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

Ding-Jiun Huang[1], Zi-Ting Chou[1], Yu-Chiang Frank Wang[1,2], and Cheng Sun[1,2]

[1] National Taiwan University
[2] NVIDIA Research

Abstract. NeRF-based methods reconstruct 3D scenes by building a radiance field with implicit or explicit representations. While NeRF-based methods can perform novel view synthesis (NVS) at arbitrary scale, the performance in high-resolution novel view synthesis (HRNVS) with low-resolution (LR) optimization often results in oversmoothing. On the other hand, single-image super-resolution (SR) aims to enhance LR images to HR counterparts but lacks multi-view consistency. To address these challenges, we propose Arbitrary-Scale Super-Resolution NeRF (ASSR-NeRF), a novel framework for super-resolution novel view synthesis (SRNVS). We propose an attention-based VoxelGridSR model to directly perform 3D super-resolution (SR) on the optimized volume. Our model is trained on diverse scenes to ensure generalizability. For unseen scenes trained with LR views, we then can directly apply our VoxelGridSR to further refine the volume and achieve multi-view consistent SR. We demonstrate quantitative and qualitatively that the proposed method achieves significant performance in SRNVS.

Keywords: neural radiance field · super-resolution · feature distillation

1 Introduction

Novel view synthesis (NVS), or 3D scene reconstruction, aims to synthesize images of a 3D scene from arbitrary viewing directions given multi-view images and camera poses. NeRF [26] achieves remarkable NVS results by employing neural network as an implicit volumetric representation, which maps 3D positions and viewing directions to view-dependent colors and occupancy. Due to its flexibility, numerous follow-up extensions, applications, and improvements had been made on top of NeRF. While current state-of-the-art NeRF-based methods can accurately synthesize geometry and appearance of a scene, high-resolution novel view synthesis (HRNVS) poses a great challenge to them, where high-resolution (HR) novel views are rendered by radiance fields constructed from low-resolution (LR) training views. Since LR training views lack details of a scene, the rendered HR novel views are blurry and noisy.

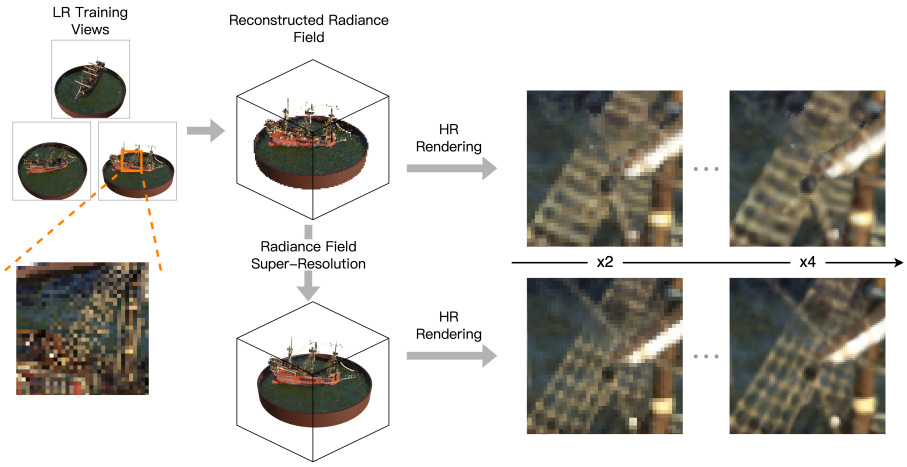

Fig. 1: Given a radiance field reconstructed from low-resolution (LR) training views, we perform radiance field super-resolution, leading to cleaner details in rendered views of high-resolution (HR).

On the other hand, single-image super-resolution (SISR) aims to synthesize an HR image from its LR counterpart. Different from mathematical up-sampling methods, e.g., bilinear interpolation, SISR methods [9, 17, 20, 21, 24, 35, 43] integrate deep learning models to enrich details and textures that are missed in LR images. Recently, generative-based methods [19, 34] shows exciting results with the advent of diffusion models.

A straightforward way of solving the above-mentioned quality issue of HRNVS 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 multiple rendered views. [33] 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 HR view through super-sampling 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 [10] and CROP [39] 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, a lengthy optimization is required for every scene, or the upscaling factor of SR is fixed, reducing the flexibility. A pre-print [1] proposes to decompose a neural radiance field to tri-plane [6], 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) for super-resolution novel view synthesis (SRNVS) without the above-mentioned issues. While HRNVS, performed by common NeRF-based models, only raise the resolution of rendered views without adding more details, SRNVS aims to enrich the details and textures in novel views and remains multi-view consistency. ASSR-NeRF consists of two main parts: a voxel-based distilled feature field reconstructed from LR views and an attention-based VoxelGridSR model that will directly super-resolve the radiance field. Inspired by [30,40,41], we first construct a distilled feature field to represent a scene with explicit voxel grids. To let VoxelGridSR perform attention on meaningful features, feature distillation is deployed to embed extracted features of 2D training views into 3D voxel grids. Since our SR is performed in 3D space, there won't be multi-view inconsistency. The design of VoxelGridSR is inspired by LIIF [8], which treats SR as a mapping problem between coordinates and the corresponding RGB values. Given the coordinate of a queried point, VoxelGridSR performs *density-distance-aware attention* on distilled features queried from its local region and outputs a refined feature representing the queried point. Since the coordinate in 3D space is continuous, VoxelGridSR is capable of optimizing SR at arbitrary scale. In addition, since VoxelGridSR is generalizable, it serves as an off-the-shelf method that can directly apply to any reconstructed feature field of unseen scene for SRNVS.

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

- We propose a novel framework, ASSR-NeRF, for super-resolution novel view synthesis (SRNVS) of radiance field.
- We distilled knowledge of low-level 2D SR priors from pre-trained feature extractor into radiance field to benefit SR in 3D space
- We design VoxelGridSR model to refine the optimized volume for SRNVS with richer textures and details.
- We train our VoxelGridSR to be generalizable so we can directly refine the radiance field of unseen scenes trained with LR views.

2 Related Work

2.1 Image Super-Resolution

Image super-resolution (SR) aims to restore a high-resolution (HR) image from its low-resolution (LR) counterpart. Early image SR methods [9,17,21,43] adopt a deep convolutional neural network (CNN) to improve performance. After the advent of the attention mechanism, methods such as SwinIR [20] and ESRT [24] achieve competitive performance using a transformer-based architecture. To further enrich details in SR results, GAN-based and diffusion based methods [19, 34, 35] generate finer details as well as rich textures through adversarial training and powerful diffusion models respectively. Although excelling in image SR 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. To perform

arbitrary-scale super-resolution (ASSR), one could first properly upscale the input image, then apply existing image SR methods. However, this approach is time-consuming and would lead to unsatisfied results when the upscaling factor is too large. Recently, several methods [4, 8, 12, 18, 36, 37] are proposed to tackle ASSR with a single model. LIIF [8] 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.

