MVImgNet2.0: A Larger-scale Dataset of Multi-view Images

XIAOGUANG HAN[*][†], SSE and FNii, CUHKSZ, China
YUSHUANG WU[*][‡], FNii and SSE, CUHKSZ, China
LUYUE SHI[*], FNii and SSE, CUHKSZ, China
HAOLIN LIU[*], FNii and SSE, CUHKSZ, China
HONGJIE LIAO, FNii and SSE, CUHKSZ, China
LINGTENG QIU, FNii and SSE, CUHKSZ, China
WEIHAO YUAN, Alibaba Group, China
XIAODONG GU, Alibaba Group, China
ZILONG DONG, Alibaba Group, China
SHUGUANG CUI, SSE and FNii, CUHKSZ, China

Fig. 1. We introduce **MVImgnet2.0**, a larger-scale dataset of multi-view images, which enjoys 3D-aware signals from multi-view consistency. MVImgnet2.0 expands its last version into a total of 520k objects from 515 categories, and also provides higher-quality annotations. The extremely rich geometry and texture information in real-world objects leads to MVImgNet2.0's great potential in supporting large-scale learning in the 3D domain.

MVImgNet is a large-scale dataset that contains multi-view images of ~220k real-world objects in 238 classes. As a counterpart of ImageNet, it introduces 3D visual signals via multi-view shooting, making a soft bridge between 2D and 3D vision. This paper constructs the MVImgNet2.0 dataset that expands MVImgNet into a total of ~520k objects and 515 categories, which derives a 3D dataset with a larger scale that is more comparable to ones in the 2D domain. In addition to the expanded dataset scale and category range, MVImgNet2.0 is of a higher quality than MVImgNet owing to four new features: (i) most shoots capture 360° views of the objects, which can support the learning of object reconstruction with completeness; (ii) the segmentation manner is advanced to produce foreground object masks of higher accuracy; (iii) a more powerful structure-from-motion method is adopted to derive the camera pose for each frame of a lower estimation error; (iv) higher-quality dense point clouds are reconstructed via advanced methods for objects captured in 360° views, which can serve for downstream applications. Extensive experiments confirm the value of the proposed MVImgNet2.0 in boosting the performance of large 3D reconstruction models. MVImgNet2.0 will be public at *luyues.github.io/mvimgnet2*, including multi-view images of all 520k

---

[*] Equal Contribution.
[†] Corresponding Author.
[‡] Work done during internship supervised by Weihao Yuan at Alibaba.

Authors' addresses: Xiaoguang Han, hanxiaoguang@cuhk.edu.cn, SSE and FNii, CUHKSZ, China; Yushuang Wu, yushuangwu@link.cuhk.edu.cn, FNii and SSE, CUHKSZ, China; Luyue Shi, 117010231@link.cuhk.edu.cn, FNii and SSE, CUHKSZ, China; Haolin Liu, 115010192@link.cuhk.edu.cn, FNii and SSE, CUHKSZ, China; Hongjie Liao, hongjieliao@link.cuhk.edu.cn, FNii and SSE, CUHKSZ, China; Lingteng Qiu, 220019047@link.cuhk.edu.cn, FNii and SSE, CUHKSZ, China; Weihao Yuan, wyuanaa@connect.ust.hk, Alibaba Group, China; Xiaodong Gu, vactor1994@gmail.com, Alibaba Group, China; Zilong Dong, zjudzl@qq.com, Alibaba Group, China; Shuguang Cui, shuguangcui@cuhk.edu.cn, SSE and FNii, CUHKSZ, China.

0730-0301/2024/12-ART $15.00
https://doi.org/10.1145/3687973

objects, the reconstructed high-quality point clouds, and data annotation codes, hoping to inspire the broader vision community.

CCS Concepts: • **Computing methodologies → Reconstruction**.

Additional Key Words and Phrases: 3D Object Dataset, 3D Object Reconstruction, Image-based Modeling

ACM Reference Format:
Xiaoguang Han, Yushuang Wu, Luyue Shi, Haolin Liu, Hongjie Liao, Lingteng Qiu, Weihao Yuan, Xiaodong Gu, Zilong Dong, and Shuguang Cui. 2024. MVImgNet2.0: A Larger-scale Dataset of Multi-view Images. *ACM Trans. Graph.* 43, 6 (December 2024), 17 pages. https://doi.org/10.1145/3687973

1 INTRODUCTION

The field of deep learning has witnessed remarkable advancements, fueled primarily by learning from vast amounts of data [Deng et al. 2009; Krishna et al. 2017; Lin et al. 2014; Miech et al. 2019]. Learning from large-scale data has proven to be a key driver in scaling up deep learning models to tackle complex understanding or generative tasks, especially for the development of large models in the fields including natural language processing [Achiam et al. 2023; Thoppilan et al. 2022; Touvron et al. 2023], computer vision [Kirillov et al. 2023; Liu et al. 2023b, 2024b; Ren et al. 2024], and multimodal learning [Li et al. 2023a; Lin et al. 2023; Liu et al. 2024a].

This learning regime also attracts great attention in the field of 3D vision. In spite of the greater difficulty in collecting and labeling 3D data compared with textual or 2D visual data, there are still some efforts contributed to constructing large-scale or high-quality 3D generic datasets [Chang et al. 2015; Deitke et al. 2023, 2022; Downs et al. 2022; Reizenstein et al. 2021; Wu et al. 2023b; Yu et al. 2023]. Among them, one line of work constructs datasets like ShapeNet [Chang et al. 2015] and Objaverse [Deitke et al. 2022] composed of synthetic data, which limits the application in real scenarios. Differently, another line of work collects 3D data of real-life objects via scanning or multi-view photogrammetry. However, such datasets like CO3D [Reizenstein et al. 2021] and GSO [Downs et al. 2022] are limited in scale and category range until Yu et al. make the first step in constructing a large-scale one, MVImgNet [Yu et al. 2023], consisting of ∼220k multi-view images of 238 classes of common objects. The massive multi-view data do not only prove valuable in 2D visual understanding [Yu et al. 2023] via learning cross-view consistency, but also support the learning of generic shape priors to benefit 3D reconstruction [Hong et al. 2023; Wang et al. 2023; Wu et al. 2023a; Xu et al. 2023]. Considering the larger scale of datasets in the 2D domain, *e.g.* ImageNet [Deng et al. 2009], containing over 1 million images of 1k categories, MVImgNet is still inferior in scale that may limit its potential to support scaling up 3D learning. Therefore, we propose MVImgNet2.0 that expands MVImgNet to **twice** its original scale and category range. With a total of **520k objects** and **515 categories** that is **half** the scale of ImageNet, MVImgNet2.0 makes a further step towards a larger real-world 3D dataset with a smaller gap to ones in the 2D domain.

In addition to the expanded data scale and category range, MVImgNet2.0 has some other new features in data acquisition and annotation to improve the dataset quality. The biggest difference in data acquisition is that MVImgNet videos usually cover 180° views of objects, while most of the videos (230k/300k) collected in MVImgNet2.0 capture **360° object views** to represent a more complete shape. On the other hand, the annotations in MVImgNet2.0 are of higher quality in three aspects: (i) the foreground **object masks** in each frame are provided of higher accuracy; (ii) the **camera poses** of each view are estimated of lower error; (iii) the **dense reconstructions** are advanced to produce object point clouds of higher accuracy and robustness. To get the masks of objects of interest in each video frame, we advance the segmentation method in MVImgNet into a new detection-segmentation-tracking pipeline, which adopts a coarse-to-fine paradigm with temporal information [Kirillov et al. 2023; Liu et al. 2023b; Yang and Yang 2022] also incorporated to generate accurate object masks finally. For the camera pose estimation, we apply a more advanced structure-from-motion (SfM) algorithm that refines keypoints and bundles using deep features [Lindenberger et al. 2021] to compute camera poses of higher accuracy, especially for objects with fewer textures. The dense reconstructions in MVImgNet are generated by the multi-view stereo method, which is also advanced in MVImgNet2.0 based on neural surface rendering with multi-resolution 3D hash grids [Li et al. 2023b; Ye 2023] to improve the reconstruction accuracy and robustness.

While MVImgNet proves valuable in various visual tasks including radiance field reconstruction, multi-view stereo, and view-consistent image understanding [Yu et al. 2023], our experiments mainly focus on the task of 3D reconstruction. We first demonstrate that the more accurate camera poses estimated in MVImgNet2.0 can better support the per-scene 3D reconstruction. Besides, our experiments pays more attention to the application of generic 3D object reconstruction. We deploy three recent large reconstruction models – LRM [Hong et al. 2023], LGM [Tang et al. 2024] and TriplaneGaussian [Zou et al. 2023], and justify the value of MVImgNet2.0 data in improving their reconstruction quality and generalizability.

The contributions of this work are then summarized as follows:

- We propose MVImgNet2.0 that expands MVImgNet to a total of 520k real-life objects and 515 categories, which makes a further step to a larger-scale 3D generic dataset.
- MVImgNet2.0 has some new features in both data acquisition and annotation: (i) most of the added videos capture objects in 360° views; (ii) frames are annotated with more accurate object masks and (iii) more accurate camera poses; (iv) objects are densely reconstructed with an advanced method.
- Extensive experiments validate MVImgNet2.0's value in the task of 3D reconstruction, especially in improving the performance of large 3D reconstruction models.

