This CVPR 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections

Ricardo Martin-Brualla*, Noha Radwan*, Mehdi S. M. Sajjadi*,
Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth

Google Research

{rmbrualla, noharadwan, msajjadi, barron, adosovitskiy, duckworthd}@google.com

Abstract

We present a learning-based method for synthesizing novel views of complex scenes using only unstructured collections of in-the-wild photographs. We build on Neural Radiance Fields (NeRF), which uses the weights of a multilayer perceptron to model the density and color of a scene as a function of 3D coordinates. While NeRF works well on images of static subjects captured under controlled settings, it is incapable of modeling many ubiquitous, real-world phenomena in uncontrolled images, such as variable illumination or transient occluders. We introduce a series of extensions to NeRF to address these issues, thereby enabling accurate reconstructions from unstructured image collections taken from the internet. We apply our system, dubbed NeRF-W, to internet photo collections of famous landmarks, and demonstrate temporally consistent novel view renderings that are significantly closer to photorealism than the prior state of the art.

1. Introduction

Synthesizing novel views of a scene from a sparse set of captured images is a long-standing problem in computer vision, and a prerequisite to many AR and VR applications. Though classic techniques have addressed this problem using structure-from-motion [11] or image-based rendering [30], this field has recently seen significant progress due to *neural rendering* techniques — learning-based modules embedded within a 3D geometric context, and trained to reconstruct observed images. The Neural Radiance Fields (NeRF) approach [25] models the radiance field and density of a scene with the weights of a neural network. Volume rendering is then used to synthesize new views, demonstrating a heretofore unprecedented level of fidelity on a range of challenging scenes. However, NeRF has only been demonstrated to work well in controlled settings: the scene is captured within a short time frame during which lighting effects remain constant, and all content in the scene is static. As we will demonstrate, NeRF's performance degrades significantly when presented with moving objects or variable illumination. This limitation prohibits direct application of NeRF to large-scale in-the-wild scenarios, where input images may be taken hours or years apart, and may contain pedestrians and vehicles moving through them.

(a) Photos (b) Renderings

Figure 1: Given only an internet photo collection (a), our method is able to render novel views with variable illumination (b). Photos by Flickr users dbowie78, vasnic64, punch / CC BY.

The central limitation of NeRF that we address here is its assumption that the world is geometrically, materially, and photometrically *static* — that the density and radiance of the world is constant. NeRF therefore requires that any two photographs taken at the same position and orientation must be identical. This assumption is severely violated in many real-world datasets, such as large-scale internet photo collections of tourist landmarks. Two photographers may stand in the same location and photograph the same landmark, but in the time between those two photographs the world can change significantly: cars and people may move, construction may begin or end, seasons and weather may change, the sun may move through the sky, etc. Even two photos

*Denotes equal contribution.

taken at the same time and location can exhibit considerable variation: exposure, color correction, and tone-mapping all may vary depending on the camera and post-processing. We will demonstrate that naively applying NeRF to in-the-wild photo collections results in inaccurate reconstructions that exhibit severe ghosting, oversmoothing, and further artifacts.

To handle these demanding scenarios, we present NeRF-W, an extension of NeRF that relaxes its strict consistency assumptions. First, we model per-image appearance variations such as exposure, lighting, weather, and post-processing in a learned low-dimensional latent space. Following the framework of Generative Latent Optimization [3], we optimize an appearance embedding for each input image, thereby granting NeRF-W the flexibility to explain away photometric and environmental variations between images by learning a shared appearance representation across the entire photo collection. The learned latent space provides control of the appearance of output renderings as illustrated in Figure 1, (b). Second, we model the scene as the union of shared and image-dependent elements, thereby enabling the unsupervised decomposition of scene content into "static" and "transient" components. Our approach models transient elements using a secondary volumetric radiance field combined with a data-dependent uncertainty field, where the latter captures variable observation noise and further reduces the effect of transient objects on the static scene representation. Because optimization is able to identify and discount transient image content, we can synthesize realistic renderings of novel views by rendering only the static component.

We apply NeRF-W to several challenging in-the-wild photo collections of cultural landmarks and show that it can produce detailed, high-fidelity renderings from novel viewpoints, surpassing the prior state of the art by a large margin on PSNR and MS-SSIM. Unlike prior work, renderings from our model exhibit smooth appearance interpolation and temporal consistency, even for wide camera trajectories. We find that NeRF-W significantly improves quality over NeRF in the presence of appearance variation and transient occluders while achieving similar quality in controlled settings.

2. Related Work

The last decade has seen the integration of physics-based multi-view geometry techniques into deep learning-based approaches for the task of 3D scene reconstruction. Here we review recent progress on novel view synthesis and neural rendering, and highlight the main differences between existing approaches and our proposed method.

