Supplementary Materials for LightCity: An Urban Dataset for Outdoor Inverse Rendering and Reconstruction under Multi-illumination Conditions

Jingjing Wang[1] ⭐ Qirui Hu[1] ⭐ Chong Bao[1] ✉ Yuke Zhu[1]
Hujun Bao[1] Zhaopeng Cui[1] Guofeng Zhang [1] ✉
[1]State Key Lab of CAD&CG, Zhejiang University

1. More for Related Work

1.1. Intrinsic Decomposition

Intrinsic decomposition aims to separate an image into reflectance (albedo), shading and sometimes additional components. Traditional intrinsic decomposition methods rely on different assumptions, leading to three main models: grayscale intrinsic models, RGB intrinsic models, and residual models. Grayscale intrinsic models were widely used in early works, with optimization-based approaches such as [2] and various data-driven methods [19] estimating reflectance and shading under a single-channel assumption. RGB intrinsic models address the limitations of grayscale models by explicitly estimating diffuse color and shading variations, leading to improved accuracy in non-uniform lighting conditions. However, both grayscale and RGB models rely on the Lambertian assumption, making them inadequate for handling specular reflections. To overcome this, residual intrinsic models [6, 27, 36], were introduced, decomposing an image into albedo A, shading S, and a residual term R to better account for specular effects. Several works have explored this decomposition for improved reflectance modeling. Despite advancements, most intrinsic decomposition and inverse rendering approaches are evaluated on simple indoor datasets due to data limitations. Expanding these methods to complex outdoor scenes remains an ongoing challenge, particularly under diverse illumination conditions.

1.2. Inverse Rendering

Inverse rendering aims to recover intrinsic scene properties such as albedo, shading, and material properties from images, enabling applications like relighting and novel view synthesis. While significant progress has been made, most existing methods and evaluations remain object-centric, with limited exploration in large-scale, complex outdoor environments.

Early works in inverse rendering relied on physics-based models and optimization techniques to estimate reflectance and shading from single images [1, 22]. With the rise of neural representations, NeRF-based approaches have been developed to jointly learn scene geometry and appearance under varying lighting conditions. NeRV [29] and NeRD [3] incorporated reflectance decomposition into NeRF, but their evaluations were limited to controlled, object-centric datasets. More recent works, such as PhySG [39]and InvRender [41], extended inverse rendering to handle non-Lambertian surfaces and indirect illumination, yet their experiments remained focused on synthetic or small-scale real-world objects.

Gaussian-based representations have also been explored for inverse rendering. GS-IR [20] and Relit3DGS [13]extended 3D Gaussian Splatting for relighting by decomposing scene appearance into intrinsic components. However, these methods are still constrained to object-level reconstructions and have not been tested on large-scale outdoor environments.

Despite these advancements, inverse rendering has yet to be widely explored in large, real-world scenes. Existing datasets are predominantly object-centric (e.g., DTU [14], NeRF Synthetic [24], OmniObject3D [32] , limiting the generalization of these methods to urban-scale outdoor environments. The lack of benchmarks with complex outdoor lighting and diverse materials remains a significant barrier to extending inverse rendering beyond object-level scenes.

1.3. Outdoor Scene Reconstruction

Outdoor scene reconstruction has been widely studied, with Neural Radiance Fields (NeRF) [24] and 3D Gaussian Splatting (3DGS) [15] enabling high-quality scene representation. Methods like CityNeRF [31] and CityGaus-

⭐ indicates equal contribution.

✉ indicates corresponding author.

sian [21] further enhance large-scale urban reconstruction. However, real-world urban-scale data collection inherently involves complex lighting variations due to weather, time of day, and environmental factors. The presence of inconsistent illumination poses significant challenges for outdoor scene reconstruction. To address illumination variations, recent works integrate appearance modeling. NeRF-W [23] first introduced latent embeddings for variational lighting appearance. Ha-NeRF [7], CR-NeRF [34] an K-Planes [12] leveraged CNN-based, cross-ray paradigm, and feature grids to modeling different lighting effects respectively. NeuralRecon [30] focused on geometry reconstruction under uncontrolled conditions. More recently, efforts to extend 3DGS with appearance modeling have merged, including wild-gaussians [33], Wild-GS [16], Gaussian-wild [38], and SWAG [10]. While these methods improve robustness on datasets like Phototourism [28], reconstructing urban scenes under extreme multi-illumination conditions remains challenging due to the lack of standardized datasets and uniform benchmarks.

2. Details for Camera Generation

To generate camera views for our urban scenes, we design two types of view sampling methods, namely uniform view sampling and adaptive sampling. And we display the coarse point cloud reconstructed by COLMAP given our camera intrinsics and extrinsics, as shown in Fig. 1.

2.1. Uniform View Sampling

For circular views, we apply two tracking constraints to the cameras and use a frame queue to record their poses. The first constraint is based on a Bezier circle path for tracking. We place 3 Bezier circles at the center of different regions, with their radius set according to the length and width of the block. The heights of the Bezier circles are determined by the maximum object height in the region, ensuring comprehensive views from both a top-down and bottom-up perspective. The second constraint is based on the standard object tracking. We place an empty object at the center of the scene, allowing the camera to maintain the correct pose while following the Bezier curve. The view density of the curve is set adaptively based on the scale of the block. For grid views, we compute the 2D bounding box of each block and divide this bounding box into grids of varyingg resolutions based on its scale hierarchy. Within each grid, we place four cameras, with pitch angles ranging from 20 to 45 degrees and yaw angles of [0, 90, 180, 270] degree, respectively.

