UrbanIR: Large-Scale Urban Scene Inverse Rendering from a Single Video

Chih-Hao Lin[1] Bohan Liu[1] Yi-Ting Chen[2] Kuan-Sheng Chen[1]
David Forsyth[1] Jia-Bin Huang[2] Anand Bhattad[1] Shenlong Wang[1]
[1]University of Illinois Urbana-Champaign [2]University of Maryland, College Park
https://urbaninverserendering.github.io/

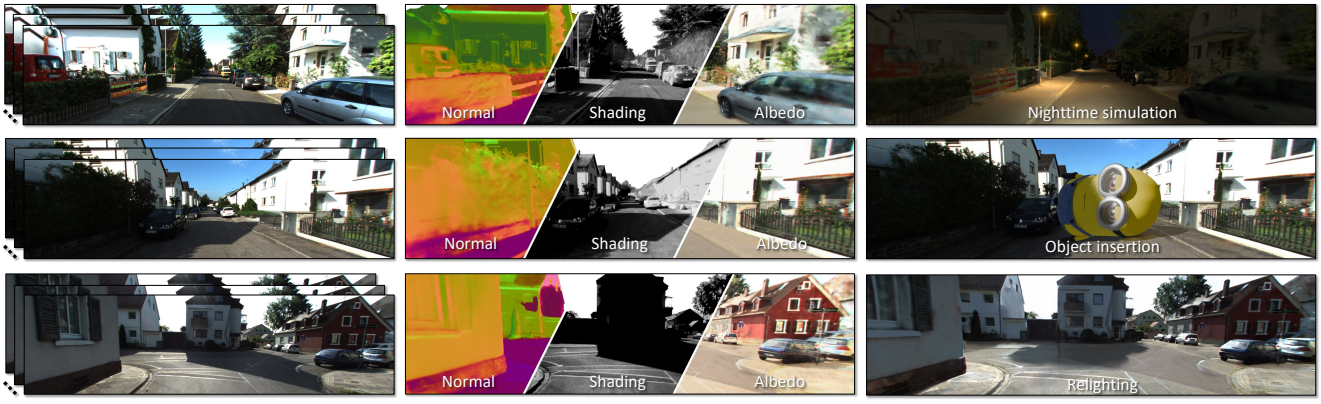

Figure 1. We present *UrbanIR* (**Urban** Scene **I**nverse **R**endering), a realistic and relightable neural scene model. UrbanIR infers accurate scene properties from *a single video* of large-scale, unbounded scenes and delivers realistic relighting, night simulation, and object insertion.

Abstract

*We present UrbanIR (***Urban** *Scene* **I***nverse* **R***endering), a new inverse graphics model that enables realistic, free-viewpoint renderings of scenes under various lighting conditions with a single video. It accurately infers shape, albedo, visibility, and sun and sky illumination from wide-baseline videos, such as those from car-mounted cameras, differing from NeRF's dense view settings. In this context, standard methods often yield subpar geometry and material estimates, such as inaccurate roof representations and numerous 'floaters'. UrbanIR addresses these issues with novel losses that reduce errors in inverse graphics inference and rendering artifacts. Its techniques allow for precise shadow volume estimation in the original scene. The model's outputs support controllable editing, enabling photorealistic free-viewpoint renderings of night simulations, relit scenes, and inserted objects, marking a significant improvement over existing state-of-the-art methods. Our code and data will be made publicly available upon acceptance.*

1. Introduction

We show how to build a model that allows realistic, free-viewpoint renderings of a scene under novel lighting conditions from a video. So, for example, a sunny afternoon video of a large urban scene can be shown at different times of day or night (as in Fig. 1), viewed from novel viewpoints, and shown with inserted objects. Our method — *UrbanIR* (**Urban** Scene **I**nverse **R**endering) — computes an inverse graphics representation from the video. UrbanIR jointly infers shape, albedo, visibility, and sun and sky illumination *from a single video of unbounded outdoor scenes* with *unknown lighting*. The resulting representations enable controllable editing, delivering photorealistic free-viewpoint renderings of relit scenes and inserted objects, as demonstrated in Fig. 1.

UrbanIR obtains its intrinsic scene representations from a video under a *single illumination condition*, but producing realistic novel views requires accurate inferences of physical parameters. UrbanIR uses a novel visibility rendering scheme and loss to precisely estimate shadow volumes in the original scene and control albedo errors. UrbanIR combines monocular intrinsic decomposition and inverse rendering with other key contributions to control errors in renderings. To our knowledge, UrbanIR is the first in its class capable of performing inverse rendering and relighting applications from a single monocular video in large-scale scenes, without requiring multiple illumination, depth sensing, or both.

UrbanIR representations are constructed from cameras mounted on cars with a narrow range of views of each scene point. Typical NeRF-style systems yield poor geometry estimates (for example, roofs) and "floaters" under these

conditions; they are usually trained with a wide range of views. Our experiments showcase that UrbanIR outperforms these baselines with significantly reduced artifacts in our sparse view setting. Finally, we show how to use UrbanIR to simulate night scenes from a single daytime-captured video, producing a controllable, realistic, physically plausible, and consistent simulation. In summary, our contributions are:

- We present UrbanIR for recovering a *relightable* neural radiance field in a constrained setting of an *unbounded scene*, using a *single monocular video* captured under a *single illumination condition*.
- We describe a novel inverse rendering framework that *builds precise shadow volumes* in large outdoor scenes with heavy shadows, resulting in significant improvements in inverse graphics estimates and relighting.
- We demonstrate a new physics-informed night simulation framework. To our knowledge, UrbanIR is the first simulation to offer realistic, *free-viewpoint night simulation* from a single daytime video capture.

2. Related Work

Inverse Graphics involves inferring illumination and intrinsic properties of a scene. The problem is underconstrained, and there is much reliance on priors [2, 3, 25, 26, 35, 48, 62, 76, 81] or on managed lighting conditions [2, 2, 19, 24, 24, 80], known geometry [16, 32, 36, 61], or material simplifications [47, 81, 86]. Recent methods use deep learning techniques to reason about material properties [44–46, 53, 75, 84]. Models trained on synthetic data [43] or pair-wise annotated data [4] have shown promising results. Learned predictors of albedo or shading are described and reviewed in [6, 18, 63]. Neural representations of material or illumination appear in [5, 38–41, 46]. Like these methods, we exploit monocular cues, such as shadows and surface normals. In contrast, we combine learning-based monocular cues and model-based relightable NeRF optimization to infer the scene's intrinsic properties and illumination.

