Cross-Ray Neural Radiance Fields for Novel-view Synthesis from Unconstrained Image Collections

Yifan Yang[1,2*] Shuhai Zhang[1,2] Zixiong Huang[1] Yubing Zhang[3] Mingkui Tan[1,2,4†]
[1]South China University of Technology [2]Pazhou Lab [3]Guangzhou Shiyuan Electronics Co., Ltd
[4]Key Laboratory of Big Data and Intelligent Robot, Ministry of Education
{seyoungyif, mszhangshuhai, sesmilhzx}@mail.scut.edu.cn, zhangyubing@cvte.com
mingkuitan@scut.edu.cn

Abstract

Neural Radiance Fields (NeRF) is a revolutionary approach for rendering scenes by sampling a single ray per pixel and it has demonstrated impressive capabilities in novel-view synthesis from static scene images. However, in practice, we usually need to recover NeRF from unconstrained image collections, which poses two challenges: 1) the images often have dynamic changes in appearance because of different capturing time and camera settings; 2) the images may contain transient objects such as humans and cars, leading to occlusion and ghosting artifacts. Conventional approaches seek to address these challenges by locally utilizing a ***single ray*** *to synthesize a color of a pixel. In contrast, humans typically perceive appearance and objects by globally utilizing information across multiple pixels. To mimic the perception process of humans, in this paper, we propose Cross-Ray NeRF (CR-NeRF) that leverages interactive information across* ***multiple rays*** *to synthesize occlusion-free novel views with the same appearances as the images. Specifically, to model varying appearances, we first propose to represent multiple rays with a novel cross-ray feature and then recover the appearance by fusing global statistics, i.e., feature covariance of the rays and the image appearance. Moreover, to avoid occlusion introduced by transient objects, we propose a transient objects handler and introduce a grid sampling strategy for masking out the transient objects. We theoretically find that leveraging correlation across multiple rays promotes capturing more global information. Moreover, extensive experimental results on large real-world datasets verify the effectiveness of CR-NeRF. The code and data can be found at* `https://github.com/YifYang993/CR-NeRF-PyTorch.git`.

1. Introduction

Novel-view synthesis is a long-standing problem in computer vision that has paved the way for numerous applications such as virtual reality and digital humans [13, 45]. More recently, the emergence of Neural Radiance Fields (NeRF) has driven the field forward, as it has shown significant performance in reconstructing 3D geometry [50] and recovering the appearance [3, 35, 1] from multi-view image sets. However, NeRF assumes that the images do not have variable appearances and moving objects [31] (called the *static scene assumption c.f.* Sec. 3), which leads to significant performance degradation on large-scale Internet image collections. To expand the scope of NeRF, we aim to exploit the collections and provide a 3D immersive experience through which we can visit international landmarks such as the Brandenburg Gate, and the Trevi Fountain from different viewpoints and times of one day.

To achieve this, we address the problem of recovering an appearance-controllable and anti-occlusion NeRF from unconstrained image collections. In other words, by reconstructing the NeRF representation, we control the appearance of the scene based on photos with various photometric conditions, while eliminating occlusions caused by the images. Although providing a sense of immersion, reconstructing NeRF with these images faces the following two challenges. 1) *Varying appearances*: Imaging two tourists who take photos in the same viewpoint but under various conditions, *e.g.,* different capturing times, diverse weather (*e.g.,* sunny, rainy, and foggy), and different camera settings (*e.g.,* aperture, shutter, and ISO). This varying condition causes that although multiple photographs are taken of the same scene, they look dramatically different. 2) *Transient occlusion*: Even with a constant appearance, transient objects such as cars and Pedestrians may obscure the scene. Since these objects are usually captured by only one photographer, it is usually impractical to reconstruct these objects in high quality. The above challenges conflict with the static-scene

[2]Corresponding author.
[1]This work was done when Yifan Yang was a research intern at Guangzhou Shiyuan Electronics Co., Ltd.

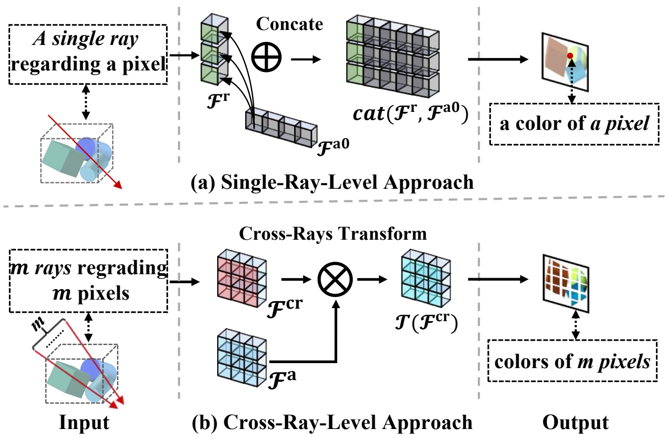


Figure 1. Illustration of single-ray-level and cross-ray-level approaches. The conventional one (a) generates each pixel regarding a *single ray* independently. In contrast, our proposed CR-NeRF (b) considers information of *multiple rays* and synthesizes a patch simultaneously. $\mathcal{F}^{a}$ and $\mathcal{F}^{a_0}$ are conditioned features.

assumption of NeRF and result in inaccurate reconstruction that leads to over-smoothing and ghosting artifacts [31].

Recently, several attempts (NeRF-W [31];Ha-NeRF [6]) have been proposed to address the aforementioned challenges. From Fig. 1(a), NeRF-W and Ha-NeRF leverage a single-ray manner, wherein *a single camera ray* (*i.e.,* a beam of light extending from a camera through a pixel on an image plane into a 3D scene) serves as input. This manner then involves considering appearance and occlusion factors and subsequently synthesizing *each color of pixel* of a novel view independently. One potential issue of this manner is its reliance on local information (*e.g.,* information of a single image pixel) of every single ray for recognizing appearance and transient objects. In contrast, humans tend to utilize global information (*e.g.,* information across multiple image pixels), which provides a more comprehensive understanding of an object to observe its appearance and handle occlusion. Motivated by this, we propose to tackle varying appearance and transient objects with a cross-ray paradigm (see Fig. 1(b)), wherein we utilize global information from multiple rays to recover the appearance and handle transient objects. Subsequently, we synthesize a region of a novel view simultaneously. Based on the cross-ray paradigm, we propose a Cross-Ray Neural Radiance Fields (CR-NeRF), which comprises two components: 1) To model variable appearances, we propose to represent information of multiple rays with a novel cross-ray feature. We then fuse the cross-ray feature and an appearance embedding via a cross-ray transformation network using global statistics, *e.g.,* feature covariance of the cross-ray. The fused feature is fed to a decoder to obtain colors of several pixels simultaneously. 2) To handle transient objects, we propose a unique perspective of handling transient objects as a segmentation problem, through which we detect transient objects by considering global information of an image region. From this perspective, we segment the unconstrained images for a visibility map of the objects. To avoid computation overhead, we introduce a grid sample strategy that samples the segmented maps to pair with the input rays. We theoretically analyze that leveraging correlation across multiple rays promotes capturing more global information.

