Syn4D: A Multiview Synthetic 4D Dataset

Zeren Jiang*[1], Yushi Lan*[1], Yihang Luo[2], Yufan Deng[1], Zihang Lai[1], Edgar Sucar[1], Christian Rupprecht[1], Iro Laina[1], Diane Larlus[3], Chuanxia Zheng[2], and Andrea Vedaldi[1]

[1] VGG, University of Oxford
[2] Nanyang Technological University
[3] Naver Labs Europe
https://jzr99.github.io/syn4d/

Abstract. Progress in tasks like 3D reconstruction and tracking of dynamic scenes from monocular video is constrained by the scarcity of high-quality datasets with dense, complete, and accurate geometric annotations. To address this limitation, we introduce Syn4D, a multiview synthetic dataset of dynamic scenes that includes ground-truth camera motion, depth maps, dense tracking, and parametric human pose annotations. A key feature of Syn4D is the ability to unproject *any* pixel into 3D to *any* time and to *any* camera. We conduct extensive evaluations across multiple downstream tasks to demonstrate the utility and effectiveness of the proposed dataset, including 4D scene reconstruction, 3D point tracking, geometry-aware camera retargeting, and human pose estimation. The experimental results highlight Syn4D's potential to facilitate research in dynamic scene understanding and spatiotemporal modeling.

Keywords: Synthetic dataset · 4D reconstruction · Multiview diffusion model

Fig. 1: Syn4D is a large-scale synthetic dataset designed for a set of 4D tasks, including camera pose estimation, depth estimation, dynamic 3D scene reconstruction, 2D/3D tracking, human pose estimation, and novel-view synthesis. On the left, we visualize two shots with two synchronized frames for each shot. On the right, we unproject depth maps and visualize the camera trajectories, geometry, as well as 3D tracking results for both shots.

* Equal contribution.

1 Introduction

Much of the progress in computer vision is driven by the development of large, general-purpose neural networks, such as transformers, with research focusing on improved learning formulations and data engineering. 3D reconstruction has long resisted this trend, but recent works have shown that feed-forward networks [59, 65, 110, 111, 114, 116] can reconstruct 3D scenes, from one, a few, or hundreds of views, with performance comparable to traditional optimization-based pipelines like structure-from-motion (SfM). Furthermore, learning-based approaches can dramatically outperform geometry-based counterparts on more ambiguous tasks, such as reconstructing 3D motion from videos (4D reconstruction) [30, 38, 49, 70, 94, 96, 135, 137, 142], which have eluded traditional approaches for half a century. These problems, in fact, *require* the use of statistical priors [103], which are better captured by learning-based approaches.

So why has machine learning taken so long to become competitive with methods like SfM? We argue that a main reason has been the lack of large-scale datasets with reliable geometric annotations. Works like the aforementioned [59, 65, 110, 111, 114, 116] dedicate significant effort to collecting, curating, and annotating synthetic and real data. However, these datasets and annotations are largely private, and often only contain *static* scenes, which are insufficient for 4D reconstruction. This is a significant bottleneck to further progress.

In this paper, we address this gap and contribute **Syn4D**, a synthetic dataset designed to advance learning-based 4D reconstruction, tracking, novel-view synthesis, and related tasks (Fig. 1). We mitigate the limitations of existing datasets in scale, quality, annotation coverage, and availability. Synthetic data has proven highly effective in 2D and 3D geometry tasks such as 3D reconstruction [126] and generation [24], human pose estimation [11], and 2D/3D tracking [37, 54, 140]. Conversely, real datasets such as Stereo4D [49], while valuable, typically provide only sparse and noisy geometric annotations.

We construct Syn4D procedurally using Unreal Engine, leveraging the large catalog of high-quality 3D environments available through the Unreal Fab store. For dynamic content, we extract 1,674 animated 3D assets from Objaverse(-XL) [23, 24] and 585 simulated dynamic humans from Bedlam2 [102]. We then place these assets at carefully curated locations within each 3D environment, and design diverse camera trajectories, ranging from simple motions to complex movements, to encourage better generalization of models trained on this data. This combination of professionally designed 3D environments and diverse animated 3D assets — humans, animals, humanoid robots, and other characters — allows producing high-quality videos of dynamic scenes, as shown in Fig. 2.

Our Syn4D also comes with comprehensive annotations, including ground-truth camera motion, depth maps, point maps, *dense* 2D and 3D tracking, and parametric human pose annotations, all in a *multiview* setting. This makes it one of the few datasets that can support a wide variety of multi-view 3D and 4D tasks, from 3D tracking to and novel view synthesis (NVS). Some datasets like Kubric [37], PointOdyssey [140], and Bedlam2 [102] contain dynamic scenes, but they lack multiview information and dense tracking annotations, which are crucial for learning 4D reconstruction. In particular, this is the first dataset to provide dense and complete 3D tracking annotations in a multiview setting, enabling the recovery of the 3D position of any point in any image at any time (and thus its 2D projection in any other frame and camera), offering an advantage over datasets like PointOdyssey [140]. Because storing this information explicitly is infeasible, we also develop an efficient representation of these dense tracks based on pixel-aligned barycentric maps paired with the corresponding mesh sequences. This design enables efficient querying of dense tracks within data loaders for training or evaluation.

Overall, SYN4D contains **4.7K multiview video clips**, totaling **1.4M frames**, with **dense geometry annotations**. This is significantly larger than existing 3D datasets that offer dense geometry and motion annotations (Tab. 1).

Empirically, we show that training on SYN4D significantly improves the performance of off-the-shelf models in geometry-aware NVS, 4D reconstruction and tracking, and human pose estimation. Because our SYN4D contains *multiview* videos with corresponding ground-truth geometry, we also introduce a new task, benchmark, and evaluation metrics for geometry-aware dynamic NVS, considering both the visual quality and geometric consistency of the generated views. Another unique property of SYN4D is its dense tracking annotations, which support new benchmarks and evaluation metrics for 3D tracking. Finally, we evaluate the performance of state-of-the-art 4D reconstruction and human pose prediction models trained on SYN4D.

To summarize our contributions, SYN4D is the first publicly available dataset containing multiview videos of dynamic scenes with dense 3D tracking annotations, along with camera images, depth maps, and human pose labels. We demonstrate that this data can be used to improve the performance of state-of-the-art 2D/3D tracking and 4D reconstruction models. We further show that the dataset enables training new models, including a geometry-aware multiview diffusion model that can generate novel-view video sequences together with consistent geometry.

2 Related Work

2.1 3D and 4D datasets

We review datasets related to SYN4D with respect to their synthetic *vs*. real nature and static *vs*. dynamic content. As shown in Tab. 1, for *realistic* dynamic scenes with *diverse* categories and *multiview dense* geometry and motion annotations, none of the existing datasets can fully meet these requirements.

Synthetic static 3D datasets. Synthetic data offers numerous benefits, including "perfect" annotations. However, creating high-quality synthetic 3D data is complex. In part, this is due to the difficulty of procuring diverse 3D assets, including scenes, characters, and animations, for building useful synthetic scenes. ShapeNet [17], Objaverse(-XL) [23,24], and others [121] collect a large number of

