DL3DV-10K: A Large-Scale Scene Dataset for Deep Learning-based 3D Vision

Supplementary Material

1. Overview

Sec. 2 discusses the data acquisition standard and the *DL3DV-10K* data distribution. Sec. 3 discusses more details of our benchmark experiments, including experiment, training details and more qualitative results, and details of the generalizable NeRF experiment.

2. Data

2.1. Data acquisition

The scene coverage for video shooting is illustrated in Fig. 2. For real-world scenes, they encompass horizontal views (180° - 360°) from different heights. We capture scenes using 360° panoramic views when the scene is accessible and well-defined, typically encompassing a diameter that can be covered on foot within 30 to 45 secs. In instances where the rear view of the scene is obstructed by larger objects, such as larger buildings, we opt for a semi-circular view (exceeding 180°) to capture the scene. To enhance scene coverage, we record videos by traversing two circular or semi-circular paths. The first traversal is conducted at overhead height, while the second is performed at approximately waist height. In data process, we apply COLMAP to calculate the camera pose for frames in the scene. Camera pose data would be released along with RGB images.

2.2. Benchmark selection

We select 140 scene as NVS benchmark by balancing scene complexity indices. Scene complexity is categorized into 16 types based on bounded (*indoor*) vs. unbounded (*outdoor*) environment, high vs. low texture frequency (*low-* vs. *high-freq*), more vs. less reflection (*more-* vs. *less-ref*), and more vs. less transparency (*more-* vs. *less-transp*). Rarer combinations like *outdoor* scenes with *low-freq*, *more-ref*, and *more-transp*, or *outdoor* scenes with *low-freq*, *less-ref*, and *more-transp* features are ignored. Therefore, a uniform distribution across the other 14 scene complexity types is selected. From each, 10 samples are chosen from various POIs to ensure statistical representativeness.

2.3. Labeling

Reflection and transparency We manually annotate reflection and transparency indices to scenes by assessing the ratio of reflective (transparent) pixels and the duration of reflectivity (transparency) observed in the video. Fig. 1 presents the reflection labeling criteria. Transparency labeling follows the same rule.

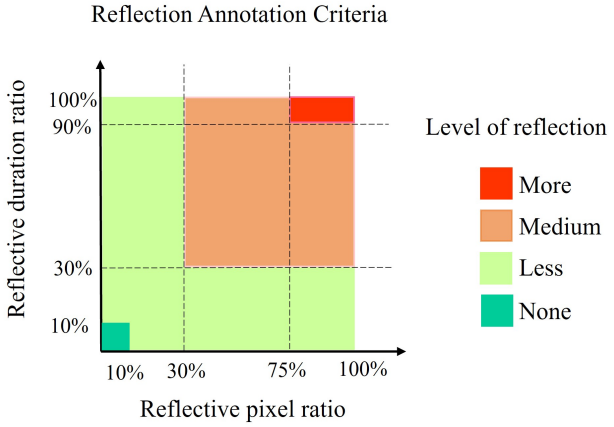

Figure 1. Reflection labeling criteria. Transparency annotation likewise.

2.4. Data Statistics

Scene summary by secondary POI category. The secondary POI categories are detailed classes within the primary POI categories. Fig. 3 shows scene statistics for each secondary POI category and the corresponding primary POI category. Fig. 4 presents scene statistics for each secondary POI category by complexity indices such as environmental setting, light condition, and level of reflection and transparency. For example, in the '*light condition*' attribute, we find that scenes from '*supermarkets*', '*shopping-malls*, and '*furniture-stores*' are mostly under artificial lighting, whereas '*hiking-trails* and '*parks-and-recreation-areas*' are under natural light. As for '*reflection*' and '*transparency*' attributes, '*shopping-malls*' are more likely to feature fully reflective scenes than other locations, while nature & outdoor scenes such as '*hiking-trails*' are predominantly non-reflective scenes. Most scenes are non-transparent. These observations align well with common expectations in real-world scenarios.

Frequency and duration estimates. The kernel density distribution of frequency metric and video duration can be found in Fig. 5. The frequency classes are delineated based on the median value of the frequency metric.