We summarize our contributions in three folds:

- A new cross-ray paradigm for novel-view synthesis from unconstrained photo collections: We find that existing methods fall short of producing satisfactory visual outcomes from unconstrained photo collections via a single-ray-level paradigm, primarily due to the neglect of the potential cooperative interaction among multiple rays. To address this, we propose a novel cross-ray paradigm, which exploits the global information across multiple rays.
- An interactive and global scheme for addressing varying appearances: Unlike existing methods that process each ray independently, we represent multiple rays by introducing a cross-ray feature, which facilitates the interaction among rays through feature covariance. This enables us to inject a global informative appearance representation into the scene, resulting in more realistic and efficient appearance modeling. Our theoretical analysis demonstrates the necessity of considering multiple rays for appearance modeling.
- A novel segmentation technique for processing transient objects: We reformulate the transient object problem as a segmentation problem. We use global information of an unconstrained image to segment a visibility map. Moreover, we apply grid sampling to pair the map with multiple rays. Empirical results show that CR-NeRF eliminates the transient objects in reconstructed images.

2. Related Works

Neural rendering. Neural rendering applies deep learning with computer graphic technologies to render images and reconstruct 3D scenes. Recent advances seek to apply learning-based technology to generate representations such as signed distance field [34, 27, 59, 19], point clouds [10, 17, 26], voxels [37, 15, 57] and occupancy fields [58, 32, 38], which are then applied for rendering novel views. With the remarkable performance, NeRF [33] has attracted attention from the neural rendering community. More recently, NeRF has been extended to represent a time-series of scenes [25, 39, 22], handle high-resolution settings [18, 52], address relighting [43], and reconstruct large-scale environments [46, 47, 48]. Notably, one limitation of NeRF is that it assumes the scene is static, which faces challenges of varying appearance and presence of transient objects in unconstrained image collections. To alleviate this, NeRF-W [31] and Ha-NeRF [6] focus on addressing the challenges by processing each ray of a scene indepen-

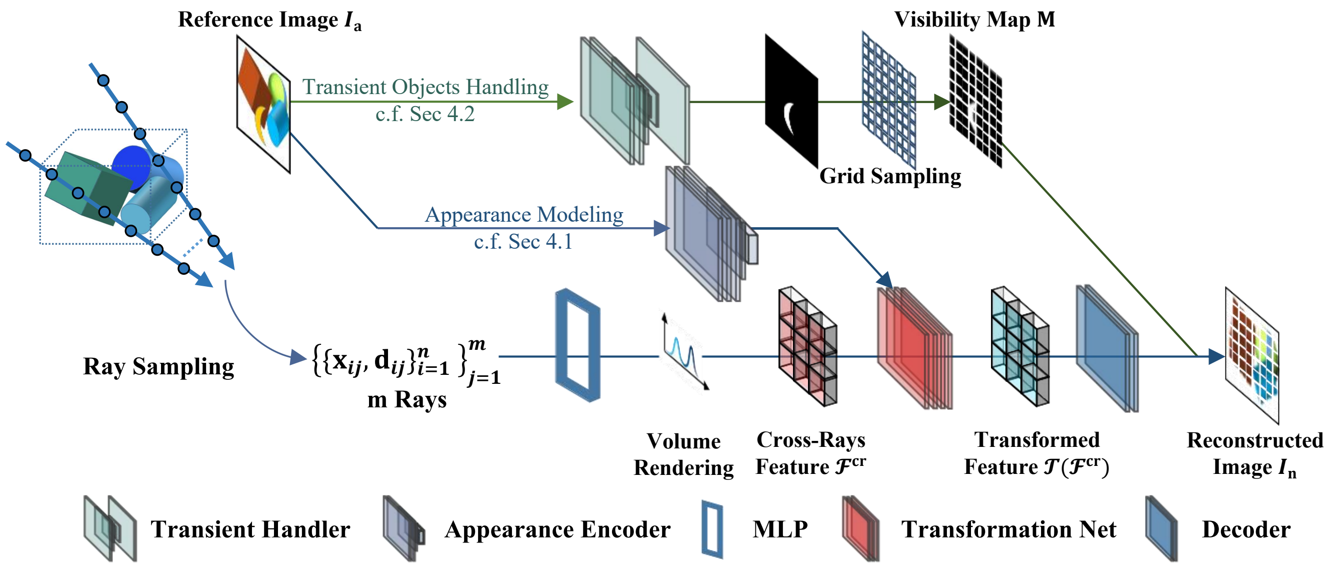


Figure 2. An overview of Cross-Ray Neural Radiance Field (CR-NeRF). Given position $\mathbf{x}$ and direction $\mathbf{d}$ of multiple rays, We first generate a cross-ray feature $\mathcal{F}^{\text{cr}}$ that accumulates multi-view information in a scene. To incorporate the appearance information of the reference image $\mathcal{I}_a$, the appearance encoder is used to learn the appearance features $\mathcal{F}_a$, the transform net to fuse $\mathcal{F}_a$ with $\mathcal{F}^{\text{cr}}$ and the decoder to synthesize the colors of multiple pixels in the reconstructed image $\mathcal{I}_{\text{n}}$ simultaneously. To eliminate transient objects in $\mathcal{I}_a$, our transient handler generates a visibility map, for which we introduce a grid sampling strategy to match the map with the rays during training.

dently. Differently, we propose to accumulate information on multiple rays for modeling appearance and eliminating transient objects.

Novel-view synthesis. Synthesizing views from a novel viewpoint has long been a fundamental problem in computer vision and computer graphics. Traditionally, novel views can be synthesized through 4D light field strategy [21, 53, 5]. However, the strategy requires a dense camera array for capturing data, which is usually impractical. Since collecting a sparse set of images are efficient, view synthesis research takes advantage of geometry structure [2, 7] to aid in constructing novel views with limited input. With the flourish of deep learning, deep neural networks have been leveraged to estimate the scene geometry (*e.g.,* point clouds [54, 40], depth map [44, 28], multiple-layer image [11, 55]). Although leveraging the geometry enhances the quality of novel views, the estimation is usually without ground truth supervision and usually is not accurate enough. To circumvent the difficulty of estimating precise geometry, we propose to utilize an implicit function, *i.e.,* neural radiance fields (NeRF) [33] for novel-view synthesis.