Dataset	**Camera**	**Track**	**HP**	**SF**	**OF**	**IS**	**Capt.**	**Engine**	**# Dyn. Obj.**	**# Clips**	**# Frames**
Sintel [13]	Stereo	×	×	×	✓	✓	×	Blender	Few others	35	1.6K
Spring [75]	Stereo	×	×	✓	✓	✓	×	Blender	Few others	47	6K
FlyingThings3D [73]	Stereo	×	×	✓	✓	✓	×	Custom	Many Rigid	2.2K	35K
Kubric (MOVi-F) [37]	Mono	Dense	×	✓	✓	✓	×	Blender	Many Rigid	4K	96K
Falling Things [104]	Stereo	×	×	×	×	✓	×	Unreal	21(Rigid)	3.5K	61K
VIPER [84]	Mono	×	×	●	✓	✓	×	GTA V	Cars/Pedestrians	N/A	254K
JTA [29]	Mono	×	✓	●	●	✓	×	GTA V	Cars/Pedestrians	512	461K
TartanAir [115]	Stereo	×	×	●	✓	✓	×	Unreal	Few others	1037	1M
SceneNet RGB-D [74]	Mono	×	×	●	✓	✓	×	Custom	None	15K	5M
Virtual KITTI [14,33]	Mono	×	×	●	✓	✓	×	Unity	Cars/Pedestrians	35	17K
Virtual KITTI 2 [14]	Stereo	×	×	✓	✓	✓	×	Unity	Cars/Pedestrians	50	21K
BlendedMVS [126]	Multi	×	×	●	●	×	×	Custom	None	113	18K
Dynamic Replica [55]	Stereo	Sparse	×	●	✓	✓	×	Blender	375(humans)+13(others)	524	169K
PointOdyssey [140]	Multi	Sparse	×	×	×	✓	×	Blender	42(humans)+7(others)	104	216K
BEDLAM2* [11,102]	Mono	×	✓	×	×	×	×	Unreal	4K(humans)	12K	3.5M
SYNTHIA [86]	Multi	×	×	×	×	×	×	Unity	Cars/Pedestrians	4	200K
SEED4D [58]	Multi	×	×	●	✓	✓	×	Unreal	Cars/Pedestrians	10.5K	16.8M
SURREAL [107]	Mono	×	✓	●	●	✓	×	Blender	930(humans)	67K	6.5M
Synscapes [119]	Mono	×	×	×	×	✓	×	Custom	None	Still images	25K
Kubric (Ours)	Multi	Dense	×	✓	✓	✓	Global	Blender	Many Rigid	5.6K	274K
Syn4D (Ours)	Multi	Dense	✓	✓	✓	✓	G.+Local	Unreal	585(humans)+1674(others)	4.7K	1.4M

Table 1: Comparisons with previous synthetic datasets with geometry annotation. *We only count the number of frames that have geometry annotation. All datasets include camera and depth/disparity annotations. Other annotations may include Optical Flow (OF), Scene Flow (SF), Instance Segmentation (IS) consistent across frames, Long-term Point Tracking (Track), Human Pose (HP), noted in the table. A ● means that the annotation can be derived in part from the ones provided (for instance SF can be derived from depth, OF and cameras up to occlusions).

3D assets from the web, and have significantly advanced the 3D and 4D generators [44,48,67,120,123,136], but they are limited mainly to *objects*. For *scene-level* datasets, such as BlendedMVS [126], Replica [91], Structured3D [139], Hypersim [85], SUN RGB-D [90], TartanAir [115], and 3D-FRONT [32], they are either relatively small in scale (*e.g.*, Replica), or lack realistic materials (*e.g.*, Structured3D and 3D-FRONT) and ground-truth geometry (*e.g.*, Hypersim and SUN RGB-D).

Some authors also explore procedural generators to create infinite 3D scenes. For example, InfiniGen [82] provides an engine to build any number of 3D scenes procedurally but contains few animations and highly stylised content. InfiniGen Indoors [83] constructs indoor environments instead.

Synthetic dynamic 3D datasets. Instead of providing only static geometry, more relevant to our work are dynamic synthetic datasets. To diversify the content, datasets like Sintel [13] and Spring [75] utilize free open source Blender movies. Others, like FlyingThings3D [73] and Kubric [37] opt for "domain randomization", tossing random objects from ShapeNet [17] and Google Scanned Object (GSO) [28], respectively. AI2-THOR [61] and ProcTHOR [25] provide instead manually-authored interactive 3D environments. BlenderProc2 [26] can also be used to create data similar to FlyingThings3D via procedural generation. Other datasets tap video games for content. A notable example is VIPER [84], which extracts depth, cameras, and semantic and instance segmentation from

GTA; another is JTA [29], which further extracts body poses. Others again, use domain-specific resources; for instance Virtual KITTI [14,33] builds on the CARLA [27] simulator, providing synthetic renderings, depth, and semantic segmentation. SYNTHIA [86], Synscapes [119], SURREAL [107], PreSIL [45] and SEED4D [58] also focus on the driving/urban domain. While they provide some dynamic content, most of them use only a few dynamic objects, or are restricted to specific categories such as cars and pedestrians, or rigid objects. Recent BEDLAM [11,102] provides more complex and realistic animations, but focuses only on humans. DynamicReplica [56] introduces other characters, but only 13 in total. In contrast, our Syn4D combines 1,674 selected animated objects from Objaverse [23,24] and 585 3D human assets from BEDLAM [11,102], along with 30 large-scale 3D scene assets we acquired from Sketch Fab, containing natural-looking clutter, complex illumination, and varied geometry and materials.

For the annotations, some datasets, such as Sintel, Spring, FlyingThings3D, TartanAir, SEED4D, and Kubric, provide all or some of the scene flow and stereo. Although scene flow is equivalent to dense 3D tracking, it is only available for consecutive frames of a video. Stereo provides two viewpoints that are close together. Others like BlendedMVS [126] and SYNTHIA [86] provide multiview geometry, but only for static scenes or a few dynamic objects. PointOdyssey [140] and Dynamic Replica [56] provide long-term tracking, but only for very *sparse* points. In contrast, our Syn4D provides *dense multiview* 3D tracking for any pair of frames, including those far apart in time and space. One direct competitor is Kubric [37], which can be used to create dynamic scenes with dense multiview geometry and tracking. However, it contains only rigid objects that fall to the ground, which lacks realism. It also lacks 3D background scenes. Our Syn4D, on the other hand, provides a realistic rendering with Unreal Engine 5 and contains a much more diverse and realistic set of 3D assets, scenes, and lighting.

Real static 3D datasets. Here, we briefly review the real 3D datasets. ABO [21], GSO [28], and OmniObject3D [121] provide object scans. For the indoor scene datasets, notable examples like Matterport3D [16], SceneNN [1] and its follow-up SceneNet RGB-D [74], ScanNet(++) [22,128] and ARKitScenes [9], provide RGB-D scans of real indoor scenes, using different RGBD sensors and scanning protocols. However, the resulting data is often noisy, and the geometry is incomplete. Besides, Gibson [122] and Habitat [87,100] build 3D environment based on scanning real 3D scenes. Although the holes are largely filled, the resulting geometry remains inaccurate. In parallel, many outdoor datasets, such as KITTI [35], SemanticKITTI [10], nuScenes [15], Waymo Open Dataset [97], KITTI-360 [64], and Argoverse2 (3D) [118] provide large-scale outdoor scans, but with very sparse geometry.

Real dynamic 3D datasets. For dynamic scenes, real datasets are even more limited. While TUM RGB-D [92], ICL-NUIM [39], and ETH3D [88] provide RGB-D videos, they do not contain real dynamic objects, but only static scenes with camera motion. Middlebury [7] builds stereo and optical flow benchmarks using real videos, but they are very small in scale and contain only a few rigid objects. To record real scenes with dynamic content, Panoptic Studio [50, 52] builds a dome with more than 500 cameras to capture multi-person interactions in a studio. 3DPW [72] scans 3D body shapes and poses of people in outdoor scenes using IMUs. PROX [40] captures humans interacting with scanned indoor scenes using a multi-camera setup. However, these datasets are limited to human interactions and do not contain other dynamic objects or complex scenes, making them less suitable for general 4D reconstruction and tracking research.

2.2 Applications

4D reconstruction and tracking. Recent contributions like DUSt3R [114], and its follow-ups [59,65,110,113,116] have made significant progress on *feed-forward* 3D reconstruction by using both real and synthetic large scale datasets for training. These methods have recently been extended for feed-forward 4D reconstruction and tracking across time [30, 49, 57, 71, 94, 96], working by extending the 3D reconstructors to dynamic scenes. In parallel, some works [47, 48, 127, 134, 143] have explored the potential of using pre-trained video generators [12, 42] for this task. Another category of methods has tackled 4D reconstruction as an extension of 2D point tracking by combining it with video depth predictions [124, 133].

Camera retargeting. Camera retargeting aims to generate content from arbitrary user-specified viewpoints. Early works, such as Zero-1-to-3 [67] and follow-ups [19, 34, 101, 108, 138, 141], present camera-pose-conditioned diffusion models for NVS, but they are limited to *static* scenes. Camera Dolly [106] extends a similar pipeline to *dynamic* scenes by finetuning SVD [12] with multiview videos rendered by Kubric [37]. Recently, there are several works [3, 4, 6, 41, 125, 131] targeting *long-term* videos and *large-scale* camera motions, but they mainly focus on different architectures and conditions. In contrast, we provide *multiview* geometry-grounded videos, which can be used not only for novel-view appearance synthesis but also for novel-geometry synthesis.

