Fusing Learned and Sensor-Based Depth for Consistent 3D Mapping

Werikson Alves

UFV - DPI - INF791

Motivation

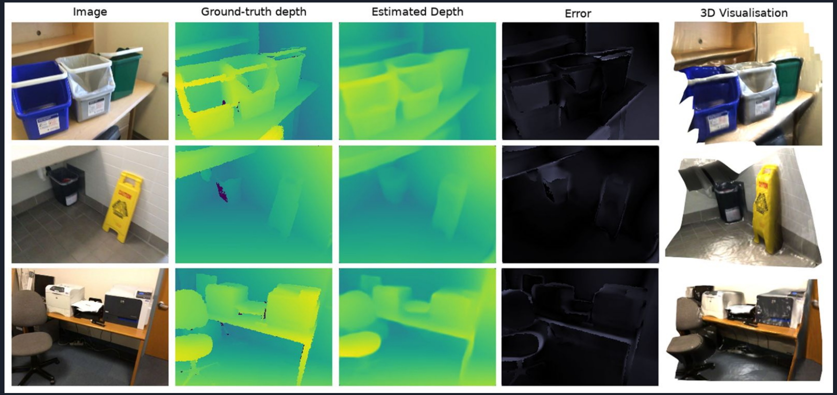
- Reconstructing 3D environments from monocular video is challenging
- Depth sensors are costly and noisy
- Deep models produce dense geometry, but lack scale
- 🧩 **Goal**: Combine both sources to leverage their strengths

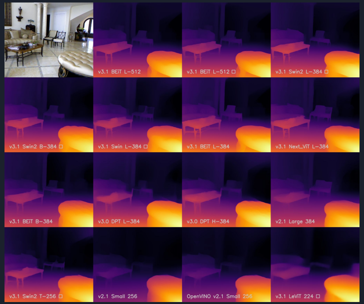
Objectives

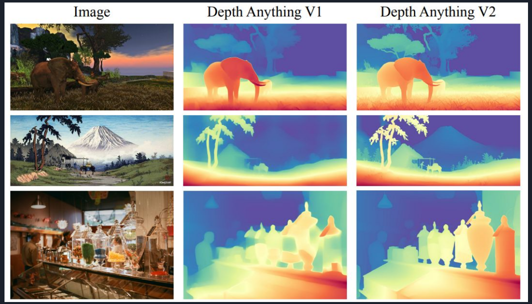
- Design a **modular pipeline** for 3D reconstruction
- Integrate **DepthAnythingV2** with **RealSense D435** depth
- Perform **depth fusion** and multiway reconstruction
- Evaluate **qualitative and quantitative** improvements

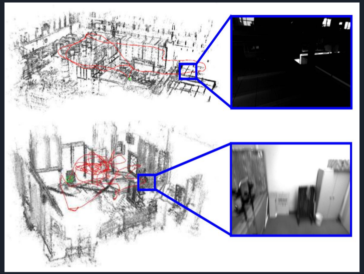
Related Work

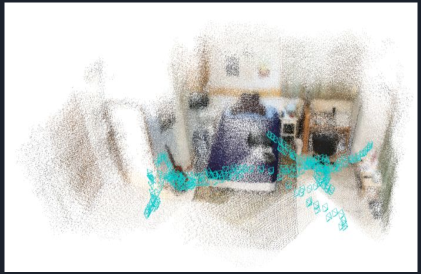
- SLAM with sensors:
  - ORB-SLAM, RTAB-Map, DeepFactors
- Monocular depth:
  - MiDaS, DepthAnythingV2
- Fusion in SLAM:
  - CNN-SLAM, D3VO, DROID-SLAM
  - Most embed depth in **tracking loop**









