

LIRM: Large Inverse Rendering Model for Progressive Reconstruction of Shape, Materials and View-dependent Radiance Fields

Supplementary Material

6. Overview

Our supplementary material consists of three parts.

- Implementation details, including multi-stage coarse-to-fine training, accelerated deferred rendering and training and testing datasets creation.
- Ablation studies on synthetic dataset, including impacts of multi-stage coarse-to-fine training and input camera trajectories, as well as qualitative and quantitative comparisons with baselines [89].
- A complementary collection of results on real data to the main paper, including qualitative results from Stanford-ORB [35] and analysis of LIRM’s limitations.
- LIRM reconstruction on 2 more challenging settings, including using LIRM update model to reconstruct a changing scene and LIRM reconstruction using egocentric data.

All of the results in the supplementary materials use the same implementation and datasets as the main paper.

In addition, we include a video for better visualization.

7. Implementation details

In this section, we summarize all implementation details to ensure the reproducibility of our method. We are seeking code release in the near future.

Coarse-to-fine training Our training consists of three stages. We first train with large batch sizes but small resolutions for fast convergence, and later train with high resolutions but small batch sizes for better details. The hyperparameters for the three stages are summarized in Tab. 6. Similar to [85], we utilize cropped regions from the original ground-truth image for supervision. β is the standard deviation that controls sharpness of the surface, as mentioned in Eq. (9). We increase $\frac{1}{\beta}$ linearly following [73]. We also tried to learn β using gradient descent; however, this approach resulted in less stable training.

Accelerated deferred rendering Deferred rendering [99] is used in all prior volume-based LRM methods [28, 36, 85] to reduce GPU memory consumption. The basic idea is to cache the gradients so that we can render an image patch-by-patch while still computing a perceptual loss like LPIPS on the whole image, which is essential for reconstructing texture details. In our setting, we find that rendering the whole image crop in a single pass can reduce time consumption for the first two stages of training, making deferred rendering unnecessary. However, in the third training stage, rendering the whole image crop becomes impractical

due to memory overflow, as we significantly increase the number of samples per ray and compute numerical normals for improved geometry reconstruction. We therefore adopt the occupancy grid acceleration developed in Nerfacc [37] in the third stage. Before we render the image crops, we first compute an occupancy grid of resolution 250 and filter out voxels with α lower than $1e^{-4}$. The computation of occupancy grid takes only 0.05 s, while significantly reducing GPU memory consumption and accelerates training. It allows us to filter out 91.2 percentage of sampled points on average and reduce the time consumption to render four image crops of size 192×192 from around 4s to 0.3s.

Training datasets creation The camera settings used to create synthetic datasets under uniform lighting are identical to those used under environmental lighting. For each 3D model, we render 32 images for training. To ensure generalizability to various camera types, the field of view is uniformly sampled between 15° and 85° . The elevation angle is uniformly sampled between $[-5^\circ, 70^\circ]$ while azimuth angle is uniformly sampled between $[0^\circ, 360^\circ]$. For data augmentation, we employ an auto-exposure algorithm that automatically adjusts the camera’s exposure settings. During training, we also on the fly apply a perturbation scale to image pixels uniformly sampled between 0.75 and 1.25. For the synthetic dataset under environmental lighting, we generate two versions: one with the original roughness values from the 3D models, and another where the roughness values are scaled by a factor between 0.3 and 0.6 to create more specular appearances. The two datasets are mixed together to train LIRM for inverse rendering.

Testing datasets creation We select input and output views for the testing datasets in a manner that closely follows MeshLRM [85]. We set elevation angles as 0° , 20° , 40° and uniformly divide azimuth angle into 16 intervals, which gives us 48 views in total. From these 48 views, we uniformly sample 8 views at elevation 20° and 40° as the input views. We then sample 12 views from the remaining 32 views as the output views. The FoV is set to be 50° . The camera always looks at the object center. Its distance to the object center is set as the minimal distance that can cover the object’s bounding sphere. Testing datasets captured under uniform lighting and environmental lighting conditions follow the same camera settings and view selection method.