Human pose estimation. Human pose and shape estimation from images typically fits parametric body models—SMPL [69] and SMPL-X [80] either from cropped single-person detections [36, 53, 60, 78, 129], achieving accurate per-person estimates but discarding scene context for occlusion and interaction, or leveraging the full image for richer, scene-aware estimation [8, 46, 93, 98, 99, 112, 117]. A key finding is that combining diverse synthetic datasets (*e.g.*, BEDLAM [11], AGORA [79]) with real-image collections is critical for accurate estimation at scale. Syn4D advances this direction by providing SMPL-X annotations rendered in high-quality, geometrically diverse 3D environments with multiview cameras.

3 Method

We first formulate the dataset content in Sec. 3.1. Then, in Sec. 3.2, we describe our key contribution to efficiently store the dense tracking annotations.

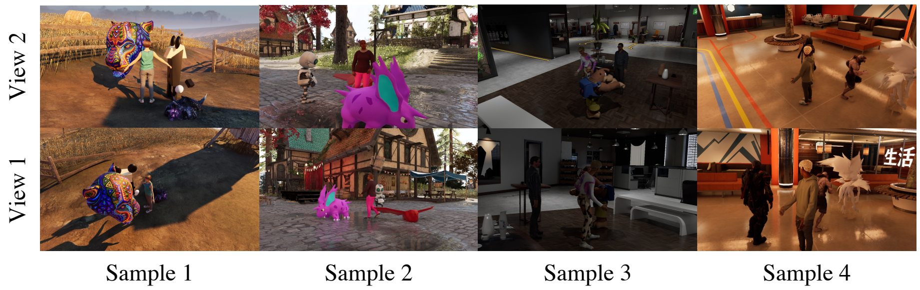


Fig. 2: Syn4D is a multiview synthetic dataset with diverse dynamic objects, including humans, animals, humanoid robots, and other characters.

In Sec. 3.3, we describe how we construct the dataset, including asset filtering and camera motion design. We generate global and local captions using a vision-language model, as discussed in Sec. 3.4. Finally, we propose a geometry-aware multiview diffusion model trained on our Syn4D dataset in Sec. 3.5.

3.1 Dataset content

Syn4D consists of a collection $\mathcal{S}$ of clips. Each clip is obtained by randomly combining a 3D environment with various dynamic 3D objects, then rendering and annotating them, as further detailed in Sec. 3.3. The clip contains information captured by C *cameras* over T frames. First, the cameras produce RGB frames $I_i \in \mathbb{R}^{3\times H\times W}$. Each frame thus has a *camera index* $c_i \in \{0, \dots, C-1\}$ and a *time index* $t_i \in \{0, \dots, T-1\}$. In addition to the RGB frame, each camera also comes with its *parameters* (extrinsics and intrinsics) π_i. Because there is one frame per camera and per time, the index i ranges from 1 to $N = TC$.

A unique feature of our dataset is the ability to *track any point*. This means that, given a pixel $u \in \{0, \dots, H-1\} \times \{0, \dots, W-1\}$ in an image I_i, we can tell the 3D position of the corresponding point at any time, as well as its 2D projection in any other camera. Formally, this information is captured by the *dynamic point maps* (DPMs) [95,96]

$$P_i(\pi_k, t_j) \in \mathbb{R}^{3\times H\times W}.$$

There is one DPM P_i for each image I_i. If $u \in \{0, \dots, H-1\} \times \{0, \dots, W-1\}$ is a pixel, then $P_i(\pi_k, t_j)(u) \in \mathbb{R}^3$ is the 3D location that the physical point corresponding to pixel $I_i(u)$ has at time t_j (which may differ from the time t_i of the image), expressed in reference frame π_k (which can also differ from the reference frame π_i of the image's camera). Finally, we also provide instance segmentation maps S_i that associate each pixel u of an image with a unique ID, distinguishing different dynamic objects in the clip.

To summarize, for each clip our dataset provides tuples $\{(I_i, P_i, S_i, t_i, \pi_i)\}_{i=1}^{N}$ describing the appearance and dense 3D geometry and motion of the scene. In addition, depth maps D_i are derived by taking the z channel of the corresponding point maps $P_i(\pi_i, t_i)$, and they are also included in the DPMs. Additional metadata includes captions for each clip, both global and local (one every 81 frames), describing the clip's content, useful for tasks like learning video generators.

3.2 An efficient dynamic point map representation

To the best of our knowledge, our dataset is the first to provide *multi-view* dense tracking annotations for general dynamic objects. In part, this can be explained by the difficulty of storing such annotations. For example, PointOdyssey [140] explicitly stores sparse tracking annotations, and it would be theoretically possible, but impractical, to store one such track for each pixel in every image.

With our notation, dense tracks are captured by DPMs $P_i(\pi_j, t_k)$. Each DPM is a $H \times W$ image with three components per pixel and is indexed by i, π_j, and t_k. The index i enumerates the N images in the clip, whereas π_j and t_k index viewpoints and times. There are N possible viewpoints π_j, one for each image (camera-time combination) in the clip, and T possible times. Hence, all DPMs together form a $(3 \times H \times W) \times N \times N \times T$ tensor, which has space complexity $\mathcal{O}(HWT^3C^2)$ (by definition $N = TC$). A saving comes from the fact that point maps that differ only by the reference frame π are related by a known rigid transformation; *i.e.*, given $Q_i(t_k) = P_i(\pi_0, t_k)$ for a fixed reference frame π_0 (*e.g.*, that of camera c_0 at time t_0), we can find $P_i(\pi_j, t_k)$ from $Q_i(t_k)$ by applying the known rigid transformation between π_0 and π_j. This reduces the space complexity to $\mathcal{O}(HWT^2C)$, which is much better, but still impractical. For instance, if $H = W = 512$, $T = 300$, and $C = 8$, and assuming four bytes per scalar, we would have to store 2.1TiB of information for just one clip! This should be compared to the 7GiB required to store the (raw) RGB frames, which is negligible in comparison.

Intuitively, this representation is highly redundant because each 3D point trajectory is represented multiple times, once for each image that observes it. Internally, Kubric [37] takes advantage of the rigidity of the objects it contains to represent tracks efficiently, but this is not possible here due to the non-rigid deformations of the objects.

We solve this problem as follows. First, we note that all 3D surfaces in the clip can be represented as an animated triangular mesh, the union of the meshes of all clip objects. Let this mesh be represented by a collection of time-varying vertices $q_v(t)$, $v = 1, \dots, V$ and (fixed) triangular faces $F \subset \{1, \dots, V\}^3$ that are triplets of such vertices. Assume that the vertices are expressed in the reference frame π_0. Given a pixel u in image I_i, we can recover the 3D point $Q_i(t)(u)$ by finding the triangular face $f = (f_1, f_2, f_3) \in F$ that contains that pixel and computing the corresponding barycentric coordinates $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \varDelta_3$. Here $\varDelta_3 \subset [0, 1]^3$ is the 3-simplex, *i.e.*, the set of triples of non-negative numbers that sum to one. Then, for all $t = 0, \dots, T - 1$, we have:

$$Q_i(t)(u) = \alpha_1 q_{f_1}(t) + \alpha_2 q_{f_2}(t) + \alpha_3 q_{f_3}(t). \tag{1}$$

Hence, for each pixel u, we only need to store four scalars (one index to identify the face f in the collection F, and three barycentric coordinates α). Together with the $3 \times V \times T$ scalars for the vertex trajectories, this has complexity $\mathcal{O}(HWTC + VT)$, which is much more manageable. In the example above, with $V = 100K$ vertices, we would only need 9.3GiB to store all DPMs. In practice, further significant reductions are possible by packing the data in fewer bits and noting that the information is required only for pixels belonging to dynamic objects, which constitute a small fraction of the total.

