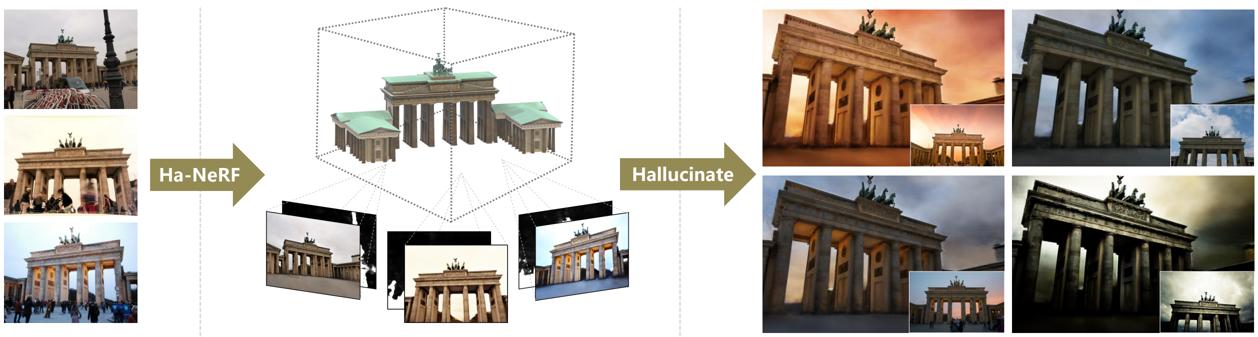
Hallucinated Neural Radiance Fields in the Wild

Xingyu Chen[1] Qi Zhang[2] Xiaoyu Li [2] Yue Chen[1]
Ying Feng[2] Xuan Wang[2] Jue Wang[2]
[1]Xi'an Jiaotong University [2]Tencent AI Lab



(a) input (b) training (c) inference

Figure 1. We recovery (b) a hallucinated neural radiance fields (Ha-NeRF) from (a) a group of tourism images with variable appearance and complex occlusions. Our method can consistently render (c) free-occlusion views which hallucinate different appearances.

Abstract

Neural Radiance Fields (NeRF) has recently gained popularity for its impressive novel view synthesis ability. This paper studies the problem of hallucinated NeRF: i.e. recovering a realistic NeRF at a different time of day from a group of tourism images. Existing solutions adopt NeRF with a controllable appearance embedding to render novel views under various conditions, but cannot render view-consistent images with an unseen appearance. To solve this problem, we present an end-to-end framework for constructing a hallucinated NeRF, dubbed as Ha-NeRF. Specifically, we propose an appearance hallucination module to handle time-varying appearances and transfer them to novel views. Considering the complex occlusions of tourism images, an anti-occlusion module is introduced to decompose the static subjects for visibility accurately. Experimental results on synthetic data and real tourism photo collections demonstrate that our method can not only hallucinate the desired appearances, but also render occlusion-free images from different views. The project and supplementary materials are available at https://rover-xingyu.github.io/Ha-NeRF/.

1. Introduction

In recent years, synthesizing photo-realistic novel views of a scene has become a research hotspot along with the rapid development of neural rendering technologies. Imagine you want to visit the Brandenburg Gate in Berlin and enjoy the landscapes at different times and weathers, but you cannot because of the coronavirus pandemic. For this hallucinated experience to be as engaging as possible, photo-realistic images from different views that can change with the weather, time, and other factors are necessary.

To achieve this, Neural Radiance Fields (NeRF) [33] and its following methods [25, 40, 56] have shown a remarkable capacity to recover the 3D geometry and appearance, giving the user an immersive feeling of physically being there. However, one significant drawback of NeRF is that they require a group of images without variable illumination and moving objects, *i.e.*, the radiance of the scene is constant and visible for each view. Unfortunately, most images of tourist landmarks are internet photos captured at different times and occluded by various objects. Most NeRF-based methods would integrate variable appearances and transient occluders into the 3D volume when they occur, which disturbs the real scene in the volume. How to synthesize the occlusion-free views from images with variable appearances and occluders remains to be solved.

Marting-Brualla *et al*. [28] attempt to tackle the aforementioned problem by proposing a NeRF in the Wild method (NeRF-W). They optimize an appearance embedding for each input image to address variable appearances and use a transient volume to decompose static components and their occlusion. Compared to NeRF, NeRF-W takes a step towards recovering a realistic world from tourism images with variable appearances and occluders. However, NeRF-W implements a controllable appearance by the optimized embeddings from train samples, making it impossible to hallucinate a novel view at an unlearned appearance. Furthermore, NeRF-W tries to optimize a transient volume for each input image with a transient embedding as input, which is highly ill-posed due to the randomness of transient occluders. And this leads to the inaccurate decomposition of the scene and further causes the entanglement of appearances and occlusion, *e.g*., results in the transient volume to remember the sunset glow.

To address these limitations, we present a hallucinated NeRF (Ha-NeRF) framework that can hallucinate the realistic radiance field from unconstrained tourist images with variable appearances and occluders, as shown in Fig. 1. For appearance hallucination, we propose a CNN-based appearance encoder and a view-consistent appearance loss to transfer photometric appearance while keeping consistent views. This design gives our method the flexibility to transfer the appearance of unlearned images. For anti-occlusion, we utilize an MLP to learn an image-dependent 2D visibility mask with an anti-occlusion loss that can automatically separate the static components with high accuracy during training. Experiments on several landmarks confirm the superior the proposed method outperforms others in terms of appearance hallucination and anti-occlusion.

Our contributions can be summarized as follows:

1. The Ha-NeRF is proposed to recover the appearance hallucination radiance fields from a group of images with variable appearances and occluders.

