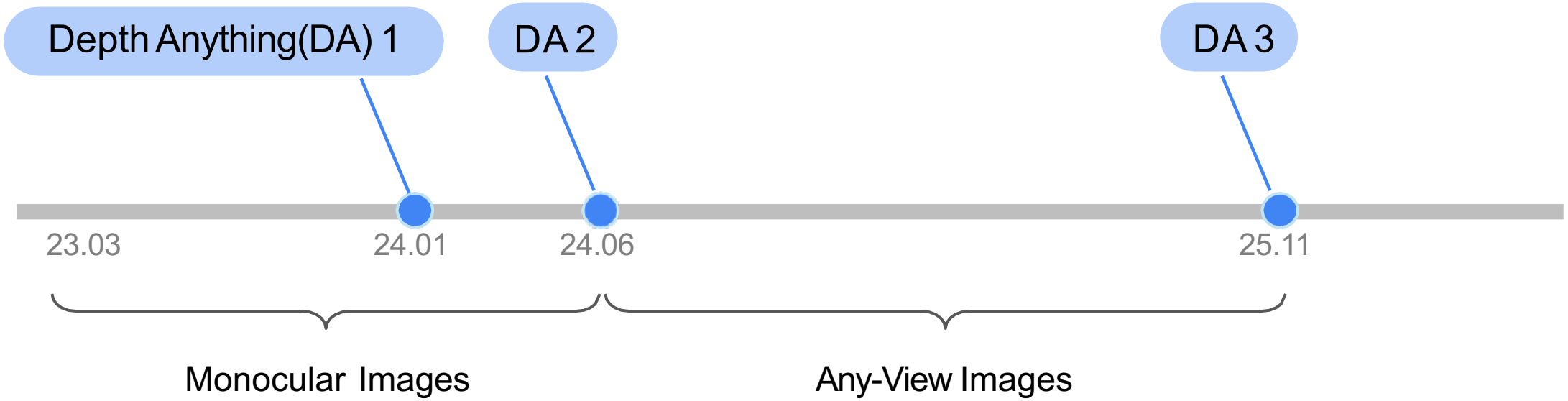
Depth Anything 3:
Recovering the Visual Space from Any Views

Haotong Lin*, Sili Chen*, Jun Hao Liew*, Donny Y. Chen*,
Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo,
Xiaowei Zhou, Guang Shi, Jiashi Feng, **Bingyi Kang*†**

Background

The DA Journey: From Monocular to Any-View

- Depth Anything (DA) 1: **data-driven**, strong generalization
- DA2: **synthetic-to-real** teacher-student learning
- **DA3: Recovering the full 3D visual space from ANY views**

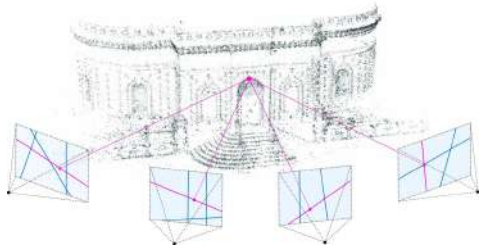
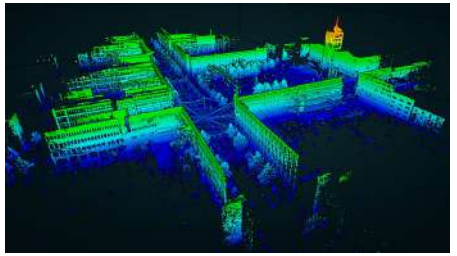


But why did it take so long from DA2 to DA3?
When we started, 3D vision was **fragmented**.