Note also that the decoding is highly efficient: to recover the track that passes through pixel u in frame I_i, *i.e.*, the sequence $(Q_i(t))_{t=0}^{T-1}$, we linearly combine vectors $(q_f(t))_{t=0}^{T-1}$ as shown in Eq. (1).

Extraction. Unreal Engine has no rendering pass that can directly produce the quantities f and α discussed above, so we compute them in pre-processing. Given the depth map $D \in \mathbb{R}^{H \times W}$ (which Unreal provides) and camera rotation $R \in \mathbb{R}^{3 \times 3}$, position $o \in \mathbb{R}^3$, and intrinsics $K \in \mathbb{R}^{3 \times 3}$ for an image I, we recover the 3D coordinate of the point at pixel u as $\bar{q}(u) = R^{-1} K^{-1} D(u)\,(u, 1)^\top + o$. Then, the face $f(u)$ and coordinates $\alpha(u)$ for that pixel are given by:

$$(f(u), \alpha(u)) = \underset{f \in F, \alpha \in \Delta_3}{\operatorname{argmin}} \|\bar{q}(u) - \alpha_1 q_{f_1}(t) - \alpha_2 q_{f_2}(t) - \alpha_3 q_{f_3}(t)\|.$$

3.3 Dataset construction

Next, we describe how we generate our synthetic clips, including their 3D content, cameras, and captions.

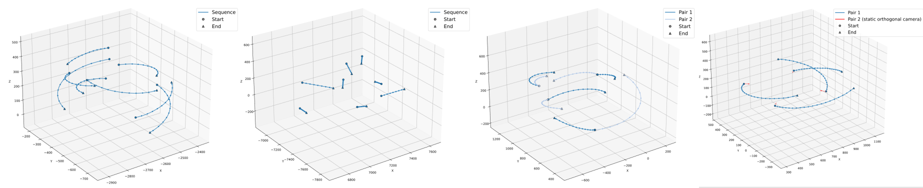
Content. We construct scenes by combining 3D environments and dynamic assets and randomly placing them according to manually defined rules.

For the *environments*, we manually select (and purchase) 30 3D scenes from the Fab store and choose an appropriate location in each scene as the root position for placing our dynamic objects and humans.

For the *objects*, unlike Bedlam2, which contains only humans, we include a more diverse set of dynamic objects, such as robots, animals, and monsters, from the Objaverse dataset. However, the 4D assets from Objaverse contain noisy, low-quality animations. We therefore first render each animated object from a bird's-eye view to filter out objects with large deformations or fast motion by analyzing the segmentation masks. Specifically, we compare the intersection-over-union (IoU) between adjacent frames to estimate motion speed and filter out frames with low IoU and extreme deformations.

For the *composition*, we randomly choose 1–3 dynamic objects from the filtered Objaverse dataset and one human from the Bedlam2 released assets. As in Bedlam2, we use the ground-occupancy map to randomly initialize the layout to avoid collisions.

For *illumination*, we use the Lumen real-time global illumination system introduced in UE5 for nearby movable light sources, and precomputed baked lightmap textures for static lights.



(a) Orbit Shots (b) Dolly Shots (c) Paired Orbit Shots (d) Paired Static-Orbit Shots

Fig. 3: Sample camera motion in our dataset.

Camera motion. Another key design decision is the camera optics, placement, and motion. For the camera *intrinsics*, we cover a broad range of horizontal fields of view from 39.6° to 90°. We dynamically adjust the focal length range depending on whether the scene is indoor or outdoor to better mimic real-world shots. For the camera *extrinsics*, we generate a variety of camera-motion combinations: static, tracking, dolly, and orbit. Similar to Bedlam2, we also apply synthetic Perlin-noise camera shake to the camera extrinsics on top of these motions. The majority of camera motions are orbits, allowing us to capture the full dynamic scene with fewer shots. We use keyframes to define the extrinsic changes in Unreal Sequencer. For the orbit shots, we randomly select the initial camera position in spherical coordinates relative to the camera root within a specified range. We then randomly generate the delta values for the radial distance, polar angle, and azimuth angle between the first and last keyframes.

To obtain a *multiview* dataset, we generate eight different cameras per clip. As shown in Fig. 3(a,b), for some scenes we generate eight independently sampled orbit and dolly shots to maximize the coverage of the dynamic scene. For the others, we generate paired shots. Specifically, in Fig. 3(c), we generate two paired orbit shots that start from a shared frame. As shown in Fig. 3(d), we first generate four orthogonal static shots and then use them as starting points for another four orbit shots. These different multiview camera patterns facilitate training the multiview diffusion model to generalize to various source and target camera configurations.

3.4 Captioning

We use Tarsier2–7B [132] to caption our SYN4D dataset. Since each video sequence contains around 300 frames, we provide captions at two levels of granularity. We sample 16 frames from the entire sequence to generate one global caption. We then split the video into clips of 81 frames and sample 32 frames from each clip to generate one local caption. These local and global captions provide users with flexibility when sampling videos of varying lengths for training.

3.5 Geometry-aware multiview diffusion model

In the experiments, we test how our new dataset can help in a variety of tasks, from 4D reconstruction to new-view synthesis. We also explore a new task uniquely supported by our data, namely the simultaneous synthesis of a novel view and the corresponding 4D geometry (*i.e.*, the 3D points and their motion).

Given a source video $I \in \mathbb{R}^{3\times H\times W\times T}$ and its corresponding point maps $(P_i(\pi_0, t_i))_{i=0}^{T-1} \in \mathbb{R}^{3\times H\times W\times T}$, we synthesize a target video $I' \in \mathbb{R}^{3\times H\times W\times T}$ and its corresponding point maps $(P'_i(\pi_0, t_i))_{i=0}^{T-1} \in \mathbb{R}^{3\times H\times W\times T}$ from a new set of cameras $\pi' = (\pi'_i)_{i=0}^{T-1}$. We start from ReCamMaster [4], which takes as input a video I and the target camera trajectory π' to generate a new video I' seen under π'. We then extend it to take P as input and to output P'. Because ReCamMaster uses latent diffusion, we adapt their video encoder $\mathcal{L}_{\text{vid}} = \mathcal{E}(V) \in \mathbb{R}^{H\times W\times T\times d}$ into a DPM encoder $\mathcal{L}_{\text{DPM}} = \mathcal{E}(P) \in \mathbb{R}^{H\times W\times T\times d}$. We then reuse their model architecture but spatially concatenate the video latents as $\mathcal{Z} = [\mathcal{L}_{\text{vid}}; \mathcal{L}_{\text{DPM}}] \in \mathbb{R}^{H\times(2W)\times T\times d}$. With spatial concatenation, we can leverage the pretrained diffusion prior for the geometry-aware novel-view synthesis task without introducing new model parameters.

4 Experiments

To evaluate the quality of SYN4D and its effectiveness for improving state-of-the-art models on a set of 4D attributes, we conduct thorough experiments on various tasks. We first report results on our proposed task of geometry-aware novel view synthesis in Sec. 4.1. Then, in Sec. 4.2, we train a state-of-the-art 4D reconstruction and tracking model, showing that our dataset supports a broad set of tasks, including 3D tracking, video depth estimation, camera pose estimation, and multi-view reconstruction. Finally, in Sec. 4.3, we fine-tune a state-of-the-art human pose estimator on our dataset, further demonstrating its utility for human pose estimation.

4.1 Geometry-aware multiview diffusion model

Metrics. Following baseline ReCamMaster [4], we report the FVD [105] and CLIP-V [81] to measure the temporal consistency of generated novel views and their alignment with the source video, respectively. Except for the visual appearance, we further estimate camera trajectories from generated views using Back on Track [18], align them to the ground-truth conditioned trajectories with the Umeyama algorithm, and report three standard metrics: Absolute Translation Error (ATE), Relative Translation Error (RPE-T), and Relative Rotation Error (RPE-R). For geometry evaluation, direct pixel-aligned absolute relative error is unsuitable because the generated novel views are not pixel-aligned with ground-truth target videos. We therefore convert point maps to RGB and evaluate them with CLIP-V and FVD, denoted as CLIP-V-P and FVD-P, respectively.

