

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

Lu Ling¹, Yichen Sheng¹, Zhi Tu¹, Wentian Zhao², Cheng Xin³, Kun Wan², Lantao Yu², Qianyu Guo¹, Zixun Yu⁴, Yawen Lu¹, Xuanmao Li⁵, Xingpeng Sun¹, Rohan Ashok¹, Aniruddha Mukherjee¹, Hao Kang⁶, Xiangrui Kong¹, Gang Hua⁶, Tianyi Zhang¹, Bedrich Benes¹, Aniket Bera¹

¹ Department of Computer Science, Purdue University, ² Adobe Inc., ³ Rutgers University

⁴ Google Inc., ⁵ Huazhong University of Science and Technology, ⁶ Wormpex AI Research



Figure 1. We introduce **DL3DV-10K**, a large-scale, scene dataset capturing real-world scenarios. **DL3DV-10K** contains **10,510** videos at 4K resolution spanning 65 types of point-of-interest (POI) locations, covering a wide range of everyday areas. With the fine-grained annotation on scene diversity and complexity, **DL3DV-10K** enables a comprehensive benchmark for novel view synthesis and supports learning-based 3D representation techniques in acquiring a universal prior at scale.

Abstract

We have witnessed significant progress in deep learning-based 3D vision, ranging from neural radiance field (NeRF) based 3D representation learning to applications in novel view synthesis (NVS). However, existing scene-level datasets for deep learning-based 3D vision, limited to either synthetic environments or a narrow selection of real-world scenes, are quite insufficient. This insufficiency not only hinders a comprehensive benchmark of existing methods but also caps what could be explored in deep learning-based 3D analysis. To address this critical gap, we present **DL3DV-10K**, a large-scale scene dataset, featuring **51.2 million frames** from **10,510** videos captured from **65** types of point-of-interest (POI) locations, covering both bounded and unbounded scenes, with different levels of reflection, transparency, and lighting. We conducted a comprehensive benchmark of recent NVS methods on **DL3DV-10K**, which revealed valuable insights for future research in NVS. In

addition, we have obtained encouraging results in a pilot study to learn generalizable NeRF from **DL3DV-10K**, which manifests the necessity of a large-scale scene-level dataset to forge a path toward a foundation model for learning 3D representation. Our **DL3DV-10K** dataset, benchmark results, and models will be publicly accessible at [Project Page](#).

1. Introduction

The evolution in deep 3D representation learning, driven by essential datasets, boosts various tasks in 3D vision. Notably, the inception of Neural Radiance Fields [19] (NeRF), offering a new approach through a continuous high-dimensional neural network, revolutionized learning-based 3D representation and novel view synthesis (NVS). NeRF excels at producing detailed and realistic views, overcoming challenges faced by traditional 3D reconstruction methods and rendering techniques. Additionally, it in-

spires waves of innovative developments such as NeRF variants [3, 4, 8, 27, 32, 37] and 3D Gaussian Splatting (3DGS) [15], significantly enhancing experiences in virtual reality, augmented reality, and advanced simulations.

However, existing scene-level datasets for NVS are restricted to either synthetic environments or a narrow selection of real-world scenes due to the laborious work for scene collection. Notably, the absence of such large-scale scene datasets hinders the potential of deep 3D representation learning methods in two pivotal aspects: 1) it is impossible to conduct a comprehensive benchmark to adequately assess NVS methods in complex real-world scenarios such as non-Lambertian surfaces. 2) It restricts the generalizability of deep 3D representation learning methods on attaining universal priors from substantial real scenes.

To fill this gap, we revisited the commonly used dataset for benchmarking NVS. i) Synthetic datasets like blender dataset [19] offer rich 3D CAD geometry but lack real-world elements, which diminishes the model’s robustness in practical applications. ii) Real-world scene datasets for NVS such as Tank and temples [16] and LLFF dataset [18] offer more variety but limited scope. They fail to capture the complex real-world scenarios such as intricate lighting and various material properties (transparency and reflectance [22]), which still challenges current SOTAs.

Moreover, existing 3D representation methods yield photorealistic views by independently optimizing NeRFs for each scene, requiring numerous calibrated views and substantial compute time. Learning-based models like PixelNeRF [38] and IBRNet [30] mitigate this by training across multiple scenes to learn scene priors. While datasets like RealEstate [42] and ScanNet++ [36] improve understanding in specific domains such as indoor layouts, their limited scope hinders the broader applicability of these models. This is primarily due to the absence of comprehensive scene-level datasets, which are crucial for learning-based methods to achieve universal representation.

Based on the above review, We introduce *DL3DV-10K* a novel dataset that captures large-scale (multi-view) MV scenes using standard commercial cameras to enable efficient collection of a substantial variety of real-world scenarios. *DL3DV-10K* comprises **51.3** million frames from **10,510** videos in 4k resolution, covers scenes from 65 types of point-of-interest [35] (POI) locations like restaurants, tourist spots, shopping malls, and natural outdoor areas. Each scene is further annotated with its complexity indices, including indoor or outdoor environments, the level of reflection and transparency, lighting conditions, and the level of texture frequency. Fig. 1 provides a glimpse of our *DL3DV-10K* dataset. Tab. 1 compares scale, quality, diversity, and annotated complexity between *DL3DV-10K* and the existing scene-level datasets.

Additionally, we present *DL3DV-140*, a comprehensive

benchmark for NVS, by sampling 140 scenes from the dataset. The diversity and fine-grained scene complexity in *DL3DV-140* will enable a fair evaluation of NVS methods. We conducted the statistical evaluation of the SOTA NVS methods on *DL3DV-140* (Sec. 4.1), including Nerfacto [27], Instant-NGP [20], Mip-NeRF 360 [3], Zip-NeRF [4], and 3DGS [15]. Leveraging the multi-view nature of the data, we attempt to showcase *DL3DV-10K*’s potential for deep 3D representation learning methods in gaining a universal prior for generating novel views. Our demonstrations indicate that pretraining on *DL3DV-10K* enhances the generalizability of NeRF (Sec. 4.2), confirming that diversity and scale are crucial for learning a universal scene prior.

To summarize, we present the following contributions:

1. We introduce *DL3DV-10K*, a real-world MV scene dataset. It has 4K resolution RGB images and covers 65 POI indoor and outdoor locations. Each scene is annotated with the POI category, light condition, environment setting, varying reflection and transparency, and the level of texture frequency.
2. We provide *DL3DV-140*, a comprehensive benchmark with 140 scenes covering the challenging real-world scenarios for NVS methods. We conduct the statistical evaluation for the SOTA NVS methods on *DL3DV-140* and compare their weaknesses and strengths.
3. We show that the pre-training on *DL3DV-10K* benefits generalizable NeRF to attain universal scene prior and shared knowledge.

2. Related Work

2.1. Novel View Synthesis

Novel View Synthesis Methods. Early novel view synthesis (NVS) work concentrated on 3D geometry and image-based rendering [17, 42]. Since 2020, Neural Radiance Fields (NeRF) [19] have been pivotal in NVS for their intricate scene representation, converting 5D coordinates to color and density, leading to various advancements in rendering speed and producing photorealistic views. Key developments include Instant-NGP [20], which speeds up NeRF using hash tables and multi-resolution grids; Mip-NeRF [2] addressed aliasing and Mip-NeRF 360 [3] expanded this to unbounded scenes with significant computational power; and Zip-NeRF [4] combines Mip-NeRF 360 with grid-based models for improving efficiency. Nerfacto [27] merges elements from previous NeRF methods to balance speed and quality. Additionally, 3D Gaussian Splatting (3DGS) [15] uses Gaussian functions for real-time, high-quality rendering.

However, those SOTA methods, building neural radiance fields for individual scenes, require dense views and extensive computation. Learning-based models like ContraNeRF [33], TransNeRF [29], PixelNeRF [38], and IBR-

Dataset	# of scene	# of POI category	Resolution	# of frames	Scene complexity annotation			NVS
					indoor/outdoor	reflection	transparency	
LLFF [18]	24	-	640 × 480	<1K	✓✓	✓	✗	✓
DTU [14]	124	5	1200 × 1600	30K	✓✗	✓	✗	✓
BlendedMVS [34]	113	-	1536 × 2048	17K	✓✓	✗	✗	✗
ScanNet [11]	1513	11	1296 × 968	2,500K	✓✗	✗	✗	✗
Matterport3D [6]	90 ¹	-	1280 × 1024	195K	✓✗	✓	✗	✗
Tanks and Temples [16]	21	14	3840 × 2160	147K	✓✓	✓	✗	✓
ETH3D [24]	25	11	6048 × 4032	<1K	✓✓	✗	✗	✗
RealEstate10K ² [42]	10,000	1	1280 × 720	10,000K	✓✗	✗	✗	✓
ARKitScenes [5]	1661	1	1920 × 1440	450K	✓✗	✗	✗	✗
ScanNet++ [36]	460	5	7008 × 4672 ³	3,980K	✓✗	✓	✗	✓
<i>DL3DV-10K</i> (ours)	10,510	65	3840 × 2160	51,200K	✓✓	✓	✓	✓

Table 1. Comparison of the existing scene-level dataset in terms of quantity, quality, diversity, and complexity, which is measured by the fine-grained surface properties, light conditions, texture frequency, and environmental setting.

