Cross-Guided Optimization of Radiance Fields with Multi-View Image Super-Resolution for High-Resolution Novel View Synthesis

Youngho Yoon and Kuk-Jin Yoon
Visual Intelligence Lab., KAIST, Korea
{dudgh1732,kjyoon}@kaist.ac.kr

Abstract

Novel View Synthesis (NVS) aims at synthesizing an image from an arbitrary viewpoint using multi-view images and camera poses. Among the methods for NVS, Neural Radiance Fields (NeRF) is capable of NVS for an arbitrary resolution as it learns a continuous volumetric representation. However, radiance fields rely heavily on the spectral characteristics of coordinate-based networks. Thus, there is a limit to improving the performance of high-resolution novel view synthesis (HRNVS). To solve this problem, we propose a novel framework using cross-guided optimization of the single-image super-resolution (SISR) and radiance fields. We perform multi-view image super-resolution (MVSR) on train-view images during the radiance fields optimization process. It derives the updated SR result by fusing the feature map obtained from SISR and voxel-based uncertainty fields generated by integrated errors of train-view images. By repeating the updates during radiance fields optimization, train-view images for radiance fields optimization have multi-view consistency and high-frequency details simultaneously, ultimately improving the performance of HRNVS. Experiments of HRNVS and MVSR on various benchmark datasets show that the proposed method significantly surpasses existing methods.

1. Introduction

Novel View Synthesis (NVS) is an approach to synthesizing an image from an arbitrary viewpoint using multi-view images and camera poses. This is an essential task in computer vision and graphics, and it can be actively used in street-view navigation, AR/VR, and robotics. Recently, Neural Radiance Fields [28] (NeRF) significantly improved the performance of NVS by learning multi-layer perceptron (MLP) from 5d coordinate input. Since then, many studies have been conducted to shorten the long learning time of NeRF [4, 10, 29, 36, 42, 43, 48], increase the performance of NVS using depth priors [5, 8, 32, 41], and enable NVS from few-shot views [13, 16, 31, 49].

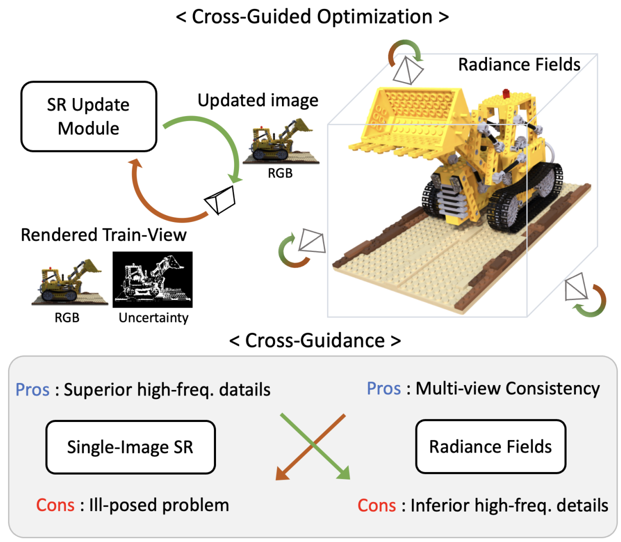

Figure 1. Cross-guided optimization between single image super-resolution and radiance fields. They complement weaknesses of one another with their respective strengths by using the SR update module, rendered train-view RGBs, and uncertainty maps.

Continuous scene representations such as NeRF [28] can be rendered at arbitrary resolution. Thus, there are many studies to improve the performance of multi-scale scene representation. Mip-NeRF [2] proposes scale-dependent positional encoding, which makes a network be trained on multiple scales. In addition, BACON [22] proposes a network capable of band-limited multi-scale decomposition by giving a constraint to the bandwidth of network outputs. Both papers showed significant down-scaling performance on volume rendering. On the other hand, NeRF-SR [39] improves the performance of high-resolution novel view synthesis (HRNVS) by learning in an unsupervised manner through super-sampling in the radiance fields optimization process.

Radiance fields have the ability to find scene geometry and optimize 5D functions simultaneously. Still, radiance fields have a low ability to perform super-resolution, and even if they synthesize high-resolution (HR) images, they only depend on the characteristic of continuous scene representation. On the other hand, single-image super-resolution (SISR) generally specializes in learning the inverse function of image degradation. Therefore, SISR could be beneficial to HRNVS by super-resolving train-view images; however, SISR is an ill-posed problem for which multiple solutions exist, and multi-view consistency cannot be maintained when multi-view images are processed separately.

To solve this problem, we propose a novel framework using cross-guided optimization between radiance fields and SISR. As shown in Fig. 1, our framework aims to ensure that radiance fields are guided by superior high-frequency details from SISR, and conversely, SISR is guided by multi-view consistency from radiance fields. We perform train-view synthesis during the radiance fields optimization process. Then, we generate voxel-based uncertainty fields to obtain uncertainty maps to find reliable regions in rendered train-view RGB images. The rendered train-view outputs and feature maps from the SISR network make it possible to do multi-view image super-resolution (MVSR) through the SR update module (SUM). Then, we continue optimizing the radiance fields using the updated SR outputs. Repeating the update process makes train-view images for radiance fields optimization have multi-view consistency and high-frequency details simultaneously, ultimately improving the performance of HRNVS.

Our method shows that the performance of HRNVS and MVSR on various benchmark datasets significantly surpasses existing methods. It also shows consistent performance improvements for various SISR models and radiance fields models in our method.

In summary, our contributions are as follows:

- We propose a novel framework for performing cross-guided radiance fields optimization using the SISR model for HRNVS.
- We propose voxel-based uncertainty fields to find reliable regions of synthesized images.
- We propose an SR update module (SUM) using voxel-based uncertainty fields and train-view synthesis outputs for MVSR.
- Experiments on various benchmark datasets show that the proposed method significantly surpasses existing methods in terms of performance for HRNVS and MVSR.