Method	Visual Quality		Camera Accuracy			Geometry Quality	
	CLIP-V ↑	FVD ↓	ATE ↓	RPE trans ↓	RPE rot ↓	CLIP-V-P ↑	FVD-P ↓
Ours (Kubric)	0.643	631	**0.064**	0.023	0.328	0.757	229
Ours (Syn4D)	**0.740**	**452**	0.070	**0.021**	**0.272**	**0.816**	**139**

Table 2: Quantitative evaluation for geometry-aware novel view synthesis on Syn4D evaluation benchmark.

Experiment setup. We fine-tune our geometry-aware multiview generator on the Kubric and Syn4D datasets. We evaluate the model on our in-house benchmark, which includes 280 video pairs with scenes and dynamic objects that are completely unseen during training.

Results. As shown in Tab. 2, our model trained on Syn4D outperforms the model trained on Kubric in all visual and geometric metric. It achieves camera accuracy comparable to that of the Kubric baseline, suggesting that future versions of the dataset would benefit from more diverse paired camera patterns.

4.2 4D reconstruction and tracking

Here, we further evaluate whether training with Syn4D improves learning-based 4D reconstruction. Quantitative results (Tabs. 3 to 5) show that co-training the 4D reconstructor with Syn4D consistently surpasses its original performance across multiple 4D tasks, including 3D tracking, camera pose estimation, multi-view 3D reconstruction, and video depth estimation. This shows that Syn4D provides valuable geometric supervision for learning spatiotemporal scene representations. Details are discussed in the following.

Metrics. We employ task-specific metrics for each dimension of 4D reconstruction. For *3D tracking*, we use the Average Percentage of Points (APD) within a threshold and the End-Point Error (EPE). For *camera pose estimation*, we report Absolute Translation Error (ATE) and Relative Pose Error (RPE) for both translation and rotation. Multi-view *point maps reconstruction* is assessed using Accuracy (Acc), Completeness (Comp), and Normal Consistency (NC). *Video depth estimation* is evaluated using the Absolute Relative Error (Rel) and threshold accuracy ($\delta < 1.25$), under both scale and scale-and-shift alignment.

Experiment setup. We adopt 4RC [71] as our baseline. It jointly predicts camera poses, dense geometry, and motion from monocular videos, in a single feed-forward pass. We use the *de facto* architecture and exactly the same training protocol as 4RC, while augmenting the training data with the Syn4D dataset.

For evaluation, we follow the standard protocols used in prior work [63, 71, 113, 114]. Sparse 3D point tracking is evaluated on Aerial Digital Twin (ADT) [77] and Panoptic Studio (PStudio) [51] from TAPVid-3D [62], which provide sparse

Method	Sparse Point Tracking				Dense Tracking			
	ADT [77]		PStudio [51]		Soviet		Warehouse	
	APD ↑	EPE ↓	APD ↑	EPE ↓	APD ↑	EPE ↓	APD ↑	EPE ↓
4RC	87.82	0.1480	87.32	0.1304	58.35	3.8458	79.07	0.3302
4RC (SYN4D)	**89.44**	**0.1258**	**88.10**	**0.1276**	**74.46**	**1.8934**	**88.79**	**0.1915**

Table 3: Quantitative evaluation for 3D tracking. We compare the baseline 4RC [70] with the model retrained with the original dataset plus SYN4D. Notably, co-training with SYN4D yields significant improvement on both *sparse* and *dense* 3D tracking, with the identical experimental setting.

Method	Camera Pose Estimation									Multi-View 3D Reconstruction					
	Sintel [13]			TUM-dynamics [92]			ScanNet [22]			7-Scenes [89]			NRGBD [2]		
	ATE ↓	RPE_t ↓	RPE_r ↓	ATE ↓	RPE_t ↓	RPE_r ↓	ATE ↓	RPE_t ↓	RPE_r ↓	Acc ↓	Comp ↓	NC ↑	Acc ↓	Comp ↓	NC ↑
4RC	0.144	0.053	0.430	**0.010**	**0.008**	**0.314**	0.032	**0.012**	0.437	0.034	0.051	0.783	0.036	0.034	0.912
4RC (SYN4D)	**0.076**	**0.040**	**0.302**	0.012	0.010	0.325	**0.032**	0.013	**0.384**	**0.031**	**0.043**	**0.791**	**0.029**	**0.032**	**0.924**

Table 4: Quantitative evaluation for camera pose estimation and multi-view 3D reconstruction. Co-training with SYN4D significantly improves the 3D reconstruction performances on all instantiations, while achieves comparable or better performances on camera pose estimation.

trajectory annotations. To further assess the performance on dense tracking, we additionally evaluate on two rendered test sets from SYN4D, Soviet and Warehouse, each containing 50 sequences. For each sequence, we sample a 24-frame clip and evaluate dense trajectories starting from the first frame. We also evaluate camera pose estimation on Sintel [13], TUM-dynamics [92], and ScanNet [22], video depth estimation on Sintel, Bonn [76], and KITTI [35], and multi-view 3D reconstruction on 7-Scenes [89] and NRGBD [2].

Results. We report quantitative results in Tabs. 3 to 5. Training with SYN4D consistently improves performance across all evaluated tasks, *without requiring any specific architecture modification.* Notably, clear gains are observed in 3D tracking, showing that improvements from training data are significant, which suggests another important direction in 3D research, *i.e.*, building large-scale datasets. A similar conclusion is also observed on camera estimation (comparable or better), multi-view 3D reconstruction, and video depth estimation, where co-training with our large-scale SYN4D consistently improves performance. These consistent improvements demonstrate the effectiveness of our SYN4D and motivate investing in large-scale 4D data creation in future work.

Qualitative results. We further provide a qualitative comparison of in-the-wild 4D reconstruction and tracking in Fig. 4. Training the 4D reconstructor with SYN4D yields more accurate object tracks (*e.g.*, the ball) and more stable distant geometry (*e.g.*, the sky and far mountains), owing to the more complete and diverse annotations in SYN4D. This indicates that SYN4D improves not only

Input 4RC 4RC (Syn4D)

Fig. 4: Qualitative comparison of in-the-wild 4RC [71] tracking trained w/o and w/ Syn4D. Training with Syn4D yields more accurate object tracks (*e.g.*, the ball) and more stable distant geometry (*e.g.*, the sky and far mountains), owing to the more complete and diverse annotations in Syn4D.

Method	**Align**	**Sintel** [13]		**Bonn** [76]		**KITTI** [35]	
		Rel ↓	$\delta < 1.25$ ↑	Rel ↓	$\delta < 1.25$ ↑	Rel ↓	$\delta < 1.25$ ↑
4RC	scale	0.311	62.2	0.051	**97.4**	0.076	95.2
4RC (Syn4D)	scale	**0.211**	**74.6**	**0.048**	97.3	**0.071**	**95.7**
4RC	scale & shift	0.249	67.0	0.048	97.3	0.058	95.5
4RC (Syn4D)	scale & shift	**0.176**	**76.6**	**0.046**	**97.3**	**0.057**	**95.7**

Table 5: Quantitative evaluation for video depth estimation. We report metrics under both scale-only and scale-and-shift alignments. Co-training with Syn4D improves the performance on almost all datasets across all metrics.

the quantitative metrics but also the perceptual quality of reconstruction and tracking on real-world videos.

4.3 Human pose estimation

Each scene in Syn4D also contains a single human with ground-truth SMPL-X [80] body pose and shape annotations. We thus convert these to SMPL [69] via the optimization-based refitting provided by SMPL-X, enabling feed-forward training with SMPL-based methods.

Metrics. We report three standard human mesh recovery metrics: Per-Joint Position Error (MPJPE), Procrustes-Aligned MPJPE (PA-MPJPE), and Per-Vertex Error (PVE) [11], all in millimeters (↓). We evaluate on the test sets of

Method	Hi4D [130]			CHI3D [31]			3DPW [109]		
	MPJPE ↓	PA-MPJPE ↓	PVE ↓	MPJPE ↓	PA-MPJPE ↓	PVE ↓	MPJPE ↓	PA-MPJPE ↓	PVE ↓
MA-HMR	58.8	43.9	73.6	47.2	31.4	55.7	63.2	40.2	73.8
MA-HMR+Cont	58.7	44.1	73.2	46.9	31.4	55.5	63.0	40.0	73.4
MA-HMR+Syn4D	**57.7**	**43.0**	**72.2**	**46.1**	**31.1**	**54.9**	**62.5**	**39.6**	**72.9**

Table 6: Quantitative evaluation for human pose estimation. +Cont denotes continuing training with the original datasets. +Syn4D denotes continuing training with the original datasets plus Syn4D.