Novel View Synthesis: Constructing novel views of a scene captured by multiple images is a long standing problem in computer vision. Structure-from-Motion [11] and bundle adjustment [39] can be used to reconstruct a sparse point cloud representation and recover camera parameters. Photo

Figure 2: Example in-the-wild photographs from the Phototourism dataset [13] used to train NeRF-W. Due to variable illumination and post-processing (top), the same object's color may vary from image to image. In-the-wild photos may also contain transient occluding subjects (bottom). Photos by Flickr users paradasos, itia4u, jblesa, joshheumann, ojotes, chyauchentravelworld / CC BY.

Tourism [33] showed how these reconstruction techniques could be scaled to unconstrained photo collections and used to perform view synthesis [1, 10]. Other approaches to view synthesis include light-field photography [17] and image-based rendering [5] but these generally require a dense capture of the scene. Recent works explicitly infer the light and reflectance properties of the objects in the scene from a set of unconstrained photo collections [16, 29, 19] using them to manipulate scene appearance and geometry. Whereas other methods utilize semantic knowledge to reconstruct transient objects [27].

Neural Rendering: More recently, neural rendering techniques [36] have been applied to scene reconstruction. Several approaches employ image translation networks [12] to re-render content more realistically using as input traditional reconstruction results [21], learned latent textures [37], point clouds [2], voxels [31], or plane sweep volumes [8, 9]. Most similar in application to our work is Neural Rerendering in the Wild (NRW) [23] which synthesizes realistic novel views of tourist sites from point cloud renderings by learning a neural re-rendering network conditioned on a learned latent appearance embedding module. Common drawbacks of these approaches are the checkerboard and temporal artifacts visible under camera motion caused by the employed 2D image translation network. Another recent approach represents the scene as camera-centric multiplane images to reconstruct captured scenes [24, 43], and internet photo collections [18]. These methods produce photorealistic renderings of novel viewpoints but the views they can interpolate are restricted to a small volume surrounding the ground truth camera poses. In contrast, volume rendering approaches [20, 25, 32] allow for accurate and consistent reconstructions even with large camera motions, as does NeRF-W. Neural Radiance Fields (NeRF) [25] use a multi-layer perceptron (MLP) to model a radiance field at an unprecedented level of fidelity, in part thanks to the use of positional encoding within the MLP [35].

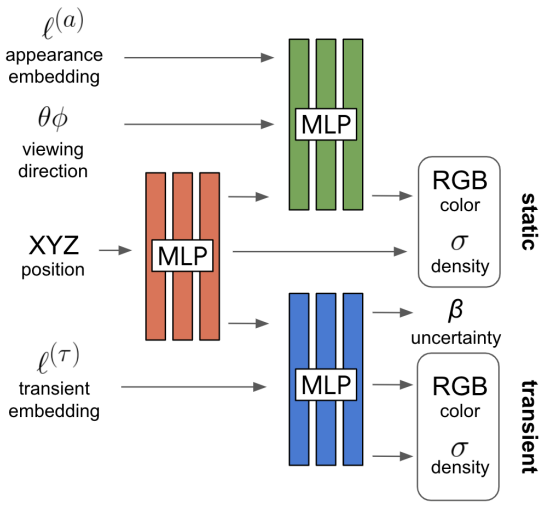


Figure 3: NeRF-W model architecture. Given a 3D position, viewing direction, and learned appearance and transient embeddings, NeRF-W produces static and transient colors and densities as well as a measure of uncertainty. Note that the static opacity is generated *before* the model is conditioned on the appearance embedding, ensuring that static geometry is shared across all images.

Our work focuses on extending NeRF to unconstrained scenarios, like internet photo collections.

3. Background

Our goal is to produce a system that takes as input a photo collection and then learns a 3D representation that is capable of generating the photos of that collection. Such a scene representation should encode the 3D structure of the scene together with appearance information so as to enable the synthesis of novel, unseen views. In the following we describe Neural Radiance Fields [25] (NeRF), the method for 3D scene reconstruction that NeRF-W extends.

NeRF represents a scene using a learned, continuous volumetric radiance field F_θ defined over a bounded 3D volume. F_θ is modeled using a multilayer perceptron (MLP) that takes as input a 3D position $\mathbf{x} = (x, y, z)$ and unit-norm viewing direction $\mathbf{d} = (d_x, d_y, d_z)$, and produces as output a density σ and color $\mathbf{c} = (r, g, b)$. To compute the color of a single pixel, NeRF approximates the volume rendering integral using numerical quadrature [22]. Let $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ be the camera ray emitted from the center of projection of a camera $\mathbf{o}$ through a given pixel on the image plane. NeRF's approximation of the expected color $\hat{\mathbf{C}}(\mathbf{r})$ of that pixel is:

$$\hat{\mathbf{C}}(\mathbf{r}) = \mathcal{R}(\mathbf{r}, \mathbf{c}, \sigma) = \sum_{k=1}^{K} T(t_k)\,\alpha(\sigma(t_k)\delta_k)\,\mathbf{c}(t_k)\,, \quad (1)$$

$$\text{where} \quad T(t_k) = \exp\left(-\sum_{k'=1}^{k-1} \sigma(t_{k'})\delta_{k'}\right), \quad (2)$$

where $\mathcal{R}(\mathbf{r}, \mathbf{c}, \sigma)$ is the volumetric rendering of color $\mathbf{c}$ with density σ, $\mathbf{c}(t)$ and $\sigma(t)$ are the color and density at point $\mathbf{r}(t)$, $\alpha(x) = 1 - \exp(-x)$, and $\delta_k = t_{k+1} - t_k$ is the distance between two quadrature points. Stratified sampling is used to select quadrature points $\{t_k\}_{k=1}^{K}$ between t_n and t_f, the near and far planes of the camera.