3. Preliminaries

Neural Radiance Fields (NeRF) [33] implicitly represents a static 3D scene with *multilayer perceptron* (MLP) and then produces a novel view via *volume rendering* (VR) [9]. NeRF generates a pixel color of a novel view from a camera ray independently. In this sense, we can describe the rendering process w.r.t. a single camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ which is cast from a camera center $\mathbf{o}$ in the direction $\mathbf{d}$ and passes through a pixel on an image plane w.r.t. the novel view. We sample n ray points $\{\mathbf{r}(t_i)\}_{i=1}^{n}$ along $\mathbf{r}$ between a given near plane t_n and a far plane t_f. For each ray point $\mathbf{r}(t_i)$, we query the MLP at a 3D position $\mathbf{x}_i = (x, y, z)$ and a viewing orientation $\mathbf{d}_i = (d_x, d_y, d_z)$ to obtain a color $\mathbf{k}_i = (r, g, b)$ and a density σ_i via equations: $\mathbf{x}_i \to \{\mathcal{F}_i^r, \sigma_i\}$, $\{\mathcal{F}_i^r\, \mathbf{d}_i\} \to \mathbf{k}_i$, where $\mathcal{F}_i^r$ denotes a ray-point-level feature regarding the ray point i . To learn high-frequency information, position encoding [33] is employed to $\mathbf{x}_i$ and $\mathbf{d}_i$. Typically, $\mathbf{o}$ and $\mathbf{d}$ are estimated by structure from motion approaches [41, 49] from multi-view images regarding the 3D scene.

To approximate the color $\hat{\mathbf{c}}(\mathbf{r})$ of the pixel of a reference image, NeRF accumulates n ray points $\{\mathbf{r}(t_i)\}_{i=1}^{n}$ along the ray $\mathbf{r}$ into the $\hat{\mathbf{c}}(\mathbf{r})$ via VR [9]:

$$\begin{aligned} \hat{\mathbf{c}}(\mathbf{r}) = \sum_{i=1}^{n} \varphi_i \alpha_i \mathbf{k}_i, \varphi_i = \exp(-\sum_{l=1}^{i-1} \sigma_l \delta_l), \\ \alpha_i = 1 - \exp\left(-\sigma_i \delta_i\right). \end{aligned} \tag{1}$$

Here, α_i is the probability of the ray that terminates at $\mathbf{r}(t_i)$; φ_i is the accumulated transmittance from the near plane t_n to $\mathbf{r}(t_i)$; $\delta_l = t_{l+1} - t_l$ is distance between two adjacent ray points. The MLP is optimized via minimizing the loss: $\mathcal{L} = ||\hat{\mathbf{c}}(\mathbf{r}) - \mathbf{c}(\mathbf{r})||_2^2$, where $\mathbf{c}(\mathbf{r})$ denotes the ground truth color of a pixel w.r.t. the ray $\mathbf{r}$.

Limitations of NeRF on novel-view synthesis from unconstrained collections. Given an unconstrained collection of a scene, we seek to reconstruct the scene whose appearance can be modified according to a new image, while removing transient objects. Since NeRF assumes the lighting in the scene is constant over time and there are no moving objects or changes in lighting during the time that the input images are captured (called *static scene assumption* [31]), NeRF is limited to effectively modeling the geometry and appearance of static scenes only. To address the limitation of NeRF, recent advances [31, 6] synthesize novel views on single-ray level (see Fig. 1 (a)) following equation:

$$\{\mathbf{x}_i, \mathbf{d}_i, \mathcal{F}^{a_0}\}_{i=1}^{n} \rightarrow \hat{\mathbf{c}}_{\mathrm{n}}, \tag{2}$$

where $\mathcal{F}^{a_0}$ is image-level conditional embedding of $\mathcal{I}_a$. From Eqn. 2, existing methods [31, 6] generate the color $\hat{\mathbf{c}}_{\mathrm{n}}$ of each pixel by processing the corresponding single ray independently. This manner ignores global information among multiple rays, leading to inaccurate appearance modeling (see our empirical studies in Fig. 3 and Fig. 4).

4. Cross-Ray Neural Radiance Fields

Given unconstrained photo collections of a scene, we seek to reconstruct the scene whose appearance can be modified based on a new image, while removing transient objects. This task is challenging due to the existence of variable appearances and transient occlusions in the photo collections. To address this, intuited by that a human usually detects appearance and transient objects by considering global information (*e.g.,* information across several image pixels) rather than local information (*e.g.,* information of a single image pixel), we propose a Cross-Ray Neural Radiance Fields (CR-NeRF) that exploits global information across multiple camera rays, which correspond to several pixels of an image, to address both challenges.

As shown in Fig. 2 and Alg. 1, CR-NeRF consists of two components: 1) *Cross ray appearance modeling* (*c.f.* Sec. 4.1). To model varying appearances, we first sample a grid of rays using a grid sampling strategy [42]. Next, we represent the rays with a novel cross-ray feature $\mathcal{F}^{\mathrm{cr}}$. We then inject an appearance embedding $\mathcal{F}^{a}$ into $\mathcal{F}^{\mathrm{cr}}$ via a learned transformation network. The fused feature is fed to a decoder for obtaining colors of multiple pixels simultaneously. We theoretically analyze the necessity of considering multiple rays and thus design an appearance loss $\mathcal{L}_a$ for cross-ray appearance modeling. 2) *Cross-ray transient objects handling* (*c.f.* Sec. 4.2). To handle transient objects, we deploy a segmentation network for generating a visibility map regarding transient objects. To pair the map with the rays, we also apply the grid sampling strategy on the maps. We devise an occlusion loss $\mathcal{L}_t$ for transient handling.

The overall optimization of our proposed CR-NeRF minimizes the following objective function:

$$\mathcal{L}_{overall} = \mathcal{L}_a + \lambda \mathcal{L}_t, \tag{3}$$

where λ is a hyper-parameter for balancing the appearance loss $\mathcal{L}_a$ (see Eqn. 8) and the occlusion loss $\mathcal{L}_t$ (see Eqn. 9).

4.1. Cross-Ray Appearance Modeling

To adapt CR-NeRF to variable appearance through a global perspective, we modify the scene by leveraging multiple rays and the appearance of the unconstrained images.