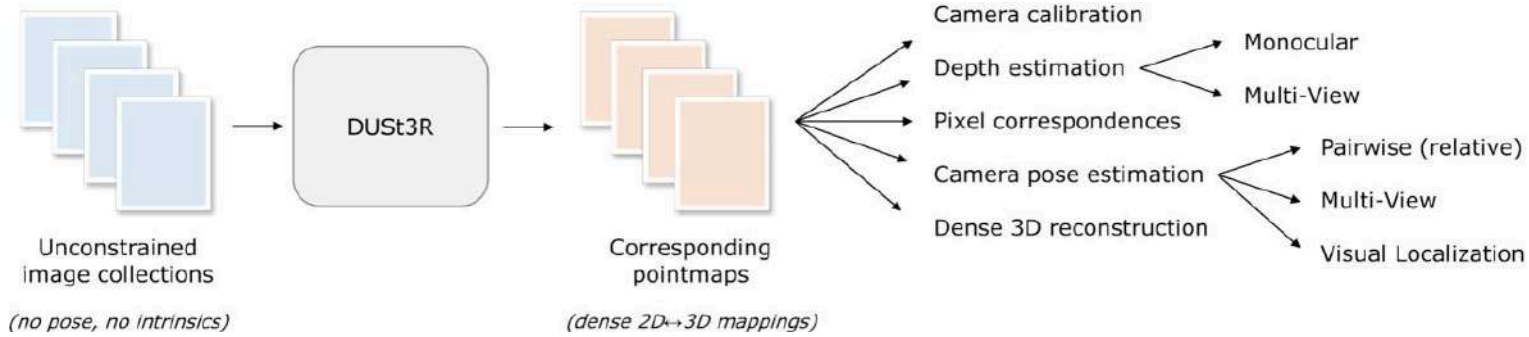
Background

The Problem: Fragmented 3D Vision

- Traditionally: **Separate specialized models** for every 3D task

- Recent unified methods: **bespoke** architectures, **redundant** outputs, **cannot leverage pretrained 2D** foundation models



Can we do it with **one simple model**, **one minimal representation**, built on **2D foundation models**?

- Image sources: Depth Anything 1; structure-from-motion-manish-joshi; laserscanning-europe.com/en/what-slam; DUSt3R

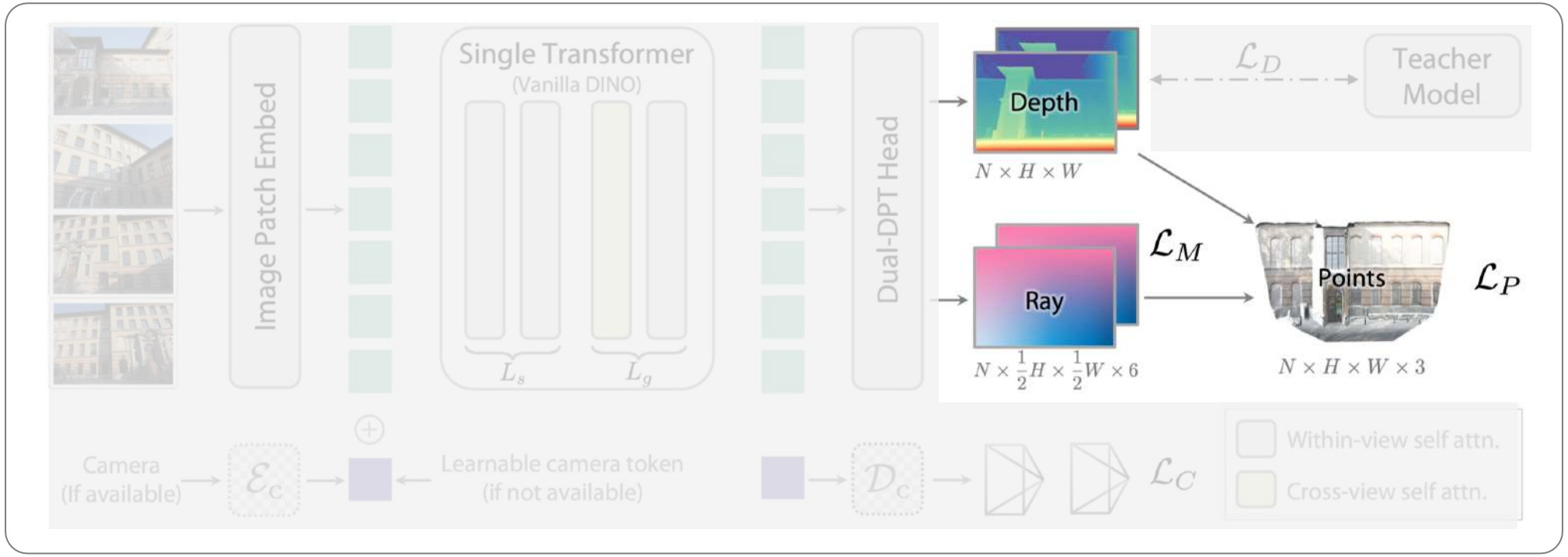
Methodology

DA3 Solution: Minimal Modelling Strategy

- **Q1:** What are the key **3D representations**?
- **Q2:** Is bespoke **architecture** necessary for 3D vision?

