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

Lu Ling[1*], Yichen Sheng[1*], Zhi Tu[1], Wentian Zhao[2], Cheng Xin[3], Kun Wan [2], Lantao Yu[2], Qianyu Guo[1], Zixun Yu [4], Yawen Lu[1], Xuanmao Li[5], Xingpeng Sun [1], Rohan Ashok[1], Aniruddha Mukherjee[1], Hao Kang [6], Xiangrui Kong[1], Gang Hua[6], Tianyi Zhang[1], Bedrich Benes[1] , Aniket Bera[1]

[1] Purdue University, [2] Adobe Inc. , [3] Rutgers University

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

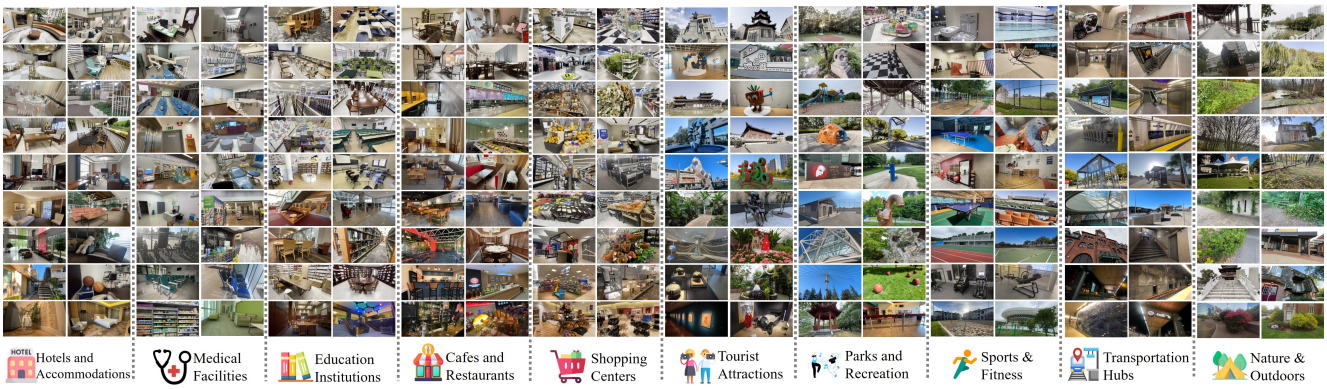

Figure 1. We introduce ***DL3DV-10K***, a large-scale, scene dataset capturing real-world scenarios. *DL3DV-10K* contains **10,510** videos at 4K resolution spanning 65 types of point-of-interest (POI) locations, covering a wide range of everyday areas. With the fine-grained annotation on scene diversity 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.*

1. Introduction

The evolution in deep 3D representation learning, driven by essential datasets, boosts various tasks in 3D vision [6, 14, 26–28, 30, 31]. Notably, the inception of Neural Radiance Fields [20] (NeRF), offering a new approach through a continuous high-dimensional neural network, revolutionized leaning-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 inspires waves of innovative developments such as NeRF variants [3, 4, 9, 32, 37, 42] and 3D Gaussian Splatting (3DGS) [16], significantly enhancing experiences in virtual reality, augmented reality, and advanced simulations.

*Joint first authors.

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 [20] 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 [17] and LLFF dataset [19] 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 [23]), 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 PixlNeRF [43], IBRNet [35], and MVSNeRF [9] mitigate this by training across multiple scenes to learn scene priors. While datasets like RealEstate [47] and ScanNet++ [41] 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 [40] (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 snippet 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 [32], Instant-NGP [21], Mip-NeRF 360 [3], Zip-NeRF [4], and 3DGS [16]. Leveraging the MV 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 scene-level dataset. It has 4K resolution videos and RGB images with camera poses, covering 65 POI locations. Each scene is annotated with the POI category, light condition, environment setting, varying reflection and transparency, and 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 [18, 47]. Since 2020, Neural Radiance Fields (NeRF) [20] 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 [21], 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 [32] merges elements from previous NeRF methods to balance speed and quality. Additionally, 3D Gaussian Splatting (3DGS) [16] 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 [38], TransNeRF [34], PixelNeRF [43], MVS-NeRF [9], and IBRNet [35] overcome this by training on numerous scenes for universal priors and sparse-view synthesis. Yet, the absence of large-scale scenes reflective of real-world diversity fails to provide adequate assessment

