, 6, 7, 8

[29] Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. Nerf-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM, 2022. 2, 3

[30] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution, 2023. 2

[31] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018. 2

[32] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data, 2021. 7, 8

[33] Min Wei and Xuesong Zhang. Super-resolution neural operator, 2023. 2

[34] Jie-En Yao, Li-Yuan Tsao, Yi-Chen Lo, Roy Tseng, Chia-Che Chang, and Chun-Yi Lee. Local implicit normalizing flow for arbitrary-scale image super-resolution, 2023. 2

[35] Y. Yoon and K. Yoon. Cross-guided optimization of radiance fields with multi-view image super-resolution for high-resolution novel view synthesis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12428–12438, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2, 3

[36] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks, 2021. 1, 2, 3, 4, 6, 7

[37] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields, 2021. 1, 2

[38] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Raymond Fu. Image super-resolution using very deep residual channel attention networks. *ArXiv*, abs/1807.02758, 2018. 2