2. Related Work

2.1. Single-Image Super-Resolution

SISR aims to learn mapping functions between LR and HR image pairs. It has improved dramatically with the advent of learning-based methods using large-scale datasets. SRCNN [9] first proposes a learning-based SR framework using CNN, and after that, EDSR [21] and RCAN [50] suggested a deeper network structure using residual blocks and an attention mechanism respectively. Also, with the advent of transformer-based architecture [38] together, researches started to solve vision problems using the corresponding architecture, and SwinIR [19] improved the performance of SISR by using swin transformer [23]. However, SISR is an inherently ill-posed problem, and there is no unique solution, which causes the SR results to produce blurry images. To address this, some studies have improved the perceptual quality of SISR using discriminative networks [20, 33] and adaptive targets [15]. Still, reconstruction accuracy and perceptual quality of SISR are a trade-off. To solve this problem, our method proposes an SR update module that receives guidance from radiance fields and refines the results from SISR features.

2.2. Multi-Image Super-Resolution

Unlike SISR, there are studies that perform SR from multiple images. Video super-resolution (VSR) has the additional problem of exploiting the information from multiple frames of video with deep correlation. Some studies propose a sliding window framework to predict the optical flow of LR frames or a framework using a recurrent model architecture [3, 11, 18, 40]. Reference-based super-resolution (RSR) is an approach to improve the details of LR images through HR images given as reference images. Some studies propose a deformable convolution or the cosine similarity between the reference and LR images [14, 24, 45, 51]. Recently, a study proposes a method to perform MVSR, which generates HR reference images using given LR inputs and paired depths [7]. However, it requires depth maps as inputs, and since each image is processed independently, it is difficult to maintain multi-view consistency. Our method improves the performance of MVSR by updating the SR outputs during the process of optimizing the radiance fields from the given LR inputs. In addition, we demonstrate that our method is superior through quantitative comparison with existing methods for performing VSR and MVSR.

2.3. Multi-Scale Representations

Through the development of implicit neural representation [28, 35] (INR), various studies have been actively conducted to represent 2D images and 3D spaces as multi-scale representations. LIIF [6] and SphereSR [47] propose continuous image representations that enable SISR with an arbitrary resolution on planar and spherical images. In radiance fields, mip-NeRF [2] proposes a scale-dependent positional encoding, allowing for multiple-scale supervision. BACON [22] enables multi-scale decomposition without multi-scale supervision through bandwidth constraints. Both studies [2, 22] improve down-scaling performance in radiance fields. In contrast, NeRF-SR [39] proposes a super-sampling strategy that improves up-scaling performance by learning in an unsupervised manner. We present

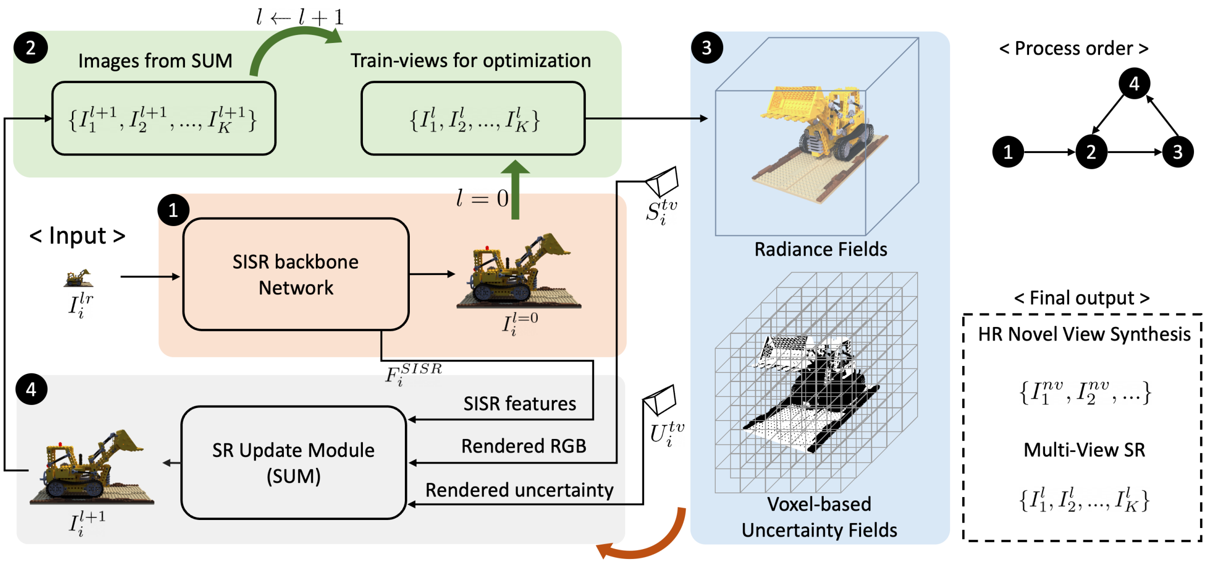

Figure 2. The overall framework for cross-guided optimization. When update step l is 0, $I_i^{l=0}$ is created through the SISR backbone network, and $\{I_i^{l=0}\}_i^K$ are used by optimizing radiance fields. During optimization, updated SR image I_i^{l+1} is generated through the SR update module (SUM) from rendered train-view RGB images S_i^{tv}, uncertainty map U_i^{tv}, and a SISR network feature F_i^{SISR}. The radiance fields optimization is continued using updated images. During optimization, the train-view image update is repeated.

a new methodology that can improve HRNVS performance through the interaction of 2D SISR and 5D radiance fields, breaking away from the current methodology that depends on the characteristics of INR.