2. An appearance hallucination module is developed to transfer the view-consistent appearance to novel views.

3. An anti-occlusion module is modeled image-dependently to perceive the ray visibility.

2. Related Work

Novel View Synthesis. Rendering photo-realistic images is at the heart of computer vision and has been the focus of decades of research. Traditionally, view synthesis could be considered as an image-based warping task combined with geometry structure [47], such as implicit geometry from dense images [4, 10, 15, 24, 29] and explicit geometry [5, 11, 17, 18, 36]. Recent works have used a set of unconstrained photo collections to explicitly infer the light and reflectance of the objects in the scene [22, 46]. Others make use of semantic information to restore transient objects [39].

With the advancement of deep learning, many approaches have applied deep learning techniques to improve the performance of view synthesis. Researchers try to combine convolutional neural networks with scene geometry to predict depth or planar homography for novel view synthesis [7, 19, 26, 35, 54, 59]. Inspired by the layered depth images [45], recent works exploit explicit scene representation (*e.g*. multi-plane images, multiple sphere images) and render novel views using alpha-compositing for novel view synthesis [3, 12, 32, 49, 53, 58]. More recently, researchers focus on the challenging problem of learning implicit function (*e.g*. encoded features, NeRF) to represent scenes for novel view synthesis [33, 43, 44, 56].

Neural Rendering. Neural rendering [51] is closely related and combines ideas from classical computer graphics and deep learning to create algorithms for synthesizing image and reconstruction geometry from real-world observations. Several works present different ways to inject learning components into the rendering pipeline, such as learned latent textures [52], point clouds [1, 9], occupancy fields [31], signed distance functions [38]. Based on the image translation network, Meshry *et al*. [30] learn a neural re-rendering network conditioned on a learned latent appearance embedding module to recovery point cloud for view synthesis. However, the utilization of an image translation network leads to temporal artifacts visible under camera motion.

With the development of volume rendering [27, 33, 48], it is easy to render realistic and consistent views. Mildenhall *et al*. [33] propose Neural Radiance Fields (NeRF) and use a multi-layer perceptron (MLP) to restore a radiance field. Many following works try to extend NeRF to the dynamic scene [6, 25, 40, 56], fast training and rendering [8, 13, 42, 55] and scene edit [2, 28, 34, 57]. Marting-Brualla *et al*. [28] propose NeRF in the wild (NeRF-W) to optimize the appearance and tackle occlusion via static volume and dynamic volume respectively, but they failed in some scenes. Their dynamic volume is often used to describe the dramatic changes in appearance, such as view-dependent lighting. Besides, while NeRF-W implements a controllable appearance, it is hard to hallucinate consistent views at an appearance that has never been seen.

Appearance Transfer. A given scene can take on dramatically diverse appearances in different weather conditions and at different times. Grag *et al*. [14] propose that the dimensionality of scene appearance in tourist images captured at the same position is relatively low, except for outliers like transient objects. One can recover appearance for a photo collection by estimating coherent albedos across the collection [22], isolating surface albedo and scene illumination from the shape recovery [21], retrieving the sun's location through timestamps and geolocation [16], or assuming a fixed view [50]. However, these methods assume simple lighting models that do not apply to nighttime scene appearance. Radenovic *et al*. [41] restore distinct day and night reconstructions, but are unable to achieve a smooth gradation of appearance from day to night. Park *et al.* [37] propose an efficient technique to optimize the appearance of a collection of images depicting a common scene. Meshry [30] use a data-driven implicit representation of appearance that is learned from the input image distribution, while Martin-Brualla *et al.* [28] extend the data-driven method to NeRF and optimize appearance latent code for each view for appearance controllable. In contrast, the proposed method try to learn the appearance features that are decomposed from views, which means it could consistently hallucinate novel views at an unlearnt appearance.

3. Preliminary

We first introduce Neural Radiance Fields (NeRF) [33] that Ha-NeRF extends. NeRF represents a scene using a continuous volumetric function F_θ that is modeled as a multilayer perceptron (MLP). It takes a 3D location $\mathbf{x} = (x, y, z)$ and 2D viewing direction $\mathbf{d} = (\alpha, \beta)$ as input and output an emitted color $\mathbf{c} = (r, g, b)$ and volume density σ as:

$$\begin{aligned} (\sigma, \mathbf{z}) &= F_{\theta_1}(\gamma_{\mathbf{x}}(\mathbf{x})), \\ \mathbf{c} &= F_{\theta_2}(\gamma_{\mathbf{d}}(\mathbf{d}), \mathbf{z}), \end{aligned} \tag{1}$$

where $\theta = (\theta_1, \theta_2)$ are the MLP parameters, $\gamma_{\mathbf{x}}(\cdot)$ and $\gamma_{\mathbf{d}}(\cdot)$ are the positional encoding functions that are applied to each of the values in $\mathbf{x}$ and $\mathbf{d}$ respectively. To render the color of a ray passing through the scene, NeRF approximates the volume rendering integral using numerical quadrature. Let $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ be the ray emitted from the camera center $\mathbf{o}$ through a given pixel on the image plane. The approximation of the color $\hat{\mathbf{C}}(\mathbf{r})$ of the pixel is:

$$\begin{aligned} \hat{\mathbf{C}}(\mathbf{r}) &= \sum_{k=1}^{K} T_k(1 - \exp(-\sigma_k \delta_k))\mathbf{c}_k, \\ T_k &= \exp(-\sum_{l=1}^{k-1} \sigma_l \delta_l), \end{aligned} \tag{2}$$

where $\mathbf{c}_k$ and σ_k are the color and density at point $\mathbf{r}(t_k)$, $\delta_k = t_{k+1} - t_k$ is the distance between two quadrature points. Stratified sampling is used to select quadrature points $\{t_k\}_{k=1}^{K}$ between the near and far planes of the camera. Intuitively, alpha compositing with alpha values $1 - \exp(-\sigma_k \delta_k)$ can be interpreted as the probability of a ray terminating at the location $\mathbf{r}(t_k)$, and function T_k corresponds to the accumulated transmittance along the ray from the near plane to $\mathbf{r}(t_k)$.