2.2 Neural Radiance Fields

NeRF [26] 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 multilayer 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 [30, 40] employ voxel grids as explicit scene representations, leading to fast convergence. TensoRF [7] 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 Zip-NeRF [3] further integrates grid-based representations, inspired by Instant-ngp [27], to enable both faster reconstruction and anti-aliased rendering, achieving state-of-the-art performance of NVS.

2.3 Super-Resolution of Neural Radiance Field

Since NeRF [26] learns a continuous volumetric representation for NVS, it can directly render novel views at arbitrary resolution. 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 image SR 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, 11] are proposed to mitigate this quality issue, but they only "preserve" details, failing to "enrich" details that are missed in LR training views. 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 [33] 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 [13] performs reference-based SR and reaches

massive speedup. [39] 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 up-scaling factor or a per-scene optimization is required.

3 Preliminaries

NeRF [26] 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 ray r casting through a pixel. Along each ray, K points are sampled to query the MLP for the corresponding color c_i and density σ_i, which is then blended by:

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

$$T_i = \prod\nolimits_{j=1}^{i-1} (1 - \alpha_j) \ , \tag{1b}$$

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

where r are sampled rays; T_i is the accumulated transmittance; 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:

$$L_{\text{photo}} = \sum\nolimits_{r \in R} \| \mathbf{C}(r) - \hat{\mathbf{C}}(r) \|_2^2 \ . \tag{2}$$

While NeRF shows appealing performance on novel view synthesis, it struggles with lengthy training and rendering time. Subsequent works [7,27,30,40] improve training efficiency by replacing the MLP with grid-based representations. We build our super-resolution algorithm based on DVGO [30], 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:

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

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. We use a density grid for geometry and a feature grid for appearance. A shallow MLP network, dubbed RGBNet, is additionally employed to map the queried voxel feature and the viewing-direction to view-dependent color.

4 Method

4.1 Overview

In this section we describe our method, dubbed ASSR-NeRF, for arbitrary-scale super-resolution NeRF. An overview of our approach is depicted in Fig. 2.

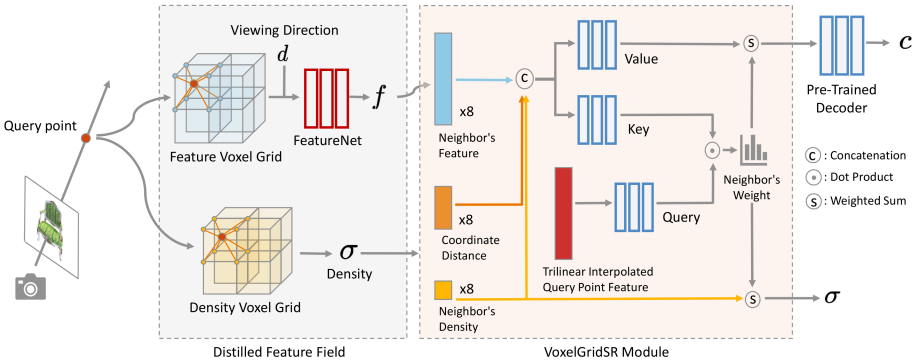

Fig. 2: Overview of ASSR-NeRF: Given a query point x along a ray, view-dependent distilled features and densities of its nearest neighbors are first sampled from a distilled feature field. Then, VoxelGridSR module aggregates the queried modalities and performs self-attention for refined feature and density. Finally, a pre-trained decoder maps the refined feature to RGB value c.

ASSR-NeRF mainly consists of two parts: *(i)* a voxel-based distilled feature field and *(ii)* a generalizable VoxelGridSR module. The distillation ensure the latent space alignment to facilitate multi-scene training and generalizability to novel scenes. The VoxelGridSR learns to utilizes the distilled SR latent for radiance field refinement. The following sections are organized as follows: we first describe our distilled feature field in Sec. 4.2 and explain why it is crucial in our approach. Then in Sec. 4.3, we introduce the generalizable VoxelGridSR module that serves as the core of radiance field super-resolution. We finally illustrate the training strategy for VoxelGridSR in Sec. 4.4.

4.2 Voxel-Based Distilled Feature Field

As shown in Fig. 2, given a query coordinate $x \in \mathbb{R}^3$, density σ and voxel feature f of each of x's nearest neighbors are queried from voxel grids. The feature and density voxel grids explicitly store appearance and occupancy information respectively, and FeatureNet further maps a voxel feature f to a view-dependent distilled feature f^d for VoxelGridSR to perform self-attention. Without special modifications, the features queried from the radiance field are nothing more than high-dimensional colors, limiting the performance of self-attention by VoxelGridSR. Applying a 3D feature extractor to the voxel feature grid is a straightforward solution, but the insufficiency of training data for a 3D feature extractor leads to poor overall performance. On the other hand, image data of large quantity as well as pre-trained models can greatly benefit tasks in 3D. Inspired by this observation, we propose distilled feature field for scene representation in ASSR-NeRF that bridges the gap between 2D and 3D data through feature distillation.

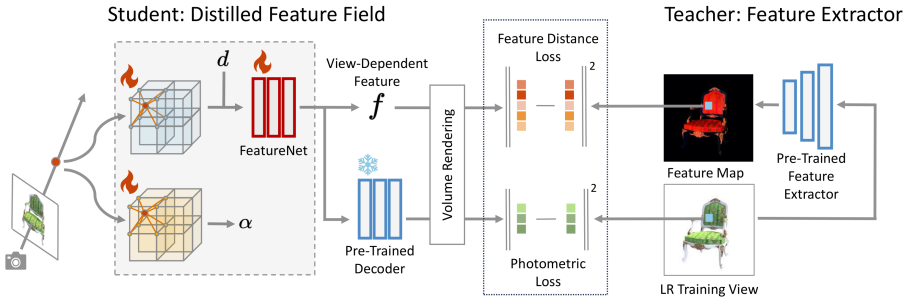

Fig. 3: Distilled feature field: In a student-teacher setting, features extracted from training views are distilled into a 3D student network. The student network is trained by minimizing the difference between rendered features and features from pre-trained image feature extractor, in addition to rendered colors and ground-truth pixel colors. FeatureNet turn voxel feature into view-dependent distilled features, and a pre-trained decoder maps view-dependent features RGB color.