Representing scene information with multiple rays. To model appearance from a multi-view observation, we first represent scene information using multiple rays. To this end, we propose a novel *cross-ray feature* $\mathcal{F}^{\mathrm{cr}}$ with equations:

$$\begin{aligned} \{\{\mathcal{F}_{ij}^{r}, \sigma_{ij}\}_{i=1}^{n}\}_{j=1}^{m} &= \{\mathrm{MLP}_{\theta_1}(\{\mathbf{x}_{ij}, \mathbf{d}_{ij}\}_{i=1}^{n})\}_{j=1}^{m}, \\ \mathcal{F}^{\mathrm{cr}} &= \{\mathrm{VR}(\{\mathcal{F}_{ij}^{r}, \sigma_{ij}, \delta_{ij}\}_{i=1}^{n})\}_{j=1}^{m}, \end{aligned} \tag{4}$$

Besides, we obtain an appearance feature $\mathcal{F}^{a}$ of an appearance image $\mathcal{I}_a$ by $\mathcal{F}^{a} = E_{\theta_2}(\mathcal{I}_a)$. With $\mathcal{F}^{\mathrm{cr}}$ and $\mathcal{F}^{a}$, it is critical to find an effective fusion manner to inject image appearance into the scene representation.

Injecting appearance into scene representation. The key of our cross-ray appearance modeling is to exploit the potential cooperative relationship among the cross-ray features $\mathcal{F}^{\mathrm{cr}}$ to facilitate appearance modeling from the given appearance image $\mathcal{I}_a$ to the scene representation. In other words, we seek a transformation operation that can transfer the style from a reference image and also retain the essential content during training. To this end, we learn a transformation $\mathcal{T}$ to align the transferred cross-ray features $\mathcal{T}(\mathcal{F}^{\mathrm{cr}})$ and the appearance feature $\mathcal{F}^{a}$ with an auxiliary identity term, which is formulated as below,

$$\min_{\mathcal{T}} \mathbb{E}_{\mathcal{F}^{\mathrm{cr}}, \mathcal{F}^{a}} \|\mathcal{T}(\mathcal{F}^{\mathrm{cr}}) - \mathcal{F}^{a}\|_2^2 + \beta \|\mathbf{P}\mathcal{T}(\mathcal{F}^{\mathrm{cr}}) - \mathcal{F}^{\mathrm{cr}}\|_2^2, \tag{5}$$

where β is a trade-off parameter and $\mathbf{P}$ is a constant matrix for matching the transformed feature $\mathcal{T}(\mathcal{F}^{\mathrm{cr}})$ and $\mathcal{F}^{\mathrm{cr}}$. Next, we theoretically analyze the necessity of considering multiple rays to solve Problem (5) for appearance modeling.

Necessity of considering multiple rays for appearance modeling. We consider a Gaussian case that can provide some insights to devise an effective approach to inject the appearance into the scene representation. To this end, we assume the two features $\mathcal{F}^{a}$ and $\mathcal{F}^{\mathrm{cr}}$ are following two Gaussian distributions and $\mathcal{T}$ is a linear transformation that rigorously matches two distributions. We provide a closed-form solution to Problem (5) under this assumption as follows.

Proposition 1. *Given an invertible constant matrix* $\mathbf{P} \in \mathbb{R}^{C \times C}$*, assuming that* $\mathcal{F}^{a} \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$*,* $\mathcal{F}^{\mathrm{cr}} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{cr}}, \boldsymbol{\Sigma}_{\mathrm{cr}})$ *and* $\mathcal{T}(\mathcal{F}^{\mathrm{cr}}) \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$*, where* $\mathcal{T}(\mathcal{F}^{\mathrm{cr}}) = \mathbf{T}(\mathcal{F}^{\mathrm{cr}} - \boldsymbol{\mu}_{\mathrm{cr}}) + \boldsymbol{\mu}_a$ *and* $\mathbf{T} \in \mathbb{R}^{C \times C}$ *is a transformation matrix, the optimal* $\mathbf{T}$ *to Problem 5 is:*

$$\mathbf{T} = \mathbf{P}^{-1} \boldsymbol{\Sigma}_{\mathrm{cr}}^{-1/2} \left(\boldsymbol{\Sigma}_{\mathrm{cr}}^{1/2} \mathbf{P}^{\top} \boldsymbol{\Sigma}_a \mathbf{P} \boldsymbol{\Sigma}_{\mathrm{cr}}^{1/2} \right)^{1/2} \boldsymbol{\Sigma}_{\mathrm{cr}}^{-1/2}. \tag{6}$$

Algorithm 1: The training pipeline of CR-NeRF.

Input: m rays, a reference image $\mathcal{I}_a$, a multilayer perceptron MLP_{θ_1}, an appearance encoder E_{θ_2}, a transformation net $\mathcal{T}_{\theta_3}$, a decoder D_{θ_4}, a content encoder E_{θ_5} and a segmentation net $\mathcal{S}_{\theta_\triangle}$.

Output: The estimated colors of m pixels of a novel view.

while not converged **do**

- Generate cross-ray features $\mathcal{F}^{\mathrm{cr}}$ and appearance feature $\mathcal{F}^{a}$ with E_{θ_2} and MLP_{θ_1} by Eqn. (4).
- Obtain the loss $\mathcal{L}_a$ for modeling appearance with E_{θ_2}, $\mathcal{T}_{\theta_3}$, D_{θ_4} and E_{θ_5} by Eqn. (8). ▷ *c.f.* Sec. 4.1
- Obtain the visibility map $\mathbf{M}$ for masking out transient objects with and $\mathcal{S}_{\theta_\triangle}$ by Eqn. (9).
- Obtain the loss $\mathcal{L}_t$ for handling transient with $\mathcal{T}_{\theta_3}$, D_{θ_4} and $\mathcal{S}_{\theta_\triangle}$ by Eqn. (10). ▷ *c.f.* Sec. 4.2
- Obtain the overall loss of $\mathcal{L}_{overall} = \mathcal{L}_a + \lambda\mathcal{L}_t$.
- Update the parameters $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_\triangle\}$ by descending the gradient: $\nabla_\Theta \mathcal{L}_{overall}$

end

Proposition 1 suggests the transformation matrix $\mathbf{T}$ is determined by the covariance of $\mathcal{F}^{\mathrm{cr}}$ and $\mathcal{F}^{a}$ given $\mathbf{P}$, which is consistent with the conclusion in [23, 29]. Inspired by this, we can construct a neural network to learn the appearance transformation $\mathcal{T}$ by feeding the covariances of $\mathcal{F}^{a}$ and $\mathcal{F}^{\mathrm{cr}}$. Specifically, we adopt an effective transformation network following Li *et al.* [23] which is defined as:

$$
\begin{gathered}
\mathcal{T}(\mathcal{F}^{\mathrm{cr}}) = \mathbf{T}\hat{\mathcal{F}}^{\mathrm{cr}}, \\
\text{where } \mathbf{T} = Cov(\bar{\mathcal{F}}^{\mathrm{cr}})Cov(\bar{\mathcal{F}}^{a})), \\
\hat{\mathcal{F}}^{\mathrm{cr}} = \phi_1(\mathcal{F}^{\mathrm{cr}}), \bar{\mathcal{F}}^{\mathrm{cr}} = \phi_2(\mathcal{F}^{\mathrm{cr}}), \bar{\mathcal{F}}^{a} = \phi_3(\mathcal{F}^{a}).
\end{gathered}
\tag{7}
$$

Here, ϕ_1, ϕ_2, and ϕ_3 are non-linear mappings parameterized by convolutional neural networks (CNN) that can express richer embedding to prepare for appearance modeling. Intuitively, we consider multiple rays when modeling appearance to employ multi-view information. The correlation between the feature maps of these different views, which can be given by the covariance, is able to capture more global texture information for a given appearance image [12, 24, 8], thus facilitating better appearance modeling for a scene.