NeRF represents the volumetric density $\sigma(t)$ and color $\mathbf{c}(t)$ using ReLU MLPs of the following form:

$$[\sigma(t), \mathbf{z}(t)] = \text{MLP}_{\theta_1}(\gamma_{\mathbf{x}}(\mathbf{r}(t)))\,, \quad (3)$$

$$\mathbf{c}(t) = \text{MLP}_{\theta_2}(\mathbf{z}(t), \gamma_{\mathbf{d}}(\mathbf{d}))\,, \quad (4)$$

with parameters $\theta = [\theta_1, \theta_2]$ and fixed encoding functions $\gamma_{\mathbf{x}}$ (for position) and $\gamma_{\mathbf{d}}$ (for viewing direction). The final activations in generating $\sigma(t)$ and $\mathbf{c}(t)$ are a ReLU and a sigmoid respectively, as density must be non-negative and color must be in $[0, 1]$. Unlike [25], we describe the neural network as two MLPs where the latter depends on one output of the former, $\mathbf{z}(t)$, to highlight the fact that volume density $\sigma(t)$ is independent of viewing direction $\mathbf{d}$.

To fit parameters θ, NeRF minimizes the sum of squared reconstruction errors with respect to an RGB image collection $\{\mathcal{I}_i\}_{i=1}^{N}$, $\mathcal{I}_i \in [0, 1]^{H \times W \times 3}$. Each image $\mathcal{I}_i$ is paired with its corresponding intrinsic and extrinsic camera parameters which can be estimated using structure-from-motion [28]. We precompute the set of camera rays $\{\mathbf{r}_{ij}\}_{j=1}^{H \times W \times 3}$ corresponding to pixel j from image i with each ray passing through the 3D location $\mathbf{o}_i$ with direction $\mathbf{d}_{ij}$, where $\mathbf{r}_{ij}(t) = \mathbf{o}_i + t\,\mathbf{d}_{ij}$.

To improve sample efficiency, NeRF simultaneously optimizes two MLPs: one coarse and one fine, where the density predicted by the coarse model determines the sampling of quadrature points for the fine model. The parameters of both models are optimized by minimizing the following loss:

$$\sum_{ij} \left\|\mathbf{C}(\mathbf{r}_{ij}) - \hat{\mathbf{C}}^c(\mathbf{r}_{ij})\right\|_2^2 + \left\|\mathbf{C}(\mathbf{r}_{ij}) - \hat{\mathbf{C}}^f(\mathbf{r}_{ij})\right\|_2^2, \quad (5)$$

where $\mathbf{C}(\mathbf{r}_{ij})$ is the observed color of ray j in image $\mathcal{I}_i$, and $\hat{\mathbf{C}}^c$ and $\hat{\mathbf{C}}^f$ are the coarse and fine models respectively.

4. NeRF in the Wild

We now present NeRF-W, a system for reconstructing 3D scenes from in-the-wild photo collections. We build on NeRF [25] and introduce two enhancements explicitly designed to handle the challenges of unconstrained imagery.

(a) Static (b) Transient (c) Composite (d) Image (e) Uncertainty

Figure 4: NeRF-W separately renders the static (a) and transient (b) elements of the scene, and then composites them (c). Training minimizes the difference between the composite and the true image (d) weighted by uncertainty (e), which is simultaneously optimized to identify and discount anomalous image regions. Photo by Flickr user vasnic64 / CC BY.

Similar to NeRF, we learn a volumetric density representation F_θ from an unstructured photo collection $\{\mathcal{I}_i\}_{i=1}^N$ for which camera parameters are known. NeRF assumes consistency in its input views: that a point in 3D space observed from the same position and viewing direction in two different images has the same intensity. But this assumption is violated by internet photos (such as those shown in Figure 2) due to two distinct phenomena:

1) Photometric variation: In outdoor photography, time of day and atmospheric conditions directly impact the illumination (and consequently, the emitted radiance) of objects in the scene. This issue is exacerbated by photographic imaging pipelines, as variation in auto-exposure settings, white balance, and tone-mapping across photographs may result in additional photometric inconsistencies [4].

2) Transient objects: Real-world landmarks are rarely captured in isolation, without moving objects or occluders around them. Tourist photos of landmarks are particularly challenging, as they often contain posing human subjects and other pedestrians.