To optimize the MLP parameters, NeRF minimizes the sum of squared errors between an image collection $\{\mathcal{I}_i\}_{i=1}^{N}$ and the corresponding rendered output. Each image $\mathcal{I}_i$ is registered with its intrinsic and extrinsic camera parameters which can be estimated using structure-from-motion algorithms. We precompute the set of camera rays $\{\mathbf{r}_{ij}\}$ at pixel

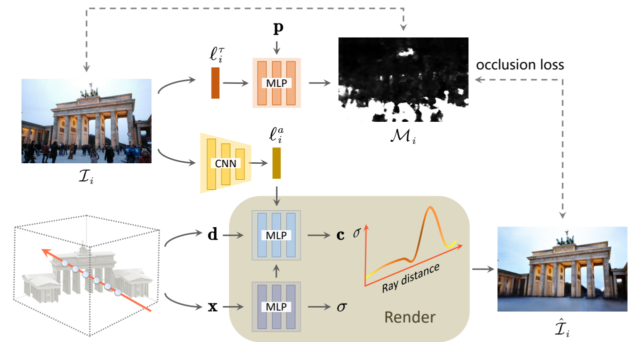


Figure 2. An overview of the Ha-NeRF architecture. Given an image $\mathcal{I}_i$, we use a CNN to encode it into an appearance latent vector ℓ_i^a. We synthesize images by sampling location $\mathbf{x}$ and viewing direction $\mathbf{d}$ of camera rays, feeding them with ℓ_i^a into MLPs to produce a color $\mathbf{c}$ and volume density σ and rendering a reconstructed image $\hat{\mathcal{I}}_i$. Given an image-dependent transient embedding ℓ_i^τ, we use an MLP to map pixel location $\mathbf{p}$ to a visible possibility $\mathcal{M}_i$, so that we can disentangle static and transient phenomena of the images with an occlusion loss.

j from image $\mathcal{I}_i$ with each ray $\mathbf{r}_{ij}(t) = \mathbf{o}_i + t\mathbf{d}_{ij}$ passing through the 3D location $\mathbf{o}_i$ with direction $\mathbf{d}_{ij}$. All the parameters are optimized by minimizing the following loss:

$$\mathcal{L} = \sum_{ij} \left\| \mathbf{C}(\mathbf{r}_{ij}) - \hat{\mathbf{C}}(\mathbf{r}_{ij}) \right\|_2^2, \tag{3}$$

where $\mathbf{C}(\mathbf{r}_{ij})$ is the observed color of ray j in image $\mathcal{I}_i$.

4. Method

Given a photo collection of a scene with varying appearances and transient occluders, the goal of our method is to reconstruct the scene that can be hallucinated from a new shot while handling the occlusion. That's to say that we can modify the appearance of the whole 3D scene according to a new view captured at a different photometric condition. Taking a photo in the wild as input, we reconstruct an appearance-independent NeRF modulated by an appearance embedding encoded by a convolutional neural network in Sec. 4.1. To address the transient occluders in the photo, we propose an occlusion handling module to separate the static scene automatically in Sec. 4.2. Fig. 2 illustrates the overview of the proposed architecture. Next, we subsequently elaborate on each module.

4.1. View-consistent Hallucination

To achieve the hallucination of a 3D scene according to a new shot from the input with varying appearances, the core problems are how to disentangle the scene geometry from appearances and how to transfer the new appearance to the reconstructed scene. NeRF-W [28] try to use an optimized

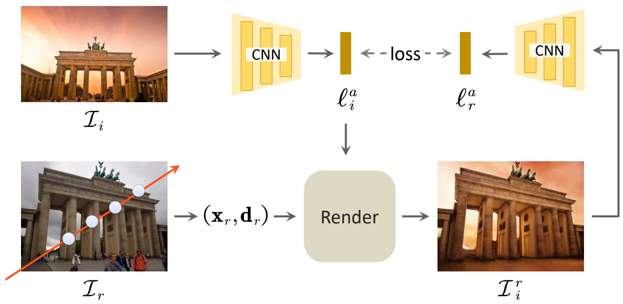


Figure 3. Illustration of view-consistent loss. Given an example image $\mathcal{I}_i$, we use a CNN to encode it into an appearance latent vector ℓ_i^a. We sample camera rays in another view to render the hallucinated image $\mathcal{I}_r^i$ together with ℓ_i^a. We encourage that the reconstructed appearance vector ℓ_r^a encoded from hallucinated image should be the same as ℓ_i^a, since it is a global representation across different views.

appearance embedding to explain the image-dependent appearances in the input. However, this embedding needs to be optimized during training, making it impossible to hallucinate the scene from a new shot beyond the training samples.

Therefore, we propose to learn the disentangled appearance representations using a convolutional neural network based encoder E_ϕ, of which parameters ϕ account for the varying lighting and photometric postprocessing in the input. E_ϕ encodes each image $\mathcal{I}_i$ into an appearance latent vector ℓ_i^a. The radiance $\mathbf{c}$ in Eq. 1 is extended to an appearance-dependent radiance $\mathbf{c}^{\ell_i^a}$, which introduces a dependency on appearance latent vector ℓ_i^a to the emitted color:

$$\mathbf{c}^{\ell_i^a} = F_{\theta_2}\left(\gamma_{\mathbf{d}}(\mathbf{d}), \mathbf{z}, \ell_i^a\right), \text{ where } \ell_i^a = E_\phi(\mathcal{I}_i). \tag{4}$$

The parameters ϕ of appearance encoder E_ϕ are learned alongside parameters θ of radiance field F_θ. This encoded appearance vector enables our method to have the flexibility to use the appearance of images beyond the training set.