2.2. Adaptive View Sampling

For street views, cameras are placed along the streets within each block at 0.5m intervals. To enhance the details of the streets and surrounding buildings, we randomly generate cameras oriented in four directions, with heights sampled within the ranges of [0.5m, 0.6m] and [0.9m, 1.3m], and pitch angles in the range of [45, 60] degree. For aerial views, note that the uniform view sampling described in the previous paragraph struggles to fully capture the complex occlusion relationships within densely clustered buildings. Motivated by this, we aim to adaptively position cameras within densely clustered buildings. Specifically, we construct an adjacency lookup table in recursion based on the heights and relative positions of all buildings within a block. This lookup table enables us to generate a simplified spatial representation of the block and efficiently identify adjacent structures in four directions. For buildings located next to streets, we sample street-facing views at an adjustable height above the building. For buildings positioned adjacent to one another, we generate camera poses based on relative height relationships, ensuring finer-grained coverage of intra-block architectural structures.

3. Details for LightCity

Our LightCity dataset contains two parts, namely the LightCity reconstruction dataset and the LightCity intrinsic dataset. The dataset for urban scene reconstruction divided into regions based on scene clusters, as shown in Fig. 3. To further illustrate our dataset's diverse diffuse color, we also visualize HSV of MatrixCity dataset in Fig. 2.

3.1. LightCity Reconstruction Dataset

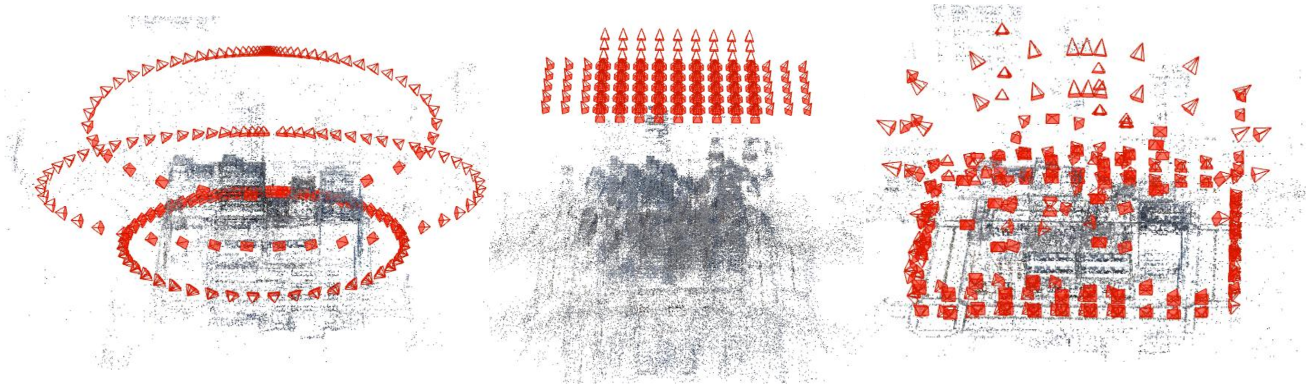
The LightCity reconstruction dataset is mainly established for task of urban scene reconstruction under multi-illuminations. Under the hierarchical-division of the city assets, we render multi-view images by uniform circle, uniform grid and adaptive sampling. Under the same viewpoints, we also construct a dataset under single-illumination.

3.2. LightCity Intrinsic Dataset

The LightCity intrinsic dataset is collected for enhancing and benchmarking outdoor intrinsic image decomposition task. To emphasize the challenge of multi-illuminations introduced in the prediction of albedo and shading, we randomly choose two sky environments, randomly rotate each fourth, randomly set the ambient lighting intensity for each view. This type of strategy enables use to simulate the complex lighting interactions within the scene across a day. For each view, we have 8 different lighted images.

3.3. Extension of LightCity Dataset

Since LightCity primarily targets the impact of diverse lighting on reconstruction and decomposition, we further include a small subset of renderings under extreme weather conditions such as fog, rain and snow (see Fig. 4). This



(a) Uniform Circle (b) Uniform Grid (c) Adaptive Sampling

Figure 1. The COLMAP coarse point clouds of block F2 under our three types of camera views sampling methods. From left to right represents uniform circle, uniform grid, and adaptive sampling. Our adaptive sampling has the most detailed and uniform point clouds, while the other two cluster on top part of the target scene.

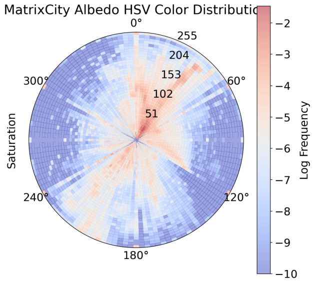


Figure 2. Visualization of HSV distribution of MatrixCity albedo images.

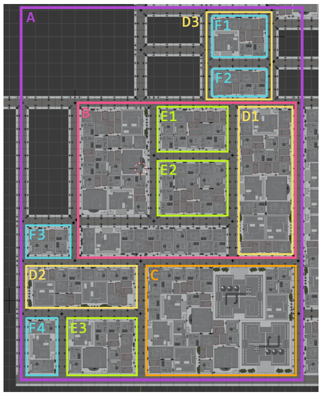


Figure 3. Different hierarchies of our LightCity reconstruction dataset, divided by different clusters of the scene. The purple rectangle represents a father node, block A, of the dataset, which contains images of the whole scene. According to different block size, block A is further split into 5 hierarchies of different scales, namely B, C, D, E, F. In total, we have 13 blocks. And we perform urban scene reconstruction on the second smallest E, F hierarchies.

enhancement aims to support further studies on weather-aware modeling. In addition, to enrich scene diversity, we also provide an extension as shown in Fig. 5 based on city assets built by the City Generator, another Blender add-on, which covers a broader range of urban layouts.