3. Experiment

3.1. NVS benchmark

Experiment Details The implementation of Nerfacto and Instant-NGP is from nerfstudio [8]. MipNeRF360 [1] and 3D gaussian splatting (3DGS) [3] codes are from the authors. ZipNeRF [2] source code is not public yet when we

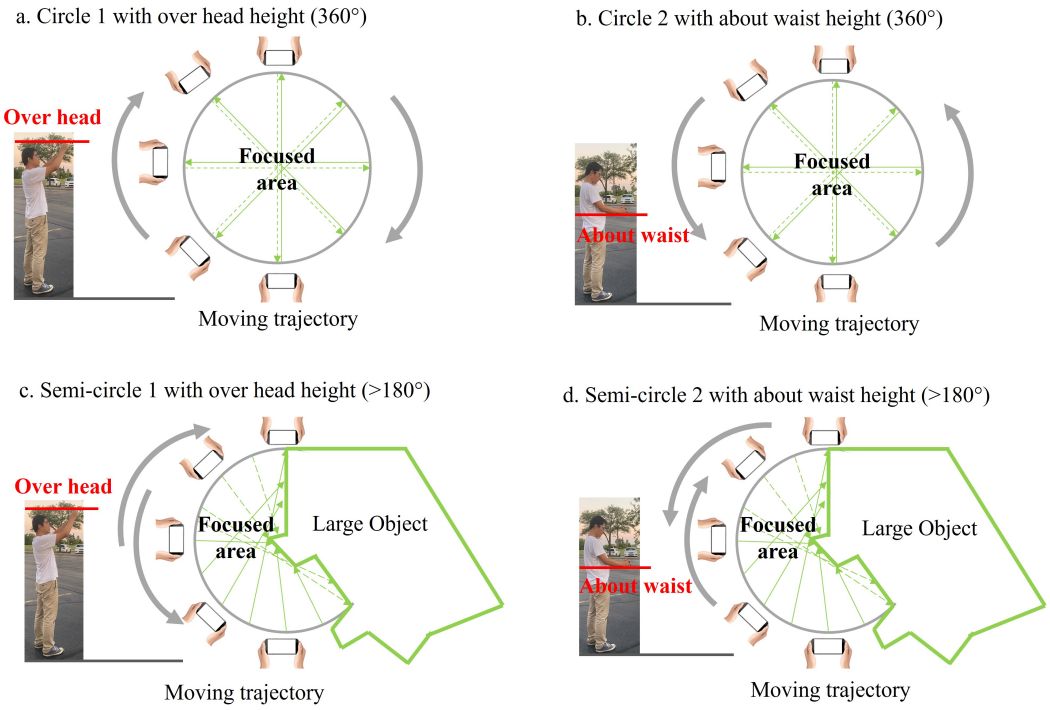

Figure 2. Video shooting examples with different heights and angles.

