Any4D: Unified Feed-Forward Metric 4D Reconstruction

Supplementary Material

Table S.1. **List of Datasets used to train `Any4D`**

Dataset	Dynamic	Scene Flow	Domain	# Scenes
BlendedMVS	✗	✗	Outdoor & object centric	500
MegaDepth	✗	✗	Outdoor	275
ScanNet++	✗	✗	Indoor	295
VKITTI2	✓	✗	AV	40
Waymo-DriveTrack	✓	✓	AV	1500
GCD-Kubric	✓	✗	Synthetic random objects	5000
CoTracker3-Kubric	✓	✓	Synthetic random objects	5000
Dynamic Replica	✓	✓	Synthetic humans & animals	500
Point Odyssey	✓	✓	Diverse Synthetic assets	159

A. Training

Datasets: We train on a combination of static and dynamic datasets with varying levels of supervision. For supervision geometric quantities - depth, intrinsics, and camra poses, all the datasets in S.1 are used. For scene flow supervision, we only rely on Kubric (from CoTracker3), PointOdyssey and Dynamic Replica, as they contain both diverse camera and scene motion crucial for learning good scene flow. We find that VKITTI-2 sequences span minimal scene motion while data from GCD lacks good camera diversity, and thus, only use them for geometry supervision.

Implementation Details: We initialize `Any4D`'s weights with the public MapAnything checkpoint. The doppler scene-flow encoder, and the scene-flow DPT decoder are initialized and learnt from scratch. We train the entire network with a learning rate of 1e-5, 5e-7 and 1e-4 for the entire network, the DINOv2 Image encoder and the Scene-flow DPT decoder respectively. We use a warmup of 10 epochs, and finetune the network for a total of 100 epochs, covering approximately 120k gradient steps in total on 8 H100 GPUs. The images and respective quantities in each batch cropped and resized to 518 image width, with a randomized height-width aspect ratio between 0.5 and 3. During each gradient step, we sample upto 4 views from each dataset, with a variable batch size of upto 24 views per GPU. As illustrated in Fig. S.1, we find that 4-view training is critical for generalizing with multi-view inference.

B. Benchmarking Setup Details

For the TAPVID-3D PStudio dataset and DriveTrack datasets, we evaluated on a uniform subset of 50 sequences from all available datasets and use the first 64 frames for evaluation. Since the dataset is extremely sparse and each sequence only contains at most a few hundred point queries, we use all points for benchmarking. For Drive-Track, we filter 50 sequences that contain non-zero allocentric motion. For the Dynamic-Replica and LSF-Odyssey datasets, we filter out static points (i.e., points with zero allocentric motion) and use dynamic points as queries for our benchmarking, to maintain homogeneity with the 2 other datasets and emphasize benchmarking of dynamic elements of a scene. We acknowledge that our evaluation is similar to [16].

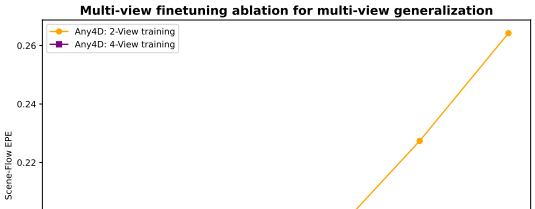

Figure S.1. **4-View training is key to enabling multi-frame generalization during inference**. Any4D trained with 2 views results in higher EPE at higher number of input views. In contrast, the 4-view model exhibits stable behaviour even at 64 views.

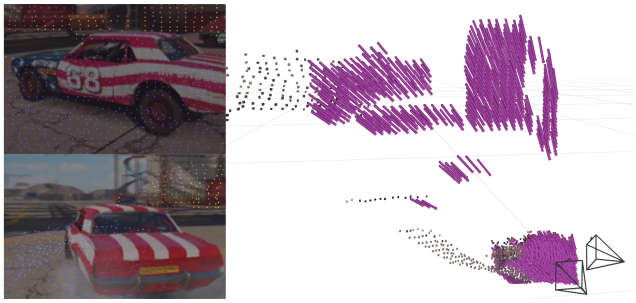

Figure S.2. **Doppler Scene Flow** is simulated as radial component of ego-centric scene flow.

C. Multi-Modal Conditioning

Simulating Doppler Velocity: As shown in Fig. S.2, we simulate the Doppler velocity from egocentric scene flow labels. More specifically, given a 3D point $\vec{p} = [x, y, z]$ and its corresponding ego scene flow vector $\vec{v} = [\Delta x, \Delta y, \Delta z]$, the simulated Doppler velocity v_r is defined as the projection of the motion vector into the radial direction of each ray. This is simply the normalized vector from the origin of the radar to the point $\vec{p}$. The Doppler (radial) velocity is computed as:

$$v_r = \frac{\vec{p} \cdot \vec{v}}{\|\vec{p}\|} = \frac{x \cdot \Delta x + y \cdot \Delta y + z \cdot \Delta z}{\sqrt{x^2 + y^2 + z^2}}$$

