Factorized and Controllable Neural Re-Rendering of Outdoor Scene for Photo Extrapolation

Boming Zhao*
bmzhao@zju.edu.cn
Zhejiang University

Bangbang Yang*
ybbbbt@gmail.com
Zhejiang University

Zhenyang Li
zhenyoungle@gmail.com
Baidu Inc

Zuoyue Li
zuli@student.ethz.ch
ETH Zürich

Guofeng Zhang
zhangguofeng@zju.edu.cn
Zhejiang University

Jiashu Zhao
jzhao@wlu.ca
Wilfrid Laurier University

Zhejiang University

Dawei Yin
yindawei@acm.org
Baidu Inc

Zhaopeng Cui†
zhpcui@gmail.com
Zhejiang University

Hujun Bao†
bao@cad.zju.edu.cn
Zhejiang University

Internet Photo Collections

Geometry / Normal

Base Appearance

Captured Photo

Synthetic Sky

Scene Re-Rendering

Extrapolated Rendering & Realism Augmentation

Extrapolated Photo

Tone Mapping

Shadow

Applicatons

Inputs Factorized Rendering Component

Figure 1: Given an Internet photo collection of an outdoor attraction, we learn a novel neural re-rendering model that encodes scenes with several factorized components, which enables the applicabilities of controllable scene re-rendering, photo extrapolation and even extrapolated 3D photo generation (see GIF in Fig. 7). All the images are from the IMC-PT dataset [14]. Photos by Flickr users astrobri, soniadal82, Fotero, scriptingnews, Devin Ford, and MikiAnn.

Abstract

Expanding an existing tourist photo from a partially captured scene to a full scene is one of the desired experiences for photography applications. Although photo extrapolation has been well studied, it is much more challenging to extrapolate a photo (i.e., selfie) from a narrow field of view to a wider one while maintaining a similar visual style. In this paper, we propose a factorized neural re-rendering model to produce photorealistic novel views from cluttered outdoor Internet photo collections, which enables the applications including controllable scene re-rendering, photo extrapolation and even extrapolated 3D photo generation. Specifically, we first develop a novel factorized re-rendering pipeline to handle the ambiguity in the decomposition of geometry, appearance and illumination. We also propose a composited training strategy to tackle the unexpected occlusion in Internet images. Moreover, to enhance photo-realism when extrapolating tourist photographs, we propose a novel realism augmentation process to complement appearance details, which automatically propagates the texture details from a narrow captured photo to the extrapolated neural rendered image. The experiments and photo editing examples on outdoor scenes demonstrate the superior performance of our proposed method in both photo-realism and downstream applications. Code and the supplementary material are available on the project webpage: https://zju3dv.github.io/neural_outdoor_rerender/.

CCS Concepts

• Computing methodologies → Computer vision; Rendering.

Keywords

neural rendering, outdoor scene re-rendering, photograph editing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Boming Zhao, Bangbang Yang, Zhenyang Li, Zuoyue Li, Guofeng Zhang, Jiashu Zhao, Dawei Yin, Zhaopeng Cui, and Hujun Bao. 2022. Factorized and Controllable Neural Re-Rendering of Outdoor Scene for Photo Extrapolation. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), Oct. 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548125
1 Introduction

When a tourist visits a famous attraction, (s)he usually likes to take a picture that can capture both the person and the whole scene. However, in many cases, only a part of the scene can be captured due to the narrow field of view (FoV) of the camera and/or crowded people, which may be frustrating to the tourist. So similar to other image post-processing (e.g., super-resolution, deblurring, etc.), it would be of great help if we can handily extrapolate the photo to obtain a wider view of the scene, while maintaining a similar visual style (i.e., lighting condition or filtering effect) between the complemented image area and the original photo, as shown in the last row of Fig. 1. Moreover, it could be more attractive to expand the extrapolated photo to full-scene dynamic 3D with vivid effects as shown in Fig. 7.

The most straightforward solution relies on 2D image processing including 2D image stitching [5], 2D image extrapolation [43, 44] or 2D image generation [16, 30, 32, 40, 52]. 2D image stitching aims to combine multi-view images of small FoVs into a panorama, however it can only generate high-quality images when the input multi-view images are captured at the same location and time, which may be impossible for the crowded environment, selfie mode or the old photos captured before. Image extrapolation or generation learns to generate visually consistent content for the extrapolated regions using library images. However, such 2D methods are uncontrollable, which makes the extrapolated parts unreal even though they look plausible, and the library images have to be carefully captured and generally occlusion-free [43, 44]. Moreover, it is also impossible for these 2D methods to produce 3D photos with immersive user experience (see Fig. 7).