In recent years, several works of neural radiance field have introduced feature distillation in their methods. [16, 32] add an additional branch to NeRF [26] to learn the semantic features from DINO [5], and performs open-set semantic segmentation of radiance fields with distilled semantic features on query points. [14] embeds multi-scale CLIP [29] features to a radiance field and performs visual grounding in 3D space. While all the above methods distilled high-level features, we propose to distill the learning-based low-level features into our radiance field as the SR priors since they represent vast information about textures and details of scenes. Feature distillation is also a crucial to our work as it guarantees the generalizability of VoxelGridSR model. By distilling features from the same teacher extractor into the radiance fields, feature voxel grids from different scenes can have aligned same latent space, i.e., the distributions of queried features from all radiance fields remain the same. In this way, VoxelGridSR can be trained in a unified latent space shared across voxel grids from all scenes and achieve generalizability.

We first follow the autoencoder training paradigm in [8] to train a residual dense network (RDN) [44] as feature extractor FE and a decoder D. Then, we follow a student-teacher setting to distill features into radiance field, as shown in Fig. 3. The distilled feature field is based on voxel grids, so VoxelGridSR module can directly perform super-resolution on scene representation, guaranteeing multi-view consistency of rendered novel views. Following [30, 40], we employ two voxel grids: a density voxel grid $V_d \in \mathbb{R}^{1\times N_x\times N_y\times N_z}$ and a feature voxel grid $V_f \in \mathbb{R}^{C\times N_x\times N_y\times N_z}$, to explicitly represent geometry and appearance of a 3D scene respectively, where C is the dimension of feature-space. Given a query point x_q, its density σ_q and voxel feature f'_q are queried from the voxel grids with trilinear interpolation, and FeatureNet further maps f'_q and viewing direction d to view-dependent features f_q. In addition, D decodes f_q to color c.

The distilled feature field is trained by minimizing the difference between rendered features $\hat{\mathbf{F}}(r)$ and teacher's extracted features $\mathbf{F}(r)$, as well as rendered colors and ground-truth pixel colors. The total loss L then becomes the sum of photometric loss L_{photo} and feature distance loss L_{feat}:

$$L = L_{\text{photo}} + \lambda L_{\text{feat}}, \tag{4a}$$

$$L_{\text{photo}} = \sum_{r \in R} \|\mathbf{C}(r) - \hat{\mathbf{C}}(r)\|_2^2 \ , \ \hat{\mathbf{C}}(r) = \sum_{i=1}^{K} T_i \, \alpha_i \, D(f_i) \tag{4b}$$

$$L_{\text{feat}} = \sum_{r \in R} \|\mathbf{F}(r) - \hat{\mathbf{F}}(r)\|_2^2 \ , \ \hat{\mathbf{F}}(r) = \sum_{i=1}^{K} T_i \, \alpha_i \, f_i \tag{4c}$$

where λ is the weight of feature distance loss, and is set to 0.5 by default. Following [16], we apply stop-gradient to density when rendering $\hat{\mathbf{F}}(r)$ since $\mathbf{F}(r)$ may not be multi-view consistent.

4.3 VoxelGridSR

We detail the architecture of VoxelGridSR for refining the radiance field. Inspired by ASSR methods [4, 8], we design our VoxelGridSR as a local 3D implicit function that maps a 3D coordinate and its nearby voxel features to a refined feature for the later color decoding. To this end, we introduce a *Density-Distance-Aware Attention* for the 3D refining procedure. Given a query point x_q, we trilinear interpolate the corresponding distilled feature f_q (Sec. 4.2) to produce the attention query. We gather the contextual information from the eight nearest neighbor grid point positions $\{x_i\}_{i=1}^{8}$, each of which comprises its distilled feature f_i, volume density σ_i, and the offset to the query $s_i = x_i - x_q$. The query, key, and value in attention operation can then be defined as:

$$\begin{cases} \mathbf{Q} = & \text{MLP}_q\left(f_q\right) \\ \mathbf{K}_i = & \text{MLP}_k\left([f_i; s_i; \sigma_i]\right) \\ \mathbf{V}_i = & \text{MLP}_v\left([f_i; s_i; \sigma_i]\right) \end{cases}, \tag{5}$$

where f_i, s_i and σ_i are concatenated before the MLPs, which then forms the key matrix $\mathbf{K} \in \mathbb{R}^{8 \times D_K}$ and the value $\mathbf{V} \in \mathbb{R}^{8 \times D_K}$ matrices. Attention can then be performed by scaled dot-product attention:

$$f_q^{(\text{refine})} = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^\intercal}{\sqrt{D_K}}\right)\mathbf{V} \tag{6}$$

where D_K is the dimension of voxel feature. The *Density-Distance-Aware Attention* allow the VoxelGridSR to take the feature relevancy and the local spatial relationship into consideration. Density information is also beneficial because it helps differentiate interfaces, *e.g.*, objects and air. The distilled feature from Sec. 4.2 enables VoxelGridSR to utilize the SR prior to enhance the textures and

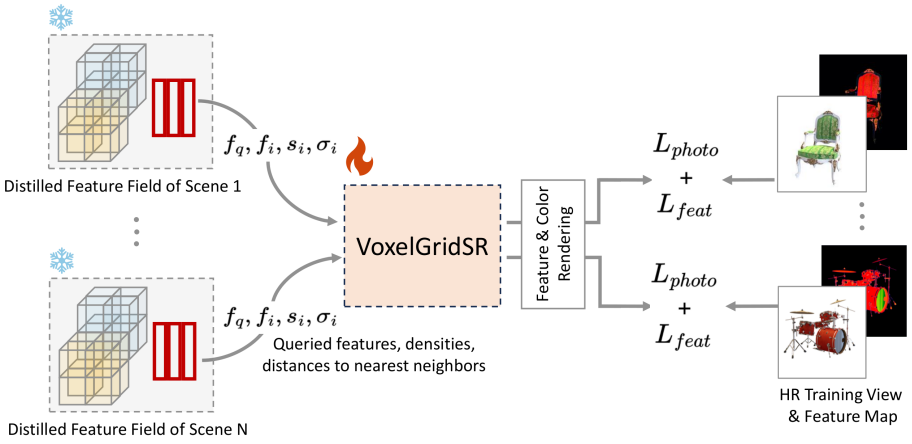

Fig. 4: Multi-scene training for VoxelGridSR: We train the generalizable VoxelGridSR module with N distilled feature fields that are reconstructed from LR training views as input and take HR training views as well as feature maps as ground-truth. In every iteration of training, a feature field of scene $i \in N$ is randomly selected. Then, VoxelGridSR maps f_q to $f_q^{(\text{refine})}$ and c, and is updated by L_{photo} and L_{feat}.