Shadow modeling using images is challenging. Methods trained to cast shadows from images [69, 82] are tailored for particular objects (pedestrians, cars, etc). Learned methods can detect and remove shadows from 2D images [20, 21, 68]. But inverse graphics require modeling the full 3D geometry, intrinsic scene properties, and ensuring temporal consistency. Model-based optimization methods can infer shadows but rely on accurate scene geometry [33, 65, 73]. Using visibility fields to model shadows results in difficulty providing consistent shadows in relation to the underlying geometry [60, 64, 74, 85]. In contrast, our method combines the strengths of learning-based monocular shadow prediction and removal and model-based inverse graphics.

Method	Scene	Illumination Conditions	RGB Only	Explicit shadow	Night Sim.
NeRFFactor [83]	Object	Multi	Yes		
TensoIR [27]	Object	Single	Yes	✓	
InvRender [85]	Object	Single	Yes		
NeRF-OSR [60]	Front-Facing	Multi	Yes		
FEGR [71]	Large Scene	Single/Multi	LiDAR	✓	
LightSim [55]	Large Scene	Single/Multi	LiDAR	✓	
UrbanIR (Ours)	Large Scene	Single	Yes	✓	✓

Table 1. Comparison of various recent relightable NeRF methods. UrbanIR is among the first to offer single-illumination and RGB-only relightable NeRF capabilities suitable for large-scale scenes.

Relightable Neural Fields: Relightable neural radiance field methods [8, 9, 23, 52, 72, 75, 79, 83] aim to factor the neural field into multiple intrinsic components and leverage neural shading equations for illumination and material modeling. These methods admit realistic and controllable rendering of scenes with varying lighting conditions and materials. However, most relightable NeRF methods focus on objects with surrounding views or small bounded indoor environments. Important exceptions are: NeRF-OSR [60], which assumes access to multiple lighting sources for decomposition, and FEGR [71], which either uses multiple lighting or exploits depth sensing, such as LiDAR.

We compare the problem setting, input requirement with recent methods in Tab. 1. UrbanIR addresses inverse rendering for large-scale urban scenes that object-centric methods [27, 83, 85] fails to reconstruct. Furthermore, our method takes videos under single illuminations as input, which is more applicable to a broader range of scenes. To estimate the geometry of large-scale driving scenes, FEGR [71] and LightSim [55] rely on captures from five to six cameras and LiDAR sensors. On the other hand, UrbanIR only needs videos from single or stereo cameras without any guidance from LiDAR. Our method also performs nighttime simulation by inserting local light sources (e.g. streetlight, vehicle light), which is not demonstrated in previous works.

3. Method

UrbanIR takes a multi-frame video of a scene under single illumination; as the camera moves, its motion is known. Write $\{I_i, E_i, K_i\}$, where $I_i \in \mathbb{R}^{H\times W\times 3}$ is the RGB image; $E_i \in \mathrm{SE}(3)$ is the camera pose; and K_i is camera intrinsic matrix. We produce a neural field model that can be viewed from *novel camera viewpoints* under *novel lighting conditions*. We do so by constructing a neural scene model that encodes albedo, normal, semantics, and visibility in a unified manner (Sec. 3.1). This model is rendered from a given camera pose with given illumination using an end-to-end differentiable volume renderer (Sec. 3.2). Our inference is by joint optimization of all properties (Sec. 3.3). Applications include changing the sun angle (Fig. 1; top right), day-to-night transitions (Fig. 1; bottom right), and object insertion

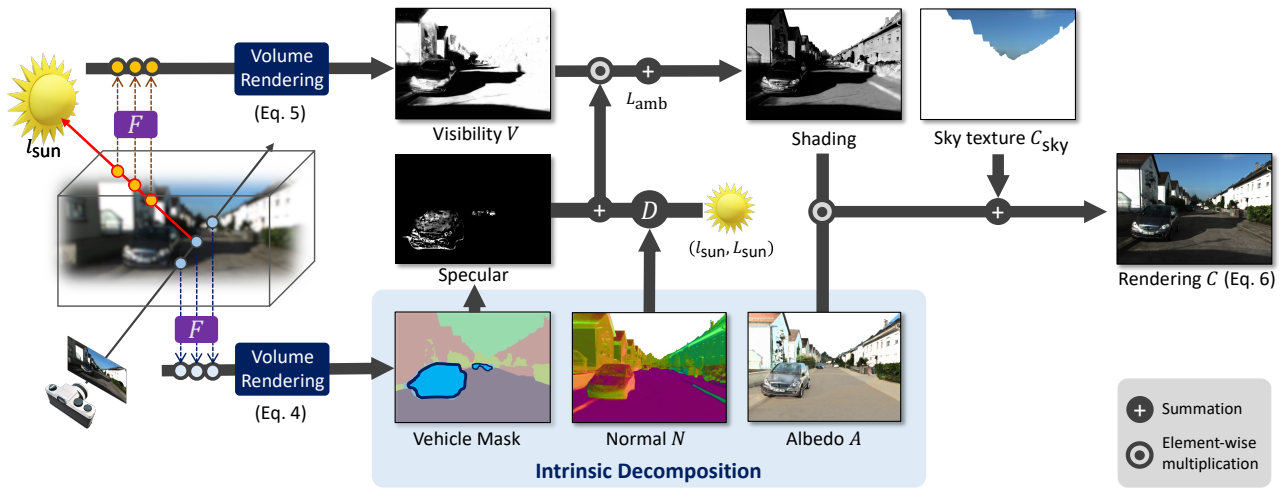

Figure 2. **Rendering Pipeline.** UrbanIR retrieves scene intrinsics (normal N, semantics S, albedo A) from camera rays, and estimate visibility V from tracing rays to the light source. The shading model computes diffuse and specular reflection and adds ambient sky light $\mathbf{L}_{\text{sky}}$ for the final shading map. We multiply shading & albedo, and render the sky appearance for final rendering. (Eq. 3.2 for more details.)

(Fig. 1; middle right). More details about applications are in Sec. 3.4. Fig. 2 provides an overview of our proposed inverse graphics and simulation framework.

3.1. Relightable Neural Scene Model

The scene representation is built on Instant-NGP [51, 57], a spatial hash-based voxel NeRF representation. Instant-NGP offers numerous advantages, including low memory consumption; high efficiency in training and rendering; and compatibility with expansive outdoor scenes. Write $\mathbf{x} \in \mathbb{R}^3$ for position in 3D, $\mathbf{d}$ for query ray direction, θ for learnable scene parameters; NeRF models, including Instant-NGP, learn a radiance field $F_\theta(\mathbf{x}, \mathbf{d}) = (\mathbf{c}, \sigma)$, where $\mathbf{c} \in \mathbb{R}^3$ and $\sigma \in \mathbb{R}$ represent observed color and opacity respectively. Standard NeRFs have view- and lighting-dependent effects, such as shading, shadow, and specularity, baked into their observed color, making them non-relightable.