We propose two model components to address these issues. In Section 4.1 we extend NeRF to allow for image-dependent appearance and illumination variations such that photometric discrepancies between images can be modeled explicitly. In Section 4.2 we further extend this model by allowing transient objects to be jointly estimated and disentangled from a static representation of the 3D world. Figure 3 shows an overview of the proposed model architecture.

4.1. Latent Appearance Modeling

To adapt NeRF to variable lighting and photometric post-processing, we adopt the approach of Generative Latent Optimization (GLO) [3] in which each image $\mathcal{I}_i$ is assigned a corresponding real-valued appearance embedding vector $\boldsymbol{\ell}_i^{(a)}$ of length $n^{(a)}$. We replace the image-independent radiance $\mathbf{c}(t)$ in Equation (1) with an image-dependent radiance $\mathbf{c}_i(t)$, which also introduces a dependency on image index i to the approximated pixel color $\hat{\mathbf{C}}_i$:

$$\hat{\mathbf{C}}_i(\mathbf{r}) = \mathcal{R}(\mathbf{r}, \mathbf{c}_i, \sigma), \tag{6}$$

$$\mathbf{c}_i(t) = \mathrm{MLP}_{\theta_2}\left(\mathbf{z}(t), \gamma_{\mathbf{d}}(\mathbf{d}), \boldsymbol{\ell}_i^{(a)}\right). \tag{7}$$

The $\{\boldsymbol{\ell}_i^{(a)}\}_{i=1}^N$ embeddings are optimized alongside θ.

Using these appearance embeddings as input to only the branch of the network that emits color grants our model the freedom to vary the emitted radiance of the scene in a particular image while still guaranteeing that the 3D geometry (predicted earlier by MLP_{θ_1}) is static and shared across all images. By setting $n^{(a)}$ to a small value, we encourage optimization to identify a continuous space in which illumination conditions can be embedded, thereby enabling smooth interpolations between conditions as demonstrated in Figure 8.

4.2. Transient Objects

We address transient phenomena using two distinct design decisions: First, we designate the color-emitting MLP (Equation (4)) used in NeRF as the "static" head of our model, and we add an additional "transient" head that emits its own color *and density*, where that density is allowed to vary across training images. This enables NeRF-W to reconstruct images containing occluders without introducing artifacts into the static scene representation. Second, instead of assuming that all observed pixel colors are equally reliable, we allow our transient head to emit a field of *uncertainty* (much like our existing fields of color and density), which allows our model to adapt its reconstruction loss to ignore unreliable pixels and 3D locations that are likely to contain occluders. We model each pixel's color as an isotropic normal distribution whose likelihood we will maximize, and we "render" the variance of that distribution using the same volume rendering approach used by NeRF. These two model components allow NeRF-W to disentangle static and transient phenomena without explicit supervision.

To construct our transient head, we build on the volume rendering formulation of Equation (6) and augment the static density $\sigma(t)$ and radiance $\mathbf{c}_i(t)$ with transient counterparts $\sigma_i^{(\tau)}(t)$ and $\mathbf{c}_i^{(\tau)}(t)$,

$$\hat{\mathbf{C}}_i(\mathbf{r}) = \sum_{k=1}^{K} T_i(t_k)\left(\alpha(\sigma(t_k)\delta_k)\mathbf{c}_i(t_k) + \alpha\left(\sigma_i^{(\tau)}(t_k)\delta_k\right)\mathbf{c}_i^{(\tau)}(t_k)\right), \tag{8}$$

$$\text{where } T_i(t_k) = \exp\left(-\sum_{k'=1}^{k-1}\left(\sigma(t_{k'}) + \sigma_i^{(\tau)}(t_{k'})\right)\delta_{k'}\right). \tag{9}$$

The expected color of $\mathbf{r}(t)$ then becomes the alpha composite of both the static and the transient components.

We employ the Bayesian learning framework of Kendall et al. [15] to model the uncertainty of the observed color. We assume that observed pixel intensities are inherently noisy (aleatoric) and further that this noise is input-dependent (heteroscedastic). We model the observed color $\mathbf{C}_i(\mathbf{r})$ with an isotropic normal distribution with image- and ray-dependent variance $\beta_i(\mathbf{r})^2$ and mean $\hat{\mathbf{C}}_i(\mathbf{r})$. Variance $\beta_i(\mathbf{r})$ is "rendered" analogously to color via alpha-compositing according to the transient density $\sigma_i^{(\tau)}(t)$:

$$\hat{\beta}_i(\mathbf{r}) = \mathcal{R}(\mathbf{r}, \beta_i, \sigma_i^{(\tau)}). \tag{10}$$

To allow the transient component of the scene to vary across images, we assign each training image $\mathcal{I}_i$ a second embedding $\boldsymbol{\ell}_i^{(\tau)} \in \mathbb{R}^{n^{(\tau)}}$ that is given as input to the transient MLP,

$$\left[\sigma_i^{(\tau)}(t), \mathbf{c}_i^{(\tau)}(t), \tilde{\beta}_i(t)\right] = \text{MLP}_{\theta_3}\left(\mathbf{z}(t), \boldsymbol{\ell}_i^{(\tau)}\right), \tag{11}$$

$$\beta_i(t) = \beta_{\min} + \log\left(1 + \exp\left(\tilde{\beta}_i(t)\right)\right), \tag{12}$$

ReLU and sigmoid activations are used for $\sigma_i^{(\tau)}(t)$ and $\mathbf{c}_i^{(\tau)}(t)$, and a softplus is used as the activation for $\beta_i(t)$ (shifted by $\beta_{\min} > 0$, a hyperparameter that ensures a minimum importance is assigned to each ray). See Figure 3 for an illustration of our complete model architecture.