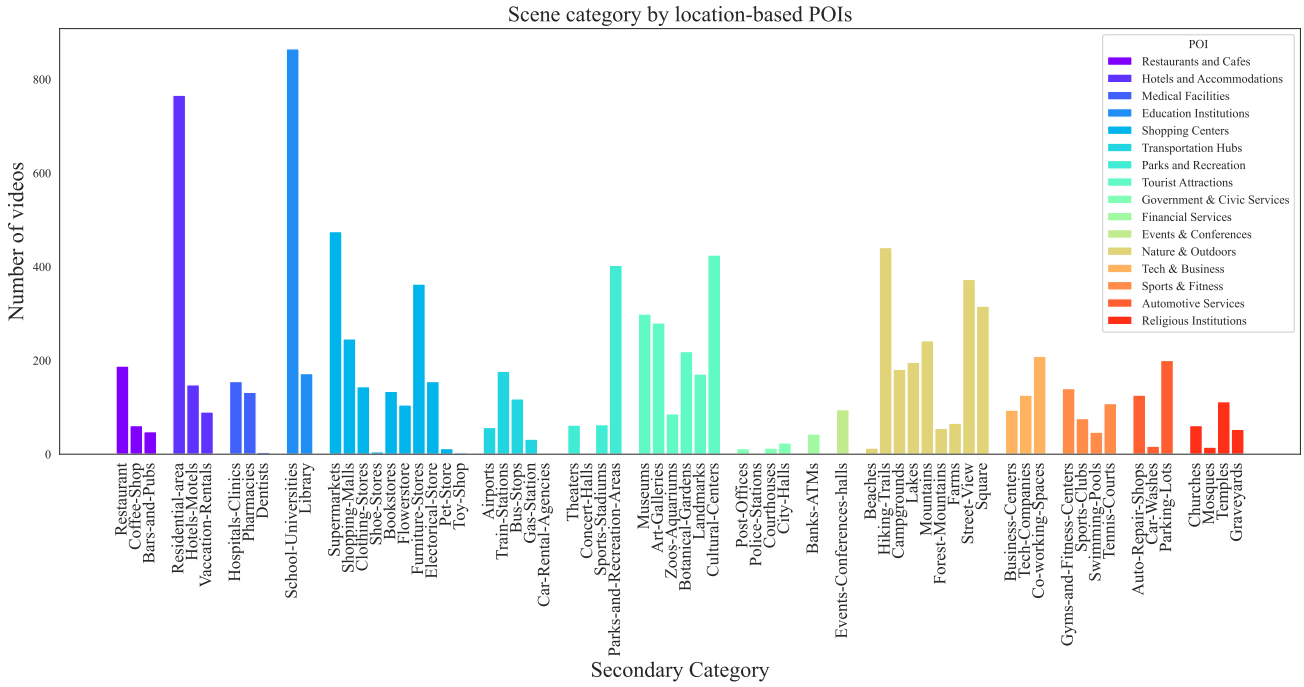

Figure 3. Number of scenes within secondary POI category. The legend contains the mapping between the primary and secondary POI categories. We observe that '*schools-universities* and '*residential-area*' are the predominant scenes in our *DL3DV-10K* dataset. In contrast, locations such as government and civic service facilities (e.g., '*post office*, '*police station*, '*court house*, and '*city hall*) are less frequently captured due to the challenges in accessing these areas for detailed video recording.

submit the paper. We used a public implementation [7] that shows the same performance results reported in the paper to test ZipNeRF.

The default ray batch is 4096. ZipNeRF is sensitive to this parameter, and we also showed 65536 (default by ZipNeRF) results. Nerfacto, Instant-NGP, ZipNeRF used half-precision fp16 while 3DGS and MipNeRF360 use full precision. All the NeRF-based methods use the same near (0.01) and the same far ($1e5$). The codes are run on A30, V100 and A100 GPUs depending on the memory they used. All the experiments took about 13,230 GPU hrs to finish.

More quantitative results. We present the performance of State-of-the-art (SOTAs) in *DL3DV-140* by scene primary POI categories in Fig. 6.

More visual results. We present more visual results for the performance of SOTAs on *DL3DV-140* by scene complexity indices. In particular, Fig. 7 describes the performance of SOTAs by environmental setting; Fig. 8 describes the performance of SOTAs by frequency; Fig. 9 describes the performance of SOTAs by transparency; and Fig. 11 describes the performance of SOTAs by reflection.

3.2. Generalizable NeRF

Experiment details We follow the default setting by IBRNet [9]. The training dataset includes LLFF [4], spaces, RealEstate10K [11] and self-collected small dataset by IBRNet authors. The evaluation dataset includes Diffuse Synthetic 360° [6], Realistic Synthetic 360° [5], part of LLFF that was not used in training. We used the official implementation. Each experiment was trained on single A100 GPU. Pretaining on Scannet++ and *DL3DV-10K* took 24 hrs. We present the sample of visual results for IBRNet experiment on Fig 10.

Besides, we report the additional IBRNet experiment of using Realistic Synthetic datase to evalcuate the potential of *DL3DV-10K*. The results can be found in Tab 1.

Method	Realistic Synthetic 360° [5] PSNR↑	SSIM↑	LPIPS↓
IBRNet	23.95	0.906	0.101
IBRNet-S	23.57	0.905	0.101
IBRNet-270	24.55	0.911	0.097
IBRNet-1K	23.58	0.906	0.102
IBRNet-2K	24.98	0.913	0.095

Table 1. IBRNet is trained from scratch, IBRNet-S, IBRNet-270, IBRNet-1K, and IBRNet-2K are IBRNet pred-trained on ScenNet++(270), DL3DV-270, DL3DV-1K, and DL3DV-2K.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 1

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *arXiv preprint arXiv:2304.06706*, 2023. 1

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 1

[4] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 3

[5] 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. 3

[6] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[7] SuLvXiangXin. zipnerf-pytorch. https://github.com/SuLvXiangXin/zipnerf-pytorch, 2023. 3

