UNIVERSITY OF MICHIGAN

Multi-Sensor SLAM: Low-cost LiDAR and Visual Fusion (LLV-SLAM)

GROUP 3
Phurithat Tangsripairoje, Meng Wu, Yikuan Xiong

EECS 568/ROB 530 - Mobile Robotics

Motivation

- Visual SLAM is low-cost and rich in semantics, but sensitive to lighting and weak in depth estimation
- LiDAR SLAM is robust and accurate, but costly and lacks semantic understanding
- Many multi-modal SLAM systems rely on expensive 3D LiDAR or pre-calibrated sensor setups
- Need for lightweight and robust SLAM that works in degraded or resource-constrained environments
- Can a low-cost sparse LiDAR help improve visual SLAM performance?

Proposed Solution

- A lightweight SLAM system combining sparse LiDAR and RGB images via adaptive depth refinement
- Designed for low-cost platforms, it improves depth accuracy and trajectory estimation in degraded scenes
- The system is cost-effective, robust, and applicable to both indoor and outdoor environments

UNIVERSITY OF MICHIGAN

Methodology

Load KITTI Dataset
Stereo RGB & Gray Images, Velodyne, Calibration matrices

Stereo Depth Estimation
$Z = fx \cdot b / d$
LiDAR fusion for depth enhancement

Initialize VO State
Feature Detection (SIFT/ORB/AKAZE)
Matching (BF/FLANN)

SLAM Loop
For each frame:
1. PnP Pose Estimation
2. Append pose to trajectory
3. Add keyframe to DB
4. Loop Closure Detection
    a. Match with earlier keyframes
    b. Estimate relative pose (PnP)
    c. Add loop constraint
5. Generate RGB-D Point Cloud

Pose Graph Optimization
GTSAM (Levenberg–Marquardt)
VO + loop constraints

Dense 3D Mapping
Back-project RGB-D, Transform, Downsample

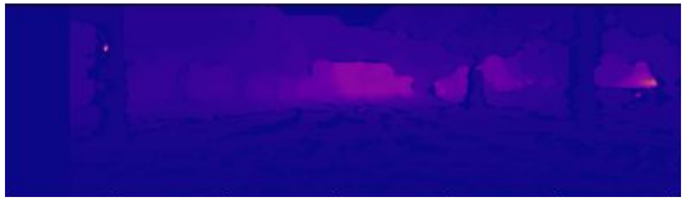
Evaluation & Visualization
ATE RMSE, 2D/3D Maps

Offline Stereo Visual (+ sparse LiDAR) SLAM Pipeline

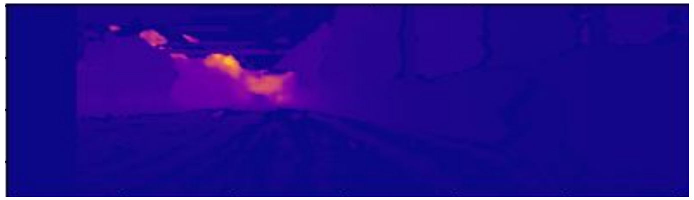
Baseline Comparison

RGB

Depth map



Seq. 0002



Seq. 0011

To evaluate the performance of the baseline stereo visual SLAM pipeline, we selected three sequences from the KITTI dataset representing varied motion and lighting conditions. Each sequence was tested using multiple combinations of feature extractors and matching strategies.

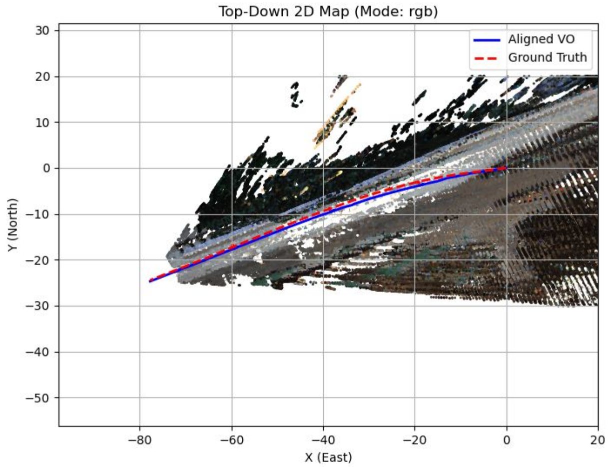
KITTI Sequences:

- `0002_sync`: Straight path, uniform lighting
- `0011_sync`: Curved road, strong lighting variation
- `0009_sync`: Sharp turns, moderate lighting variation