In contrast, UrbanIR learns a model of the intrinsic scene attributes field independent of viewing angles and lighting conditions. Write diffuse albedo $\mathbf{a}$, surface normal $\mathbf{n}$, semantic vector $\mathbf{s}$, and density σ; then UrbanIR learns:

$$F_\theta(\mathbf{x}) = (\mathbf{a}, \mathbf{n}, \mathbf{s}, \sigma) \tag{1}$$

where θ is learnable parameters. The diffuse albedo represents the intrinsic color and texture of the material; the normal represents the intrinsic surface geometry; density encodes the spatial opacity, and semantics is used as a key to query surface reflectance. Following Instant-NGP [51], we learn a dense feature hash table to represent the scene, and an individual MLP header is used to decode each attribute given a queried feature at point $\mathbf{x}$. We provide the details of the architecture in the supplementary. The geometry of the scene is implicitly encoded in σ. In contrast to existing relightable outdoor scene models that demand coupled explicit geometry [60, 71], our scene model is implicit, providing compactness and consistency to appearance modeling.

The lighting model is a parametric sun-sky model [34, 78]. This encodes outdoor illumination as:

$$\mathbf{L} = \{(\mathbf{L}_{\text{sun}}, \psi_{\text{sun}}, \phi_{\text{sun}}), \mathbf{L}_{\text{amb}}, \mathbf{L}_{\text{sky}}\}. \tag{2}$$

Our sun model is a 5-DoF representation, encoding sun color $\mathbf{L}_{\text{sun}}$ along with the azimuth and zenith $\psi_{\text{sun}}, \phi_{\text{sun}}$. The $\mathbf{L}_{\text{amb}}$ model is represented as a 3-DoF ambient light. The sky dome model infers the sky texture from the viewing direction: $\mathbf{C}_{\text{sky}} = \mathbf{L}_{\text{sky}}(\mathbf{r})$. We chose this minimalist sun-sky model as it is more compact than other alternatives (e.g., HDR dome or Spherical Gaussians) yet has proven highly effective in modeling various outdoor illumination effects [34, 78].

3.2. Rendering

Given the scene model F_θ and a lighting model $\mathbf{L}$, rendering involves two steps: 1) volume rendering of the scene's intrinsic properties and visibility map onto the image plane, and 2) a shading process to produce the final result with view-dependent and lighting-dependent effects:

$$\mathbf{C} = \texttt{Shade}(\texttt{Intrinsic}(F_\theta, \mathbf{r}), \texttt{Shadow}(F_\theta, \mathbf{r}, \mathbf{L}), \mathbf{L}) \tag{3}$$

where $\mathbf{L}$ is the lighting model, $\mathbf{C}$ is the final RGB color.

Intrinsics images are obtained by volume rendering. We accumulate predictions from $F(\cdot;\theta)$ along the query ray. Multiple points are sampled along the ray, and intrinsics at the query pixel along the ray [28, 49]. In particular, the albedo $\mathbf{A}$, normal $\mathbf{N}$, and semantics $\mathbf{S}$ are predicted as:

$$\mathbf{A}(\mathbf{r}) = \sum_{i=1}^{N} w_i \mathbf{a}_i, \mathbf{N}(\mathbf{r}) = \sum_{i=1}^{N} w_i \mathbf{n}_i, \mathbf{S}(\mathbf{r}) = \sum_{i=1}^{N} w_i \mathbf{s}_i, \tag{4}$$

where $w_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)\,(1 - \exp(-\sigma_i \delta_i))$ is alpha-composition weight, $\delta_i = t_i - t_{i-1}$. We perform rendering for each camera ray and get the final semantic map, albedo map, and the normal map.

Shadow modeling and rendering are essential for obtaining realistic-looking outdoor images. Modeling the visibility of the sun with a per-scene optimized MLP head (as in [83, 85]) is impractical because we need to change the sun's position in relighting but can learn from only one position. An alternative is to construct an explicit geometry model to cast shadows [71], but this model might not be consistent with the other neural fields, and imposing consistency is difficult. Instead, we first compute an estimate $\mathbf{x}(\mathbf{r})$ of the 3D point being shaded, then estimate the visibility $V(\mathbf{x}, \text{sun})$. Our key insight is that shadows in outdoor scenes are primarily due to the visibility of a single directional sunlight.

We obtain $\mathbf{x}(\mathbf{r})$ for each ray by volume rendering depth (so substitute $\hat{t} = \sum w_i t_i$ into the equation for the ray being rendered). Now, to check whether $\mathbf{x}$ is visible to the light source, we compute the transmittance along the ray segment between $\mathbf{x}$ and the light source using volume rendering:

$$V(\mathbf{x}, \text{sun}) = \exp\left(-\sum_i \sigma_i(\mathbf{x}_i)\delta_i\right) \text{ where } \mathbf{x}_i = \mathbf{x} + t_i \mathbf{l}_{\text{sun}} \tag{5}$$

Lower transmittance along a ray from a surface point to a light source suggests fewer obstacles between the point and the light source. Eq. 5 establishes a strong link between transmittance, lighting, and visibility fields used in training. In particular, a point in a training image known as shadowed (resp. out of shadow) should have large (resp. small) accumulated transmittance. We use this constraint to adjust distant geometry during training. Compared to other alternatives [71, 85], our proposed visibility test is simple to compute, flexible for relighting, and aligns with intrinsic properties with a few mild assumptions for outdoor scenes.

Shading is performed by a Blinn-Phong model [7] that incorporates sun and sky terms for the foreground scene and an MLP query for the background sky. For $\mathbf{S}(\mathbf{r}) \in$ sky, we use $\mathbf{C}(\mathbf{r}) = \mathbf{L}_{\text{sky}}(\mathbf{r})$ and otherwise, we use

$$\mathbf{C}(\mathbf{r}) = \mathbf{A}(\mathbf{r})\,(\mathbf{L}_{\text{sun}}\mathbf{D}\mathbf{V} + \mathbf{L}_{\text{amb}}) \tag{6}$$

where $\mathbf{D} = \max(\mathbf{N}(\mathbf{r}) \cdot \mathbf{l}_{\text{sun}}, 0)$ is the diffuse lighting at the surface, $\mathbf{l}_{\text{sun}}$ is the sunlight direction (derived from $\psi_{\text{sun}}, \phi_{\text{sun}}$). The visibility $V(\mathbf{x}, \text{sun})$ is 1 if $\mathbf{x}(\mathbf{r})$ can see the sun and 0 otherwise. This shading model is capable of producing a realistic appearance with shadows following varying lighting conditions. The model can readily be extended with additional lighting sources at the relighting stage, as later shown in the night simulation.