Another possible solution is to first reconstruct the corresponding outdoor scene landmark from a collection of images and then re-render the scene with a large field of view and lighting effect close to the original photos, which has been widely studied in computer vision and graphics [15, 33, 36, 49]. Traditional approaches use multi-view stereo and photometric blending to reconstruct a textured scene mesh [15, 42], so as to support the virtual presence of the scene but require high-quality image sets [28, 42] and are not feasible for adapting lighting conditions to a specific photograph. Recently, neural rendering shows promising results on surface mesh reconstructions, novel view synthesis and scene relighting. It also enables to model appearance variations of outdoor scene landmarks from cluttered Internet photo collections [6, 23, 25, 39], which liberates the restriction of the data requirement and delivers high flexibility for lighting effect adaptation. However, existing works either disentangle lighting variations in a latent space while being regardless of explicitly controlling illumination changes [6, 23, 39], or require carefully captured images to ensure a smooth factorization of material appearances [3, 4, 37, 57]. Besides, due to the nature of network smoothness, the rendered images from the neural implicit rendering tend to average the observations and inevitably lose some appearance details (e.g., gushing springs, etc.), which largely degrades the user experience of photo extrapolation.

In this paper, we propose a novel factorized and controllable neural re-rendering pipeline that enables realistic outdoor photo extrapolation from the readily accessible but cluttered Internet photo collections. Instead of modeling appearance variations as a whole in a latent space [6, 23, 25, 39], our rendering model factorizes scene representation into several components (see Fig. 1), including base appearance, scene geometry (and normal), synthetic sky and an explicitly explainable illumination condition (with data-driven HDR environment map, affine tone mapping and learnable shadow). Once the scene has been encoded in the rendering model, we can easily perform controllable scene re-rendering under novel views with user-selected lighting conditions, and conduct photo extrapolation or even extrapolated 3D photo generation that extends a captured tourist selfie from a narrow FoV to a wider view, while maintaining similar lighting effect and appearance details by utilizing photo adaptation and a novel realism augmentation mechanism.

However, it is non-trivial to learn a factorized scene representation from cluttered outdoor photos and conduct realistic photo extrapolation with neural rendering models. 1) Due to the ill-posed nature of the problem, naive solutions of NeRF-based inverse rendering [3, 4, 37, 57] are no longer applicable for such cluttered images and unbounded outdoor scenes, hence we propose a novel rendering pipeline for this challenging task. Specifically, at the rendering stage, we utilize a data-driven sky HDR decoder from Gillan et al. [13] to constrain the HDR map optimization in a reasonable latent space and resolve the scale ambiguity between the recovered HDR environment map and the base appearance. Then, to model the unobserved shadow caster (e.g., buildings behind the attraction), we introduce a learnable shadow branch that provides spatial shadow value during volume rendering. Finally, to handle dramatic color distortion that is beyond physically explainable lighting (e.g., user’s filtering effect), we apply a learnable affine tone mapping [29, 39] to the rendered pixels. 2) As the training images are collected in a crowd-sourcing paradigm from the Internet, there might be unexpected occluders (e.g., tourists or birds) that affect the learning of factorized rendering even with transient modeling [23] (see Sec 4.6). To tackle this challenge, we employ a composited training scheme to first train the geometry model and then train the rendering model with distilled occlusion-free images from NeRF-W [23]. This process can be regarded as transferring the latent appearance embedding into an explicit and controllable illumination parameterization. 3) Since the neural implicit model tends to fuse appearances from multi-view observation, the re-rendered scene is somehow more blurry than the user-captured photo and also lacks some fine details such as water of the fountain splashes and clouds. To bridge the gap between the neural rendering and the tourist photos during photo extrapolation, we propose a novel realism augmentation by fully exploiting rich textures from the given photo and propagating them into the rendered view. In this way, the extrapolated photograph can be more visually coherent to the captured one.