LRM -VolSDF	Learning rate	Batch size	Input num.	Samples per ray	Input res.	GT res.	Crop. res.	$\frac{1}{\beta}$	Epochs	Update
Stage 1	$4e^{-4} \rightarrow 2e^{-5}$	768	[3, 6]	128	512	256	128	$1e \rightarrow 2e^2$	30	2
Stage 2	$2e^{-5} \rightarrow 1e^{-6}$	320	[3, 6]	512	512	384	128	$2e^2 \rightarrow 2.5e^2$	5	3
Stage 3	$1e^{-6} \rightarrow 0$	256	[3, 6]	1024	512	512	192	$2.5e^2$	2	3

Table 6. Training settings for LIRM. Our learning rate decreases following a cosine scheduling while $\frac{1}{\beta}$ increases following a linear scheduling.

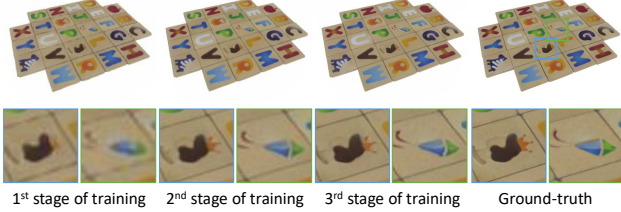


Figure 9. Qualitative comparisons of view synthesis results after different stages of our coarse-to-fine training paradigm.

Table 7. Quantitative comparisons of different stages of training for view synthesis under uniform lighting on GSO dataset

LIRM-hexa 4 th	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)	CD (\downarrow)
Stage 1	27.56	0.924	0.113	0.120
Stage 2	30.80	0.950	0.060	0.118
Stage 3	30.56	0.948	0.054	0.115

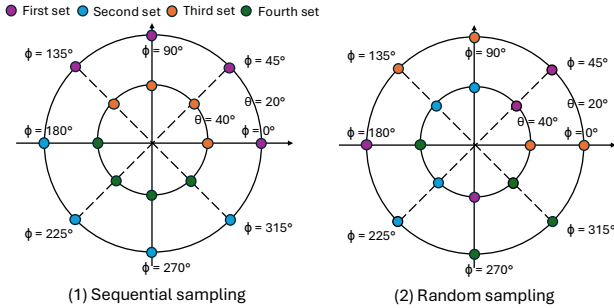


Figure 10. Two different camera trajectories to test LIRM's update model. Random sampling is the default camera trajectory used in the main paper. θ and ϕ are elevation and azimuth angles.

8. Experiments on Synthetic Data

Impacts of coarse-to-fine training We test the network's reconstruction quality after different stages of training. We run the experiments on the GSO dataset rendered with uniform lighting. The quantitative results are summarized in Tab. 7. We report the view synthesis metrics and chamfer distance after the 4th update. The second stage of training significantly enhances texture details, while the second and third stages exhibit similar texture quality. However, the geometry quality in the third stage is better due to the incorporation of numerical normal loss. Fig. 9 visualizes view synthesis results from different stages of training.

Table 8. Quantitative comparisons of different camera trajectories for view synthesis under uniform lighting on GSO dataset. "Rd" and "Sq" represent random sampling and sequential sampling. The numbers of rows (1st to 4th) represent the number of updates performed by our model.

	PSNR (\uparrow)		SSIM (\uparrow)		LPIPS (\downarrow)	
	Rd	Sq	Rd	Sq	Rd	Sq
1 st	29.27	27.70	0.941	0.933	0.061	0.081
2 nd	30.48	30.09	0.947	0.946	0.056	0.060
3 rd	30.65	30.66	0.949	0.949	0.054	0.055
4 th	30.56	30.56	0.948	0.948	0.054	0.055

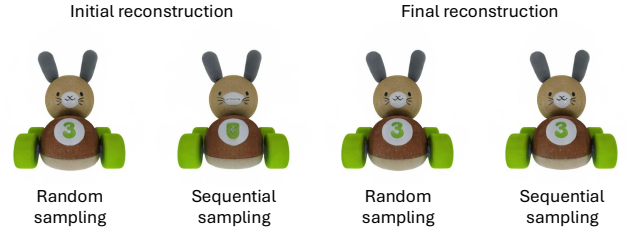


Figure 11. Comparisons of reconstruction results under different camera trajectories. LIRM is robust to the order of input images and can converge to similar reconstruction results.