3.3. Inverse graphics

We train scene $F(\cdot)$ (Eq. 1) and lighting $\mathbf{L}$ (Eq. 2) models jointly using a loss:

$$\min_{\theta, \mathbf{L}} \mathcal{L}_{\text{render}} + \lambda_1 \mathcal{L}_{\text{visibility}} + \lambda_2 \mathcal{L}_{\text{normal}} + \lambda_3 \mathcal{L}_{\text{semantics}} + \lambda_4 \mathcal{L}_{\text{reg}}, \tag{7}$$

where individual loss terms are described below.

Rendering loss measures the agreement between observed images and images rendered from the model using the training view and lighting, yielding $\mathcal{L}_{\text{render}} = \sum_{\mathbf{r}} \|\mathbf{C}_{\text{gt}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_2^2$, where $\mathbf{C}$ is rendered color per ray, as defined in Eq. 3, and $\mathbf{C}_{\text{gt}}$ is the observed "ground-truth" color. Minimizing the rendering loss ensures our scene model can reproduce the observed images.

Visibility loss recovers unseen geometry with shadow guidance, improving shadow synthesis for relighting. Specifically, a pixel that is known to be in shadow must be at a point that *cannot see the sun*, so constraining geometry along a ray from that pixel to the sun. This loss could be computed by simply comparing visibility $V(\mathbf{x}, \text{sun})$ with the shadow mask detection [12]. However, the 2D shadow detection is not consistent across different frames, making optimization unstable if visibility is supervised with the masks directly. Therefore, we construct an intermediate "guidance" visibility estimate $\hat{V}(\mathbf{r})$ which is an MLP head trained to reproduce the shadow masks, and compute

$$\mathcal{L}_{\text{visibility}} = \sum_{\mathbf{r} \in \mathcal{R}} \text{CE}\left(M(\mathbf{r}), \hat{V}(\mathbf{r})\right) + \text{CE}\left(V(\mathbf{r}), \hat{V}(\mathbf{r})\right),$$

where $M(\mathbf{r})$ is the shadow mask at pixel $\mathbf{r}$, , and CE$(.,.)$ is a cross-entropy loss. Here, the first term forces the $\hat{V}$ to generate consistent shadow masks, and the second forces V to agree with $\hat{V}$, recovering scene geometry that is not captured in the images but still cast shadows (e.g. top of the buildings).

Normal loss is computed by comparing results N_{gt} from an off-the-shelf normal estimator [17, 29] to the output of the normal MLP. An alternate estimate of the normal follows from the density field: $\hat{N}(\mathbf{r}) = -\frac{\nabla\sigma(\mathbf{x})}{\|\nabla\sigma(\mathbf{x})\|}$. We found that enforcing the consistency between the normal estimation improves the geometry, thus enhances relighting quality significantly. Then our normal loss is given by:

$$\mathcal{L}_{\text{normal}} = \sum_{\mathbf{r} \in \mathcal{R}} \left(\|N_{\text{gt}}(\mathbf{r}) - N(\mathbf{r})\|^2 + \|N(\mathbf{r}) - \hat{N}(\mathbf{r})\|^2\right).$$

We also adopt normal regularization from Ref-NeRF [67] to produce smoother geometry.

Semantic loss is computed by comparing predicted semantics $\mathbf{s}$ with labels in the dataset [42] or detected with [14]. We use an additional loss to encourage high-depth values in

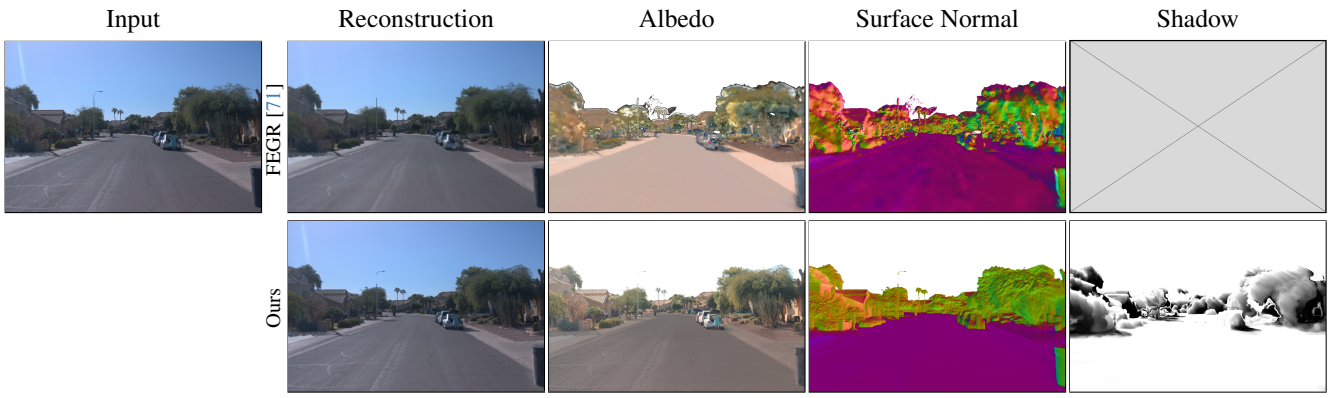

Figure 3. **Intrinsic Decomposition of Waymo Open Dataset [66].** We thank the FEGR authors for sharing the results of their Waymo testing sequence with us for comparison. UrbanIR not only decomposes albedo and shadow better but also produces smoother and more detailed albedo and normal. We recommend readers zoom in to view the difference in the intrinsic images.