2.4. Image Enhancement with Radiance Fields

Some studies improve NVS performance through radiance fields using physics-based multi-view geometry techniques for train-view images requiring image enhancement. NeRF-W [26] solves the problem of inputs with variable illumination and transient occluders by relaxing strict consistency assumptions. Deblur-NeRF [25] solves the problem of blurry input by developing a module that models spatially-varying blur kernels. RawNeRF [27] enables high-dynamic range (HDR) novel view synthesis by learning NeRF from raw data inputs and synthesizing raw output images. HDR-NeRF [12] makes exposure control and HDR image rendering possible by learning two implicit functions, radiance field and tone mapper. We propose a new method that finally enables HRNVS by allowing SR input images to be appropriately super-resolved during radiance fields optimization simultaneously. Unlike the image enhancement method using the existing radiance fields, we solve the problem by repeatably updating the train-view images which is the source of the radiance fields optimization.

3. Proposed Method

3.1. Preliminary

NeRF [28] trains an MLP network to estimate density σ and view-dependent color c from 3D position x and 2D direction d to perform 3D scene representation. It performs ray casting to estimate one-pixel value $\hat{C}(r)$ for any camera viewpoint. For each ray passing through the view-point and pixel, a total of N points are sampled, and the corresponding density and color are obtained through the MLP network. With this, we can get $\hat{C}(r)$ using the following equation:

$$\hat{C}(r) = \sum_{i=1}^{N} T_i \alpha_i c_i = \sum_{i=1}^{N} T_i (1 - e^{-\sigma_i \delta_i}) c_i \quad (1)$$

where $T_i = \Pi_{j=1}^{i-1} e^{-\sigma_i \delta_i}$ which is the accumulated transmittance along the ray from starting point to point i. Finally, we can train the NeRF model with a photometric loss as follows:

$$L_{photo} = \frac{1}{|R|} \sum_{r \in R} ||C(r) - \hat{C}(r)||_2^2 \quad (2)$$

Even after NeRF, a lot of studies have been conducted to improve the performance of volume rendering through various modelings such as voxel-based [4, 10, 36, 42], octree-based [48], and point-based [43] models. All models that use volume rendering also follow Eq. 1 and Eq. 2 by default.

3.2. Overview

As shown in Fig. 2, We propose a novel framework for cross-guided optimization (**CROP**) that simultaneously improves both performances by performing MVSR and radiance fields optimization complementary to each other. We propose voxel-based uncertainty fields to increase reliability in the process of performing cross-guided optimization (Sec. 3.3). We also propose an SR update module

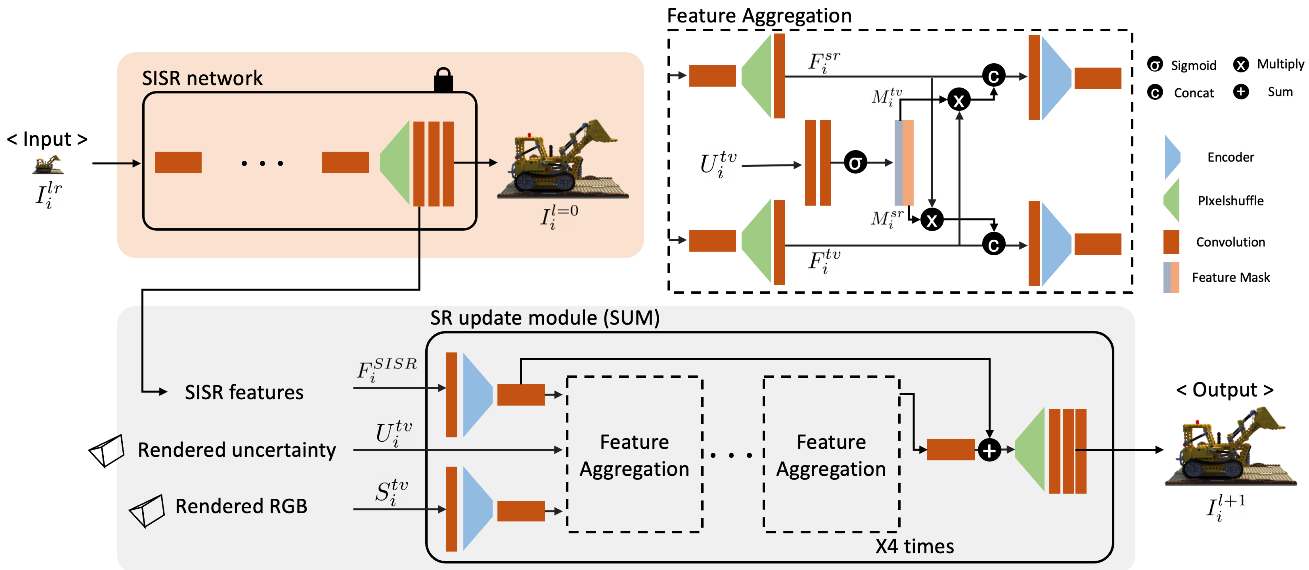

Figure 3. SR update module (SUM). SUM generates updated SR output through feature aggregation of the SISR feature map and rendered train-view RGB image using the rendered uncertainty map.

(SUM) that generates updated SR images through feature aggregation of rendered train-view RGB images from radiance fields and train-view images from the SISR network (Sec. 3.4). Finally, we introduce an optimization strategy for cross-guidance of SISR network and radiance fields using uncertainty fields and SUM (Sec. 3.5).