4. More Results For Intrinsic Image Decomposition

4.1. Baseline Details

We display a brief summary of methods we used for evaluation of Intrinsic Image Decomposition.

DPF. [8] DPF (Dense Prediction Fields) is a novel approach for dense prediction tasks using weak point-level supervision. It leverages point-level supervision for dense prediction by predicting values at queried coordinates, inspired by implicit representations. It enables high-resolution outputs and performs well in semantic parsing and intrinsic image decomposition.

dmp. [17] DMP leverages pre-trained text-to-image (T2I) diffusion models as priors for dense prediction tasks. It reformulates the diffusion process with interpolations to create a deterministic mapping between input images and predictions. Using low-rank adaptation for fine-tuning, DMP achieves strong generalizability across tasks like 3D property estimation and intrinsic image decomposition.

IntrinsicAny. [9] IntrinsicAnything addresses the challenge of recovering object materials from posed images under unknown lighting. Instead of relying solely on differentiable rendering, it introduces a generative material prior using diffusion models for albedo and specular components. This helps resolve ambiguities in inverse rendering. A coarse-to-fine training strategy further enforces multi-view

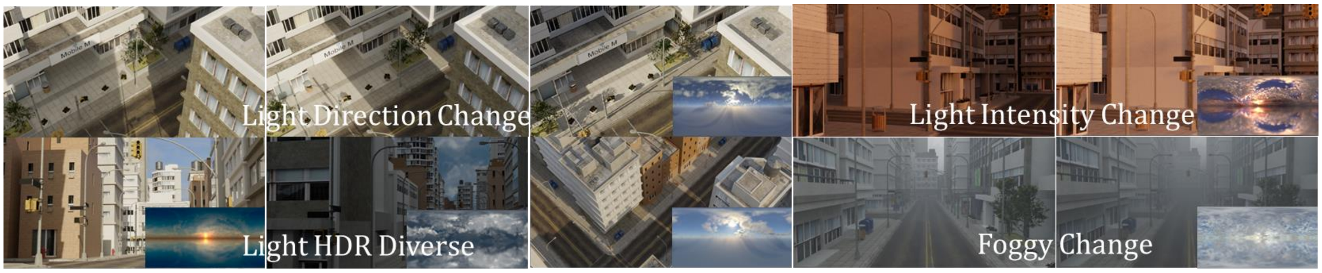


Figure 4. More examples on condition changing of LightCity.

Figure 5. More examples on expansion of LightCity built by the City Generator.

consistency, leading to more accurate material recovery.

CDID. [5] CDID tackles intrinsic image decomposition by separating an image into diffuse albedo, colorful diffuse shading, and specular residuals. Unlike prior methods assuming single-color illumination and a Lambertian world, it progressively removes these constraints, enabling more realistic and flexible illumination-aware editing.

PIENet. [11] PIE-Net is a deep learning method for detecting feature edges in 3D point clouds by representing them as parametric curves (lines, circles, B-splines). It follows a region proposal approach, first identifying edge and corner points, then ranking them for selection.

4.2. Detailed Dataset for Evaluations

We use multiple indoor and outdoor datasets for a through evaluation on our mixed-finetuning mechanism. And we provide a brief summary of all datasets we used.

Hypersim. Hypersim is a large-scale synthetic dataset featuring photorealistic indoor scenes with multi-view RGB images, depth maps, surface normals, and intrinsic decomposition (albedo, shading). It serves as a benchmark for tasks like indoor intrinsic decomposition, depth estimation, and inverse rendering.

IIW. IIW is a real-world dataset for intrinsic image decomposition, containing over 5,000 images with human-annotated pairwise reflectance comparisons. It provides a diverse set of unconstrained scenes, making it a key benchmark for evaluating intrinsic decomposition methods.

EDEN. EDEN is a multimodal synthetic dataset designed for nature-oriented applications, such as agriculture and gardening. It contains over 300K images from 100+ garden models, annotated with various vision modalities, including semantic segmentation, depth, surface normals, intrinsic colors, and optical flow. The dataset can be used for semantic segmentation and monocular depth prediction.

4.3. Indoor Scenes

We display the evaluation results of image intrinsic decomposition of indoor scenes of Hypersim and IIW in Tab. 2 and Tab. 1, respectively. For Hypersim dataset, the DNN-based CDID has the best averaged performance on si-PSNR, si-MSE and si-LMSE for albedo decomposition. However, the diffusion-based DMP tends have better visual fidelity with SSIM for albedo higher than 0.53, shading higher than 0.62. It aligns with the high quality of generated images of diffusion models. Besides, the DMP mixfine-tuned with LightCity tends to get higher si-PSNR and LPIPS for shading estimation. This findings aligns with previous in outdoor datasets. For IIW dataset, the DMP fine-tuned on Hypersim has the best WHDR score, there is a little quality drop for DMP mixfine-tuned with LightCity, we attribute this to the domain gap between the two datasets, which brings chanllenge for diffusion models to learn. However, DPF mixfine-tuned with LightCity is 5% lower on WHDR metrics, exhibiting improved performance.

4.4. Sim-to-real Discussion

Synthetic data plays a vital role in computer and robotic vision, particularly for tasks like scene understanding and inverse rendering. It allows precise control over lighting, materials, and geometry through engines such as Blender or Unreal. However, low-quality synthetic datasets can suffer from a large sim-to-real gap, negatively impacting generalization to real-world images. To mitigate this, we have used the best open-sourced rendering engine, Blender Cycles, for photo-realism. This high realism reduces the domain gap

Table 1. Performance of alebdo estimation on IIW datasets. The first , second and third values are highlighted.