Impacts of inference camera trajectories For all the synthetic data experiments in the main paper, all the input views are fed to LIRM follow the same random order, which is shown in Fig. 10 (2). We test the impact of camera trajectory by feeding input images into LIRM sequentially, as shown in Fig. 10 (1). Qualitative and quantitative results of the new camera trajectory are summarized in Fig. 11 and Tab. 8 respectively. We observe that the initial reconstruction results are worse when we follow the new sequential order because those initial input images only observe one side of objects. However, our reconstruction errors converge to similar numbers after using all 16 images. It shows that LIRM update module is very robust to camera trajectories.

Comparisons with a MeshLRM baseline Prior state-of-the-art LRM-based mesh reconstruction method [85] has not been open sourced yet. To compare with this strong baseline, we trained an LRM-VolSDF model with the same network architecture as [85] using our newly created synthetic dataset built on Shutterstock [1]. Tab. 9 compares the baseline model with LIRM on GSO dataset rendered with uniform lighting. We observe that LIRM consistently

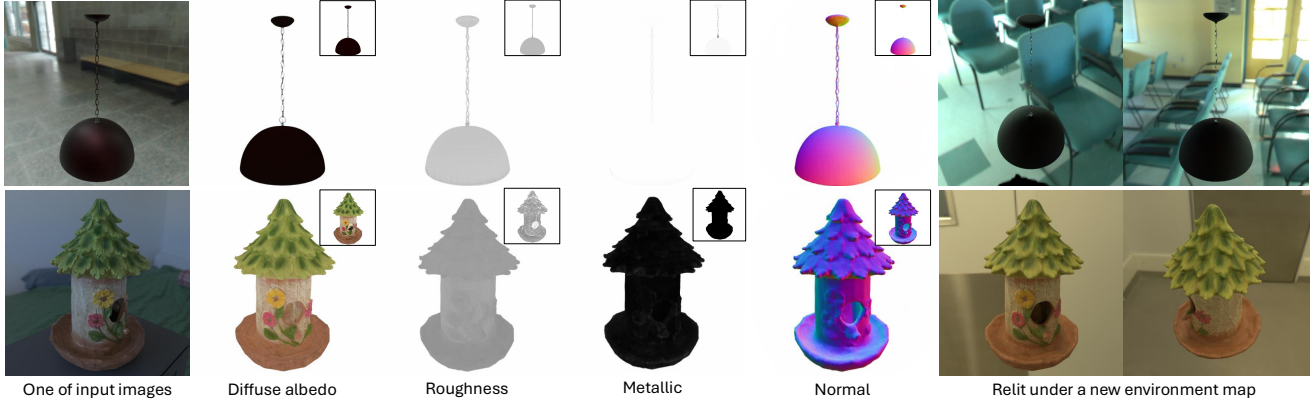


Figure 12. More inverse rendering results from ABO [12] (first row) and DTC [2] (second row) datasets. Ground-truths material maps are included in insets.

Table 9. Quantitative comparisons for view synthesis under uniform lighting on **GSO** dataset with different number of images.

4 images	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)
Baseline	28.72	0.940	0.070
LIRM-hexa 1 st	29.27	0.941	0.061
8 images	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)
Baseline	30.19	0.947	0.061
LIRM-hexa 2 nd	30.48	0.947	0.056
12 images	PSNR (\uparrow)	SSIM (\uparrow)	LPIPS (\downarrow)
Baseline	30.50	0.948	0.059
LIRM-hexa 3 rd	30.65	0.949	0.054

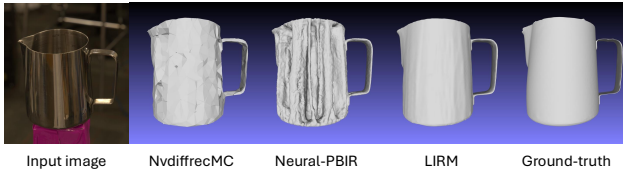


Figure 13. Qualitative comparisons of geometry reconstruction quality. Meshes reconstructed by LIRM have much less artifacts compared to optimization-based methods.

performs better compared to the baseline model with different number of input images. Moreover, our update model enables us to utilize more input images without increasing GPU memory consumption, whereas the baseline model requires sending all images to the transformer simultaneously.