3.3. Voxel-based Uncertainty Fields

Our framework receives guidance from the train-view synthesis of radiance fields and performs updating SR results using SUM. Although RGB images reflecting the scene geometry of multi-view images can be obtained from train-view synthesis $\{S_i^{tv}\}_{i=1}^K$, train-view images $\{I_i^l\}_{i=1}^K$ for optimizing radiance fields are predicted SR outputs, not ground truth (GT) $\{I_i^{gt}\}_{i=1}^K$. Thus, synthesized train-view outputs cannot be trusted completely. Therefore, we try to generate the uncertainty map of the synthesized output through the following inference. If the integrated error of pixels rendered from rays passing through a certain 3D point is high, the rgb values sampled at that point have high uncertainty. We use this inference to generate voxel-based uncertainty fields $V^{(unc)} \in^{1\times N_x \times N_y \times N_z}$. As shown in Fig. 4, for a specific grid v_i inside a voxel-grid $V^{(unc)}$, there is a set of sampled ray points $P_i^{tv} = \{p_{i1}^{tv}, p_{i2}^{tv}, \dots\}$ inside the neighborhood voxels of v_i. If we set a corresponding rays of P_i^{tv} as $R_i^{tv} = \{r_{i1}^{tv}, r_{i2}^{tv}, \dots\}$, we can get the rendered rgb values $C_i^{tv} = \{c_{i1}^{tv}, c_{i2}^{tv}, \dots\}$ from $\{S_i^{tv}\}_{i=1}^K$ and train-view rgb values $\hat{C}_i^{tv} = \{\hat{c}_{i1}^{tv}, \hat{c}_{i2}^{tv}, \dots\}$ from $\{I_i^l\}_{i=1}^K$. Then, referring to the Eq. 1, we can derive the error e_{ij}^{tv} of the sampled point p_{ij}^{tv} as the following equation:

$$e_{ij}^{tv} = T_{ij}\alpha_{ij}||c_{ij}^{tv} - \hat{c}_{ij}^{tv}||_1 \tag{3}$$

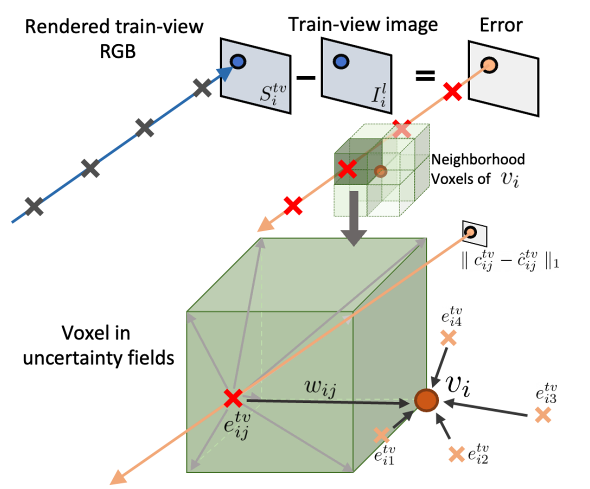

Figure 4. Voxel-based uncertainty fields. Each voxel-grid integrates the errors of adjacent sampling points.

where T_{ij} and α_{ij} are calculated by the radiance fields. Then, we propagate all estimated e_{ij}^{tv} into 8 neighborhood voxel grids. At this time, the value propagated by e_{ij}^{tv} to v_i is obtained as follows:

$$v_i = \sum_j e_{ij}^{tv} w_{ij} \; / \; \sum_j w_{ij} \tag{4}$$

where w_{ij} is the trilinear interpolation weight of p_{ij}^{tv} with respect to v_i. Eq. 4 can be derived for all points of the voxel grids at once through the gradient backward process. (More details can be found in the supplementary.) Based on the uncertainty fields obtained through Eq. 4, we can finally obtain the train-view uncertainty map using the following equation:

$$u^{tv} = \sum_{i=1}^{k} T_i \alpha_i e_i^k, \text{ where } e_i = f_{tri}(p_i, V^{(unc)}) \quad (5)$$

where u^{tv} is one pixel-uncertainty of one ray, and f_{tri} is a trilinear interpolation to the point p_i.

3.4. SR update Module (SUM)

As shown in Fig. 3, we try to improve SR results of SISR model by receiving guidance from the feature F_i^{SISR} of the SISR model with an LR image I_i^{lr} as an input, the rendered train-view RGB output S_i^{tv} of the radiance fields, and the rendered train-view uncertainty map U_i^{tv}. We propose an SR update module (SUM) that aims to derive the SR output I_i^{l+1} of higher quality than SISR output I_i^0 by delivering information about the reliable region in S_i^{tv} to F_i^{SISR} using U_i^{tv}. Therefore, we perform the feature aggregation module (FAM) serially a total of four times to achieve the goal of SUM. FAM initially makes two feature masks M_i^{tv} and M_i^{sr} from U_i^{tv}, and performs the feature aggregation as the following equations:

$$F_{i,out}^{sr} = F_i^{sr} \ \| \ (F_i^{tv} * M_i^{tv}) \quad (6)$$

$$F_{i,out}^{tv} = F_i^{tv} \ \| \ (F_i^{sr} * M_i^{sr}) \quad (7)$$

where $\|$ is concatenation, and $\{F_i^{sr}, F_i^{tv}\}$ are the intermediate features of $\{F_i^{SISR}, S_i^{tv}\}$. This makes it possible to share information in regions that need each other. In SUM, convolution operations for S_i^{tv} and F_i^{SISR} are performed with low-resolution scale features, but FAM for information sharing is performed with high-resolution scale features. The reason is to make it possible to transmit spatial information of S_i^{tv} including information about scene geometry to F_i^{SISR} exactly.