IIW-Indoor			
Method		D_{train}	WHDR / %
DNN Based	PIE-Net	/	32.77
	DPF	H	43.14
		H+L	38.502
	Intrinsic 2024	objects	21.33
Diffusion Based	DMP	H	19.08
		H+L	20.37
	IntrinsicAnything	/	27.08

and improves transferability, similar to how datasets like Hypersim [26] leveraged PBR to boost real-world performance.

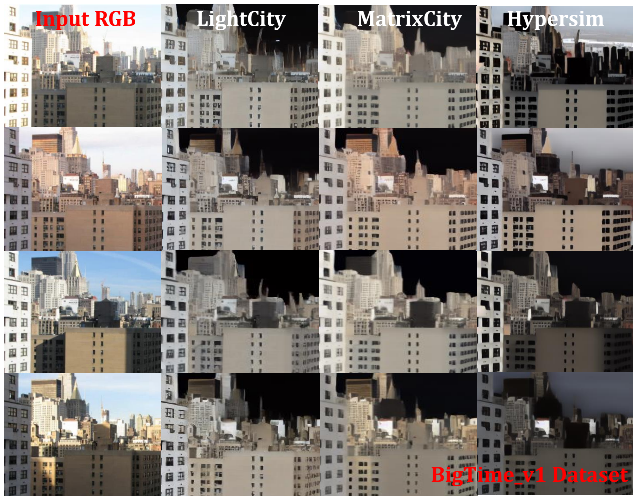


Figure 6. Albedo Decomposed from BigTime_v1 dataset.

To further evaluate the real-world generalization, we leverage the strong generalization ability of generative models. Recent studies have shown that diffusion models exhibit impressive generalization across domains, including tasks like normal prediction (e.g., StableNormal [35]). Building on DMP, a diffusion-based model, we assess intrinsic decomposition performance of both indoor and outdoor real-world scenes. For indoor evaluation, we report results on the IIW dataset(Sec. 4.3). For outdoor scenes, we use BigTime_v1 and the Waymo Open dataset. BigTime_v1 captures outdoor environments under varying illumination throughout a day, while the Waymo Open dataset offers diverse urban scenes collected under different lighting and weather conditions by Waymo autonomous vehicles. For albedo consistency as shown in Fig. 6, DMP mix-finetuned with LightCity presents lowest average variance of 0.015, while mix-finetuned with MatrixCity-mix and purely Hypersim have higher variance of 0.036 and 0.042, respectively. DMP mix-finetuned with LightCity also generalizes well to real-world urban scenes, as shown in Fig. 7.

Besides, to further minimize the sim-to-real gap, generative models can also be treated as domain transfer models for sim-to-real transfer. Established works have demonstrated the efficacy of such approaches in bridging synthetic-real discrepancies [42]. Leveraging diffusion-based image-to-image pipelines such as img2img-turbo [25] and InstructPix2Pix [4] offers a promising future direction to make synthetic datasets more applicable to real-world scenarios.

5. More Results for Multi-image Inverse Rendering

5.1. Baseline Details

We present a short description for the baseline methods (NeRF-OSR and GS-IR) for our inverse rendering.

NeRF-OSR. is the first approach to learning a neural representation that explicitly decomposes scene geometry, diffuse albedo, and shadows from multi-view and multi-illumination input images, thereby enabling more flexible scene editting.

GS-IR. first extends 3DGS for inverse rendering, leveraging a PBR framework to jointly reconstruct scene geometry, material properties, and unknown natural illumination from multi-view captured images at both object-level and scene-level tasks.

5.2. Novel View Synthesisi and Geometry Quality

We also provide geometry ground-truth for multi-view inverse rendering. And evaluate the geometry quality of both used baselines. As shown in Tab. 4, GS-IR performs better in urban scene inverse rendering than NeRF-based NeRF-OSR both in novel view synthesis and geometry reconstruction.

5.3. Material Estimation

As an important component in PBR-based inverse rendering, GS-IR also optimizes per-Gaussian metallic and roughness attribute to produce photo-realistic lighting effect. So we evaluate the decomposed material properties with our ground-truth properties, there are still large step to improve accuracy of material estimation in urban inverse rendering.

6. More Results for Multi-illumination Outdoor Reconstruction

6.1. Baseline Details

We present a short description for the baseline methods (NeRF-W, wild-gaussians and Gaussian-wild) for our outdoor reconstruction under multi-illumination.

Table 2. Single image intrinsic decomposition results under Hypersim Indoor dataset. The first , second and third values are highlighted.

