

MVImgNet: A Large-scale Dataset of Multi-view Images

Xianggang Yu* Mutian Xu*,† Yidan Zhang* Haolin Liu* Chongjie Ye*
 Yushuang Wu Zizheng Yan Chenming Zhu Zhangyang Xiong Tianyou Liang
 Guanying Chen Shuguang Cui Xiaoguang Han‡

*equal technical contribution †part of project lead ‡corresponding author

SSE, CUHK SZ FNii, CUHK SZ

gaplab.cuhk.edu.cn/projects/MVImgNet



Figure 1. It is time to embrace **MVImgNet**! We introduce MVImgnet, a large-scale dataset of *multi-view images*, which is efficiently collected by shooting videos of real-world objects. It enjoys 3D-aware signals from multi-view consistency, being a soft bridge between 2D and 3D vision. Through dense reconstruction on MVImgNet, we also present a large-scale real-world 3D object *point cloud* dataset – MVPNet. **Exterior:** Examples of various multi-view images in MVImgNet (see Fig. 3 more intuitively). **Interior:** Instances of colorful point clouds from MVPNet (see Fig. 7 more clearly) are assembled into a stereo sign of ‘MVImgNet’.

Abstract

Being data-driven is one of the most iconic properties of deep learning algorithms. The birth of ImageNet [24] drives a remarkable trend of ‘learning from large-scale data’ in computer vision. Pretraining on ImageNet to obtain rich universal representations has been manifested to benefit various 2D visual tasks, and becomes a standard in 2D vision. However, due to the laborious collection of real-world 3D data, there is yet no generic dataset serving as a counterpart of ImageNet in 3D vision, thus how such a dataset can impact the 3D community is unraveled. To remedy this defect, we introduce **MVImgNet**, a large-scale dataset of multi-view images, which is highly convenient to gain by shooting videos of real-world objects in human daily life. It contains **6.5 million** frames from **219,188**

videos crossing objects from **238** classes, with rich annotations of object masks, camera parameters, and point clouds. The multi-view attribute endows our dataset with 3D-aware signals, making it a soft bridge between 2D and 3D vision.

We conduct pilot studies for probing the potential of MVImgNet on a variety of 3D and 2D visual tasks, including radiance field reconstruction, multi-view stereo, and view-consistent image understanding, where MVImgNet demonstrates promising performance, remaining lots of possibilities for future explorations.

Besides, via dense reconstruction on MVImgNet, a 3D object point cloud dataset is derived, called **MVPNet**, covering **87,200** samples from **150** categories, with the class label on each point cloud. Experiments show that MVPNet can benefit the real-world 3D object classification while posing new challenges to point cloud understanding.

MVImgNet and MVPNet will be public, hoping to inspire the broader vision community.

Author contributions listed at end of the paper.

1. Introduction

Being data-driven, also known as data-hungry, is one of the most important attributes of deep learning algorithms. By training on large-scale datasets, deep neural networks are able to extract rich representations. In the past few years, the computer vision community has witnessed the bloom of such ‘*learning from data*’ regime [42, 43, 54], after the birth of ImageNet [24] – the pioneer of large-scale real-world image datasets. Notably, pretraining on ImageNet is well-proven to boost the model performance when transferring the pretrained weights into not only high-level [35, 40, 41, 59, 60] but also low-level visual tasks [15, 56], and becomes a *de-facto* standard in 2D. Recently, various 3D datasets [5, 9, 11, 23, 34, 95, 103] are produced to facilitate 3D visual applications.

However, due to the non-trivial scanning and laborious labeling of real-world 3D data (commonly organized in point clouds or meshes), existing 3D datasets are either synthetic or their scales are not comparable with ImageNet [24]. Consequently, unlike in 2D vision where models are usually pretrained on ImageNet to gain universal representation or commonsense knowledge, most of the current methods in 3D area are directly trained and evaluated on particular datasets for solving specific 3D visual tasks (*e.g.*, NeRF dataset [63] and ShapeNet [11] for novel view synthesis, ModelNet [95] and ScanObjectNN [87] for object classification, KITTI [34] and ScanNet [23] for scene understanding). Here, two crucial and successive issues can be induced: **(1) There is still no generic dataset in 3D vision, as a counterpart of ImageNet in 2D.** **(2) What benefit such a dataset can endow to 3D community is yet unknown.** In this paper, we focus on investigating these two problems and set two corresponding targets: Build the primary dataset, then explore its effect through experiments.

Milestone 1 – Dataset:

For a clearer picture of the first goal, we start by carefully revisiting existing 3D datasets as well as ImageNet [24]. **i)** 3D synthetic datasets [11, 95] provide rich 3D CAD models. However, they lack *real-world* cues (*e.g.*, context, occlusions, noises), which are indispensable for model robustness in practical applications. ScanObjectNN [87] extracts real-world 3D objects from indoor scene data, but is limited in scale. For 3D scene-level dataset [5, 10, 23, 34, 38, 80], their scales are still constrained by the laborious scanning and labeling (*e.g.*, millions of points per scene). Additionally, they contain specific inner-domain knowledge such as a particularly intricate indoor room or outdoor driving configurations, making it hard for general transfer learning. **ii)** Although ImageNet [24] contains the most comprehensive real-world objects, it only describes a 2D world that misses *3D-aware* signals. Since humans live in a 3D world, 3D consciousness is vitally important for realizing human-like intelligence and solving real-life visual problems.

Based on the above review, our dataset is created from a new insight – **multi-view images**, as a soft bridge between 2D and 3D. It lies several benefits to remedying the aforementioned defects. Such data can be *easily gained in considerable sizes* via shooting an object around different views on common mobile devices with cameras (*e.g.*, smartphones), which can be collected by crowd-sourcing in *real world*. Moreover, the multi-view constraint can bring natural 3D visual signals (later experiments show that this not only benefits 3D tasks but also 2D image understanding). To this end, we build **MVImgNet**, containing **6.5** million frames from **219,188** videos crossing real-life objects from **238** classes, with rich annotations of object masks, camera parameters, and point clouds. You may take a glance at our MVImgNet from Fig. 1.

Milestone 2 – Experimental Exploration:

Now facing the second goal of this paper, we attempt to probe the power of our dataset by conducting some pilot experiments. Leveraging the multi-view nature of the data, we start by focusing on the view-based 3D reconstruction task and demonstrate that pretraining on MVImgNet can not only benefit the *generalization ability of NeRF* (Sec. 4.1), but also *data-efficient multi-view stereo* (Sec. 4.2). Moreover, for image understanding, although humans can easily recognize one object from different viewpoints, deep learning models can hardly do that robustly [26]. Considering MVImgNet provides numerous images of a particular object from different viewpoints, we verify that MVImgNet-pretrained models are endowed with decent *view consistency* in general *image classification* (supervised learning in Sec. 5.1, self-supervised contrastive learning in Sec. 5.2) and *salient object detection* (Sec. 5.3).

Bonus – A New 3D Point Cloud Dataset – MVPNet:

Through dense reconstruction on MVImgNet, a new 3D object point cloud dataset is derived, called **MVPNet**, which contains **87,200** point clouds with 150 categories, with the class label on each point cloud (see Fig. 7). Experiments show that MVPNet not only benefits the real-world 3D object classification task but also poses new challenges and prospects to point cloud understanding (Sec. 6).

MVImgNet and MVPNet will be public, hoping to inspire the broader vision community.

2. Related Work

Single-view image datasets. The MNIST database [1] is one of the most pioneering datasets, composed of 70k monochrome images of handwritten digits. The CIFAR10 and CIFAR100 datasets proposed by [53], respectively collect 60k tiny color images (32×32) of various common objects or animals in 10 and 100 classes. ImageNet [24] is presented with a large scale, high accuracy, large diversity, and hierarchical structure, which provides opportunities for training deep neural networks. In detection

and segmentation tasks, MSCOCO [57] is one of the most popular datasets containing 328k images with rich annotations. Some other datasets include PASCAL VOC [28], Visual Genome [52], Cityscapes [22], MPII [4], etc. Although these datasets facilitate the development of deep visual learning, they mainly serve for 2D single-view image understanding, which limits their applications in 3D vision.

Video datasets. Another line is video datasets. Pioneering works construct the HMDB-51 [55] and UCF-101 [81] dataset. Afterward, the ActivityNet [8] and Kinetics [48] datasets are constructed of a larger scale and variation, of which the latter has collected 650k video clips that cover 700 classes. Besides, some datasets are built for human pose estimation [47, 89] and object detection/segmentation/tracking [68, 94, 111]. IEMOCAP [7] provides video data for the task of multimodal emotion recognition. MSVD [14] and MSR-VTT [102] annotate videos with extra captions. Further, HowTo100M [62] proposes a larger-scale one of 136 million samples from narrated instructional videos. ActivityNet Captions [51] introduces the task of dense-captioning events and constructs a dataset with 20k videos. These datasets are primarily for video understanding which is different from our objective.

3D datasets. As the applications in 3D vision attract increasing attention, various 3D datasets are proposed. One line of works focuses on indoor scenes [5, 10, 23, 46, 77, 80, 99], where S3DIS [5] and ScanNet [23] are two of most popular datasets. In addition, some works provide 3D object point clouds for contextual object surface reconstruction [6, 30]. 3D outdoor scenes are scanned via LiDAR sensors [9, 34, 38, 82] for autonomous driving. ShapeNet [11] and ModelNet [95] are two object-centric datasets that provide rich 3D Computer-Aided Design (CAD) models for shape analysis, followed by similar datasets [50, 97], which are usually low-quality, untextured, and have a domain gap with real-world objects. Another line of works [20, 21, 67, 78, 88, 103] advocate real 3D objects but are still limited in scale. We close this gap by shooting multi-view images of real-world objects, which capture the 3D awareness while allowing a scalable collection.

