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

Anonymous CVPR submission

Paper ID

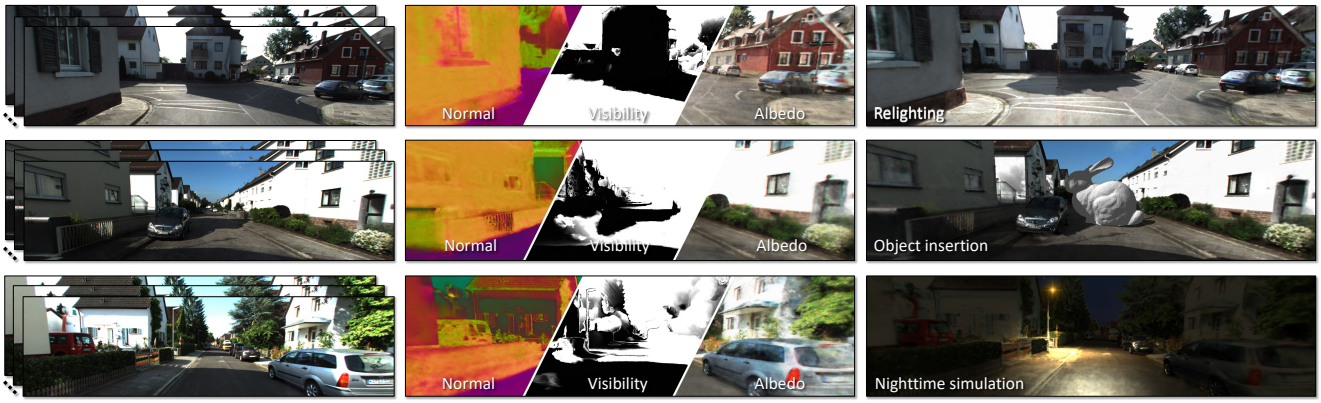

Figure 1. We present *UrbanIR* (**Urban** Scene **I**nverse **R**endering), a novel, realistic, and relightable neural scene model. UrbanIR concurrently infers shape, albedo, visibility, and more from *a single video* of large-scale, unbounded scenes. The resulting representation facilitates realistic and controllable editing, delivering photorealistic free-viewpoint renderings (last column) of relit scenes (top row), inserted objects (middle row), and nighttime simulation (bottom row).

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 make precise estimates of shadow volumes in the original scene and so 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 conditions, depth sensing, or both.

UrbanIR representations are constructed from cameras

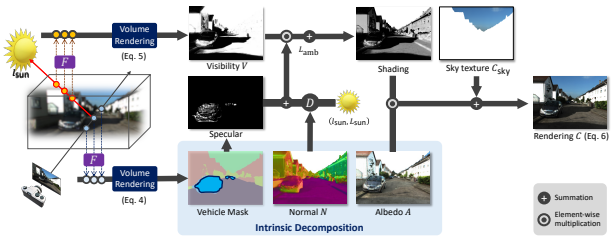

Figure 2. **Rendering Pipeline.** UrbanIR retrieves scene intrinsics (normal N, semantics S, albedo A) from camera rays, and estimate visibility V from tracing rays to the light source. The shading model computes diffuse and specular reflection and adds ambient sky light $\mathbf{L}_{\text{sky}}$ for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

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 numerous "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 also 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 key 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. Method

The illustration of the rendering pipeline is shown in Fig. 2. Please refer to our supplementary materials for the complete method description.

3. Experiment Results

3.1. Datasets

We evaluate UrbanIR on two datasets: the KITTI-360 dataset [3] and the Waymo Open Dataset [8]. The KITTI-360 dataset [3] 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 dataset includes RGB images from stereo cameras, semantic labels, camera poses, and RTK-GPS poses. On the other hand, the Waymo Open Dataset (WOD) [8] 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.

3.2. Baselines

We compare UrbanIR with scene relighting and editing methods: FEGR [9], Colmap MVS [7], Instruct NeRF2NeRF [2], NeRF-OSR [5], RelightNet [10]. Implementation details are in supplementary.

3.3. Decomposition Quality

We evaluate intrinsic decomposition on the Waymo Open Dataset [8] and present the comparison in Fig 3. NeRF-OSR [4] requires multi-illumination as input and fails to decompose albedo and shadow, leaving severe artifacts due to noisy normal estimation. FEGR [9] 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. By integrating monocular prior in the optimization process, it successfully decomposes clean albedo, normal, and shadow maps under single illumination.