details of the scene. Additionally, we also refine the geometry by aggregating the grid point densities with the attention weights:

$$\sigma_q = \text{softmax}\left(\frac{\mathbf{Q}\cdot\mathbf{K}^\intercal}{\sqrt{D_K}}\right)\cdot[\sigma_1,\cdots,\sigma_8]^\intercal \tag{7}$$

4.4 Multi-Scene Training for VoxelGridSR

Training VoxelGridSR on a single scene is less useful. As the distilled 3D feature is aligned to the SR latent space of the 2D teacher, we train VoxelGridSR on multiple scenes for generalizability. Once trained, the VoxelGridSR module can serve as an off-the-shelf enhancer to any distilled feature field under the same latent space. Fig. 4 depicts the multi-scene training procedure of VoxelGridSR. We first pre-trained N scenes with LR views and distilled feature fields. Subsequently, during each iteration of the cross-scene training, a feature field of scene $i \in N$ is randomly selected and refined by VoxelGridSR through *Density-Distance-Aware Attention*. We then can minimize the photometric loss L_{photo} between the rendered SR view and the ground-truth HR view. Feature matching loss L_{feat} is also applied to ensure the latent space to remain consistent.

5 Experiments

5.1 Implementation Details

We impelment ASSR-NeRF with PyTorch [28]. 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. FeatureNet in our distilled feature field is based on RDN [44], and the pre-trained decoder consists of 5 MLP layers with dimensions of 256. FeatureNet and decoder are trained together in an autoencoder paradigm on DIV2K [31] dataset. To train VoxelGridSR, we take the BlendedMVS [38] dataset at resolution 768×576 as HR ground-truth training views and downsample with a factor $\times 4$ to generate corresponding LR training views. Distilled feature field of every scene is trained for $20k$ iterations and a batchsize 4096. We first reconstructed distilled feature fields of 40 scenes from BlendedMVS. After reconstructing distilled feature fields of 40 scenes, we train the VoxelGridSR module with the distilled feature fields for $240K$ and a batchsize 2048.

5.2 Comparisons and Discussions

We compare ASSR-NeRF and other state-of-the-art NeRF-based NVS methods [3, 30] with a few different settings. Among them, DVGO [30] serves as a baseline since its architecture, consisting voxel grids and shallow MLP layers, is similar to us. Zip-NeRF [3] is optimized for anti-aliasing so that their rendered novel views remain clear with large difference in training and testing resolutions. We also compare our method with image SR methods [20, 34] and a radiance field SR method [39].

NeRF-based NVS methods In this experiment, we compare ASSR-NeRF with state-of-the-art NeRF-based methods for high-resolution novel view synthesis (HRNVS). For every scene, we train a distilled feature field with LR training views, and directly apply pre-trained VoxelGridSR on the distilled feature field to perform super-resolution novel view synthesis (SRNVS). We also train other methods with LR training views and perform HRNVS. We compare all the methods on two datasets, Synthetic-NeRF [26] and BlendedMVS [38] with three different scales. Note that there is no overlapping between scenes used to train VoxelGridSR and scenes used for testing in all experiments. In Tab. 1, we show the result in PSNR, SSIM and LPIPS. We can see that ASSR-NeRF surpasses all the other methods at every scale. Although Zip-NeRF gains some advantage when the scale is low, its performance declines greatly when the scale increases, revealing common NeRF-based methods' shortcomings. This result matches our initial observation that NVS methods trained on LR training views have non-ideal performance when rendering HR novel views. We also show the qualitative results in Fig. 4. While having cleaner rendered views, ASSR-NeRF can also generate finer details as well as textures. For example, ASSR-NeRF generates cleaner patterns on figure's clothes and sharper edge of microphone.

Table 1: Quantitative results on Synthetic-NeRF [26] and BlendedMVS [38]: We compare ASSR-NeRF with other NeRF-based NVS methods. All methods are trained with LR training views of a downsampling factor x4, and render at three different resolutions. ASSR-NeRF achieves the best performance in all resolutions and datasets.

Synthetic-NeRF	x1.6			x2			x4		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DVGO [30]	29.89	0.945	0.061	28.80	0.933	0.077	27.33	0.910	0.115
TensoRF [7]	**31.36**	0.950	0.060	29.96	0.947	0.078	28.07	0.916	0.118
Instant-ngp [27]	28.55	0.933	0.095	28.23	0.926	0.101	27.26	0.902	0.128
Zip-NeRF [3]	30.95	**0.962**	**0.041**	29.56	0.951	0.057	27.73	0.923	0.102
ASSR-NeRF (ours)	31.09	0.961	0.048	**30.57**	**0.954**	**0.057**	**29.02**	**0.932**	**0.093**

BlendedMVS	x2			x2.5			x4		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DVGO [30]	26.88	0.909	0.111	26.96	0.902	0.124	24.74	0.845	0.187
TensoRF [7]	27.72	0.921	0.114	27.78	0.912	0.131	25.35	0.852	0.182
Instant-ngp [27]	26.63	0.894	0.150	26.64	0.882	0.166	25.14	0.835	0.188
Zip-NeRF [3]	28.48	0.929	0.148	28.41	0.918	0.173	25.97	0.854	0.237
ASSR-NeRF (ours)	**28.52**	**0.931**	**0.080**	**28.56**	**0.926**	**0.093**	**26.38**	**0.873**	**0.148**

Table 2: Quantitative results on BlendedMVS [38]: We compare our method with hyrbrid approaches where LR rendered views are super-resolved by image SR methods with factor x4 and a NeRF SR method that shares the same setting with us.