3.5. Cross-Guided Optimization Strategy

Dataset Generation for SUM. We use two large-scale NVS datasets, RTMV [37] and BlendedMVS [46] dataset, to train SUM. We conduct radiance fields optimization using DVGO [36] that dramatically reduces the optimization time and inference time of NeRF to use a large-scale NVS dataset. We created the radiance fields dataset by optimizing 60 scenes of RTMV and 40 scenes of BlendedMVS in advance. And then, we synthesized the rendered train-view RGB images, uncertainty maps, and paired LR/HR images $\{(S_i^{tv}, U_i^{tv}, I_i^{lr}, I_i^{gt})\}_i$ required for training the SUM.

Training SUM. Training SUM is performed using the dataset created by RTMV and BlendedMVS datasets. The SISR backbone network is frozen in order not to lose the generalizability of the model pretrained by a large-scale SISR dataset. The loss function used for training is:

$$L_{SUM} = ||I_i^{gt} - f_{sum}(S_i^{tv}, U_i^{tv}, I_i^{lr})||_1 \quad (8)$$

where $f_{sum}(*)$ is the estimated SR output through the SISR backbone network and SUM.

Optimization for the test set. After completing the training SUM, we finally proceed with the optimization of the radiance fields for test sets. As shown in Fig. 2, we firstly obtain SR images $\{I_i^0\}_{i=1}^K$ from the input LR images using the SISR backbone network and optimize the radiance fields by using $\{I_i^0\}_{i=1}^K$ as train-view images. During optimization, we update the train-view images from $\{I_i^l\}_{i=1}^K$ to $\{I_i^{l+1}\}_{i=1}^K$ using SUM and continue the radiance fields optimization from updated train-view images. The photometric loss used for optimization is as follows.

$$L_{photo} = \frac{1}{|R|} \sum_{r \in R} ||c^{tv}(r) - \hat{c}^{tv}(r)||_2^2 \quad (9)$$

where R is a set of rays for every pixel in every train-view image $\{I_i^l\}_{i=1}^K$. After optimization, we finally get high-quality HRNVS outputs $\{I_1^{nv}, I_2^{nv}, ...\}$ and MVSR outputs $\{I_i^l\}_{i=1}^K$ corresponding to $\{I_i^{lr}\}_{i=1}^K$.

4. Experiments

4.1. Datasets

Train and Validation set. We use a subset of 60 scenes of google scanned environment setting in RTMV [37] and a subset of 40 scenes in BlendedMVS [46] as a training dataset for the SR update module (SUM), a total of 100 scenes. Among 100 scenes, we split it into 86 scenes for training and 14 scenes for validation. We preprocess the resolution of the RTMV dataset to 800×800 and use it as GT, and we use preprocessed BlendedMVS dataset at 768×576 as GT. In addition, the scale factor for generating LR images is set to 4. The down-scaling process is performed by bicubic interpolation (imresize Matlab function) commonly used in the bicubic image SR datasets [1, 30, 44]. We make LR images into SR images using the SISR-backbone model and then perform optimization using the DVGO [36] model. Finally, we obtain the train-view synthesis to get RGBs and uncertainty outputs required for training the SUM.

Test set. We use a total of three datasets as test sets to evaluate HRNVS and MVSR. Syntheic-NeRF [28] consists of 8 scenes, and the resolution of each image is 800×800. BlendedMVS [46] is a synthetic dataset using realistic ambient lightning. We use a subset of 4 scenes in BlendedMVS, and the resolution of each image is 768×576. Finally, Tanks and Temple [17] is a real-world dataset. The resolution is 1920×1080, and we use a subset of 5 scenes. We generate train-view LR images by performing bicubic interpolation on scale factor 4 for train-view images as in the test set. Also, HR images are used as novel-view images GT for HRNVS and train-view images GT for MVSR.

Method	Type	Novel View Synthesis			Multi-View SR		
		PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
Synthetic NeRF Dataset							
LR	Unsupervised	28.5821	0.9211	0.1019	-	-	-
NeRF-SR [39]	Unsupervised	28.8960	0.9266	0.0992	-	-	-
EDSR [21]	Single Image	29.7931	0.9361	0.0820	31.7865	0.9500	0.0776
SwinIR [19]	Single Image	30.4878	0.9431	<u>0.0708</u>	33.1234	<u>0.9611</u>	0.0600
MVSRnet [7]	Multi-View Image	30.0039	0.9388	0.0773	31.9639	0.9544	0.0687
VRT [18]	Video	<u>30.5072</u>	<u>0.9435</u>	<u>0.0708</u>	<u>33.2345</u>	0.9544	<u>0.0599</u>
Ours(CROP)+EDSR	Multi-View Image	30.0351	0.9403	0.0729	32.1515	0.9563	0.0636
Ours(CROP)+SwinIR	Multi-View Image	**30.7140**	**0.9459**	**0.0671**	**33.7725**	**0.9644**	**0.0565**
GT	Unsupervised	31.9228	0.9562	0.0538	-	-	-
BlendedMVS Dataset							
LR	Unsupervised	25.2950	0.8475	0.1775	-	-	-
NeRF-SR [39]	Unsupervised	26.4342	0.8747	0.1635	-	-	-
EDSR [21]	Single Image	26.1210	0.8723	0.1585	28.7489	0.8941	0.1570
SwinIR [19]	Single Image	26.5167	0.8787	0.1492	29.2811	0.9034	0.1476
MVSRnet [7]	Multi-View Image	26.4726	0.8798	0.1474	29.0628	0.9034	0.1475
VRT [18]	Video	<u>26.5795</u>	<u>0.8848</u>	0.1458	<u>29.6916</u>	<u>0.9112</u>	<u>0.1405</u>
Ours(CROP)+EDSR	Multi-View Image	26.2594	0.8793	<u>0.1456</u>	28.7968	0.9015	0.1488
Ours(CROP)+SwinIR	Multi-View Image	**26.6914**	**0.8874**	**0.1405**	**29.7451**	**0.9126**	**0.1378**
GT	Unsupervised	28.0466	0.9225	0.1052	-	-	-
Tanks and Temples Dataset							
LR	Unsupervised	27.3093	0.9033	0.1578	-	-	-
NeRF-SR [39]	Unsupervised	26.7621	0.8869	0.1920	-	-	-
EDSR [21]	Single Image	28.4840	0.9144	0.1484	34.4941	0.9523	0.0938
SwinIR [19]	Single Image	28.5877	0.9157	0.1462	35.6951	<u>0.9604</u>	<u>0.0841</u>
MVSRnet [7]	Multi-View Image	28.5191	0.9148	0.1470	34.7496	0.9526	0.0975
VRT [18]	Video	<u>28.5973</u>	0.9159	0.1459	<u>35.7584</u>	0.9595	0.0854
Ours(CROP)+EDSR	Multi-View Image	28.5539	<u>0.9164</u>	<u>0.1444</u>	35.2112	0.9568	0.0897
Ours(CROP)+SwinIR	Multi-View Image	**28.6490**	**0.9176**	**0.1425**	**36.2430**	**0.9613**	**0.0808**
GT	Unsupervised	28.6749	0.9184	0.1427	-	-	-