Methodology

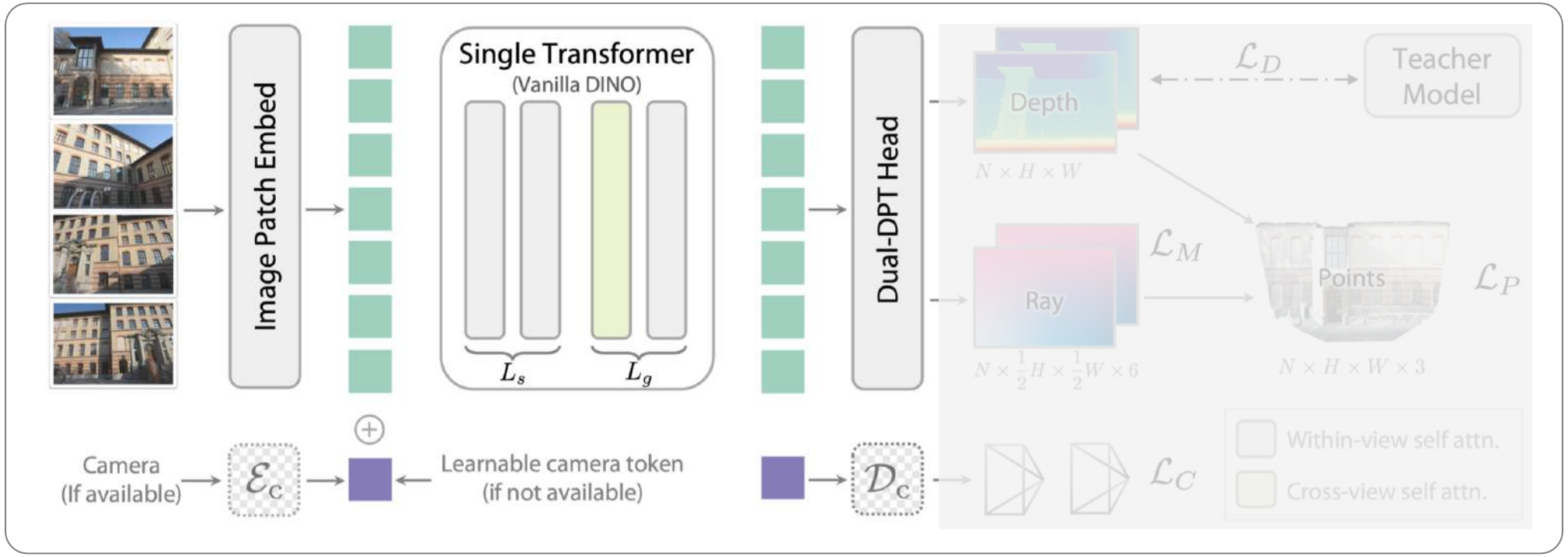
Insight I: You Only Need Depth + Rays in 3D Vision



- You only need **depth and rays**, everything else in 3D vision can be derived
  - E.g., Points $P = t + D(u, v) \times d$
  - **No** need to predict them **separately**.
- Ablation: DA3 vs. point maps or camera-only alternatives, **~2× pose accuracy** (Table 6)