3.4. Relighting Quality

Relighting under various sunlight conditions is evaluated in Fig. 5, **??**. NeRF-OSR [5] cannot simulate shadows under novel light conditions. While Blender [1] can change the lighting parameters explicitly, they either cast bad shadows due to incomplete geometry or do not cast new shadows at all; further, the original shadow remains unchanged in the image. We implement a mesh-based visibility baseline which extracts mesh with marching cubes for visibility calculation. It generates different shadows according to light conditions, but the mesh on the edges of and outside the training views is poor because there are few observations. This leads to noisy geometry and incomplete shadows on the ground. 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. **??**) by inserting car headlights and streetlights, without distracting effects

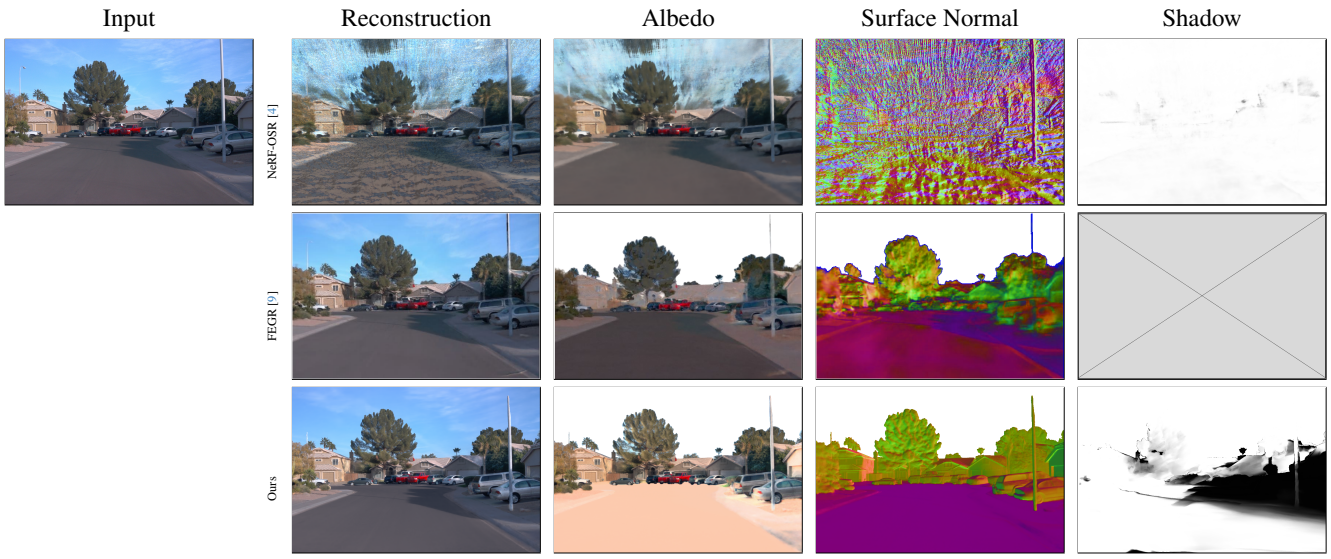

Figure 3. **Intrinsic Decomposition of Waymo Open Dataset [8].** FEGR does not release code, so we directly use the images from their paper, and the shadow map of this viewpoint is not available.

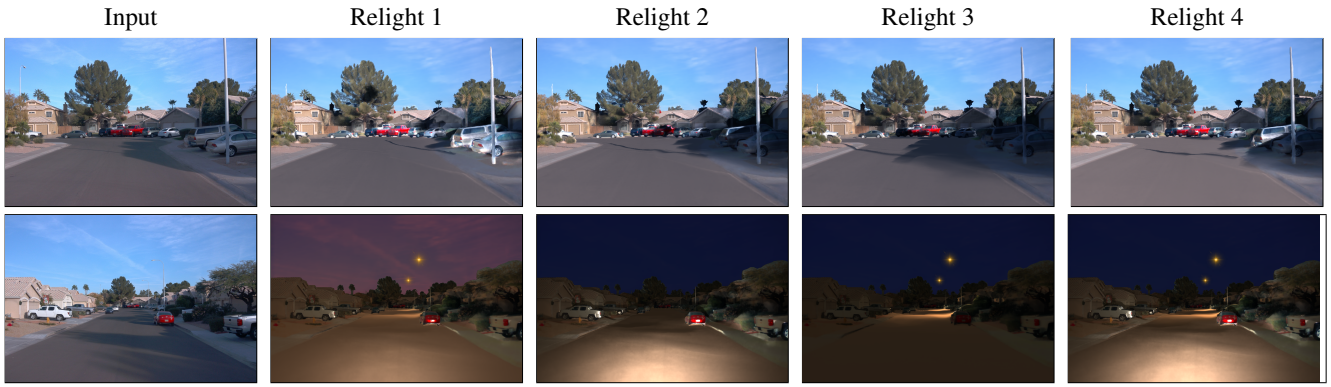

Figure 4. **Controllable Relighting of Waymo Open Dataset [8].The first row shows different lighting during a day, and the second row changes the input image into night-time with different lighting configurations**.

from the original shadows. Moreover, the relighting results obtained from UrbanIR are highly controllable, as demonstrated in Fig. 4. Different light directions and intensities were used to adjust the relighting outcomes. Light sources were also added and turned on and off.

