

# NeRF-SR: High-Quality Neural Radiance Fields using Super-Sampling

Chen Wang<sup>1</sup>      Xian Wu<sup>2</sup>      Yuan-Chen Guo<sup>1</sup>      Song-Hai Zhang<sup>1</sup>  
 Yu-Wing Tai<sup>2,3\*</sup>      Shi-Min Hu<sup>1</sup>

<sup>1</sup> BNRIst, Department of Computer Science and Technology, Tsinghua University

<sup>2</sup>Kuaishou Technology      <sup>3</sup>The Hong Kong University of Science and Technology

## Abstract

We present NeRF-SR, a solution for high-resolution (HR) novel view synthesis with mostly low-resolution (LR) inputs. Our method is built upon Neural Radiance Fields (NeRF) [33] that predicts per-point density and color with a multi-layer perceptron. While producing images at arbitrary scales, NeRF struggles with resolutions that go beyond observed images. Our key insight is that NeRF has a local prior, which means predictions of a 3D point can be propagated in the nearby region and remain accurate. We first exploit it by a super-sampling strategy that shoots multiple rays at each image pixel, which enforces multi-view constraint at a sub-pixel level. Then, we show that NeRF-SR can further boost the performance of super-sampling by a refinement network that leverages the estimated depth at hand to hallucinate details from related patches on an HR reference image. Experiment results demonstrate that NeRF-SR generates high-quality results for novel view synthesis at HR on both synthetic and real-world datasets.

## 1. Introduction

Synthesizing photorealistic views from a novel viewpoint given a set of posed images, known as *novel view synthesis*, has been the long-standing problem in computer vision community, and an important technique for VR and AR applications such as navigation, and telepresence. More recently, *neural rendering* has made significant progress on novel view synthesis by leveraging learnable components with 3D geometry context to reconstruct input images. As the current state-of-the-art method, neural radiance fields (NeRF) [33] have emerged as a promising direction for scene representation even on sparse image sets of complex real-world scenes. NeRF uses the weights of multi-layer perceptrons (MLPs) to encode the radiance field and volume density of a scene. Most importantly, the implicit neural representation is continuous, which enables NeRF to take as input any position in the volume at inference time

and render images at any arbitrary resolution.

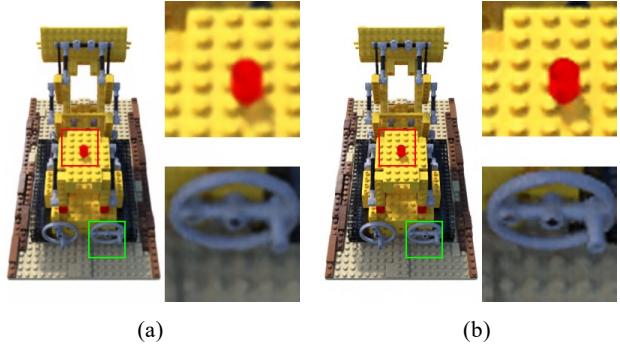


Figure 1. NeRF can synthesize photorealistic outputs with regard to the ground truth at the resolution of training images but struggles at synthesizing higher resolutions as shown in (a), while NeRF-SR produces high-quality novel views (b) even with low-resolution inputs.

A high-resolution 3D scene is essential for many scenarios, e.g. a prerequisite to providing an immersive virtual environment in VR. However, a NeRF trained on low-resolution (LR) images usually generates blurry views at higher resolution (See Figure 1). To resolve this problem, we present NeRF-SR, a technique that extends NeRF and creates high-resolution (HR) novel views with better quality even with LR inputs. First, we observe there is a sampling gap between the training and testing phase for super-resolving input images. To this end, we propose a super sampling strategy to better enforce the multi-view consistency embedded in NeRF and propagate the results of a single pixel to sub-pixels, enabling the generation of both SR images and SR depth maps. Second, when HR images are limited and unable for predicting camera parameters, we improve the outputs from the last stage and adopt a patch-wise warp-and-refine strategy with only one HR reference, which benefits from the estimated depth and extracts patches with fine details from the reference. To the best of our knowledge, we are the first to produce visually

pleasing results for novel view synthesis under mainly low-resolution images. Our method requires only the images of the target scene from which we dig into the internal statistics and does not rely on any external priors.

Our contributions are summarized as follows:

- an extension of NeRF that produces decent super-resolution results with mostly LR input images
- a super sampling strategy that bridges the sampling gap between training and testing stages in NeRF
- a refinement network that blends details from HR references by finding relevant patches with available depth maps

## 2. Related Work

**Novel View Synthesis.** Novel view synthesis can be categorized into image-based, learning-based, and geometry-based methods. Image-based methods warp and blend relevant patches in the observation frames to generate novel views based on measurements of quality [12, 22]. Learning-based methods predict blending weights and view-dependent effects via neural networks and/or other hand-crafted heuristics [6, 13, 43, 58]. Deep learning has also facilitated methods that can predict novel views from a single image, but they often require a large amount of data for training [36, 45, 51, 60, 66]. Different from image-based and learning-based methods, geometry-based methods first reconstruct a 3D model [48] and render images from target poses. For example, Aliev *et al.* [1] assigned multi-resolution features to point clouds and then performed neural rendering, Thies *et al.* [57] stored neural textures on 3D meshes and then render the novel view with traditional graphics pipeline. Other geometry representations include multi-planes images [10, 23, 24, 32, 55, 76], voxel grids [14, 18, 41], depth [10, 43, 44, 66] and layered depth [51, 61]. These methods, although producing relatively high-quality results, the discrete representations require abundant data and memory and the rendered resolutions are also limited by the accuracy of reconstructed geometry.

**Neural Radiance Fields.** Implicit neural representation has demonstrated its effectiveness to represent shapes and scenes, which usually leverages multi-layer perceptrons (MLPs) to encode signed distance fields [9, 37], occupancy [5, 31, 40] or volume density [33, 35]. Together with differentiable rendering [19, 27], these methods can reconstruct both geometry and appearance of objects and scenes [26, 35, 46, 52, 53]. Among them, Neural Radiance Fields (NeRF) [33] achieved remarkable results for synthesizing novel views of a static scene given a set of posed input images. The key idea of NeRF is to construct a continuous radiance field with MLPs and obtain images by differentiable volumetric rendering, thus the optimization process

can be done by minimizing photometric loss. There are a growing number of NeRF extensions emerged, *e.g.* reconstruction without input camera poses [25, 64], modelling non-rigid scenes [29, 38, 39, 42], unbounded scenes [72] and object categories [16, 59, 70]. NeRF has also been investigated to enable relighting [3, 54], generation [34, 49], editing [28, 68], 3D reconstruction [4, 65, 71].