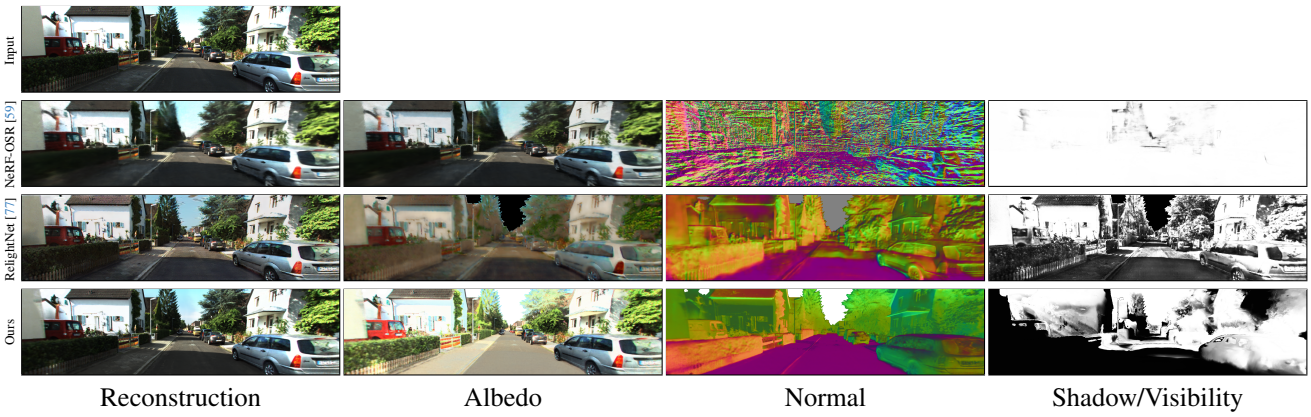

Figure 4. **Intrinsic Decomposition Comparison.** Please note that NeRF-OSR [59] fails to decompose intrinsic, and RelightNet [59] tends to bake shadow in the albedo.

the sky region, reducing floaters in the sky:

$$\mathcal{L}_{\text{semantics}} = \sum_{\mathbf{r}\in\mathcal{R}} \text{CE}\left(S_{\text{gt}}(\mathbf{r}), S(\mathbf{r})\right) - \sum_{\mathbf{r}\in\text{sky}} D(\mathbf{r}).$$

A regularization term is used to regularize the albedo of the scene and ambient light intensity. This is necessary due to the ill-posed nature of our optimization process. However, removing the hard shadow from the sunlight in the albedo field $\mathbf{A}$ remains a challenge, particularly in urban driving sequences. To address this challenge, we introduce a prior that ensures the ground albedo is homogeneous. This is important because the ground region typically shares a similar albedo value. More specifically, we first compute the average ground albedo $\bar{\mathbf{A}}_{\text{g}}$ from albedo $\mathbf{A}$ and semantic S_{gt} and regularize the albedo using $\mathcal{L}_{\text{albedo}} = \sum_{\mathbf{r}\in\text{ground}} \|\mathbf{A}(\mathbf{r}) - \bar{\mathbf{A}}_{\text{g}}\|_2$.

We also calculate an *ambient regularization* term as $\|\mathbf{L}_{\text{amb}}\|_2$. We regularize the intensity of ambient light to avoid unnatural color shifts in the recovered albedo caused by a large intensity of ambient light. Our regularization term is thus $\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{albedo}} + \|\mathbf{L}_{\text{amb}}\|_2$.

3.4. Applications

As the geometry, lighting, albedo, and semantics are recovered, UrbanIR enalbes numerous scene-editing applications, including (1) change sunlight direction and cast the corresponding shadow; (2) turn off sunlight and introduce new light sources (e.g. streetlights) for nighttime simulation; and (3) insert virtual objects and synthesize realistic shading. We encourage the readers to read the supplementary material for implementation details.

4. Experiment Results

4.1. Datasets

We evaluate UrbanIR on two datasets: the KITTI-360 dataset [42] and the Waymo Open Dataset [66]. The KITTI-360 dataset [42] consists of 9 stereo video sequences showcasing urban scenes. For our analysis, we selected 7 non-overlapping clipped sequences, each containing around 100 images. These sequences cover various light directions, vehicle trajectories, and layouts of buildings and vegetation. The

Figure 5. **Shadow Removal in Albedo.** Our method correctly recovers albedo under a shadow while ShadowFormer [21] fails to.

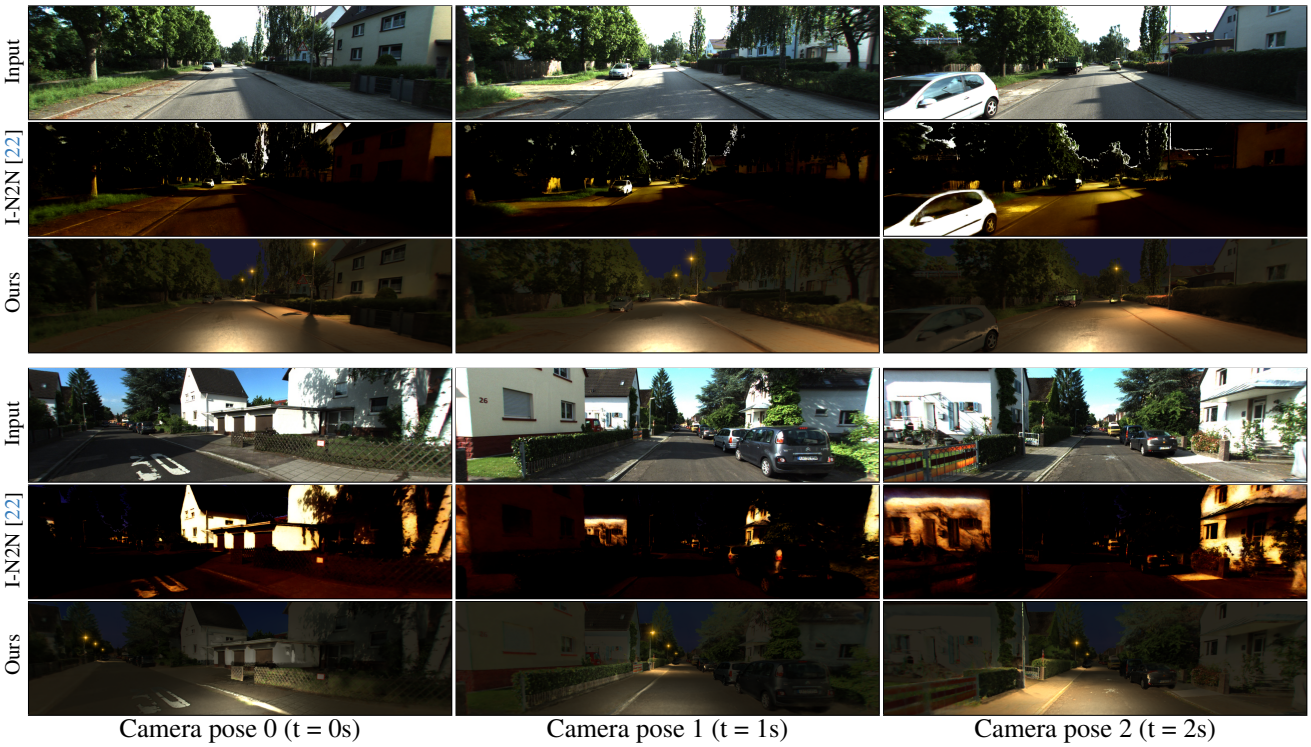

Figure 6. **Nighttime rendering.** The scene changes from daytime to nighttime by introducing new light sources, such as headlights on a car and streetlights. The top three and bottom three rows are from the same driving video but at different times. UrbanIR successfully removes dark shadows with sharp boundaries, resulting in a more realistic rendering of new light sources (such as streetlights and headlights) during night-time simulations. Our method is superior to Instruct-NeRF2NeRF [22], which relies on generative prior.