Our contribution can be summarized as follows. First, we propose a novel factorized neural rendering model which learns to encode unbounded outdoor scenes from cluttered Internet photo collections, and delivers the capability of controllable scene re-rendering, photo extrapolation and even extrapolated 3D photo generation. Second, to tackle the challenges of learning outdoor scene representation, our factorized rendering pipeline enables to handle varying lighting effects and color distortions by utilizing a composited training scheme to guide the training process. Moreover, a novel realism augmentation mechanism is also proposed to effectively complement details from a narrow-view real photo.
to a wide-view synthesized image. At last, the experiments and photography editing examples on several outdoor attractions show the superiority of our method in scene re-rendering, photo extrapolation, and extrapolated 3D photo generation.

2 Related Works

Outdoor scene reconstruction and rendering. Traditional methods generally use SFM [33] and MVS techniques to reconstruct surface mesh [15, 49], and blend colored images [28, 42] to obtain a textured mesh for visualization. But they require high-quality images with consistent illumination condition, and cannot handle Internet photo collections [36] with varying lighting and frequent occlusions. Recently, researchers use neural rendering techniques for outdoor scene rendering [6, 20, 23, 25, 26, 29, 39, 48]. Li et al. [20] uses multi-plane images to render outdoor attractions from photo collections, but cannot produce reasonable views when looking from the tilted side view due to the limitation of MPIs [26], NRW [25], NeRF-W [23] and their following works [6, 29, 39, 48, 50] model outdoor lightings with variations with a latent appearance code, which enable novel view synthesis with customizable camera trajectories and support appearance transition with code interpolation. However, as these methods learn appearance variations in a stand-alone latent space, they do not support controllable re-rendering with user-selected lighting effects.

Scene rendering with controllable illumination. Early methods mainly rely on optical equipment to measure the geometry [21, 53], reflectance [24, 41] and environment lighting [8, 38] for relightable scene rendering. In recent years, researchers propose to solve the scene relighting (or inverse rendering) with neural networks [13, 18, 19, 22, 54, 55, 58], but only support static photograph. Very recently, some works [3, 4, 10, 17, 37, 57] build up a relightable neural rendering framework upon object-centric neural volume rendering [2, 26, 51], which produce a “self-occlusion style” shadow effect by utilizing visibility from learned density field. However, they cannot be extended to large-scale outdoor scenes and are also not capable of handling noisy observations such as Internet photo collections. In parallel to our works, Rudnev et al. [31] proposes to learn a neural radiance field for outdoor scene relighting, but shows limited ability of representing and re-rendering on Internet photo collections due to the coarse learned geometry (or surface normal) and simplified lighting model. Instead, as our surface normal is derived from the SDF field and we utilize a more flexable external lighting model, our approach can be applied to a broader range of outdoor scenes with cluttered photo collections.

Photo extrapolation. Photo extrapolation (a.k.a. image outpainting/expansion) can extend a given image with a narrow FoV to a wide view. Early approaches either build up a photo library [43] or a short video clip [44] of the surrounding scene, and perform image-montage or stitching to outpaint the images. Since these approaches use reliable reference images to ensure consistent appearance in the extended area, they require laborious capturing of the scene and cannot adapt to the lighting conditions at different times of the day. Recent works attempt to conduct the extrapolation with generative neural models [16, 30, 32, 40, 52], which shows plausible results for natural landscapes (e.g., mountain valley, beach) or daily scenes (e.g., cars, corridors), but might not look reasonable for tourist attractions with specific shapes and appearances (see Sec. 4.3). Besides, all these image-based photo extrapolation methods are not designed to support rendering novel views with given camera trajectories, which hinders downstream applications like the extrapolated 3D photo generation.

3 Method

We propose a novel factorized neural rendering framework that learns to encode outdoor scenes from Internet photo collection, which enables controllable scene re-rendering with user-desired lighting condition and photo extrapolation or extrapolated 3D photo generation that extends a narrow-view image to a broader field of view. We show an overview of our method in Fig. 2. Unlike previous neural implicit methods [6, 23, 25, 39] that encode all the appearance variations (e.g., lighting condition, auto exposure, white balancing and filtering effects, etc.) in one latent space, we present the first attempt to model outdoor scenes with a more controllable and explainable re-rendering pipeline (Sec. 3.1). To survive from the training with noisy Internet photos, we utilize a composited training scheme (Sec. 3.2), which learns scene geometry with transient removal strategy from Martin-Brualla et al. [23], and supervises re-rendering with distilled occlusion-free images. Moreover, we apply a novel realism augmentation technique that propagates appearance details from tourist photos to the rendered views (Sec. 3.3), which efficiently improves the photo-realism of the rendering results. Please refer to our supplementary material for more technical background.