Dataset	# of scene	# of POI category	Resolution	# of frames	Scene complexity annotation		
					indoor/outdoor	reflection	transparency
LLFF [19]	24	-	640 × 480	<1K	✓✓	✓	✗
DTU [15]	124	5	1200 × 1600	30K	✓✗	✓	✗
BlendedMVS [39]	113	-	1536 × 2048	17K	✓✓	✗	✗
ScanNet [12]	1513	11	1296 × 968	2,500K	✓✗	✗	✗
Matterport3D [7]	90[1]	-	1280 × 1024	195K	✓✗	✓	✗
Tanks and Temples [17]	21	14	3840 × 2160	147K	✓✓	✓	✗
ETH3D [25]	25	11	6048 × 4032	<1K	✓✓	✗	✗
RealEstate10K [47]	10,000	1	1280 × 720	10,000K	✓✗	✗	✗
ARKitScenes [5]	1661	1	1920 × 1440	450K	✓✗	✗	✗
ScanNet++ [41]	460	5	7008 × 4672[2]	3,980K	✓✗	✓	✗
DL3DV-10K (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.

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) [20], ShapeNet [8] and Objaverse [13], 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 [15] and Tanks and Temples [17], offer limited variety. While ScanNet [12] 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 [19] offers cellphone-captured 24 forward-facing scenes; Mip-NeRF 360 dataset [20] provides 9 indoor and outdoor scenes with uniform distance around central subjects; Nerfstudio dataset [32] 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 [8] and Pix2Vox++ [36], to foundational real-world datasets capturing both object-level and scene-level images.

[1] 90 building-scale scenes covering 2056 rooms

[2] 7008×4672 in 270 scenes and 1920×1440 in 190 scenes

Object-level datasets like Objectron [1], CO3D [22], and MVimgnet [44] 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 [43].

Scene-level datasets are essential for NVS and scene understanding, yet offerings like LLFF [19], Matterport3D [7], and BlendedMVS [39] encompass limited scenes. DTU [15], despite its use in developing generalized NeRF [43], is limited by its scale, affecting models' ability to generalize. RealEstate10k [47], ARKitScenes [5], ScanNet [12], and the high-resolution ScanNet++ [41] improve this with a broad range of detailed indoor scenes. Yet, their applicability remains less comprehensive for 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.

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.

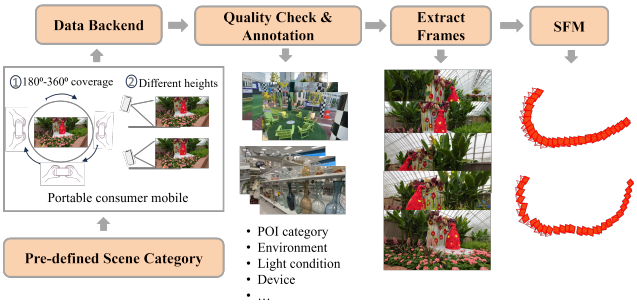

Figure 2. The efficient data acquisition pipeline of *DL3DV-10K*. Refer to *supplementary materials* for more visual illustrations.

3.1. Data Acquisition

Diversity. Our collection focuses on various commonly accessible static scenes, adhering to the point-of-interest (POI) [40] 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 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 [45] and normals [46].

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. 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.

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.

3.2. Data Processing

Data Screening. The dataset is targeted to static scene and are collected under the permission of agents. Besides, we removed voice and any additional metadata from videos and screen from sensitive content. We detect and mosaic any personal identifiers such as faces, names, and numbers.

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 [11] 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 frequency 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