	PSNR ↑	SSIM ↑	LPIPS ↓
Zip-NeRF [3] + SwinIR [20]	26.21	0.866	0.159
Zip-NeRF [3] + StableSR [34]	24.56	0.839	0.169
CROP [39]	26.25	**0.879**	**0.145**
ASSR-NeRF (ours)	**26.38**	0.873	0.148

Super-resolution methods In this experiment, we compare our method with image SR methods as well as radiance field SR methods. Following the training setting from Sec. 5.2, we train Zip-NeRF [3] with LR training views. Then, we render LR novel view at the same scale, and super-resolve the novel views with state-of-the-art image SR methods [20, 34]. SwinIR [20] integrates Swin Transformer [23] into its architecture for better feature extraction on input images. StableSR [34] utilized pre-trained diffusion models to generate finer details and textures on SR results. Besides using image SR on rendered novel views, we compare with a radiance field SR method [39] that shares a similar setting with us. CROP [39] first super-resolves LR training views with pre-trained image SR methods, and uses the super-resolved views to train a NeRF-based NVS model. Unlike other radiance field SR methods [13, 33] that requires an HR reference image of the same scene, CROP only needs LR training views when testing. Our method differs from CROP that our VoxelGridSR module is optimized at arbitrary scale while CROP is trained at a fixed upscaling factor. All methods perform HRNVS with a upscaling factor $\times 4$. In Tab. 2, we show the quantati-

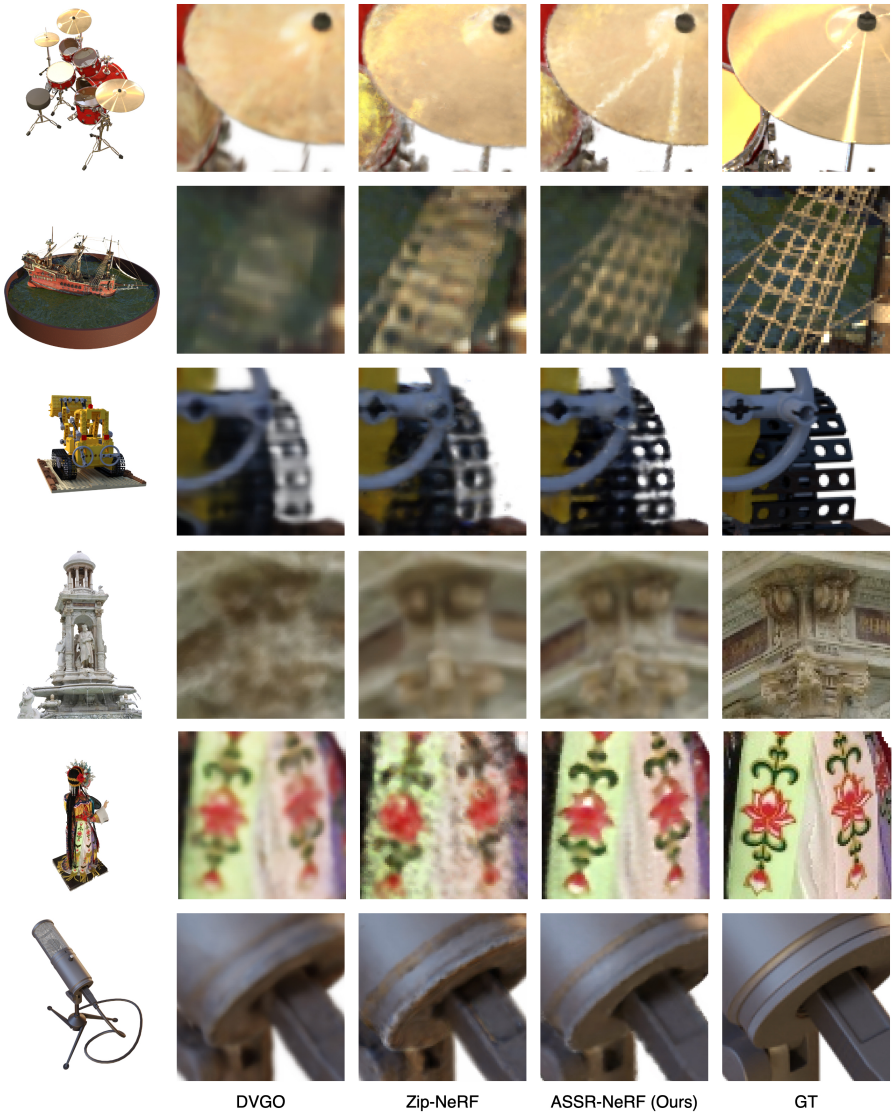

Fig. 5: Qualitative results on Synthetic-NeRF [42] and BlendedMVS [38]: All other baselines are first reconstructed from LR training views, then perform HRNVS. For ASSR-NeRF, pre-trained VoxelGridSR model is applied to achieve SRNVS. The results show that ASSR-NeRF generates cleaner edges as well as richer details than other baselines.

tive results on BlendedMVS [38]. While reaching comparable SSIM and LPIPS scores as CROP, our method outperforms in terms of PSNR.

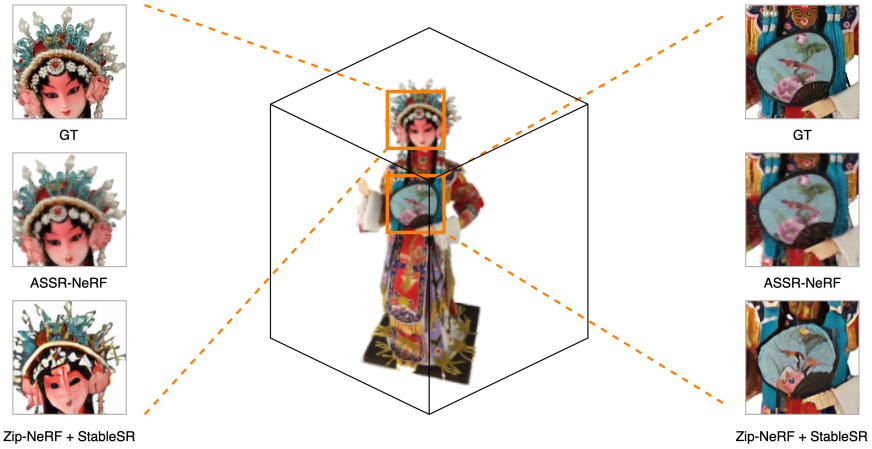

Fig. 6: Qualitative results of view consistency. Applying image SR method, e.g., StableSR [34] to rendered LR novel views from Zip-NeRF [3] leads to sharp but distorted results, and loses multi-view consistency. On the other hand, our method generates clean HR novel views with consistent details. We encourage readers to view our supplementary videos where our method achieves better multi-view consistency than other baselines.