3DPW [109], Hi4D [130], and the training set of CHI3D [31], three challenging real-image benchmarks. Note that, none of the models is trained or fine-tuned on the benchmark datasets.

Experiment setup. We adopt MA-HMR [112] as our baseline, a state-of-the-art single-stage multi-person SMPL recovery method. It is trained with AGORA [79], BEDLAM [11], and DTO-Humans [112]. We fine-tune MA-HMR's final checkpoint on the original three datasets plus Syn4D (MA-HMR+Syn4D) for six epochs. To ensure a fair comparison, we also fine-tune on the original three datasets alone (MA-HMR+Cont) for exactly the same number of epochs. We filter out frames in which the human is heavily occluded using the segmentation masks. We use a learning rate of 10^{-5} and a total batch size of 128.

Results. As shown in Tab. 6, MA-HMR+Cont yields similar performance to the baseline, confirming that additional training epochs alone does not account for the gain. Adding Syn4D (MA-HMR+Syn4D) consistently improves all three metrics across all three benchmarks, demonstrating that our synthetic human annotations provide a useful complement to existing training corpora. We attribute the improvement to the diverse occlusion patterns in Syn4D, which are well-suited to MA-HMR's one-stage design.

5 Conclusions

We have introduced Syn4D, a large multiview synthetic 4D dataset of dynamic scenes with accurate and complete geometric annotations, including camera motion, depth, dense tracking, and human pose. Experiments show that Syn4D improves performance across many tasks, including 3D tracking, video depth estimation, camera pose estimation, multi-view reconstruction, and human pose estimation. It also supports new tasks, such as geometry-aware novel view synthesis, which we introduce and explore. These results demonstrate that Syn4D provides effective supervision for learning spatiotemporal scene representations and offers a strong foundation for future research in 4D vision.

Acknowledgements The authors of this work were supported by Clarendon Scholarship, NTU SUG-NAP, NRF-NRFF17-2025-0009, ERC 101001212-UNION, and EPSRC EP/Z001811/1 SYN3D.

References

1. Ankur, H., Viorica, P., Vijay, B., Simon, S., Roberto, C.: SceneNet: understanding real world indoor scenes with synthetic data. In: Proc. CVPR (2015) 5
2. Azinović, D., Martin-Brualla, R., Goldman, D.B., Nießner, M., Thies, J.: Neural RGB-D surface reconstruction. In: CVPR (2022) 13
3. Bahmani, S., Skorokhodov, I., Siarohin, A., Menapace, W., Qian, G., Vasilkovsky, M., Lee, H.Y., Wang, C., Zou, J., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Vd3d: Taming large video diffusion transformers for 3d camera control. Proc. ICLR (2025) 6
4. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., Zhang, D.: ReCamMaster: camera-controlled generative rendering from a single video. In: Proc. ICCV (2025) 6, 11
5. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14834–14844 (2025) 26
6. Bai, J., Xia, M., Wang, X., Yuan, Z., Fu, X., Liu, Z., Hu, H., Wan, P., Zhang, D.: Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. arXiv preprint arXiv:2412.07760 (2024) 6
7. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. IJCV **92**(1) (2011) 5
8. Baradel, F., Armando, M., Galaaoui, S., Brégier, R., Weinzaepfel, P., Rogez, G., Lucas, T.: Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In: European Conference on Computer Vision. pp. 202–218. Springer (2024) 6
9. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., Shulman, E.: ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In: Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021), https://openreview.net/forum?id=tjZjv_qh_CE 5
10. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In: Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV) (2019) 5
11. Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: a synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proc. CVPR (2023) 2, 4, 5, 6, 14, 15
12. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv.cs **abs/2311.15127** (2023) 6
13. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: Proc. ECCV (2012) 4, 13, 14, 29, 30
14. Cabon, Y., Murray, N., Humenberger, M.: Virtual KITTI 2. arXiv **2001.10773** (2020) 4, 5
15. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 11621–11631 (2020) 5

16. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. International Conference on 3D Vision (3DV) (2017) 5
17. Chang, A.X., Funkhouser, T.A., Guibas, L.J., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An information-rich 3D model repository. arXiv.cs **abs/1512.03012** (2015) 3, 4
18. Chen, W., Zhang, G., Wimbauer, F., Wang, R., Araslanov, N., Vedaldi, A., Cremers, D.: Back on track: Bundle adjustment for dynamic scene reconstruction. In: Proceedings of the International Conference on Computer Vision (ICCV) (2025) 11
19. Chen, Y., Zheng, C., Xu, H., Zhuang, B., Vedaldi, A., Cham, T.J., Cai, J.: MVSplat360: Benchmarking 360 generalizable 3D novel view synthesis from sparse views. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2024) 6
20. Chen, Z., Liu, T., Zhuo, L., Ren, J., Tao, Z., Zhu, H., Hong, F., Pan, L., Liu, Z.: 4dnex: Feed-forward 4d generative modeling made easy. arXiv preprint arXiv:2508.13154 (2025) 27
21. Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., Guillaumin, M., Malik, J.: ABO: Dataset and benchmarks for real-world 3D object understanding. In: Proc. CVPR. pp. 21126–21136 (2022) 5
22. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: Proc. CVPR (2017) 5, 13
23. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., VanderBilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A.: Objaverse-XL: A universe of 10M+ 3D objects. CoRR **abs/2307.05663** (2023) 2, 3, 5
24. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3D objects. In: Proc. CVPR (2023) 2, 3, 5
25. Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Ehsani, K., Salvador, J., Han, W., Kolve, E., Kembhavi, A., Mottaghi, R.: ProcTHOR: large-scale embodied AI using procedural generation. In: Proc. NeurIPS (2022) 4
26. Denninger, M., Winkelbauer, D., Sundermeyer, M., Boerdijk, W., Knauer, M., Strobl, K.H., Humt, M., Triebel, R.: BlenderProc2: A procedural pipeline for photorealistic rendering. Journal of Open Source Software **8**(82), 4901 (2023) 4
27. Dosovitskiy, A., Ros, G., Codevilla, F., López, A.M., Koltun, V.: CARLA: an open urban driving simulator. In: Conference on Robot Learning (2017) 5
28. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google Scanned Objects: A high-quality dataset of 3D scanned household items. In: Proc. ICRA (2022) 4, 5
29. Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., Cucchiara, R.: Learning to detect and track visible and occluded body joints in a virtual world. In: Proc. ECCV (2018) 4, 5
30. Feng, H., Zhang, J., Wang, Q., Ye, Y., Yu, P., Black, M.J., Darrell, T., Kanazawa, A.: St4RTrack: simultaneous 4D reconstruction and tracking in the world. In: Proc. ICCV (2025) 2, 6
31. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Three-dimensional reconstruction of human interactions. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7214–7223 (2020) 15, 29
32. Fu, H., Cai, B., Gao, L., Zhang, L.X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., et al.: 3d-front: 3d furnished rooms with layouts and semantics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10933–10942 (2021) 4
33. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual KITTI. In: Proc. CVPR (2016) 4, 5
34. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: CAT3D: create anything in 3d with multi-view diffusion models. arXiv **2405.10314** (2024) 6
35. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. International Journal of Robotics Research (IJRR) (2013) 5, 13, 14
36. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. arXiv.cs **abs/2305.20091** (2023) 6
37. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., Kipf, T., Kundu, A., Lagun, D., Laradji, I., Liu, H.T.D., Meyer, H., Miao, Y., Nowrouzezahrai, D., Oztireli, C., Pot, E., Radwan, N., Rebain, D., Sabour, S., Sajjadi, M.S.M., Sela, M., Sitzmann, V., Stone, A., Sun, D., Vora, S., Wang, Z., Wu, T., Yi, K.M., Zhong, F., Tagliasacchi, A.: Kubric: a scalable dataset generator. In: Proc. CVPR (2022) 2, 4, 5, 6, 8, 28
38. Han, J., An, H., Jung, J., Narihira, T., Seo, J., Fukuda, K., Kim, C., Hong, S., Mitsufuji, Y., Kim, S.: D^{2}USt3R: Enhancing 3D reconstruction with 4d pointmaps for dynamic scenes. arXiv **2504.06264** (2025) 2
39. Handa, A., Whelan, T., McDonald, J., Davison, A.: A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In: IEEE Intl. Conf. on Robotics and Automation, ICRA. Hong Kong, China (May 2014) 5
40. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: International Conference on Computer Vision (ICCV). pp. 2282–2292 (2019) 6
41. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation (2024) 6
42. Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., Shan, Y.: DepthCrafter: generating consistent long depth sequences for open-world videos. In: Proc. CVPR (2025) 6
43. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: DeepMVS: Learning multi-view stereopsis. In: Proc. CVPR (2018) 28
44. Hunyuan3D, T., Yang, S., Yang, M., Feng, Y., Huang, X., Zhang, S., He, Z., Luo, D., Liu, H., Zhao, Y., Lin, Q., Lai, Z., Yang, X., Shi, H., Zhao, Z., Zhang, B., Yan, H., Wang, L., Liu, S., Zhang, J., Chen, M., Dong, L., Jia, Y., Cai, Y., Yu, J., Tang, Y., Guo, D., Yu, J., Zhang, H., Ye, Z., He, P., Wu, R., Wei, S., Zhang, C., Tan, Y., Sun, Y., Niu, L., Huang, S., Zheng, B., Liu, S., Chen, S., Yuan, X., Yang, X., Liu, K., Zhu, J., Chen, P., Liu, T., Wang, D., Liu, Y., Linus, Jiang, J., Huang, J., Guo, C.: Hunyuan3D 2.1: From images to high-fidelity 3D assets with production-ready PBR material. arXiv **2506.15442** (2025) 4
45. Hurl, B., Czarnecki, K., Waslander, S.: Precise synthetic image and LiDAR (PreSIL) dataset for autonomous vehicle perception. arXiv (2019) 5

46. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5579–5588 (2020) 6
47. Jiang, Z., Zheng, C., Laina, I., Larlus, D., Vedaldi, A.: Geo4D: Leveraging video generators for geometric 4D scene reconstruction. In: Proceedings of the International Conference on Computer Vision (ICCV) (2025) 6
48. Jiang, Z., Zheng, C., Laina, I., Larlus, D., Vedaldi, A.: Mesh4D: 4D mesh reconstruction and tracking from monocular video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2026) 4, 6
49. Jin, L., Tucker, R., Li, Z., Fouhey, D., Snavely, N., Holynski, A.: Stereo4D: Learning how things move in 3D from internet stereo videos. arXiv **2412.09621** (2024) 2, 6
50. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proc. ICCV (2015) 6
51. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social interaction capture. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(1), 190–204 (2019) 12, 13
52. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B.C., Matthews, I.A., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social interaction capture. PAMI **41**(1) (2019) 6
53. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proc. CVPR (2018) 6
54. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: CoTracker: It is better to track together. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024) 2
55. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Dynamicstereo: Consistent dynamic depth from stereo videos. CVPR (2023) 4, 28
56. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: DynamicStereo: Consistent dynamic depth from stereo videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 5
57. Karhade, J., Keetha, N.V., Gupta, T., Zhang, Y., Sharma, A., Scherer, S., Ramanan, D.: Any4D: Towards unified feed-forward metric 4d reconstruction. In: Proc. ICCV Workshop (2025) 6
58. Kästingschäfer, M., Gieruc, T., Bernhard, S., Campbell, D., Insafutdinov, E., Brox, T.: SEED4D: a synthetic ego–exo dynamic 4D data generator, driving dataset and benchmark. In: Proc. WACV (2025) 4, 5
59. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S.R., Richardt, C., Ramanan, D., Scherer, S., Kontschieder, P.: MapAnything: universal feed-forward metric 3D reconstruction. In: Proc. 3DV (2026) 2, 6
60. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: ICCV (2019) 6

61. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., Kembhavi, A., Gupta, A., Farhadi, A.: AI2-THOR: An interactive 3D environment for visual ai. arXiv (2017) 4
62. Koppula, S., Rocco, I., Yang, Y., Heyward, J., Carreira, J., Zisserman, A., Brostow, G., Doersch, C.: TAPVid-3D: a benchmark for tracking any point in 3D. arXiv **2407.05921** (2024) 12
63. Lan, Y., Luo, Y., Hong, F., Zhou, S., Chen, H., Lyu, Z., Yang, S., Dai, B., Loy, C.C., Pan, X.: STream3R: Scalable sequential 3D reconstruction with causal transformer. In: ICLR (2026) 12
64. Liao, Y., Xie, J., Geiger, A.: Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **45**(3), 3292–3310 (2022) 5
65. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv **2511.10647** (2025) 2, 6, 31
66. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., Li, X., Sun, X., Ashok, R., Mukherjee, A., Kang, H., Kong, X., Hua, G., Zhang, T., Benes, B., Bera, A.: DL3DV-10K: a large-scale scene dataset for deep learning-based 3d vision. In: Proc. CVPR (2024) 28
67. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3D object. In: Proc. ICCV (2023) 4, 6
68. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: Proc. ICLR (2023) 27
69. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015) 6, 14
70. Luo, Y., Zhou, S., Lan, Y., Pan, X., Loy, C.C.: 4RC: 4D reconstruction via conditional querying anytime and anywhere. arXiv preprint arXiv:2602.10094 (2026) 2, 13, 28, 29, 30
71. Luo, Y., Zhou, S., Lan, Y., Pan, X., Loy, C.C.: 4RC: 4D reconstruction via conditional querying anytime and anywhere. In: Proc. ICML (2026) 6, 12, 14
72. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: Proc. ECCV (2018) 6
73. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proc. CVPR (2016) 4
74. McCormac, J., Handa, A., Leutenegger, S., Davison, A.J.: SceneNet RGB-D: can 5M synthetic images beat generic ImageNet pre-training on indoor segmentation? In: Proc. ICCV (2017) 4, 5
75. Mehl, L., Schmalfuss, J., Jahedi, A., Nalivayko, Y., Bruhn, A.: Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In: Proc. CVPR (2023) 4
76. Palazzolo, E., Behley, J., Lottes, P., Giguère, P., Stachniss, C.: ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. arXiv (2019) 13, 14
77. Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O., Newcombe, R., Ren, C.Y.: Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In: ICCV (2023) 12, 13
78. Patel, P., Black, M.J.: Camerahmr: Aligning people with perspective. In: 2025 International Conference on 3D Vision (3DV). pp. 1562–1571. IEEE (2025) 6

79. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: Agora: Avatars in geography optimized for regression analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13468–13478 (2021) 6, 15
80. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proc. CVPR (2019) 6, 14
81. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proc. ICML. vol. 139 (2021) 11
82. Raistrick, A., Lipson, L., Ma, Z., Mei, L., Wang, M., Zuo, Y., Kayan, K., Wen, H., Han, B., Wang, Y., Newell, A., Law, H., Goyal, A., Yang, K., Deng, J.: Infinite photorealistic worlds using procedural generation. In: Proc. CVPR (2023) 4
83. Raistrick, A., Mei, L., Kayan, K., Yan, D., Zuo, Y., Han, B., Wen, H., Parakh, M., Alexandropoulos, S., Lipson, L., Ma, Z., Deng, J.: Infinigen indoors: Photorealistic indoor scenes using procedural generation. In: Proc. CVPR (2024) 4
84. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for benchmarks: Structured dataset generation by play. In: Proc. ICCV (2017) 4
85. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proc. ICCV (2021) 4
86. Ros, G., Sellart, L., Materzynska, J., Vázquez, D., López, A.M.: The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proc. CVPR (2016) 4, 5
87. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: Proc. ICCV (2019) 5
88. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 3260–3269 (2017) 5
89. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in RGB-D images. In: CVPR (June 2013) 13
90. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 567–576 (2015) 4
91. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C.Y., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H.M., De Nardi, R., Goesele, M., Lovegrove, S., Newcombe, R.A.: The Replica dataset: A digital replica of indoor spaces. arXiv (2019) 4
92. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) pp. 573–580 (2012) 5, 13
93. Su, C., Ma, X., Su, J., Wang, Y.: Sat-hmr: Real-time multi-person 3d mesh estimation via scale-adaptive tokens. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16796–16806 (2025) 6