Relevant to our work, Mip-NeRF [2] also considers the issue of *resolution* in NeRF. They showed that NeRFs rendered at various resolutions would introduce aliasing artifacts and resolved it by proposing an integrated positional encoding that featurize conical frustums instead of single points. Yet, Mip-NeRF only considers rendering with downsampled resolution. To our knowledge, there has not been any work that studies how to increase the resolution of NeRF.

**Image Super-Resolution** Our work is also related to image super-resolution. Classical approaches in single-image super-resolution (SISR) utilize priors such as image statistics [20, 77] or gradients [56]. CNN-based methods aim to learn the relationship between HR and LR images in CNN by minimizing the mean-square errors between SR images and ground truths [7, 8, 62]. Generative Adversarial Networks (GANs) [11] are also popular in super-resolution which hallucinates high resolution details by adversarial learning [21, 30, 47]. These methods mostly gain knowledge from large-scale datasets or existing high-resolution and low-resolution pairs for training. Besides, these 2D image-based methods, especially GAN-based methods do not take the view consistency into consideration and are sub-optimal for novel view synthesis.

Reference-based image super-resolution (Ref-SR) up-scales input images with additional reference high-resolution (HR) images. Existing methods match the correspondences between HR references and LR inputs with patch-match [74, 75], feature extraction [67, 69] or attention [69]. Inspired by their works, we also aim to learn HR details from the given reference image. However, these methods mostly super-resolve only one SR input based on the reference, while ours can refine details from all novel views with only one reference image.

## 3. Background

Neural Radiance Fields (NeRF) [33] encodes a 3D scene as a continuous function which takes as input 3D position  $\mathbf{x} = (x, y, z)$  and observed viewing direction  $\mathbf{d} = (\theta, \phi)$ , and predicts the radiance  $\mathbf{c}(\mathbf{x}, \mathbf{d}) = (r, g, b)$  and volume density  $\sigma(\mathbf{x})$ . The color depends both on viewing direction  $\mathbf{d}$  and  $\mathbf{x}$  to capture view dependent effects, while the density only depends on  $\mathbf{x}$  to maintain view consistency. NeRF is typically parametrized by a multilayer perceptron (MLP)  $f : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$ .

NeRF is an emission-only model (the color of a pixel only depends on the radiance along a ray with no other lighting factors). Therefore, according to volume rendering [17], an image can be rendered from a virtual camera at arbitrary positions by integrating color along the camera ray  $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$  that shots from the camera center  $\mathbf{o}$  in direction  $\mathbf{d}$ :

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt \quad (1)$$

where

$$T(t) = \exp\left(-\int_{t_n}^t \sigma(\mathbf{r}(t))dt\right) \quad (2)$$

is the accumulated transmittance that indicates the probability that a ray travels from  $t_n$  to  $t$  without hitting any particle.

NeRF is trained minimizing the mean-squared error (MSE) between the renderings of a ray  $\mathbf{r}$  and the corresponding ground-truth color:

$$\mathcal{L}_{\text{MSE}} = \sum_{p \in \mathcal{P}} \|\hat{\mathbf{C}}(\mathbf{r}(p)) - \mathbf{C}(\mathbf{r}(p))\|_2^2 \quad (3)$$

where  $\mathcal{P}$  denotes all pixels of the training set images,  $\mathbf{r}(p)$  denotes the ray shooting from camera center to the corners (or centers in some variants [2]) of pixel  $p$ .  $\hat{\mathbf{C}}(\mathbf{r}(p))$  and  $\mathbf{C}(\mathbf{r}(p))$  denotes the ground truth and predicted color.

In practice, the integral in Equation (1) is approximated by numeric quadrature that samples a finite number of points along with the rays and computes the summation of radiances according to the estimated per-point transmittance. The sampling in NeRF follows a *coarse-to-fine* mechanism with two MLPs, *i.e.* coarse network is queried on equally spaced samples whose outputs are utilized to sample another group of points for more accurate estimation and fine network is then queried on both groups of samples.

## 4. Approach

In this section, we introduce the details of NeRF-SR. The overall structure is presented in Figure 2. The super-sampling strategy and patch refinement network will be introduced in Section 4.1 and Section 4.2.

### 4.1. Super Sampling

NeRF optimizes a radiance field by enforcing multi-view color consistency and samples rays based on camera poses and pixels locations in the training set. Compared to the infinity possible incoming ray directions in the space, the sampling is quite sparse given limited input image observations. However, NeRF can still create plausible novel views because the output resolution is the same as the input one and it relies on the interpolation property of neural networks. This becomes a problem when we render an image at a higher resolution than training images, specifically,

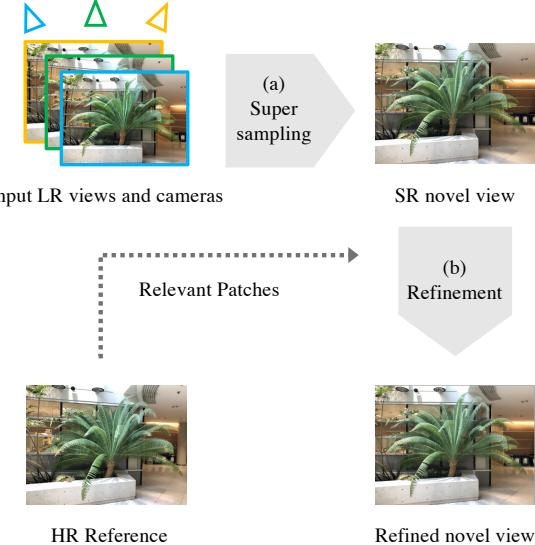


Figure 2. An overview of the proposed NeRF-SR that includes two components. (a), we adopt a super sampling strategy to produce super-resolution novel views from only low-resolution inputs. (b) Given an high-resolution reference at any viewpoint from which we utilize the depth map at hand to extract relevant patches, NeRF-SR generates more details for synthesized images.

there is a gap between the training and testing phase. Suppose a NeRF was trained on images of resolution  $H \times W$ , the most straightforward way to reconstruct a training image on scale factor  $s$ , *i.e.* an image of resolution  $sH \times sW$  is sampling a grid of  $s^2$  rays in an original pixel. Obviously, not only the sampled ray directions were never seen during training, but the pixel queried corresponds to a smaller region. Regarding this issue, we propose a super sampling strategy that tackles the problem of rendering SR images for NeRF. The intuition of super sampling is explained as follows and illustrated in Figure 3.