[8] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023. 1

[9] 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

[10] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 7

[11] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 3

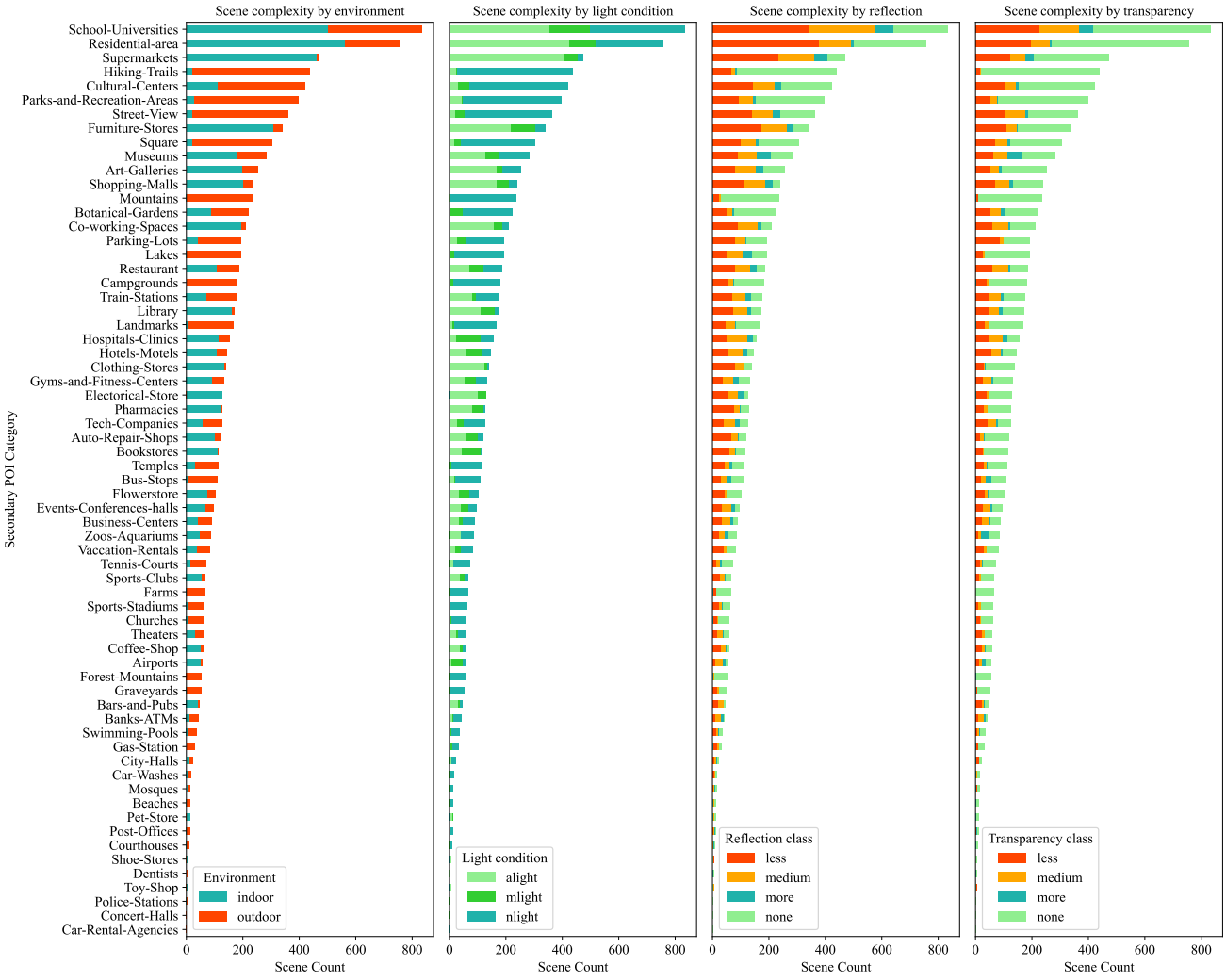

Figure 4. We show the distribution of scenes captured in secondary POI categories by complexities, including environmental setting, light conditions, reflective surfaces, and transparent materials.