However, the problem that disentangles the appearance from viewing direction with unpaired images is inherently ill-posed and requires additional constraints. Inspired by recent works [20, 23, 60] that exploit latent regression loss to encourage invertible mapping between image space and latent space, we propose a view-consistent loss $\mathcal{L}_{\text{v}}$ to achieve the disentanglement of appearance and view by taking an appearance vector $\ell_i^{(a)}$ from the the appearance encoder E_ϕ and attempt to reconstruct it in different views, which is formulated as:

$$\mathcal{L}_{\text{v}} = \left\|\mathrm{E}_\phi(\mathcal{I}_i^r) - \ell_i^a\right\|_1, \tag{5}$$

where $\mathcal{I}_i^r$ is the rendered image whose view is randomly generated and appearance vector is conditioned on the image $\mathcal{I}_i$ as shown in Fig. 3. Here we assume that the reconstructed appearance vector $\mathrm{E}_\phi(\mathcal{I}_i^r)$ should be the same as the original appearance vector ℓ_i^a, since the appearance vector is a global representation cross different views. Owing to the view-consistent loss, we can perform view-consistent appearance rendering, given the same appearance vector as input. In addition, we prevent encoding the image geometry content into the appearance vector with the help of view-consistent loss, which encodes the render images from different views (also content) to the same vector when conditioning the volume on the same vector.

To improve efficiency, we sample a grid of rays and combine them as the image $\mathcal{I}_i^r$ instead of rendering a whole image during training. And this is based on the assumption that the global appearance vector of an image will remain unchanged after sampling using a random grid.

4.2. Occlusion Handling

Instead of using a 3D transient field to reconstruct the transient phenomena which is only observed in an individual image as in [28], we eliminate the transient phenomena using an image-dependent 2D visibility map. This simplification makes our method has a more accurate segmentation between static scene and transient objects. To model the map, we employ an implicit continuous function F_ψ which maps a 2D pixel location $\mathbf{p} = (u, v)$ and an image-dependent transient embedding ℓ_i^τ to a visible possibility $\mathcal{M}$:

$$\mathcal{M}_{ij} = \mathrm{F}_\psi\left(\mathbf{p}_{ij}, \ell_i^\tau\right). \tag{6}$$

We train the visibility map to disentangle static and transient phenomena of the images which indicate the visibility of rays originated from the static scene in an unsupervised manner with an occlusion loss $\mathcal{L}_{\text{o}}$:

$$\mathcal{L}_{\text{o}} = \mathcal{M}_{ij}\left\|\mathbf{C}\left(\mathbf{r}_{ij}\right) - \hat{\mathbf{C}}\left(\mathbf{r}_{ij}\right)\right\|_2^2 + \lambda_{\text{o}}(1 - \mathcal{M}_{ij})^2. \tag{7}$$

The first term is the reconstruction error considering pixel visibility between the rendered and ground truth colors. Larger values of visible possibility $\mathcal{M}$ enhance the importance assigned to a pixel, under the assumption that it belongs to the static phenomena. The first term is balanced by the second, which corresponds to a regularizer with a multiplier λ_o on invisible probability, and this discourages the model from turning a blind eye to static phenomena.

4.3. Optimization

To achieve Ha-NeRF, we combine the aforementioned constraints and jointly train the parameters (θ, ϕ, ψ) and the per-image transient embedding $\{\ell_i^\tau\}_{i=1}^N$ to optimize the full objective:

$$\mathcal{L} = \lambda \sum_i \mathcal{L}_{\text{v}} + \sum_{ij} \mathcal{L}_{\text{o}}. \tag{8}$$

Figure 4. Qualitative results of experiments on constructed dataset. Ha-NeRF is able to encode the appearances and transfer them to novel views photo realistically (*e.g.* blue sky and sunshine in "Sacre Coeur", plants in "Brandenburg Gate", light reflection in "Trevi Fountain"). Besides, Ha-NeRF removes transient occlusions to render a consistent 3D scene geometry (*e.g.* square and pillars in "Brandenburg Gate").

	Brandenburg Gate			Sacre Coeur			Trevi Fountain		
	PSNR ↑	SSIM ↑	LIPIS ↓	PSNR ↑	SSIM ↑	LIPIS ↓	PSNR ↑	SSIM ↑	LIPIS ↓
NeRF [33]	18.90	.8159	.2316	15.60	.7155	.2916	16.14	.6007	.3662
NeRF-W [28]*	24.17	.8905	.1670	19.20	.8076	.1915	18.97	.6984	.2652
NeRF-A	22.93	.8517	.1727	19.57	.7864	.1839	19.89	.6798	.2377
NeRF-T	19.84	.8368	.1835	16.66	.7657	.2267	15.92	.6186	.2830
Ha-NeRF	24.04	.8773	.1391	20.02	.8012	.1710	20.18	.6908	.2225

* NeRf-W optimizes their appearance vector on the left half of each test image while Ha-NeRF hallucinate appearance that never seen before.

Table 1. Quantitative results of experiments on our constructed dataset. Ha-NeRF achieves competitive PSNR and SSIM while outperforming the others on LIPIS across all datasets, even with the unfair experiment settings when compared with NeRF-W.

5. Experiments

5.1. Implementation Details

The static neural radiance field F_θ consists of 8 fully-connected layers with 256 channels followed by ReLU activations to generate σ and one additional 128 channels fully-connected layer with sigmoid activation to output the appearance-dependent RGB color $\mathbf{c}$. The appearance encoder E_ϕ consists of 5 convolution layers followed by an adaptive average pooling and a fully-connected layer to get the appearance vector $\ell_i^{(a)}$ with 48 dimensions. The image-dependent 2D visibility mask F_ψ is modeled by 5 fully-connected layers of 256 channels followed by sigmoid activation to generate the visible possibility $\mathcal{M}$ conditioned on transient embedding ℓ_i^τ with 128 dimensions. We set λ to 1×10^{-3} and λ_o to 6×10^{-3}.