We start from the image formation process. Let  $\mathcal{R}(p)$  denotes the set of all possible ray directions for pixel  $p$  in training image, the color of  $p$  can be expressed as:

$$\mathbf{C}(p) = \mathbf{C}(\mathbf{r}(p)) = \underset{\mathbf{r}' \in \mathcal{R}(p)}{\text{Comp}}(\mathbf{r}'(p)) \quad (4)$$

where  $\text{Comp}$  is the composition and rasterization process for radiances of all incoming rays included in  $\mathcal{R}(p)$ . Although ideally the training ray directions should be sampled from  $\mathcal{R}(p)$ , it is challenging for the network to fit this huge amount of data. In practice, to super-resolve images at the scale of  $s$ , a pixel is evenly split into a  $s \times s$  grid sub-pixels  $\mathcal{S}(p)$  and ray directions for pixel  $p$  will be sampled from  $\mathcal{R}'(p) = \{\mathbf{r}(j) \mid j \in \mathcal{S}(p)\} \subset \mathcal{R}(p)$  for training. Thus, a  $sH \times sW$  image can be directly obtained by rendering the sub-pixels, erasing the sampling gap between the training

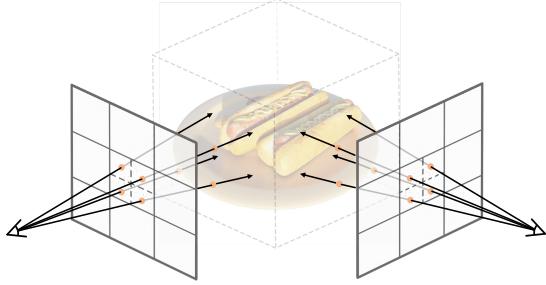


Figure 3. The super-sampling strategy splits a pixel (solid line) into multiple sub-pixels (dash line) and draws a ray for each sub-pixel. Thus, compared to vanilla NeRF, more 3D points in the scene can be corresponded and constrained.

and testing phase.

Another concern is how to perform supervision with only ground truth images at dimension  $H \times W$ . Similar to the blind-SR problem, the degradation process from  $sH \times sW$  is unknown and may be affected by many factors. To abort any assumptions on the downsampling operation and make it suitable for all situations, we averaged the radiances rendered from  $\mathcal{R}'(p)$  to perform supervision (“average” kernel):

$$\mathcal{L}_{\text{MSE}} = \sum_{p \in \mathcal{P}} \left\| \frac{1}{|\mathcal{R}'(p)|} \sum_{\mathbf{r}' \in \mathcal{R}'(p)} \hat{\mathbf{C}}(\mathbf{r}') - \mathbf{C}(\mathbf{r}(p)) \right\|_2^2 \quad (5)$$

Super-sampling fully exploits the cross-view consistency introduced by NeRF to a sub-pixel level, *i.e.* a position can be corresponded through multiple viewpoints. While NeRF only shoots one ray for each pixel and optimizes points along that ray, super-sampling constraints more positions in the 3D space and better utilize the multi-view information in input images. In other words, super sampling directly optimizes a denser radiance field at training time.

## 4.2. Patch-Based Refinement

With super-sampling, the synthesized image achieves much better visual quality than vanilla NeRF. However, when the images for a scene don’t have enough sub-pixel correspondence, the results of super-sampling cannot find enough details for high-resolution synthesis. On the other hand, often there are limited high-resolution images, *e.g.* a HR panoptic image or shots from devices, from which HR content are available for further improving the results.

Therefore, we present a patch-based refinement network to recover high-frequency details introduced by an HR reference input, as shown in Figure 4. The core design consideration focus on how to “blend” details on the reference image  $\mathcal{I}_{\text{REF}}$  into NeRF synthesized SR images that already captured the overall structure. We adopt a patch-by-patch

refine strategy that turns an SR patch  $\tilde{P}$  into the refined patch  $P$ . Other than  $\tilde{P}$ , the input should also include an HR patch from  $\mathcal{I}_{\text{REF}}$  that reveals how the objects or textures in  $\tilde{P}$  presents in high-resolution. However, due to occlusion and inaccuracy of depth estimation, multiple HR patches are required to cover the region in  $\tilde{P}$  and we use  $K$  patches  $\{P^{\text{REF}}\}_{k=1}^K$  for reference. Also, patches in  $\{P^{\text{REF}}\}_{k=1}^K$  cover larger regions than  $\tilde{P}$  and contain less relevant information.

We use a U-Net based convolutional architecture for the refinement network, which has demonstrated its efficacy in several existing novel view synthesis methods [6, 43, 44]. In earlier attempts, we model the refinement procedure as an image-to-image translation [15] and find channel-wise stack  $\tilde{P}$  and  $\{P^{\text{REF}}\}_{k=1}^K$  were unable to fit the training set perfectly. Therefore, inspired by [6, 43], we instead encode each patch respectively with an encoder consisting of seven convolutional layers. The decoder of the network takes as input the nearest-neighbor upsampled features from previous layers concatenated with both the encoded features of  $\tilde{P}$  and maxpooled features of  $\{P^{\text{REF}}\}_{k=1}^K$  at the same spatial resolution. All convolutional layers are followed by a ReLU activation.

**Training** The training of the refinement network requires SR and HR patch pairs, which are only available at the camera pose of  $\mathcal{I}_{\text{REF}}$ . Therefore,  $\tilde{P}$  is randomly sampled from the SR image and  $P$  is the patch on  $\mathcal{I}_{\text{REF}}$  at the same location. We perform perspective transformations to  $\tilde{P}$  and  $P$  as during testing, the input patches are mostly from different camera poses. Moreover, to account for the inaccuracy of reference patches at testing time, we sample  $\{P^{\text{REF}}\}_{k=1}^K$  within a fixed window around  $P$ . In order to preserve the spatial structure of  $\tilde{P}$  while improving its quality, our objective function combines reconstruction loss  $\mathcal{L}_{\text{rec}}$  and perceptual loss  $\mathcal{L}_{\text{per}}$ , where

$$\mathcal{L}_{\text{rec}} = \|\tilde{P} - P\|_1 \quad (6)$$

Note that we adopt  $l_1$ -norm instead of MSE because it is already minimized in super-sampling and  $l_1$ -norm will sharpen the results.