Net [30] overcome this by training on numerous scenes for universal priors and sparse-view synthesis. Yet, the absence of large-scale scene-level datasets reflective of real-world diversity fails to provide adequate assessment for the SOTAs. Additionally, it hinders the potential for learning-based 3D models to gain a universal prior.

Novel View Synthesis Benchmarks. NVS benchmarks are generally split into synthetic benchmarks like the NeRF-synthetic (Blender) [19], ShapeNet [7] and Objaverse [12], featuring 3D CAD models with varied textures and complex geometries but lacking real-world hints such as noise and non-Lambertian effects.

In contrast, real-world NVS benchmarks, originally introduced for multi-view stereo (MVS) tasks like DTU [14] and Tanks and Temples [16], offer limited variety. While ScanNet [11] has been used for benchmarking NVS, its motion blur and narrow field-of-view limit its effectiveness. Later benchmarks for outward- and forward-facing scenes emerged, but they vary in collection standards and lack diversity, particularly in challenging scenarios like lighting effects on reflective surfaces. For example, LLFF [18] offers cellphone-captured 24 forward-facing scenes; Mip-NeRF 360 dataset [19] provides 9 indoor and outdoor scenes with uniform distance around central subjects; Nerf-studio dataset [27] proposes 10 captures from both phone and mirrorless cameras with different lenses.

Inspired by the capture styles of these datasets, *DL3DV-140* provides a variety of scenes to comprehensively evaluate NVS methods, including challenging view-dependent effects, reflective and transparent materials, outdoor (unbounded) environments, and high-frequency textures. We also offer extensive comparative analyses, demonstrating the efficacy of *DL3DV-140* in assessing NVS techniques.

2.2. Multi-view Scene Dataset

Multi-view (MV) datasets are commonly used for NVS tasks in the 3D vision field. These datasets range from synthetic, like ShapeNet [7] and Pix2Vox++ [31], to foundational real-world datasets capturing both object-level and scene-level images.

Object-level datasets like Objectron [1], CO3D [21], and MVimgnet [39] offer substantial scale for learning-based reconstruction. While they facilitate the learning of spatial regularities, reliance solely on object-level MV datasets impedes prediction performance on unseen objects and complex scenes containing multiple objects [38].

Scene-level datasets are essential for NVS and scene understanding, yet offerings like LLFF [18], Matterport3D [6], and BlendedMVS [34] encompass limited scenes. DTU [14], despite its use in developing generalized NeRF [38], is limited by its scale, affecting the models’ ability to generalize. RealEstate10k [42], ARKitScenes [5], ScanNet [11], and the high-resolution ScanNet++ [36] improve this with a broad range of detailed indoor scenes. Yet, their applicability remains less comprehensive for diverse indoor settings like shopping centers and restaurants, or outdoor scenarios. Although RealEstate10k, focusing on YouTube real estate videos, provides comparable scale with us, they comprise low resolution and lack of diversity. Overall, the limited scale and diversity of these datasets pose challenges for the robust and universal training of 3D deep learning models. We close this gap by introducing *DL3DV-10K*, encompassing multifarious real-world scenes from indoor to outdoor environments, enhancing 3D spatial perception, and paving the way for more robust learning-based 3D models.

¹90 building-scale scenes covering 2056 rooms

²YouTube video

³7008 × 4672 in 270 scenes and 1920 × 1440 in 190 scenes



Figure 2. The efficient data acquisition pipeline of *DL3DV-10K*. Refer to *supplementary materials* for more visual illustrations of scene coverage.

3. Data Acquisition and Processing

Our data acquisition goal is to gather large-scale, high-quality scenes reflective of real-world complexity and diversity. We develop a pipeline that integrates video capture, pre-processing, and analysis, leveraging widely available consumer mobiles and drones to ensure the coverage of everyday accessible areas. We designed a detailed guideline to train the collectors to minimize motion blur, exclude exposure lights, and avoid moving objects. This collection process, illustrated in Fig. 2, is user-friendly and enhances both quality and efficiency.

3.1. Data Acquisition

Diversity. Our collection focuses on various commonly accessible scenes, adhering to the point-of-interest (POI) [35] categories. *DL3DV-10K* captures scenes from 16 primary and 65 secondary POI categories. These categories include varied settings such as educational institutions, tourist attractions, restaurants, medical facilities, transportation hubs, etc. The diversity, spanning from indoor to outdoor environments and different cities, is instrumental in enriching the dataset with a broader spectrum of cultural and architectural variations. Such variety is essential for benchmarking NVS techniques as it challenges and refines their ability to generalize across a multitude of real-world scenarios and fosters robustness and adaptability in 3D models. Sec. 3.3 summarizes the POI category of collected scenes.

Complexity. Real-world scenes challenge state-of-the-art techniques (SOTAs) with their diverse environments (indoor vs. outdoor), view-dependent lighting effects, reflective surfaces, material properties, and high-frequency textures. The high detail frequency in textures, from delicate fabric to coarse stone, requires meticulous rendering. Outdoor (unbounded) scenes, with varying details between near and distant objects, challenge the robustness of NVS methods in handling scale differences. Complex shadows and view-dependent highlights from natural and artificial

Category	Device		Quality by moving objects	
	Consumer mobile	Drone	<3s	3s - 10s
# of scene	10,407	103	8064	2446

Table 2. Number of scenes by devices and level of quality.

lights, interacting with reflective and transparent materials like metal and glass, require precise handling for realistic depiction. Additionally, we provide multiple views of the scene from various angles at different heights and distances. This multiplicity and complexity of views enable 3D methods to predict view-dependent surface properties. It also aids in separating viewpoint effects in learning view-invariant representations like patch descriptors [40] and normals [41].

Quality. The quality of the video is measured by the content and coverage density of the scene and video resolution. We have formulated the following requirements as guidelines for recording high-quality scene-level videos: 1) The scene coverage is in the circle or half-circle with a 30 secs-45 secs walking diameter and has at least five instances with a natural arrangement. 2) The default focal length of the camera corresponds to the 0.5x ultra-wide mode for capturing a wide range of background information. 3) Each video should encompass a horizontal view of at least 180° or 360° from different heights, including overhead and waist levels. It offers high-density views of objects within the coverage area. 4) The video resolution should be 4K and have 60 fps (or 30 fps). 5) The video's length should be at least 60 secs for mobile phone capture and 45 secs for drone video recording. 6) We recommend limiting the duration of moving objects in the video to under 3 secs, with a maximum allowance of 10 secs. 7) The frames should not be motion-blurred or overexposed. 8) The captured objects should be stereoscopic. Post-capture, we meticulously inspect videos based on the above criteria, such as moving objects in the video, to guarantee the high quality of videos captured on mobile devices. Tab. 2 summarizes the number of scenes recorded by different devices and the associated quality levels in terms of moving objects' duration in the videos.

3.2. Data Processing

Frequency Estimation. To estimate the frequency metric of the scene over the duration of the captured video, we first sample 100 frames from each video, as texture frequency is typically calculated based on RGB images. Then, we convert RGB images to grayscale and normalize the intensities. To extract the high-frequency energy, we apply a two-dimensional bi-orthogonal wavelet transform [10] and compute the Frobenius norm of the ensemble of LH, HL and HH subbands. The Frobenius norm is finally normalized by the number of pixels, and the average quotient over 100 frames is the frequency metric. The distribution of fre-

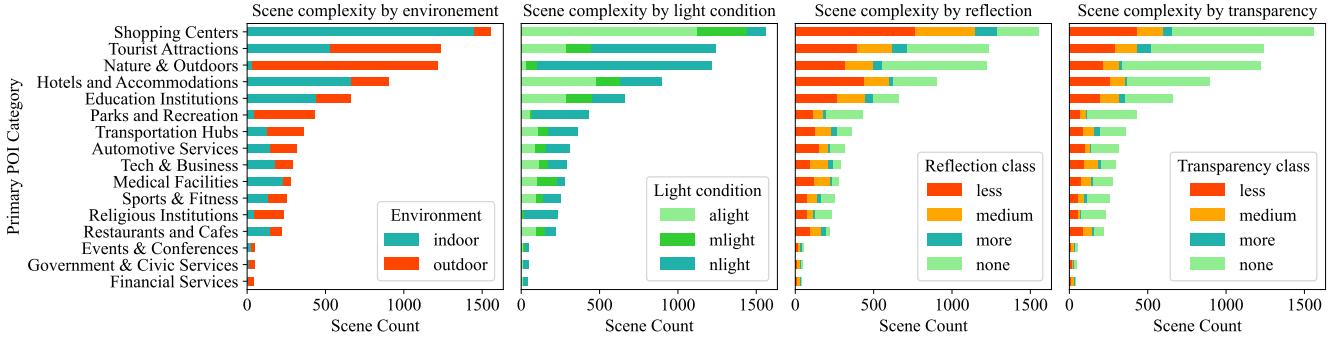


Figure 3. We show the distribution of scene category (the primary POI locations) by complexity indices, including environmental setting, light conditions, reflective surface, and transparent materials. Attributes in light conditions include: natural light (*'nlight'*), artificial light (*'alight'*), and a combination of both (*'mlight'*). Reflection class includes *'more'*, *'medium'*, *'less'*, and *'none'*. Transparency class likewise.

quency metrics is shown in *supplementary materials*.