Multi-view image datasets. Multi-view image data is recently regarded as the source of 3D reconstruction or novel view synthesis. Early works collect multi-view images of real objects but only provide 3D models that are approximated [98] or for only a few instances [20]. Henzler *et al.* [44] contribute a larger video dataset to benchmark the task of 3D reconstruction. Another concurrent dataset, Objectron [3], annotates 3D bounding boxes and ground planes for all objects but lacks camera poses or point clouds for all 15k videos. GSO [27] gets clean 3D models with textures via scanning common household items but includes only a limited number of samples. Some works also construct synthetic multi-view datasets [11, 33, 85, 100, 107]. Neverthe-

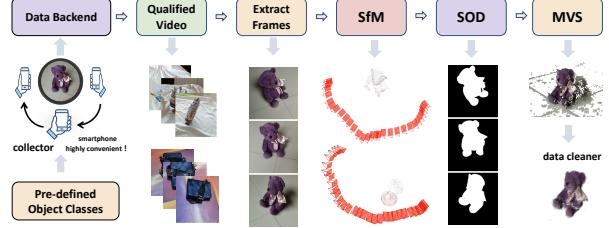


Figure 2. The efficient data acquisition pipeline of MVImgNet.

less, these datasets have a small scale and category range for special tasks, which limits the robust and generic learning of 3D deep models. We construct a large-scale multi-view image dataset – MVImgNet, which contains 219k videos for real-world objects in a wide range of 238 categories.

Special discussions on CO3D [73]. Very recently, CO3D extends [44] to 19k videos covering 50 object categories. Although we share a similar idea in the light of object-centric multi-view data collections, our MVImgNet enjoys conspicuously larger scales.

More importantly, we have fundamentally different *motivations* and *insights*. The primary motivation of CO3D is *exactly pre-set*, which is to transfer the 3D reconstruction training/evaluation from synthetic datasets into the real-world setup. In contrast, MVImgNet examines existing 3D datasets by re-walking the past development journey in the 2D domain. Concretely, through reflecting on how ImageNet [24] demonstrates its generic impact on data-hungry algorithms, we attempt to build a 3D counterpart of ImageNet and choose multi-view images as a soft bridge between 2D and 3D. Instead of nailing down a special target, we conduct a variety of *pilot* studies on how MVImgNet can benefit miscellaneous visual tasks, aiming to inspire and retain lots of possibilities for the broader vision community.

In the later experiments, our datasets indeed show greater power than CO3D on different visual challenges.

3. The Basis – MVImgNet Dataset

As shown in Fig. 2, the whole data acquisition pipeline of MVImgNet is highly efficient, which is illustrated below.

3.1. Raw Data Preparation

Building a large-scale dataset is always challenging, due to not only the laborious data collection but also the non-trivial annotation, which is especially critical for real-world 3D data [5, 10, 23]. Thanks to the rapid development of mobile devices, shooting a video around an object in the wild becomes highly convenient and accessible in our daily life, which makes multi-view images be easily gained by crowd-sourcing.

Composition setup. We set up some constraints on the ratio. Depending on the category, each class is initially set with the different expected number of video cap-

	Annotation	Collected	Valid	GPU hours
Sparse	219,188	215,755	3,806.8	
Segmentation	104,261	✓	2,316.9	
Dense	98,899	80,000	25,122.4	

Table 1. **Data statistics**, including **collected** amount, **valid** amount after cleaning, and the **GPU hours** for processing.

Dataset	Real	# of objects	# of categories	Multi-view	3D-GT
ShapeNet [11]	X	51k	55	render	CAD model
ModelNet [95]	X	12k	40	render	CAD model
Choi <i>et al.</i> [20]	✓	2k	9	360° captured	RGB-D scan
Objectron [3]	✓	15k	9	limited	3D bbox, pcl
GoogleScan [27]	✓	2k	NA	360° captured	RGB-D scan
Henzler <i>et al.</i> [44]	✓	2k	7	360° captured	pcl
ScanObjectNN [87]	✓	14k	15	limited	pcl
CO3D [73]	✓	19k	50	360° captured	-
CO3D-pcl [73]	✓	5k	50	360° captured	pcl
MVImgNet (ours)	✓	220k	238	180° captured	-
MVPNet (ours)	✓	80k	150	180° captured	pcl

Table 2. **Comparison** between our datasets and related ones. “pcl” denotes point clouds, “bbox” means bounding boxes.

tures regarding their *generality*, *e.g.*, the number of captures for “bottles”, “bags” and “snacks” is planned to be around 2000, while the number of “chips”, “apples” and “guitars” is set to about 1000. This setting is closer to real life.

Video capture. How to capture videos directly affect the quality of our data, so we draw up the following requirements as guidance for the captured videos: 1) The length of each video should be about 10 seconds. 2) The frames in the video should not be blurred. 3) Each video should capture 180° or 360° view of the object as much as possible. 4) The proportion of the object in the video should be above 15%. 5) Each video can only contain one class of principal object. 6) The captured object should be stereoscopic.

Crowdsourcing. We employ around 1000 normal collectors from different professions (*e.g.*, teacher, doctor, student, cook, babysitter) and ages (20~50). Each of them is asked to take several videos in their daily life, (*i.e.*, denoting diverse real-world environments) and upload them to the backend. Meanwhile, about 200 new well-trained expert data cleaners are responsible to review each submission and assure it fulfills the aforementioned capture requirements, when they may report some feedback or directly delete the unqualified submissions. The whole procedure ensures both the *diversity* and *quality* of the raw videos.

3.2. Data Processing

For each qualified video submission, we exploit an automatic process to obtain the common 2D and 3D annotations, including object masks, camera intrinsic and extrinsic, depth maps, and point clouds.

Sparse reconstruction. Following the procedure of [63, 64], the sparse reconstruction aims to reconstruct the camera intrinsic and extrinsic for each video, by applying the COLMAP Structure-from-Motion (SfM) algorithm [74] on a series of equal-time-interval chosen frames.



Figure 3. Some **frames** sampled from **MVImgNet**.

Foreground object segmentation. Each frame extracted from the original video is fed to the CarveKit [69] for generating the binary foreground object mask, not only benefiting the dense reconstruction but also contributing to the further step of salient object detection (Sec. 5.3).

Dense reconstruction. With the sparse model output from COLMAP SfM, we employ multi-view stereo [75] of COLMAP to generate the dense depth and normal maps for each frame. We extract the depths of an object using the binary foreground masks, which are then back-projected and fused according to the normal information, yielding a densely reconstructed point cloud for each video. Finally, the point cloud is manually cleaned by: 1) Delete the object with obvious noisy or extremely sparse reconstruction. 2) Remove all backgrounds. The final derived 3D point cloud dataset – MVPNet, is illustrated in Sec. 6.1.

3.3. Dataset Summary

Statistics. The statistics of MVImgNet are shown in Tab. 1, and Tab. 2 compares our datasets with other alternatives. MVImgNet includes 238 object classes, from 6.5 million frames of 219,188 videos. Fig. 3 shows some samples of MVImgNet. The annotation comprehensively covers object masks, camera parameters, and point clouds.

Category. We leverage WordNet [65] taxonomy that is used by ImageNet [24] to describe multi-hierarchy categories of objects and define 238 common classes. Unlike ImageNet which contains various plants and animals (nature-centric), the objects in our MVImgNet are *found* or *used* in human daily life (human-centric), where 65 classes overlap with ImageNet. The detailed category taxonomy, per-category data distributions, and more sample visualizations of MVImgNet are illustrated in the supplementary material.

4. 3D Reconstruction

4.1. Radiance Field Reconstruction

Pre-review. Recently, a series of generalizable Neural Radiance Fields (NeRF) variants [13, 86, 92, 110] have been proposed to reconstruct radiance field on-the-fly from one/few-shot source views for novel view synthesis.

Method	Diffuse Synthetic 360° [79]			Realistic Synthetic 360° [63]			Real-world 360° Objects [73]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Train from scratch	37.17	0.990	0.017	25.49	0.916	0.100	22.18	0.714	0.365
MVImgNet-pretrained	37.66	0.990	0.014	27.26	0.930	0.071	24.67	0.740	0.310

Table 3. NeRF quantitative results on three different levels of test sets.

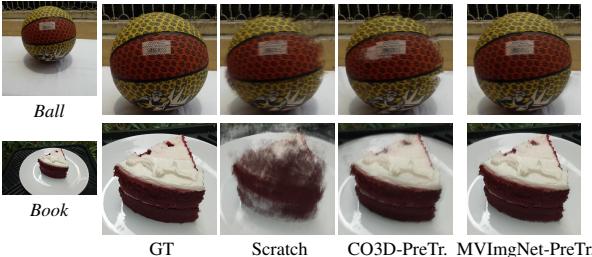


Figure 4. NeRF qualitative comparison of **train-from-scratch** IBRNet, CO3D [73]-pretrained IBRNet and **MVImgNet-pretrained** IBRNet [92].