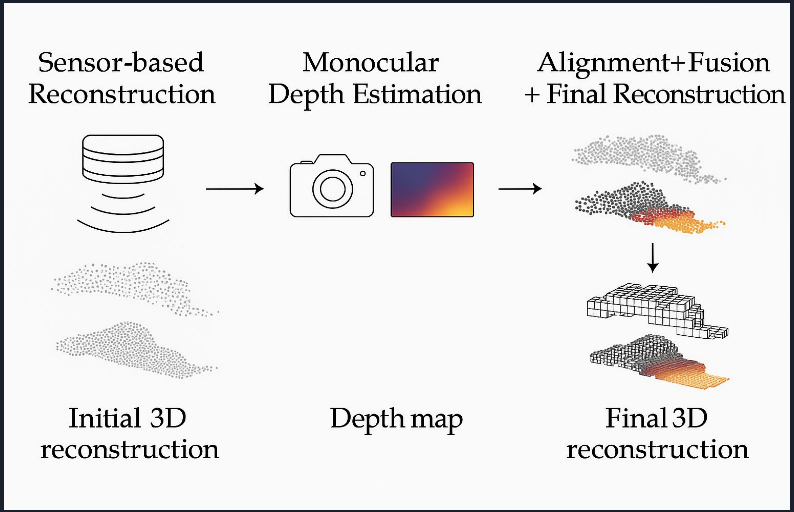


Related Work

- External fusion instead of integration in tracking
- Operates offline, frame-wise
- Focus on reconstruction fidelity, not real-time pose

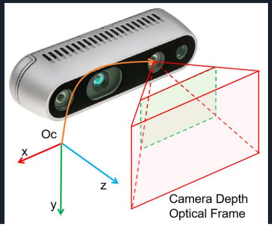
Method Overview

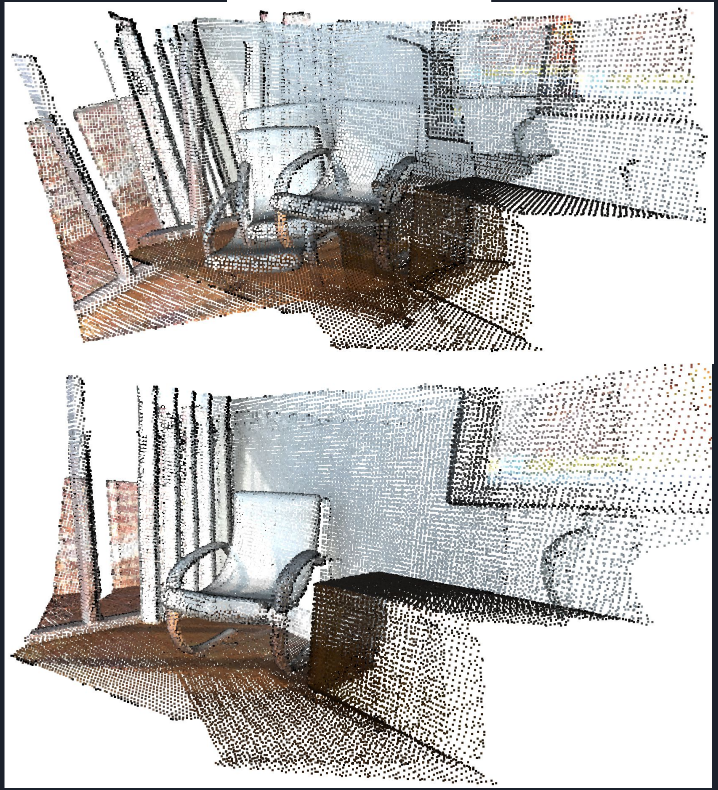
- 3 main stages:
    - Sensor-based reconstruction
    - Monocular depth estimation
    - Alignment + fusion + final reconstruction



Stage 1: Sensor-Based Reconstruction

- RealSense D435: RGB + Depth
- Multiway registration (Open3D):
  - Point cloud → ICP → Pose graph → Optimization
- Output: Sensor-based point cloud





Stage 2: Monocular Depth Estimation

- Model: **DepthAnythingV2** (vits encoder)
- Inference over RGB folder
- Same pipeline (registration, global merge)
- Issue: no absolute **scale**, causing distortion