Multi-View Consistency In this section, we discuss the multi-view consistency issue. In Sec. 5.2, we conduct the experiments of super-resolving rendered LR novel view with image SR methods. Although this approach can also generate cleaner HR novel views with finer details, multi-view inconsistency remains as a serious problem. As shown in Fig. 6, super-resolving rendered LR novel views with image SR methods lead to distorted geometry and textures across different views. On the other hand, our method generate consistent views from different camera poses. This advantage of our method is contributed from the design of ASSR-NeRF's SR module. Instead of applying SR on 2D feature maps, VoxelGridSR directly applies SR on 3D volume, i.e., the distilled feature fields, guaranteeing consistency of geometry and appearance across every viewing direction.

5.3 Ablation Studies

We provide ablation studies to analyze our proposed method in this section. We analyze the design of VoxelGridSR module to verify the effectiveness of *Density-Distance-Aware Attention*.

Analysis of VoxelGridSR We follow the same training setting described in Sec. 5.1 and train different variants of VoxelGridSR. Tab. 3 shows the variants' performance of SRNVS on BlendedMVS dataset with an upscaling factor $\times 4$.

Table 3: Ablation study on VoxelGridSR module. We verify the effectiveness of density-aware and distance-aware attention on voxel grids. While considering both denstiy and relative distance of nearest neighbors improves performance, we highlight that density information is more crucial to SR in 3D space.

Model	Feature-aware	Density-aware	Distance-aware	PSNR ↑	SSIM ↑	LPIPS ↓
A	✓	✓	✓	26.38	0.873	0.148
B	✓	✓		26.20	0.873	0.157
C	✓		✓	26.14	0.872	0.147

Density-aware means whether to consider densities of nearest neighbors when performing self-attention, and *Distance-aware* indicates whether the relative distance to each nearest neighbor is considered. The metrics show that both density and relative distance can benefit the performance of self-attention. Interestingly, we find that *Distance-aware* brings more improvements for 3D SR. This is reasonable, since density contains information about the local geometry and SR in 3D space can thus lead to sharper edges of objects in rendered views.

5.4 Discussions and limitations

The experiments show that our framework achieves competitive performance in SRNVS. However, one of limitations is the increased rendering time since VoxelGridSR performs self-attention on every sampled point. Reducing rendering time while keeping the same quality will be our future research direction. We also notice that currently there isn't an effective benchmark to assess multi-view consistency, and we can only compare our methods with others through frame-wise metrics such as PSNR and qualitative presentations. We think video quality assessment (VQA) might be an interesting research direction that can serve as an assessment metric for multi-view consistency.

6 Conclusions

In this work, we propose ASSR-NeRF, a novel framework for radiance field super-resolution. ASSR-NeRF consists of a distilled feature field for scene representation and a generalizable VoxelGridSR module for raidance field SR. Once a distilled feature field is reconstructed from any set of LR training views, a pre-trained generalizable VoxelGridSR module can be directly applied for super-resolution novel view synthesis (SRNVS). Our approach can greatly benefit real-world applications. For example, low-resolution training views captured by cheap devices can be efficiently utilized for high-quality novel view synthesis, reducing cost and time required by views capturing. Experiments on various benchmarks as well as qualitative comparisons show that our framework is strongly effective in improving rendering quality.

References

1. Bahat, Y., Zhang, Y., Sommerhoff, H., Kolb, A., Heide, F.: Neural volume super-resolution (2023)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields (2023)
4. Cao, J., Wang, Q., Xian, Y., Li, Y., Ni, B., Pi, Z., Zhang, K., Zhang, Y., Timofte, R., Gool, L.V.: Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution (2023)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers (2021)
6. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3d generative adversarial networks (2022)
7. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields (2022)
8. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function (2021)
9. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**, 295–307 (2014), https://api.semanticscholar.org/CorpusID:6593498
10. Han, Y., Yu, T., Yu, X., Wang, Y., Dai, Q.: Super-nerf: View-consistent detail generation for nerf super-resolution (2023)
11. Hu, W., Wang, Y., Ma, L., Yang, B., Gao, L., Liu, X., Ma, Y.: Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields (2023)
12. Hu, X., Mu, H., Zhang, X., Wang, Z., Tan, T., Sun, J.: Meta-sr: A magnification-arbitrary network for super-resolution (2019)
13. Huang, X., Li, W., Hu, J., Chen, H., Wang, Y.: Refsr-nerf: Towards high fidelity and super resolution view synthesis. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8244–8253. IEEE Computer Society, Los Alamitos, CA, USA (jun 2023). https://doi.org/10.1109/CVPR52729.2023.00797, https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00797
14. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields (2023)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023)
16. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation (2022)
17. Ledig, C., Theis, L., Huszár, F., Caballero, J., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 105–114 (2016), https://api.semanticscholar.org/CorpusID:211227
18. Lee, J., Jin, K.H.: Local texture estimator for implicit representation function (2022)

19. Li, H., Yang, Y., Chang, M., Feng, H., hai Xu, Z., Li, Q., ting Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2021), https://api.semanticscholar.org/CorpusID:233476433
20. Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L.V., Timofte, R.: Swinir: Image restoration using swin transformer. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) pp. 1833–1844 (2021), https://api.semanticscholar.org/CorpusID:237266491
21. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 1132–1140 (2017), https://api.semanticscholar.org/CorpusID:6540453
22. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection (2023)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021)
24. Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., Zeng, T.: Transformer for single image super-resolution. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 456–465 (2021), https://api.semanticscholar.org/CorpusID:248366743
25. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines (2019), https://arxiv.org/abs/1905.00889
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis (2020)
27. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics **41**(4), 1–15 (Jul 2022). https://doi.org/10.1145/3528223.3530127, http://dx.doi.org/10.1145/3528223.3530127
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library (2019)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
30. Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction (2022)
31. Timofte, R., Agustsson, E., Gool, L.V., Yang, M.H., Zhang, L., Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M., Wang, X., Tian, Y., Yu, K., Zhang, Y., Wu, S., Dong, C., Lin, L., Qiao, Y., Loy, C.C., Bae, W., Yoo, J., Han, Y., Ye, J.C., Choi, J.S., Kim, M., Fan, Y., Yu, J., Han, W., Liu, D., Yu, H., Wang, Z., Shi, H., Wang, X., Huang, T.S., Chen, Y., Zhang, K., Zuo, W., Tang, Z., Luo, L., Li, S., Fu, M., Cao, L., Heng, W., Bui, G., Le, T., Duan, Y., Tao, D., Wang, R., Lin, X., Pang, J., Xu, J., Zhao, Y., Xu, X., Pan, J., Sun, D., Zhang, Y., Song, X., Dai, Y., Qin, X., Huynh, X.P., Guo, T., Mousavi, H.S., Vu, T.H., Monga, V., Cruz, C., Egiazarian, K., Katkovnik, V., Mehta, R., Jain, A.K., Agarwalla, A., Praveen, C.V.S., Zhou, R., Wen, H., Zhu, C., Xia, Z., Wang, Z., Guo, Q.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1110–1121 (2017). https://doi.org/10.1109/CVPRW.2017.149