Method		D_{train}	Hypersim-Indoor									
			Albedo					Shading				
			si-PSNR↑	SSIM↑	LPIPS↓	si-MSE↓	si-LMSE↓	si-PSNR↑	SSIM↑	LPIPS↓	si-MSE↓	si-LMSE↓
DNN	PIE-Net	Outdoor	12.55	0.449	0.479	0.101	0.095	14.81	0.513	0.454	0.025	0.024
	DPF	H	15.43	0.445	0.576	0.033	0.031	14.59	0.570	0.531	0.070	0.063
		H+L	13.48	0.403	0.599	0.089	0.082	14.85	0.520	0.570	0.037	0.034
	CDID	E+Indoor.etc	16.70	0.487	0.372	0.023	0.020	14.85	0.190	0.515	0.011	0.010
Diffusion	DMP	H	16.48	0.534	0.369	0.036	0.034	15.59	0.624	0.353	0.046	0.043
		H+L	16.56	0.531	0.371	0.033	0.031	15.73	0.621	0.352	0.043	0.040
	IntriAny	Objects	12.33	0.407	0.510	0.195	0.182	/				

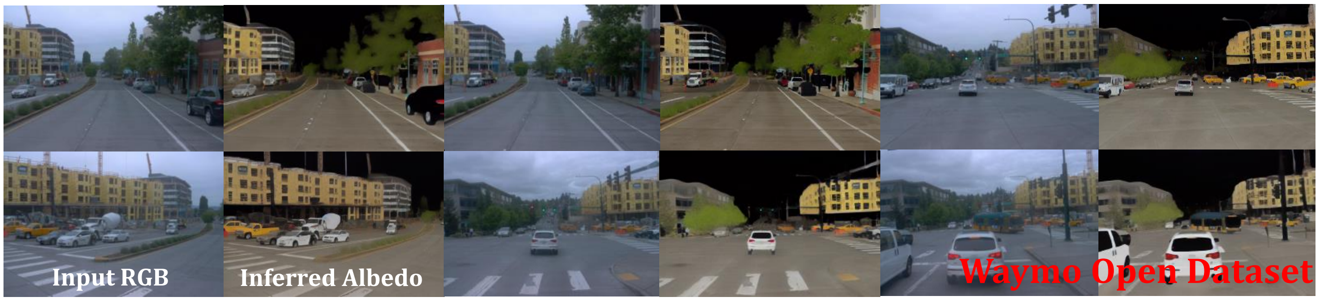


Figure 7. Albedo Decomposed from Waymo Open dataset.

NeRF-W. extends the implicit NeRF to unconstrained multi-illumination reconstruction by introducing a per-image learned low-dimensional latent appearance embeddings as shared MLP conditions utilizing GLO, thus disentangling scene geometry from illumination inconsistencies.

wild-gaussians. adapts the explicit 3D Gaussian Splatting (3DGS) representation for real-world scene reconstruction under varying lighting conditions. It incorporates an MLP-based appearance modeling module with affine color mapping to capture image-dependent Gaussian colors while preserving rendering efficiency.

Gaussian-wild. further enhances local high-frequency changes of the scene by separating each Gaussian's appearance into intrinsic and dynamic features based on 3DGS, to better capture fine-grained scene details while adapting to varying lighting conditions.

NexusSplats. utilizes an neural network to represent image-specific global lighting conditions and Gaussian-specific localized response to global lighting variations, to effectively capture complex illumination changes across scenes.

6.2. Novel View Synthesis

To provide a baseline for our LightCity reconstruction dataset. We also train Gaussian-wild (GS-W) under the single-illumination dataset. The result is shown in Tab. 3. Compared with that trained under multi-illumination dataset, the performance dropped, which further indicating the strong influence of multi-illumination on performance of urban reconstructions.

In previous sections, we display the visualization evaluation results under test set of multi-illumination dataset. We also display the results under the test set of single-illumination dataset in Fig. 9. Compared with 3DGS, methods for modeling appearance embedding has a quality degradation. Although, NeRF-W is able to restore the shadow of the image (col1), it's performance under other unseen views remain worse. GS-W tends to restore a more clear structure of the never-seen input GT, but there are floaters in some part. This further illustrate the challenge on our multi-illumination reconstruction dataset.

We also perform a deep analysis of the performane between the best GS-W and NeRF-W, as visualized in fig. 8. The first row illustrates blurred detail of GS-W, floaters covering the building leading to visual artifacts. The second row illustrates blurred detail of NeRF-W, which tends to blur the detail of complex scenes.

6.3. Geometry Quality

For thoroughly evaluate the reconstruction geometry of the LightCity dataset, we render the normal map of all used methods under multi-illumination conditions. The error metrics are displayed in Tab. 6. Across all blocks, Gaussian-wild has the lowest MeaAE and MedAE, indicating its prior

Table 3. Performance of novel view synthesis of Gaussian-Wild trained under single-illumination dataset for block F2

F2	Gaussian-wild
PSNR	28.74
SSIM	0.878
LPIPS	0.174

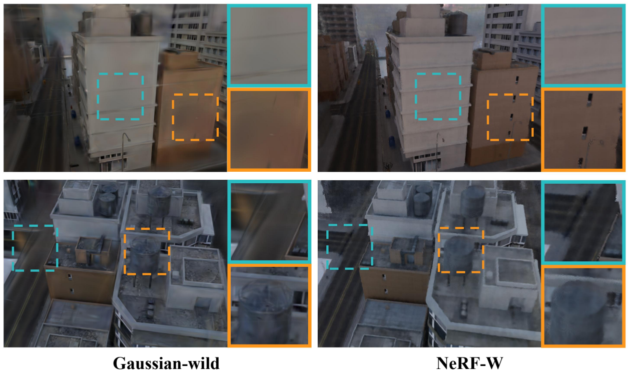


Figure 8. Comparison of novel view synthesis under multi-illumination between Gaussian-wild and Nerf-W.

Table 4. Performance of novel view synthesis for multi-view inverse rendering.

		Datasets			
Methods	Metrics	F2	F3	E1	E2
NeRF-OSR	PSNR	17.35	20.15	21.11	20.95
	SSIM	0.562	0.600	0.622	0.597
	LPIPS	0.461	0.413	0.400	0.438
	MeaAE	32.81	31.39	30.96	33.98
GS-IR	PSNR	**26.35**	**27.26**	**27.29**	**26.76**
	SSIM	**0.862**	**0.90**	**0.858**	**0.861**
	LPIPS	**0.233**	**0.186**	**0.196**	**0.200**
	MeaAE	**28.07**	**23.60**	**27.73**	**28.78**

Table 5. Performance of material estimation for multi-view inverse rendering.