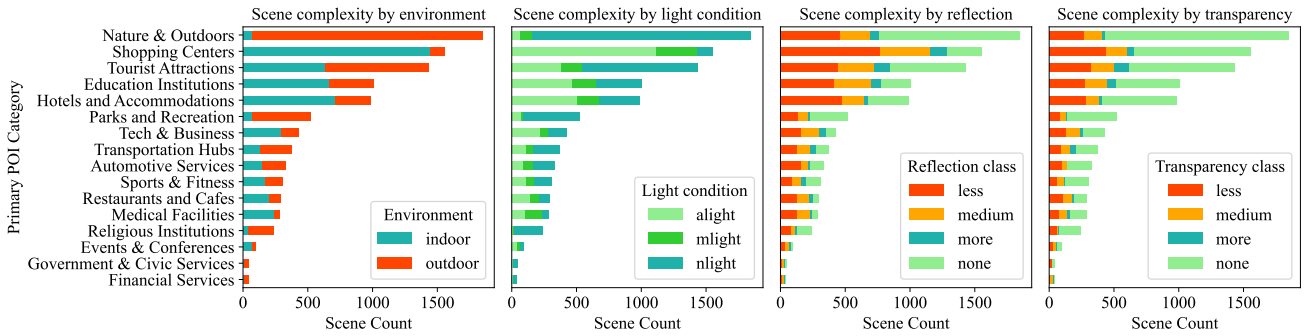

Figure 3. We show the distribution of primary POI scene category by complexity indices. Attributes in light conditions include: natural light ('*nlight*'), artificial light ('*alight*'), and a combination of both ('*mlight*'). Reflection class includes '*more*', '*medium*', '*less*', and '*none*'. Transparency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

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 '*Parks and Recreation*'. Its secondary POI category includes '*theaters*', '*concert halls*', '*sports stadiums*', 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.

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 categories into two classes 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*). More details related to the benchmark selection refer to *supplementary materials*. *DL3DV-140* offers challenging scenes with a rich mix of diversity and complexity for a comprehensive evaluation of SOTA methods.

	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). Training time and memory usage may be different depending on various configurations.

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 [32], Instant-NGP [21], Mip-NeRF 360 [3], Zip-NeRF [4], and 3D Gaus-

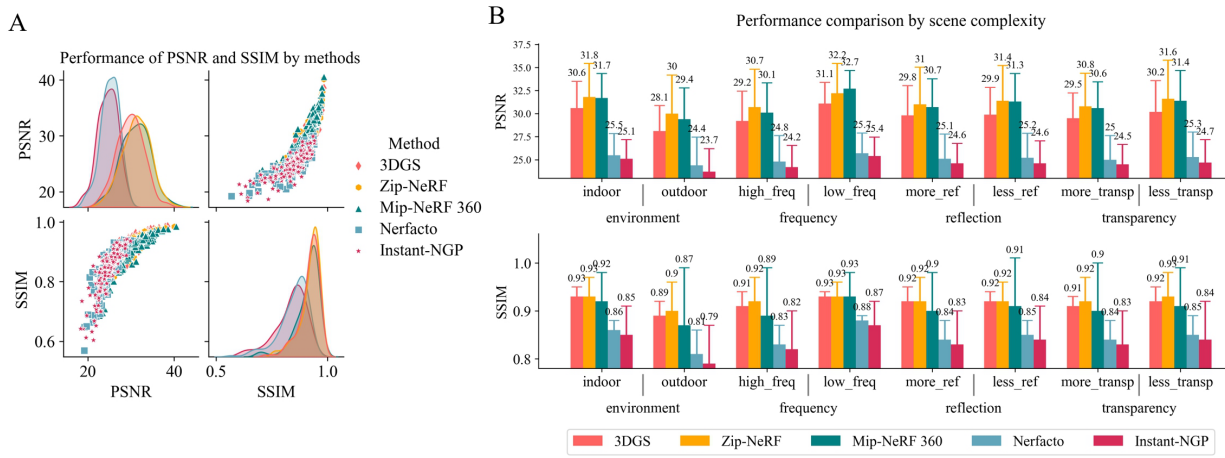

Figure 4. A presents the probability density of PSNR (top diagonal) and SSIM (bottom diagonal) on *DL3DV-140* for each method. Higher density indicates better performance. The sub-diagonal scatter plots reveal the correlation between PSNR and SSIM. B describes the performance comparison by scene complexity. The text above the bar plot is the mean value of the methods on the attribute.

sian Splatting (3DGS) [16].

Experiment details. The SOTAs have different assumptions on the image resolution. For fairness, we use 960×560 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. Refer to *supplementary materials* for details. 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 [16].