32. Tschernezki, V., Laina, I., Larlus, D., Vedaldi, A.: Neural feature fusion fields: 3d distillation of self-supervised 2d image representations (2022)
33. Wang, C., Wu, X., Guo, Y.C., Zhang, S.H., Tai, Y.W., Hu, S.M.: Nerf-sr: High quality neural radiance fields using supersampling. In: Proceedings of the 30th ACM International Conference on Multimedia. MM '22, ACM (Oct 2022). https://doi.org/10.1145/3503161.3547808, http://dx.doi.org/10.1145/3503161.3547808
34. Wang, J., Yue, Z., Zhou, S., Chan, K.C.K., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution (2023)
35. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C.C., Qiao, Y., Tang, X.: Esrgan: Enhanced super-resolution generative adversarial networks. In: ECCV Workshops (2018), https://api.semanticscholar.org/CorpusID:52154773
36. Wei, M., Zhang, X.: Super-resolution neural operator (2023)
37. Yao, J.E., Tsao, L.Y., Lo, Y.C., Tseng, R., Chang, C.C., Lee, C.Y.: Local implicit normalizing flow for arbitrary-scale image super-resolution (2023)
38. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks (2020)
39. Yoon, Y., Yoon, K.: 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). pp. 12428–12438. IEEE Computer Society, Los Alamitos, CA, USA (jun 2023). https://doi.org/10.1109/CVPR52729.2023.01196, https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01196
40. Yu, A., Fridovich-Keil, S., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks (2021)
41. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenoctrees for real-time rendering of neural radiance fields (2021)
42. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields (2020)
43. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.R.: Image super-resolution using very deep residual channel attention networks. ArXiv **abs/1807.02758** (2018), https://api.semanticscholar.org/CorpusID:49657846
44. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2472–2481 (2018). https://doi.org/10.1109/CVPR.2018.00262

Supplementary Material

1 Additional Implementation Details

1.1 Feature extractor and pre-trained decoder

As described in Sec.4 of the main paper, we use a pre-trained feature extractor FE to provide SR priors to be distilled into radiance field and a pre-trained decoder D to map voxel features to distilled features. We follow the training procedure from [8] to train FE and D together. Under an autoencoder paradigm, an input image I of shape $\mathrm{H} \times \mathrm{W} \times 3$ is first encoded into a pixel-aligned feature map F of shape $\mathrm{H} \times \mathrm{W} \times \mathrm{C}$ by FE, where H and W is the height and width of the input image and C is the dimension of the feature. The decoder D then maps pixel-aligned feature to RGB value. The models can then be trained to minimize the L1 loss between the decoder output and the input image I. We choose RDN [44] as FE for its good trade-off of training efficiency and performance compared to other architectures and a decoder D of 5 MLP layers with ReLU activation function.

1.2 Preprocessing of dataset

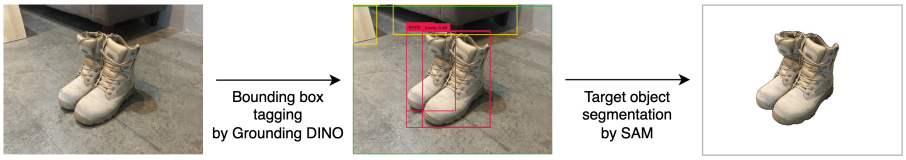

Fig. 1: Pipeline for dataset preprocessing: Given a raw training view, we first use Gounding DINO [22] to locate the target obejct, then utilize a segment anything model (SAM) [15] to segment and generate training view with object mask.

We train the generalizable VoxelGridSR model for all experiments with BlendedMVS [38] dataset, and test our method on a subset of 5 scenes, following all previous works. Before training VoxelGridSR model, we first reconstructed 40 radiance fields of scenes from BlendedMVS. We found that BlendedMVS mostly contain scenes of complex objects as well as complicated backgrounds, which may affect the quality of the reconstructed radiance field. To ensure that VoxelGridSR is trained with well-reconstructed radiance fields, we design a preprocessing pipeline to refine BlendedMVS data. Given a raw training view of scene s, we first use Grounding DINO [22] to tag the bounding box of target object. Then, a segment anything model (SAM) [15] is used to segment the object and generate training view with object mask. The processed images and camera poses are then used to reconstruct radiance fields for training VoxelGridSR. Pre-trained Grounding DINO and SAM are directly utilized for the procedure.

2 Multi-View Consistency

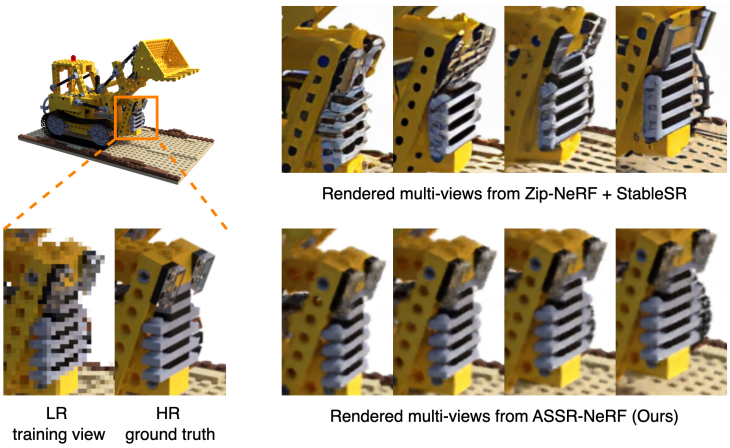

Fig. 2: Comparison of multi-view consistency: Super-resolving LR novel views from Zip-NeRF [3] by StableSR [34] leads to serious inconsistency across views from different camera poses. ASSR-NeRF can render HR novel views of consistent geometry and appearance. We encourage readers to visit our video showing the consistency issue at `https://drive.google.com/file/d/1h8WjmN7r1R79Cd4Q-dRLgbhToMZNR3pz/view`.