Datasets	Metrics	GS-IR
A2	metallic mse	0.1168
	roughness mse	0.2643
A3	metallic mse	0.1813
	roughness mse	0.2964
B1	metallic mse	0.2888
	roughness mse	0.2833
B2	metallic mse	0.2599
	roughness mse	0.2466

geometry reconstruction quality compared with other methods. However, under the constraint of multi-illumination, those methods presents a quality decay compared with the origin 3DGS. Besides, we presents the normal map of different methods in fig. 12. Although Gaussian-wild has the highest normal accuracy, it tends to be blurred in some flat areas, this may due to the extra floaters introduced by multi-illumination input. However, NeRF-W has a relatively sharp normal except for some roughness. This might be attributed to its discrete sampling of rays. Another two 3DGS-based methods, i.e., wild-gaussian and NexusSplats, can hardly reconstruct normals of the scene, with a wide area of Gaussian surfels covering the screen space (column 3).

Besides, we also investigate the consistency of normal between different views. As illustrated in Fig. 11, NeRF-W tends to reconstruct different normal maps between different views of the same scene, exhibiting strong inconsistencies. This problem is not found for GS-based methods since they disentangle apperance and location of each Gaussian.

7. More Results on Relighting

Relighting is a vital and real-world task in computer vision, enabliing applications such as content editing, lighting transfer and scene manipulation. In our multi-illumination reconstruction experiments, we perform 3D-based relighting by optimizing the reconstructed representation under test lighting conditions. Visual examples are shown in right of Fig. 10. In addition to 3D-driven methods like NeRF-OSR, we also evaluate image-based relighting techniques-single-image models that directly manipulate input images to match target lighting. We evaluated three models: IC-Light [40], Self-OSR [37] and ColorTransfer [18]. As shown in left of Fig. 10, IC-Light struggled to relight complex outdoor scenes while Self-OSR and ColorTransfer showed only limited performance. These results indicate that current image-based relighting methods generalize poorly to outdoor urban scenes. Thus, LightCity offers promising potential to support future work in image-based relighting for outdoor environments.

References

[1] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014. 1

[2] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4): 1–12, 2014. 1

[3] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of*

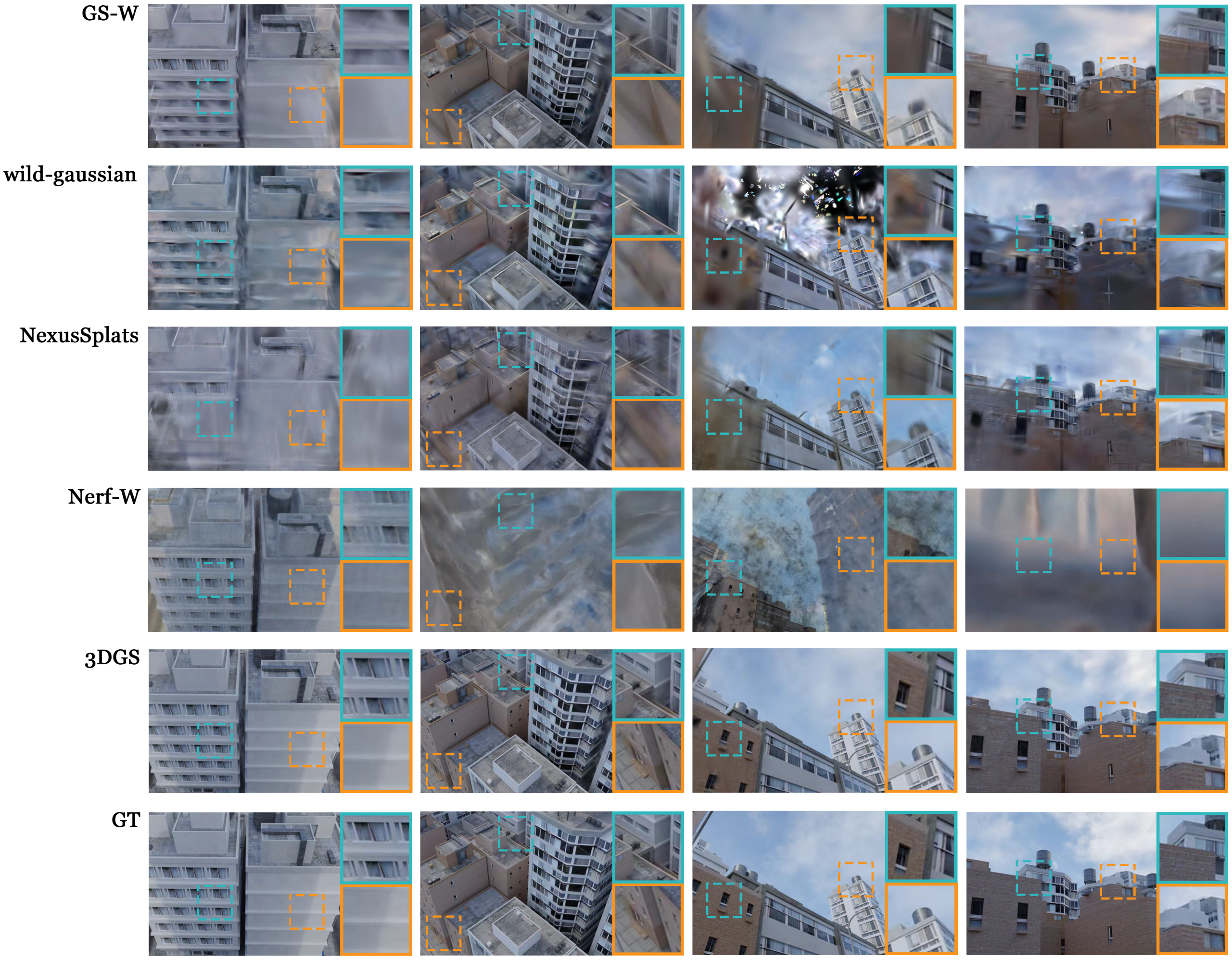


Figure 9. Novel-view rendering results under test set of single-illumination dataset.