Methodology

Insight II: A Single Transformer Can Handle 3D Vision

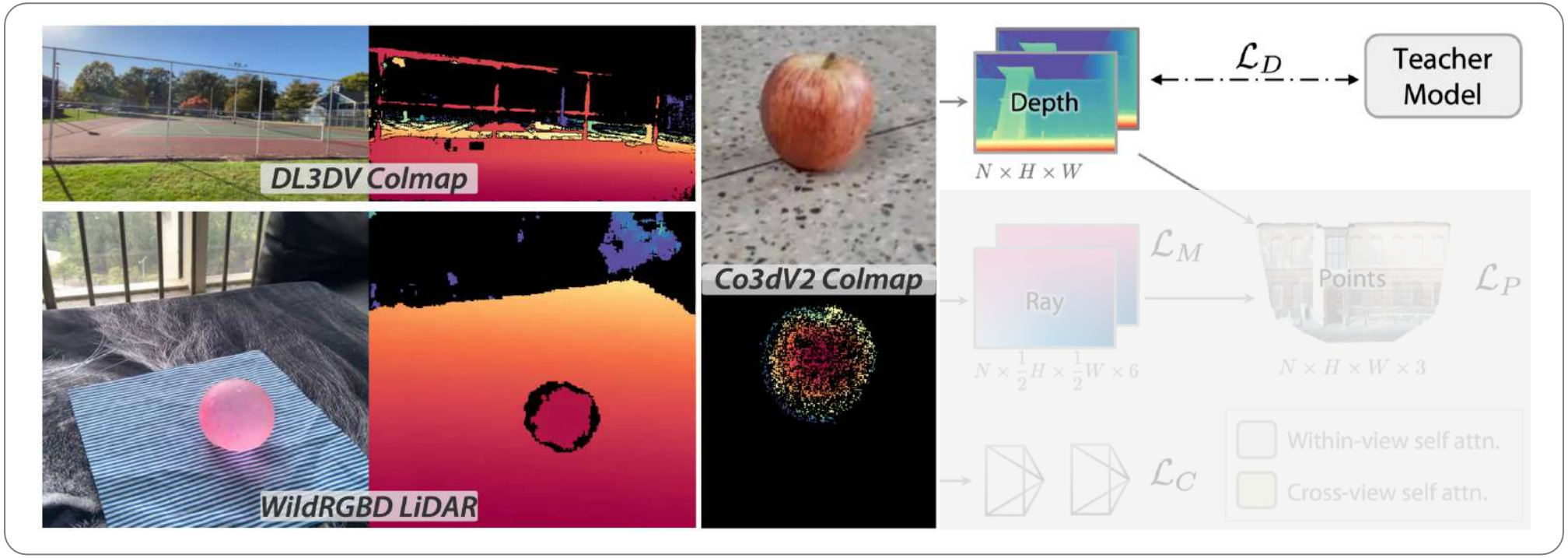


- **Vanilla DINO + Dual-DPT Head**: no bespoke architecture needed
- Purely through **token reordering,** no new modules
- Directly benefits **from pretrained 2D foundation models** like DINOv2

- Ablation: DA3 vs. VGGT-style architecture, ∼**10**× **pose accuracy** (Table 7)

Methodology

Teacher-Student Training: Tackling 3D Data Challenges

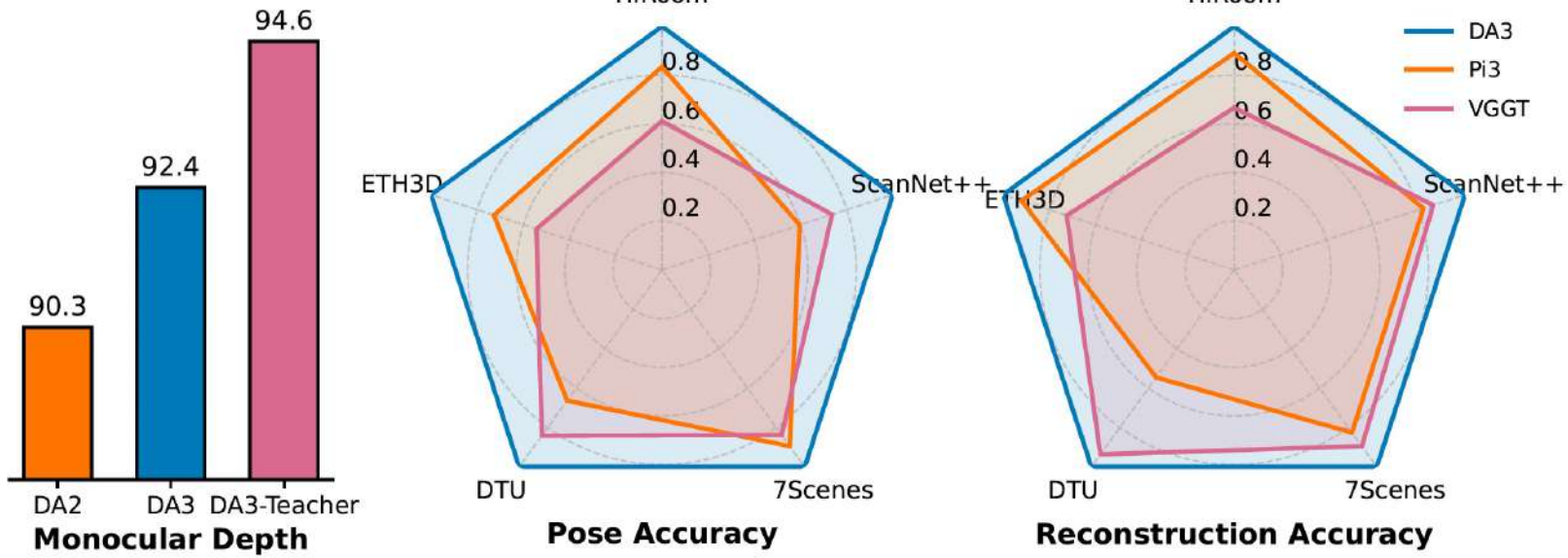


- Real-world data is essential, but labels are often **noisy and sparse**
- Teacher generates dense **relative** pseudo-depth, aligned to GT via **scale-shift fitting**