Table 1. HR novel view synthesis results and multi-view image SR results on the Synthetic NeRF, BlendedMVS, Tanks and Temples dataset for X4 SR. **Bold** indicates the best results, and <u>underline</u> indicates the second best results.

4.2. Experimental Setup

HRNVS and MVSR performances of our framework depend on the performance of radiance fields. Therefore, we use DVGO [36] in all experiments to unify the radiance fields model. Experimental setups are largely divided into four types: unsupervised, SISR, MVSR, and VSR. Except for the unsupervised setup, all setups perform super-resolution on train-view images and then optimize the radiance fields.

Unsupervised setup. In this setup, there are three methods: LR, NeRF-SR [39], and GT. LR optimizes the radiance fields from train-view LR images and then performs HRNVS. NeRF-SR [39] uses the super-sampling loss proposed here. Additionally, we perform HRNVS using GT to obtain upper bounds for the performance of this task.

SISR setup. In this setup, there are two methods: EDSR [21] and SwinIR [19]. We obtain SR results for train-view LR images using the EDSR [21] model pretrained with DIV2K (800 images) and the SwinIR [19] model pretrained with the DIV2K+ Flickr2K dataset (2650 images). After that, the radiance fields optimization is performed using the corresponding results.

VSR setup. In this setup, one VRT [18] method is executed. In advance, we use the poses of the train-view images to form the video frame order of the images. We obtain the SR result of the LR images from the VRT [18] model pretrained with the Vimeo90K [44] (64612 seven-frame videos) and optimize the radiance fields using it.

MVSR setup. In this setup, three methods are performed: MVSRnet [7] and our two models using EDSR-backbone and SwinIR-backbone. Since MVSRnet [7] needs LR depth maps of train-view images, we use COLMAP [34] to extract depth maps or GT depth maps of the dataset. Through this, we train MVSRnet [7] using RTMV and BlendedMVS dataset as used in training SUM, and after training, we optimize the radiance fields using the SR results of the test set. Our framework using EDSR-backbone and SwinIR-backbone follows the training and cross-guided optimization strategy as described in Sec. 3.5.