dataset includes RGB images from stereo cameras, semantic labels, camera poses, and RTK-GPS poses. On the other hand, the Waymo Open Dataset (WOD) [66] captures driving sequences from five cameras and one 64-beam LiDAR sensor at 10 Hz. However, we only used the single camera from the front view and did not use any LiDAR information for our evaluation.

Quantitative evaluation of relighting sequences is difficult as most datasets only capture the same location under a single illumination, and no ground truth for relighting is available. Therefore, we recorded a scene at different times of the day, covering different illuminations. The images were captured by a stereo camera, and the poses were estimated using RTK-GPS information.

4.2. Baselines

We compare UrbanIR with scene relighting and editing methods: FEGR [71], Instruct NeRF2NeRF [22], NeRF-OSR [60], RelightNet [77]. Implementation details are in the supplementary material.

4.3. Decomposition Quality

We evaluate intrinsic decomposition on the Waymo Open Dataset [66] and present the comparison in Fig 3. NeRF-OSR [59] requires multi-illumination as input and fails to decompose albedo and shadow, leaving severe artifacts due to noisy normal estimation. FEGR [71] uses five cameras and LiDAR for reconstruction but still bakes shadow patterns into the albedo and normal. However, UrbanIR only requires a single camera as input without any LiDAR information. Integrating monocular prior in optimization successfully decomposes clean albedo, normal, and shadow maps under single illumination.

We also compare with NeRF-OSR [60] and RelightNet [77] on KITTI-360 [42] in Fig. 4. NeRF-OSR reconstructs a noisy normal map and cannot capture the scene

Figure 7. **Rendering and relighting comparison**. UrbanIR leverages optimization to enable realistic and controllable relighting effects, demonstrating effectiveness in simulating different sunlight directions from a single video input.

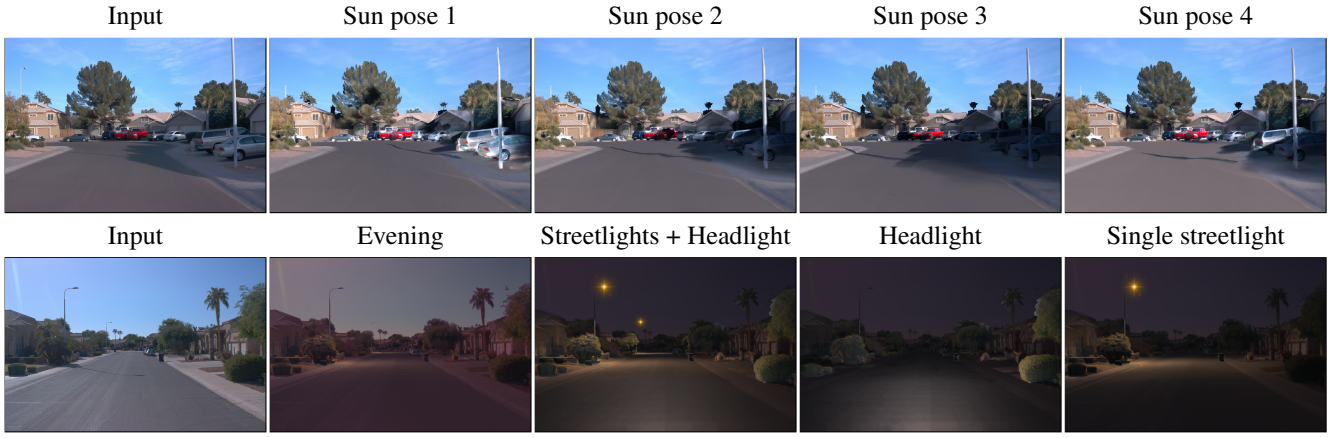

Figure 8. **Controllable Relighting of Waymo Open Dataset [66].** The first row shows different lighting during the day, and the second row changes the input image into night-time with different lighting configurations.

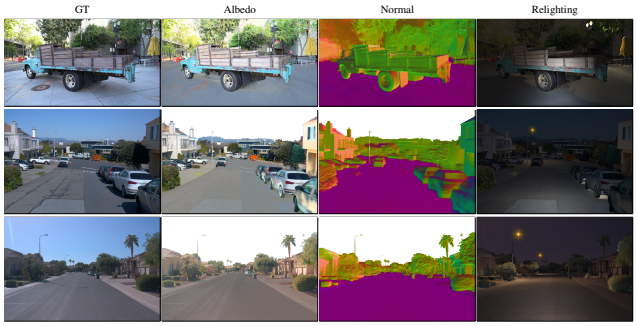

Figure 9. **Decomposition and relighting results of Tanks and Temples [31] and Waymo Open Dataset [66].**

shadows from a single lighting condition, leaving dark shadow patterns in the albedo. RelightNet predicts better normals but still bakes shadows into the albedo. UrbanIR generates clean and sharp albedo and normal fields and also produces a geometry-aware shadow from the input video sequence. In Fig. 5, we compare the learned albedo with the output of shadow removal network [21]. ShadowFormer [21] recovers albedo well on the ground but cannot estimate the correct albedo for the building and vehicles. Our optimization process uses albedo regularization ($\mathcal{L}_{\text{reg}}$). This helps UrbanIR recover a cleaner albedo field on most surfaces.

4.4. Relighting Quality

Relighting under various lighting conditions is evaluated in Fig. 6, 7. NeRF-OSR [60] cannot simulate shadows under novel light conditions. Instruct-NeRF2NeRF [22] leverages generative model [10] to update the training views with text prompt and edits the neural field gradually. While it makes the overall color darker for night simulation, it fails to remove existing shadows and add new light sources.