**Testing** At inference time, given a patch  $\tilde{P}$  on synthesized image  $\mathcal{I}_n$ , we can find a high-resolution reference patch on reference image  $\mathcal{I}_{\text{REF}}$  for each pixel on  $\tilde{P}$ :

$$P_{i,j}^{\text{REF}} = W(\tilde{P}_{i,j}, d_{i,j}) \quad (7)$$

where  $i, j$  denotes a location on patch  $\tilde{P}$ ,  $d$  is the estimated depth and  $W$  is the warping operation.  $W$  first computes the 3D coordinate of  $i, j$  based on  $d_{i,j}$ , project it on  $\mathcal{I}_{\text{REF}}$  and extract the reference patch according to camera intrinsics and extrinsics. Thus, to obtain  $P$ , we first sample  $K$  patches from  $\{P_{i,j}^{\text{REF}}\}$  to construct  $\{P^{\text{REF}}\}_{k=1}^K$  and then input them together with  $\tilde{P}$  into the network.

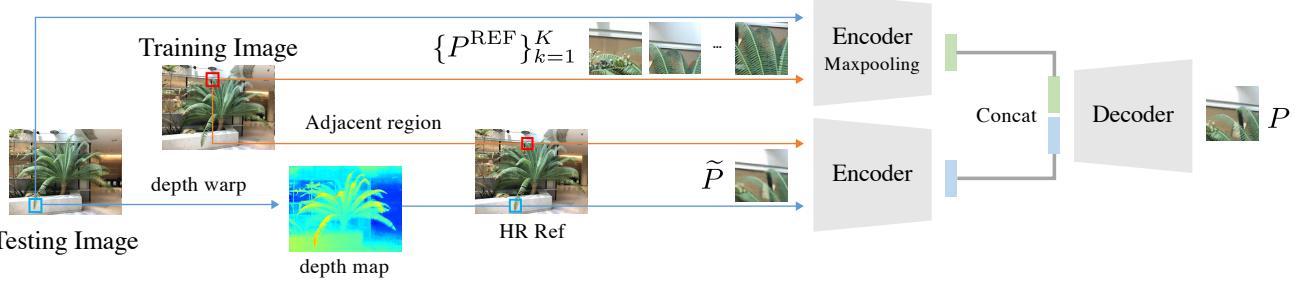


Figure 4. Our refinement module encodes synthesized patches  $\tilde{P}$  from images produced by super-sampling and reference patches  $\{P^{\text{REF}}\}_{k=1}^K$  from  $\mathcal{I}_{\text{REF}}$ . The encoded features of  $\mathcal{I}_{\text{REF}}$  are maxpooled and concatenated with that of  $\tilde{P}$ , which is then decoded to generate the refined patch. In the training phase,  $\tilde{P}$  is sampled from synthesized SR image at the camera pose of  $\mathcal{I}_{\text{REF}}$  and  $\{P^{\text{REF}}\}_{k=1}^K$  is sampled at adjacent regions. When testing,  $\{P^{\text{REF}}\}_{k=1}^K$  is obtained via depth warping. (The input and output patches are zoomed for better illustration, zoom in to see the details on leaves after refinement)

## 5. Experiments

In this section, we provide both quantitative and qualitative comparisons to demonstrate the advantages of the proposed NeRF-SR. We first show results and analysis of super-sampling, and then demonstrate how the refinement network adds more details to it. Our result only with super-sampling is denoted as Ours-SS and our result after patch-based refinement is denoted as Ours-Refine.

### 5.1. Dataset and Metrics

To evaluate our methods, we train and test our model on two datasets. We evaluate the quality of view synthesis with respect to ground truth from the same pose using three metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [63] and LPIPS [73].

**Blender Dataset** The Realistic Synthetic 360° of [32] (known as Blender dataset) contains 8 detailed synthetic objects with 100 images taken from virtual cameras arranged on a hemisphere pointed inward. As in NeRF [33], for each scene we input 100 views for training and hold out 200 images for testing.

**LLFF Dataset** LLFF dataset [32, 33] consists of 8 real-world scenes that contain mainly forward-facing images. We train on all of the images and report the average metrics on the whole set.

### 5.2. Training Details

We implement all experiments on top of NeRF [33] using PyTorch. As we train on different image resolutions independently, for fair comparison we train blender dataset and llff dataset for respectively 20 epochs and 30 epochs, where each epoch contains an iteration of the whole training set. We choose Adam as the optimizer (with hyperparameters  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ) with batch size set to 2048 (2048 rays a batch for all experimented scales) and learning rate

decayed exponentially from  $5 \cdot 10^{-4}$  to  $5 \cdot 10^{-6}$ . Following NeRF, NeRF-SR also uses a hierarchical sampling with a “coarse” and a “fine” MLP. The coarse samples and fine samples are both set to 64.

### 5.3. Effectiveness of super-sampling

For blender dataset, we super-sample on two resolutions:  $100 \times 100$  and  $200 \times 200$ , and test scales  $\times 2$  and  $\times 4$ . For the LLFF dataset, the input resolution is  $504 \times 378$  and we also upscale by  $\times 2$  and  $\times 4$ . The downscaling of images in the dataset from original resolution to training resolution is done by the default lanczos method in the Pillow package.

Figure 10 shows qualitative results for all methods on a subset of blender scenes. Renderings from bicubic exhibit correct global shapes but lack high-frequency details. Vanilla NeRF produces renderings that have more details than bicubic if the scene is already well-reconstructed at input resolution. However, it is still restricted by the information in the input image. NeRF-SR find sub-pixel level correspondence through super-sampling, which means missing details in the input can be found from other views that lie in the neighboring region in 3D space.

Quantitative results of blender dataset are summarized in Table 1. Directly training a NeRF on input images and rendering high-resolution outputs leads to quite poor results that cannot compete with bicubic interpolation. NeRF-SR outperforms NeRF and bicubic in all scenarios except slightly inferior SSIM for  $200 \times 200$  input. We urge the readers to inspect the visual results and see how bicubic tends to be heavily smoothed.