Loss function $\mathcal{L}_a$ for varying appearance modeling: To generate a novel-view image with a satisfactory appearance from the transformed feature $\mathcal{T}_{\theta_3}(\mathcal{F}^{\mathrm{cr}})$), we need to enforce a decoder D_{θ_4} into the training process of appearance modeling. Inspired by the formulation in Problem (5), we provide the loss function for appearance modeling as:

$$
\begin{aligned}
\mathcal{L}_a = &||E_{\theta_2}[D_{\theta_4}(\mathcal{T}_{\theta_3}(\mathcal{F}^{\mathrm{cr}}))] - \mathcal{F}^{\mathrm{a}}||_2^2 \\
&+ \beta||E_{\theta_5}[D_{\theta_4}(\mathcal{T}_{\theta_3}(\mathcal{F}^{\mathrm{cr}}))] - E_{\theta_5}[D_{\theta_4}(\mathcal{F}^{\mathrm{cr}})]||_2^2,
\end{aligned}
\tag{8}
$$

where $\mathcal{F}^{\mathrm{cr}}$ is obtained with an MLP_{θ_1} by Eqn. 4. Here, we use a tailored encoder E_{θ_5} to model the transformed feature $\mathbf{P}\mathcal{T}(\mathcal{F}^{cr})$ so that the content of the transformed image closely matches its original counterpart. In this way, we can synthesize a novel-view image by $\mathcal{I}_{\mathrm{n}} = D_{\theta_4}(\mathcal{T}_{\theta_3}(\mathcal{F}^{\mathrm{cr}}))$.

4.2. Transient Objects Handling

To deal with transient objects caused by unconstrained photo collections for novel-view synthesis, we propose a new perspective, *i.e.,* obtaining the visibility map of transient objects by segmenting the reference image $\mathcal{I}_a$. With the receptive fields of a deep segmentation network [30], the interactions of different pixels and rays are facilitated, thus introducing more global information.

To accurately detect transient objects, we start by exploring a pre-trained Mask R-CNN model [14] and a pre-trained DeepLabV3 model [4] that are capable of effectively segmenting common objects such as tourists and cars, *etc.* We observe that although the models properly segment the common objects, the reconstruction error is amplified (see empirical studies in Sect. 5.2). The possible reason is that the target transient objects are not limited to common objects, more objects (*e.g.,* shadows of tourists in Fig. 6) should also be taken into consideration.

In this sense, we choose a learning-based manner to select which objects to segment and therefore deploy a light-weight segmentation network $\mathcal{S}_{\theta_\triangle}$ following [56]. Since we cannot sample all rays that interact with $\mathcal{I}_a$ due to limited GPU memory in the training phase, naively processing all rays of transient objects (*i.e.,* $\mathcal{S}_{\theta_\triangle}(\mathcal{I}_a)$) is therefore not applicable. Hence, we apply a grid sampling strategy (GS) [42] which samples $\mathcal{S}_{\theta_\triangle}(\mathcal{I}_a)$ to pair with m rays (see Fig. 2). The whole process for estimating $\mathbf{M}$ is:

$$
\mathbf{M} = \mathrm{GS}(\mathcal{S}_{\theta_\triangle}(\mathcal{I}_a)),
\tag{9}
$$

where $\mathcal{S}_{\theta_\triangle} : \mathbb{R}^{3\times h_{cr1}h_{cr2}} \to \mathbb{R}^{3\times h_{cr1}h_{cr2}}$, h_{cr1} and h_{cr2} are heights and width of $\mathcal{I}_a$. Here, $\mathcal{S}_{\theta_\triangle}$ learns a visibility map $\mathbf{M}$ without the supervision of ground truth segmentation masks. During training, we set m to be smaller than $h_{cr1}h_{cr2}$ for saving computational overhead.

Loss function $\mathcal{L}_t$ for eliminating transient objects: The loss function for handling transient objects is:

$$
\mathcal{L}_t = ||(1-\mathbf{M}) \odot (\mathcal{I}_{\mathrm{n}} - \mathcal{I}_a)^2||_1 + \lambda_0\|\mathbf{M}\|^2,
\tag{10}
$$

where $\odot$ denotes element-wise multiplication. The loss $\mathcal{L}_t$ aims to mask out transient objects via $\mathbf{M}$. To prevent our transient network from masking everything, we follow Ha-NeRF to add $\lambda_0\|\mathbf{M}\|^2$ as a regularization term.

4.3. Difference of CR-NeRF with Existing Methods

To model varying appearances, Ha-NeRF and NeRF-W process each single ray independently by Eqn. 2. To handle transient objects, NeRF-W implements an additional MLP for rendering transient objects by Eqn. 2. Ha-NeRF

		Brandenburg Gate			Sacre Coeur			Trevi Fountain		
		PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
R / 4	NeRF	19.62	0.8200	0.1455	16.21	0.7197	0.2181	16.40	0.6189	0.2422
	NeRF-W*	24.00	0.8758	0.1332	21.07	**0.8422**	<u>0.1119</u>	19.75	0.7207	0.2029
	Ha-NeRF	24.58	0.8829	0.0927	20.36	0.7947	0.1317	<u>20.27</u>	<u>0.7270</u>	<u>0.1628</u>
	CR-NeRF-R (Ours)	<u>26.18</u>	<u>0.8937</u>	<u>0.0840</u>	<u>21.64</u>	0.8206	0.1160	20.08	0.6538	0.2372
	CR-NeRF (Ours)	**26.86**	**0.9069**	**0.0733**	**22.03**	<u>0.8369</u>	**0.1060**	**22.02**	**0.7488**	**0.1354**
R / 2	NeRF	18.90	0.8159	0.2316	15.60	0.7155	0.2916	16.14	0.6007	0.3662
	NeRF-W*	24.17	0.8905	0.1670	19.20	0.8076	0.1915	18.97	0.6984	0.2652
	Ha-NeRF	24.04	0.8773	0.1391	20.02	0.8012	0.1710	20.18	0.6908	0.2225
	CR-NeRF-R (Ours)	<u>25.94</u>	<u>0.8929</u>	<u>0.1378</u>	<u>21.66</u>	<u>0.8171</u>	<u>0.1646</u>	<u>21.37</u>	<u>0.7111</u>	<u>0.2212</u>
	CR-NeRF (Ours)	**26.53**	**0.9003**	**0.1060**	**22.07**	**0.8233**	**0.1520**	**21.48**	**0.7117**	**0.2069**