The loss for ray $\mathbf{r}$ in image i with true color $\mathbf{C}_i(\mathbf{r})$ is

$$L_i(\mathbf{r}) = \frac{\left\|\mathbf{C}_i(\mathbf{r}) - \hat{\mathbf{C}}_i(\mathbf{r})\right\|_2^2}{2\beta_i(\mathbf{r})^2} + \frac{\log \beta_i(\mathbf{r})^2}{2} + \frac{\lambda_u}{K}\sum_{k=1}^{K} \sigma_i^{(\tau)}(t_k). \tag{13}$$

The first two terms are the (shifted) negative log likelihood of $\mathbf{C}_i(\mathbf{r})$ according to a normal distribution with mean $\hat{\mathbf{C}}_i(\mathbf{r})$ and variance $\beta_i(\mathbf{r})^2$. Larger values of $\beta_i(\mathbf{r})$ attenuate the importance assigned to a pixel, under the assumption that it belongs to some transient object. The first term is balanced by the second, which corresponds to the log-partition function of the normal distribution and precludes the trivial minimum at $\beta_i(\mathbf{r}) = \infty$. The third term is an L_1 regularizer with a multiplier λ_u on (non-negative) transient density $\sigma_i^{(\tau)}(t)$, and this discourages the model from using transient density to explain away static phenomena.

At test time we omit the transient and uncertainty fields, and render only $\sigma(t)$ and $\mathbf{c}(t)$. See Figure 4 for an illustration of static, transient, and uncertainty components.

4.3. Optimization

Like NeRF, we simultaneously optimize two copies of F_θ: A fine model that uses the model and losses described above, and a coarse model that uses only the latent appearance modeling component. Alongside parameters θ we optimize per-image appearance embeddings $\{\boldsymbol{\ell}_i^{(a)}\}_{i=1}^N$ and transient embeddings $\{\boldsymbol{\ell}_i^{(\tau)}\}_{i=1}^N$. NeRF-W's loss function is then,

$$\sum_{ij} L_i(\mathbf{r}_{ij}) + \frac{1}{2}\left\|\mathbf{C}(\mathbf{r}_{ij}) - \hat{\mathbf{C}}_i^c(\mathbf{r}_{ij})\right\|_2^2, \tag{14}$$

λ_u, $\beta_{\min}$, and embedding dimensionalities $n^{(a)}$ and $n^{(\tau)}$ form the set of additional hyperparameters for NeRF-W.

As optimization only produces appearance embeddings $\{\boldsymbol{\ell}_i^{(a)}\}$ for images in the training set, the embeddings of test-set images are unspecified. For test-set visualizations, we choose $\boldsymbol{\ell}^{(a)}$ to best fit a target image (*e.g.* Figure 8) or set it to an arbitrary value.

(a) NeRF-W w/o opt. (b) NeRF-W (c) Reference

Figure 5: Because optimization only yields appearance embeddings $\boldsymbol{\ell}^{(a)}$ for images in the training set, when evaluating error metrics on test-set images we optimize $\boldsymbol{\ell}^{(a)}$ to match the appearance of the true image using only the left half of each image. Error metrics are evaluated on only the right half of each image, so as to avoid information leakage. Photo by Flickr user eadaoinflynn / CC BY.

5. Experiments

Here we provide an evaluation of NeRF-W on unconstrained (*e.g.* "in-the-wild") internet photo collections of cultural landmarks. We select six landmarks from the Phototourism dataset [13]. Inspired by prior work [23], we reconstruct the *Trevi Fountain* and *Sacre Coeur* as well as four novel scenes, the *Brandenburg Gate*, *Taj Mahal*, *Prague Old Town Square*, and *Hagia Sophia*. Empirical performance for these scenes can be found in Table 1, but we urge the reader to visually inspect the video results in the supplement.

Baselines: We evaluate our proposed method against Neural Rerendering in the Wild (NRW) [23], NeRF [25], and two ablations of NeRF-W: NeRF-A (appearance), wherein the "transient" head is eliminated; and NeRF-U (uncertainty), wherein the appearance embedding $\boldsymbol{\ell}_i^{(a)}$ is eliminated. NeRF-W is the composition of NeRF-A and NeRF-U. While other recent work such as [18] is employed on a similar domain, we restrict baselines to those capable of extrapolating significantly beyond the views represented in the dataset.