The qualitative and quantitative results for LLFF dataset are demonstrated in Figure 11 and Table 2 respectively. NeRF-SR fill in the details even on the complex scenes and outperforms other baselines significantly.

In Section 4.1, we mentioned that the supervision is performed by comparing the average color of sub-pixels due

Method	Blender×2 (100 × 100)			Blender×4 (100 × 100)			Blender×2 (200 × 200)			Blender×4 (200 × 200)		
	PSNR↑	SSIM↑	LPIPS↓									
Bicubic	27.83	0.938	0.135	26.03	0.894	0.222	30.43	<b>0.959</b>	0.088	28.44	<b>0.924</b>	0.154
NeRF [33]	24.92	0.865	0.122	23.28	0.846	0.194	25.97	0.883	0.114	24.60	0.877	0.161
Ours-SS	<b>29.77</b>	<b>0.946</b>	<b>0.045</b>	<b>28.07</b>	<b>0.921</b>	<b>0.071</b>	<b>31.00</b>	0.952	<b>0.038</b>	<b>28.46</b>	0.921	<b>0.076</b>

Table 1. Quality metrics for novel view synthesis on blender dataset. We report PSNR/SSIM/LPIPS for scale factors  $\times 2$  and  $\times 4$  on two input resolutions ( $100 \times 100$  and  $200 \times 200$ ) respectively. Though Bicubic achieves slightly better SSIM for resolution  $200 \times 200$ , we recommend readers to look into the renderings where our method produces better structure and details.

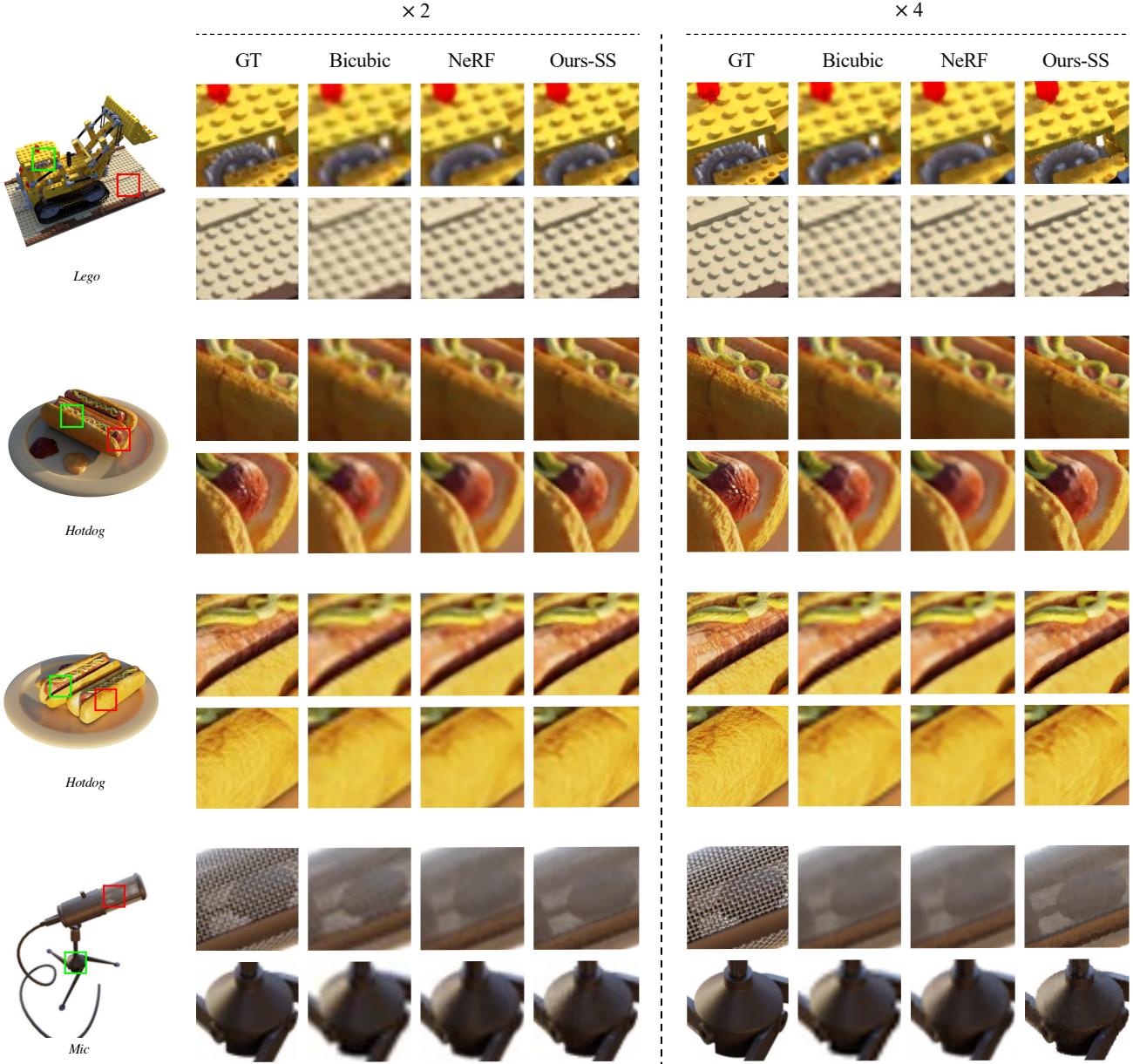


Figure 5. Qualitative comparison on blender dataset when the input images are  $200 \times 200$  and upscale by 2 and 4. Note how NeRF-SR recovers correct details through super-sampling even when inputting low-resolution images, such as *Lego*'s gears, *Hotdog*'s sausage and sauce, *Mic*'s magnets and shiny brackets. Note NeRF-SR is able to synthesize consistently over different viewpoints, here we provide two for *hotdog*, videos can be found in the supplementary material. Please zoom in for a better inspection of the results.



Figure 6. Qualitative comparison on LLFF dataset at an upscale of 4 between bicubic, NeRF, Ours-SS and Ours-Refine. NeRF-SR presents correct and clear texture in the leaves of *Leaves* and *Flowers* and fissures on *T-rex*'s ears and noses, which can further be enhanced using the refinement network. Please zoom in for better inspection of the results.

to the unknown nature of the degradation process. We further experiment on the condition that the degradation from high-resolution to input images is also average for blender data at resolution  $100 \times 100$ . Results show this symmetric downscale and upscale operation provides better renderings than asymmetric one (See Figure 7). PSNR, SSIM, LPIPs are all improved to 30.94 dB, 0.956, 0.023 for scale  $\times 2$  and 28.28 dB, 0.925 and 0.061 for  $\times 4$  respectively. The sensitivity to the degradation process is similar to that exhibited in single-image super-resolution.