3.1 Factorized Outdoor Scene Re-Rendering

Factorized rendering formulation. We first introduce our factorized scene re-rendering pipeline, as shown in the middle part of Fig. 2. To allow controllable re-rendering of neural implicit model with user-specific external lighting, previous methods generally require empirical normal regularization [37, 57] or post-processing [17] to obtain a smooth surface normal, which inevitably hurts geometry details. In contrast, we select SDF functions as the representation of scene geometry [45], since it offers exact surface and well-defined normal (by computing gradient w.r.t. query point) to facilitate the explicit relighting process. Specifically, we represent scene geometry and basic color with a geometry MLP and base appearance MLP, and use explicit external lighting (i.e., 16 × 32 HDR map from HDR decoder and affine tone mapping), learnable shadow MLP and a stand-alone sky generator to re-render the scene with appearance variations. The rendering of pixel $C(r)$ with point samples $\{x_k\}$ along the ray $r$ is defined as follows:

$$C(r) = \sum_k T_k a_k \Gamma(c_k, I_t) + (1 - \sum_k T_k a_k) e^v(v, L_e),$$

$$T_k = \prod_{j=1}^{k-1} (1 - \alpha_j), \quad \alpha_j = \max \left( \Phi_x(SDF_j) - \Phi_x(SDF_{j+1}), 0 \right),$$

where $e^b$ is the relit scene color (introduced later), $e^v$ is the generated sky color along the ray direction $v$ and conditioned by environment code $L_e$, $\Gamma(\cdot)$ is the tone mapping conditioned by tone code $I_t$, $T$ is accumulated transmittance, $\Phi_x$ is the cumulative distribution of logistic distribution, and $\alpha$ is opacity derived from adjacent SDF.
More specifically, we define the relit scene color $c^b$ as the following:

$$c^b = \sum_{\text{mp}} a(s) L_v(x, \omega_\text{e}, l_e)(\omega_\text{e} \cdot n) / \Delta \omega_\text{e},$$

where $a$ is the basic color from the base appearance MLP, $s$ is the spatial varying shadow value from the shadow MLP conditioned by shadow code $l_s$, $\omega_\text{e}$ indicate incoming light direction, $n$ is the surface normal derived from the gradient w.r.t. query point $x$. $L_v$ is the incoming HDR lighting along $\omega_\text{e}$ and conditioned by environment code $l_e$. $\Delta \omega_\text{e}$ is the solid angle to the light sample. In summary, the external lighting condition (or appearance variation) is implicitly encoded as environment code $l_e$, shadow code $l_s$ and tone code $v$. Note that we adopt the Lambertian reflectance assumption as pre-processing.

**Data-driven HDR decoder.** Though it is technically sound to re-render outdoor scenes with HDR maps, the problem of disentangling external environment lighting from photo collections is highly ill-posed. The reasons behind it include the unconstrained freedom of HDR and scale ambiguity between base appearance and HDR intensities. For example, one might learn an HDR map that embraces main colors of the scene while leaving a degenerated base appearance, or produces a reasonable relit with a brighter base appearance and a darker HDR map. To this end, we propose to use a data-driven outdoor HDR prior, which constrains the optimization of environment maps in a pre-trained latent space. Practically, we first train a panoramic HDR sky network [13] with Laval sky dataset [13], and pick up the sky decoder as a prior. During the re-rendering stage, we fixed the weight of this decoder, and take as input a per-frame latent environment code $l_e$, and then down-scale the decoder’s output to obtain a $16 \times 32$ environment map. In this way, our model can search for a proper HDR map while avoiding color leaking from buildings into light maps (e.g., similar observations for the indoor scenes [57]).