Labeling. We categorize and annotate the diversity and complexity of scenes based on our established criteria, including key attributes of the scene, such as POI category, device model, lighting conditions, environmental setting, surface characteristics, and high-frequent textures. The POI category and device model depict the location where the video was collected and the equipment used during its capture. Lighting conditions are differentiated into natural, artificial, or a combination of both, influencing the ambient illumination of scenes. The environmental setting annotation distinguishes between indoor and outdoor settings, which is crucial for evaluating the NVS performance in both bounded and unbounded spaces. We measure the surface properties by the level of reflectivity, ranging from more and medium to less and none. It is estimated by the ratio of reflective pixels in the image and its present duration in the video. Material property, measured by transparency, follows a similar rule. Refer to *supplementary materials* for more details on reflection and transparency labeling criteria.

3.3. Data Statistics

Scale. *DL3DV-10K* aims to provide comprehensive and diverse coverage of scene-level datasets for 3D vision. It covers scenes collected from 16 primary POIs and 65 secondary POIs and comprises 51.3 million frames of **10,510** videos with 4K resolution. As shown in Tab. 1, it enjoys fine-grained annotation for scene complexity indices.

Hierarchy. We classify the scene category following the POI category taxonomy. For example, the primary POI category is the ‘*entertainment area*’. Its secondary POI category includes ‘*theaters*’, ‘*concert halls*’, ‘*sports stadiums*’, and ‘*parks and recreation areas*’. The statistics of primary POI-based scene categories by annotated complexity are presented in Fig. 3. The distribution of scenes captured in these POI locations follows: 1) their generality in nature.

2) the probability of no moving objects appearing within 60 sec in the locations. 3) the accessibility to these locations. For example, the government and civic services locations usually do not allow video shooting in high-level details. 4) the collected video is under the permission of the agents.

3.4. Benchmark

To comprehensively assess the SOTAs, the benchmark needs to cover the inherent complexity of real-world scenarios with varying reflectance, texture, and geometric properties. To achieve our goal, we present *DL3DV-140* as a comprehensive benchmark, sampling 140 static scenes from our dataset. Additionally, we simplified the reflection and transparency classes into two categories for better interpretation: more reflection (including more and medium reflection) and less reflection (including less and no reflection); the same approach applies to transparency. *DL3DV-140* with scenes collected from diverse POIs, maintains a balance in each annotated scene complexity indices. This means *DL3DV-140* are categorized as *indoor* (bounded) scenes vs. *outdoor* (unbounded) scenes, high vs. low texture frequency (*low-freq* vs. *high-freq*), more vs. less reflection (*more-ref* vs. *less-ref*), and more vs. less transparency (*more-transp* vs. *less-transp*). *DL3DV-140* offers challenging scenes with a rich mix of diversity and complexity for a comprehensive evaluation of existing SOTA methods.

4. Experiment

4.1. Evaluation on the NVS benchmark

Methods for comparison. We examine the current relevant state-of-the-art (SOTA) NVS methods on *DL3DV-140*, including NeRF variants such as Nerfacto [27], Instant-NGP [20], Mip-NeRF 360 [3], Zip-NeRF [4], and 3D Gaussian Splatting (3DGS) [15].

Experiment details. The SOTAs have different assumptions on the image resolution. For fairness, we use 960×560

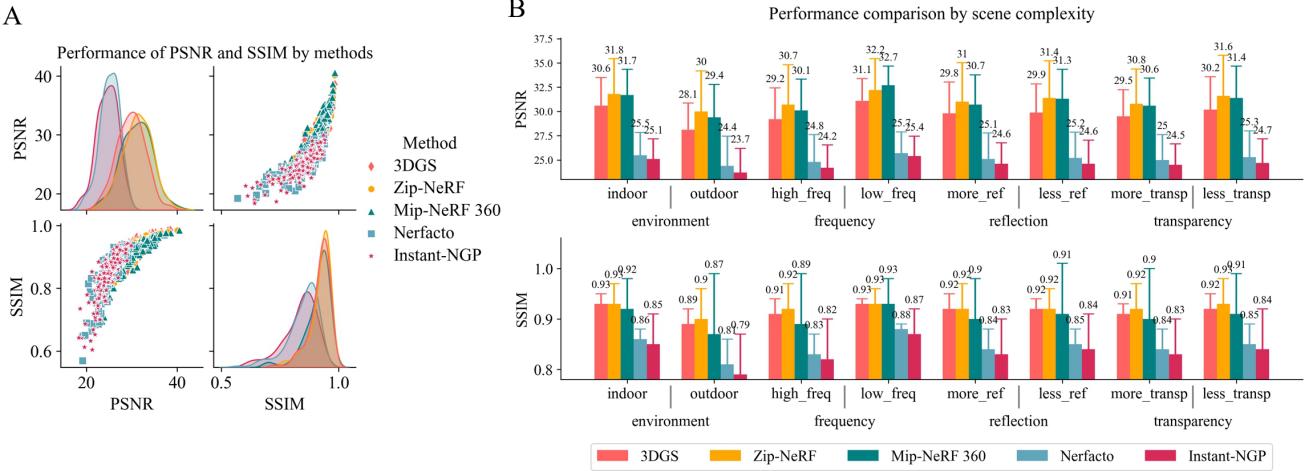


Figure 4. A presents the density plot of PSNR and SSIM and their relationship on *DL3DV-140* for each method. B describes the performance comparison by scene complexity. The text above the bar plot is the mean value of the methods on the attribute.

	PSNR ↑	SSIM ↑	LPIPS ↓	Train ↓	Mem ↓
Instant-NGP	25.01	0.834	0.228	1.2 hr	3.9GB
Nerfacto	24.61	0.848	0.211	2.6 hr	3.7GB
Mip-NeRF 360	30.98	0.911	0.132	48.0 hr	23.6GB
3DGS	29.82	0.919	0.120	2.1 hr	16.8GB
Zip-NeRF*	29.07	0.878	0.169	2.5 hr	23.8GB
Zip-NeRF	31.22	0.921	0.112	4.0 hr	38.2GB

Table 3. Performance on *DL3DV-140*. The error metric is calculated from the mean of 140 scenes on a scale factor of 4. Zip-NeRF uses the default batch size (65536) and Zip-NeRF* uses the identical batch size as other methods (4096). Note, the training time and memory usage may be different depending on various configurations. Refer to *supplementary materials* for details.

resolution to train and evaluate all the methods. Each scene in the benchmark has 300-380 images, depending on the scene size. We use 7/8 of the images for training and 1/8 of the images for testing. Different methods vary in their method configurations. We use most of the default settings, like network layers and size, optimizers, etc. But to ensure fairness, we fix some standard configurations. Each NeRF-based method (Nerfacto, Instant-NGP, Mip-NeRF 360, Zip-NeRF) has the same 4,096 ray samples per batch (equivalent to chunk size or batch size), the same near 0.05 and far 1e6. Note that Mip-NeRF 360 and Zip-NeRF use a much higher number of rays (65,536) per batch by default. We modify the learning rate to match the change of ray samples as suggested by the authors. We notice that Zip-NeRF performance is sensitive to the ray samples. So, we add one more experiment for Zip-NeRF with the same ray samples of 4,096 as other methods. For all methods, we train enough iterations until they converge.

Quantitative results. Tab. 3 summarizes the average PSNR, SSIM, and L-PIPS metrics across all scenes in

DL3DV-140 along with training hours and memory consumption for each method. Furthermore, Fig. 4 provides detailed insights into the metric density functions and their correlations.