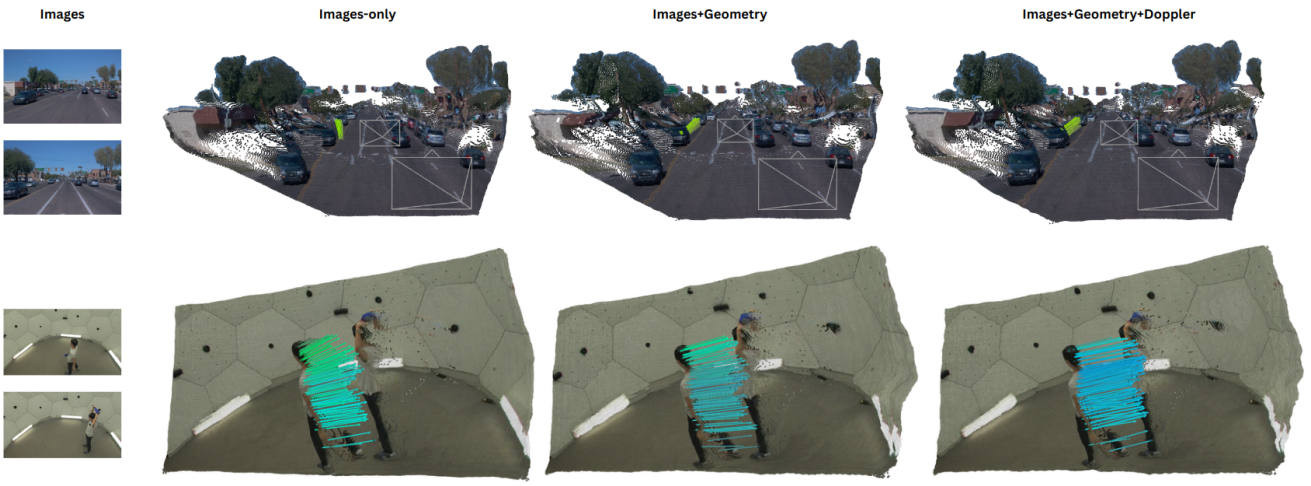

Figure S.3. **Qualitative visualizations of `Any4D` estimating 3D geometry and point tracking on TAPVID-3D Waymo Drive-Track sequences.** As visible, the image-only variant (`column 1`) sometimes produces an offset to the scene flow at the edges. However, the predictions improve whenever sparse geometry (`column 2`) and doppler annotations are available (`column 3`).

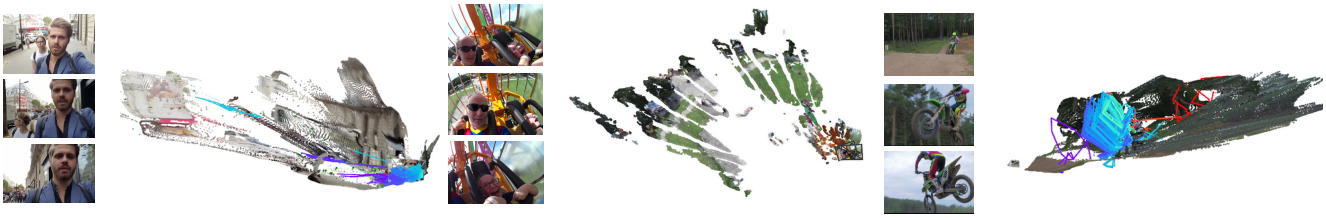

Figure S.4. **Qualitative visualizations of `Any4D` limitations.** Videos with large camera motion inducing no visual overlap of background or scene motion dominating the image space are common failure modes for Any4D. We believe that the availability of large-scale dense scene flow and 3D tracking datasets and integrating real-time optimization is key to overcoming these limitations.

D. Obtaining Long-Range 3D Tracks

Any4D naturally supports long-range 3D point tracking within a single forward pass, without any temporal chaining or post-hoc optimization. Given an N-frame sequence $\{I_1, \ldots, I_N\}$, Any4D predicts: (i) dense per-frame geometry $\{G_i\}_{i=1}^{N} \in \mathbb{R}^{H\times W\times 3}$, relative to a reference image frame I_{ref}, and (ii) allocentric scene flow from the reference frame to all other frames, $\{M_{\text{ref}\to i}\}_{i=1}^{N}$, where $M_{\text{ref}\to i} \in \mathbb{R}^{H\times W\times 3}$ denotes the 3D displacement of each reference-frame pixel to frame i. Then, the dense long-range 3D tracks of points in reference frame $T_{3D} \in \mathbb{R}^{N\times H\times W\times 3}$ can be trivially computed:

$$T_{3D} = \{\, G_{\text{ref}} + M_{\text{ref}\to i} \,\}_{i=1}^{N},$$

Table S.2. **Camera Pose Benchmarking of multi-view feed-forward methods on Sintel.**

Method	ATE ↓	RPE trans ↓	RPE rot ↓
VGGT	0.305	0.1276	**0.44**
MapAnything	0.236	0.095	2.48
Any4D	**0.176**	**0.082**	1.22

Acknowledgments

We thank Tarasha Khurana and Neehar Peri for their initial discussions in the project. We appreciate the help from Jeff Tan with setting up Stereo4D (which we ended up not using due to poor dataset quality). Lastly, we thank Bardienus Duisterhof and members of the AirLab & Deva's Lab at CMU for insightful discussions and feedback on the paper.

This work was supported by Defense Science and Technology Agency contract `#DST000EC124000205`, Bosch Research, and the IARPA via Department of Interior/Interior Business Center (DOI/IBC) contract `140D0423C0074`. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government. Lastly, this work was supported by a hardware grant from Nvidia and used PSC Bridges-2 through allocation `cis220039p` from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program.