**Learnable shadow.** Following standard rendering pipeline [1, 35], recent neural implicit rendering and relighting approaches [10, 37, 57] tend to learn a visibility mapping from self-occlusions of density field, which is used to synthesize shadow effect on the rendered views but requires knowing the shape and position of all the object occluders in the scenes. However, for outdoor photo collections, we found that the shadow might come from unobserved buildings behind the capturing positions, which beyond the visibility from self-occlusion of the reconstructed scene geometry and make it much more complicated. Rather than pursuing a physically correct shadow mapping, inspired by 2D inverse rendering [54], we propose to model the shadow effect with a spatial variant shadow MLP. Specifically, the shadow MLP is conditioned with a per-frame latent shadow code $l_s$, and learns a shadow value $s \in (0, 1)$ for each query point along volume rendering rays. To alleviate undesired scaling between base appearance and shadow, we add a shadow regularization during the training stage, which encourages the shadow value close to 1 and is defined as the following:

$$L_{\text{rs}} = \sum_k ||s_k - 1||_{2}^{2}.$$  

In our experiment, we find this shadow modeling successfully simulates shadow effects for cluttered outdoor photo collections while being regardless of issues with unseen shadow casters (Sec. 4.6).

**Affine tone mapping.** Because the output of data-driven HDR decoder is a physically plausible sky lighting that is robust to sensor variation such as white balancing and exposure, we need to take additional tone mapping to handle a large variant of all these color distortions (even including extreme filtering effect by users) in photo collections. In practice, we adopt the strategy from Rematas et al. [29] by learning a $3 \times 4$ affine tone mapping matrix $[A, b]$ (only upper 3 rows) for each frame, where the matrix is the output of a lightweight tone mapper with per-frame latent tone code $l_e$ as input. To avoid color shifting of base appearance due to unconstrained freedom of mapping matrix, we append an affine regularization into the training loss, which encourages the affine tone mapping to be “zero-mean” and is defined as the following:

$$L_{\text{aff}} = e^{T}(R \odot Re) + b^{T}b, \text{ where } R = A - I.$$  

where $\odot$ is the Hadamard product, $e$ is the $3 \times 1$ column vector whose entries are all 1.

**Neural sky generator.** Since we build up the scene geometry with an SDF-based model, it is not applicable to model background sky along with the buildings [23, 29, 45]. Motivated by Rematas et al. and Hao et al. [11, 29], we use a neural sky generator to simulate the sky dome of the scene, which directly maps the viewing direction $r$ to a 3-channel sky color with the condition of environment code $l_e$. As demonstrated in Eq. (1), sky colors are blended according to the
remaining transmittance, so we can jointly train the sky generator at the re-rendering stage.

### 3.2 Composited Training with Photo Collections

Compared to previous methods, our scene representation is much harder to train due to the ill-posed nature of the factorization and noisy Internet data collections. Therefore, we develop a composited training, which learns scene geometry and re-rendering in a two-staged fashion.

**Learning scene geometry from Internet photos.** In the first stage, we learn scene geometry from photo collections with a geometry MLP and radiance MLP following Wang et al. [45]. To handle occasional object occlusions and appearance variations, we adopt the appearance embedding and transient MLP from NeRF-W [23] (only for this stage). Unlike radiance field methods [23, 39] that can render the sky with scattered far sampling, the SDF-based method is inclined to learn an exact surface, which results in a sky dome stitching to the building that is not desired for external relighting. So, we additionally apply a sky segmentation loss to encourage the sky area to be empty, which is defined as:

$$ L_m = BCELoss \left( \sum_i T_i \alpha_k \cdot M_{sky} \right), $$

where the sky mask $M_{sky}$ is annotated with Mask-RCNN [12]. Now, we define the training loss of the scene geometry as:

$$ L_{gen} = \lambda_c L_c + \lambda_m L_m + \lambda_{re} L_{re}. $$

where $L_c$ is the photometric loss with transient modeling following NeRF-W [23], $L_{re} = || \nabla_q \text{SDF}(x) || - 1||^2$ is the Eikonal loss as suggested by Gropp et al. [9]. We set $\lambda_c = 1.0$, $\lambda_m = 0.1$ and $\lambda_{re} = 0.1$. Note that we omit the form of pixels summation in this section for brevity. After the first training stage, we only keep the geometry MLP, while discarding radiance and transient MLP.