What can MVImgNet do? Training NeRFs that can generalize to unseen objects requires learning 3D priors from a huge amount of multi-view images. Existing state-of-the-arts either resort to learning on large-scale synthetic data [13, 86, 110], or adopt a mixed use of synthetic data and a small self-collected real dataset [92]. However, the synthetic data introduces a big domain gap with real-world objects, only a few real scenes cannot remedy this defect. We argue that: *our MVImgNet perfectly fits the huge data demand of learning-based generalizable NeRF methods.*

To verify this, we choose IBRNet [92] as the baseline and conduct an empirical study. We pretrain IBRNet on the full MVImgNet dataset then finetune on the training datasets used in IBRNet [92] for a few iterations, and compare with the original train-from-scratch IBRNet model in terms of generalization capability (*i.e.*, generalizing to unseen scenes with only few-shot inputs). For fairness, we evaluate all methods under the same protocol of IBRNet.

Here comes a challenge about *how to evaluate* the generalization ability of NeRFs, since there is no official benchmark for this. To this end, we employ three different *third-party* object-centric datasets to form the test set. 1) The diffuse synthetic 360° object dataset [79] which contains 4 Lambertian objects. 2) The realistic synthetic 360° object dataset [63] which consists of 8 realistically fabricated objects. 3) The real-world 360° object dataset [73] which includes 88 real-world scenes from different lighting conditions. To summarize, the whole test set comprises 100 objects from 56 distinct categories, ranging from the synthetic domain to the real-world domain, which is considered to be an impartial evaluation set for generalization ability.

The quantitative and qualitative results are respectively shown in Tab. 3 and Fig. 4. Evidently, pretraining on MVImgNet improves the generalization ability of the model by a large margin. Moreover, we perform the same pretraining on CO3D [73] and MVImgNet-small (a random subset

Where to pretrain	Real-world 360° Objects [73]		
	PSNR↑	SSIM↑	LPIPS↓
CO3D [73]	24.01	0.732	0.339
MVImgNet-small	24.08	0.736	0.316
MVImgNet	24.67	0.740	0.310

Table 4. NeRF quantitative comparison with CO3D [73].

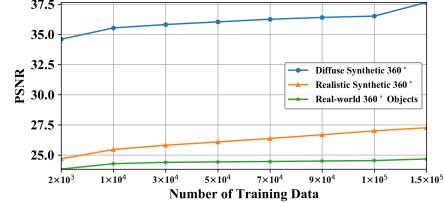


Figure 5. PSNR of IBRNet pretrained by a different number of MVImgNet data (higher is better).

of MVImgNet, which owns the same scale as CO3D). As shown in Tab. 4 and Fig. 4, MVImgNet shows greater power than CO3D [73]. The implementation details can be found in the supplementary material.

More data, more power. Fig. 5 shows that an apparent rising trend of generalization metrics can be observed with the increase of training data.

4.2. Multi-view Stereo

Pre-review. Multi-view Stereo (MVS) [31] is a classical task in 3D computer vision, with the goal of reconstructing 3D geometry from multi-view images. Conventional methods [32, 45, 76, 113] reconstruct 3D geometry by finding the patches matched with different images and estimating the depth according to the extrinsic camera. In recent years, deep learning methods are introduced into MVS to solve the issues such as weak textures, and non-laplacian spheres. Two popular end-to-end methods MVSNet [108] and R-MVSNet [109] propose to encode the multi-view images and build a cost volume for predicting depth maps. Under the supervision of large-scale RGB-D datasets [2], they outperform conventional patch-matched-based approaches. However, these methods require massive amounts of RGB-D data, which is always difficult to acquire. This drives the emergence of self-supervised MVS methods [12, 25, 101, 106].

What can MVImgNet do? We demonstrate that our MVImgNet is capable of benefiting the *data-efficient* MVS with limited training examples, which is practically meaningful considering MVS always requires the burdensome collection of RGB-D data.

We pretrain a self-supervised MVS method, JDACS [101], on MVImgNet. Our implementation follows the original settings of JDACS. Then, we select DTU [2] dataset with 79 training samples and 22 test samples. We perform the *data-efficient* evaluation with limited training data, where the MVImgNet-pretrained MVSNet is finetuned on 5%, 15%, and 25% of DTU training samples, and eval-

Ratio	2mm ↑	4mm ↑	8mm ↑
5%	48.61 / 50.10	65.18 / 67.97	75.69 / 78.58
15%	54.74 / 54.68	70.96 / 72.04	80.32 / 81.73
25%	57.29 / 58.63	74.41 / 75.20	83.68 / 84.28

Table 5. **MVS depth map accuracy** on DTU evaluation set under different ratios of DTU training samples, in terms of **training from scratch / MVIImgNet-pretrained** (higher is better).

Where to pretrain	2mm ↑	4mm ↑	8mm ↑
CO3D	46.18	66.72	78.39
MVIImgNet-small	47.73	66.53	78.15
MVIImgNet	50.10	67.97	78.58

Table 6. **MVS depth map accuracy** comparison with CO3D [73] on DTU evaluation set. Pretrain on MVIImgNet, MVIImgNet-small or CO3D, and finetune using 5% of DTU training samples.

uated on the DTU test set. Tab. 5 reports the accuracy given different thresholds of the error between the predicted and ground truth depth map, which indicates that the MVIImgNet-pretrained model is capable of improving the model trained from scratch by a large margin under the data-efficient setup. Furthermore, we perform pretraining using CO3D [73] and MVIImgNet-small. Tab. 6 indicates that our MVIImgNet is stronger than CO3D, and MVIImgNet-small shows comparable power as CO3D. For more implementation details, please see the supplementary material.

We advocate benchmarking NeRF and MVS methods with the help of pretraining on MVIImgNet.

5. View-consistent Image Understanding

5.1. View-consistent Image Classification

Pre-review. As indicated by Dong *et al.* [26], although humans can easily recognize one object from different views, deep learning models can hardly do that robustly. MVIImgNet provides numerous images from different viewpoints, so we hope to enhance the model’s view consistency with the help of MVIImgNet, which is significantly important for realizing human-like intelligence.

What can MVIImgNet do? One naive approach is to finetune the ImageNet-pretrained ResNet-50 [43] on our MVIImgNet. However, such an approach will be problematic due to the categories of two datasets are very distinct, which may cause catastrophic forgetting issues [49].

For this reason, we create a *new training set*, namely **MVI-Mix**, by mixing the original ImageNet data with MVIImgNet. Specifically, we randomly sample 5 *consecutive* frames of each video in MVIImgNet, and mix the multi-view images with original ImageNet data. We also build another two artificial training sets, namely MVI-Gap and MVI-Aug, that only differ MVI-Mix with image views. **MVI-Gap** samples the 5 *non-consecutive* frames that have much larger view differences. **MVI-Aug** samples 1 frame from each video of MVI-Mix and applies 4 *different data augmentations* (*i.e.*, random color jittering, grid mask, ro-

Dataset	Confidence Var	Accuracy
ImageNet-only	0.207	53.09%
MVI-Aug	0.105	71.48%
MVI-Gap	0.103	77.23%
MVI-Mix	0.102	77.31%

Table 7. **View-consistent image classification results** on MVIImgNet test set using **fully-supervised ResNet-50** [43]. **Adding MVIImgNet for training improves the view consistency** (smaller Var and higher Acc indicate better view consistency).

	Train: CO3DI-Mix	Train: MVI-Mix
Test: CO3D	Var: 0.104 Acc: 91.25%	Var: 0.155 (+0.051) Acc: 84.83% (-6.42%)
Test: MVIImgNet	Var: 0.188 (+0.086) Acc: 57.93% (-19.38%)	Var: 0.102 Test: 77.31%

Table 8. **View-consistent image classification** comparison with CO3D [73]. The *performance drops* when train on CO3DI-Mix and test on MVIImgNet (**blue number**) are much larger than the opposite (**red number**), which means that MVI-Mix pretrained model is more robust to multiview consistency than CO3DI-Mix pretrained model.

tation and erase, respectively) to it, which aims to differentiate the multi-view augmentation from the normal data augmentations. We compare the *variance* of the softmax confidence and the accuracy on MVIImgNet *test set* for examining the view consistency. As illustrated in Tab. 7, adding our MVIImgNet data for training can effectively improve the model’s view consistency and accuracy. The test dataset used for all experiments is MVIImgNet test dataset. In addition to using the convolution-based network (*i.e.*, ResNet-50), we also apply the *Transformer*-based architecture, DeiT-Tiny [84]. The results are 48.76% accuracy and 0.225 Var on ImageNet-only, VS 0.122 Var (**0.103↓**) and 73.88% (**25.12↑**) accuracy on MVI-Mix. This leads to a unanimous conclusion.

We further construct CO3DI-Mix (corresponding to MVI-Mix) as the mixer of CO3D [73] & ImageNet under the same ratio of MVIImgNet & ImageNet in MVI-Mix. As illustrated in Tab. 8, MVIImgNet brings more benefits than CO3D for view-consistent image recognition.

5.2. View-consistent Contrastive Learning

Pre-review. Contrastive Learning (CL) is one mainstream of self-supervised training techniques [16–19, 36, 39, 83]. One key factor of CL is the construction of positive/negative pairs, *e.g.*, MoCo-v2 [17] treats an image with different augmentations, *a.k.a.* **views**, as the positive pair.

What can MVIImgNet do? One natural question is: *can the viewpoints of MVIImgNet serve as the positive pairs for CL?* To answer this question, we finetune the off-the-shelf ImageNet-pretrained MoCo-v2 on MVIImgNet. For each iteration, we randomly sample two frames from the same video as the positive pair and apply the original data augmentations used in MoCo-v2 [17]. Meanwhile, the frames from other videos will be treated as their negatives.

Training scheme	Confidence Var	Accuracy
w/o finetune	0.098	70.26%
w/ finetune	0.086	71.22%

Table 9. **View-consistency classification results** under the **self-supervised contrastive learning** regime on MVImgNet test set. Finetuning MoCo-v2 [17] on MVImgNet improves the view consistency.