To evaluate the performance of Ha-NeRF in the wild, We construct three datasets called "Brandenburg Gate", "Sacre Coeur" and "Trevi Fountain" using the Phototourism dataset, which consists of internet photo collections of cultural landmarks. We downsample all the images by 2 times during training. For synthetic dataset, we build a Lego dataset with color and occlusion perturbation used for experiments requiring ground truth.

Figure 5. Hallucination in the "Brandenburg Gate" dataset with the global color shifts, such as weather, season and postprocessing filters. There are the images whose viewing direction is the same as the leftmost column content images, and the appearance is conditioned on the top line example appearance images.

5.2. Evaluation

Baselines. We evaluate our proposed method against NeRF, NeRF-W, and two ablations of Ha-NeRF: NeRF-A and NeRF-T. NeRF-A (appearance) builds upon our full model by eliminating the visibility network F_ϕ, while NeRF-T (transient) removes the appearance encoder E_ψ from the full model. Ha-NeRF is the complete model of our method.

Comparisons. We evaluate our method and baselines on the task of novel view synthesis. All methods use the same set of input views to train the parameters and embedding for each scene except NeRF-W, which uses the left half of each test image to optimize the appearance embedding for the test set since they can not hallucinate new appearance without optimizing during training. We present rendered images for visual inspection and report quantitative results based on PSNR, SSIM, LIPIS.

Fig. 4 shows qualitative results for all models and baselines on a subset of scenes. NeRF suffers from ghosting artifacts and global color shifts. NeRF-W produces more accurate 3D reconstructions and is able to model varying photometric effects. However, it still suffers from blur artifacts like the fog effect around the peristele of "Brandenburg Gate". This fog effect is the consequence of NeRF-W's attempt to estimate a 3D transient field to reconstruct the transient phenomena, while the transient objects are only observed in a single image. At the same time, renderings from NeRF-W also tend to exhibit different appearances compared to the ground truth, such as sunshine and blue sky in"Sacre Coeur" and light reflection in "Trevi Fountain".

NeRF-A has a more consistent appearance, such as the blue sky at the top of "Sacre Coeur". However, it is unable to reconstruct high-frequency details due to the occlusion. In contrast, NeRF-T is able to reconstruct structures with occlusion such as the square of "Brandenburg Gate", but is

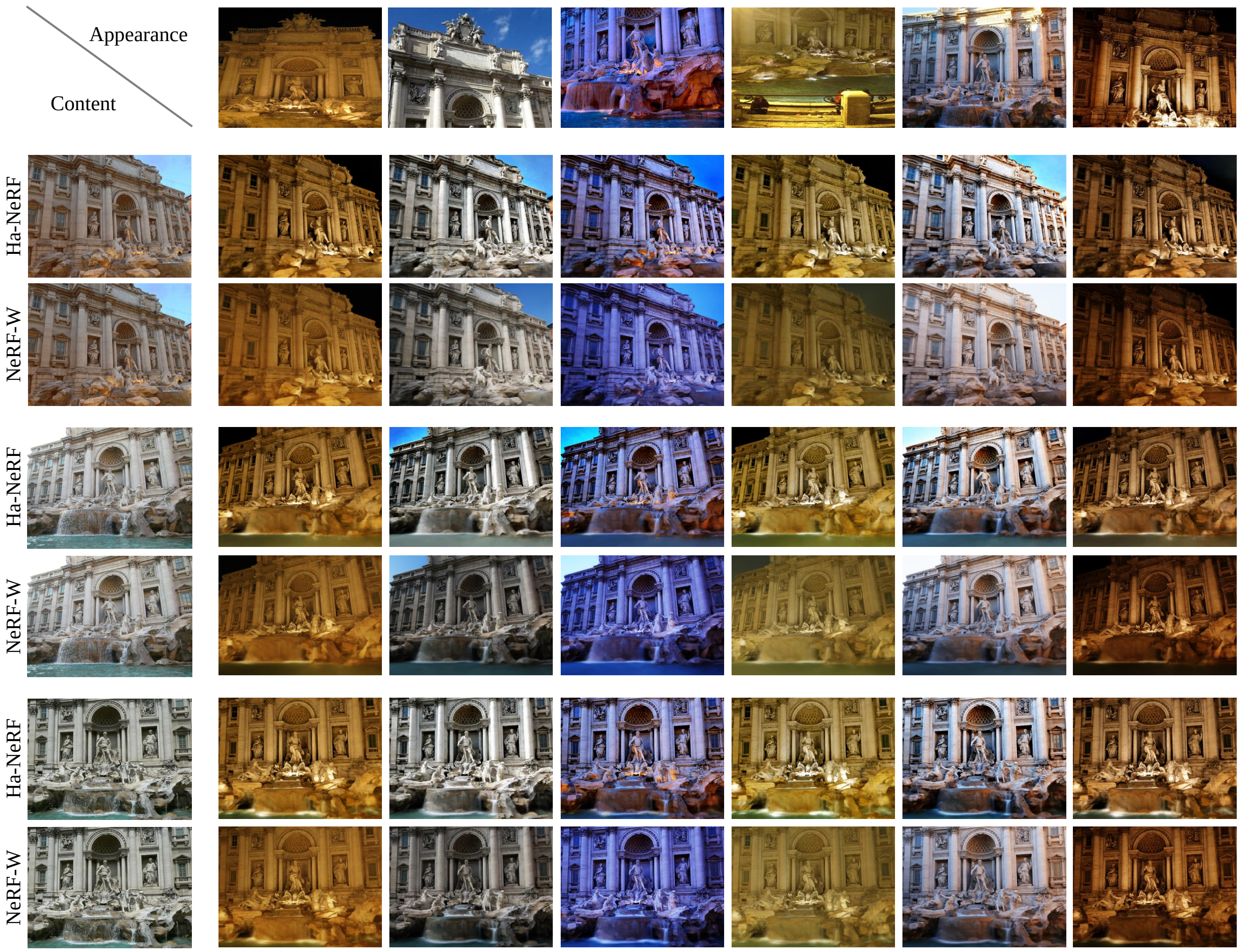


Figure 6. Hallucination in the “Trevi Fountain” dataset with high-frequency information of appearance, such as sunshine and colored light reflection. There are the images whose viewing direction is the same as the leftmost column content images, and the appearance is conditioned on the top line example appearance images.