Stage 3: Scale Alignment & Fusion

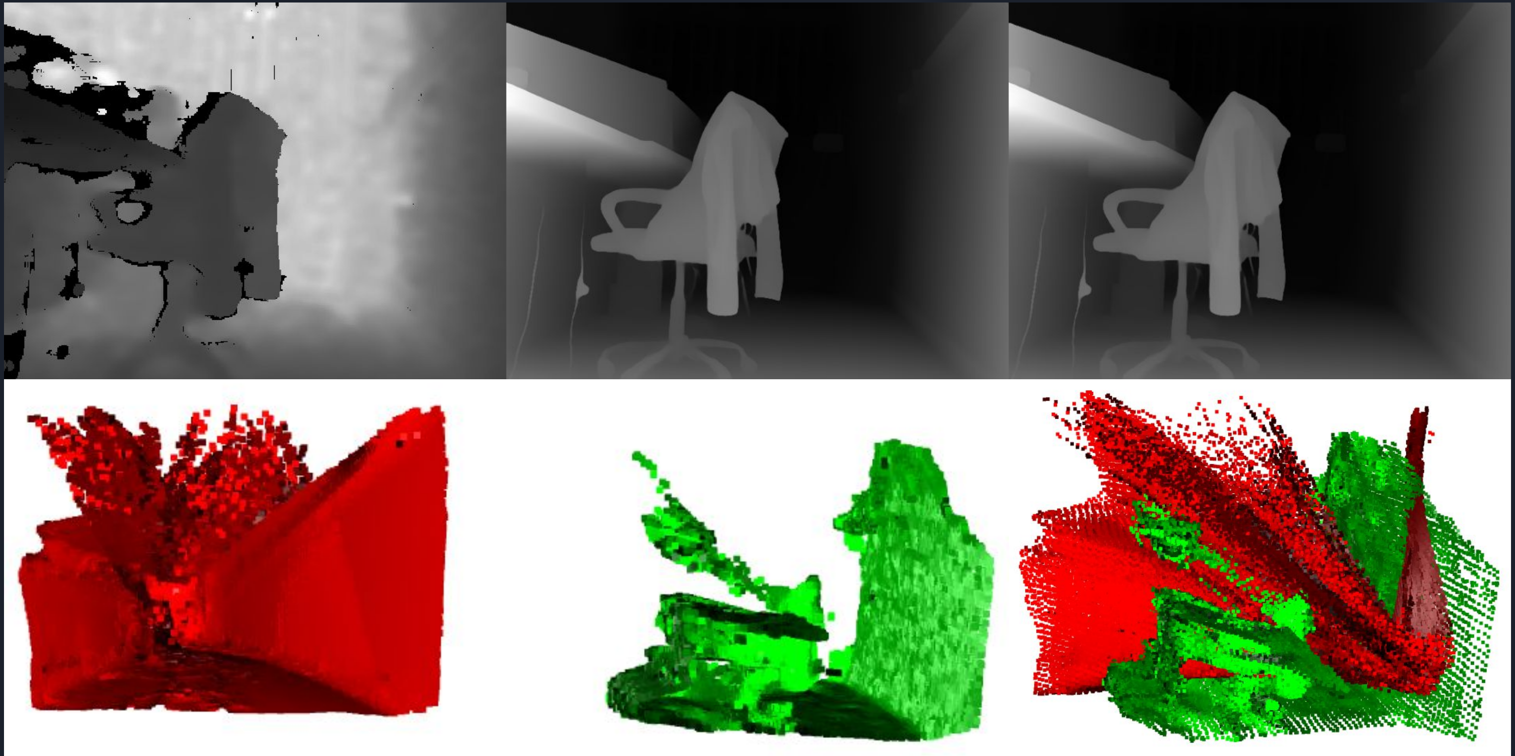
- Estimate **scale factor** (inverse-depth regression)
- Apply **ICP** to align clouds frame-wise
- Depth fusion:
  - z-score based blending
  - Combine accuracy (sensor) + completeness (monocular)

$$d_f = \frac{d_e - \mu(d_e)}{\sigma(d_e)} \cdot \sigma(d_s) + \mu(d_s)$$

Experimental Setup

- 3 indoor sequences: U, L, I shaped paths
- 60 cm height, 4 FPS
- Hardware: i7 CPU + RTX 2050
- Voxel size: 0.05 m