**Joint optimization of re-rendering.** In the second stage, we learn factorized scene re-rendering with geometry MLP frozen. Instead of training with raw cluttered photos or manually masking out occluders, we take a distilled fashion by exploiting occlusion-free images from NeRF-W’s static branch for a more steady supervision. Our experiment shows that this strategy efficiently eases the learning of factorized re-rendering and improves the rendering quality both quantitatively and qualitatively (Sec. 4.6). The training loss is then defined as the following:

$$ L_{train} = \lambda_{cr} L_{cr} + \lambda_{rs} L_{rs} + \lambda_{it} L_{it}, $$

where $L_{cr}$ is the MSE loss between the re-rendered pixel and the occlusion-free images. $\lambda_{cr}$, $\lambda_{rs}$, and $\lambda_{it}$ are the loss weights for the MSE loss, shadow regularization (see Eq. (3)), and affine tone mapping regularization (see Eq. (4)) respectively. We empirically set $\lambda_{cr} = 2.0$, $\lambda_{rs} = 0.01$ and $\lambda_{it} = 0.1$.

### 3.3 Photo Adaptation & Realism Augmentation Optimization based photo adaptation.** Once the factorized scene representation has been trained, our model can be adapted to re-captured photos with novel lighting conditions, i.e., minimizing photometric error between rendered pixels and the captured photo with latent optimization on shadow code $I_s$, environment code $I_e$ and tone code $I_t$. Besides, when adapting to photos with a large portion of people like tourist selfies in photo extrapolation applications, we empirically mask out these part and only optimize pixels labelled as sky and attraction (e.g., buildings and sculptures).

**Realism augmentation.** Even though the optimization-based photo adaptation can achieve a rendering result close to the real photo, the detail of the synthetic view is still less realistic. For example, some live details such as water splashes of the fountain and shapes of clouds are missing, which is mainly due to the fact that the neural implicit rendering tends to average texture details from multi-view observations. Fortunately, for tasks like photo extrapolation, there is still an opportunity to enhance rendering details if we can fully exploit information from the given photos. To this end, we design a novel realism augmentation strategy, which significantly propagates texture details from a narrow-view real photo to a wide-view rendered image. We show the pipeline of this strategy in Fig. 3, which illustrates the on-the-fly learning and inference fashion of this augmentation process. Specifically, we adopt an encoder-decoder based network structure from a super-resolution work LiFF [7] as the realism augmentation network since it uses an implicit representation of the image and is flexible to support arbitrary scale and aspect ratio. At the learning stage, we set the network input as a downscale rendered image, and fine-tune the network with the target of aligned and masked (without occluders such as tourists) real photo, so the network learns to compensate details from a blurry neural rendering to the real one. Then, at the propagating stage, we fix the network and simply forward the network with the complete rendered image, where the learned “detail-compensating” knowledge would be propagated to the full view of the rendering result.

### 4 Experiments

In this section, we first evaluate the outdoor scene re-rendering quality of our method (Sec. 4.2), and then conduct photo extrapolation (Sec. 4.3), controllable scene re-rendering (Sec. 4.4) and extrapolated 3D photo generation (Sec. 4.5) on several outdoor attractions. At last, we perform ablation studies to analyse the effectiveness of the training strategy and factorized re-rendering components (Sec. 4.6).
4.1 Datasets
Following the previous work [23], we use Internet photo collections of outdoor attractions from the Phototourism (IMC-PT) 2020 dataset [14, 36], where the image poses are recovered by COLMAP [33]. Specifically, we select 5 famous tourist attractions, including Trevi Fountain, Sacre Coeur, Westminster Abbey, Pantheon and Notre Dame. For Trevi Fountain and Sacre Coeur, we follow the split of NeRF-W for training and testing. For the other three scenes, we also take a similar pre-processing by discarding training images with a large portion of object occlusions, and only select occlusion-free images for metric evaluation.

4.2 Comparison of Scene Re-Rendering Quality
We first compare the scene re-rendering quality with the evaluation protocol from NeRF-W [23], i.e., giving a left half image for optimization, the neural network is asked to render the full view of the image. The metrics of PSNR, SSIM and LPIPS [56] are used to measure the rendering quality. Specifically, we adopt the baseline method NeRF-W [23] and the SOTA image extrapolation method PixelSynth [30] for comparison. The other relightable neural implicit rendering methods (e.g., NeRV [37], NeRD [3], NerFactor [57], and etc.) are not applicable here, because they are not feasible for unbounded outdoor scenes or learning from cluttered photo collections. Note that since NeRF-W has not released the official source code, we adopt an alternative implementation 1 in our experiment, thus the reported result is different from [23]. For PixelSynth, we overfit the network to each individual scene with occlusion-free training images as introduced in Sec. 3.2. We report the quantitative results in Tab. 1 and present the quantitative visualization in Fig. 4.