The results indicate that Zip-NeRF, Mip-NeRF 360, and 3DGS consistently outperform Instant-NGP and Nerfacto across all evaluation metrics. Remarkably, Zip-NeRF demonstrates superior performance in terms of average PSNR and SSIM, although it consumes more GPU memory using the default batch size. Besides, we notice that reducing the default batch size for Zip-NeRF significantly decreases its PSNR, SSIM, and LPIPS, see Zip-NeRF* in Tab. 3. Mip-NeRF 360 achieves a PSNR of 30.98 and SSIM of 0.91, yet it shows relatively lower computational efficiency, with an average training time of 48 hours. The density functions of PSNR and SSIM, depicted in Fig. 4A, underscores Zip-NeRF and 3DGS's robust performance across all scenes. Moreover, we observe that 3DGS, with an SSIM of 0.92, surpasses Mip-NeRF 360's SSIM of 0.91, consistent with the findings from 3DGS's evaluation using the Mip-NeRF 360 dataset [15].

Fig. 4 B illustrates the performance across scene complexity indices. Among all indices, outdoor (unbounded) scenes appear to be the most challenging, as all methods yield the lowest PSNR and SSIM scores in this setting. Conversely, low-frequency scenes are the easiest to generate. Furthermore, more transparent scenes present higher challenges compared to less transparent ones. In terms of method comparison, Zip-NeRF outperforms others in most scenes, except in low-frequency scenarios where Mip-NeRF 360 demonstrates superior performance. Additionally, Mip-NeRF 360's smaller standard deviation in low-frequency scenes indicates its robustness in this scenario. We also present the findings of SOTAs' performance in

terms of scene diversity, which are described in the *supplementary materials*.

Visual results. We show the visual result for the SOTAs on *DL3DV-140* in Fig. 5. Overall, the artifact pattern for NeRF variants is the amount of “grainy” microstructure, while 3DGS creates elongated artifacts or “splotchy” Gaussians.

NeRF variants exhibit high sensitivity to distance scale, often generating blurry backgrounds and less-detailed foregrounds. For instance, Instant-NGP produces floating artifacts in the far-distance background of unbounded scenes. Although Zip-NeRF and Mip-NeRF 360 output fine-grained details compared to other NeRF variants, they also struggle with aliasing issues in sharp objects with high-frequent details, such as grasses and tiny leaves, as shown in Fig. 5 with ‘*high-freq*’ case. In contrast, 3DGS performs better against aliasing issues than NeRF variants; it suffers from noticeable artifacts in the far-distance background, such as the sky and far-distance buildings, as shown in Fig. 5 with ‘*outdoor*’ case.

Regarding view-dependent effects in reflective and transparent scenes, 3DGS excels at rendering finely detailed and sharp lighting, such as strong reflections on metal or glass surfaces, and effectively captures subtle edges of transparent objects, a challenge for other methods. However, it tends to oversimplify softer reflective effects, like cloud reflections on windows or subtle light on the ground, as shown in Fig. 5 with ‘*more-ref*’ and ‘*more-transp*’ cases. In contrast, Zip-NeRF and Mip-NeRF 360 are less sensitive to the intensity of reflective light, capturing reflections more generally. On the other hand, Nerfacto and Instant-NGP struggle with these complex lighting effects, often producing floating artifacts.

4.2. Generalizable NeRF

Recent NeRFs and 3DGS aim to only fit the training scene. Using these NVS methods to generalize to unseen real-world scenarios requires training on a large set of real-world multi-view images. Due to the lack of real-world scene-level multi-view images, existing works either resort to training on large-scale object-level synthetic data [8, 9, 23, 28, 38] or a hybrid of synthetic data and a small amount of real data [29, 30, 33]. The limited real-world data cannot fully bridge the domain gap. In this section, we conduct a pilot experiment to show that our large-scale real-world scene *DL3DV-10K* dataset has the potential to drive the learning-based generalizable NeRF methods by providing substantial real-world scenes for training.

Experiment details. We choose IBRNet [30] as the baseline to conduct an empirical study. To demonstrate the effectiveness of *DL3DV-10K*, we pre-train the IBRNet on our *DL3DV-10K* to obtain a general prior and fine-tune on the

Method	Diffuse Synthetic 360° [13]			Realistic Synthetic 360° [19]			Real Forward-Facing [18]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
IBRNet	34.72	0.983	0.024	23.95	0.906	0.101	24.82	0.808	0.178
IBRNet-S	34.22	0.979	0.024	23.57	0.905	0.101	24.86	0.807	0.183
IBRNet-270	35.18	0.984	0.024	24.55	0.911	0.097	25.00	0.812	0.180
IBRNet-1K	35.13	0.984	0.023	23.58	0.906	0.102	25.02	0.814	0.175
IBRNet-2K	35.34	0.984	0.024	24.98	0.913	0.095	25.08	0.815	0.176

Table 4. IBRNet: IBRNet trained from scratch, IBRNet-S: IBRNet trained from ScanNet++, IBRNet-270, IBRNet-1K, and IBRNet-2K: IBRNet trained from 270 scenes, 1,000 scenes, and 2,000 scenes from *DL3DV-10K*. Refer to *supplementary materials* for more samples on *DL3DV-10K*.

training dataset used by IBRNet and compare the performance with the train-from-scratch IBRNet on the evaluation datasets used by IBRNet. ScanNet++ [36] is another recent high-quality real-world scene dataset that focuses on indoor scenarios. We conduct a similar experiment on ScanNet++ to further show that the richer diversity and larger scale of *DL3DV-10K* significantly improve the generalizable NeRFs results.

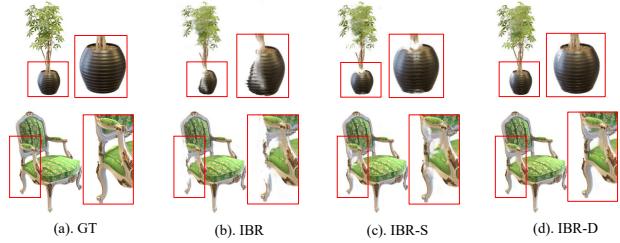


Figure 6. Qualitative comparison of (a). Ground truth. (b). IBRNet trained from scratch. (c). IBRNet pretrained on ScanNet++ and (d). IBRNet pretrained on *DL3DV-2K*

Results. The quantitative and qualitative results are shown in Tab. 4 and Fig. 6. The knowledge learned from ScanNet++ does not help IBRNet perform better on these existing benchmarks. However, the prior learned from a subset of our *DL3DV-10K* help IBRNet perform better on all the evaluation benchmarks. Also, when an increased data input from *DL3DV-10K* IBRNet consistently performs better in all the benchmarks.

5. Conclusion

We introduce *DL3DV-10K*, a large-scale multi-view scene dataset, gathered by capturing high-resolution videos of real-world scenarios. The abundant diversity and the fine-grained scene complexity within *DL3DV-140*, featuring 140 scenes from *DL3DV-10K*, create a challenging benchmark for Neural View Synthesis (NVS). Our thorough statistical evaluation of SOTA NVS methods on *DL3DV-140* provides a comprehensive understanding of the strengths and weaknesses of these techniques. Furthermore, we demonstrate that leveraging *DL3DV-10K* enhances the generalizability