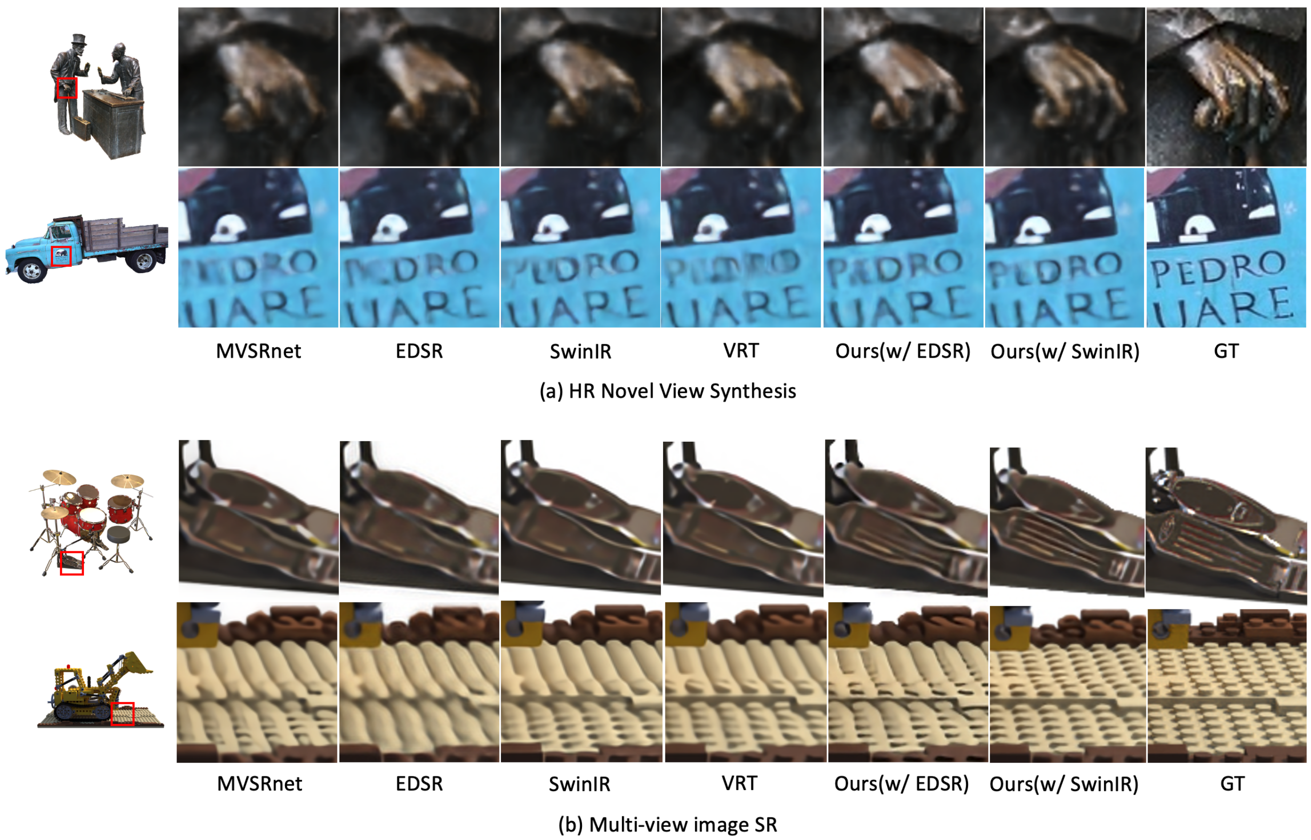

Figure 5. Qualitative comparisons of HR novel view synthesis (a) and multi-view image SR with scale factor as 4 (b) of different methods.

4.3. Quantitative Analysis

Table 1 shows the HRNVS and MVSR results obtained through various methods from the X4 LR inputs. We use the Synthetic NeRF, BlendedMVS, and Tanks and Temples dataset for evaluation. We use PSNR, SSIM, and LPIPS(VGG) as evaluation metrics.

HR novel view synthesis. The 3rd column of Table 1 shows the HRNVS results. Our method (SwinIR-backbone) has the highest performance in the three datasets and all three metrics. Also, for the two backbone models (EDSR/SwinIR) used in our method, there are performance improvements in all cases compared to when only EDSR/SwinIR is used. In particular, in the synthetic nerf dataset, both EDSR/SwinIR PSNR values increased by more than 0.2dB.

Multi-view image SR. The 4th column of Table 1 shows the MVSR results. In MVSR, our method (SwinIR-backbone) has the highest performance in all three datasets and three metrics. For the two backbone models (EDSR/SwinIR) used in our method, there are performance improvements in all cases compared to when only EDSR/SwinIR is used. In particular, our method (SwinIR-backbone) increases the PSNR value by more than 0.46dB for all datasets.

4.4. Qualitative Comparison and Analysis

HR novel view synthesis. As shown in Fig. 5 (a), our models generate accurate text images for novel views and synthesize clear textures. On the other hand, other models cannot create a clear text image and synthesize a blurry texture.

Multi-view image SR. As shown in Fig. 5 (b), our models generate a perfect texture and a small repeating pattern even though a small amount of information. On the other hand, other models generate irregular patterns and produce images with little or no texture.

Uncertainty map and error with GT. We qualitatively compare the uncertainty map and the error map with train-view RGB and GT for various views in the Lego scene of the synthetic nerf dataset. As shown in Fig. 6, high errors occur in the highly activated uncertainty region.

4.5. Ablation Study and Discussion

In Table 2, we perform an ablation study on the rendered train-view RGB output, the uncertainty map, and the number of SUM updates on the synthetic NeRF dataset. Also, in Table 3, we check generalizability by conducting experiments on other radiance fields model as TensoRF [4] other than DVGO [36].

Model	Rendered RGB	Rendered uncertainty	# Update	Novel View Synthesis			Multi-View SR		
				PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
1	x	x	0	30.4878	0.9431	0.0708	33.1234	0.9611	0.0600
2	x	x	1	30.4241	0.9428	0.0713	32.9417	0.9606	0.0617
3	✓	x	1	30.6501	0.9449	0.0686	33.6402	0.9631	0.0651
4	✓	x	2	30.6852	0.9455	0.0679	33.6279	0.9634	0.0641
5	✓	x	3	30.6824	0.9458	**0.0671**	33.5945	0.9635	0.0635
6	✓	✓	1	30.6554	0.9450	0.0686	33.7112	0.9637	0.0583
7	✓	✓	2	30.6978	0.9456	0.0679	33.7497	0.9641	0.0574
8	✓	✓	3	**30.7140**	**0.9459**	**0.0671**	**33.7725**	**0.9644**	**0.0565**

Table 2. Ablation study on Synthetic NeRF dataset for X4 SR. **Bold** indicates the best, and underline indicates the second best results.

Method	Novel View Synthesis			Multi-View SR		
	PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
Synthetic NeRF Dataset						
SwinIR	30.8246	0.9441	0.0693	33.1234	0.9611	0.0600
Ours(CROP)	**31.0368**	**0.9459**	**0.0664**	**33.8953**	**0.9649**	**0.0556**
BlendedMVS Dataset						
SwinIR	26.4551	0.8805	0.1422	29.2811	0.9034	0.1476
Ours(CROP)	**26.5315**	**0.8864**	**0.1329**	**29.5891**	**0.9109**	**0.1393**
Tanks and Temples Dataset						
SwinIR	28.5118	0.9151	0.1415	35.6951	0.9604	0.0841
Ours(CROP)	**28.5406**	**0.9161**	**0.1388**	**35.9486**	**0.9611**	**0.0813**

Table 3. Quantitative Results with TensoRF [4] on Synthetic NeRF, BlendedMVS, and Tanks and Temples datasets. **Bold** indicates the best results. Ours uses SwinIR [19] as the SISR backbone model.