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 visual results for 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

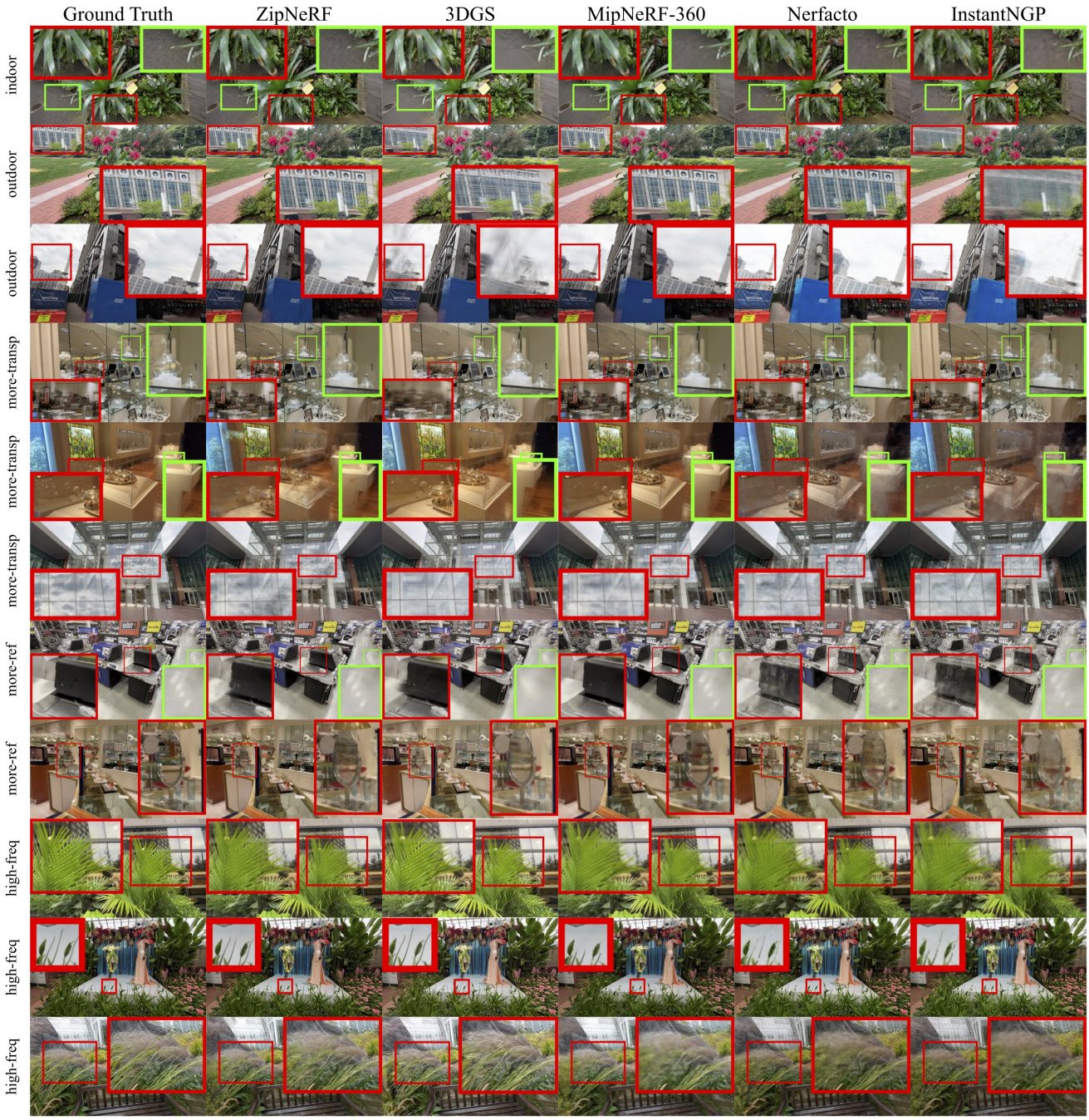

Figure 5. We compare the SOTA NVS methods and the corresponding ground truth images on *DL3DV-140* from held-out test views. More examples can be found in *supplementary materials*. The scenes are classified by complexity indices: *indoor vs. outdoor*, *more-ref vs. less-ref*, *high-freq vs. low-freq*,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

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

Method	Diffuse Synthetic 360° [29]			Real Forward-Facing [19]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
IBRNet	34.72	0.983	0.024	24.82	0.808	0.178
IBRNet-S	34.22	0.979	0.024	24.86	0.807	0.183
IBRNet-270	35.18	0.984	0.024	25.00	0.812	0.180
IBRNet-1K	35.13	0.984	0.023	25.02	0.814	0.175
IBRNet-2K	35.34	0.984	0.024	25.08	0.815	0.176

Table 4. IBRNet is trained from scratch. IBRNet-S, IBRNet-270, IBRNet-1K, and IBRNet-2K are pre-trained on ScanNet++(270), DL3DV-270, DL3DV-1K, and DL3DV-2K.