94. Sucar, E., Insafutdinov, E., Lai, Z., Vedaldi, A.: V-DPM: 4D video reconstruction with dynamic point maps. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2026) 2, 6
95. Sucar, E., Insafutdinov, E., Lai, Z., Vedaldi, A.: V-DPM: Video reconstruction with dynamic point maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2026) 7
96. Sucar, E., Lai, Z., Insafutdinov, E., Vedaldi, A.: Dynamic Point Maps: A versatile representation for dynamic 3D reconstruction. In: Proceedings of the International Conference on Computer Vision (ICCV) (2025) 2, 6, 7
97. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proc. CVPR (2020) 5, 28
98. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11179–11188 (2021) 6
99. Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: Putting people in their place: Monocular regression of 3d people in depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13243–13252 (2022) 6
100. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., Batra, D.: Habitat 2.0: Training home assistants to rearrange their habitat. In: Proc. NeurIPS (2021) 5
101. Szymanowicz, S., Zhang, J.Y., Srinivasan, P., Gao, R., Brussee, A., Holynski, A., Martin-Brualla, R., Barron, J.T., Henzler, P.: Bolt3D: Generating 3D scenes in seconds. arXiv **2503.14445** (2025) 6
102. Tesch, J., Becherini, G., Achar, P., Yiannakidis, A., Kocabas, M., Patel, P., Black, M.J.: BEDLAM2.0: Synthetic humans and cameras in motion. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025) 2, 4, 5
103. Torresani, L., Hertzmann, A., Bregler, C.: Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. PAMI **30**(5) (2008) 2
104. Tremblay, J., To, T., Birchfield, S.: Falling things: A synthetic dataset for 3d object detection and pose estimation. In: Proc. ICRA (2018) 4
105. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges (2019), https://arxiv.org/abs/1812.01717 11
106. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: European Conference on Computer Vision (ECCV). pp. 313–331. Springer (2024) 6
107. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proc. CVPR (2017) 4, 5
108. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: European Conference on Computer Vision (ECCV). pp. 439–457. Springer (2024) 6

109. Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European conference on computer vision (ECCV). pp. 601–617 (2018) 15, 28
110. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025) 2, 6
111. Wang, J., Chen, M., Zhang, S., Karaev, N., Schönberger, J., Labatut, P., Bojanowski, P., Novotny, D., Vedaldi, A., Rupprecht, C.: VGGT-Ω. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2026) 2
112. Wang, K., Zheng, K., Shi, Y., Guo, C., Wu, J.: Towards metric-aware multi-person mesh recovery by jointly optimizing human crowd in camera space. arXiv preprint arXiv:2511.13282 (2025) 6, 15, 28
113. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3D perception model with persistent state. In: Proc. CVPR (2025) 6, 12
114. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D vision made easy. In: Proc. CVPR (2024) 2, 6, 12
115. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: TartanAir: a dataset to push the limits of visual SLAM. In: Proc.IROS (2020) 4
116. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv **2507.13347** (2025) 2, 6
117. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: Prompthmr: Promptable human mesh recovery. In: Proceedings of the computer vision and pattern recognition conference. pp. 1148–1159 (2025) 6
118. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., Ramanan, D., Carr, P., Hays, J.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021) (2021) 5
119. Wrenninge, M., Unger, J.: Synscapes: A photorealistic synthetic dataset for street scene parsing. arXiv **1810.08705** (2018) 4, 5
120. Wu, T., Zheng, C., Guan, F., Vedaldi, A., Cham, T.J.: Amodal3R: Amodal 3D reconstruction from occluded 2D images. In: Proceedings of the International Conference on Computer Vision (ICCV) (2025) 4
121. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., Lin, D., Liu, Z.: OmniObject3D: Large-vocabulary 3D object dataset for realistic perception, reconstruction and generation. In: Proc. CVPR. pp. 803–814 (2023) 3, 5
122. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: Proc. CVPR (2018) 5
123. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3D latents for scalable and versatile 3d generation. In: Proc. CVPR (2025) 4
124. Xiao, Y., Wang, Q., Zhang, S., Xue, N., Peng, S., Shen, Y., Zhou, X.: Spatialtracker: Tracking any 2d pixels in 3d space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20406–20417 (2024) 6

125. Xie, M., Khan, N., Wang, T., Dhingra, N., Nam, S., Yang, H., Hui, Z., Metzler, C., Vedaldi, A., Pirsiavash, H., Luo, L.: LaVR: Latent space conditioned video re-rendering using large 4D reconstruction models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2026) 6
126. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: BlendedMVS: a large-scale dataset for generalized multi-view stereo networks. In: Proc. CVPR (2020) 2, 4, 5
127. Yenphraphai, J., Mirzaei, A., Chen, J., Zou, J., Tulyakov, S., Yeh, R.A., Wonka, P., Wang, C.: Shapegen4d: Towards high quality 4d shape generation from videos. arXiv preprint (2025) 6
128. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: ScanNet++: a high-fidelity dataset of 3d indoor scenes. In: IEEE International Conference on Computer Vision (ICCV) (2023) 5, 28
129. Yin, W., Cai, Z., Wang, R., Zeng, A., Wei, C., Sun, Q., Mei, H., Wang, Y., Pang, H.E., Zhang, M., et al.: Smplest-x: Ultimate scaling for expressive human pose and shape estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 6
130. Yin, Y., Guo, C., Kaufmann, M., Zarate, J.J., Song, J., Hilliges, O.: Hi4d: 4d instance segmentation of close human interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17016–17027 (2023) 15, 28
131. Yu, M., Hu, W., Xing, J., Shan, Y.: TrajectoryCrafter: redirecting camera trajectory for monocular videos via diffusion models. In: Proc. ICCV (2025) 6
132. Yuan, L., Wang, J., Sun, H., Zhang, Y., Lin, Y.: Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding (2025), https://arxiv.org/abs/2501.07888 10
133. Zhang, B., Ke, L., Harley, A.W., Fragkiadaki, K.: TAPIP3D: tracking any point in persistent 3D geometry. arXiv **2504.14717** (2025) 6
134. Zhang, B., Xu, S., Wang, C., Yang, J., Zhao, F., Chen, D., Guo, B.: Gaussian variation field diffusion for high-fidelity video-to-4d synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12502–12513 (October 2025) 6
135. Zhang, C., Le Moing, G., Koppula, S., Rocco, I., Momeni, L., Xie, J., Sun, S., Sukthankar, R., Barral, J.K., Hadsell, R., et al.: Efficiently reconstructing dynamic scenes one d4rt at a time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7382–7392 (2026) 2
136. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: CLAY: A controllable large-scale generative model for creating high-quality 3D assets. In: Proc. SIGGRAPH (2024) 4
137. Zhang, S., Ge, Y., Tian, J., Xu, G., Chen, H., Lv, C., Shen, C.: POMATO: marrying pointmap matching with temporal motion for dynamic 3D reconstruction. arXiv **2504.05692** (2025) 2
138. Zheng, C., Vedaldi, A.: Free3D: Consistent novel view synthesis without 3D representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 6
139. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: European Conference on Computer Vision (ECCV). pp. 519–535. Springer (2020) 4
140. Zheng, Y., Harley, A.W., Shen, B., Wetzstein, G., Guibas, L.J.: PointOdyssey: A large-scale synthetic dataset for long-term point tracking. In: Proc. CVPR (2023) 2, 3, 4, 5, 8, 28

141. Zhou, J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12405–12414 (2025) 6
142. Zhu, R., Lu, J., Hu, W., Han, X., Cai, J., Shan, Y., Zheng, C.: Motioncrafter: Dense geometry and motion reconstruction with a 4d vae. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026) 2
143. Zhu, R., Lu, J., Hu, W., Han, X., Cai, J., Shan, Y., Zheng, C.: Motioncrafter: Dense geometry and motion reconstruction with a 4d vae. In: CVPR (2026) 6