Complementary qualitative inverse rendering results

In Fig. 12, we show more qualitative inverse rendering and relighting results from DTC [2] and ABO [12] datasets. The ground-truths materials are shown in the insets. For examples from both datasets, LIRM can accurately reconstruct detailed geometry and spatially varying materials, such as roughness and metallic maps. These high-quality inverse rendering results enable us to create realistic rendering under novel lighting conditions.

9. More Experiments on Real Data

Complementary qualitative results on Stanford-ORB

Fig. 14 includes more inverse rendering and relighting results from Stanford-ORB [35] dataset. We compared with 4 prior works as described in the main paper. MetaLRM [70] is a concurrent LRM-based inverse rendering method that takes sparse views as inputs and reconstructs geometry and material maps in a feed-forward manner. We obtained the relighting results from authors of [70]. We can see LIRM significantly improves both the geometry and material reconstruction quality compared to the prior work. We also compared LIRM with the three best optimization-based methods on Stanford-ORB’s leaderboard. All three methods take dense views as inputs and need at least close to an hour to finish optimization. In contrast, LIRM achieves reconstruction quality comparable to state-of-the-art methods using only sparse inputs and requiring less than 1 second for reconstruction. We observe that LIRM better recovers specular highlights and exhibits more robust geometry reconstruction for glossy materials compared to optimization-based methods. Fig. 13 shows geometry reconstruction results of LIRM and two leading optimization-based inverse rendering methods. For shiny objects, LIRM can still generate smooth geometry while optimization-based methods either lose geometry details or generate a lot of artifacts near the highlight regions. This reveals that our LIRM model has learned high-quality geometry prior from the large collection of 3D models.

Limitations of LIRM

We observe two major limitations of LIRM. First, even though our NDE module can model specular highlights and view dependent effects, as shown in Fig. 6, Fig. 5 and Tab. 5, it fails to model mirror reflection as shown in Fig. 15 (a). We argue this is an extremely challenging problem as it requires the network to reconstruct the full 3D scene from sparse observation of reflection of an unknown object. In addition, compared to optimization-

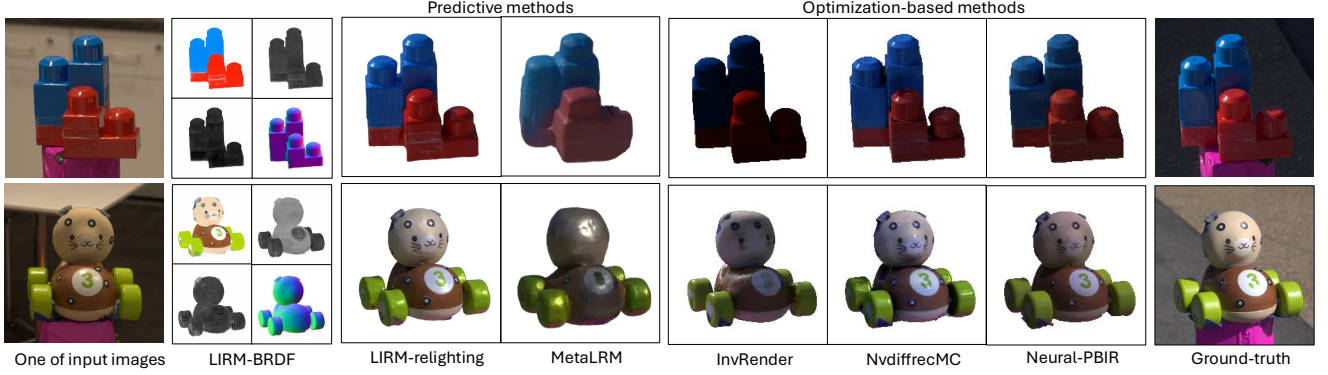


Figure 14. More inverse rendering and relighting results on Stanford-ORB dataset [35].