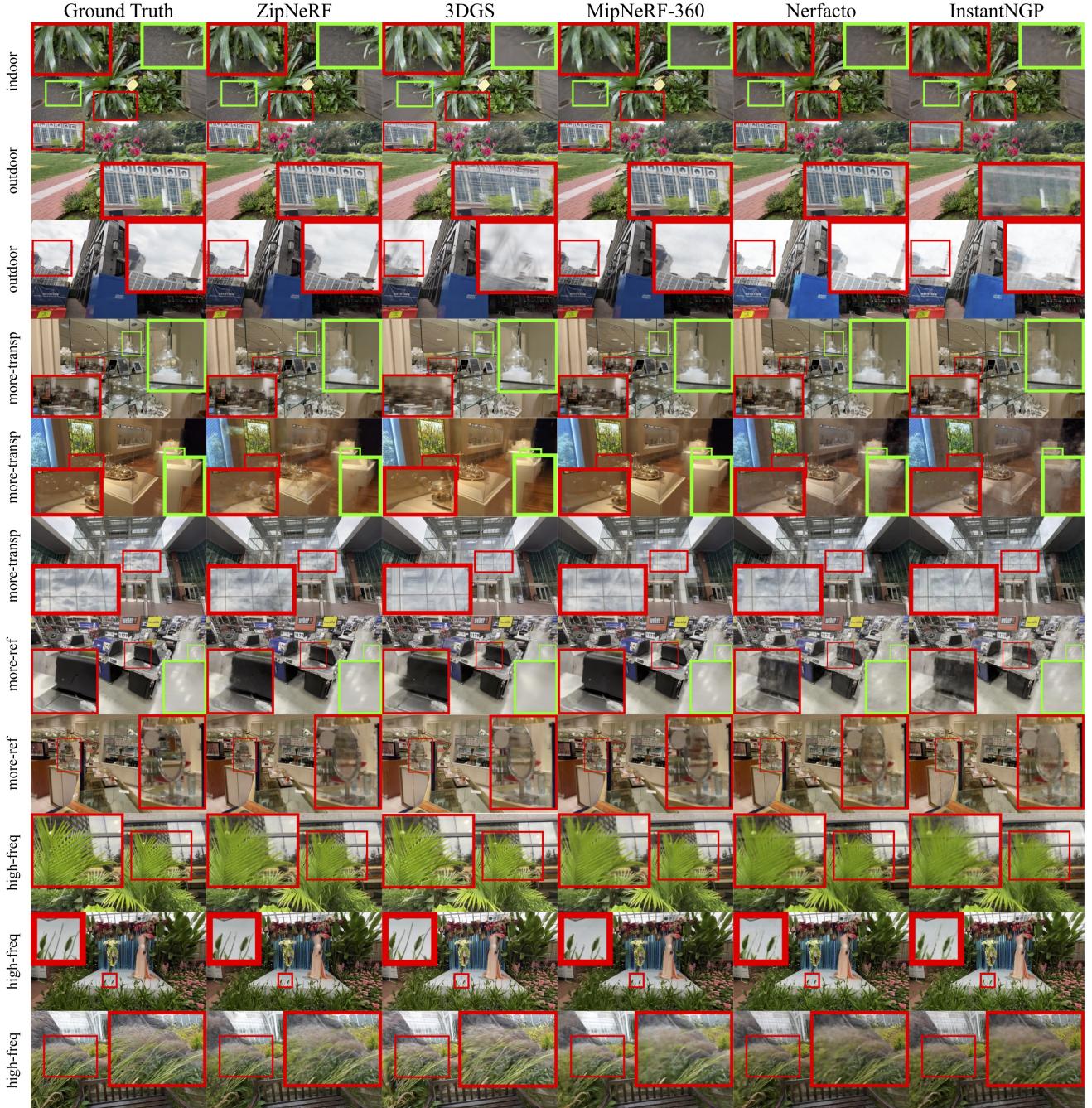


Figure 5. We compare the SOTA NVS methods and the corresponding ground truth images on *DL3DV-140* from held-out test views. More examples can be found in *supplementary materials*. The scenes are classified by complexity indices: *indoor* vs. *outdoor*, *more-ref* vs. *less-ref*, *high-freq* vs. *low-freq*, and *more-transp* vs. *less-transp*. Best view by zooming in.

of NeRF, enabling the development of a universal prior. This underscores the potential of *DL3DV-10K* in paving the way for the creation of a foundational model for learning 3D representations.

Limitations. *DL3DV-10K* encompasses extensive real-world scenes, enjoying the coverage of everyday accessible areas. This rich diversity and scale provide valuable insights for exploring deep 3D representation learning. However, there are certain limitations. While we demon-

strate *DL3DV-10K*'s potential in static view synthesis, some scenes include moving objects due to the nature of mobile phone video scene collection, as classified in Tab. 2, thereby introducing additional challenges for NVS. Nonetheless, such challenges may provide insights into exploring the robustness of learning-based 3D models. Moreover, these challenges may be solved by future learning-based 3D models for dynamic NVS.

References

- [1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7822–7831, 2021. 3
- [2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2
- [3] 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. 2, 5, 1
- [4] 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. 2, 5, 1
- [5] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 3
- [6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 3
- [7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3
- [8] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2, 7
- [9] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7911–7920, 2021. 7
- [10] Albert Cohen, Ingrid Daubechies, and J-C Feauveau. Biorthogonal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 45(5):485–560, 1992. 4
- [11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3
- [12] 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. *arXiv preprint arXiv:2307.05663*, 2023. 3
- [13] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (ToG)*, 37(6):1–15, 2018. 7
- [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. 3
- [15] 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. 2, 5, 6, 1
- [16] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 2, 3
- [17] Marc Levoy and Pat Hanrahan. Light field rendering. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 441–452. 2023. 2
- [18] 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. 2, 3, 7
- [19] 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, 3, 7
- [20] 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. 2, 5
- [21] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 3
- [22] Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Efstratios Gavves, and Tinne Tuytelaars. Deep reflectance maps. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4508–4516, 2016. 2

- [23] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. *arXiv preprint arXiv:2102.08860*, 2021. 7
- [24] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 3
- [25] 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
- [26] SuLvXiangXin. zipnerf-pytorch. <https://github.com/SuLvXiangXin/zipnerf-pytorch>, 2023. 1
- [27] Matthew Tancik, Ethan Weber, Evonne Ng, Rui long 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. 2, 3, 5, 1
- [28] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 7
- [29] Dan Wang, Xinrui Cui, Septimiu Salcudean, and Z Jane Wang. Generalizable neural radiance fields for novel view synthesis with transformer. *arXiv preprint arXiv:2206.05375*, 2022. 2, 7
- [30] 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. 2, 3, 7
- [31] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *International Journal of Computer Vision*, 128(12):2919–2935, 2020. 3
- [32] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 2
- [33] Hao Yang, Lanqing Hong, Aoxue Li, Tianyang Hu, Zhen-guo Li, Gim Hee Lee, and Liwei Wang. Contranerf: Generalizable neural radiance fields for synthetic-to-real novel view synthesis via contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16508–16517, 2023. 2, 7
- [34] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020. 3
- [35] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 325–334, 2011. 2, 4
- [36] 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. 2, 3, 7
- [37] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2
- [38] 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. 2, 3, 7
- [39] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: 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
- [40] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1802–1811, 2017. 4
- [41] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5287–5295, 2017. 4
- [42] 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. 2, 3

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

Supplementary Material

6. Overview

Our supplementary materials include this [pdf](#), [video demo](#) and a [HTML](#) for more qualitative results.

Sec. 7 discusses the data acquisition standard and the *DL3DV-10K* data distribution. Sec. 8 discusses more details of our benchmark experiments, including experiment, training details and more qualitative results, and details of the generalizable NeRF experiment.

7. Data

7.1. Data acquisition

The scene coverage for video shooting is illustrated in Fig. 8. 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.

7.2. 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. 7 presents the reflection labeling criteria. Transparency labeling follows the same rule.

7.3. Data Statistics

Scene summary by secondary POI category. The secondary POI categories are detailed classes within the primary POI categories. Fig. 9 shows scene statistics for each secondary POI category and the corresponding primary POI category. Fig. 10 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

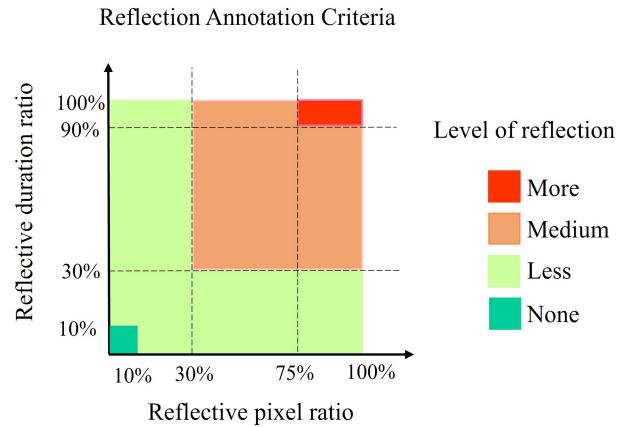


Figure 7. Reflection labeling criteria. Transparency annotation likewise.

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. We present a sample of *DL3DV-10K* in the [video demo](#).

Frequency and duration estimates. The kernel density distribution of frequency metric and video duration can be found in Fig. 11. The frequency classes are delineated based on the median value of the frequency metric.

8. Experiment

8.1. NVS benchmark

Experiment Details The implementation of Nerfacto and Instant-NGP is from nerfstudio [27]. MipNeRF360 [3] and 3D gaussian splatting (3DGS) [15] codes are from the authors. ZipNeRF [4] source code is not public yet when we submit the paper. We used a public implementation [26] 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.