Figure 6. Qualitative comparison between **MVImgNet-finetuned** U2Net [72] and **original** U2Net for salient object detection (SOD). Left: **finetuning on MVImgNet** improves the performance on a *hard* view. Right: **finetuning on MVImgNet** improves the performance on two consecutive hard views.

The MVImgNet-finetuned model is finally evaluated on the MVImgNet test dataset for examining the view consistency, where we compare the variance and mean of the softmax confidence and the accuracy. As Tab. 9 shows, finetuning on MVImgNet with CL can also improve both the model’s view consistency and accuracy. More implementation details of view-consistent image classification is described in the supplementary material.

In the future, it is highly recommended to regard view-consistency as a criterion for evaluating the image recognition task, and utilize MVImgNet to pretrain the models.

5.3. View-consistent SOD

Pre-review. Salient Object Detection (SOD) aims to segment the most visually prominent objects in an image. Although remarkable progress has been made recently, it remains lots of challenges.

What can MVImgNet do? We test a state-of-the-art SOD model U2Net [72] on our MVImgNet. As Fig. 6 shows, U2Net failed to segment “hard” views, even though some views can be segmented with few flaws. Despite such a disappointment, the inconsistent predictions of different views caught our attention: *can we improve the SOD models with multi-view consistency?*

We propose to leverage the multi-view consistency to improve SOD with the help of optical flows. Specifically, given two consecutive frames, we first calculate their optical flow and warp the flow to the segmentation mask of one of the frames. Then, a consistency loss can be calculated between the warped mask and the mask of another frame. For ease of implementation, we directly finetune U2Net on our MVImgNet. To prevent the catastrophic forgetting issue [49], in addition to the MVImgNet, we also use the original training data DUTS-TR [91]. For the “hard” views



Figure 7. Some 3D point clouds sampled from **MVPNet**.

(IoU ≤ 0.7) on our MVImgNet test set, the model finetuned on our MVImgNet can bring a 4.1% IoU improvement (see Fig. 6 for qualitative comparison). The implementation details of view-consistent SOD can be found in the supplementary material.

6. MVPNet for 3D Understanding

6.1. MVPNet Dataset

Derived from the dense reconstruction on MVImgNet (as mentioned in Sec. 3.2), a new large-scale real-world 3D object point cloud dataset – MVPNet, is born, which contains 87,200 point clouds with 150 categories. Fig. 7 shows some examples of MVPNet. As listed in Tab. 2, compared with existing 3D object datasets, our MVPNet contains a conspicuously richer amount of real-world object point clouds, with abundant categories covering many common objects in the real life. The detailed category taxonomy per-category data distributions of MVPNet are illustrated in the supplementary material.

6.2. 3D Point Cloud Classification

In this work, we focus on point cloud classification. We believe that future works may also utilize our dataset to help much more 3D understanding tasks such as indoor-scene parsing, outdoor-environment perception, pose estimation, and robotics manipulation.

Pre-review. We advocate paying more attention to *real-world* setup, which is more feasible for real applications. ScanObjectNN [88] has been manifested as the most challenging point cloud classification benchmark so far, so we choose it for the major comparison with our MVPNet.

What can MVPNet do? We show that pretraining on MVPNet is able to aid the performance of real-world point cloud classification. We pretrain several models [37, 61, 70, 71, 93, 96, 104, 105] on MVPNet, and finetune them on ScanObjectNN.

Two settings are considered for evaluation. First is PB_T50_RS in ScanObjectNN with small perturbation, translation and rotation on point cloud. Another is adding heavy rotation on PB_T50_RS, to create a more challenging setting. The results are shown in Tab. 10, where pretraining on MVPNet is able to increase the classification accuracy under most circumstances.

Method	Add Random Rotation		PB_T50.RS	
	from scratch	pretrained	from scratch	pretrained
PointNet [70]	60.57 / 55.20	64.25 / 59.29	70.63 / 67.28	67.73 / 64.12
PointNet++ [71]	76.50 / 73.42	78.76 / 76.54	78.80 / 75.70	80.22 / 76.91
DGCNN [93]	80.50 / 78.45	80.42 / 78.20	79.44 / 76.24	82.36 / 80.08
PointMLP [61]	83.69 / 82.54	84.87 / 83.71	85.64 / 84.14	85.98 / 84.38
CurveNet [96]	73.96 / 69.96	78.99 / 76.59	74.27 / 69.43	83.68 / 81.17
GDANet [105]	80.33 / 79.14	83.59 / 82.29	79.01 / 75.91	83.90 / 82.51
PACConv [104]	70.91 / 65.70	76.21 / 72.47	72.88 / 68.60z	76.91 / 73.45
PCT [37]	81.32 / 79.17	82.08 / 80.41	77.46 / 73.64	84.20 / 81.94
PointMAE [66]	83.17 / 80.75	86.19 / 84.60	77.34 / 73.52	84.13 / 81.92

Table 10. **ScanObjectNN real-world point cloud classification results.** The comparison is between the **train-from-scratch** model and **MVPNet pretrained** model. The metric is **overall / average accuracy**.

Method	PB_T50.RS		
	PreTr. on MVPNet	PreTr. on MVPNet-small	PreTr. on CO3D
PointNet++	78.90 / 77.11	79.00 / 76.78	78.79 / 77.20
CurveNet	83.68 / 81.17	73.92 / 68.92	73.78 / 69.62
PointMAE	84.13 / 81.92	81.85 / 79.25	81.40 / 78.87

Table 11. **ScanObjectNN real-world point cloud classification** comparison with CO3D [73]. Pretrain on MVImgNet, MVImgNet-small or CO3D, and finetune on ScanObjectNN. The metric is **overall / average accuracy**.

6.3. Self-supervised Point Cloud Pretraining

Pre-review. Self-supervised learning has been exploited for 3D object point cloud understanding [58, 66, 90, 112]. Nevertheless, they are all pretrained on synthetic datasets [11, 95], making it hard to obtain rich real-world representations. To solve this, our MVPNet becomes a natural choice.

What can MVPNet do? In Tab. 10, PointMAE pretrained on our MVPNet outperforms the state-of-the-art methods when finetuned on ScanObjectNN, proving the benefit of MVPNet on the self-supervised learning regime for real-world point cloud classification. Tab. 11 also shows that pretraining on MVPNet is more powerful than CO3D [73] for real-world point cloud classification.

6.4. MVPNet Benchmark Challenge

We present the MVPNet benchmark challenge for real-world point cloud classification, which contains 64,000 training and 16,000 testing samples. The results of various methods are shown in Tab. 12.

MVPNet is more challenging than ScanObjectNN. We firstly train models on ScanObjectNN, then conduct the test on MVPNet, which is concluded in Tab. 13. On the opposite, we also train on MVPNet and test on ScanObjectNN, which is listed in Tab. 14. Comparing Tab. 13 and Tab. 14, the *accuracy drops* are significantly *larger* when training on ScanObjectNN and testing on MVPNet, which verifies the greater challenge of our MVPNet. All the experiments in 3D understanding strictly follow the original settings of the selected backbone networks.

MVPNet is suggested for investigating 3D point cloud understanding in the future.

Method	Overall / Average Accuracy
PointNet [70]	70.72 / 54.46
PointNet++ [71]	79.15 / 58.24
DGCNN [93]	86.49 / 63.98
PointMLP [61]	88.89 / 73.64
CurveNet [96]	88.88 / 75.37
GDANet [105]	89.54 / 68.41
PACConv [104]	83.35 / 59.13
PCT [37]	91.49 / 75.41

Table 12. Quantitative results on our **new MVPNet benchmark for real-world point cloud classification.**

Method	train: MVPNet	train: ScanObjectNN
	test: MVPNet	test: MVPNet
PointNet++ [71]	79.15 / 58.24	15.13 (-64.06) / 7.96 (-50.28)
CurveNet [96]	88.88 / 75.37	46.95 (-41.93) / 28.24 (-47.13)

Table 13. Quantitative comparison while **training on ScanObjectNN, testing on MVPNet**. The evaluation metric is **overall / average accuracy**.

Method	train: ScanObjectNN	train: MVPNet
	test: ScanObjectNN	test: ScanObjectNN
PointNet++ [71]	76.50 / 73.42	40.35 (-36.15) / 33.59 (-39.83)
CurveNet [96]	73.96 / 69.96	51.84 (-22.12) / 46.27 (-23.69)

Table 14. Quantitative comparison while **training on MVPNet, testing on ScanObjectNN**. The evaluation metric is **overall / average accuracy**.

7. Conclusion

We have introduced MVImgNet, a large-scale dataset of multi-view images, which is efficiently collected by shooting videos of real-world objects. The multi-view nature endows our dataset with 3D-aware visual signals, making MVImgNet a soft bridge to link 2D and 3D vision. To probe the power of MVImgNet, we conduct a host of pilot experiments on various visual tasks, including radiance field reconstruction, multi-view stereo, and view-consistent image understanding, where MVImgNet demonstrates promising effectiveness, expecting more future explorations. As a bonus of MVImgNet, a point cloud dataset – MVPNet is derived. Experiments show MVPNet can benefit real-world 3D object classification. As a broader impact on society, our datasets *delineates* a world – that is closer to a colorful and vivid real 3D world – where we human lives.

Limitations. We mainly focus on human-centric classes, and our data can definitely satisfy the understanding of common objects in human daily life. As a result, the category richness of MVImgNet is less than ImageNet [24], which may cause inferior performance when serving for recognition on nature-centric classes such as turtle, bear, etc.