2 RELATED WORK

Large-scale datasets. Expanding the size and breadth of training datasets has proven to be a highly effective strategy for enhancing the performance and robustness of deep learning models. In computer vision, the introduction of large-scale datasets such as ImageNet [Deng et al. 2009] and MS-COCO [Lin et al. 2014] has driven significant advancements across various tasks, including image classification, object detection, and captioning. This trend has persisted, with the diversity and scale of available datasets growing exponentially. Notable examples include image datasets like OpenImages [Kuznetsova et al. 2020] and Visual Genome [Krishna et al. 2017], video datasets like Kinetics [Kay et al. 2017] and HowTo100M [Miech et al. 2019], and multi-modal datasets like Conceptual Captions [Sharma et al. 2018], YFCC100M [Thomee et al. 2016], and LAION [Schuhmann et al. 2022], which support comprehensive vision-language correlations. By increasing the scale and coverage of these datasets, researchers and practitioners have achieved substantial improvements in the capabilities of computer vision and multi-modal systems [Dosovitskiy et al. 2020; Jia et al. 2021; Liu et al. 2023b; Radford et al. 2021].

3D datasets. 3D datasets encompass a broad spectrum, ranging from indoor to outdoor scenes and from human subjects to various objects. This paper primarily focuses on generic object 3D datasets. These datasets can be categorized into two main groups. The first group consists of synthetic 3D objects, such as those found in ShapeNet [Chang et al. 2015], ModelNet [Wu et al. 2015], ABO [Collins et al. 2022], and Objaverse [Deitke et al. 2023, 2022]. These datasets offer high-quality computer-aided design (CAD) models as 3D ground truths and can render 2D views in simulation to support 3D reconstruction learning. Though Objaverse collects over 800k CAD models, the quality of models is hard to guarantee, of which only ~170k (21%) have textures and less are of high quality for training, *e.g.* ~80k (10%) are used in LGM training [Tang et al. 2024]. Besides, the main limitation of such datasets lies in the intrinsic domain gap between synthetic and real objects. Another group focuses on collecting real-world 3D data through scanning or multi-view shooting. Dedicated scanning methods can produce high-quality 3D assets from real-life objects, as seen in ScanObjectNN [Uy et al. 2019], NAVI [Jampani et al. 2023], GSO [Downs et al. 2022], and OmniObj3D [Wu et al. 2023b], which collect 14k, 8k, and 6k 3D object data, respectively. Given the high cost of scanning, which limits scalability, MVImgNet [Yu et al. 2023] collects 220k 3D objects from 238 categories of real-life objects using multi-view shooting, marking a significant step toward constructing large-scale generic 3D datasets comparable to extensive 2D visual datasets. Another related dataset, CO3D [Reizenstein et al. 2021], employs a similar data collection method but on a smaller scale, with 19k objects and 38k objects in its new version. The proposed MVImgNet2.0 dataset advances this approach by expanding the scale and category range of MVImgNet to 520k objects in 515 categories, potentially facilitating the learning of large models for 3D understanding and generation.

3D reconstruction. 3D reconstruction from a single view or multiple views is a challenging task and also an important application for 3D datasets. Significant advancements have been made in single image to 3D reconstruction, starting with early methods focusing on point clouds [Fan et al. 2017; Wu et al. 2020], voxels [Chen and Zhang 2019; Choy et al. 2016; Tulsiani et al. 2017], meshes [Gkioxari et al. 2019; Wang et al. 2018], and introducing shape priors like 3D templates [Goel et al. 2020; Kanazawa et al. 2018; Kulkarni et al. 2020; Roth et al. 2016], semantics [Li et al. 2020], and poses [Bogo et al. 2016; Novotny et al. 2019] has also been extensively researched. With the emerging techniques based on implicit representations like SDFs [Mittal et al. 2022; Park et al. 2019], occupancy networks [Mescheder et al. 2019], and NeRF [Jang and Agapito 2021; Mildenhall et al. 2020; Müller et al. 2022b; Yu et al. 2021], some category-agnostic methods show great generalization potential [Niemeyer et al. 2020; Yan et al. 2016] but suffer from the lack of fine-grained details [Xu et al. 2019; Yu et al. 2021]. The field of 3D reconstruction from multiple views has also been a major focus in computer vision and graphics for decades. Traditional approaches to this task include structure-from-motion (SfM) methods for sparse reconstruction and calibration [Agarwal et al. 2011; Pollefeys et al. 2004; Schonberger and Frahm 2016a; Snavely et al. 2006], as well as multi-view stereo (MVS) techniques for dense reconstruction [Furukawa and Ponce 2009; Pollefeys et al. 2008; Schönberger et al. 2016]. More recently, deep learning-based MVS methods have emerged [Cheng et al. 2020; Gu et al. 2020; Shen et al. 2021; Yao et al. 2018, 2019], providing efficient, high-quality reconstruction through a feed-forward process. More recently, the use of pre-trained image/language models has introduced semantics and multi-view guidance [Li et al. 2023a, 2022; Radford et al. 2021; Rombach et al. 2022; Saharia et al. 2022] for image-to-3D reconstruction [Anciukevičius et al. 2023; Deng et al. 2023; Li et al. 2023c; Liu et al. 2023a; Shen et al. 2023; Tang et al. 2023]. Further, with the emergence of large-scale 3D datasets [Deitke et al. 2022; Reizenstein et al. 2021; Yu et al. 2023], some works explore a purely data-driven approach that learns a large model to reconstruct generic objects in the wild from 3D datasets [Hong et al. 2023; Tang et al. 2024; Wang et al. 2023; Xu et al. 2024; Zhang et al. 2024; Zou et al. 2023]. The proposed dataset has great potential in supporting these large reconstruction models to scale up their 3D reconstruction capability.

Application and impact of MVImgNet. The MVImgNet dataset [Yu et al. 2023], which provides a vast collection of multi-view images of real objects, has been extensively utilized in a variety of downstream tasks. In addition to fundamental 2D/3D understanding tasks as demonstrated in [Aubret et al. 2024; Chen et al. 2024a; Ke et al. 2024a; Lee et al. 2024], MVImgNet has significantly influenced the field of 3D object reconstruction. It offers a robust 3D prior through its extensive multi-view captures of diverse objects and also enhances the robustness of models in real-world reconstruction scenarios, as shown in [Hong et al. 2023; Jiang et al. 2024; Ntavelis et al. 2023; Wang et al. 2023; Wu et al. 2024]. Furthermore, researchers have explored employing MVImgNet data to train or finetune the multi-view diffusion models [Gao et al. 2024; Xu et al. 2023] and video diffusion models [Chen et al. 2024b; Han et al. 2024; Xie et al. 2023; Zuo et al. 2024] for high-quality 3D reconstruction. MVImgNet is also applied in some other related tasks, such as scene-level reconstruction and generation [Anciukevicius et al. 2024], generalizable novel view synthesis [Jang and Agapito 2024; Zhu et al. 2023], video generation [He et al. 2024], and 3D super resolution [Shen et al. 2024]. With MVImgNet2.0's larger scale, increased categories, 360-degree capturing, and higher annotation quality, it is anticipated to provide an even stronger 3D prior for 3D reconstruction and to better support other downstream tasks.

3 DATASET

MVImgNet2.0 is a large-scale dataset of multi-view images, which is efficiently collected via shooting 360°-view videos with phone cameras surrounding objects in the wild. In this section, we introduce the data acquisition and annotation pipeline in constructing

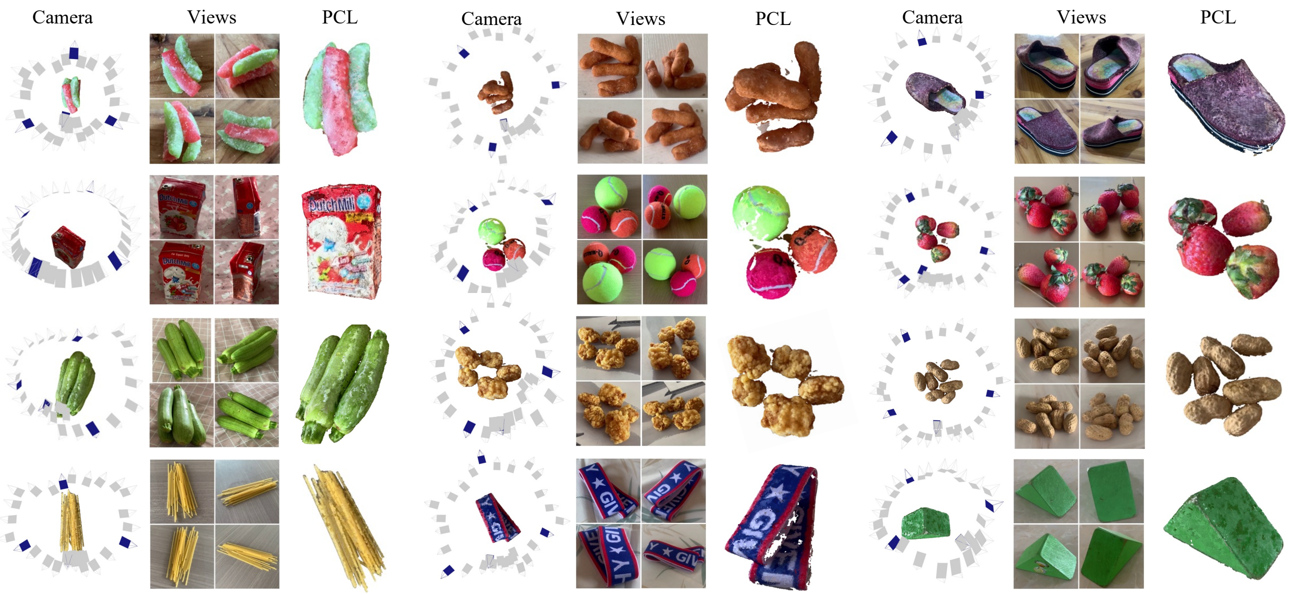

Fig. 2. MVImgNet2.0 data visualization. Objects in MVImgNet2.0 are in a wide range. For each object, we visualize the estimated camera poses and then sample 4 views to present images (whose corresponding camera poses are highlighted in dark color). We also visualize the point cloud annotations (PCL).