Table 1. Quantitative experimental results on three real-world datasets under two resolution settings, *i.e.,* downscaling original image resolution by 2 (R/2) and 4 (R/4). The **bold** and the <u>underlined</u> numbers indicate the best and second-best results, respectively.

estimates a visibility map by separately utilizing a UV coordinate and a conditional feature of a reference image. Differently, CR-NeRF considers information across multiple rays. Specifically, CR-NeRF takes m rays as input, fuses them with a conditional feature, and generates a region of an image simultaneously (see Fig. 1 (b)). A recent work *i.e.,*, 4K-NeRF To capture ray correlation, leverages depth-modulated convolutions. In contrast, CR-NeRF captures the covariance of different rays. We theoretically (*c.f.* Sec. 4.1) and empirically (see details in the appendix) analyze the necessity of considering multiple rays.

5. Experiments

Implementation details. We implement our approach using PyTorch [36] and train our networks with Adam [20] optimizer. For a fair comparison, we follow all common hyper-parameter settings same as Ha-NeRF [6], *e.g.,* setting the number of input rays, learning rate, λ and height and width of fully connected layers to 1024, 5×10^{-4}, 1×10^{-3}, 8 and 256, respectively. We set β to 1×10^{-5}. For a thorough study, we downscale the original image of each dataset by 2 times (R/2) and 4 times (R/4). During inference, we omit the segmentation network $\mathcal{S}_{\theta_\Delta}$, see more details about the inference of CR-NeRF in the appendix.

Datasets, metrics, and comparison methods. Following Ha-NeRF [6], we evaluate our proposed method on three datasets: Brandenburg Gate, Sacre Coeur, and Trevi Fountain. For visual inspection, we present rendered images generated from the same set of input views. We also report quantitative results based on PSNR, SSIM [51], and LPIPS [60, 16]. We evaluate our proposed method against NeRF [33], NeRF-W [31], Ha-NeRF [6]. For ablation studies, we construct several variants of our CR-NeRF: 1) CR-NeRF-R replaces the cross-ray features from CR-NeRF with a ray-point-level features, *i.e.,* features of ray points along multiple rays; 2) CR-NeRF-B is constructed upon CR-NeRF without the cross-rays appearance modeling module and transient handling module; 3) CR-NeRF-A eliminates the cross-rays appearance modeling module only and 4) CR-NeRF-T removes the transient handling module.

5.1. Comparison Experiments

Quantitative experiments. We conduct extensive experiments on Brandenburg Gate, Sacre Coeur, and Trevi Fountain datasets. We follow Ha-NeRF with the image resolution setting of 2× downscaling (R/2) and further evaluate the effectiveness of our CR-NeRF on 4× downscaling (R/4). As demonstrated in Tab. 1, we observe that vanilla NeRF performs worst among all methods, since NeRF assumes the scene behind the training images is static. By modeling the style embedding and handling the transient objects, NeRF-W and Ha-NeRF achieve competitive performance in terms of PSNR, SSIM, and LPIPS. Note that NeRF-W optimizes its style embedding on test images since NeRF-W can not transfer to unseen test images directly. Thus, the comparison with NeRF-W is unfair. Even with the unfair comparison, thanks to the cross-ray manner, our CR-NeRF outperforms NeRF-W and Ha-NeRF on Brandenburg and Trevi under two downscaling settings.

Qualitative experiments. We summarize the qualitative results of all comparison methods in Fig. 3. We observe that NeRF produces foggy artifacts and inaccurate appearance. NeRF-W and Ha-NeRF are able to reconstruct a more promising 3D geometry and model appearance from the ground truth image. However, the reconstructed geometry is not accurate enough, *e.g.,* the shape of the green plant and ghost effects around the pillar in Brandenburg, the cavity in Sacre. Besides, the transferred appearance is not realistic enough, *e.g.,* sunshine on statues in Sacre, and the color of blue sky and grey roof in Trevi. Differently, our CR-NeRF introduces a cross-ray paradigm and therefore achieves more

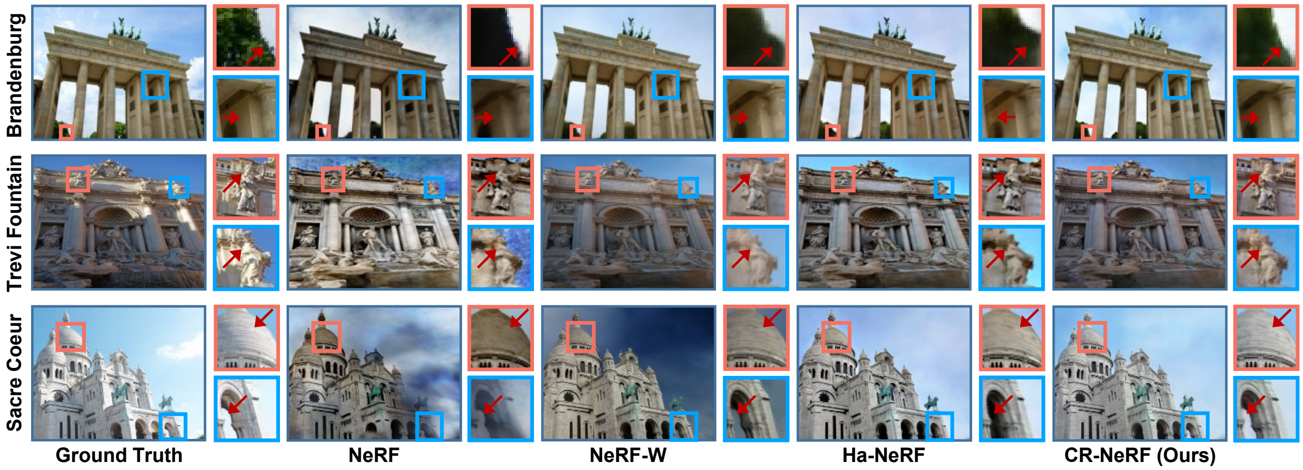


Figure 3. Qualitative experimental results on three unconstrained datasets. CR-NeRF recovers realistic appearance (*e.g.,* green plant in Brandenburg, sunshine on statues in Trevi, and light blue sky in Sacre.). Moreover, CR-NeRF removes transient objects for a consistent geometry (*e.g.,* less ghost effects around pillars of Brandenburg and Sacre).