Moreover, our data do not consider very complex backgrounds, so they can not be *straightforwardly* adopted for complicated scene-level understanding. Yet we believe that such an issue can be solved by future methods via utilizing some mediate methods such as domain adaptation or knowledge distillation, with the help of *commonsense* knowledge and *universal* representations gained from our data.

Contributions

Xianggang Yu contributes to the whole building pipeline of the dataset, including data acquisition and data processing. He conducted experiments on radiance field reconstruction. He also designed and advised the experiment on view-consistent SOD.

Mutian Xu advised the exploration of view-consistent image understanding and 3D understanding. He also organized and wrote the whole paper, and led most part of the research.

Yidan Zhang worked on data processing. He implemented experiments on view-consistent image classification and view-consistent SOD. He also collected and provided illustrations of dataset statistics and dataset details.

Haolin Liu was responsible for point cloud labeling and cleaning. He implemented experiments on 3D point cloud classification.

Chongjie Ye conducted data labelling. He implemented experiments on multi-view stereo and 3D point cloud classification.

Yushuang Wu worked on the data collection and annotation. He also collected and wrote related works.

Zizheng Yan suggested the experiments on view-consistent image understanding.

Chenming Zhu conducted the experiment on view-consistent contrastive learning.

Zhangyang Xiong worked on data collection and processing.

Tianyou Liang worked on data cleaning and assisted experiments on view-consistent image understanding.

Guanying Chen, Shuguang Cui advised the project.

Xiaoguang Han proposed and led the whole research.

References

- [1] *Gradient-based learning applied to document recognition*, 1998. 2
- [2] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *IJCV*, 2016. 5, 14, 15
- [3] 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 *CVPR*, 2021. 3, 4
- [4] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 3
- [5] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 2, 3
- [6] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *CVPR*, 2019. 3
- [7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 2008. 3
- [8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 3
- [9] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *CVPR*, 2020. 2, 3
- [10] 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. In *3DV*, 2017. 2, 3
- [11] 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. 2, 3, 4, 8
- [12] Di Chang, Aljaž Božič, Tong Zhang, Qingsong Yan, Yingcong Chen, Sabine Süsstrunk, and Matthias Nießner. Rc-mvsnet: Unsupervised multi-view stereo with neural rendering. *arXiv preprint arXiv:2203.03949*, 2022. 5
- [13] 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 *ICCV*, 2021. 4, 5
- [14] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *The 49th annual meeting of the association for computational linguistics: human language technologies*, 2011. 3
- [15] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 2
- [16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 6
- [17] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 6, 7
- [18] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 6
- [19] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 6
- [20] Sungjoon Choi, Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. A large dataset of object scans. *arXiv preprint arXiv:1602.02481*, 2016. 3, 4
- [21] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *CVPR*, 2022. 3
- [22] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 3
- [23] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2, 3
- [24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 3, 4, 8, 14, 15
- [25] Yikang Ding, Qingtian Zhu, Xiangyue Liu, Wentao Yuan, Haotian Zhang, and Chi Zhang. Kd-mvs: Knowledge distillation based self-supervised learning for multi-view stereo. In *ECCV*, 2022. 5
- [26] Yinpeng Dong, Shouwei Ruan, Hang Su, Caixin Kang, Xingxing Wei, and Jun Zhu. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. *arXiv preprint arXiv:2210.03895*, 2022. 2, 6
- [27] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. *arXiv preprint arXiv:2204.11918*, 2022. 3, 4, 15
- [28] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 3
- [29] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *CVPR*, 2019. 15
- [30] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia,

- Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *ICCV*, 2021. 3
- [31] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2015. 5
- [32] David Gallup, Jan-Michael Frahm, Philippos Mordohai, and Marc Pollefeys. Variable baseline/resolution stereo. In *CVPR*, 2008. 5
- [33] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeanette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. Objectfolder 2.0: A multisensory object dataset for sim2real transfer. In *CVPR*, 2022. 3
- [34] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2, 3
- [35] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 2
- [36] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheorghiciuc Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *NeurIPS*, 2020. 6
- [37] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 2021. 7, 8
- [38] Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan Wegner, Konrad Schindler, and Marc Pollefeys. Semantic3d.net: A new large-scale point cloud classification benchmark. In *ISPRS*, 2017. 2, 3
- [39] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, June 2020. 6
- [40] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. *arXiv preprint arXiv:1811.08883*, 2018. 2
- [41] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2
- [42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *ICCV*, 2015. 2
- [43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 6, 14, 15
- [44] Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotny. Unsupervised learning of 3d object categories from videos in the wild. In *CVPR*, 2021. 3, 4
- [45] Xiaoyan Hu and Philippos Mordohai. Least commitment, viewpoint-based, multi-view stereo. In *International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, 2012. 5
- [46] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *3DV*, 2016. 3
- [47] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, 2013. 3
- [48] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3
- [49] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 2017. 6, 7
- [50] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. Abc: A big cad model dataset for geometric deep learning. In *CVPR*, 2019. 3
- [51] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 3
- [52] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 3
- [53] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2
- [54] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 2
- [55] Hildegarde Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 3
- [56] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021. 2
- [57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3
- [58] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *ECCV*, 2022. 8
- [59] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2
- [60] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2
- [61] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In *ICLR*, 2022. 7, 8
- [62] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 3

- [63] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 4, 5
- [64] 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 Trans. Graph.*, 2019. 4
- [65] George A. Miller. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. 4
- [66] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, 2022. 8, 15
- [67] Keunhong Park, Konstantinos Rematas, Ali Farhadi, and Steven M Seitz. Photoshape: photorealistic materials for large-scale shape collections. *ACM Trans. Graph.*, 2018. 3
- [68] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 3
- [69] Open Source Project. *CarveKit - image-background-remove-tool*. 4
- [70] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 7, 8
- [71] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 7, 8, 15
- [72] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 2020. 7, 14, 22
- [73] 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 *ICCV*, 2021. 3, 4, 5, 6, 8, 21
- [74] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 4
- [75] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 4
- [76] Shuhan Shen. Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes. *TIP*, 2013. 5
- [77] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 3
- [78] Arjun Singh, James Sha, Karthik S Narayan, Tudor Achim, and Pieter Abbeel. Bigbird: A large-scale 3d database of object instances. In *ICRA*, 2014. 3
- [79] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019. 5
- [80] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgbd scene understanding benchmark suite. In *CVPR*, 2015. 2, 3
- [81] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3
- [82] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 3
- [83] Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022. 6
- [84] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 6
- [85] Jonathan Tremblay, Moustafa Meshry, Alex Evans, Jan Kautz, Alexander Keller, Sameh Khamis, Charles Loop, Nathan Morrical, Koki Nagano, Towaki Takikawa, et al. Rtmv: A ray-traced multi-view synthetic dataset for novel view synthesis. *arXiv preprint arXiv:2205.07058*, 2022. 3
- [86] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *ICCV*, 2021. 4, 5
- [87] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, 2019. 2, 4, 15
- [88] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, 2019. 3, 7
- [89] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 3
- [90] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matthew J. Kusner. Unsupervised point cloud pre-training via occlusion completion. In *ICCV*, 2021. 8
- [91] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 7, 16
- [92] 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 *CVPR*, 2021. 4, 5, 15, 21
- [93] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph.*, 2019. 7, 8

- [94] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 3
- [95] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015. 2, 3, 4, 8
- [96] Tiange Xiang, Chaoyi Zhang, Yang Song, Jianhui Yu, and Weidong Cai. Walk in the cloud: Learning curves for point clouds shape analysis. In *ICCV*, 2021. 7, 8, 15
- [97] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. Objectnet3d: A large scale database for 3d object recognition. In *ECCV*, 2016. 3
- [98] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 3
- [99] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, 2013. 3
- [100] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. *IJCV*, 2020. 3
- [101] Hongbin Xu, Zhipeng Zhou, Yu Qiao, Wenxiong Kang, and Qiuxia Wu. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. In *AAAI*, 2021. 5, 15
- [102] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 3
- [103] Mutian Xu, Pei Chen, Haolin Liu, and Xiaoguang Han. To-scene: A large-scale dataset for understanding 3d tabletop scenes. In *ECCV*, 2022. 2, 3
- [104] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *CVPR*, 2021. 7, 8
- [105] Mutian Xu, Junhao Zhang, Zhipeng Zhou, Mingye Xu, Xiaojuan Qi, and Yu Qiao. Learning geometry-disentangled representation for complementary understanding of 3d object point cloud. In *AAAI*, 2021. 7, 8
- [106] Jiayu Yang, Jose M Alvarez, and Miaomiao Liu. Self-supervised learning of depth inference for multi-view stereo. In *CVPR*, 2021. 5
- [107] Zhenpei Yang, Zaiwei Zhang, and Qixing Huang. Hm3d-abo: A photo-realistic dataset for object-centric multi-view 3d reconstruction. *arXiv preprint arXiv:2206.12356*, 2022. 3
- [108] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 5
- [109] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*, 2019. 5
- [110] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 4, 5
- [111] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 3
- [112] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, 2022. 8
- [113] Enliang Zheng, Enrique Dunn, Vladimir Jojic, and Jan-Michael Frahm. Patchmatch based joint view selection and depthmap estimation. In *CVPR*, 2014. 5
- [114] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *SIGGRAPH*, 2018. 15

Supplement for MVIImgNet

A Per-category Data Distribution	14
B More Visualizations of Data Samples	14
C More Visualizations of Qualitative Results	14
D More Experiments of Data Scalability	14
E More Discussions about Our Datasets	15
F. Implementation Details	15
F.1. 3D Reconstruction	15
F.2. View-consistent Image Understanding	15
F.3. 3D Understanding	16

A. Per-category Data Distribution

The category taxonomy is shown in Fig. I for MVIImgNet, and Fig. II for MVPNet. The per-category data distribution is illustrated in Fig. III for MVIImgNet, and Fig. IV for MVPNet. The average size is 921 per class for MVIImgNet and 581 per class for MVPNet.