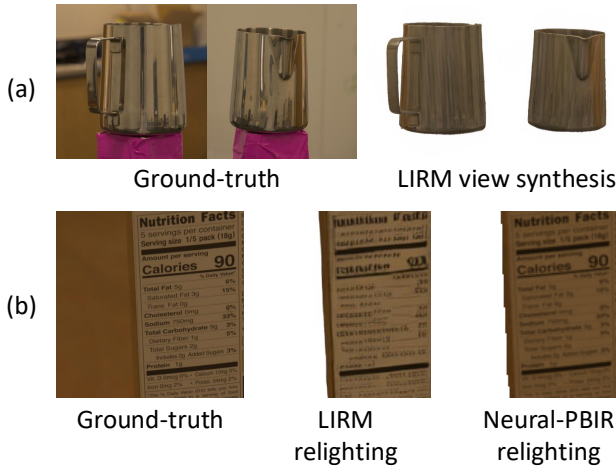


Figure 15. Two limitations of LIRM. (a) LIRM-NDE cannot handle high-frequency reflection of mirror surfaces. (b) Compared to optimization-based method, LIRM still traces behind in reconstructing texture details.

based methods, LIRM still fails to recover more detailed texture. In Fig. 15 (b), LIRM can reconstruct writings for the brand name printed on the box, but not ingredient list, unlike an optimization-based method [73]. We attribute this limitation to the network capacity, as our hexa-plane representation should have a sufficiently high resolution. A larger model may be required to achieve higher-quality reconstructions.

Challenging scenarios We test our LIRM’s generalization ability on two more challenging scenarios. The first scenario is updating reconstruction of a changing scene, where we first capture one set of images, then change the scene configuration by adding a new object and capture the second set of images. We use the two sets of images as inputs to LIRM’s update model. The input images and reconstruction results are shown in Fig. 16. Despite that this scenario never occurs in the training data, LIRM manages to reconstruct the added object accurately while still preserv-

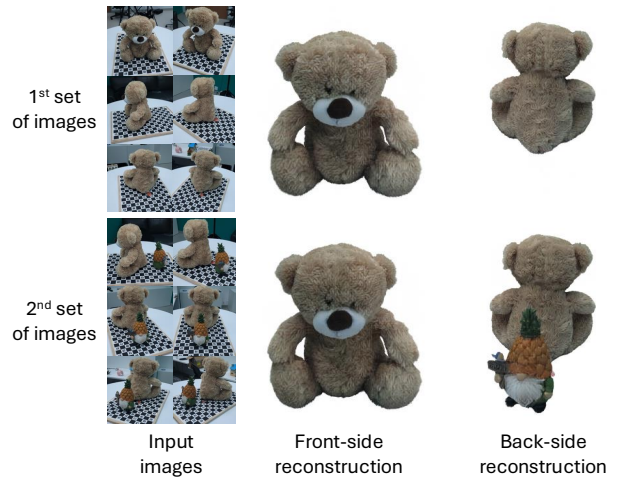


Figure 16. Reconstructing a changing scene with the LIRM update model. Even if we change the scene configuration after capturing the first set of input images, LIRM’s update model can still achieve accurate reconstruction of the newly added object while preserving old reconstruction.

ing the initial reconstruction of the first object. Note that the front face of the “teddy bear” is not shown in the second set of input images and yet our model keeps all the facial details unchanged through the updating process. This suggests that our model may be applied to dynamic scene reconstruction or large-scale scene reconstruction when one set of images cannot cover the whole scene.