- **Best of both:** teacher's detail + GT's geometric accuracy

Experiments

Results: State-of-the-Art Across All 3D Tasks



+35.7% pose accuracy | **+23.6%** geometry accuracy over prior SOTA (VGGT)

Simple yet effective!

A very strong backbone for downstream tasks

Applications

Application I: Feed-Forward 3DGS

- Feed-forward
- Pose-free
- In-the-wild

- **GS-DPT head** + small dataset fine-tuning → high-quality feed-forward 3DGS rendering
- Take-away: better geometry → better novel view synthesis

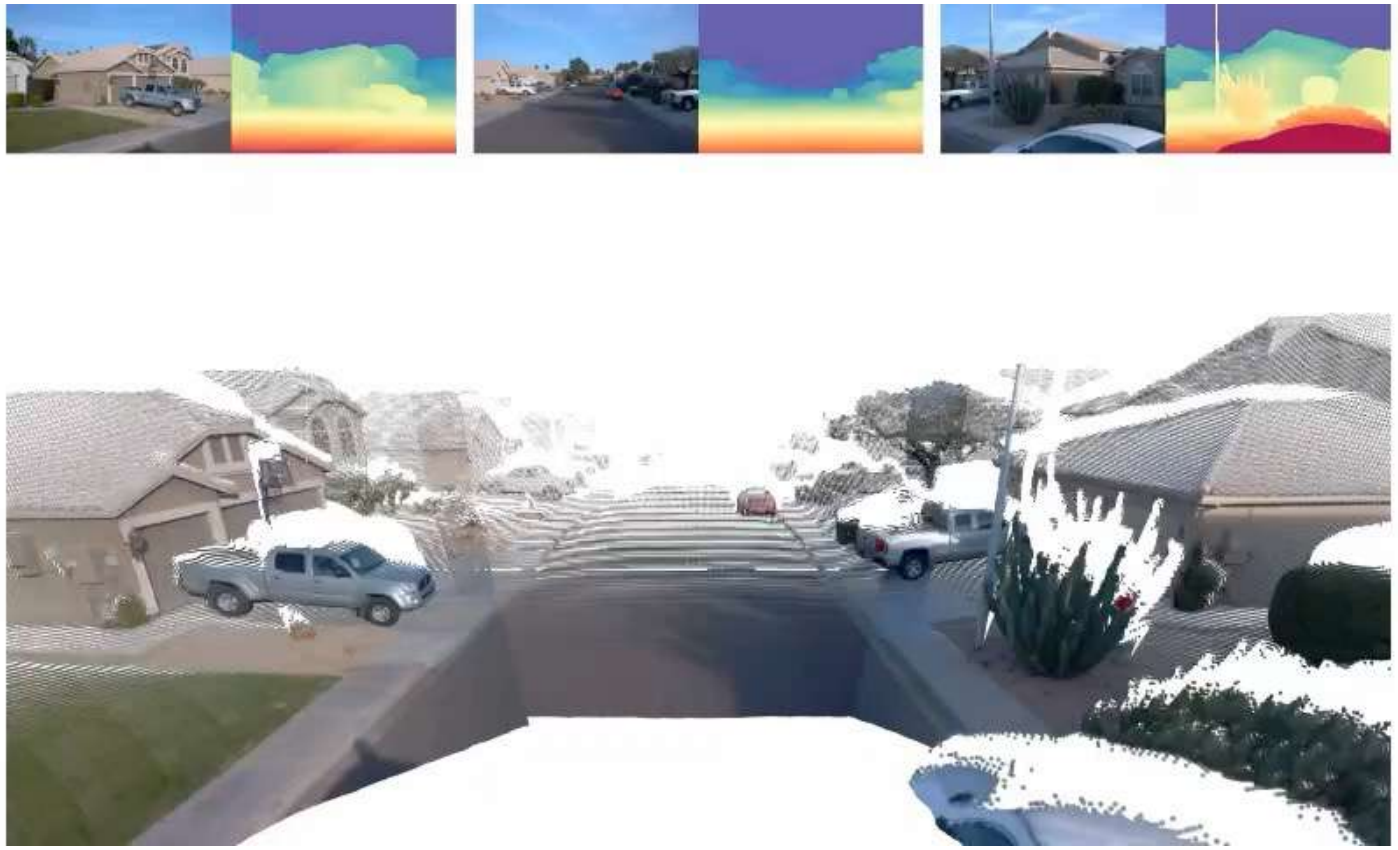
Applications

Application II: SLAM for Large-Scale Scenes

- **No modification** to DA3: Chunk-based processing + overlapping alignment + loop closure
- 3D reconstruction to **kilometer-scale**, **unbounded outdoor** environments

Applications

Application III: Spatial Perception for Autonomous Driving



- Natively supports **outward-facing** multi-viewpoint inputs, works even **without overlap**
- Tuned on massive and diverse multi-view data, thanks to teacher-student training

Discussion

Community Impact

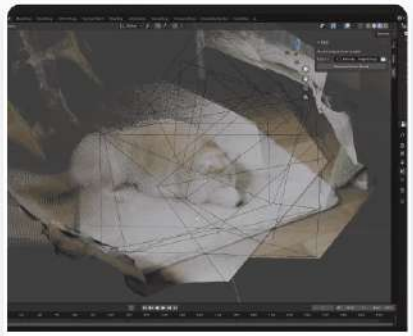
Open-source: code, model, and benchmarks