#### 5.4. Refinement network

LLFF dataset contains real-world pictures that have a much more complex structure than the blender dataset, and super-sampling isn't enough for photorealistic renderings. We further boost its outputs with a refinement network introduced in Section 4.2. We use a fixed number of reference patches ( $K = 8$ ) and the dimensions of patches are set to  $64 \times 64$ . While inferencing, the input images are divided into non-overlapping patches and stitched together after re-

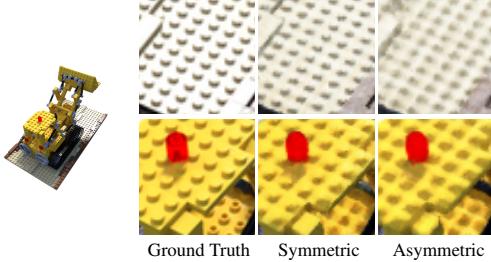


Figure 7. Visual results when both the downscale method and supervision signal are “average” (Input resolution is  $100 \times 100$  and upscaled by 4). Compared to asymmetric operation, symmetric averaged downscale and upscale produce more detailed super-resolution novel view synthesis.

finement. Without the loss of generosity, we set the reference image is to the first image in the dataset for all scenes, which is omitted when calculating the metrics.

The quantitative results of refinement can be found in Table 2. After refinement, metrics are improved substantially at the scale of 4. For the scale of 2, PSNR slightly decreases, a possible reason is that super-sampling already learns a decent high-resolution neural radiance fields for small upsampling and the refinement only improves subtle details *i.e.* noise points as in Figure 8, which influences PSNR negatively, but LPIPS is still promoted. However, the problem doesn’t occur for larger magnifications such as 4 since super-sampling derives much fewer details from low-resolution inputs.

We also demonstrate the renderings qualitatively before and after refining in Figure 11. It is clear to see that the refinement network boosts the outputs from super-sampling by a large margin.

Method	LLFF $\times 2$			LLFF $\times 4$		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Bicubic	26.67	<b>0.845</b>	0.234	24.75	0.727	0.437
NeRF [33]	23.30	0.663	0.296	24.47	0.701	0.388
Ours-SS	<b>27.30</b>	0.838	0.134	<b>25.13</b>	<b>0.730</b>	<b>0.244</b>
Ours-Refine	27.26	0.842	<b>0.103</b>	<b>25.59</b>	<b>0.759</b>	<b>0.165</b>

Table 2. Quality metrics for view synthesis on LLFF dataset. We report PSNR/SSIM/LPIPS for scale factors  $\times 2$  and  $\times 4$  on input resolutions ( $504 \times 378$ ).

## 6. Limitation

While NeRF-SR demonstrated its superior performance for constructing high-resolution neural radiance fields, it only achieves this with a pre-determined scale and not in an arbitrary scale manner, which limits its practical use. Super-sampling brings advantages for better view correspondence, but comes at the cost of computational efficiency. The refinement network also has deficiencies in that it can only provide details for regions already captured in the reference



Figure 8. An example of refinement comparison for magnification of 2 for flower. Although the PSNR is decreased from 22.50 dB to 22.22 dB after refinement, see how it mitigates the noisy points and improves visual quality.

image, although to provide strong priors often requires massive external data.

## 7. Conclusion

In this paper, we have presented NeRF-SR for novel view synthesis at high resolution. The motivation of our procedure is derived from our observation of local-prior in NeRF that comes in two aspects. The first is sub-pixel level multi-view consistency that encourages sub-pixel predictions to remain accurate, for which we propose a super sampling strategy that finds corresponding points through multi-views in sub-pixels. Second, we employ the estimated local depth to find relevant patches from an HR reference image from which a refinement network can blend details and produce better results. NeRF-SR is the first to tackle the problem of HR novel view synthesis and achieve more photorealistic renderings without any external data.

Our work also opens up a plethora of avenues for future work. First, we believe that the super-sampling strategy can be extended to be region sensitive to accelerate training. Second, it is also interesting to extend our approach to a generalized version that super-resolves novel views without

requiring additional pre-scene training.

We encourage the readers to watch the video to see multi-view renderings.

## A. Experimental Details

**Refinement Network** For the refinement network, during training, we first apply the same random perspective transformation to both the reference image  $\mathcal{I}_{\text{REF}}$  and synthesized SR image  $\mathcal{I}_{\text{SR}}$  at the same pose to generate additional 199 pairs to form the set  $\{(\mathcal{I}'_{\text{REF}}, \mathcal{I}'_{\text{SR}})\}_{i=1}^{200}$  for data augmentation. In each iteration, we first randomly select one pair  $(\mathcal{I}'_{\text{REF}}, \mathcal{I}'_{\text{SR}})$  and crop the input patch and ground truth patch both at position  $p$ . The reference patches are randomly sampled from the square centered at  $p$  with side length 128 at  $\mathcal{I}_{\text{REF}}$ . For upscale of  $\times 2$  and  $\times 4$ , the refinement network is trained for 3 and 8 epochs, each contains 500,000 patches.

## B. Additional Results

**Average Kernel** We present additional comparison between “average” kernel on blender dataset in Figure 9. It is clear that symmetric “average” kernel produces better results than asymmetric kernel.

**Additional Renderings** Figure 10 and Figure 11 shows additional static results on blender and LLFF dataset.

## C. Adversarial Training

The goal of the refinement process is to learn high-frequency details of reference images. Therefore, we have also considered using adversarial training [11] which has been employed in image-to-image translation [15], image super-resolution [50] for its ability to align one distribution to another. We treat NeRF as a generator and add an additional patch discriminator that takes as input patches from generator and reference images, expecting NeRF to generate images indistinguishable from HR references. However, the size of input patches are restricted by memory limitations and GAN struggles to provide meaningful guidance while introducing unignorable artifacts. The attempt to train a separate discriminator and freeze the weights when training with NeRF is also not successful.