Figure 7. Images rendered from a fixed camera position with interpolated appearance between appearance 1 and appearance 2. We denote NeRF-W w/T as simultaneously rendering the transient field and the static field of NeRF-W.

unable to model varying photometric effects. Ha-NeRF has the benefits of both ablations and thereby produces better appearance and anti-occlusion renderings.

Quantitative results are summarized in Table 1. Optimizing NeRF on photo collections in the wild leads to particularly poor results that cannot compete with NeRF-W. In contrast, Ha-NeRF achieves competitive PSNR and SSIM compared to NeRF-W while outperforming the others on LIPIS across all datasets. Actually, this comparison is unfair to us. To transfer the appearance from test images, NeRF-W needs to optimize the appearance vectors on a subset of the test images during training. While Ha-NeRF does not use any test images during training. When testing, Ha-NeRF can directly encode the image appearance by a learned encoder. Despite this, our method still can produce competitive results compared with NeRF-W. Moreover, NeRF-W exhibits view inconsistency. As the camera moves, renderings conditioned on the same appearance embedding appear to have an inconsistent appearance, which can not be reflected by current metrics.

Appearance Hallucination. By conditioning the color on the latent vector $\ell_i^{(a)}$, we can modify the lighting and appearance of a rendering without altering the underlying 3D geometry. In the meantime, encoding appearance with an encoder E_ϕ allows our framework to perform example-guided appearance transfer. The appearances transferred to the output are controlled by a user-provided example image even never been seen before.

In Fig. 5, we see rendered images produced by Ha-NeRF using different appearance vectors extracted from example images. We also show the results of NeRF-W where appearance vectors are optimized during training. Notice that Ha-NeRF hallucinates realistic images while NeRF-W suffers from global color shifts (such as weather, season and postprocessing filters) compared with the example images. Moreover, Fig. 6 shows that Ha-NeRF can capture the high-frequency information of appearance and hallucinate the sunshine and colored light reflection of the scene.

Ha-NeRF can also interpolate the appearance vectors to get other hallucinations. In Fig. 7, we present five images rendered from a fixed camera position, where we interpolate the appearance vectors encoded from the leftmost and rightmost images. Note that the appearance of the rendered images is smoothly transitioned between the two endpoints by Ha-NeRF. However, the interpolated results of NeRF-W completely ignore the sunset glow. Furthermore, We render the transient field of NeRF-W (NeRF-W w/T), which shows the sunset glow. It reveals that NeRF-W could not disentangle the variable appearance (sunset glow) from transient phenomena (people) well.

The components of Ha-NeRF are designed to deal with specific problems, *i.e.* color hallucination and anti-occlusion. Unfortunately, the uncontrolled captures of the Phototourism dataset make it challenging to demonstrate the effectiveness of each component. For this reason, we present an ablation study in which we construct variations (*e.g.* color and occlusion) of a synthetic dataset used in [28]. We manually introduce the phenomena we expect to find in in-the-wild imagery. For fair comparisons, the inputs of NeRF-W are the same as Ha-NeRF, and thus the appearances of test images are obtained from another image with the same appearance in the training data. Tab. 2 and Fig. 8 shows the quantitative and qualitative results of ablation study about color hallucination and anti-occlusion. The proposed method outperforms the baselines. Specifically, for anti-occlusion, our method tackles the occlusions very well, which verifies the accuracy of our image-dependent occlusion. For color hallucination, the performance of the proposed method is better than baselines, especially on the color shifts and highlight regions with high-frequency information, as shown in Fig. 8. For the experiment on the combination of color and occlusion, the performance of the proposed method shows that our method could decompose the occlusion and appearance well compared with NeRF-W. All of the results demonstrate the superior performance of the proposed method.

Figure 8. Example dataset perturbations and rendering from NeRF, NeRF-W, Ha-NeRF and ground-truth for fair comparisons. **Better viewed on screen with zooming in.**

Cross-Appearance Hallucination. Furthermore, we can perform appearance transfer by a user-provided example image from a different dataset. As shown in Fig. 9, we can hallucinate new appearance for "Trevi Fountain" condition on the example image of "Brandenburg Gate", vice versa (Fig. 10). We note that NeRF-W inherently can not hallucinate an appearance from other datasets because NeRF-W needs to optimize the appearance vectors on the example images, which must depict the same place.

View-consistent Hallucination. As the camera moves, ren-

Figure 9. Cross dataset hallucination in the "Trevi Fountain" condition on the example image of "Brandenburg Gate" with the global color shifts, such as weather, season, and postprocessing filters. There are the images whose content is the same as the leftmost column content images, and the appearance is conditioned on the top line example appearance images of another dataset.