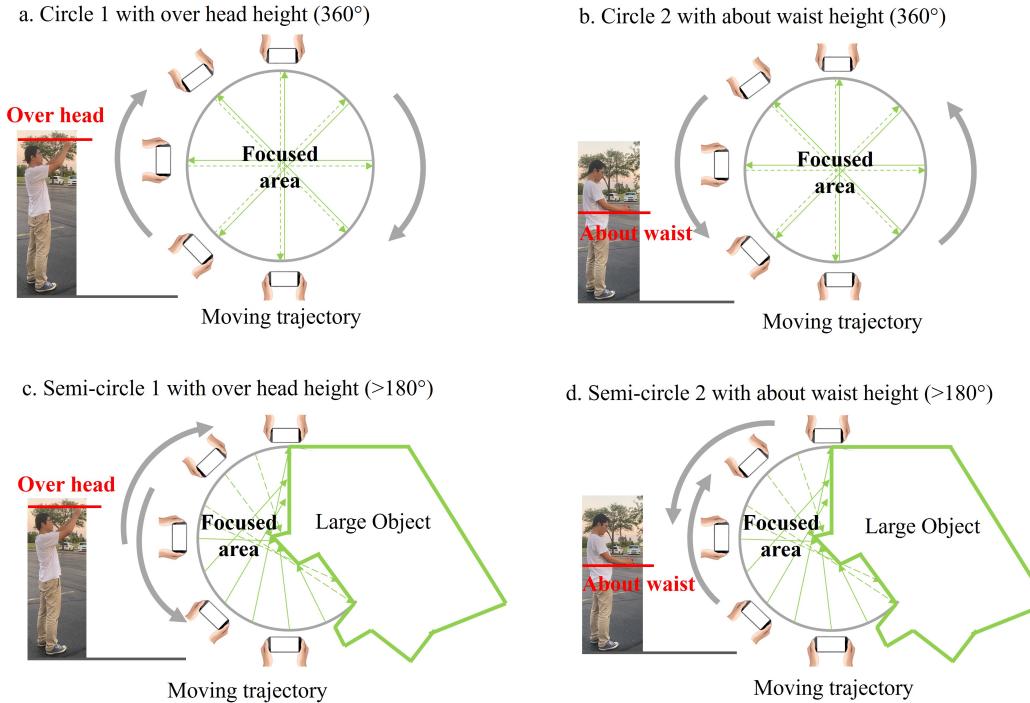


Figure 8. Video shooting examples with different heights and angles.

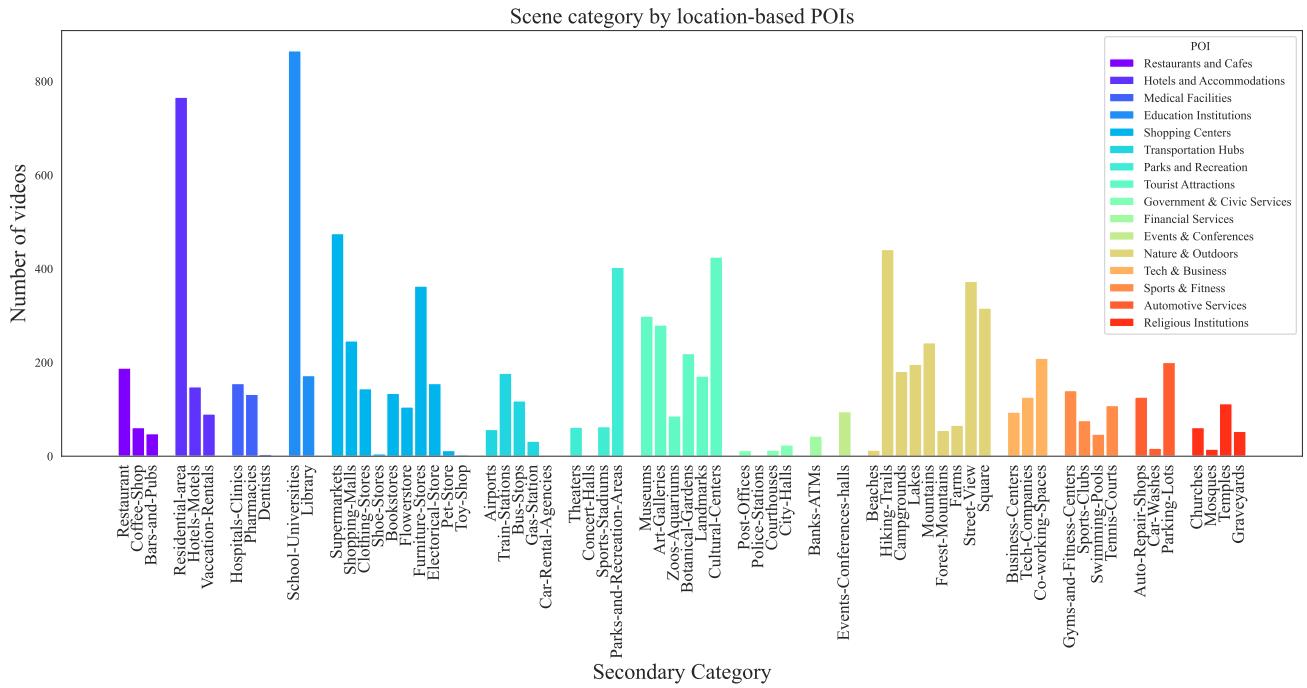


Figure 9. Number of scenes within secondary POI category. The legend contains the mapping between the primary and secondary POI categories. We observe that '*schools-universities*' and '*residential-area*' are the predominant scenes in our *DL3DV-10K* dataset. In contrast, locations such as government 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.

More quantitative results. We present the performance of State-of-the-art (SOTAs) in *DL3DV-140* by scene primary POI categories in Fig. 12.

More visual results. We present more visual results for the performance of SOTAs on *DL3DV-140* by scene complexity indices. In particular, Fig. 13 describes the performance of SOTAs by environmental setting; Fig. 14 describes the performance of SOTAs by frequency; Fig. 15 describes the performance of SOTAs by transparency; and Fig. 17 describes the performance of SOTAs by reflection. Due to the file size limitation, we provide visual results for 70 scenes in *DL3DV-140* in the **HTML** supplementary submission.

8.2. Generalizable NeRF

Experiment details We follow the default setting by IBR-Net [30]. The training dataset includes LLFF [18], spaces, RealEstate10K [42] and self-collected small dataset by IBRNet authors. The evaluation dataset includes Diffuse Synthetic 360° [25], Realistic Synthetic 360° [19], 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.

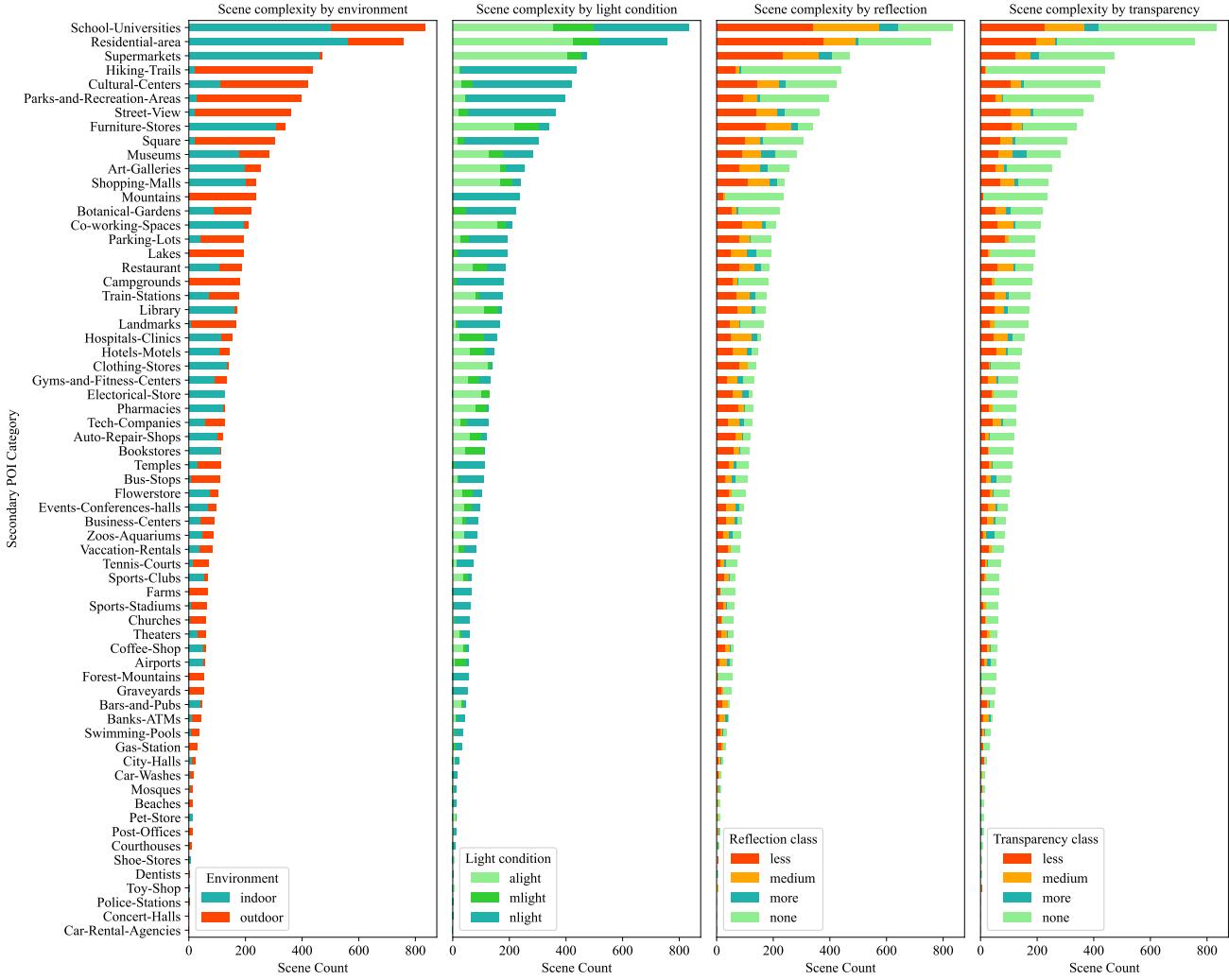


Figure 10. We show the distribution of scenes captured in secondary POI categories by complexities, including environmental setting, light conditions, reflective surfaces, and transparent materials.