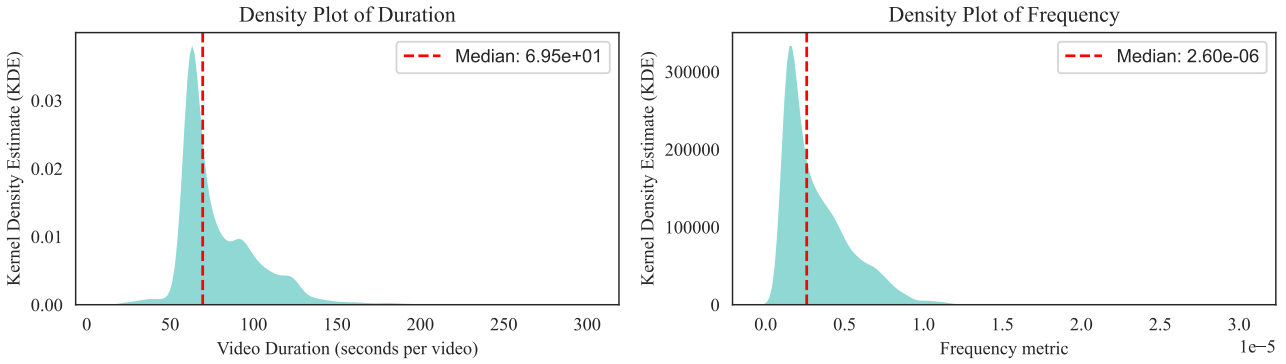

Figure 5. We show the distribution of video duration and frequency metric in **10,510** videos. The minimum duration for video shooting with consumer mobile devices is set at 60 secs, while for drone cameras, it's at least 45 secs. In our dataset, the median video duration is 69.5 secs. Furthermore, the median value of the frequency metric, determined by the average image intensity, stands at 2.6e-06. Based on this median value, we categorize scenes into high frequency ('*high_freq*') and low frequency ('*low_freq*') classes.

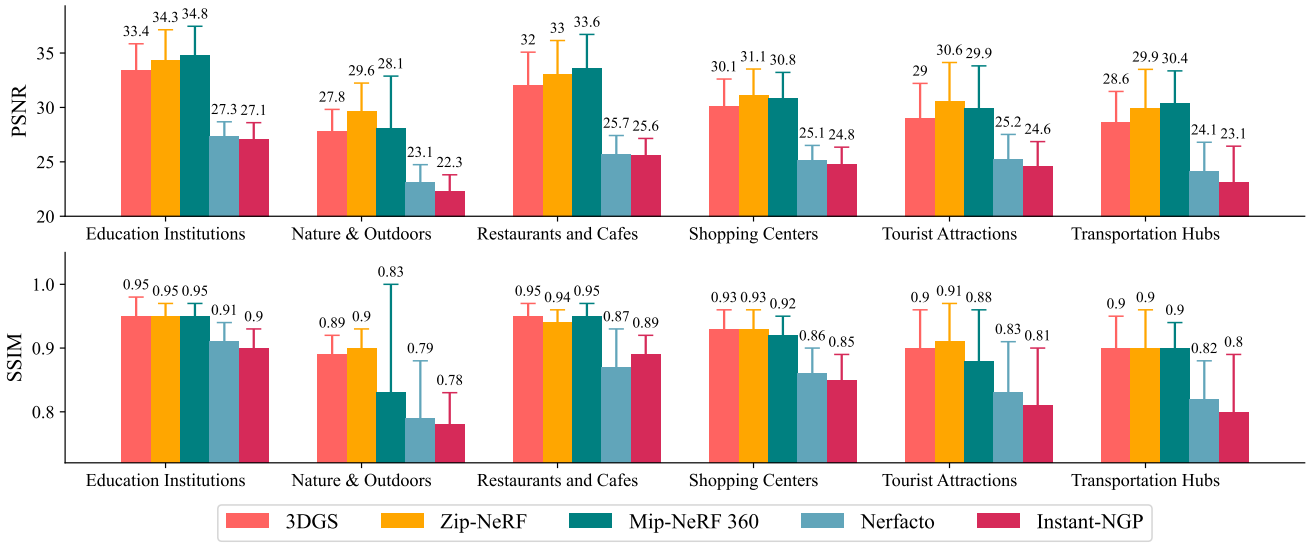

Figure 6. We present the average performance on 6 primary POI categories (*Education institutions, Nature & Outdoors, Restaurants and Cafes, Shopping Centers, Tourist Attractions, and Transportation Hubs*) in the *DL3DV-140*. The text above the bar plot is the mean value of the methods on the primary POI categories. As shown in the figure, NVS methods have better performance on scenes captured in *Education institutions*, *Restaurants and Cafes*, *Shopping Centers* than *Tourist Attractions*, *Transportation Hubs*, and *Nature & Outdoors*. Because majority scenes in *Education institutions*, *Restaurants and Cafes*, and *Shopping Centers* are indoor scenes. Additionally, the performance on *Shopping Centers* is worse than *Education institutions* and *Restaurants and Cafes*.