B. More Visualizations of Data Samples

MVIImgNet. Fig. V presents a larger set of examples in MVIImgNet. Several multi-view images and the corresponding class label are illustrated for each sample. It clearly shows the differences between each view, and comprehensive categories in our dataset.

MVPNet. Fig. VI shows various 3D point clouds from MVPNet. It can be seen that each sample has a distinct texture, noise, and pose, indicating real-world signals.

C. More Visualizations of Qualitative Results

Radiance field reconstruction. We visualize more results of generalizable NeRF reconstruction in Fig. VII, where the MVIImgNet-pretrained model performs consistently much better than the train-from-scratch model.

View-consistent SOD. Fig. VIII illustrates more results of the view-consistent salient objection detection (SOD) task on our MVIImgNet test set, where finetuning U2Net [72] on MVIImgNet gains better result than the original U2Net.

D. More Experiments of Data Scalability

As indicated in the main paper, more power can be gained with more data utilized from our datasets. In this section, we provide more experimental results following such rules.

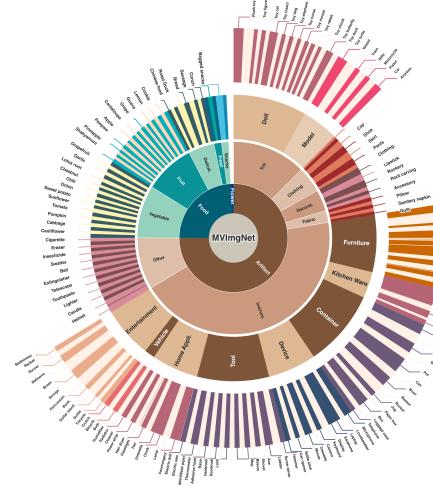


Figure I. **Category taxonomy of MVIImgNet**, where the angle of each class denotes its actual data proportion. **Interior**: Parent class. **Exterior**: Children class.

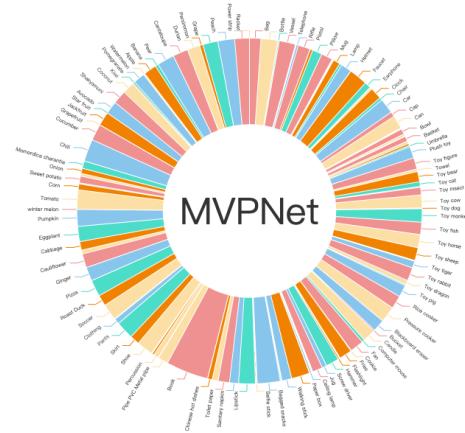


Figure II. **Category distribution of MVPNet**.

Multi-view stereo. Tab. I lists the MVS depth map accuracy on DTU [2] evaluation set. It shows that using larger amounts of videos from MVIImgNet for pretraining yields higher accuracy.

View-consistent image classification. Similar conclusions are also found in the view-consistent image classification task. We progressively add more MVIImgNet training data into MVI-Mix data (mixing the original ImageNet [24] data with MVIImgNet data as stated in the main paper) to train ResNet-50 [43] and evaluate on MVIImgNet test set. Tab. II demonstrates that adding more MVIImgNet training data brings better view consistency for the image recognition task.

Real-world point cloud classification. Besides, as shown in Tab. III and Tab. IV, when employing larger ratios of

Method	$2mm \uparrow$	$4mm \uparrow$	$8mm \uparrow$
pretrained with 10k videos	52.96	72.25	83.79
pretrained with 50k videos	56.86	73.79	83.42
pretrained with 100k videos	58.63	75.20	84.28

Table I. **MVS depth map accuracy** on DTU [2] evaluation set, using **different amounts (10k, 50k, 100k) of videos** (one video may contain several multi-view images / frames) from MVImgNet for pretraining.

Scale	Confidence Var	Accuracy
ImageNet-only	0.207	53.09%
MVI-Mix with 20k videos	0.119	75.03%
MVI-Mix with 40k videos	0.114	76.88%
MVI-Mix with 80k videos	0.104	77.03%
MVI-Mix with 100k videos	0.102	77.31%
MVI-Mix with 120k videos	0.101	77.47%

Table II. **View-consistency image classification results** on MVImgNet test set, using **different amounts (20k, 40k, 80k, 100k, 120k) of videos** (one video may contain several multi-view images / frames) from MVImgNet for training ResNet-50 [43] (smaller Confidence Var and higher Accuracy indicate better view consistency).

data from MVPNet for pretraining both supervised (*i.e.*, PointNet++ [71], CurveNet [96]) and self-supervised models (*i.e.*, PointMAE [66]), the better performance can be achieved when fine-tuning them on ScanObjectNN dataset [87] for real-world point cloud classification task.

E. More Discussions about Our Datasets

Data filter. Our $\sim 219k$ videos are screened from $\sim 260k$ raw videos, where the videos with bad camera estimations are filtered. When building MVPNet, we select 90k (the most common 150 categories are chosen) videos, yielding 87k point clouds to remain after the manual cleaning.

Real-world captures. Note that when we capture the object videos, we maintain the *original* status of objects in *real-world* environments, *i.e.*, objects will *not be intentionally* displayed standalone for ideal 360° captures (*e.g.*, the sofa is against the wall). By doing so: **1)** The capture is easy to conduct, making it possible to build a very large-scale dataset. **2)** The produced data better matches the *real-world applications*, *e.g.*, our obtained point clouds are usually of partial views which are more like real-captured. **3)** The produced images usually contain the diverse scene-level *background*, instead of the 360° capture of single objects on a *clean* supporter. This better provides the potential for *in-the-wild* scene-level visual tasks.

Method	from scratch	Add Random Rotation		
		25%	50%	100%
PointNet++ [71]	76.50 / 73.42	77.82 / 75.98	78.11 / 76.13	78.76 / 76.54
CurveNet [96]	73.96 / 69.96	73.75 / 69.86	75.83 / 72.48	78.99 / 76.59
PointMAE [66]	83.17 / 80.75	83.83 / 81.94	85.22 / 83.34	86.19 / 84.60

Table III. ScanObjectNN [87] real-world point cloud classification results of using **different ratio (25%, 50%, 100%) of data from MVPNet for pretraining** under the setting of Add Random Rotation. The metric is **overall / average accuracy**.

Method	from scratch	PB_T50_RS		
		25%	50%	100%
PointNet++ [71]	78.80 / 75.70	79.67 / 76.63	81.36 / 79.33	80.22 / 76.91
CurveNet [96]	74.27 / 69.43	77.26 / 72.65	81.32 / 78.03	83.68 / 81.17
PointMAE [66]	77.34 / 73.52	82.75 / 79.90	84.18 / 81.41	84.13 / 81.92

Table IV. ScanObjectNN [87] real-world point cloud classification results of using **different ratio (25%, 50%, 100%) of data from MVPNet for pretraining** under the setting of PB_T50_RS. The metric is **overall / average accuracy**.)

F. Implementation Details

F.1. 3D Reconstruction

Radiance field reconstruction.

We choose IBR-Net [92] as the baseline method, and use the original training datasets of IBRNet [92], which include Google Scanned Objects [27], RealEstate10K [114], the Spaces dataset [29], and 102 real scenes from handheld cellphone captures. We pretrain IBRNet on the full MVImgNet dataset and finetune on the aforementioned IBRNet training datasets for 10k iterations. For each object, 8~12 views are used for training and 10 views for inference. #views is independent on #objects. The raw input resolution of each sample is used for computing, and it varies. The finetuning takes 10k iterations, and the scratch model is exactly the same as the author-released IBRNet model for a fair comparison. The pretraining takes about 3 days on 8 RTX3090 GPUs.

Multi-view stereo. Multi-view stereo (MVS) aims at recovering 3D scenes from multi-view images and calibrated cameras. As for the data preprocessing, 200K frames are randomly sampled from 100K videos in MVImgNet, and are resized to 640×360 or 360×640 . We choose JDACS [101] to perform self-supervised pretraining on MVImgNet. JDACS takes multi-view images and corresponding poses as input, and uses MVSNet as the backbone to output the synthetic/pseudo depth, where the self-supervision signal is provided by multi-view consistency.

F.2. View-consistent Image Understanding

View-consistent image classification.

As mentioned in the main paper, we mix MVImgNet and original ImageNet [24] for creating a new training set. The hybrid datasets contain 1, 100 categories (after removing

the overlapping classes), coming from 500k frames of 100k MVIImgNet videos and 200k ImageNet images.

View-consistent contrastive learning. We follow the original MoCo v2 to conduct experiments. For reducing view redundancy, we randomly sample 5 frames of each video from MVIImgNet for finetuning. For each iteration, we randomly sample two view images from the same video as positive pair and apply random data augmentation to increase the generalization capability of the model, images from other videos will be treated as negative pairs

View-consistent SOD. We propose to leverage the multi-view consistency to improve SOD with the help of *optical flows*. The two adjacent frames should be the same after warping the optical flow to one of the other frames, yielding the loss of the optical flow as:

$$Loss_{OF} = \mathcal{M}(f_i) - \mathcal{M}(f_{i-1}) \cdot \mathcal{F}(f_i), \quad (1)$$

where i denotes the frame index, \mathcal{M} means the mask, and \mathcal{F} is the optical flow between f_i and f_{i-1} calculated before training. By adding $Loss_{OF}$ into the original SOD loss, the final loss is:

$$Loss = \tau * Loss_{OF} + (1 - \tau) * Loss_{SOD}, \quad (2)$$

where τ is set to 0.15 in our experiments.