In contrast, UrbanIR synthesizes sharp shadows and varying surface shading following the sun's direction. Further, the original scene shadows are largely absent. This allows synthesizing images at night (Fig. 6) by inserting car headlights and streetlights, without distracting effects from the original shadows. Moreover, the relighting results obtained from UrbanIR are highly controllable, as demonstrated in Fig. 8. Different light directions and intensities were used to adjust the relighting outcomes. Light sources were also added and turned on and off. UrbanIR not only handles driving sequences but also performs well on multi-view datasets such as Tanks and Temples [31]. In Fig. 9, our method estimates accurate albedo and normal and simulates realistic nighttime images by inserting streetlights into the scene, showing that our method can generalize to diverse scenes and camera trajectories.

4.5. Quantitative Evaluation

The quantitative evaluation results can be found in Tab. 2. We tested the novel view synthesis on KITTI-360 [42] using 10 images as the novel views for all 7 sequences. UrbanIR outperforms baselines such as NeRF-OSR [60] and Instruct-NeRF2NeRF [22] in all metrics, indicating that our model not only decomposes intrinsic well but also produces high-quality images. To evaluate the relighting in novel views,

Figure 10. **Dynamic Object Insertion with Shadow Volume.** We insert a simple object (yellow cube) into the scene and move it along the road for evaluating object insertion. Without visibility loss, the geometry of the unseen region is noisy and casts wrong shadows. In contrast, our full model recovers geometry and produces accurate estimates of shadow according to the inserted object position.

	Novel View Synthesis			NVS + Novel light		
	PSNR ↑	SSIM↑	LPIPS↓	PSNR ↑	SSIM↑	LPIPS↓
NeRF-OSR [60]	18.66	0.527	0.388	12.49	0.543	0.459
Instruct-N2N [22]	20.55	0.688	0.169	13.93	**0.707**	0.320
UrbanIR (Ours)	**22.95**	**0.796**	**0.135**	**17.43**	0.683	**0.218**

Table 2. **Quantitative evaluation.** We evaluate novel view synthesis (NVS) on KITTI-360 [42] and evaluate NVS + Novel light on the real-world outdoor data.

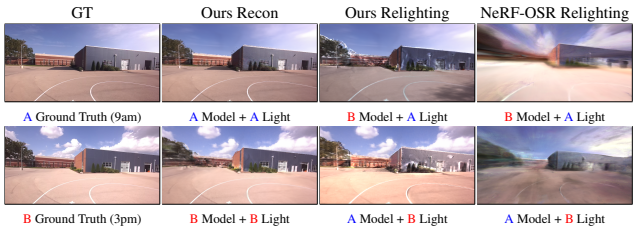

Figure 11. **Novel view and novel light synthesis.**

we captured videos of the outdoor scenes in the morning and afternoon. After individually optimizing models at both sequences, we performed relighting by exchanging lighting parameters and camera poses, and the image metrics were calculated with the ground truth capture. Our method outperformed all baselines and demonstrated the effectiveness of our intrinsic decomposition and lighting parameterization. The qualitative results can be seen in Fig. 11. UrbanIR was successful in removing existing shadows, changing the shading on the building, and modifying the sky texture during different times of the day. Please note that we selected and compared with the most competitive baseline methods that are open-sourced, and other methods such as FEGR [71] and LightSim [54] do not have codebase available publicly, making it impossible to make a fair comparison with them.

4.6. Object Insertion

Following [70, 71], we build the object insertion pipeline with Blender [13], and the results are shown in Fig. 10 and 12. By tracing the rays from the object surface toward light sources (i.e. the sun), UrbanIR estimates the visibility with volume rendering (Eq. 5). As a result, our full model can cast scene shadows on the inserted objects and also weaken the object shadow on the ground if it overlaps with the existing scene shadow. The visibility modeling (Sec. 3.3) recovers

Figure 12. **Object Insertion Qualitative Results.**Without visibility modeling (middle column), the scenes do not cast shadows on the inserted objects, and the original object shadow looks unrealistic in the existing shadow. Our full method (right column) simulates the better interaction between the reconstructed scenes and inserted objects with the help of visibility modeling.

the geometry that is not captured well in the input views (e.g. building top), enabling UrbanIR to simulate shadows better and to enhance the insertion realism significantly.

5. Limitation and Discussion

In this work, we investigated the task of inverse rendering of unbounded outdoor scenes under single illumination. This task is ill-posed and extremely challenging due to the sparsity of observations across space and time. To overcome this challenge and successfully decompose various scene intrinsic properties, we utilized prior knowledge such as pre-trained networks and regularization to reduce the uncertainty space and improve the performance of downstream applications like relighting and object insertion. However, there are limitations. Our optimization process can be affected by the noisy predictions from prior models and requires careful tuning of our losses. Sometimes, shadows cannot be removed entirely in the albedo field, and they may still appear in the final images. Additionally, the visibility optimization refines only the geometry along the light direction, which means that large changes in the sun's direction can lead to poor shadows when the geometry estimates are not accurate.

References

[1] Tomas Akenine-Moller, Eric Haines, and Naty Hoffman. *Real-time rendering*. AK Peters/crc Press, 2019.

[2] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2014. 2

[3] H.G. Barrow and Joan M. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, 1978. 2

[4] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. 2014. 2

[5] Anand Bhattad and D. A. Forsyth. Stylitgan: Prompting stylegan to produce new illumination conditions, 2023. 2

[6] Anand Bhattad, Daniel McKee, Derek Hoiem, and DA Forsyth. Stylegan knows normal, depth, albedo, and more. *arXiv preprint arXiv:2306.00987*, 2023. 2

[7] James F Blinn. Models of light reflection for computer synthesized pictures. In *Proceedings of the 4th annual conference on Computer graphics and interactive techniques*, pages 192–198, 1977. 4

[8] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, 2021. 2

[9] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan T. Barron, and Hendrik P.A. Lensch. Neural-PIL: Neural Pre-Integrated Lighting for Reflectance Decomposition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[10] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 7

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[12] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *CVPR*, 2020. 4

[13] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 8

[14] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 4

[15] Dawson-Haggerty et al. trimesh.