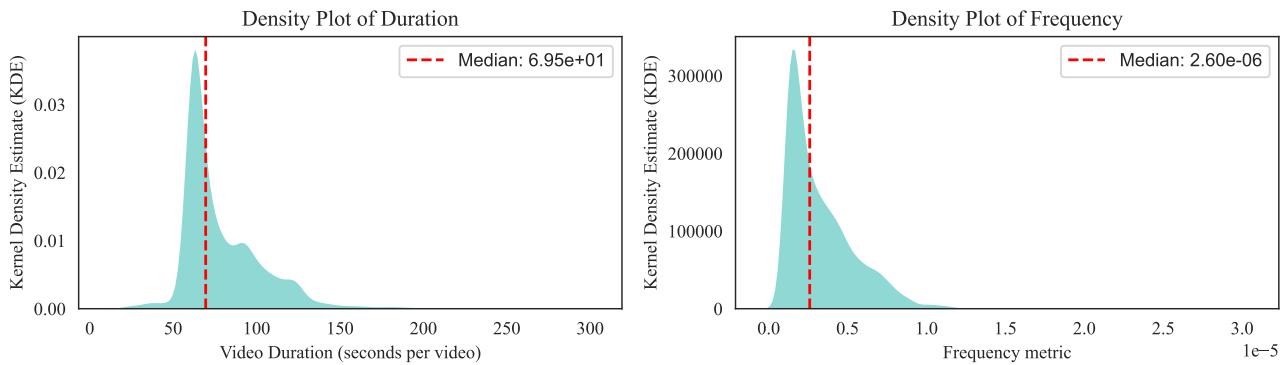


Figure 11. We show the distribution of video duration and frequency metric in **10,510** videos. The minimum duration for video shooting with consumer mobile devices is set at 60 secs, while for drone cameras, it's at least 45 secs. In our dataset, the median video duration is 69.5 secs. Furthermore, 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.

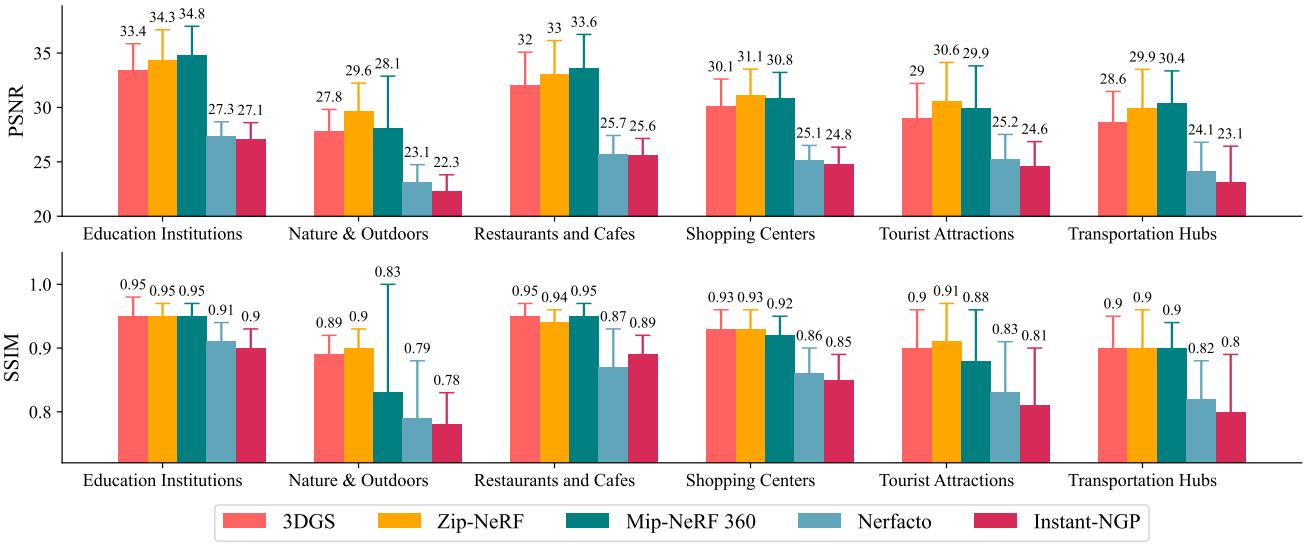


Figure 12. We present the average performance on 6 primary POI categories (*Education institutions, Nature & Outdoors, Restaurants and Cafes, Shopping Centers, Tourist Attractions, and Transportation Hubs*) in the *DL3DV-140*. The text above the bar plot is the mean value of the methods on the primary 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 textit{Education institutions} and *Restaurants and Cafes*.

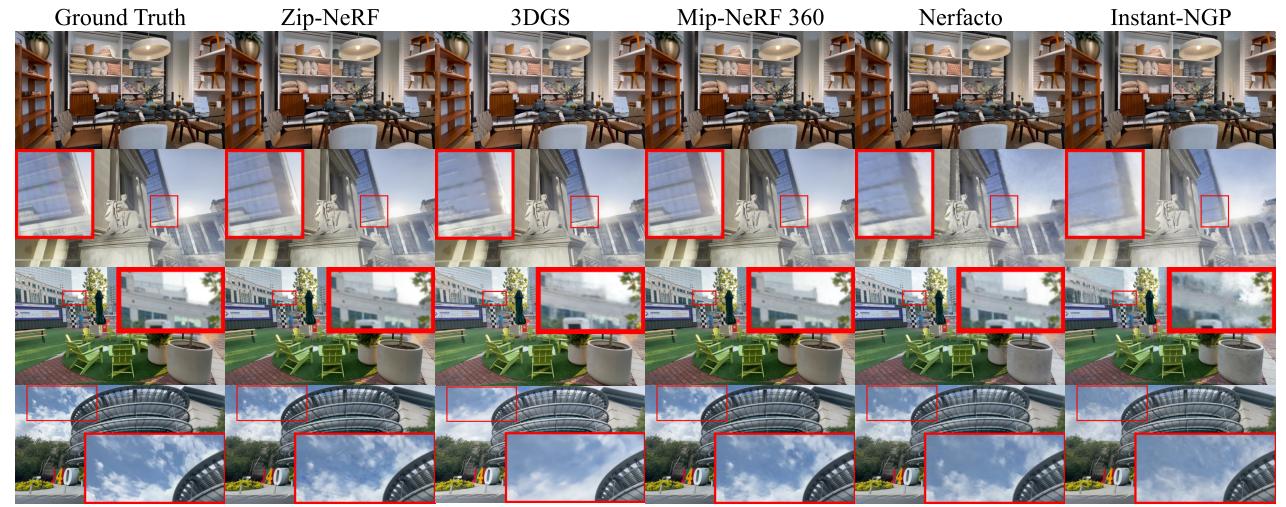


Figure 13. We compare the SOTAs for indoor (bounded) and outdoor (unbounded) environments on *DL3DV-140* from held-out test views. As illustrated in the figure, indoor scenes pose fewer challenges compared to outdoor scenes, where SOTAs demonstrate varying levels of performance. We observe that outdoor scene is more challenging for 3D Gaussian Splatting (3DGS), Nerfacto, and Instant-NGP than Zip-NeRF and Mip-NeRF 360.