For fast training, we sample 10 frames uniformly from each video of 100k MVIImgNet and 10, 553 training images from DUTS-TR [91].

F.3. 3D Understanding

All the experiments in 3D understanding strictly follow the original settings of the selected backbone networks.

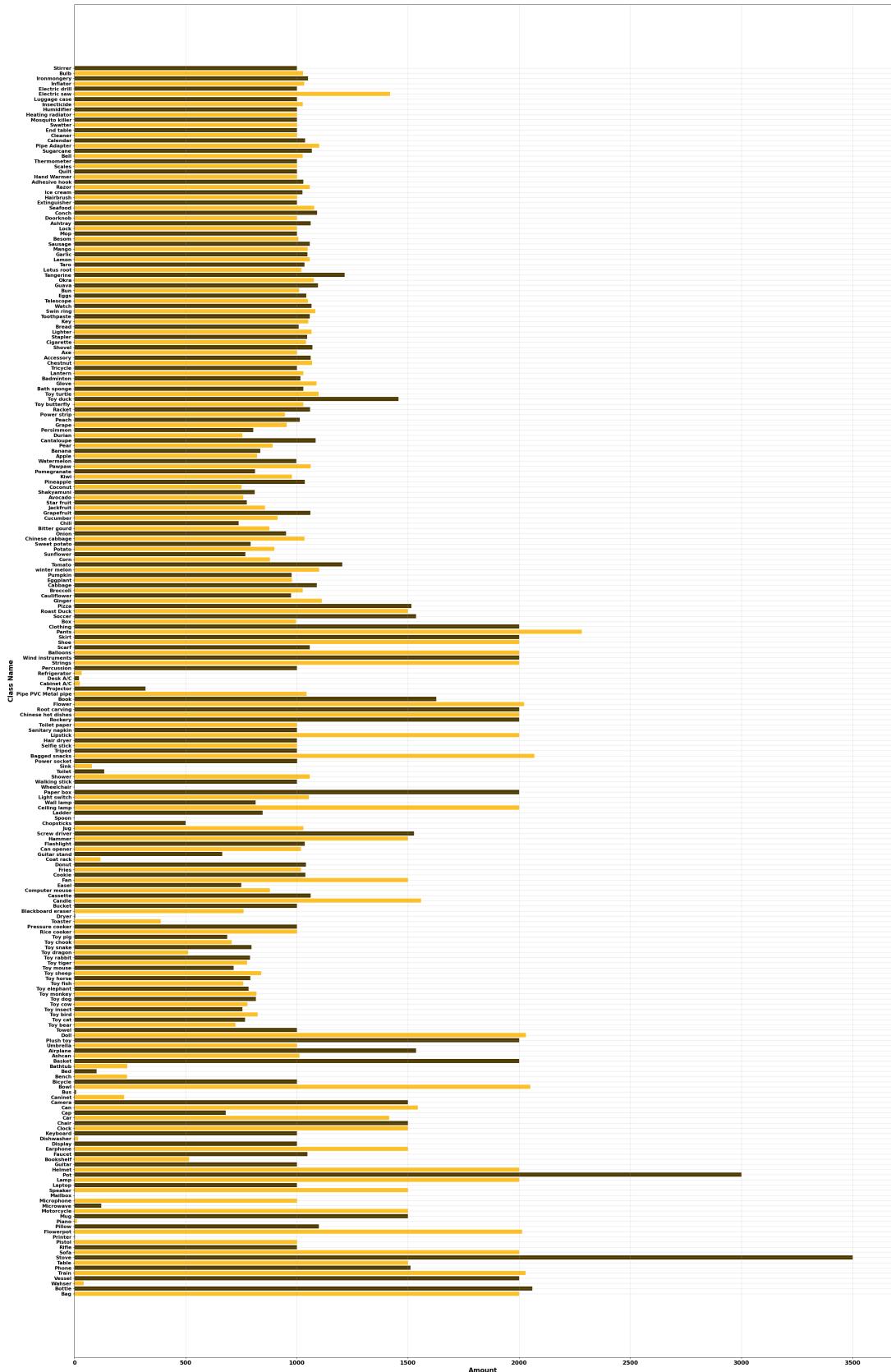


Figure III. Data amount of each category of **MVImgNet**.

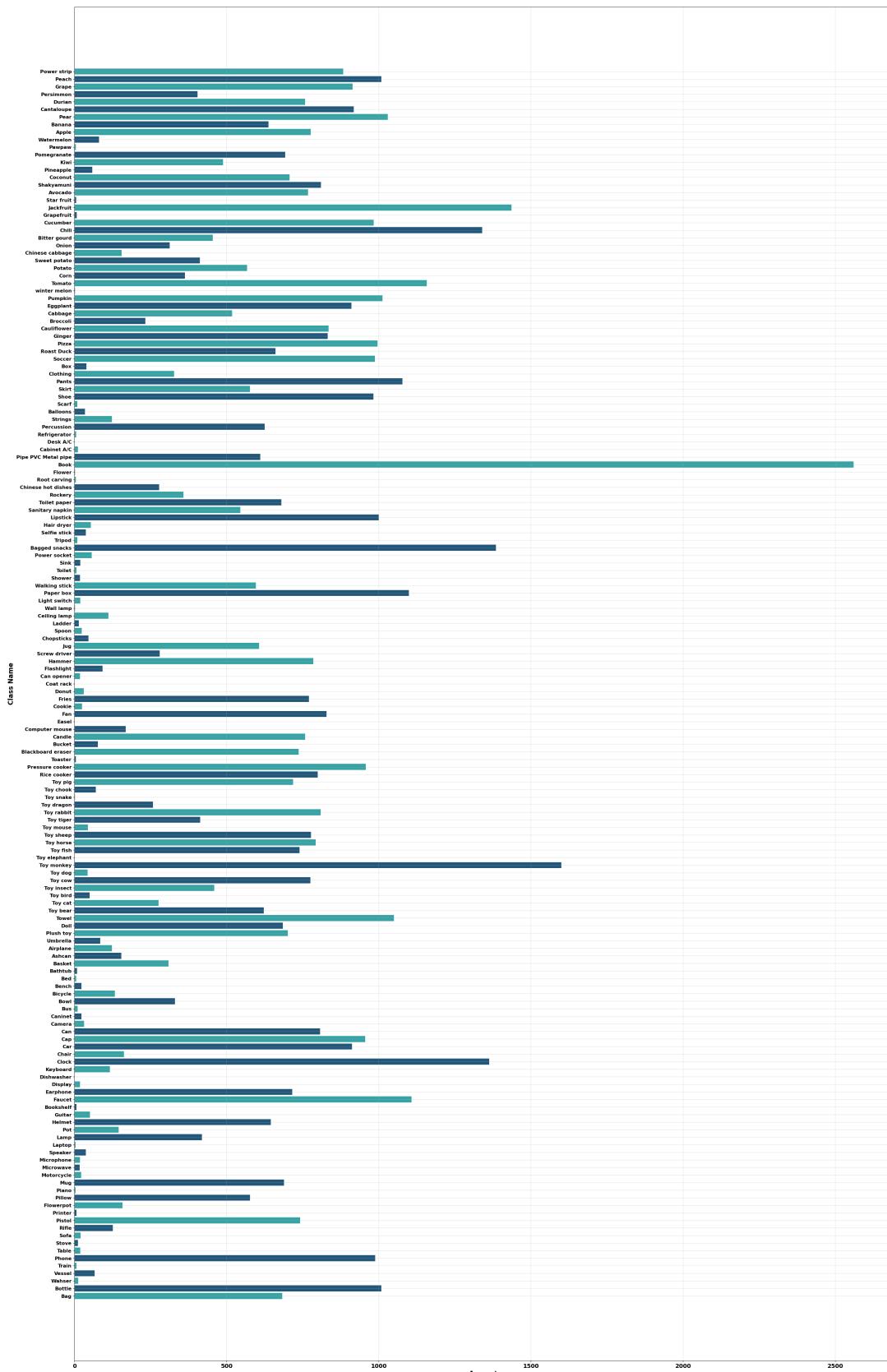


Figure IV. Data amount of each category in MVPNet.