It is obvious that even though we overfit each PixelSynth model to a specific scene, the complete rendering view is still far from satisfactory (e.g., the geometry structure differs a lot to the ground-truth), which proves that the GAN network of the PixelSynth does not ensure a consistent rendering output. NeRF-W achieves much better results for scene re-rendering, but some live details (e.g., water splash in Trevi Fountain and the sky clouds in Notre Dame) are still missing due to the smooth nature of neural implicit field [57]. Generally, a disentangled rendering pipeline is usually more challenging to render high-quality images [37, 57]. Thanks to the factorized rendering pipeline and realism enhancement in our method, we still achieve on-par or even better rendering quality both quantitatively and qualitatively while successfully maintaining live details close to

---

Table 1: We compare the scene re-rendering quality with PixelSynth [30] and NeRF-W [23] on five outdoor scenes of the Internet photo collections [14, 36]. Note that we use an alternative implementation of NeRF-W. See the text for details.

| Methods        | Trevi Fountain | Sacre Coeur | Pantheon Exterior | Westminster Abbey | Notre Dame Front Facade |
|----------------|----------------|-------------|-------------------|-------------------|-------------------------|
|                | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| PixelSynth [30]| 14.73 | 0.587 | 0.693 | 14.19 | 0.659 | 0.566 | 12.09 | 0.603 | 0.607 | 14.00 | 0.664 | 0.618 |
| NeRF-W * [23] | 21.31 | 0.764 | 0.380 | 21.23 | 0.850 | 0.283 | 24.78 | 0.875 | 0.265 | 21.35 | 0.801 | 0.364 |
| Ours           | 23.09 | 0.792 | 0.345 | 21.51 | 0.849 | 0.162 | 24.46 | 0.867 | 0.237 | 24.66 | 0.854 | 0.238 |

1https://github.com/kwea123/nerf_pl/tree/nerfw
Figure 5: We compare photo extrapolation with Auto-Stitch [5], PixelSynth [5] and NeRF-W [5] on four outdoor scenes [14, 36]. Photos by Flickr users Hugão Cota, Legalv1, Foster’s Lightroom, and stobor.

Figure 6: We exhibit the results of controllable re-rendering by changing tone mapping and HDR environment maps.

Figure 7: We show two examples of extrapolated 3D photo generation, which transfers tourist photos into extrapolated and dynamic 3D photos with camera moving effect. Please use Adobe Reader or check our project webpage to see animations. Photos by Flickr users MikiAnn and Chris Devers.

4.3 Comparison of Photo Extrapolation

We now conduct the comparison on the photo extrapolation task in Fig. 5. In this part, we also perform an extrapolation test on a 2D image-based method [5], which is denoted as Auto-Stitch. Specifically, we first retrieve ~30 nearby captured position from photo collections based on the SfM camera poses and perform panoramic stitching with OpenPano” [5]. Then, we warp the stitched image to the given photo view and blend the front tourists into the image. It is clear that the extrapolated views of this pipeline are full of stitching artifacts such as human shadows, and there are some vacancy parts near the border of extended images due to the lack of observations near the captured position, which indicates that the 2D image-based approach is not suitable for photo extrapolation with Internet photo collections, as they might need a carefully collected data library to achieve a clean result [28, 43]. For the photo extrapolation results of NeRF-W and PixelSynth, similar to what we have analyzed in Sec. 4.2, they are faced with the issues of the lack of live texture details and distorted 3D structure and color (e.g., in Pantheon of Fig. 5, due to the generalizability issue of appearance embedding, NeRF-W correctly simulates the appearance of the building, but the blue sky color is severely distorted), which inevitably degrades the user experience of this functionality. Thanks to the factorized scene representation and realism augmentation, our results show better photo-realism with vivid details.