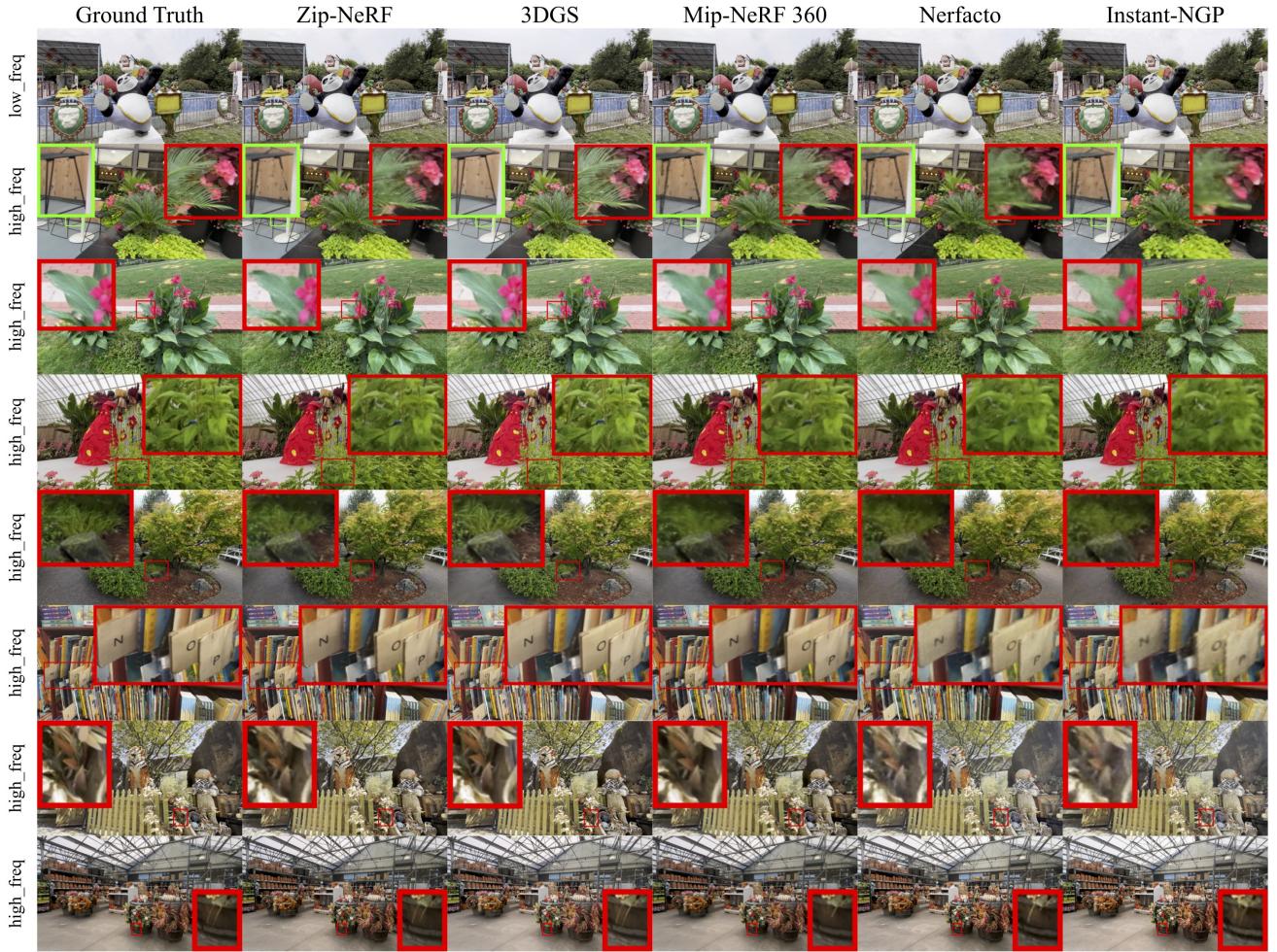


Figure 14. We compare the performance of SOTAs in frequency (*low_freq* vs. *high_freq*) on *DL3DV-140* from held-out test views. As shown in the figure, high frequency (*high_freq*) scene is more challenging than low frequency (*low_freq*) scene. We observe that 3DGS consistently captures scenes with 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.

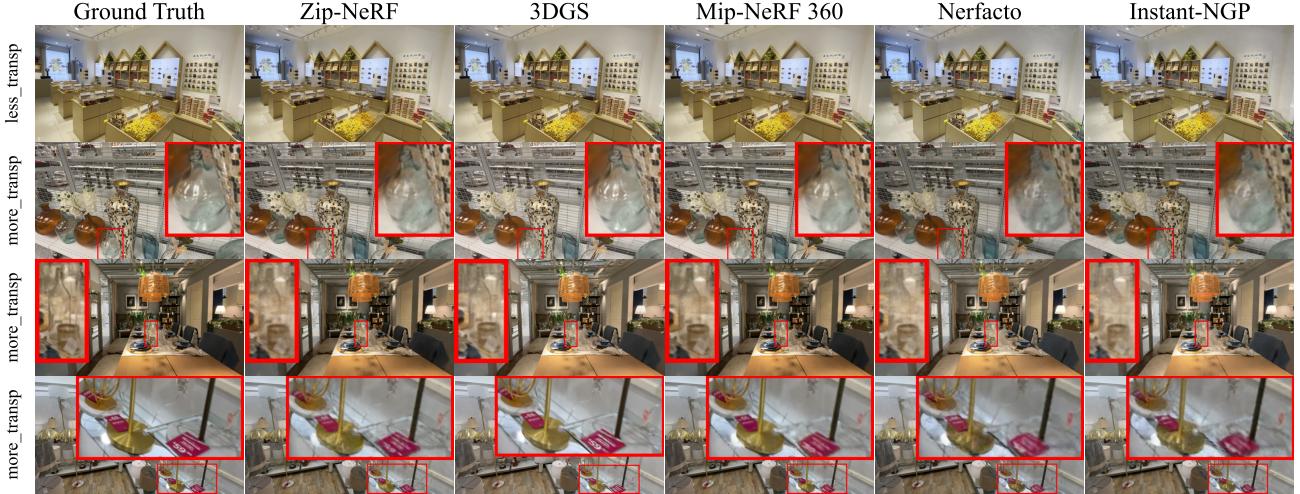


Figure 15. We compare the performance of SOTAs for transparency classes (*less_transp* vs. *more_transp*) on *DL3DV-140* from held-out test views. As shown in the figure, scenes with more transparent materials (*more_transp*) are more challenging than scenes with less transparent materials (*less_transp*). 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.

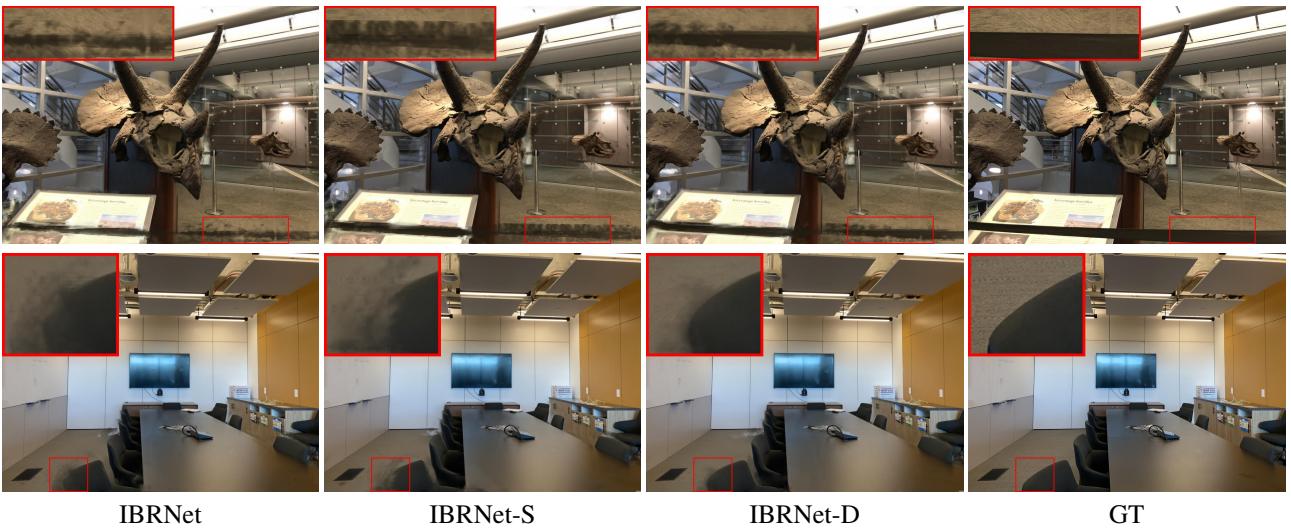


Figure 16. **More qualitative results for generalizable NeRF.** IBRNet-S: pretrain IBRNet on Scannet++ [36]. IBRNet-D: pretrain IBRNet on *DL3DV-10K*. Priors learned from *DL3DV-10K* help IBRNet perform the best on the evaluation.



Figure 17. 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.