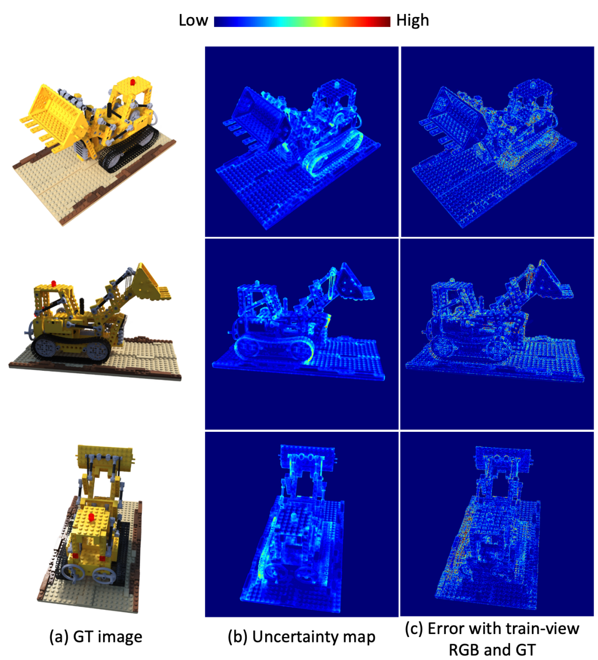

Figure 6. GT images (a), uncertainty maps (b), and error maps with train-view RGBs and GT images (c).

Rendered train-view RGB. As shown in models 1, 2, and 3 of Table 2, if there is no train-view rendered output, the performance is rather degraded when updated using SUM like model 2. On the other hand, if there is a train-view rendered output, both tasks have performance improvement as in model 3.

Rendered uncertainty map. Looking at models 5 and 8 of Table 2, by using the uncertainty map, there is a performance improvement of 0.03 dB in HRNVS performance and 0.20 dB in MVSR performance. In addition, when the uncertainty map is not used, it can be seen that the PSNR performance of MVSR drops sharply as the update proceeds, as shown in models 3, 4, and 5.

Number of updates. Looking at models 6, 7, and 8 of Table 2, when the rendered uncertainty map is used, the performances of HRNVS and MVSR are steadily improved as the update progresses. Through these results, we interpret that SUM extracts high-performance MVSR results through appropriate feature fusion using the uncertainty map at every update. In addition, it can be said that the performance of HRNVS is improved by maintaining multi-view consistency between MVSR results.

Generalizability for other radiance fields model. In Table 3, we can see the results of using TensoRF [4] instead of DVGO [36] for the radiance fields model for three test datasets. We did not perform any additional training SUM for TensoRF [4]. As can be seen in Table 3, it can be seen that there is a performance improvement for three metrics in both HRNVS and MVSR tasks. From these results, we analyze that our framework using SUM can be generalized to other radiance fields model without being biased to DVGO [36].

5. Conclusion

This paper proposes a novel framework for performing cross-guided radiance fields optimization using the SISR model for high-resolution novel view synthesis. We also propose an SR update module using voxel-based uncertainty fields and train-view synthesis results. Experiments on various benchmark datasets show that the proposed method significantly surpasses existing methods in terms of performance for high-resolution novel view synthesis and multi-view image super-resolution.

Acknowledgement. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636).

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 5

[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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 1, 2

[3] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021. 2

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 3, 7, 8

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 1

[6] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 2

[7] Ri Cheng, Yuqi Sun, Bo Yan, Weimin Tan, and Chenxi Ma. Geometry-aware reference synthesis for multi-view image super-resolution. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6083–6093, 2022. 2, 6

[8] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 1

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 2

[10] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 1, 3

[11] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019. 2

[12] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18398–18408, 2022. 3

[13] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 1

[14] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2103–2112, 2021. 2

[15] Younghyun Jo, Seoung Wug Oh, Peter Vajda, and Seon Joo Kim. Tackling the ill-posedness of super-resolution through adaptive target generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16236–16245, 2021. 2

[16] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12912–12921, 2022. 1

[17] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 5

[18] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 2, 6

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 2, 6, 8

[20] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2022. 2

[21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2, 6

[22] David B Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. Bacon: Band-limited coordinate networks for multiscale scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16252–16262, 2022. 1, 2

[23] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[24] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6368–6377, 2021. 2

[25] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12861–12870, 2022. 3

[26] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 3

[27] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022. 3

[28] 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. In *ECCV*, 2020. 1, 2, 3, 5

[29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 1

[30] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 5

[31] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 1

[32] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022. 1

[33] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE international conference on computer vision*, pages 4491–4500, 2017. 2

[34] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[35] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 2

[36] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 1, 3, 5, 6, 7, 8

[37] Jonathan Tremblay, Moustafa Meshry, Alex Evans, Jan Kautz, Alexander Keller, Sameh Khamis, Charles Loop, Nathan Morrical, Koki Nagano, Towaki Takikawa, and Stan Birchfield. Rtmv: A ray-traced multi-view synthetic dataset for novel view synthesis. *IEEE/CVF European Conference on Computer Vision Workshop (Learn3DG ECCVW), 2022*, 2022. 5

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[39] 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*, pages 6445–6454, 2022. 1, 2, 6

[40] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[41] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021. 1

[42] Liwen Wu, Jae Yong Lee, Anand Bhattad, Yu-Xiong Wang, and David Forsyth. Diver: Real-time and accurate neural radiance fields with deterministic integration for volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16200–16209, 2022. 1, 3

[43] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 1, 3

[44] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 5, 6

[45] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020. 2

[46] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020. 5

[47] Youngho Yoon, Inchul Chung, Lin Wang, and Kuk-Jin Yoon. Spheresr: 360deg image super-resolution with arbitrary projection via continuous spherical image representation. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5677–5686, 2022. 2

[48] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 1, 3

[49] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1

[50] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2

[51] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7982–7991, 2019. 2