## References

- [1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 696–712. Springer, 2020. [2](#)
- [2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. [2](#)
- [3] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020. [2](#)
- [4] 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. *arXiv preprint arXiv:2103.15595*, 2021. [2](#)
- [5] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. [2](#)
- [6] Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. Extreme view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7781–7790, 2019. [2, 4](#)
- [7] 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](#)



Figure 9. Additional visual results when both the downscale method and supervision signal are “average” (Input resolution is  $100 \times 100$  and upscaled by 4). Compared to asymmetric operation, symmetric averaged downscale and upscale produce more detailed super-resolution novel view synthesis.

Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *arXiv preprint arXiv:2103.13415*, 2021. [2, 3](#)

- [3] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020. [2](#)
- [4] 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. *arXiv preprint arXiv:2103.15595*, 2021. [2](#)
- [5] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. [2](#)
- [6] Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. Extreme view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7781–7790, 2019. [2, 4](#)
- [7] 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](#)

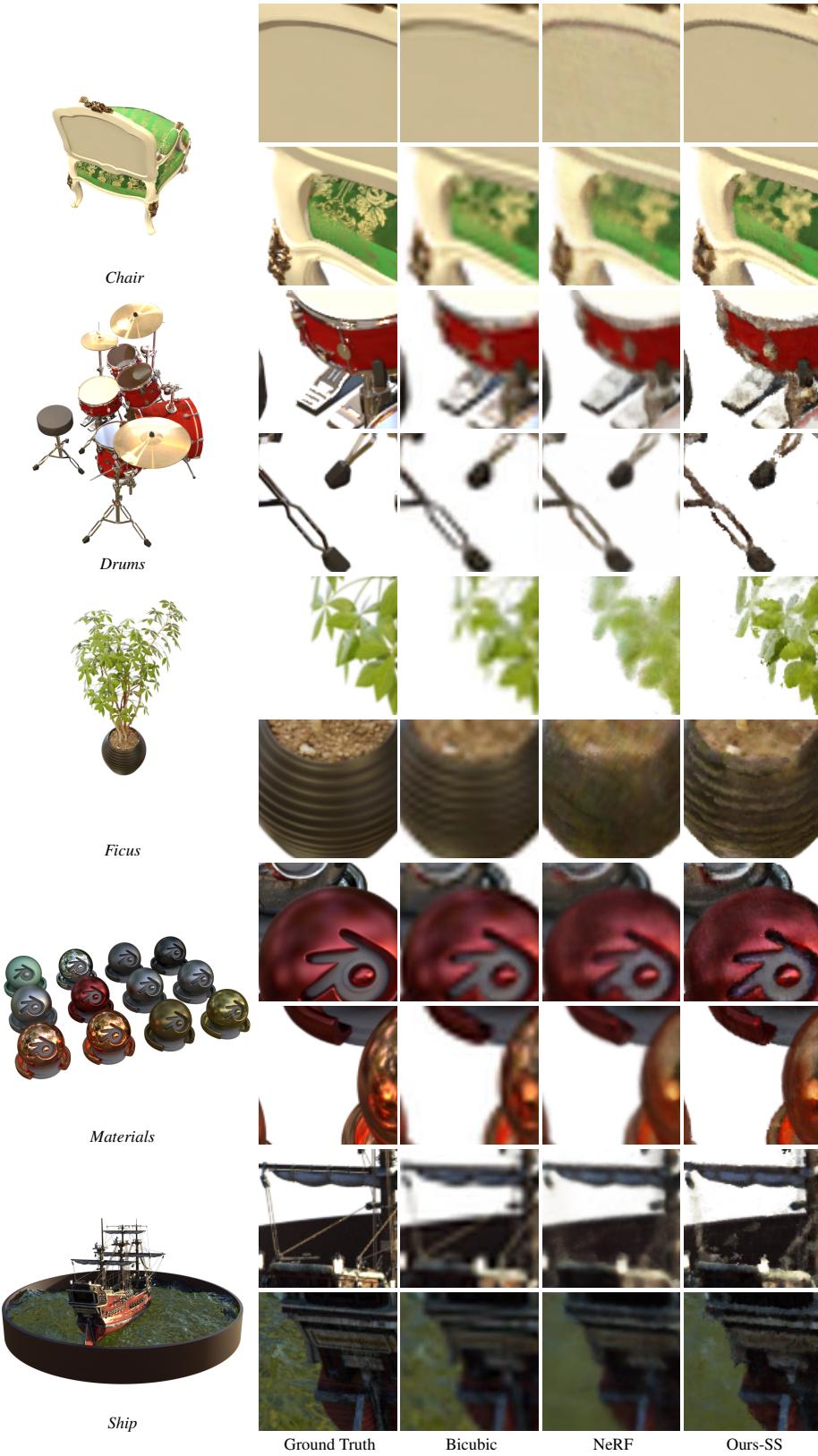


Figure 10. Additional Renderings on blender dataset for magnification of 4.