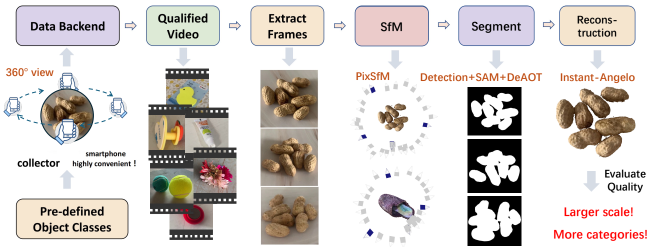

Fig. 3. The data acquisition and annotation pipeline in MVImgNet2.0. One video is first collected, uploaded, and qualified by collectors and annotators, then we extract frames from the video to conduct annotation including camera pose estimation via PixSfM, then object segmentation via a detection-segmentation-tracking pipeline, and lastly dense point cloud reconstruction via Instant-Angelo. All annotations are qualified by human annotators finally to filter out failure cases. New features in the MVImgNet2.0 pipeline are highlighted in brown or red color.

MVImgNet2.0, as shown in Fig. 3, mainly focusing on showing the differences between MVImgNet [Yu et al. 2023] and MVImgNet2.0.

3.1 Raw Data Acquisition

Similar to MVImgNet, the raw video data is gained through crowdsourcing. We first specify the diverse data categories to collect and the maximum amount for each category. The categories are chosen following the WordNet [Miller 1994] taxonomy and also from the common objects encountered or utilized in human daily life, and the maximum amount is determined by their generality and the complexity involved in capturing them. In addition to quantitatively expanding some of MVImgNet's categories with a small number of videos (70 categories), we have also collected 277 new categories to expand MVImgNet's collection of categories. Then, we draw up the requirements for the captured videos: (i) The length of each video must be around 10 seconds; (ii) The frames in the video must not be blurred; (iii) The presence proportion of the object in the video frames must be above 15%; (iv) Each video can only contain one class of principal object; (v) The captured object must be rather "three-dimensional" (excluding ones that are too flat and thin, or lacking in depth); (vi) Each video must capture 360° view of the object as much as possible. A visualization of camera poses in data collection is shown in Fig. 2. After setting up the requirements, similar to MVImgNet, we employ around a thousand normal collectors to take videos and upload them to the backend. Meanwhile, well-trained expert data cleaners are responsible for reviewing each submission and ensuring it fulfills the aforementioned capture requirements. The whole procedure ensures both the diversity and quality of the raw videos.

To sum up, data in MVImgNet2.0 has two main differences from ones in MVImgNet: 1) the data scale and category range are expanded, which allows for learning a more generalizable model; 2) videos are collected by capturing 360° view of objects, which allows for learning a better shape prior.

3.2 Data Annotation

For each qualified video submission, we exploit a similar data processing procedure as in reconstructing the MVImgNet dataset to conduct semi-automated annotation, as shown in Fig. 3. At first, around 30 frames are extracted from each video for sparse reconstruction, which derives the estimated camera poses of each view. Then, we generate object masks via segmentation methods for each extracted frame. Finally, given the camera poses and the masks in each view, we conduct dense reconstruction to produce the object point clouds. The main differences in MVImgNet2.0 lie in the advanced approaches to achieve higher-quality annotations, including 1) camera poses, 2) foreground object masks, and 3) point clouds.

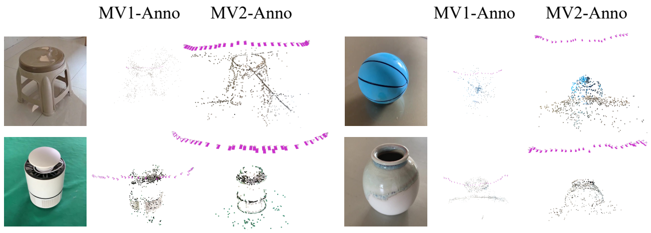

Fig. 4. The sparse reconstruction results comparison between using the original approach in MVImgNet (MV1-Anno) and using our advanced approach (MV2-Anno) for camera pose estimation (cameras are visualized in purple).

Sparse reconstruction. The sparse reconstruction aims to reconstruct the camera intrinsic and extrinsic for each video, by applying the Structure-from-Motion (SfM) algorithm [Schonberger and Frahm 2016b] on a series of equal-time-interval chosen frames. In MVImgNet2.0, we apply the Pixel-Perfect Structure-from-Motion (PixSfM) algorithm [Lindenberger et al. 2021] to obtain the sparse reconstruction results, which can estimate more precise camera parameters via two steps of keypoint and bundle adjustment based on dense features. The sparse reconstruction quality is improved by PixSfM, especially for objects with smooth surfaces and fewer textures, where classical SfM usually fails to produce reasonable estimation, as shown in Fig. 4.

Foreground object segmentation. MVImgNet uses the open-source segmentation tool CarveKit [Selin 2024] to generate the foreground object masks, which often results in ambiguous boundaries or incorrect masks, especially for those with a bit complex background. To obtain accurate object masks, we apply an advanced detection-segmentation-tracking pipeline, based on an open-set object detector Grounding-DINO [Liu et al. 2023b], a segmentation tool Segment-Anything (SAM) [Kirillov et al. 2023] [Ke et al. 2024b], and a video object tracker DeAOT [Yang and Yang 2022]. Given a sequence of video frames, we first apply Grounding-DINO to generate bounding box (bbox) candidates of foreground objects, where the category name is used as the text prompt. With the detection results giving coarse indications, we then apply SAM to generate an object mask for each bbox by using the image as input and the bbox as the prompt. We also initially filter out masks that may be inaccurate, according to their size, distance from the image boundary, number of connected components, etc. Finally, we further take the temporal information, the relation between neighboring frames, into consideration, where one mask is selected as the input of the video object tracker DeAOT to generate the final accurate masks. In addition, to obtain more precise object masks, we also manually check some results of each category and adjust the segmentation pipeline for some categories. We visualize some segmentation results in Fig. 5 to show the improved mask quality in MVImgNet2.0. A quantitative comparison between the segmenting performance by the original and advanced approach is presented in the SupMat's Tab. R.1.

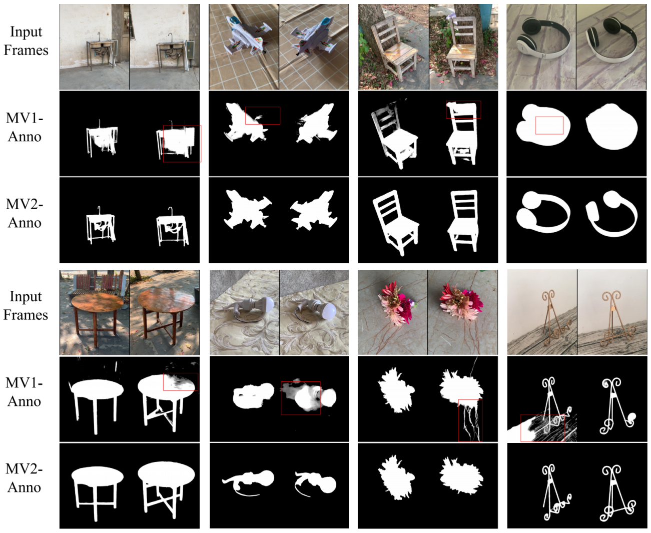

Fig. 5. The comparison of the foreground object segmentation results between using the original approach in MVImgNet (MV1-Anno) and using our advanced approach (MV2-Anno).

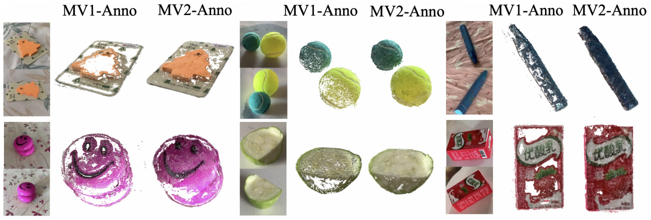

Fig. 6. The dense point cloud reconstruction results comparison between using the original approach in MVImgNet (MV1-Anno) and using our advanced approach (MV2-Anno).

Dense reconstruction. Different from MVImgNet which employs multi-view stereo (MVS) [Schönberger et al. 2016] of COLMAP to generate the densely reconstructed point cloud, we advance the point cloud reconstruction approach based on a neural surface reconstruction method, Neural-Angelo [Li et al. 2023b]. It incorporates multi-resolution hash encoding into neural SDF representations that allows for high-fidelity dense 3D reconstruction. In implementation, we adopt the open-source Instant-Angelo project [Ye 2023] to achieve fast point cloud reconstruction. Given the Neural-Angelo reconstruction outputs, similar to MVImgNet, we also manually clean the point clouds to delete the objects with obvious noisy, extremely sparse reconstructions, or backgrounds. Example comparisons between point clouds produced by the approach used in MVImgNet and MVImgNet2.0 are visualized in Fig. 6, which demonstrates that the advanced method used in MVImgNet2.0 usually leads to more accurate and complete reconstructions.

3.3 Dataset Statistics

Tab. 1 shows the statistics of MVImgNet2.0 and other alternatives and Fig. 2 and Fig. 11 shows some samples of MVImgNet2.0. In summary, MVImgNet2.0 includes 300k videos with 9 million frames and 347 object classes, of which 277 are new categories not covered by MVImgNet, and the annotations comprehensively cover object masks, camera pose parameters, and point clouds. The categories are organized in a taxonomic manner in SupMat's Fig. R.4, and we recommend our project page for a more detailed display. In addition, we also give more detailed per-category statistics in SupMat's Fig. R.3. With the construction of MVImgNet2.0, the total statistics

Table 1. **Comparison** between MVImgNet2.0 and related datasets. "pcl" denotes point clouds. *Note that only parts of data in CO3D (~30%) and MVImgNet (~40%) are annotated with point cloud GT.