Optimization: Building on NeRF[1], we implement all experiments in TensorFlow 2 using Keras. For each scene, we

[1] https://github.com/bmild/nerf

(a) Reference (b) NeRF (c) NeRF-W

Figure 6: Depth maps from NeRF and NeRF-W, rendered by computing the expected termination depth of each ray. NeRF's geometry is corrupted by appearance variation and occluders, while NeRF-W is robust to such phenomena and produces accurate 3D reconstructions. Photos by Flickr users burkeandhare, photogreuhphies / CC BY.

use COLMAP [28] with two radial and two tangential distortion parameters enabled to estimate each image's camera parameters. As in NeRF, for each scene we train a model initialized to random weights. We optimize all NeRF variants for 300,000 steps with a batch size of 2048 on 8 Nvidia V100 GPUs using Adam [7] (with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$), which takes approximately 2 days. Hyperparameters shared by all NeRF variants are chosen to maximize PSNR on the Brandenburg Gate dataset and are fixed to those values in all other scenes. Additional hyperparameters for variants of NeRF-W are chosen via grid search to maximize PSNR on a held-out validation set on the Brandenburg Gate scene and are fixed to those values for all other scenes. See the supplement for additional details on hyperparameters.

Evaluation: We evaluate on the task of novel view synthesis: given a held-out image with accompanying camera parameters, we render an image from the same pose and compare it to the ground truth. As measuring perceptual image similarity is challenging [26, 38, 40, 42], we present rendered images for visual inspection and report quantitative results based on PSNR, MS-SSIM [41], and LPIPS [42]. Because optimization only produces appearance embeddings for training-set images, when computing error metrics on test-set images we optimize an appearance embedding $\boldsymbol{\ell}^{(a)}$ on the left half of each image and report metrics on the right half (Figure 5). See the supplement for additional discussion of error metrics.

Results: Figure 7 shows qualitative results for all models and baselines on a subset of scenes. NRW produces renderings with checkerboard artifacts characteristic of 2D rerendering methods [14]. NRW is also sensitive to upstream errors in 3D geometry such as incomplete point clouds, as can be seen in the smaller towers of the church in the Prague Old Town. NeRF produces a consistent 3D geometry, but large parts of the scene have ghosting artifacts and occlusions, which are particularly noticeable on Sacre Coeur and Prague Old Town. Renderings from NeRF also tend to exhibit strong global color shifts when compared to the ground truth. These artifacts are the direct consequence of NeRF's static-world assumption — NeRF attempts to explain away all photometric variation and transient occlusion using a single scene representation. This static assumption impairs not only NeRF's renderings but also its underlying geometry, while NeRF-W produces accurate 3D reconstructions (Figure 6).

The NeRF-A ablation produces less "foggy" renderings than NeRF, as shown in Figure 7. However, NeRF-A is unable to reconstruct high-frequency details such as the brickwork on Sacre Coeur's dome. In contrast, the NeRF-U ablation is better able to capture fine detail, but is unable to model varying photometric effects. NeRF-W has the benefits of both ablations, and thereby produces sharper and more accurate renderings.

Quantitative results are summarized in Table 1. Optimizing NeRF on in-the-wild photo collections leads to particularly poor results that are unable to compete with NRW. In contrast, NeRF-W outperforms the baselines on PSNR and MS-SSIM across all datasets. In particular, NeRF-W improves over the previous state of the art NRW by an average margin of 4.4dB in PSNR, and with up to 40% improvements in MS-SSIM. In spite of minimizing only a per-pixel squared error during training, NeRF-W improves upon the prior state of the art on LPIPS in 3 of 6 scenes and remains competitive in the remainder. Lacking a perceptual loss, NeRF-W is not incentivized to produce the high-frequency textures favored by perceptual metrics such as LPIPS. However, NRW exhibits temporal instability — as the camera moves, renderings appear to flicker and wobble unrealistically, and this is not captured by the single-image metrics or figures used in this paper. We strongly encourage the reader to inspect the supplemental video to observe the temporal instability of NRW compared to NeRF and NeRF-W.

Controllable Appearance: One consequence of modeling appearance with a latent embedding space $\boldsymbol{\ell}^{(a)} \in \mathbb{R}^{n^{(a)}}$ is that it enables the modification of lighting and appearance of a rendering without altering the underlying 3D geometry. In Figure 1 (right), we see slices of four rendered images produced by NeRF-W using appearance embeddings associated with four training set images. In addition to the embeddings associated with images in the training set, one may also apply NeRF-W to arbitrary vectors in the same space. In Figure 8, we present five images rendered from a fixed camera