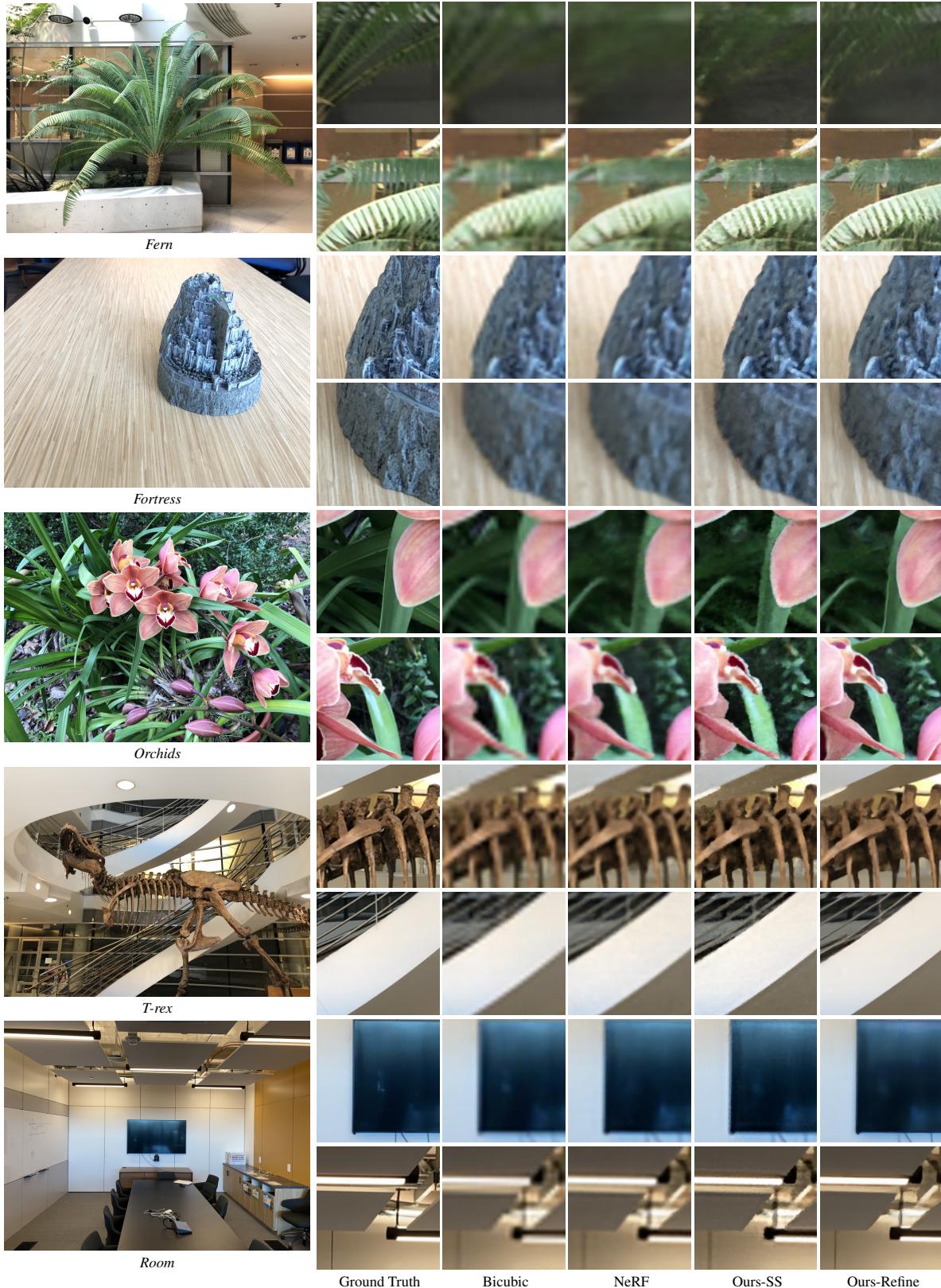


Figure 11. Additional Renderings on LLFF dataset for magnification of 4.

- [8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. [2](#)
- [9] Yueqi Duan, Haidong Zhu, He Wang, Li Yi, Ram Nevatia, and Leonidas J Guibas. Curriculum deepsdf. In *European Conference on Computer Vision*, pages 51–67. Springer, 2020. [2](#)
- [10] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019. [2](#)
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. [2, 9](#)
- [12] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54, 1996. [2](#)
- [13] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. [2](#)
- [14] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Learning a neural 3d texture space from 2d exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8356–8364, 2020. [2](#)
- [15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. [4, 9](#)
- [16] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. [2](#)
- [17] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. [3](#)
- [18] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)*, 35(6):1–10, 2016. [2](#)
- [19] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. [2](#)
- [20] Kwang In Kim and Younghée Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE transactions on pattern analysis and machine intelligence*, 32(6):1127–1133, 2010. [2](#)
- [21] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. [2](#)
- [22] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996. [2](#)
- [23] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021. [2](#)
- [24] Zhengqi Li, Wenqi Xian, Abe Davis, and Noah Snavely. Crowdsampling the plenoptic function. In *European Conference on Computer Vision*, pages 178–196. Springer, 2020. [2](#)
- [25] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. *arXiv preprint arXiv:2104.06405*, 2021. [2](#)
- [26] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *arXiv preprint arXiv:2007.11571*, 2020. [2](#)
- [27] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. [2](#)
- [28] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. *arXiv preprint arXiv:2105.06466*, 2021. [2](#)
- [29] 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. [2](#)
- [30] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. [2](#)
- [31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. [2](#)
- [32] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortíz-Cayón, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. [2, 5](#)
- [33] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. [1, 2, 5, 6, 8](#)
- [34] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields.

- In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2
- [35] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2
- [36] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics (TOG)*, 38(6):1–15, 2019. 2
- [37] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2
- [38] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2
- [39] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2
- [40] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 2
- [41] Eric Penner and Li Zhang. Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics (TOG)*, 36(6):1–11, 2017. 2
- [42] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2
- [43] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision*, pages 623–640. Springer, 2020. 2, 4
- [44] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12216–12225, 2021. 2, 4
- [45] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixel-synth: Generating a 3d-consistent experience from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14104–14113, 2021. 2
- [46] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2
- [47] 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
- [48] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2
- [49] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *arXiv preprint arXiv:2007.02442*, 2020. 2
- [50] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019. 9
- [51] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020. 2
- [52] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2
- [53] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *arXiv preprint arXiv:1906.01618*, 2019. 2
- [54] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2
- [55] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–184, 2019. 2
- [56] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2
- [57] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2
- [58] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. Image-guided neural object rendering. In *8th International Conference on Learning Representations*. OpenReview.net, 2020. 2
- [59] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 2
- [60] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF*

- Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 2
- [61] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 302–317, 2018. 2
- [62] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *Proceedings of the IEEE international conference on computer vision*, pages 370–378, 2015. 2
- [63] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multi-scale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 5
- [64] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2
- [65] 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. 2
- [66] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 2
- [67] Yanchun Xie, Jimin Xiao, Mingjie Sun, Chao Yao, and Kaizhu Huang. Feature representation matters: End-to-end learning for reference-based image super-resolution. In *European Conference on Computer Vision*, pages 230–245. Springer, 2020. 2
- [68] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13779–13788, 2021. 2
- [69] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Bainng 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
- [70] 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. 2
- [71] Jason Y Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. *arXiv preprint arXiv:2110.07604*, 2021. 2
- [72] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2
- [73] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5
- [74] 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
- [75] Haitian Zheng, Minghao Guo, Haoqian Wang, Yebin Liu, and Lu Fang. Combining exemplar-based approach and learning-based approach for light field super-resolution using a hybrid imaging system. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2481–2486, 2017. 2
- [76] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 2
- [77] M Zontak and M Irani. Internal statistics of a single natural image. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 977–984, 2011. 2