Model variants: monocular, multi-view, 3DGS

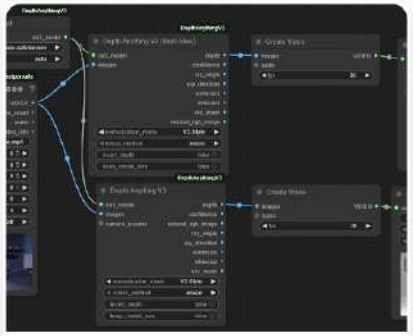
Model sizes: small, base, large, giant

All models are trained exclusively on public academic datasets.

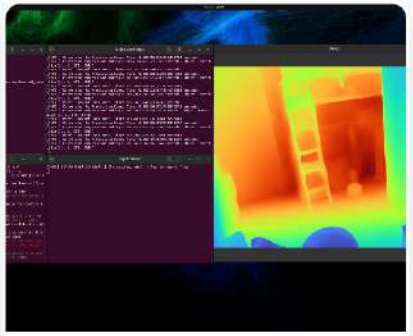
Community integration:



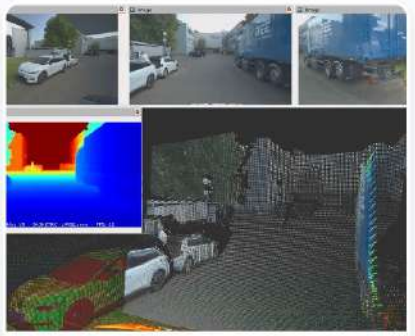
DA3-blender



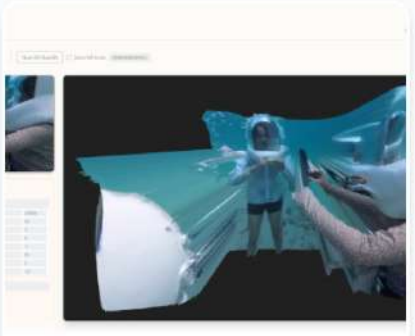
ComfyUI-DepthAnythingV3



DA3-ROS2-Wrapper



DA3-ROS2-CPP-TensorRT



VideoDepthViewer3D

Summary & Looking Forward

DA3: minimal modelling strategy

- Minimal representation: **depth + rays**
- Minimal architecture: **a single plain transformer**
- **State-of-the-art** across pose, geometry, monocular depth, and feed-forward NVS

Not the end of the DA series:

- **Better capability:** Can we unify 3D + 4D vision?
- **3D World models:** Can we unify 3D reconstruction + 3D generation?
- **Spatial intelligence:** LLM-friendly architecture and representation?

Discussion

Thanks for Your Attention

Depth Anything 3

Trace Anything

One More Thing:

- **Trace Anything:** Representing Any Video in 4D via Trajectory Fields; Unify 3D + 4D vision
- Come chat with our team at: **P4-#3310, Apr 25, 10:30 AM**