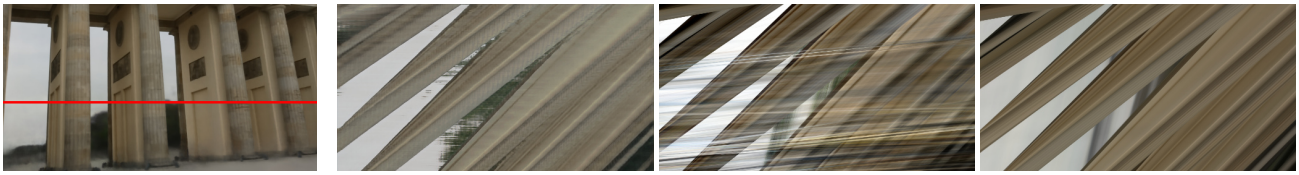
Figure 7: Qualitative results from experiments on the Phototourism dataset. NeRF-W is simultaneously able to model appearance variation (top), remove transient occluders (flag, middle), and reconstruct fine details in the scene (bottom). Further datasets are shown in Figure 14 (supplementary). Photos by Flickr users firewave, clintonjeff, leoglenn_g / CC BY.

position, where we interpolate between the appearance embeddings associated with the left and right training images. Note that the appearance of the rendered images smoothly transitions between the two end points without introducing artifacts to the 3D geometry. We encourage readers to view the supplementary video to better appreciate the naturalness

	Brandenburg Gate			Sacre Coeur			Trevi Fountain			Taj Mahal			Prague			Hagia Sophia		
	PSNR	MS-SSIM	LPIPS	PSNR	MS-SSIM	LPIPS	PSNR	MS-SSIM	LPIPS	PSNR	MS-SSIM	LPIPS	PSNR	MS-SSIM	LPIPS	PSNR	MS-SSIM	LPIPS
NRW [23]	23.85	0.914	0.141	19.39	0.797	0.229	20.56	0.811	0.242	21.24	0.844	**0.201**	19.89	0.803	**0.216**	20.75	0.796	**0.231**
NeRF	21.05	0.895	0.208	17.12	0.781	0.278	17.46	0.778	0.334	15.77	0.697	0.427	15.67	0.747	0.362	16.04	0.749	0.338
NeRF-A	27.96	0.941	0.145	24.43	0.923	0.174	26.24	0.924	0.211	25.99	0.893	0.225	22.52	0.870	0.244	21.83	0.820	0.276
NeRF-U	19.49	0.921	0.174	15.99	0.826	0.223	15.03	0.795	0.277	10.23	0.778	0.373	15.03	0.787	0.315	13.74	0.706	0.376
NeRF-W	**29.08**	**0.962**	**0.110**	**25.34**	**0.939**	**0.151**	**26.58**	**0.934**	**0.189**	**26.36**	**0.904**	0.207	**22.81**	**0.879**	0.227	**22.23**	**0.849**	0.250

Table 1: Quantitative results on the Phototourism dataset [13] for NRW [23], NeRF [25], and two ablations of the proposed model. Best results are **highlighted**. NeRF-W outperforms the previous state of the art across all datasets on PSNR and MS-SSIM and achieves competitive results in LPIPS. Note that LPIPS generally favours methods such as NRW trained with an adversarial or perceptual loss and it is less sensitive to typical GAN artifacts, see Figures 7 and 14 (supplementary).

Figure 8: Interpolations between the appearance embeddings $\ell^{(a)}$ of two training images (left, right), which results in renderings (middle) where color and illumination are interpolated but geometry is fixed. Note that the training images contain people (left) and lights (right) that do not appear in the renderings. Photos by Flickr users mightyohm, blatez / CC BY.



Reference NRW NeRF NeRF-W

Figure 9: Epipolar plane images (EPI) synthesized from videos rendered by different models for the Brandenburg Gate scene. The camera is translated from left to right along a straight path, and the horizontal line at the same position (red line, reference) is taken across all video frames and stacked vertically, producing the EPIs shown above. A temporally consistent video results in a clean and smooth EPI, while noise in an EPI indicates temporal flickering artifacts. NRW's video contains heavy flickering with transient objects popping in and out of the frame while NeRF produces severe ghosting artifacts in front of the landmarks. NeRF-W produces highly temporally consistent videos. We strongly encourage the readers to watch the video in the supplementary material.

of such interpolations.

View-consistency: Figure 9 shows "flatland" light field renderings for NRW, NeRF, and NeRF-W with the camera panning along a straight path. Renderings from NeRF-W are more view-consistent (the Lambertian scene content is correctly reconstructed as being constant across viewing directions) and exhibits significantly less flickering than NRW or NeRF. NRW is unable to model temporal consistency between frames for transient objects, while NeRF is forced to embed view-dependent effects as colored fog into its scene representation.

Limitations: While NeRF-W is able to produce photorealistic and temporally consistent renderings from unstructured photographs, rendering quality degrades in areas of the scene that are rarely observed in the training images, or only observed at very oblique angles, like the ground, as shown in Figure 10. Similar to NeRF, NeRF-W is also sensitive to camera calibration errors, which can lead to blurry reconstructions on the parts of the scene that have been imaged by incorrectly-calibrated cameras.

Figure 10: Limitations of NeRF-W on the Phototourism dataset. Rarely-seen parts of the scene (ground, left) and incorrect camera poses (lamp post, right) can result in blur.