[16] Yue Dong, Guojun Chen, Pieter Peers, Jiawan Zhang, and Xin Tong. Appearance-from-motion: Recovering spatially varying surface reflectance under unknown lighting. *ACM Transactions on Graphics (TOG)*, 33 (6):1–12, 2014. 2

[17] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, 2021. 4

[18] David Forsyth and Jason J Rock. Intrinsic image decomposition using paradigms. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7624–7637, 2021. 2

[19] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2009. 2

[20] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. *arXiv preprint arXiv:2212.04711*, 2022. 2

[21] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *AAAI*, 2023. 2, 6, 7

[22] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 6, 7, 8

[23] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. *arXiv:2206.03380*, 2022. 2

[24] Daniel Hauagge, Scott Wehrwein, Kavita Bala, and Noah Snavely. Photometric ambient occlusion. In *CVPR*, 2013. 2

[25] Berthold KP Horn. Determining lightness from an image. *Computer graphics and image processing*, 1974. 2

[26] Berthold KP Horn. Obtaining shape from shading information. *The psychology of computer vision*, 1975. 2

[27] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. *CVPR*, 2023. 2

[28] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 1984. 3

[29] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *CVPR*, 2022. 4

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2014.

[31] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM TOG*, 2017. 7

[32] Pierre-Yves Laffont, Adrien Bousseau, and George Drettakis. Rich intrinsic image decomposition of outdoor scenes from multiple views. *IEEE transactions on visualization and computer graphics*, 2012. 2

[33] Samuli Laine, Timo Aila, Ulf Assarsson, Jaakko Lehtinen, and Tomas Akenine-Möller. Soft shadow volumes for ray tracing. In *ACM SIGGRAPH 2005 Papers*, pages 1156–1165. 2005. 2

[34] Jean-François Lalonde and Iain Matthews. Lighting estimation in outdoor image collections. In *International Conference on 3D Vision (3DV)*. IEEE, 2014. 3

[35] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 1971. 2

[36] Hendrik PA Lensch, Jan Kautz, Michael Goesele, Wolfgang Heidrich, and Hans-Peter Seidel. Image-based reconstruction of spatial appearance and geometric detail. *TOG*, 2003. 2

[37] Yuan Li, Zhi-Hao Lin, David Forsyth, Jia-Bin Huang, and Shenlong Wang. Climatenerf: Extreme weather synthesis in neural radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3227–3238, 2023.

[38] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: SVBRDF acquisition with a single mobile phone image. In *ECCV*, pages 72–87, 2018. 2

[39] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018.

[40] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *CVPR*, 2020.

[41] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. *ECCV*, 2022. 2

[42] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *in arXiv*, 2021. 4, 5, 6, 7, 8

[43] Daniel Lichy, Jiaye Wu, Soumyadip Sengupta, and David W Jacobs. Shape and material capture at home. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6123–6133, 2021. 2

[44] Stephen Lombardi and Ko Nishino. Reflectance and illumination recovery in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 38(1): 129–141, 2015. 2

[45] Stephen Lombardi and Ko Nishino. Radiometric scene decomposition: Scene reflectance, illumination, and geometry from rgb-d images. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 305–313. IEEE, 2016.

[46] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)*, 38(4):65, 2019. 2

[47] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In *ECCV*, 2018. 2

[48] Stephen Robert Marschner. *Inverse rendering for computer graphics*. Cornell University, 1998. 2

[49] 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*, 2020. 3

[50] Thomas Müller. tiny-cuda-nn, 2021.

[51] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 2022. 3

[52] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. *arXiv:2111.12503 [cs]*, 2021. 2

[53] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *CVPR*, 2022. 2

[54] Ava Pun, Gary Sun, Jingkang Wang, Yun Chen, Ze Yang, Sivabalan Manivasagam, Wei-Chiu Ma, and Raquel Urtasun. Lightsim: Neural lighting simulation for urban scenes. *NeurIPS*, 2023. 8

[55] Ava Pun, Gary Sun, Jingkang Wang, Yun Chen, Ze Yang, Sivabalan Manivasagam, Wei-Chiu Ma, and Raquel Urtasun. Neural lighting simulation for urban scenes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[56] Yi-Ling Qiao, Alexander Gao, Yiran Xu, Yue Feng, Jia-Bin Huang, and Ming C Lin. Dynamic mesh-aware radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 385–396, 2023.

[57] Chen Quei-An. ngp_pl: a pytorch-lightning implementation of instant-ngp, 2022. 3

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[59] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Neural radiance fields for outdoor scene relighting. *ECCV*, 2022. 5, 6

[60] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3, 6, 7, 8

[61] Imari Sato, Yoichi Sato, and Katsushi Ikeuchi. Illumination from shadows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003. 2

[62] Yoichi Sato, Mark D Wheeler, and Katsushi Ikeuchi. Object shape and reflectance modeling from observation. In *SIGGRAPH*, 1997. 2

[63] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8598–8607, 2019. 2

[64] 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 *CVPR*, 2021. 2

[65] Jon Story. Hybrid ray traced shadows. In *Game Developer Conference*, 2015. 2

[66] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 5, 6, 7

[67] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022. 4

[68] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 361–378. Springer, 2022. 2

[69] Yifan Wang, Andrew Liu, Richard Tucker, Jiajun Wu, Brian L Curless, Steven M Seitz, and Noah Snavely. Repopulating street scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5110–5119, 2021. 2

[70] Zian Wang, Wenzheng Chen, David Acuna, Jan Kautz, and Sanja Fidler. Neural light field estimation for street scenes with differentiable virtual object insertion. In *ECCV*, 2022. 8

[71] Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In *CVPR*, 2023. 2, 3, 4, 5, 6, 8

[72] Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representation for inverse rendering of urban scenes. *arXiv*, 2023. 2

[73] Tai-Pang Wu, Chi-Keung Tang, Michael S Brown, and Heung-Yeung Shum. Natural shadow matting. *ACM Transactions on Graphics (TOG)*, 26(2):8–es, 2007. 2

[74] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. S3-nerf: Neural reflectance field from shading and shadow under a single viewpoint. *arXiv preprint arXiv:2210.08936*, 2022. 2

[75] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3164, 2019. 2

[76] Yizhou Yu, Paul Debevec, Jitendra Malik, and Tim Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999. 2

[77] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William A. P. Smith. Self-supervised outdoor scene relighting. In *ECCV*, 2020. 5, 6

[78] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *CVPR*, 2019. 3

[79] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR*, 2021. 2

[80] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *CVPR*, 2022. 2

[81] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape from shading: A survey. *IEEE TPAMI*, 1999. 2

[82] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 2019. 2

[83] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM TOG*, 2021. 2, 4

[84] Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. *arXiv*, 2020. 2

[85] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *CVPR*, 2022. 2, 4

[86] Tinghui Zhou, Philipp Krahenbuhl, and Alexei A Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *ICCV*, 2015. 2