We provide additional qualitative comparisons about multi-view consistency in Fig. 2. While super-resolving rendered LR novel views from Zip-NeRF [3] leads to sharper results, it results in multi-view inconsistency, i.e., geometry and appearance across views from adjacent camera poses varies a lot. On the other hand, our method performs super-resolution novel view synthesis (SRNVS) with great consistency.

3 Analysis of Feature Distillation

In Sec.4.2. of the main paper, we mention that feature voxel grids from different scenes can have aligned same latent space by distilling features from the same teacher extractor into the radiance fields, VoxelGridSR can thus be trained in a unified latent space shared across voxel grids from all scenes and achieve generalizability. Fig. 3 shows the importance of distilled feature fields to training a generalizable VoxelGridSR. Trained with distilled feature fields, VoxelGridSR can reach generalizability and perform SR successfully on unseen scenes, as shown in the upper row. Without feature distillation, VoxelGridSR is trained with voxel-based radiance fields of diverse voxel feature distribution, and eventually fails to gain SR ability, as shown in the lower row of Fig. 3.

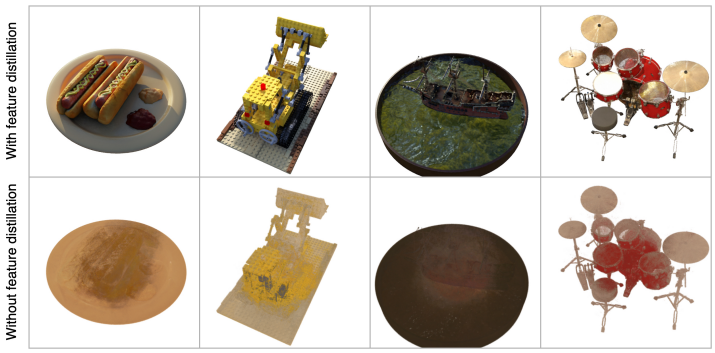

Fig. 3: Effectiveness of training VoxelGridSR with distilled feature fields: Shown in the upper row, VoxelGridSR model trained with distilled feature fields can achieve generalizability and perform SR on radiance fields of unseen scenes. Trained with radiance fields without feature distillation, VoxelGridSR fails to super-resolve and leads to corrupted novel views with incorrect colors, as shown in the lower row.

4 Additional Results

4.1 Bounded scenes

In Sec.5.2. of the main paper, we provide qualitative comparison with DVGO [30] and Zip-NeRF [3]. Here we provide additional qualitative results from more models, including TensoRF [7] and Instant-ngp [27], on Synthetic-NeRF [26] and BlendedMVS [38]. Fig. 4 shows that our method can render high-resolution novel views with richer and cleaner details.

4.2 Forward-facing scenes

In main paper, we conduct experiments on datasets of bounded scenes. These scenes are object-centric and have simple backgrounds [26, 38]. In this section, we provide results on LLFF [25], a dataset containing forward-facing scenes with complex objects and backgrounds. Following the same experiment settings from Sec.5.2. of the main paper, we compare our method with Zip-NeRF on LLFF. As shown in Tab. 1, ASSR-NeRF outperforms Zip-NeRF with high upscaling factor (x4). Fig. 5 also shows that our method can effectively improve the geometry and achieve cleaner appearances even when reconstructing scenes with complex backgrounds and objects.

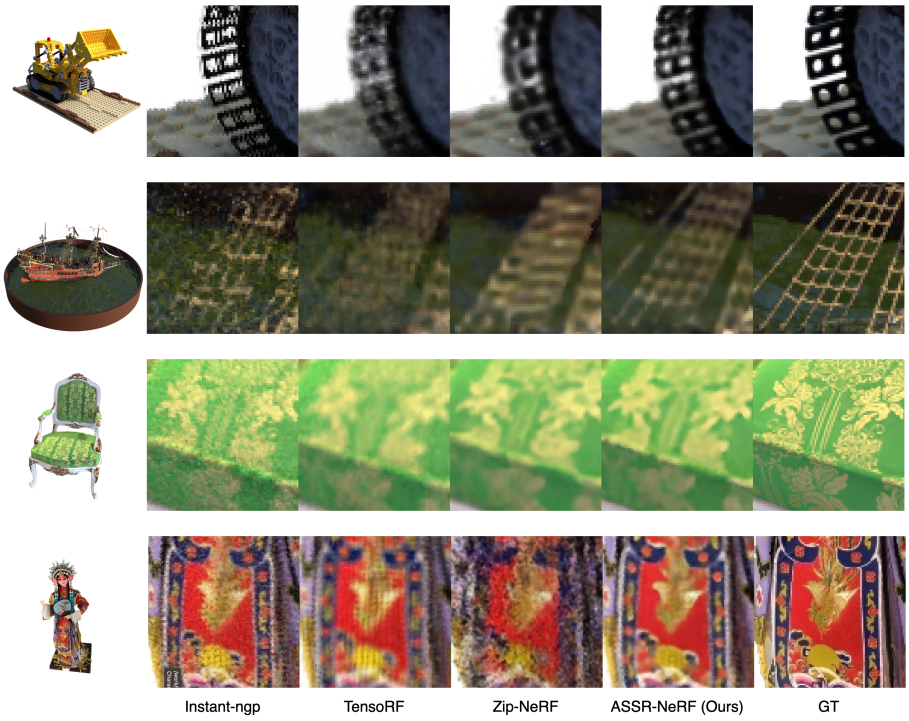

Fig. 4: Additional qualitative results: All other baselines are first reconstructed from LR training views, then perform HRNVS. For ASSR-NeRF, pre-trained VoxelGridSR model is applied to achieve SRNVS.

x4	PSNR↑	SSIM↑	LPIPS↓
Zip-NeRF	23.351	0.690	0.419
Ours(ASSR-NeRF)	**23.801**	**0.725**	**0.361**

Table 1: Quantitative results on LLFF. The experiment was trained on 252x189 image resolutions and tested on 1008x756.

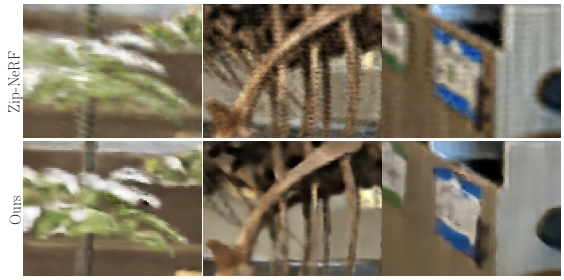

Fig. 5: Qualitative results on LLFF. Our method renders more realistic HR images in scenes with complex backgrounds.