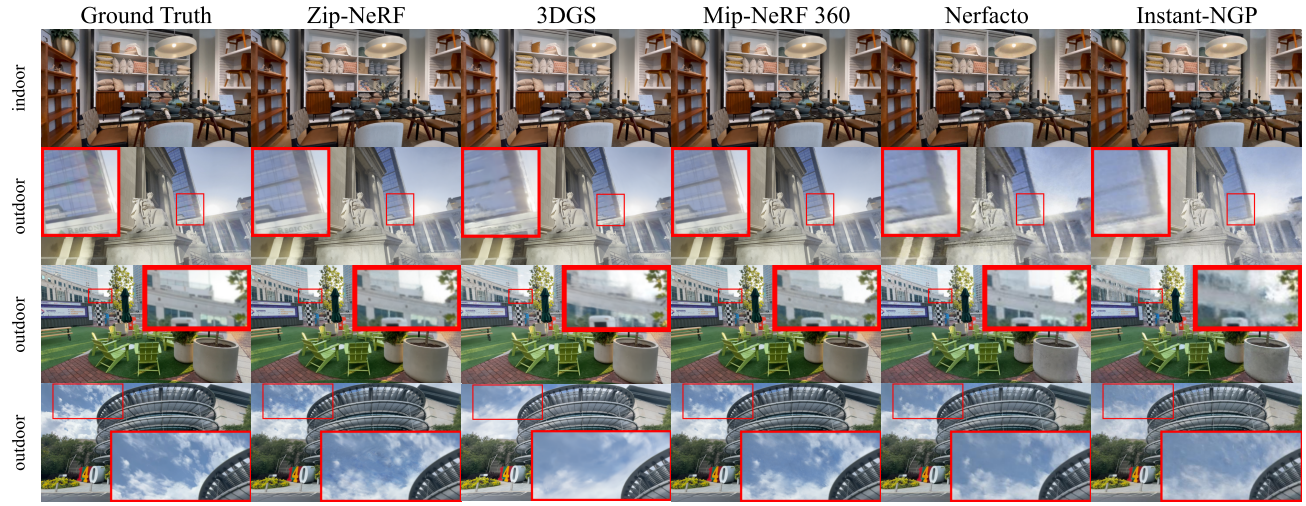

Figure 7. We compare the SOTAs for indoor (bounded) and outdoor (unbounded) environments on *DL3DV-140* from held-out test views. As illustrated in the figure, indoor scenes pose fewer challenges compared to outdoor scenes, where SOTAs demonstrate varying levels of performance. We observe that outdoor scene is more challenging for 3D Gaussian Splatting (3DGS), Nerfacto, and Instant-NGP than Zip-NeRF and Mip-NeRF 360.