Dataset	Real	# of objects	# of classes	Multi-view	3D-GT
ShapeNet [Chang et al. 2015]	✗	51k	55	render	CAD model
ModelNet [Wu et al. 2015]	✗	12k	40	render	CAD model
ScanObjectNN [Uy et al. 2019]	✓	14k	15	limited	pcl
CO3D [Reizenstein et al. 2021]	✓	19k	50	360° views	pcl*
GSO [Downs et al. 2022]	✓	1k	17	360° views	RGB-D scan
ABO [Downs et al. 2022]	✗	8k	63	render	CAD model
Objaverse [Deitke et al. 2022]	✗	818k	21k	render	CAD model
OmniObj3D [Wu et al. 2023b]	✓	6k	190	360° views	RGB-D scan
MVImgNet1.0 [Yu et al. 2023]	✓	220k	238	180° views	pcl*
MVImgNet2.0	✓	300k	347	180°/360° views	pcl
MVImgNet1.0+2.0	✓	520k	515	180°/360° views	pcl

of MVImgNet datasets reach 520k objects in 515 categories, which is closer to the scale of 2D large-scale datasets, *e.g.* ImageNet with ~1 million data in 1000 categories.

4 EXPERIMENTS

This section aims to validate the value of the proposed MVImgNet2.0 in the application of 3D reconstruction. We first introduce the experiment setup, then conduct per-scene 3D reconstruction to validate the value of camera pose annotations with higher accuracy, and finally we pay the main focus on justifying the value of new features in MVImgNet2.0 in improving the performance of large reconstruction models, including the larger data scale, the expanded category range, the 360°-view videos, and the higher-quality annotations.

4.1 Experiment Setup

Datasets. We adopt three datasets in experiments, including the synthetic dataset Objaverse [Deitke et al. 2022], the original data in MVImgNet [Yu et al. 2023] (MV1-Data), and the newly added data in MVImgNet2.0 (MV2-Data). The training set includes multiple views captured by videos with estimated camera poses or obtained via rendering the synthetic models from random camera poses. Each object has over 30 views to support training. The test set consists of 1k data sampled from 20 held-out categories in MVImgNet2.0 that are unseen in training. Each test sample contains one or more views with estimated camera poses as input and 8 posed novel-view images with the resolution of 512×512 as the ground truths. Besides, we also provide high-quality dense point cloud reconstructions with manual cleaning by annotators for each sample in the test set as their shape ground truths.

Evaluation metrics. As MVImgNet2.0 can provide 2D multi-view ground truths, we mainly employ PSNR/SSIM (higher is better) and LPIPS [Zhang et al. 2018] (lower is better) as the evaluation metrics to measure the reconstruction quality in projected views. In the evaluation of category-agnostic reconstruction, all backgrounds of test views are masked out as in the training set to focus on the reconstruction accuracy of foreground objects, but preserved in per-scene reconstruction experiments. As TriplaneGaussian also outputs the reconstructed point cloud shape, we also employ Chamfer Distance (CD) as the measurement of the reconstructed shape quality.

Baselines. Our baselines attempt to cover a wide range of reconstruction models. For per-scene 3D reconstruction, we utilize two baselines, Instant-NGP (INGP) [Müller et al. 2022a] and 3D Gaussian Splatting (3DGS) [Kerbl et al. 2023], which are based on the technique of neural radiance field (NeRF) and Gaussian splatting [Kerbl et al. 2023], respectively. Furthermore, we adopt three large reconstruction models as the baselines of category-agnostic 3D reconstructions, which conduct data-driven shape learning from large-scale 3D data for reconstructing generic real-life objects: (i) Large Multi-View Gaussian Model (LGM) [Tang et al. 2024] that can serve for multi-view reconstruction based on the 3D Gaussian representation; (ii) Large Reconstruction Model (LRM) [Hong et al. 2023] that addresses single-view reconstruction based on the NeRF representation; (iii) Triplane Meets Gaussian Splatting (Triplane-Gaussian) [Zou et al. 2023] that addresses single-view reconstruction requires point cloud supervision in training.

Implementation details. The implementations of INGP, LGM, and TriplaneGaussian follow the official codes, while 3DGS and LRM are implemented following two open-source projects, GauStudio [Ye et al. 2024] and OpenLRM [He and Wang 2023]. We implement the base version of each 3D reconstruction model if not specified. All hyper-parameters and training strategies follow the recommended or default setting in the papers or released projects unless specified. We use NVIDIA A100 GPUs to train these baselines. In the experiments for investigating the training data factors, we apply LGM-tiny, following the ablation study setting in the LGM paper [Tang et al. 2024]. Note that LGM, in the original paper, is adopted to reconstruct from four views generated by multi-view diffusion models to address the task of image-to-3D or text-to-3D generation, in our experiments we directly feed four object views into LGM to perform multi-view reconstruction.

4.2 Per-scene 3D Reconstruction

We begin our investigation with per-scene 3D reconstruction of object-centric scenarios, utilizing two baseline methods: INGP and 3DGS. We randomly choose 50 objects from 25 categories (2 scenes for each category) to conduct experiments. Among a total of 30 views for each object, we randomly select 20 of them for optimizing the parameters in INGP [Müller et al. 2022a] and 3DGS [Kerbl et al. 2023], and use the remaining 10 views for evaluation. We design two sets of controlled experiments, each differentiated by a single variable: the camera poses utilized during training. For the first group, the camera poses are estimated via the annotation approach used in MVImgNet (MV1-Anno), while the second group employs the advanced approach in MVImgNet2.0 (MV2-Anno).

Results. We compute the results averaged across the selected 50 scenes. The quantitative results are shown in Tab. 2, where both INGP and 3DGS can get a more accurate reconstruction using the camera poses estimated via the advanced approach in MVImgNet2.0. By using the camera poses estimated by the advanced approach in MVImgNet2.0, INGP can achieve a higher average PSNR by ~1.1dB, and 3DGS can achieve a significant improvement of ~5.8dB in PSNR. It validates the higher quality of camera poses estimated in MVImgNet2.0. These results also indicate that MVImgNet2.0 can better support the learning-based reconstruction methods in the task of per-scene reconstruction or novel view synthesis.

Table 2. **Per-scene 3D reconstruction quality** comparison when using the estimated camera poses by the annotation manners in MVImgNet (MV1-Anno) and MVImgNet (MV2-Anno) for training. Two baselines are used: Instant-NGP (INGP) and 3D Gaussian splatting (3DGS).

Baseline	MV1-Anno	MV2-Anno	PSNR↑	SSIM↑	LIPIS↓
INGP [Müller et al. 2022a]	✓		36.05	0.980	0.023
		✓	**37.17**	**0.984**	**0.018**
3DGS [Huang et al. 2024]	✓		36.44	0.982	0.027
		✓	**42.19**	**0.991**	**0.015**

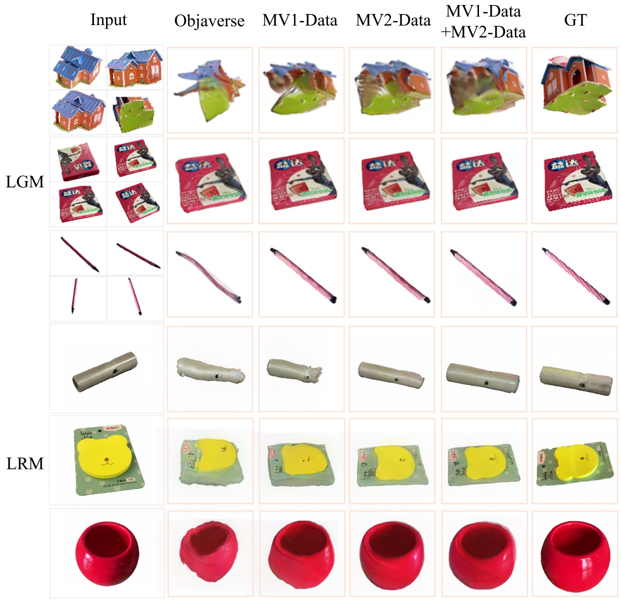

Fig. 7. Qualitative results of LGM and LRM trained on different data from Objaverse, MV1-Data, and MV2-Data.

4.3 Category-agnostic 3D Reconstruction

On the task of category-agnostic 3D reconstruction, we perform experiments to evaluate the performance of large reconstruction models trained on different kinds of data to validate the value of MVImgNet2.0. We apply an LGM [Tang et al. 2024] for the task of multi-view reconstruction, and an LRM [Hong et al. 2023] and a TriplaneGaussian [Zou et al. 2023] to address the single-view reconstruction. By using different kinds of multi-view data in training, we investigate their effect on the learning of large reconstruction models. Besides, we also use different kinds of point cloud supervision in training TriplaneGaussian to further validate the value of the point cloud annotations in MVImgNet2.0. Finally, we deeply investigate the factor of data scale, category range, and view range in training data by evaluating their effects on the performance of a tiny LGM (LGM-tiny).

Experiments on LGM and LRM. We train LGM and LRM on three kinds of data: (i) synthetic data from Objaverse [Deitke et al. 2022]; (ii) real data from MVImgNet [Yu et al. 2023] (MV1-Data); and (iii) added data in MVImgNet2.0 (MV2-Data). In training, we use 4 randomly selected views as input, and 20 views as supervision to train an LGM-base model, and use 1 randomly selected input view to train an LRM-base model. As shown in Tab. 3, the quantitative results of LGM and LRM consistently demonstrate: 1) Using real data for training can derive a stronger large reconstruction model when applied to real-world objects. Compared with using Objaverse data for training, using MV1-Data for training can achieve a higher PSNR by ~1.9dB for LGM and by ~0.7dB for LRM; 2) Though with a close scale, using MV2-Data only (300k objects) can lead to a higher reconstruction quality (~0.5dB↑ for both LRM and LGM) compared with using MV1-Data only (220k objects), thanks to the higher quality of MV2-Data; 3) Further incorporating MV2-Data into MV1-Data in training can further bring performance gains, *i.e.*, ~0.4dB and ~0.3dB for LGM and LRM, respectively, which additionally confirm the value of MVImgNet2.0 in benefiting data-driven 3D shape learning. We also provide some qualitative results in Fig. 7 to visualize the differences in reconstruction quality.