In the second scenario, we test our LIRM model on images casually captured by egocentric Aria glasses. Users were asked to wear a pair of Aria glasses, walking towards the object, causally look around the object and then walk away. This egocentric capturing setting better mimics how common people may take photos for 3D reconstruction. However, it also presents unique challenges, such as large field-of-view, motion blur, sensor noise, etc. We directly test our LIRM model on the challenging egocentric captured images without any fine-tuning. Example inputs and

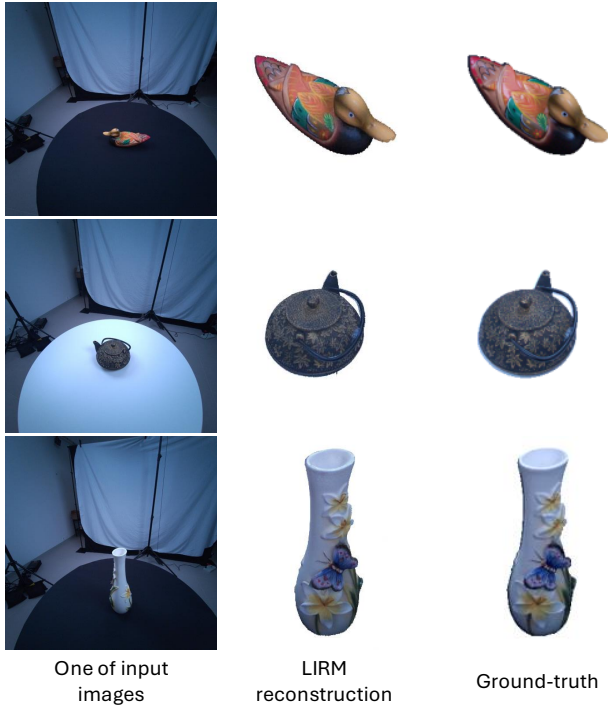


Figure 17. LIRM reconstruction from images casually captured using egocentric Aria glasses [19].

reconstruction results are shown in Fig. 17. For each video sequence, we extract 16 images as inputs. Even though there are clear domain gaps between testing inputs and our training data, our LIRM still reconstructs the object appearance that is very close to the ground-truth.

References

- [1] Shutterstock. <https://www.shutterstock.com/search/3d>. 7, 2
- [2] Digital twin catalog. <https://www.projectaria.com/datasets/dtc/>, 2024. 7, 3
- [3] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013. 2
- [4] 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. 2
- [5] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 1, 2
- [6] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *Advances in Neural Information Processing Systems*, 34: 10691–10704, 2021.
- [7] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan Barron, Hendrik Lensch, and Varun Jampani. Samurai: Shape and material from unconstrained real-world arbitrary image collections. *Advances in Neural Information Processing Systems*, 35:26389–26403, 2022. 1, 2
- [8] Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint*, 2024. 3
- [9] Chris Careaga and Yağız Aksoy. Intrinsic image decomposition via ordinal shading. *ACM Transactions on Graphics*, 43(1):1–24, 2023. 2
- [10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3
- [11] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnr: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14124–14133, 2021. 3
- [12] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 7, 3
- [13] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 4
- [14] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 1
- [15] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 3
- [16] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [17] Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (ToG)*, 37(4):1–15, 2018. 2
- [18] Sam Earle, Filippos Kokkinos, Yuhe Nie, Julian Togelius, and Roberta Raileanu. Dreamcraft: Text-guided generation of functional 3d environments in minecraft. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, pages 1–15, 2024. 3
- [19] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brigid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eickenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tasos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreeses, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for ego-centric multi-modal ai research, 2023. 5
- [20] Andreas Engelhardt, Amit Raj, Mark Boss, Yunzhi Zhang, Abhishek Kar, Yuanzhen Li, Deqing Sun, Ricardo Martin Brualla, Jonathan T Barron, Hendrik Lensch, et al. Shinobi: Shape and illumination using neural object decomposition via brdf optimization in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19636–19646, 2024. 1, 2