Table 6. Performance comparison of geometry quality for urban reconstruction under multi-illuminations.

		Methods				
Datasets	Metrics	3DGS	Gaussian-wild	wild-gaussian	NexusSplats	Nerf-w
A2	MeaAE	21.57	**26.91**	32.23	32.94	27.88
	MedAE	21.81	**26.38**	31.30	31.81	27.49
A3	MeaAE	23.50	**25.27**	30.85	32.83	29.81
	MedAE	23.33	**26.00**	31.08	32.22	28.45
B1	MeaAE	24.34	**31.15**	37.02	35.76	49.76
	MedAE	23.52	**29.65**	36.31	35.45	50.55
B2	MeaAE	23.99	**29.16**	40.63	37.84	37.99
	MedAE	23.98	**28.13**	40.34	36.23	36.21

the IEEE/CVF International Conference on Computer Vision, pages 12684–12694, 2021. 1

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 5

[5] Chris Careaga and Yağız Aksoy. Intrinsic image decomposition via ordinal shading. *ACM Transactions on Graphics*, 43

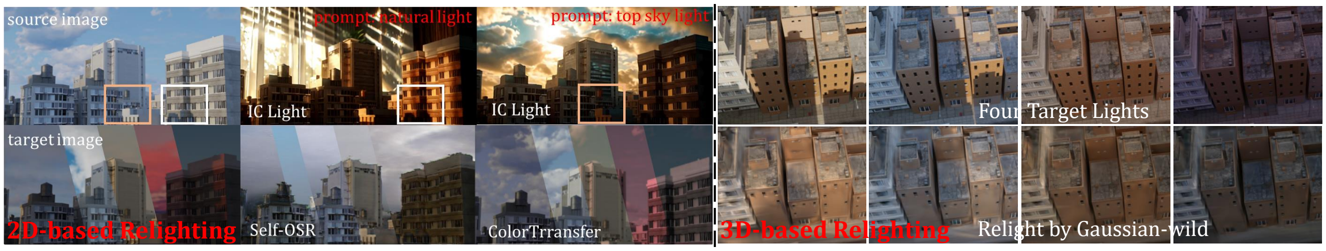


Figure 10. Visual results for 3D-based relighting.

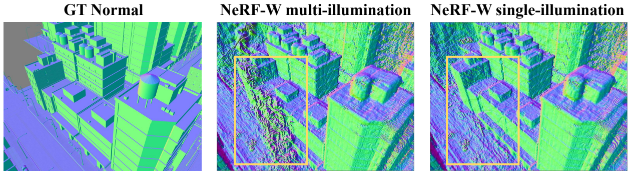


Figure 11. Comparisons of normal maps under different views of NeRF-W for urban reconstruction.

(1):1–24, 2023. 4

[6] Chris Careaga and Yağız Aksoy. Colorful diffuse intrinsic image decomposition in the wild. *ACM Trans. Graph.*, 43 (6), 2024. 1

[7] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12952, 2022. 2

[8] Xiaoxue Chen, Yuhang Zheng, Yupeng Zheng, Qiang Zhou, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Dpf: Learning dense prediction fields with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15347–15357, 2023. 3

[9] Xi Chen, Sida Peng, Dongchen Yang, Yuan Liu, Bowen Pan, Chengfei Lv, and Xiaowei Zhou. Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination. In *European Conference on Computer Vision*, pages 450–467. Springer, 2024. 3

[10] Hiba Dahmani, Moussab Bennehar, Nathan Piasco, Luis Roldao, and Dzmitry Tsishkou. Swag: Splatting in the wild images with appearance-conditioned gaussians. In *European Conference on Computer Vision*, pages 325–340. Springer, 2024. 2

[11] Partha Das, Sezer Karaoglu, and Theo Gevers. Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19790–19799, 2022. 4

[12] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2

[13] Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*, pages 73–89. Springer, 2024. 1

[14] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 1

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[16] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. WildGaussians: 3D gaussian splatting in the wild. *NeurIPS*, 2024. 2

[17] Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7861–7871, 2024. 3

[18] Junyong Lee, Hyeongseok Son, Gunhee Lee, Jonghyeop Lee, Sunghyun Cho, and Seungyong Lee. Deep color transfer using histogram analogy. *The Visual Computer*, 36(10): 2129–2143, 2020. 7

[19] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European conference on computer vision (ECCV)*, pages 371–387, 2018. 1

[20] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21644–21653, 2024. 1

[21] Yang Liu, Chuanchen Luo, Lue Fan, Naiyan Wang, Junran Peng, and Zhaoxiang Zhang. Citygaussian: Real-time high-quality large-scale scene rendering with gaussians. In *European Conference on Computer Vision*, pages 265–282. Springer, 2024. 2