4.4 Controllable Scene Re-Rendering

We show our controllable scene re-rendering capability with user-selected tone mapping and HDR environment maps in Fig. 6. Note that we cannot find proper competitors for this task, since existing methods either only support appearance changing through the latent space while lack of explicit controlling of lighting effect [6, 23, 25, 39], or only support inverse rendering and relighting for 2D images but cannot synthesize novel lighting effect with given camera trajectories [54, 55]. As shown in Fig. 6, with the factorized scene representation, we can freely control the lighting effect through tone mapping and even user-selected HDR maps, e.g., the appearance of the re-rendered building naturally exhibits the lighting effect with a dusk and cool tone from the given HDR maps in Fig. 6 (c) and (d).
4.6 Ablation Studies

We first inspect the effectiveness of the occlusion-free supervision in the re-rendering stage during composited training (Sec. 3.2). As shown in Fig. 8, when training with raw images that contains occluders, we might end up with a neural model that brings some shadows at the lower part of the rendered image (Fig. 8 (a)), while the rendered result (Fig. 8 (b)) from the full model is free of such artifacts. This proves the necessity of occlusion-free supervision for learning factorized re-rendering.

Affine tone mapping. We then analyze the impact of affine tone mapping in the factorized re-rendering. As shown in Fig. 9 and Tab. 2, the rendered scene without tone mapping shows pale lighting effects compared to the ground-truth image, which demonstrates that relying on the HDR decoder alone cannot guarantee faithful modeling of various lighting effects. By introducing affine tone mapping, we mitigate the pressure of the HDR decoder, and achieve better photo adaptation ability (e.g., Fig. 9 (c) shows the scene lighted by yellow sunlight as the ground-truth while Fig. 9 (a) fails).

Shadow modeling. We also study the effectiveness of shadow modeling in Fig. 10 and Tab. 2. It is noticeable that when introducing shadow modeling, our method can better simulate shadows in the outdoor scene (e.g., dimming appearance at the highlighted area in Fig. 10), even if the shadow caster (e.g., the shadow caused by some building behind the fountain, see highlighted green rectangle in Fig. 10) is not observed before. Meanwhile, when equipped with shadow modeling, the metric of the rendering quality is also improved (see the second and the last row of Tab. 2).

Realism augmentation. We finally inspect the efficacy of the realism augmentation mechanism. As shown in Fig. 9 and Fig. 3, the texture details of buildings and scenes (e.g., springs of fountain and curves of sculptures) have been enhanced with rich patterns, and the lighting effect is also much closer to the ground-truth image. Besides, the metric result is also improved by a large margin as show in Tab. 2. These results demonstrate the value of realism augmentation in broader applications such as photo adaptation and extrapolation where a single neural model is required to adapt to different illumination conditions and dynamic scene details.

5 Conclusion

We propose a novel factorized neural re-rendering model, which encodes the appearance and geometry of outdoor scenes from Internet photo collections in a factorized paradigm, and delivers controllable scene re-rendering, photo extrapolation and even extrapolated 3D photo generation. One limitation is that we take the Lambertian reflectance assumption for modeling, which is not capable of representing shiny and mirrored materials such as glass walls of buildings. Second, our method is agnostic to unobserved shadow casters (e.g., building behind the tourists), so the shadow effect is not controllable through external lighting. Third, we do not extrapolate photos for uncaptured part of human bodies. A possible workaround is to adopt portrait image completion techniques [47] to complete the bodies, which can be directly incorporated with our pipeline in future works.

Acknowledgement. This work was partially supported by the NSFC (No. 62102356) and Zhejiang Lab (2021PETOAC01).
[50] Bangbang Yang, Yinda Zhang, Yijin Li, Zhaopeng Cui, Sean Fanello, Hujun Bao, and Guofeng Zhang. 2022. Neural Rendering in a Room: Amodal 3D Understanding and Free-Viewpoint Rendering for the Closed Scene Composed of Pre-Captured Objects. ACM Trans. Graph. 41, 4, Article 101 (July 2022), 10 pages. https://doi.org/10.1145/3528223.3530163

[55] Ye Yu and William Alfred Peter Smith. 2021. Outdoor Inverse Rendering from a Single Image Using Multiview Self-Supervision. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In CVPR.

[57] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. 2021. NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination. ACM Transactions on Graphics (TOG) 40, 6 (2021), 1–18.

[58] Yongjie Zhu, Yinda Zhang, Si Li, and Boxin Shi. 2021. Spatially-Varying Outdoor Lighting Estimation from Intrinsics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 12834–12842.