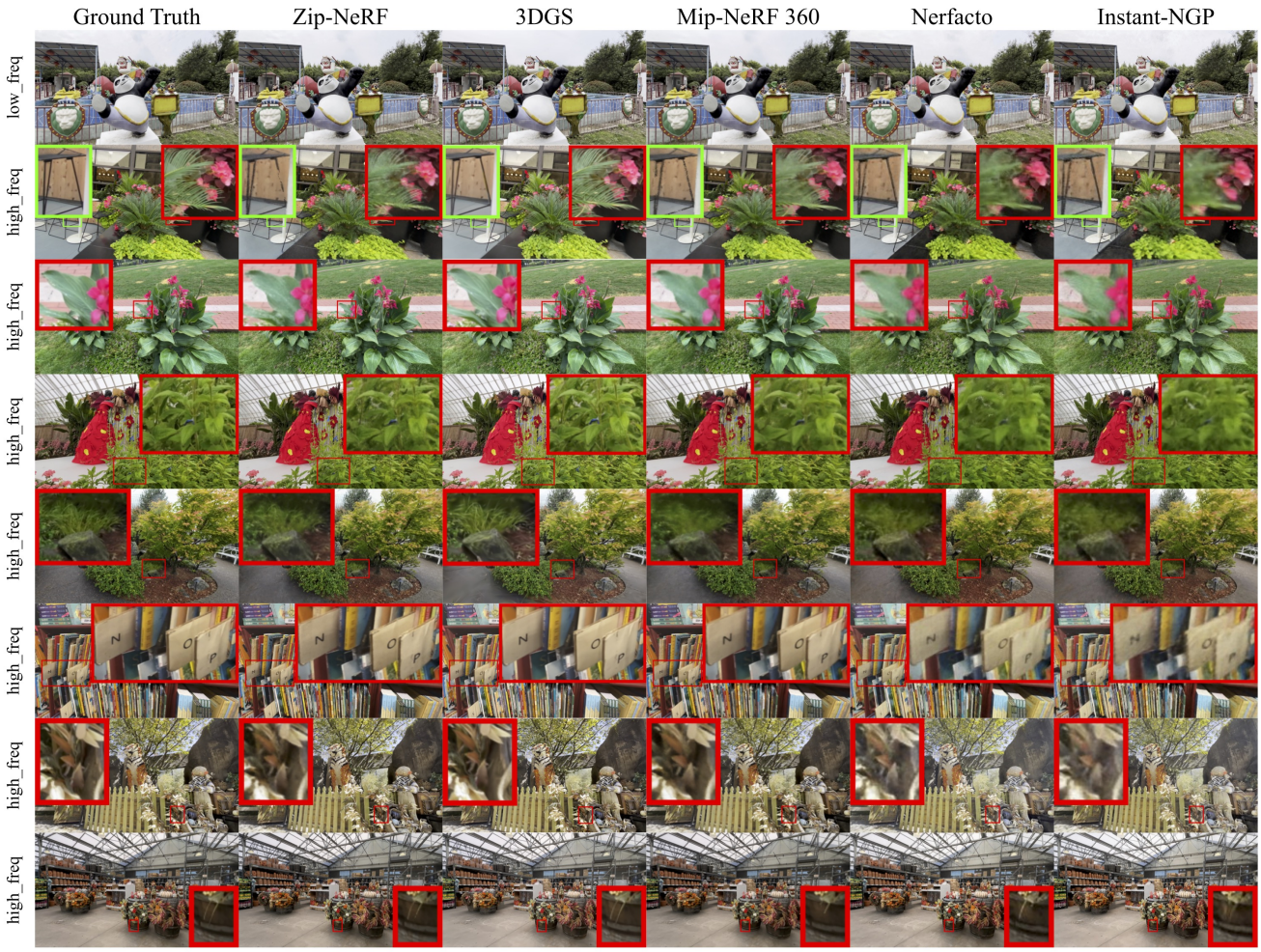

Figure 8. We compare the performance of SOTAs in frequency (low_freq vs. $high\_freq$) on *DL3DV-140* from held-out test views. As shown in the figure, high frequency (*high_freq*) scene is more challenging than low frequency (*low_freq*) scene. We observe that 3DGS consistently captures scenes with high-frequent details and renders the shape edge for the scene details. As for NeRF variants, it is more challenging for Nerfacto and Instant-NGP to handle scenes with high-frequent details than Zip-NeRF and Mip-NeRF 360. Besides, NeRF variants suffer aliasing issues.