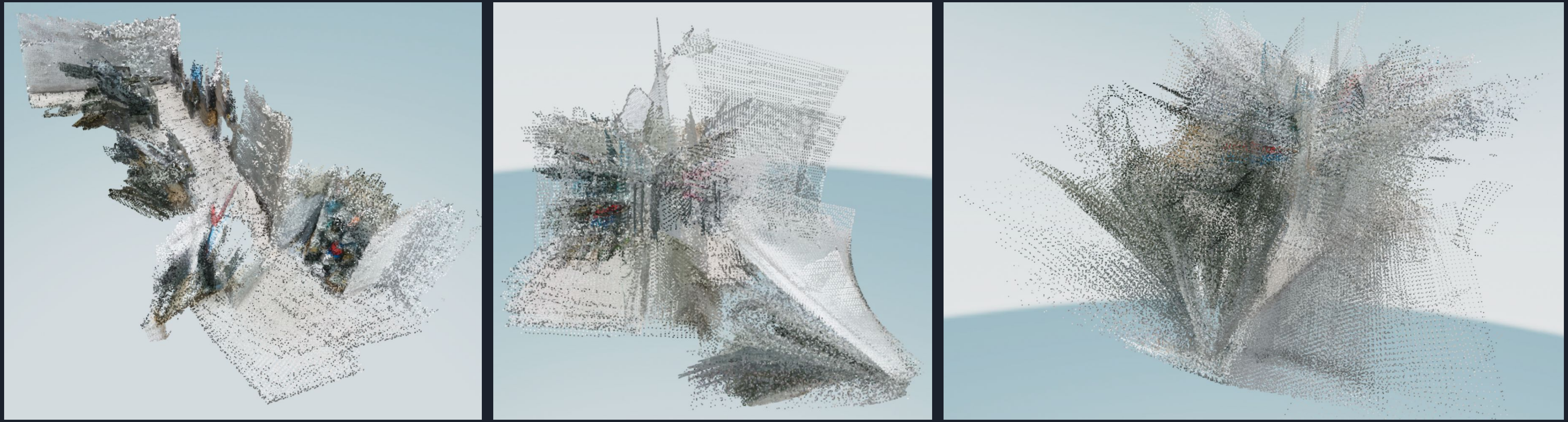
Depth Map Visualization



Reconstruction Results (I)

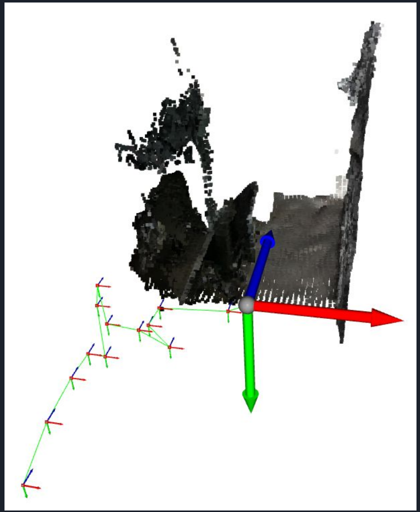
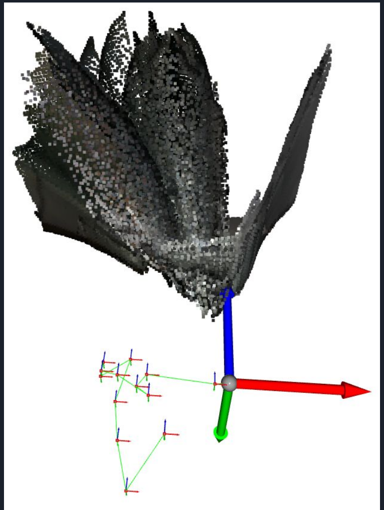
Reconstruction Results (L)

Reconstruction Results (U)



Trajectory Estimation (I)

- Trajectory from sensor vs. fused
- Note: Fused trajectory more complete, but **scale drift**

Quantitative Results (I)

Method	Points	Volume (m^3)	Density
Sensor	40,431	18.84	2145.68
Estimated	76,990	24.48	3145.58
Fused	71,247	28.51	2499.23

Discussion

- Fused results are smoother and denser
- Sensor maps are accurate but noisy
- Monocular maps are rich but misaligned
- Fusion reduces noise and completes missing regions
- 🔴 Scale drift remains a major issue

Limitations & Challenges

- No **absolute scale** in monocular depth
- Drift accumulates over time
- Multiway registration sensitive to depth inconsistency
- Processing time grows with sequence length

Future Directions

- Global **scale correction** (e.g., SLAM-assisted)
- **Pose-depth joint optimization**
- Temporal depth fusion (across frames)
- Real-time adaptation for mobile robotics

Conclusion

- Presented a **modular and scalable pipeline**
- Combines deep monocular and sensor depth
- Fused reconstructions show clear improvements
- Foundation for robust monocular 3D SLAM

Questions?

- Contact: werikson.alves@ufv.br
- Code available on GitHub