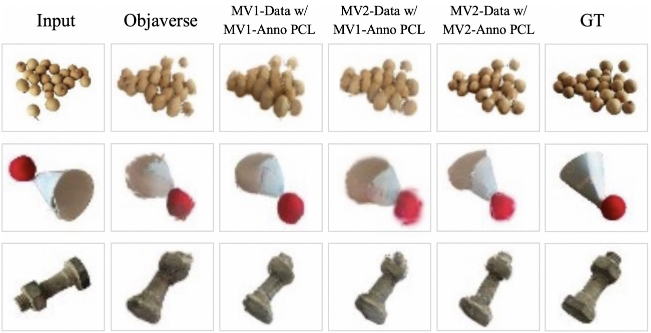

Fig. 8. Qualitative results of TriplaneGaussian trained on different view data and point cloud supervision.

Experiments on TriplaneGaussian. We train TriplaneGaussian with different multi-view data and point cloud supervision to further investigate the value of 360°-view data and the proposed point cloud annotations in MVImgNet2.0. As shown in Tab. 4, training on the Objaverse synthetic data results in poor generalization on real data, getting the lowest PSNR in rendering quality. However, since Objaverse can provide perfect point cloud supervision sampled from the ground-truth object surface, training on Objaverse can lead to an acceptable level of shape quality. As MVImgNet only collects 180° views for each object, training on MV1-Data with incomplete point cloud supervision leads to low quality in both 2D rendering and 3D shape. Training on 360° views (MV2-Data) but with point cloud supervision obtained via the annotation approach in MVImgnet (MV1-Anno) can bring better reconstruction results in rendering quality but poor performance in shape quality. Further using higher-quality point cloud supervision (MV2-Anno) can lead to improvements in the overall reconstruction quality. To sum up, the model trained on 360° real-world data (MV2-Data) with more complete point cloud supervision (MV2-Anno) can achieve a higher rendering quality by ~1.0dB in PSNR and also a lower Chamfer distance by 4×10^{-4} than the one trained on Objaverse. Qualitative results in Fig. 8 also confirm our claim. Thus, the experiment results validate the value of 360°-view data and the higher quality of point cloud annotations provided in MVImgNet2.0.

Training data factors. We further investigate the effects of three factors in training data on training large reconstruction models:

Table 3. **Generalizable 3D reconstruction quality** comparison when using different set of multi-view data for training. Note that when neither MVImgNet data (MV1-Data) nor MVImgNet2.0 data (MV2-Data) is used, the baseline model is trained on synthetic data from the Objaverse dataset. Two baselines are used: LRM and LGM, for single-view and multi-view reconstruction, respectively.

Baseline	MV1-Data	MV2-Data	PSNR↑	SSIM↑	LIPIS↓
LGM [Tang et al. 2024]			23.59	0.932	0.050
	✓		25.49	0.951	0.035
		✓	26.01	0.953	0.034
	✓	✓	**26.41**	**0.956**	**0.032**
LRM [Hong et al. 2023]			20.76	0.929	0.065
	✓		21.42	0.932	0.059
		✓	21.97	0.935	0.056
	✓	✓	**22.27**	**0.936**	**0.033**

Table 4. **Generalizable 3D reconstruction quality** comparison when using different multi-view data and point cloud supervision to train a Triplane-Gaussian model. The first row uses only Objaverse synthetic data in training, while the second row uses MVImgNet real data (MV1-Data) for training with point clouds obtained via the annotation manner in MVImgNet (MV1-Anno). The last two rows use added MVImgNet2.0 data (MV2-Data) for training, with point cloud supervision (PCL super.) are obtained via different annotation approaches used by MVImgnet (MV1-Anno) and MVImgNet2.0 (MV2-Anno).

MV1-Data	MV2-Data	PCL super.	PSNR↑	SSIM↑	LIPIS↓	CD↓ ($\times 10^{-2}$)
		Objaverse	21.79	0.923	0.062	0.40
✓		MV1-Anno	22.43	0.924	0.060	0.82
	✓	MV1-Anno	22.51	0.928	0.056	0.89
	✓	MV2-Anno	**22.77**	**0.929**	**0.053**	**0.36**

data scale, category range, and view range. The baseline model is based on an LGM-tiny for multi-view reconstruction, and the image resolution for training and testing is 256×256 for efficiency. In the first group of experiments, we basically use 40k MV1-Data and progressively increase the number of MV2-Data from 20k to 140k added for training. As shown in Fig. 9(a), LGM achieves increasingly better performance with the added MV2-Data grows in scale.In the second group, we hold the added MV2-Data volume constant as 100k but increase the number of categories covered in these data. As the category range expanded, LGM can also achieve iterative performance gains, as shown in Fig. 9(b). Next, we control the ratio of 360°-view data in a total of 100k MV2-Data added, and find that a higher ratio of 360°-view data is positive to the final reconstruction quality of LGM (see Fig. 9(c)). These results indicate that training on data of a larger scale, covering richer categories, and capturing a wider view range are important to improve the performance of large reconstruction models. Finally, we use a constant scale of 100k data (MV1-Data + MV2-Data) in total for training, but gradually increase the ratio of MV2-Data from 0% to 100%. It leads to a gap of ~0.5dB between using 0% and 100% of MV2-Data, as shown in Fig 9(d), which demonstrates that with higher quality, MV2-Data is of larger value than MV1-Data for the learning of large reconstruction models. Some qualitative results of LGM-tiny are included in SupMat's Fig. R.2.

5 CONCLUSION

In this paper, we propose MVImgNet2.0, a larger-scale dataset with multi-view images for generic real-world objects. It expands the

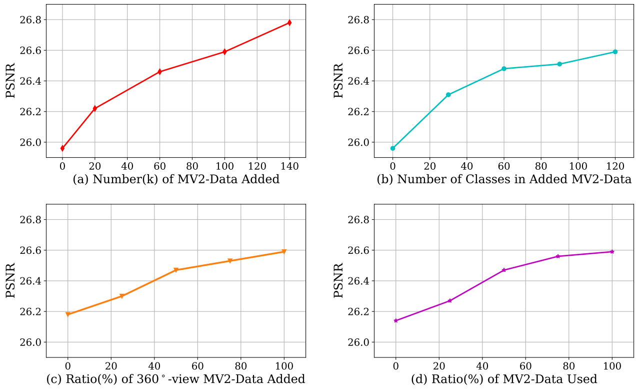

Fig. 9. The experiments for analyzing the effects of three factors in training data to the learning of LGM, including data scale (a), category range (b), and view range (c).

MVImgNet dataset and doubles the original scale and category range. Besides, MVImgNet2.0 also advances the data processing approaches to provide annotations of higher quality, including the foreground object masks, the estimated camera poses, and the reconstructed point clouds. To validate the value of MVImgNet2.0 data, we conduct extensive experiments on the task of 3D reconstruction and demonstrate that MVImgNet2.0 data not only can be used as high-quality per-scene reconstruction data but also is promising to provide a stronger 3D prior for generalizable object reconstruction. All data and annotations will be released to the public to inspire the computer vision and graphics communities.

Limitations and future work. MVImgNet2.0, while impressive, does have some limitations. Firstly, due to the challenges in capturing, we have excluded large-scale objects such as buildings and dynamic subjects like animals that are difficult to stabilize for imaging. Future multi-view datasets could aim to capture these "hard categories" to facilitate learning a more comprehensive 3D prior. Secondly, there is significant potential for improvement in annotation quality, particularly in dense reconstructions, which is an important area for future work, especially with the ongoing advancements in dense reconstruction techniques. Thirdly, the majority of the videos focus on a single central object, resulting in camera trajectories that are relatively monotonous, *i.e.* orbiting around the main subject. We plan to expand our collection to include videos with complex arrangements of multiple objects and to explore alternative camera trajectory modes to support more sophisticated scene-level reconstructions.

ACKNOWLEDGEMENTS

The work was supported in part by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, Guangdong Provincial Outstanding Youth Fund(No. 2023B1515020055), NSFC-61931024, and Shenzhen Science and Technology Program No. JCYJ20220530143604010. It is also partly supported by the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104,by Key Area R&D Program of Guangdong

Province (Grant No. 2018B030338001), by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and by Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055).