Figure 4. Modeling appearance in Brandenburg and Trevi datasets using various viewing directions and appearance images. The viewing directions of the synthesized images are the same as that of the nearest content images on the left. (a, b) appearance images from Brandenburg and Trevi, respectively. (c, e) Content images from Trevi. (g, i) Content image from Brandenburg. (d, f) Synthesized images in Trevi. (h, j) Synthesized images in Brandenburg.

realistic appearance modeling and reconstructs a consistent geometry by suppressing transient objects.

Comparison of appearance modeling. We investigate the appearance modeling ability of our CR-NeRF in Fig. 4. We observe that 1) CR-NeRF captures appearance information more accurately than Ha-NeRF, especially towards recovering appearances from images with high-frequency information, *e.g.,* green sky, blue sky, red building, sunlight on the gate. 2) CR-NeRF successfully removes transient objects such as tourists and cars while retaining static objects such as roads and buildings.

5.2. Ablation Studies

Ablation of appearance module and transient module. We summarize the ablation studies of CR-NeRF on Brandenburg, Sacre, and Trevi dataset in Tab. 2. We observe that CR-NeRF-A and CR-NeRF-T outperform CR-NeRF-B. and CR-NeRF exceeds all variants, indicating the effectiveness of our Appearance Module and Transient Module.

Cross-ray manner and fusing level. We study the ef-

	Brandenburg Gate			Sacre Coeur			Trevi Fountain		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
CR-NeRF-B	19.58	0.8216	0.1470	16.11	0.7145	0.2196	16.37	0.6206	0.2493
CR-NeRF-A	26.38	0.8929	0.0885	21.67	0.8182	0.1127	21.85	0.7473	0.1388
CR-NeRF-T	20.46	0.8361	0.1300	16.28	0.7650	0.1799	16.55	0.6446	0.2230
CR-NeRF	**26.86**	**0.9069**	**0.0733**	**22.03**	**0.8369**	**0.1060**	**22.02**	**0.7488**	**0.1354**

Table 2. Ablation studies of CR-NeRF on three datasets. The performance of our baseline (CR-NeRF-B) is progressively improved by adding the appearance modeling module (CR-NeRF-A) and the transient handler (CR-NeRF-T). The bold numbers indicate the best result.

fectiveness of the cross-ray manner and the fusing level by comparing with our baseline CR-NeRF-R quantitatively in Tab. 1 and qualitatively in the appendix. From Tab. 1, CR-NeRF-R achieves a competitive performance on three datasets, which shows the superiority of leveraging various rays. Moreover, our proposed CR-NeRF outperforms CR-NeRF-R consistently on all datasets. We assume that compared with the cross-ray-point features, the granularity of the cross-ray features $\mathcal{F}^{\text{cr}}$ is closer to that of the image-level conditional features. Therefore, feature fusion is more effective. we provide qualitative results in the appendix.

5.3. Further Experiments

Unseen appearance modeling. Our proposed CR-NeRF is able to deal with unseen appearance images thanks to the ability of our cross-ray appearance modeling handler. As shown in Fig. 5, our CR-NeRF captures the whole range appearance (*e.g.,* the blue and purple appearance in the last two columns in Brandenburg and Trevi fountain datasets) of the given style image more accurately compared with Ha-NeRF. Moreover, our CR-NeRF synthesizes a more consistent appearance than images generated by Ha-NeRF (*e.g.,* the sudden bright light on the sky of the second column in Brandenburg dataset). Note that NeRF-W needs to optimize its appearance embedding on each test image by pixel-level supervision, thus NeRF-W cannot be directly applied to unseen appearance modeling.

Inference time on multiple images. When dealing with multiple images of various appearances with fixed camera position, the inference efficiency of our CR-NeRF exceeds Ha-NeRF significantly (*i.e.,* 2.12 seconds vs 24.09 seconds in Tab. 3). The reason is that our CR-NeRF generates cross-ray features $\mathcal{F}^{\text{cr}}$ only once by using a NeRF backbone and synthesizes various appearances by fusing $\mathcal{F}^{\text{cr}}$ and appearance embedding of each image. In contrast, Ha-NeRF requires the use of its NeRF backbone for each estimation. For efficiency, we modify Ha-NeRF by saving its interim results. However, since the interim results of Ha-NeRF occupy a large amount of GPU memory beyond the capacity of the single TITAN Xp GPU, moving the results to the host memory requires additional I/O time.

Transient objects handling. We observe that simply masking common objects harms reconstruction performance.

(Seconds per Image)	A Single Image		Multiple Images	
	R/4	R/2	R/4	R/2
Ha-NeRF	4.92s	14.01s	3.88s	24.09s
CR-NeRF (Ours)	4.02s	15.07s	0.92s	2.12s

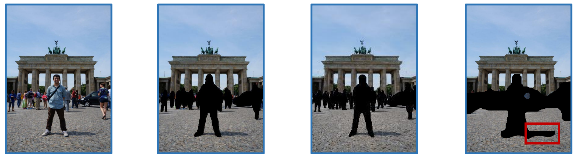
Table 3. Inference time of CR-NeRF and Ha-NeRF with one TITAN Xp GPU on Brandenburg with two downscaling ratios.

Figure 5. Modeling appearance from unseen images with high-frequency information to Brandenburg and Trevi.

	PSNR (↑)	SSIM (↑)	LPIPS (↓)
CR-NeRF-A	26.38	0.8929	0.0885
CR-NeRF-A + DeepLabV3	24.89$_{(\downarrow 1.49)}$	0.8781$_{(\downarrow 0.0148)}$	0.1065$_{(\uparrow 0.0180)}$
CR-NeRF-A + Mask R-CNN	25.46$_{(\downarrow 0.92)}$	0.8885$_{(\downarrow 0.0044)}$	0.0919$_{(\uparrow 0.0034)}$
CR-NeRF	**26.86**$_{(\uparrow 0.48)}$	**0.9069**$_{(\uparrow 0.0140)}$	**0.0733**$_{(\downarrow 0.0152)}$

Table 4. Discussion on transient handlers on Brandenburg dataset.