[22] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 1

[23] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021. 2

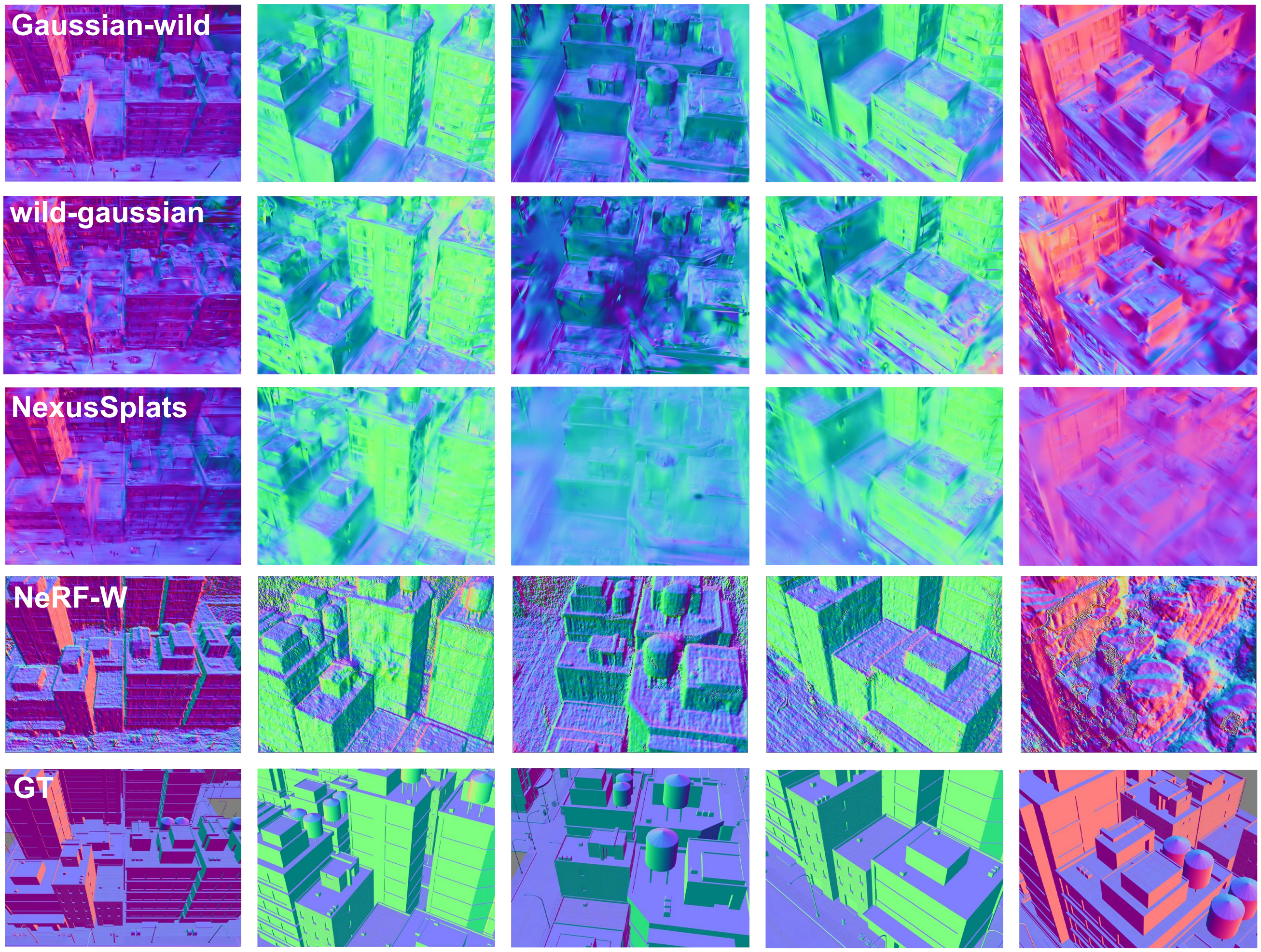


Figure 12. Visualization of normal maps for urban reconstruction under multi-illuminations.

[24] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1

[25] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024. 5

[26] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 5

[27] Jian Shi, Yue Dong, Hao Su, and Stella X Yu. Learning non-lambertian object intrinsics across shapenet categories. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1685–1694, 2017. 1

[28] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. Association for Computing Machinery, 2006. 2

[29] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7495–7504, 2021. 1

[30] Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. Neural 3d reconstruction in the wild. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022. 2

[31] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12922–12931, 2022. 1

[32] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for

realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 1

[33] Jiacong Xu, Yiqun Mei, and Vishal Patel. Wild-gs: Real-time novel view synthesis from unconstrained photo collections. *Advances in Neural Information Processing Systems*, 37:103334–103355, 2024. 2

[34] Yifan Yang, Shuhai Zhang, Zixiong Huang, Yubing Zhang, and Mingkui Tan. Cross-ray neural radiance fields for novel-view synthesis from unconstrained image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15901–15911, 2023. 2

[35] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 43(6):1–18, 2024. 5

[36] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 41(6): 1–21, 2022. 1

[37] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William AP Smith. Self-supervised outdoor scene relighting. In *European Conference on Computer Vision*, pages 84–101. Springer, 2020. 7

[38] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In *European Conference on Computer Vision*, pages 341–359. Springer, 2024. 2

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021. 1

[40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 7

[41] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18643–18652, 2022. 1

[42] Haonan Zhao, Yiting Wang, Thomas Bashford-Rogers, Valentina Donzella, and Kurt Debattista. Exploring generative ai for sim2real in driving data synthesis. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 3071–3077. IEEE, 2024. 5