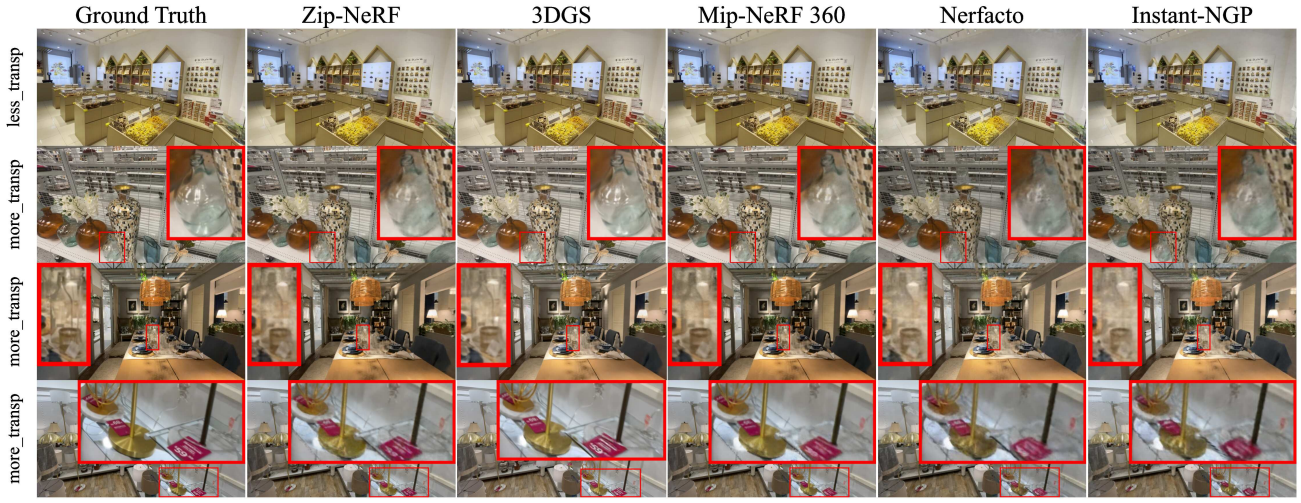

Figure 9. We compare the performance of SOTAs for transparency classes ($less\_transp$ vs. $more\_transp$) on *DL3DV-140* from held-out test views. As shown in the figure, scenes with more transparent materials (*more_transp*) are more challenging than scenes with less transparent materials (*less_transp*). In our analysis of the selected scenes, we noted that 3DGS, Zip-NeRF, and Mip-NeRF 360 effectively capture the subtle edges of transparent objects. Conversely, Nerfacto and Instant-NGP tend to consistently generate artifacts.

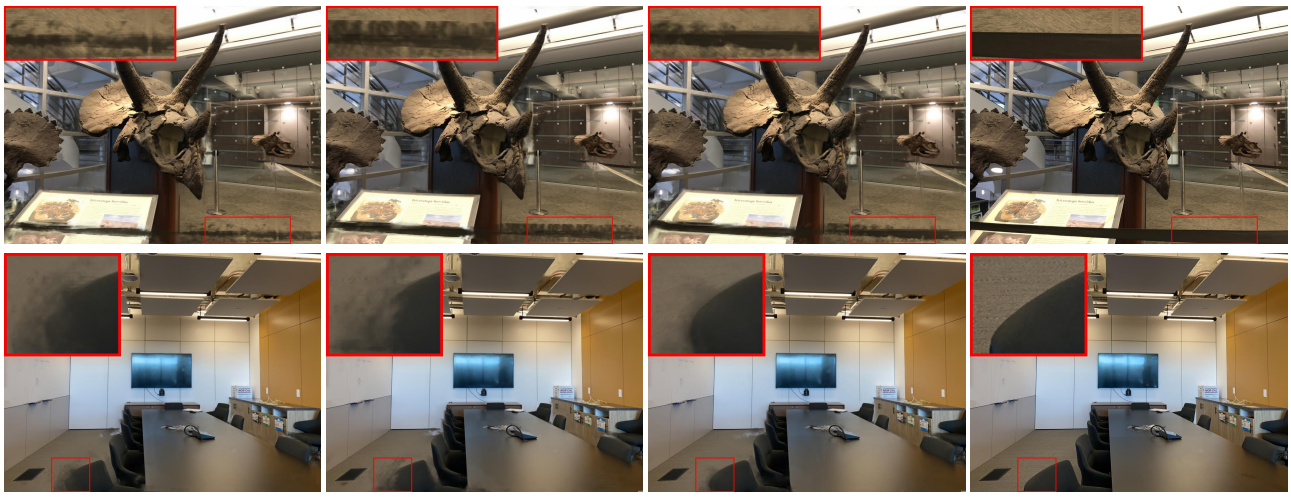

Figure 10. **More qualitative results for generalizable NeRF.** IBRNet-S: pretrain IBRNet on Scannet++ [10]. IBRNet-D: pretrain IBRNet on DL3DV-2K. Priors learned from DL3DV-2K help IBRNet perform the best on the evaluation.

Figure 11. We compare the SOTAs for reflection classes ($less\_ref$ vs. $more\_ref$) on *DL3DV-140* from held-out test views. As shown in the figure, scenes with more reflective surfaces (*more_ref*) are more challenging than scenes with less reflective surfaces (*less_ref*). Among SOTAs, Zip-NeRF and Mip-NeRF 360 are adept at capturing subtle reflections and highlights. On the other hand, 3DGS tends to overly smooth out less intense reflections. Nerfacto and Instant-NGP struggle to effectively manage scenes with highly reflective surfaces, often resulting in the generation of artifacts.