Specifically, we use a pre-trained DeepLabV3 and a pre-trained Mask R-CNN that produce promising segmentation results for common objects such as pedestrians and cars (we carefully choose the categories for estimation to avoid masking out static objects). However, performance degrades when combining CR-NeRF-A with these two networks (see Tab. 4). Considering that the transient handler of CR-NeRF is trained without the supervision of ground truth visibility maps, our estimated visibility maps are inevitably less accurate than the pre-trained network on the common objects (see the appendix for more details). We assume the definition of transient objects is still an open question and we leave it to our future work.



Reference Image DeepLabV3 Mask R-CNN CR-NeRF

Figure 6. Transient objects from pre-trained DeeplabV3 [4], pre-trained Mask R-CNN [14] and CR-NeRF. CR-NeRF captures a shadow of a tourist without using segmentation labels for training.

6. Conclusion

In this paper, we address novel-view synthesis from unconstrained images by considering the information of multiple rays within a scene. The unconstrained scenario introduces the varying appearances and transient objects in the images. We propose a novel cross-ray paradigm for the task by leveraging global interactive information across multiple rays. Guided by the paradigm, to address the variable appearance, we propose to represent information of multiple rays with cross-ray features and then inject an appearance of each image via fuse feature covariance of the rays and the image appearance. To handle transient objects, we propose a novel perspective of handling transient objects via image segmentation on multiple rays. Based on this, we estimate and grid sample a visibility map to pair with the rays. Extensive experimental results on large real-world datasets show the effectiveness of our proposed method.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (NSFC) (62072190), National Natural Science Foundation of China (NSFC) 61836003 (key project), Program for Guangdong Introducing Innovative and Enterpreneurial Teams 2017ZT07X183.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, pages 5470–5479, 2022. 1

[2] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *SIGGRAPH*, pages 425–432, 2001. 3

[3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, pages 14124–14133, 2021. 1

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *CVPR*, page 4067–4075, 2017. 5, 9

[5] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *SIGGRAPH*, pages 279–288, 1993. 3

[6] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *CVPR*, pages 12943–12952, 2022. 2, 4, 6

[7] Chia-Ming Cheng, Shu-Jyuan Lin, Shang-Hong Lai, and Jinn-Cherng Yang. Improved novel view synthesis from depth image with large baseline. In *ICPR*, pages 1–4, 2008. 3

[8] Gang Dai, Yifan Zhang, Qingfeng Wang, Qing Du, Zhuliang Yu, Zhuoman Liu, and Shuangping Huang. Disentangling writer and character styles for handwriting generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5977–5986, 2023. 5

[9] Robert A Drebin, Loren Carpenter, and Pat Hanrahan. Volume rendering. In *SIGGRAPH*, pages 65–74, 1988. 3

[10] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, pages 605–613, 2017. 2

[11] 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 *CVPR*, pages 2367–2376, 2019. 3

[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 5

[13] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *CVPR*, pages 5784–5794, 2021. 1

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 5, 9

[15] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *ICCV*, pages 9984–9993, 2019. 2

[16] Hongxiang Huang, Daihui Yang, Gang Dai, Zhen Han, Yuyi Wang, Kin-Man Lam, Fan Yang, Shuangping Huang, Yongge Liu, and Mengchao He. Agtgan: Unpaired image translation for photographic ancient character generation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5456–5467, 2022. 6

[17] Li Jiang, Shaoshuai Shi, Xiaojuan Qi, and Jiaya Jia. Gal: Geometric adversarial loss for single-view 3d-object reconstruction. In *ECCV*, pages 802–816, 2018. 2

[18] Yifan Jiang, Peter Hedman, Ben Mildenhall, Dejia Xu, Jonathan T Barron, Zhangyang Wang, and Tianfan Xue. Alignerf: High-fidelity neural radiance fields via alignment-aware training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 46–55, 2023. 2

[19] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *CVPR*, pages 1251–1261, 2020. 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[21] Marc Levoy and Pat Hanrahan. Light field rendering. In *SIGGRAPH*, pages 31–42, 1996. 3

[22] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 2

[23] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *CVPR*, pages 3809–3817, 2019. 5

[24] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *NeurIPS*, pages 386–396, 2017. 5

[25] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 2

[26] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. In *AAAI*, pages 7114–7121, 2018. 2

[27] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *CVPR*, pages 2019–2028, 2020. 2

[28] Frank P-W Lo, Yingnan Sun, Jianing Qiu, and Benny Lo. Food volume estimation based on deep learning view synthesis from a single depth map. *Nutrients*, 10(12):2005, 2018. 3

[29] Ming Lu, Hao Zhao, Anbang Yao, Yurong Chen, Feng Xu, and Li Zhang. A closed-form solution to universal style transfer. In *ICCV*, pages 5952–5961, 2019. 5

[30] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NeurIPS*, 2016. 5

[31] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, pages 7210–7219, 2021. 1, 2, 4, 6

[32] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, pages 4460–4470, 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, 2021. 2, 3, 6

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

[35] 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 *ICCV*, pages 5865–5874, 2021. 1

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 6

[37] Kebin Peng, Rifatul Islam, John Quarles, and Kevin Desai. Tmvnet: Using transformers for multi-view voxel-based 3d reconstruction. In *CVPR*, pages 222–230, 2022. 2

[38] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *ECCV*, pages 523–540, 2020. 2

[39] 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

[40] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *ICCV*, pages 14104–14113, 2021. 3

[41] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 3

[42] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, pages 20154–20166, 2020. 4, 5

[43] 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

[44] Pratul P Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. In *ICCV*, pages 2243–2251, 2017. 3

[45] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, pages 12278–12291, 2021. 1

[46] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. 2

[47] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12922–12931, 2022. 2

[48] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12375–12385, 2023. 2

[49] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 3

[50] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit

surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, pages 27171–27183, 2021. 1

[51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6

[52] Zhongshu Wang, Lingzhi Li, Zhen Shen, Li Shen, and Liefeng Bo. 4k-nerf: High fidelity neural radiance fields at ultra high resolutions. *arXiv preprint arXiv:2212.04701*, 2022. 2

[53] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. *ACM TOG*, 24(3):765–776, 2005. 3

[54] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, pages 7467–7477, 2020. 3

[55] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *CVPR*, pages 8534–8543, 2021. 3

[56] Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Yongdong Zhang. Cgnet: A light-weight context guided network for semantic segmentation. *IEEE TIP*, 30:1169–1179, 2020. 5

[57] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *ICCV*, pages 2690–2698, 2019. 2

[58] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. In *NeurIPS*, pages 20683–20695, 2021. 2

[59] Jingyang Zhang, Yao Yao, and Long Quan. Learning signed distance field for multi-view surface reconstruction. In *ICCV*, pages 6525–6534, 2021. 2

[60] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6