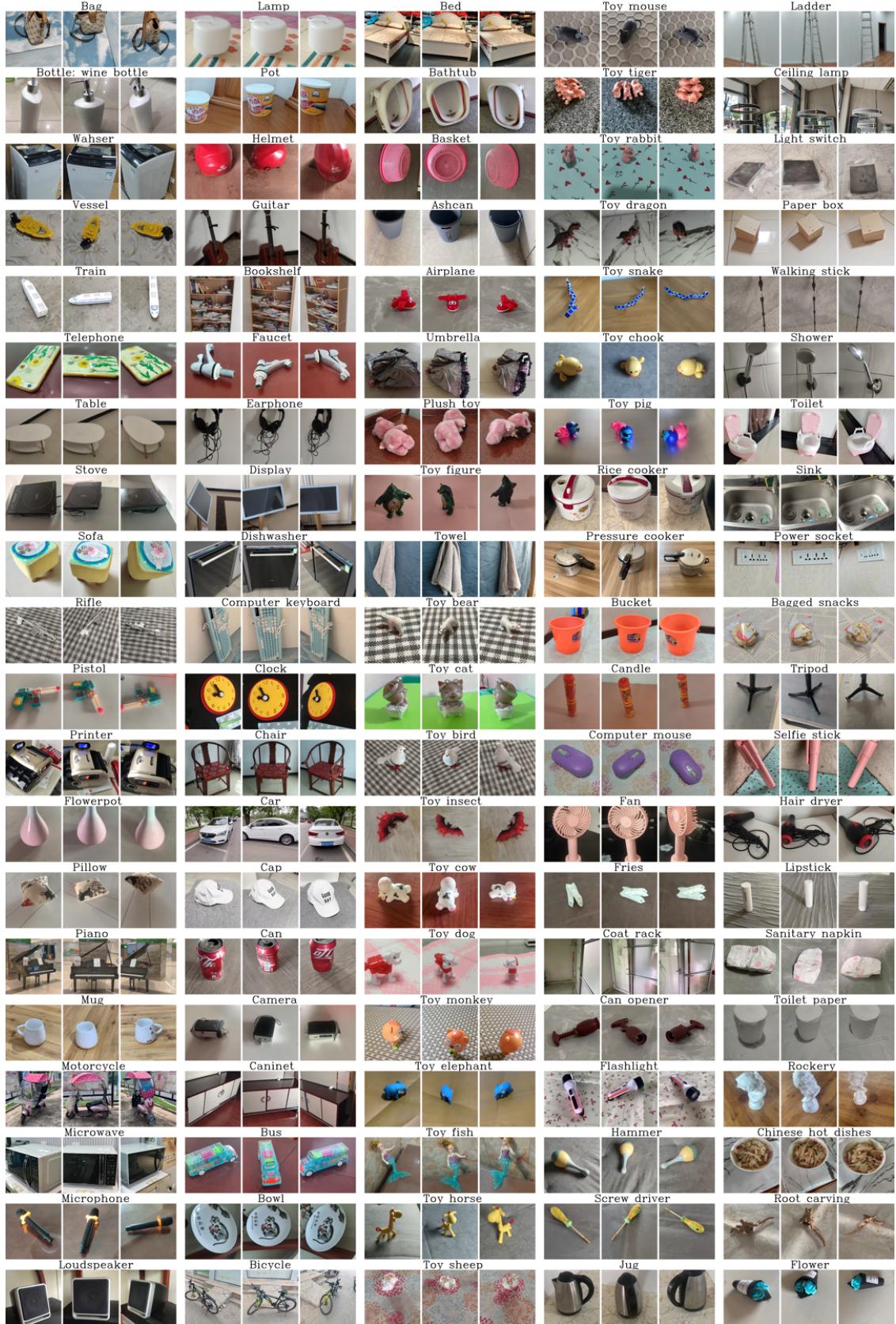


Figure V. A variety of multi-view images in **MVImgNet**.



Figure VI. A variety of 3D object point clouds in **MVPNet**.

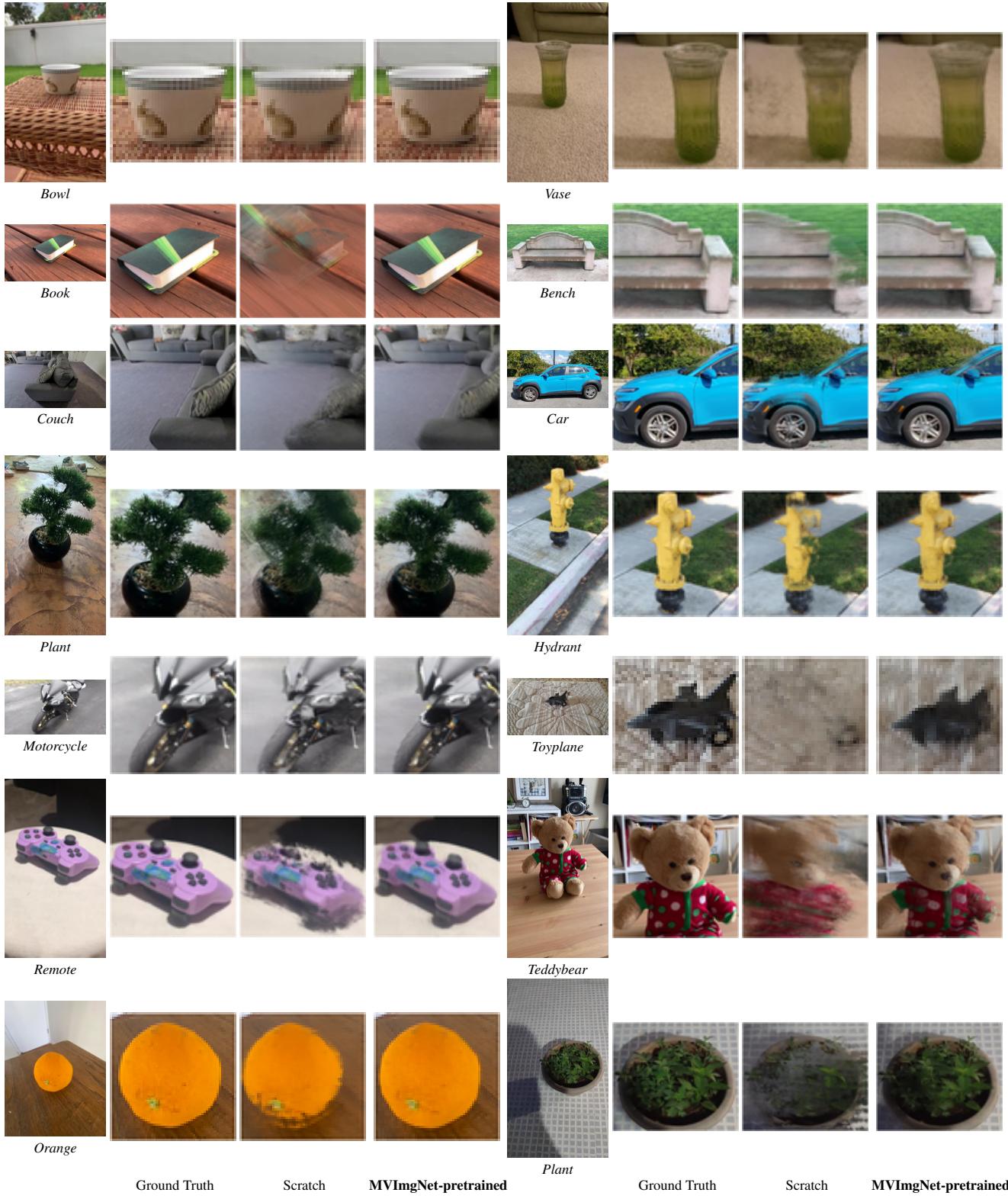


Figure VII. More qualitative comparison on real-world 360° objects [73] of **MVIImgNet-pretrained** IBRNet [92] model and the **train-from-scratch** model.

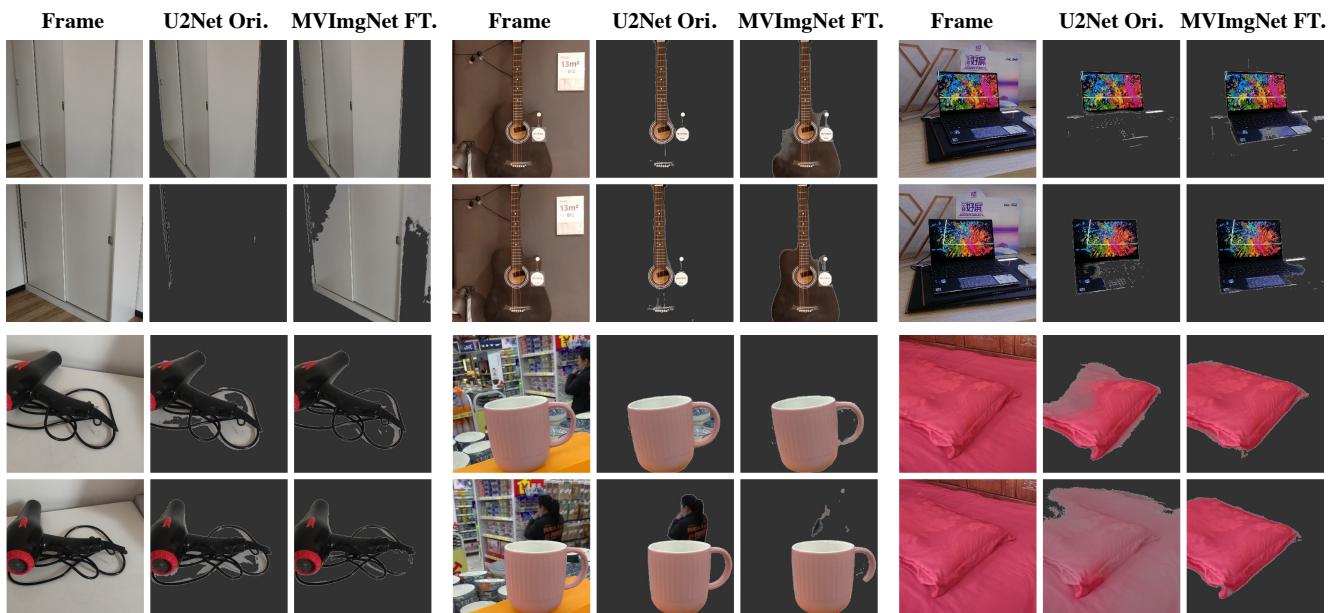


Figure VIII. More qualitative results of view-consistent salient object detection. **Finetuning U2Net [72] on MVImgNet** improves the performance.