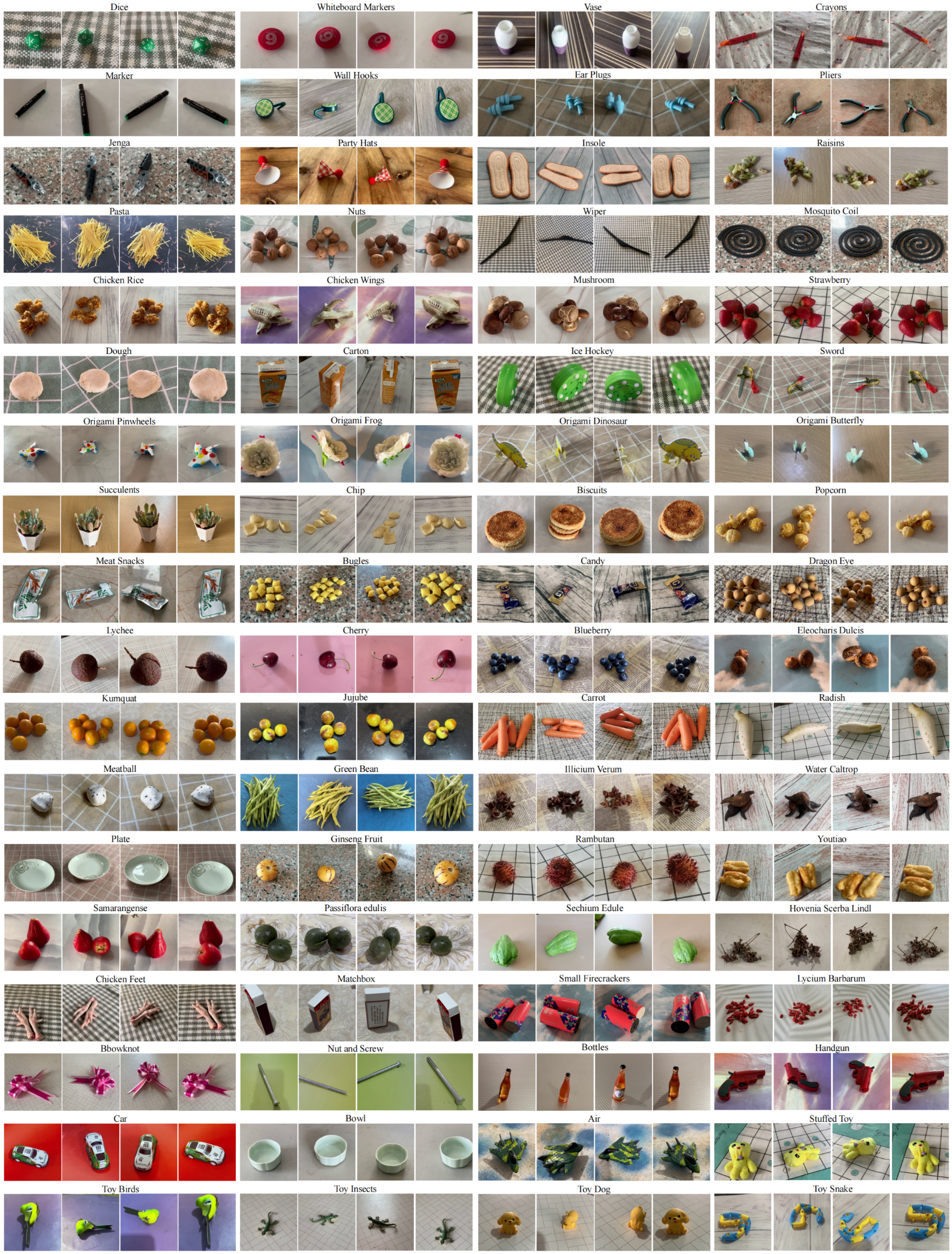

Fig. 10. Visualization of more multi-view data from different object categories in MVImgNet2.0.

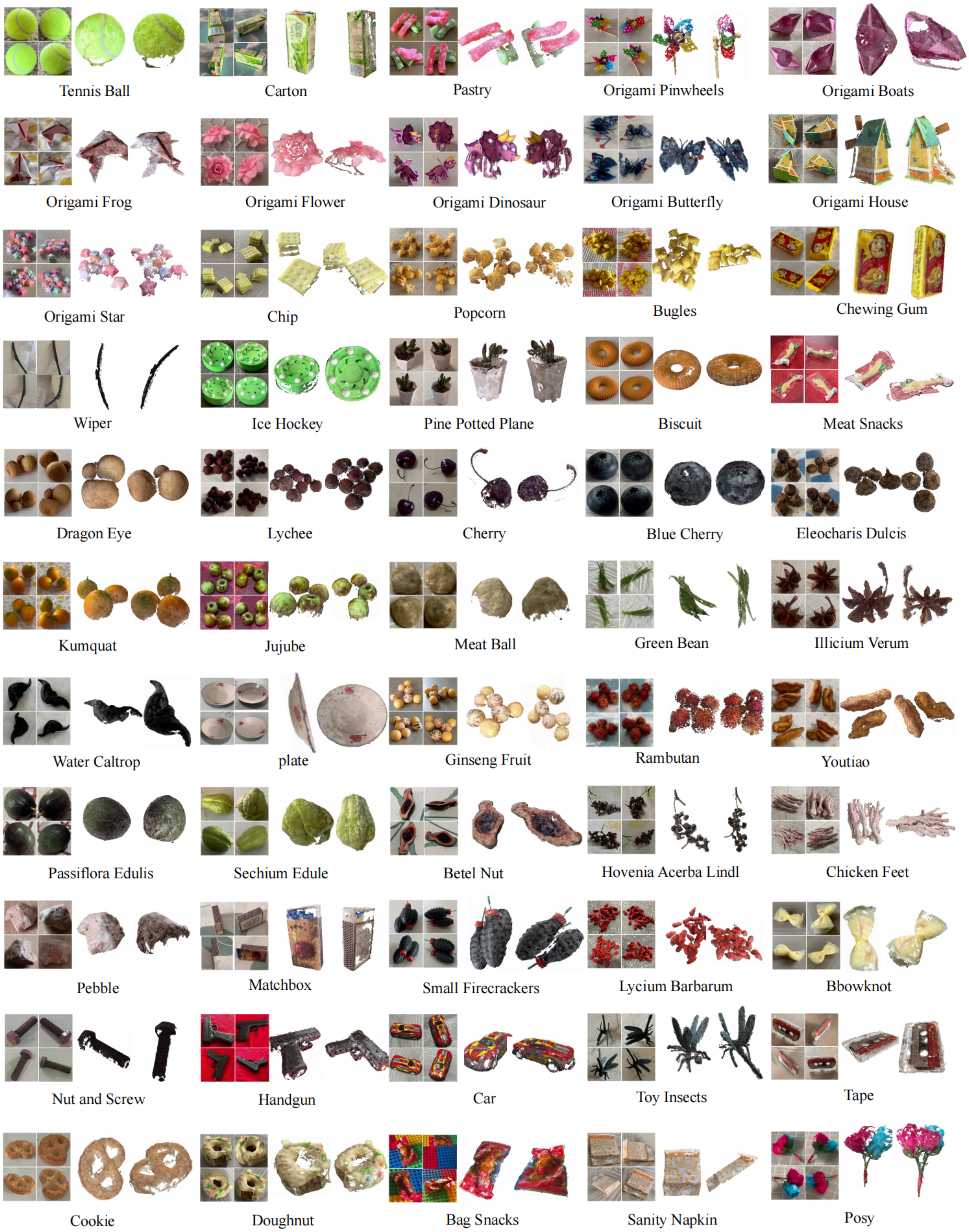

Fig. 11. Visualization of more reconstructed point cloud annotations from different object categories in MVImgNet2.0.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. 2011. Building rome in a day. *Commun. ACM* 54, 10 (2011), 105–112.

Titas Anciukevicius, Fabian Manhardt, Federico Tombari, and Paul Henderson. 2024. Denoising diffusion via image-based rendering. *arXiv preprint arXiv:2402.03445* (2024).

Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. 2023. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12608–12618.

Arthur Aubret, Céline Teulière, and Jochen Triesch. 2024. Self-supervised visual learning from interactions with objects. *arXiv preprint arXiv:2407.06704* (2024).

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 561–578.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. 2024a. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6593–6602.

Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. 2024b. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738* (2024).

Zhiqin Chen and Hao Zhang. 2019. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5939–5948.

Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2524–2534.

Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 628–644.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. 2022. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21126–21136.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663* (2023).

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2022. Objaverse: A Universe of Annotated 3D Objects. *arXiv preprint arXiv:2212.08051* (2022).

Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. 2023. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20637–20647.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 248–255.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. 2022. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2553–2560.

Haoqiang Fan, Hao Su, and Leonidas J Guibas. 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 605–613.

Yasutaka Furukawa and Jean Ponce. 2009. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence* 32, 8 (2009), 1362–1376.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. 2024. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314* (2024).

Georgia Gkioxari, Jitendra Malik, and Justin Johnson. 2019. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9785–9795.

Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. 2020. Shape and viewpoint without keypoints. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 88–104.

Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2495–2504.

Junlin Han, Filippos Kokkinos, and Philip Torr. 2024. Vfusion3d: Learning scalable 3d generative models from video diffusion models. *arXiv preprint arXiv:2403.12034* (2024).

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2024. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101* (2024).

Zexin He and Tengfei Wang. 2023. OpenLRM: Open-Source Large Reconstruction Models. https://github.com/3DTopia/OpenLRM.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400* (2023).

Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024. 2d gaussian splatting for geometrically accurate radiance fields. *arXiv preprint arXiv:2403.17888* (2024).

Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin Brualla, Kaushal Patel, et al. 2023. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* 36 (2023), 76061–76084.

Wonbong Jang and Lourdes Agapito. 2021. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12949–12958.

Wonbong Jang and Lourdes Agapito. 2024. NViST: In the Wild New View Synthesis from a Single Image with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10181–10193.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the ACM International Conference on Machine Learning (ICML)*. PMLR, 4904–4916.

Hanwen Jiang, Qixing Huang, and Georgios Pavlakos. 2024. Real3D: Scaling Up Large Reconstruction Models with Real-World Images. *arXiv preprint arXiv:2406.08479* (2024).

Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. 2018. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 371–386.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

Junlong Ke, Zichen Wen, Yechenhao Yang, Chenhang Cui, Yazhou Ren, Xiaorong Pu, and Lifang He. 2024a. Integrating Vision-Language Semantic Graphs in Multi-View Clustering. IJCAI.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. 2024b. Segment anything in high quality. *Advances in Neural Information Processing Systems* 36 (2024).

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)* 123 (2017), 32–73.

Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. 2020. Articulation-aware canonical surface mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 452–461.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)* 128, 7 (2020), 1956–1981.