Synthetic Experiments: The components of NeRF-W were designed to deal with specific forms of photometric inconsistency, such as color shifts and occluders. Unfortunately, the uncontrolled nature of the Phototourism dataset means that it is challenging to demonstrate that each model component does indeed address the confounding factor that it was designed to address. For this reason, in the supplement we present a controlled ablation study in which we construct variations of a synthetic dataset used in [25] wherein we manually introduce the phenomena we expect to find in in-the-wild imagery. As can be seen in the supplement, the results of this ablation study are consistent with our expectations.

6. Conclusion

We have presented NeRF-W, a novel approach for 3D scene reconstruction of complex environments from unstructured internet photo collections that builds upon NeRF. We learn a per-image latent embedding capturing photometric appearance variations often present in in-the-wild data, and we decompose the scene into image-dependent and shared components to allow our model to disentangle transient elements from the static scene. Experimental evaluation on real-world (and synthetic) data demonstrates significant qualitative and quantitative improvement over previous state-of-the-art approaches.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz, and Richard Szeliski. Building rome in a day. *Commun. ACM*, 2011.

[2] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. *arXiv preprint arXiv:1906.08240*, 2019.

[3] Piotr Bojanowski, Armand Joulin, David Lopez-Pas, and Arthur Szlam. Optimizing the latent space of generative networks. *ICML*, 2018.

[4] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing images for learned raw denoising. *CVPR*, 2019.

[5] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. *SIGGRAPH*, 2001.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] Jimmy Ba Diederik P. Kingma. Adam: A method for stochastic optimization. *ICLR*, 2015.

[8] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. *CVPR*, 2019.

[9] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. *CVPR*, 2016.

[10] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, and Marc Pollefeys. Building rome on a cloudless day. *ECCV*, 2010.

[11] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[13] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *arXiv preprint arXiv:2003.01587*, 2020.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *ECCV*, 2016.

[15] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS*, 2017.

[16] Pierre-Yves Laffont, Adrien Bousseau, Sylvain Paris, Frédo Durand, and George Drettakis. Coherent intrinsic images from photo collections. *SIGGRAPH Asia*, 2012.

[17] Marc Levoy and Pat Hanrahan. Light field rendering. *SIGGRAPH*, 1996.

[18] Zhengqi Li, Wenqi Xian, Abe Davis, and Noah Snavely. Crowdsampling the plenoptic function. *ECCV*, 2020.

[19] Andrew Liu, Shiry Ginosar, Tinghui Zhou, Alexei A Efros, and Noah Snavely. Learning to factorize and relight a city. In *European Conference on Computer Vision*, 2020.

[20] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *SIGGRAPH*, 2019.

[21] Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, Adarsh Kowdle, Christoph Rhemann, Dan B Goldman, Cem Keskin, Steve Seitz, Shahram Izadi, and Sean Fanello. Lookingood: Enhancing performance capture with real-time neural re-rendering. *SIGGRAPH Asia*, 2018.

[22] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1995.

[23] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. *CVPR*, 2019.

[24] 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. *SIGGRAPH*, 2019.

[25] 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. *ECCV*, 2020.

[26] Thrasyvoulos N Pappas, Robert J Safranek, and Junqing Chen. Perceptual criteria for image quality evaluation. *Handbook of image and video processing*, 110, 2000.

[27] True Price, Johannes L Schönberger, Zhen Wei, Marc Pollefeys, and Jan-Michael Frahm. Augmenting crowd-sourced 3d reconstructions using semantic detections. *CVPR*, 2018.

[28] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. *CVPR*, 2016.

[29] Qi Shan, Riley Adams, Brian Curless, Yasutaka Furukawa, and Steven M Seitz. The visual turing test for scene reconstruction. *3DV*, 2013.

[30] Heung-Yeung Shum, Shing-Chow Chan, and Sing Bing Kang. *Image-based rendering*. Springer Science & Business Media, 2008.

[31] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. *CVPR*, 2019.

[32] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *NeurIPS*, 2019.

[33] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo Tourism: Exploring photo collections in 3D. *SIGGRAPH*, 2006.

[34] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE TIP*, 2018.

[35] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains, 2020.

[36] Ayush Tewari, Christian Theobalt, Dan B Goldman, Eli Shechtman, Gordon Wetzstein, Jason Saragih, Jun-Yan Zhu, Justus Thies, Kalyan Sunkavalli, Maneesh Agrawala, Matthias Niessner, Michael Zollhöfer, Ohad Fried, Ricardo Martin Brualla, Rohit Kumar Pandey, Sean Fanello, Stephen Lombardi, Tomas Simon, and Vincent Sitzmann. State of the art on neural rendering. *Computer Graphics Forum*, 2020.

[37] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *SIGGRAPH*, 2019.

[38] Kim-Han Thung and Paramesran Raveendran. A survey of image quality measures. *TECHPOS*, 2009.

[39] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. *International Workshop on Vision Algorithms*, 1999.

[40] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 2002.

[41] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. *Asilomar Conference on Signals, Systems & Computers*, 2003.

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*, 2018.

[43] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018.