Figure 10. Cross dataset hallucination in the "Brandenburg Gate" condition on the example image of "Trevi Fountain" with the global color shifts, such as weather, season, and postprocessing filters. There are the images whose content is the same as the leftmost column content images, and the appearance is conditioned on the top line example appearance images of another dataset.

derings of NeRF-W conditioned on the same appearance embedding appear to have an inconsistent appearance. On the contrary, we can perform view-consistent appearance rendering thanks to our view-consistent loss. Because the aligned image pairs are challenging to collect (tuples of images captured in different views under the same appearance conditions), we use the synthetic dataset [28] with ground truth and manually introduce color perturbation for training. The left-top of Fig. 11 shows that both NeRF-W and Ha-NeRF can render images with the correct appearance. However, the right and left-down of Fig. 11 framed in red indicate that the appearance of images rendered by NeRF-W has green bias and red bias with respect to ground truth. The experimental results demonstrate that NeRF-W suffers from inconsistent appearance when the camera moves, while Ha-NeRF is consistently consistent with ground truth.

Occlusion Handling. We eliminate the transient phenomena using an image-dependent 2D visibility map, while

	Dataset	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF	Original	32.23	.9566	.0167
NeRF	Color Perturbation	20.89	.9006	.1195
NeRF-W		28.85	.9370	.0337
Ha-NeRF		29.60	.9375	.0294
NeRF	Occlusion	19.04	.8627	.1394
NeRF-W		28.12	.9336	.0434
Ha-NeRF		30.93	.9494	.0233
NeRF	Color Perturbation & Occlusion	17.17	.8385	.1959
NeRF-W		27.10	.9211	.0525
Ha-NeRF		28.41	.9349	.0360

Table 2. Quantitative results from experiments on the synthetic dataset in different cases. Ha-NeRF outperforms the others on all evaluation metrics.

Figure 11. Results of the experiment on synthetic Lego dataset with ground truth and manually introduce color perturbation for training. The left-top of the figure shows that both NeRF-W and Ha-NeRF can render images with the correct appearance. However, the right and left-down of the figure framed in red shows that the appearance of images rendered by NeRF-W has green bias and red bias with respect to ground truth. It demonstrates that NeRF-W suffers from inconsistent appearance when the camera moves, while Ha-NeRF is consistently consistent with ground truth. **Better viewed on screen with zooming in.**

NeRF-W uses a 3D transient field to reconstruct the transient objects. As illustrated in Fig. 12, our occlusion handling method generates an accurate segmentation between static scene and transient objects, which allows us to render occlusion-free images. However, NeRF-W inaccurately decomposes the scene (*e.g.*, board, people, and fence still leave on the renderings of NeRF-W) and further entangles the variable appearance and transient occlusion in the 3D transient field (*e.g.*, results in the transient volume to remember the white cloud of "Brandenburg Gate").

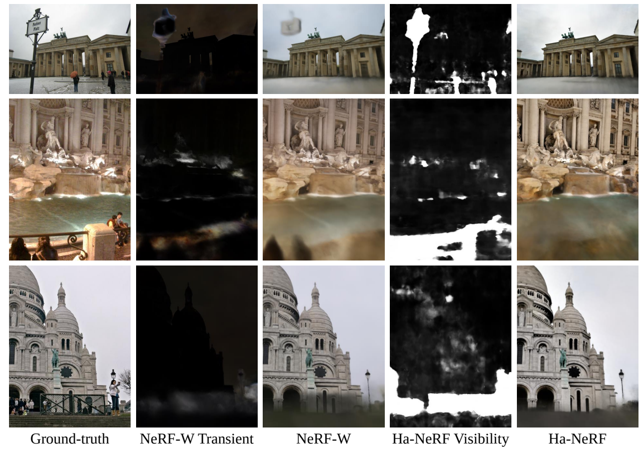


Figure 12. Anti-occlusion renderings of Ha-NeRF and NeRF-W. NeRF-W Transient is the renderings of the 3D transient field of NeRF-W, which tries to reconstruct the transient objects only observed in an individual image. We denote Ha-NeRF Visibility as our 2D visibility map that learned to disentangle static and transient phenomena of the images, indicating the visibility of rays originated from the static scene.

Limitations. Without exception, the proposed Ha-NeRF suffers from the noisy camera extrinsic parameters, similar to most NeRF based approaches. Additionally, the quality of synthesized images degrades while the input images are either motion-blurred or defocused. Specific techniques have to be developed to handle these issues.

6. Conclusion

NeRF has grown in prominence and has been utilized in various applications, including the recovery of NeRF from tourism images. While NeRF-W works effectively with a train-data optimized appearance embedding, it is hard to hallucinate novel views consistently at an unlearnt appearance. To overcome this challenging problem, we present the Ha-NeRF, which can hallucinate the realistic radiance field under variable appearances and complex occlusions. Specifically, we propose an appearance hallucination module to handle time-varying appearances and transfer them to novel views. Furthermore, we employ an anti-occlusion module to learn an image-dependent 2D visibility mask capable of accurately separating static subjects. Experimental results using synthetic data and tourism photo collections demonstrate that our method can render free-occlusion views and hallucination of the appearance. Codes and models will be made publicly available to the research community to facilitate reproducible research.

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *ECCV*, pages 696–712. Springer, 2020. 2

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

[3] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM TOG*, 39(4):86–1, 2020. 2

[4] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *SIGGRAPH*, pages 425–432, 2001. 2

[5] Rodrigo Ortiz Cayon, Abdelaziz Djelouah, and George Drettakis. A bayesian approach for selective image-based rendering using superpixels. In *3DV*, pages 469–477, 2015. 2

[6] 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

[7] Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. Extreme view synthesis. In *IEEE ICCV*, pages 7781–7790, 2019. 2

[8] Forrester Cole, Kyle Genova, Avneesh Sud, Daniel Vlasic, and Zhoutong Zhang. Differentiable surface rendering via non-differentiable sampling. In *=IEEE CVPR*, pages 6088–6097, 2021. 2

[9] Peng Dai, Yinda Zhang, Zhuwen Li, Shuaicheng Liu, and Bing Zeng. Neural point cloud rendering via multi-plane projection. In *IEEE CVPR*, pages 7830–7839, 2020. 2