Han-Hung Lee, Yiming Zhang, and Angel X Chang. 2024. Duoduo CLIP: Efficient 3D Understanding with Multi-View Images. *arXiv preprint arXiv:2406.11579* (2024).

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the ACM International Conference on Machine Learning (ICML)*. PMLR, 19730–19742.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. 2023c. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214* (2023).

Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. 2020. Self-supervised single-view 3d reconstruction via semantic consistency. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 677–693.

Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. 2023b. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8456–8465.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. VILA: On Pre-training for Visual Language Models. arXiv:2312.07533 [cs.CV]

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 740–755.

Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. 2021. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5987–5997.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* 36 (2024).

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023a. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9298–9309.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. 2024b. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. arXiv:2402.17177 [cs.CV]

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4460–4470.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2630–2640.

B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

George A. Miller. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. 2022. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 306–315.

Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. 2022b. Autorf: Learning 3d object radiance fields from single view observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3971–3980.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022a. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–15.

Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. 2020. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3504–3515.

David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. 2019. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7688–7697.

Evangelos Ntavelis, Aliaksandr Siarohin, Kyle Olszewski, Chaoyang Wang, Luc V Gool, and Sergey Tulyakov. 2023. Autodecoding latent 3d diffusion models. *Advances in Neural Information Processing Systems* 36 (2023), 67021–67047.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 165–174.

Marc Pollefeys, David Nistér, J-M Frahm, Amir Akbarzadeh, Philippos Mordohai, Brian Clipp, Chris Engels, David Gallup, S-J Kim, Paul Merrell, et al. 2008. Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision* 78 (2008), 143–167.

Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. 2004. Visual modeling with a hand-held camera. *International Journal of Computer Vision* 59 (2004), 207–232.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the ACM International Conference on Machine Learning (ICML)*. PMLR, 8748–8763.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. 2021. 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 (ICCV)*.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159 [cs.CV]

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Joseph Roth, Yiying Tong, and Xiaoming Liu. 2016. Adaptive 3D face reconstruction from unconstrained photo collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4197–4206.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494.

Johannes L Schonberger and Jan-Michael Frahm. 2016a. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4104–4113.

Johannes L Schonberger and Jan-Michael Frahm. 2016b. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4104–4113.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 501–518.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), 25278–25294.

Nikita Selin. 2019–2024. CarveKit: Image Background Remove Tool. https://github.com/OPHoperHPO/image-background-remove-tool.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

Qiuhong Shen, Xingyi Yang, and Xinchao Wang. 2023. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261* (2023).

Yuan Shen, Duygu Ceylan, Paul Guerrero, Zexiang Xu, Niloy J Mitra, Shenlong Wang, and Anna Früstück. 2024. SuperGaussian: Repurposing Video Models for 3D Super Resolution. *arXiv preprint arXiv:2406.00609* (2024).

Zhelun Shen, Yuchao Dai, and Zhibo Rao. 2021. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13906–13915.

Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. 2015. Hierarchical image saliency detection on extended CSSD. *IEEE transactions on pattern analysis and machine intelligence* 38, 4 (2015), 717–729.

Noah Snavely, Steven M Seitz, and Richard Szeliski. 2006. Photo tourism: exploring photo collections in 3D. In *ACM siggraph 2006 papers*. 835–846.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2024. LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation. *arXiv preprint arXiv:2402.05054* (2024).

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22819–22829.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. 2017. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2626–2634.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1588–1597.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. 2018. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*. 52–67.

Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. 2023. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024* (2023).

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. 2023a. Reconfusion: 3d reconstruction with diffusion priors. *arXiv preprint arXiv:2312.02981* (2023).

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. 2024. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21551–21561.

Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. 2020. Pq-net: A generative part seq2seq network for 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 829–838.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. 2023b. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 803–814.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1912–1920.

Shaoan Xie, Yang Zhao, Zhisheng Xiao, Kelvin CK Chan, Yandong Li, Yanwu Xu, Kun Zhang, and Tingbo Hou. 2023. Dreaminpainter: Text-guided subject-driven image inpainting with diffusion models. *arXiv preprint arXiv:2312.03771* (2023).

Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. 2019. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems* 32 (2019).

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. 2024. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621* (2024).

Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. 2023. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217* (2023).

Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. 2016. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. *Advances in neural information processing systems* 29 (2016).

Zongxin Yang and Yi Yang. 2022. Decoupling features in hierarchical propagation for video object segmentation. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), 36324–36336.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*. 767–783.

Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5525–5534.

Chongjie Ye. 2023. Instant-angelo: Build high-fidelity Digital Twin within 20 Minutes. https://github.com/hugoycj/Instant-angelo.

Chongjie Ye, Yinyu Nie, Jiahao Chang, Yuantao Chen, Yihao Zhi, and Xiaoguang Han. 2024. GauStudio: A Modular Framework for 3D Gaussian Splatting and Beyond. *arXiv preprint arXiv:2403.19632* (2024).

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4578–4587.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. 2023. MVImgNet: A Large-scale Dataset of Multi-view Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. 2024. GS-LRM: Large Reconstruction Model for 3D Gaussian Splatting. *arXiv preprint arXiv:2404.19702* (2024).

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 586–595.

Haidong Zhu, Tianyu Ding, Tianyi Chen, Ilya Zharkov, Ram Nevatia, and Luming Liang. 2023. CaesarNeRF: Calibrated Semantic Representation for Few-shot Generalizable Neural Rendering. *arXiv preprint arXiv:2311.15510* (2023).

Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. 2023. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. *arXiv preprint arXiv:2312.09147* (2023).

Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Rui Peng, Siyu Zhu, Zilong Dong, Liefeng Bo, et al. 2024. Videomv: Consistent multi-view generation based on large video generative model. *arXiv preprint arXiv:2403.12010* (2024).

A MORE DETAILS ABOUT MVIMGNET2.0 DATA

Per-category data distributions. We count the number of object videos in each category in the proposed MVImgNet2.0 dataset. Note that among all 347 classes, 70 are old classes from MVImgNet [Yu et al. 2023], and these classes are not required to collect more than 1000 videos in the data acquisition process of MVImgNet2.0. Excluding them, ~60% of classes (164/277) cover 1000 or more objects. We provide a histogram for the number of objects in each MVImgNet2.0 new class in Fig. R.1. As shown, categories can be divided into three groups: the first group is "hard classes", where less than 500 videos can be collected, while another two groups of categories can get around 1000 and even 2000 videos collected, respectively. A more detailed statistics of the number of objects in each category is shown in Fig. R.3.

Category taxonomy. The category taxonomy is shown in Fig. R.4 to better exhibit the categories and their hierarchical relationships in MVImgNet2.0.

Table R.1. Quantitative segmentation results (MSE↓ $\times 10^{-1}$) on the ECCSD dataset, the DAVIS dataset, and a subset of 500 MVImgNet images (MV1-500) with ground-truth object masks.

Methods	ECCSD	DAVIS	MV1-500
MV1-Anno	0.143	0.195	0.243
MV2-Anno (ours)	0.103	0.143	0.172

B MORE EXPERIMENTS

Mask annotation quality. In the annotation process, we adopt a detection-segmentation-tracking pipeline to generate the foreground object mask in each view. To better demonstrate the superiority of the used segmentation manner over the original one in MVImgNet, we further evaluate the performance of this pipeline on a subset of MVImgNet data where 500 frames (randomly selected from different categories) are manually annotated with object segmentation masks, and also on other object-centric datasets with ground-truth object masks, *i.e.*, ECSSD [Shi et al. 2015] and DAVIS [Pont-Tuset et al. 2017]. The segmentation results of CarveKit [Selin 2024] (the original manner, denoted as MV1-Anno) and the advanced one (MV2-Anno) are presented in Tab. R.1. As shown, the performance of MV2-Anno surpasses MV1-Anno by a considerable margin.

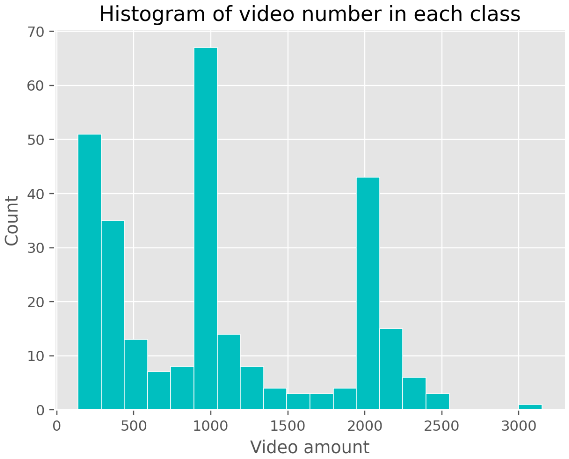

Fig. R.1. The histogram of object number in each class.

More qualitative results. We visualize more results of LGM-tiny for multi-view object reconstruction. We mainly show the reconstruction quality of LGM-tiny when trained with MVImgNet1.0 data (MV1-Data), MVImgNet2.0 data (MV1-Data), and both of them. As shown in Fig. R.2, LGM trained with MV2-Data can achieve a higher reconstruction ability, and the use of both of them can further lead to improvements.