3.5. Quantitative Evaluation

The results of the quantitative evaluation can be found in Tab. 1. We tested the novel view synthesis on KITTI-360 [3] using 10 images as the novel views for all 7 sequences. UrbanIR is better than NeRF-OSR in all metrics, indicating that our model can decompose intrinsic well and produce high-quality images. To evaluate the relighting in novel views, we captured videos of the same scene in the morning and afternoon. After optimizing models at both sequences individually, we performed relighting by exchanging lighting parameters and calculating image metrics with the ground truth capture. Our method also outperformed NeRF-OSR in all metrics significantly. The qualitative results can be seen in Fig. 7. UrbanIR was successful in removing existing shadows, changing the shading on the building, and modifying the sky texture during different times of the day.

3.6. Object Insertion

The object insertion pipeline is described in Sect. **??**. In Fig. 6, we compare with baselines by inserting a yellow cube and moving along the road. Mesh+Blender cannot synthesize complete geometry and shadow. Without visibility optimization in row (B), the scene shadow on objects is noisy. Our complete model has an accurate shadow volume so that shadows cast on the object by the environment are well represented.

4. Discussion

In this work, we investigated the task of inverse rendering of unbounded outdoor scenes under single illumination. This

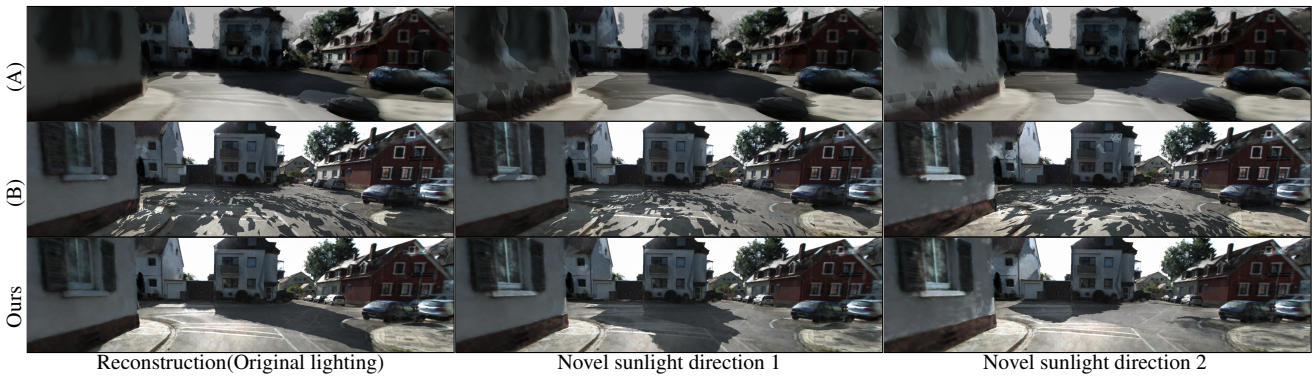

Figure 5. **Rendering and relighting comparison**. We show a set of two scenes comparing different methods. Each column shows a different sun position, the first column showing original images. For each set and from top to bottom, we have (a) COLMAP [7] + Blender [1] (b) Mesh-based visibility (NeRF-Mesh) and (c) UrbanIR (Ours). In COLMAP, Shadows are present in the scene; however, they are "baked-in" and cannot be manipulated or relit independently. For NeRF-Mesh, weak observations of the scene geometry result in poor quality visibility estimation during direct observation-based reconstruction. In contrast, for our approach, UrbanIR, our visibility optimization enables realistic and controllable relighting effects.

Figure 6. **Dynamic Object Insertion with Shadow Volume.** Our method produces accurate estimates of shadow volumes where others cannot. This can be visualized by inserting a simple object into the scene, and then looking at shadows cast onto that object. (A) COLMAP dense reconstruction [6, 7] + Blender [1] (B) Ours without visibility optimization and (C) Ours with visibility optimization improves shadows.

	Novel View Synthesis			NVS + Novel light		
	PSNR ↑	SSIM↑	LPIPS↓	PSNR ↑	SSIM↑	LPIPS↓
NeRF-OSR [5]	18.66	0.527	0.388	12.49	0.543	0.459
Ours	**22.95**	**0.796**	**0.135**	**17.43**	**0.683**	**0.218**

Table 1. **Quantitative evaluation.**

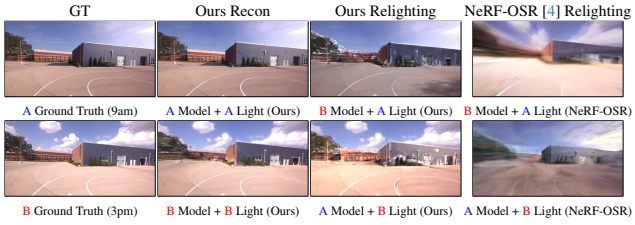

Figure 7. **Novel view and novel light synthesis.**

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] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2, 4

[2] 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. 2

[3] 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. 2, 3

[4] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Neural radiance fields for outdoor scene relighting. *ECCV*, 2022. 2, 3, 4

[5] 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, 4

[6] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 4

[7] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2, 4

[8] 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. 2, 3

[9] 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

[10] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William A. P. Smith. Self-supervised outdoor scene relighting. In *ECCV*, 2020. 2