- [21] 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. 4
- [22] Clement Fuji Tsang, Maria Shugrina, Jean Francois Lafleche, Towaki Takikawa, Jiehan Wang, Charles Loop, Wenzheng Chen, Krishna Murthy Jatavallabhula, Edward Smith, Artem Rozantsev, Or Perel, Tianchang Shen, Jun Gao, Sanja Fidler, Gavriel State, Jason Gorski, Tommy Xiang, Jianing Li, Michael Li, and Rev Lebedev. Kaolin: A pytorch library for accelerating 3d deep learning research. <https://github.com/NVIDIAGameWorks/kaolin>, 2022. 2
- [23] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017. 2
- [24] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2019. 2
- [25] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, light, and material decomposition from images using monte carlo rendering and denoising. *Advances in Neural Information Processing Systems*, 35:22856–22869, 2022. 1, 2, 8
- [26] Zexin He and Tengfei Wang. Openlrn: Open-source large reconstruction models. <https://github.com/3DTopia/OpenLRM>, 2023. 2, 3
- [27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3
- [28] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2, 3, 1
- [29] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 3
- [30] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussian-shader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5322–5332, 2024. 1, 2
- [31] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensor: Tensorial inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2023. 1, 2
- [32] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022. 3
- [33] 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, 2, 5
- [34] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12912–12921, 2022. 3
- [35] Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Elliott Wu, Jiajun Wu, et al. Stanford-orb: a real-world 3d object inverse rendering benchmark. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 7, 8, 1, 4
- [36] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 2, 3, 4, 1
- [37] Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: A general nerf acceleration toolbox. *arXiv preprint arXiv:2210.04847*, 2022. 6, 1
- [38] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 2
- [39] Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Transactions on Graphics (ToG)*, 36(4):1–11, 2017. 2
- [40] 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. 2
- [41] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: Svbrdf acquisition with a single mobile phone image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 72–87, 2018. 2
- [42] 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. 2
- [43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020. 2
- [44] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhao Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, et al. Openrooms: An end-to-end open framework for photorealistic indoor scene datasets. *arXiv preprint arXiv:2007.12868*, 2020. 2

- [45] 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. In *European Conference on Computer Vision*, pages 555–572. Springer, 2022. 2
- [46] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 6
- [47] 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, 2
- [48] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 3
- [49] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10083, 2024. 3
- [50] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 3
- [51] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7708–7717, 2019. 2
- [52] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3
- [53] 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
- [54] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 1, 2
- [55] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, pages 210–227. Springer, 2022. 3
- [56] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 3
- [57] Stephen R Marschner, Stephen H Westin, Eric PF Lafor-tune, Kenneth E Torrance, and Donald P Greenberg. Image-based brdf measurement including human skin. In *Rendering Techniques’ 99: Proceedings of the Eurographics Workshop in Granada, Spain, June 21–23, 1999* 10, pages 131–144. Springer, 1999. 1
- [58] Wojciech Matusik. *A data-driven reflectance model*. PhD thesis, Massachusetts Institute of Technology, 2003. 1
- [59] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. Lime: Live intrinsic material estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6315–6324, 2018. 2
- [60] 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, 2
- [61] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 1, 2
- [62] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 1, 2
- [63] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 3
- [64] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3
- [65] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3
- [66] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 3
- [67] Ruoxi Shi, Xinyue Wei, Cheng Wang, and Hao Su. Zerorf: Fast sparse view 360deg reconstruction with zero pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21114–21124, 2024. 3
- [68] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3