[10] Abe Davis, Marc Levoy, and Fredo Durand. Unstructured light fields. In *Computer Graphics Forum*, volume 31, pages 305–314, 2012. 2

[11] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *SIGGRAPH*, pages 11–20, 1996. 2

[12] 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 *IEEE CVPR*, pages 2367–2376, 2019. 2

[13] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. *arXiv preprint arXiv:2103.10380*, 2021. 2

[14] Rahul Garg, Hao Du, Steven M Seitz, and Noah Snavely. The dimensionality of scene appearance. In *IEEE ICCV*, pages 1917–1924. IEEE, 2009. 2

[15] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *SIGGRAPH*, pages 43–54, 1996. 2

[16] Daniel Cabrini Hauagge, Scott Wehrwein, Paul Upchurch, Kavita Bala, and Noah Snavely. Reasoning about photo collections using models of outdoor illumination. In *BMVC*, 2014. 2

[17] Peter Hedman, Suhib Alsisan, Richard Szeliski, and Johannes Kopf. Casual 3D photography. *ACM TOG*, 36(6):1–15, 2017. 2

[18] Peter Hedman and Johannes Kopf. Instant 3D photography. *ACM TOG*, 37(4):1–12, 2018. 2

[19] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM TOG*, 37(6):1–15, 2018. 2

[20] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018. 4

[21] Kichang Kim, Akihiko Torii, and Masatoshi Okutomi. Multi-view inverse rendering under arbitrary illumination and albedo. In *ECCV*, pages 750–767. Springer, 2016. 2

[22] Pierre-Yves Laffont, Adrien Bousseau, Sylvain Paris, Fredo Durand, and George Drettakis. Coherent intrinsic images from photo collections. *ACM TOG*, 31(6), 2012. 2

[23] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, pages 35–51, 2018. 4

[24] Marc Levoy and Pat Hanrahan. Light field rendering. In *SIGGRAPH*, pages 31–42, 1996. 2

[25] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE CVPR*, pages 6498–6508, 2021. 1, 2

[26] Miaomiao Liu, Xuming He, and Mathieu Salzmann. Geometry-aware deep network for single-image novel view synthesis. In *IEEE CVPR*, pages 4616–4624, 2018. 2

[27] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 2

[28] 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 *IEEE CVPR*, pages 7210–7219, 2021. 1, 2, 3, 4, 5, 8, 9

[29] Leonard McMillan and Gary Bishop. Plenoptic modeling: An image-based rendering system. In *SIGGRAPH*, pages 39–46, 1995. 2

[30] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *IEEE CVPR*, pages 6878–6887, 2019. 2, 3

[31] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *IEEE CVPR*, pages 4743–4752, 2019. 2

[32] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM TOG*, 38(4):1–14, 2019. 2

[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 *ECCV*, pages 405–421. Springer, 2020. 1, 2, 3, 5

[34] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *IEEE CVPR*, pages 11453–11464, 2021. 2

[35] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM TOG*, 38(6):1–15, 2019. 2

[36] Ryan S Overbeck, Daniel Erickson, Daniel Evangelakos, Matt Pharr, and Paul Debevec. A system for acquiring, processing, and rendering panoramic light field stills for virtual reality. *ACM TOG*, 37(6):1–15, 2018. 2

[37] Jaesik Park, Yu-Wing Tai, Sudipta N Sinha, and In So Kweon. Efficient and robust color consistency for community photo collections. In *IEEE CVPR*, pages 430–438, 2016. 3

[38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE CVPR*, pages 165–174, 2019. 2

[39] True Price, Johannes L Schönberger, Zhen Wei, Marc Pollefeys, and Jan-Michael Frahm. Augmenting crowd-sourced 3d reconstructions using semantic detections. In *IEEE CVPR*, pages 1926–1935, 2018. 2

[40] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE CVPR*, pages 10318–10327, 2021. 1, 2

[41] Filip Radenovic, Johannes L Schonberger, Dinghuang Ji, Jan-Michael Frahm, Ondrej Chum, and Jiri Matas. From dusk till dawn: Modeling in the dark. In *IEEE CVPR*, pages 5488–5496, 2016. 2

[42] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. *arXiv preprint arXiv:2103.13744*, 2021. 2

[43] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *ECCV*, pages 1–17, 2020. 2

[44] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *IEEE CVPR*, pages 12216–12225, 2021. 2

[45] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *SIGGRAPH*, pages 231–242, 1998. 2

[46] Qi Shan, Riley Adams, Brian Curless, Yasutaka Furukawa, and Steven M Seitz. The visual turing test for scene reconstruction. In *3DV*, pages 25–32. IEEE, 2013. 2

[47] Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. In *Visual Communications and Image Processing*, volume 4067, pages 2–13, 2000. 2

[48] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NIPS*, pages 1121–1132, 2019. 2

[49] 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 *IEEE CVPR*, pages 175–184, 2019. 2

[50] Kalyan Sunkavalli, Wojciech Matusik, Hanspeter Pfister, and Szymon Rusinkiewicz. Factored time-lapse video. In *SIGGRAPH*, pages 101–es. 2007. 2

[51] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. *Computer Graphics Forum*, 39(2):701–727, 2020. 2

[52] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM TOG (TOG)*, 38(4):1–12, 2019. 2

[53] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *IEEE CVPR*, pages 551–560, 2020. 2

[54] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *IEEE CVPR*, pages 7467–7477, 2020. 2

[55] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. *arXiv preprint arXiv:2103.14024*, 2021. 2

[56] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE CVPR*, pages 4578–4587, 2021. 1, 2

[57] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM TOG*, 40(4):1–18, 2021. 2

[58] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM TOG*, 37(4):1–12, 2018. 2

[59] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, pages 286–301, 2016. 2

[60] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in neural information processing systems*, pages 465–476, 2017. 4