Method	DTU Validation Samples [15]			Real Forward-Facing [19]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MVSNeRF	18.49	0.598	0.476	16.16	0.350	0.549
MVSNeRF-S	19.74	0.660	0.428	17.50	0.442	0.541
MVSNeRF-270	20.00	0.671	0.397	17.64	0.449	0.492
MVSNeRF-1K	20.42	0.671	0.396	17.89	0.458	0.492
MVSNeRF-2K	20.85	0.671	0.394	17.90	0.459	0.505

Table 5. MVSNeRF is trained from scratch. MVSNeRF-S, MVSNeRF-270, MVSNeRF-1K and MVSNeRF-2K are pre-trained on ScanNet++(270), DL3DV-270, DL3DV-1K, and DL3DV-2K.

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 [9, 10, 24, 33, 43] or a hybrid of synthetic data and a small amount of real data [34, 35, 38]. The limited real-world data cannot fully bridge the domain gap. In this section, we conduct a pilot experiment to show that our *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 [35] and MVSNet [9] as two baselines to conduct empirical studies. To demonstrate the effectiveness of *DL3DV-10K*, we pre-train the IBRNet and MVSNeRF separately on *DL3DV-10K* to obtain a general prior and fine-tune on the training dataset used by themselves and compare the performance with the train-from-scratch IBRNet/MVSNeRF on the their evaluation datasets. ScanNet++ [41] is another recent real-world scene dataset that focuses on indoor scenarios. We select 270 high-quality scenes from ScanNet++ and conduct a similar experiment on ScanNet++(270) to further show that the richer diversity and larger scale of *DL3DV-10K* significantly improve the generalizable NeRFs results.

Results. The quantitative results are shown in Tab. 4 and Tab. 5 . The knowledge learned from ScanNet++ does not help IBRNet and MVSNeRF perform better on their benchmarks. However, the prior learned from a subset of our *DL3DV-10K* help them perform better on their evaluation benchmarks. Also, when increase data input from *DL3DV-10K* IBRNet and MVSNeRF consistently performs better in all their benchmarks. Refer to *supplementary materials* for more samples on *DL3DV-10K*.

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 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 demonstrate *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.

Acknowledgments. We express our sincere thanks to the volunteers for their invaluable contribution to the DL3DV-10K dataset, especially *Zhaopeng Wang*, *Jinghua Wu*, *Yueting Zhao*, *Haomeng Zhang*, *Aaditya Kharel*, *Izel Avila*, *Rahul Nahar*, *Mayesha Monjur*, *Neel Acharya*, *Xindi Tang*, *Wanhai Sheng*, and *Chunfang Chen*.

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.

[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.

[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.

[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.

[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.

[6] Manush Bhatt, Rajesh Kalyanam, Gen Nishida, Liu He, Christopher May, Dev Niyogi, and Daniel Aliaga. Design and deployment of photo2building: A cloud-based procedural modeling tool as a service. In *Practice and Experience in Advanced Research Computing*, pages 132–138. 2020.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Yuchun Huang, Ping Ma, Zheng Ji, and Liu He. Part-based modeling of pole-like objects using divergence-incorporated 3-d clustering of mobile laser scanning point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3): 2611–2626, 2020.

[15] 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.

[16] 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.

[17] 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.

[18] Marc Levoy and Pat Hanrahan. Light field rendering. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 441–452. 2023.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. *arXiv preprint arXiv:2102.08860*, 2021.

[25] 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.

[26] Yichen Sheng, Jianming Zhang, and Bedrich Benes. Ssn: Soft shadow network for image compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4380–4390, 2021.

[27] Yichen Sheng, Yifan Liu, Jianming Zhang, Wei Yin, A Cengiz Oztireli, He Zhang, Zhe Lin, Eli Shechtman, and Bedrich Benes. Controllable shadow generation using pixel height maps. In *European Conference on Computer Vision*, pages 240–256. Springer, 2022.

[28] Yichen Sheng, Jianming Zhang, Julien Philip, Yannick Hold-Geoffroy, Xin Sun, He Zhang, Lu Ling, and Bedrich Benes. Pixht-lab: Pixel height based light effect generation for image compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16643–16653, 2023.

[29] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019.

[30] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023.

[31] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. *arXiv preprint arXiv:2403.10701*, 2024.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] Hao Yang, Lanqing Hong, Aoxue Li, Tianyang Hu, Zhenguo 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.