- [69] Yahao Shi, Yanmin Wu, Chenming Wu, Xing Liu, Chen Zhao, Haocheng Feng, Jingtuo Liu, Liangjun Zhang, Jian Zhang, Bin Zhou, et al. Gir: 3d gaussian inverse rendering for relightable scene factorization. *arXiv preprint arXiv:2312.05133*, 2023. 1, 2
- [70] Yawar Siddiqui, Tom Monnier, Filippas Kokkinos, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, et al. Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. *arXiv preprint arXiv:2407.02445*, 2024. 3, 7, 8
- [71] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 3
- [72] Pratul P Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2020. 2
- [73] Cheng Sun, Guangyan Cai, Zhengqin Li, Kai Yan, Cheng Zhang, Carl Marshall, Jia-Bin Huang, Shuang Zhao, and Zhao Dong. Neural-pbr reconstruction of shape, material, and illumination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18046–18056, 2023. 1, 2, 8, 4
- [74] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdifffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 175–191. Springer, 2025. 3
- [75] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdifffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 175–191. Springer, 2025. 3
- [76] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2, 3
- [77] Benjamin Ummenhofer, Sanskar Agrawal, Rene Sepulveda, Yixing Lao, Kai Zhang, Tianhang Cheng, Stephan Richter, Shenlong Wang, and German Ros. Objects with lighting: A real-world dataset for evaluating reconstruction and rendering for object relighting. In *2024 International Conference on 3D Vision (3DV)*, pages 137–147. IEEE, 2024. 2
- [78] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 4
- [79] Peihao Wang, Zhiwen Fan, Dejia Xu, Dilin Wang, Sreyas Mohan, Forrest Iandola, Rakesh Ranjan, Yilei Li, Qiang Liu, Zhangyang Wang, et al. Steindreamer: Variance reduction for text-to-3d score distillation via stein identity. *arXiv preprint arXiv:2401.00604*, 2023. 3
- [80] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 2, 3
- [81] Peihao Wang, Dejia Xu, Zhiwen Fan, Dilin Wang, Sreyas Mohan, Forrest Iandola, Rakesh Ranjan, Yilei Li, Qiang Liu, Zhangyang Wang, et al. Taming mode collapse in score distillation for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9037–9047, 2024. 3
- [82] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021. 3
- [83] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12538–12547, 2021. 2
- [84] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [85] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024. 2, 3, 4, 5, 6, 7, 8, 1
- [86] Liwen Wu, Sai Bi, Zexiang Xu, Fujun Luan, Kai Zhang, Iliyan Georgiev, Kalyan Sunkavalli, and Ravi Ramamoorthi. Neural directional encoding for efficient and accurate view-dependent appearance modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21157–21166, 2024. 2, 5
- [87] Tong Wu, Zhibing Li, Shuai Yang, Pan Zhang, Xingang Pan, Jiaqi Wang, Dahua Lin, and Ziwei Liu. Hyperdreamer: Hyper-realistic 3d content generation and editing from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 3
- [88] Chao Xu, Ang Li, Linghao Chen, Yulin Liu, Ruoxi Shi, Hao Su, and Minghua Liu. Sparp: Fast 3d object reconstruction and pose estimation from sparse views. *arXiv preprint arXiv:2408.10195*, 2024. 3
- [89] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 3, 4, 5, 1
- [90] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein,

- Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023. [2](#), [3](#)
- [91] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. [2](#)
- [92] Ziyi Yang, Yanzhen Chen, Xinyu Gao, Yazhen Yuan, Yu Wu, Xiaowei Zhou, and Xiaogang Jin. Sire-ir: Inverse rendering for brdf reconstruction with shadow and illumination removal in high-illumination scenes. *arXiv preprint arXiv:2310.13030*, 2023. [1](#), [2](#)
- [93] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. [4](#)
- [94] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenotrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. [5](#)
- [95] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. [3](#)
- [96] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. [3](#)
- [97] Cheng Zhang, Lifan Wu, Changxi Zheng, Ioannis Gkioulekas, Ravi Ramamoorthi, and Shuang Zhao. A differential theory of radiative transfer. *ACM Transactions on Graphics (TOG)*, 38(6):1–16, 2019. [2](#)
- [98] 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](#), [2](#)
- [99] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *European Conference on Computer Vision*, pages 717–733. Springer, 2022. [1](#)
- [100] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5565–5574, 2022. [1](#), [2](#)
- [101] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2025. [2](#), [3](#), [5](#), [6](#)
- [102] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. [6](#)
- [103] Tianyuan Zhang, Zhengfei Kuang, Haian Jin, Zexiang Xu, Sai Bi, Hao Tan, He Zhang, Yiwei Hu, Milos Hasan, William T Freeman, et al. Relitlm: Generative relightable radiance for large reconstruction models. *arXiv preprint arXiv:2410.06231*, 2024. [3](#)
- [104] 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 Transactions on Graphics (ToG)*, 40(6):1–18, 2021. [1](#), [2](#)
- [105] 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. [8](#)
- [106] Youjia Zhang, Teng Xu, Junqing Yu, Yuteng Ye, Yanqing Jing, Junle Wang, Jingyi Yu, and Wei Yang. Nemf: Inverse volume rendering with neural microflake field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22919–22929, 2023. [1](#), [2](#)