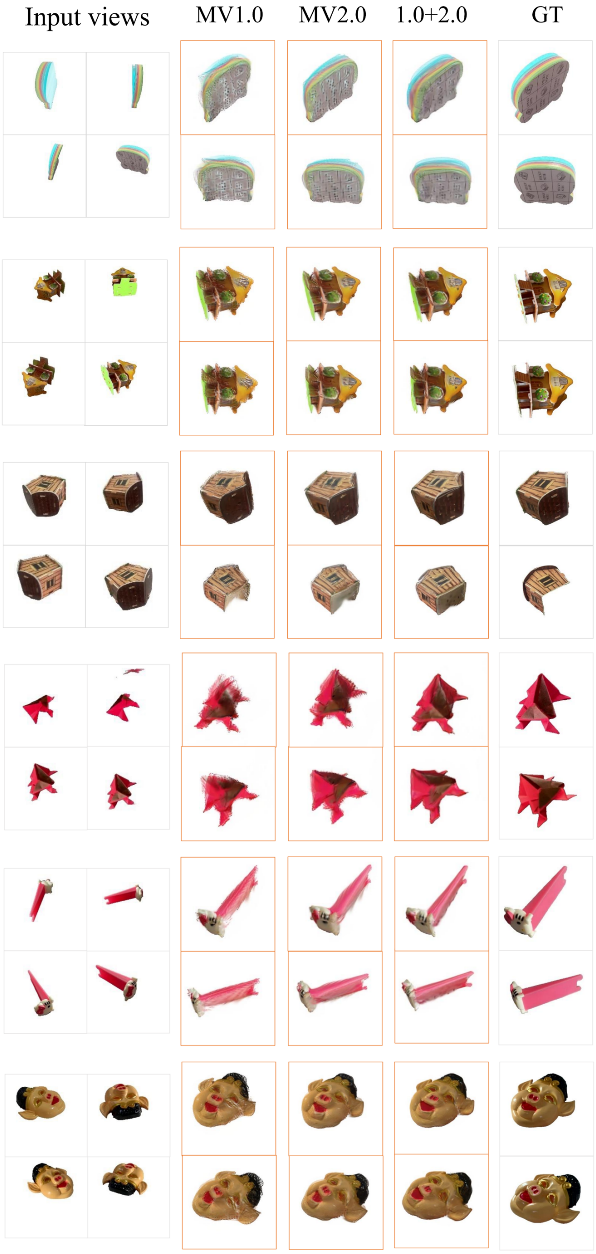

Fig. R.2. More qualitative results of LGM-tiny when trained with different kinds of data: MVImgNet1.0 (MV1.0), MVImgNet2.0 (MV2.0), and both.

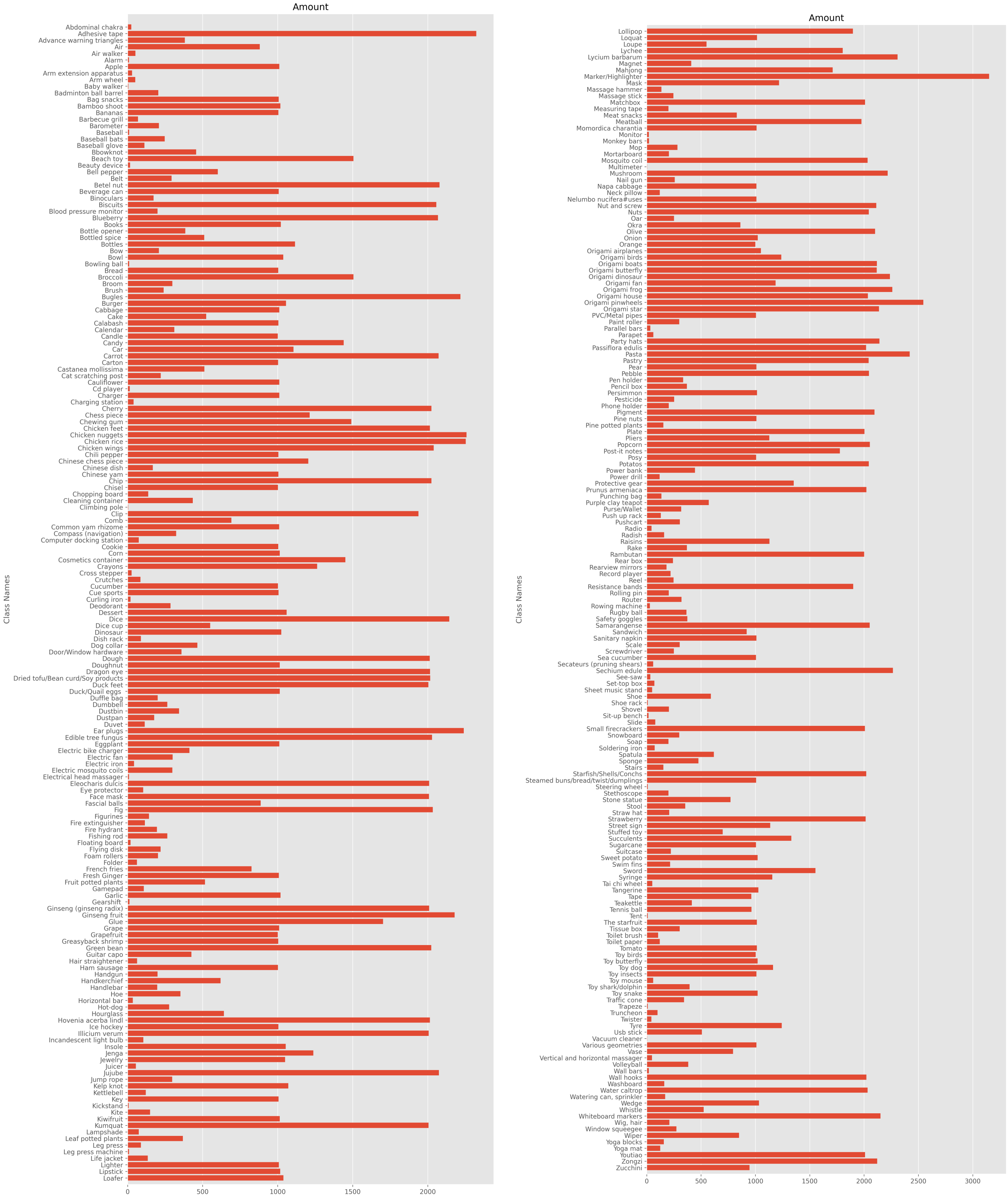

Fig. R.3. Amounts of objects in each category in the proposed MVImgNet2.0 dataset.

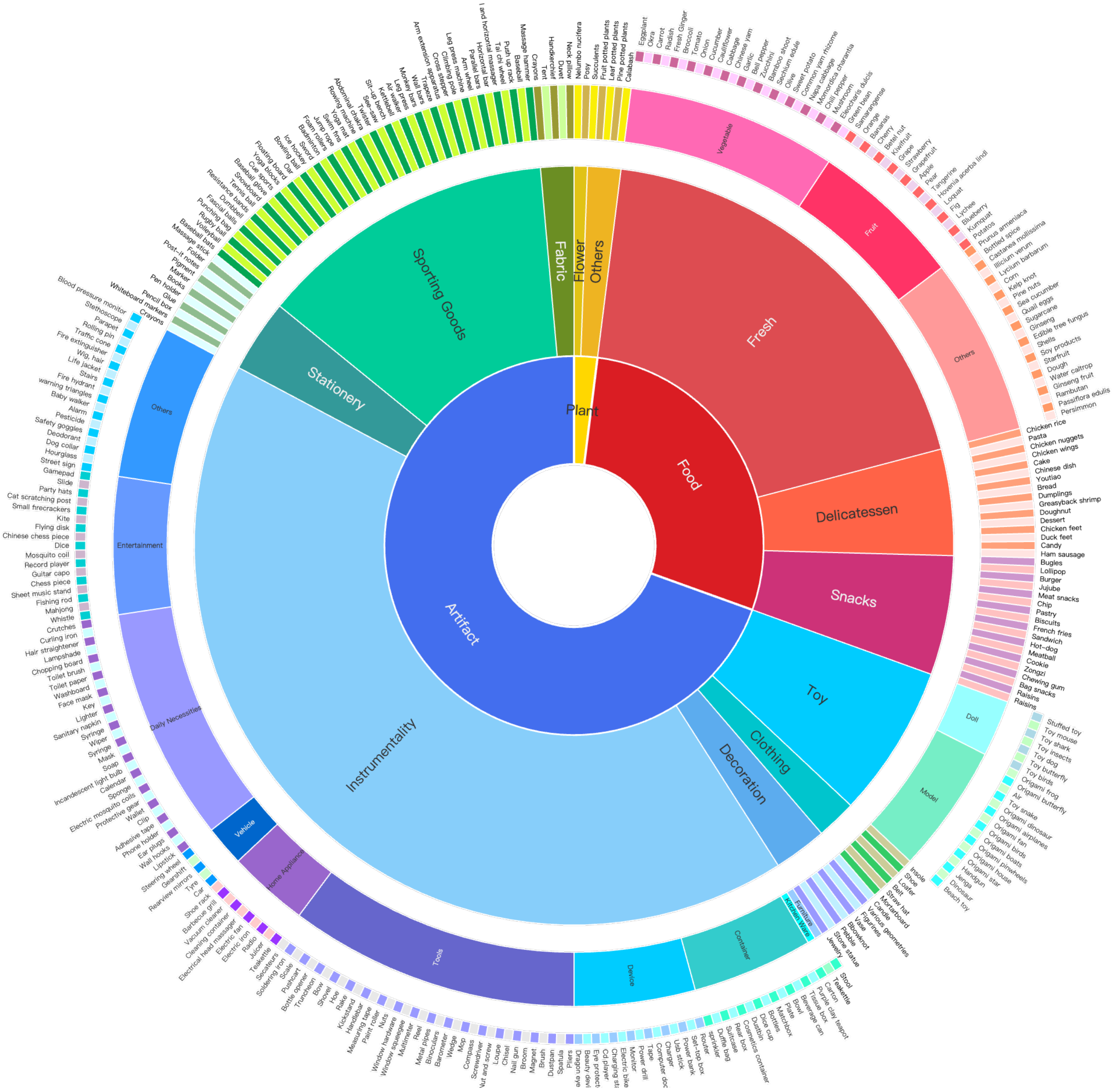

Fig. R.4. The category taxonomy of the proposed MVImgNet2.0 dataset.