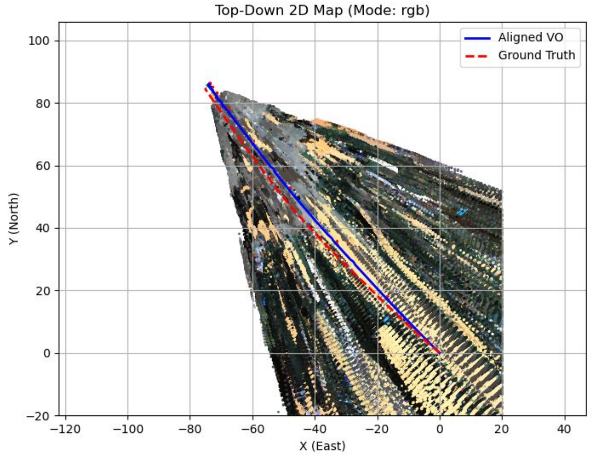
Configurations Tested:

- **Feature extractors:** ORB, SIFT, AKAZE
- **Matchers:** Brute-Force with Cross-Check, Brute-Force with KNN, FLANN with KNN

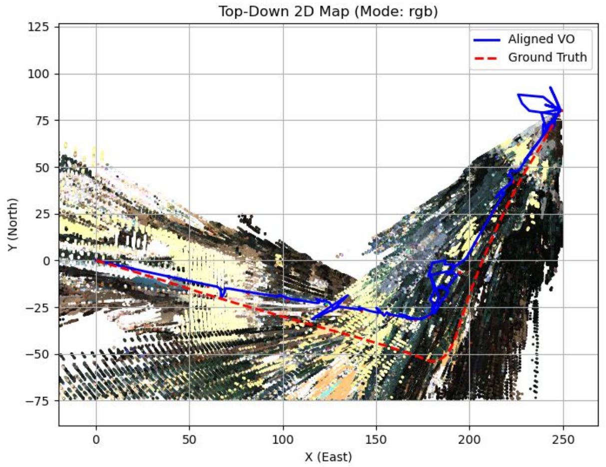
Key Observations for baseline cases



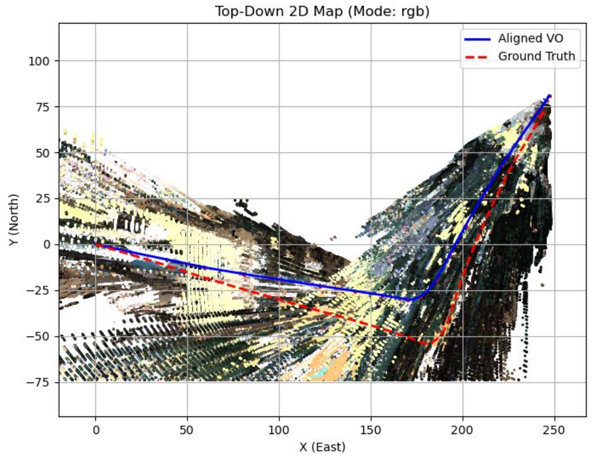
AKAZE-BF_crosscheck on 0002



SIFT-BF_crosscheck on 0011



SIFT-BF_crosscheck on 0009



SIFT-FLANN_KNN on 0009

Key findings from the baseline cases

- **AKAZE + BF cross-check**: Best ATE on straight path (`0002`)
- **SIFT + BF cross-check**: Best on curves (`0011`, `0009`), worse on `0002`
- **SIFT + FLANN-KNN**: Smoother but slightly less accurate
- **Sharp turns (`0009`)**: Caused higher VO drift
- **Lighting variation**: Degraded point cloud clarity
- **Conclusion**: Visual-only SLAM is sensitive to motion and illumination — highlights need for sensor fusion

Sequence	Best Config	ATE RMSE (m)
0002	AKAZE + BF_crosscheck	0.8046
0011	SIFT + BF_crosscheck	1.9058
0009	SIFT + BF_crosscheck	16.3576
0009	SIFT + FLANN_KNN	16.4777

Methodology

- Stereo + LiDAR Depth Refinement
- Monocular Scale Alignment via LiDAR

Methodology

Stereo + LiDAR Depth Refinement

LiDAR Projection into Image Plane

Step 1: Transform LiDAR point P to camera frame:

$$P_c = R_{vl} \cdot P + t_{vl} \quad \text{then} \quad P_r = R_{\text{rect}} \cdot P_c$$

Step 2: Project to image:

$$u \sim K \cdot P_r$$

Fusion Strategy:

For each pixel $u \in \Omega$:

- If LiDAR is valid and stereo is missing or far off:

$$D_{\text{fused}}(u) = D_{\text{lidar}}(u)$$

- If the two are close:

$$D_{\text{fused}}(u) = \alpha D_{\text{lidar}}(u) + (1 - \alpha) D_{\text{stereo}}(u)$$

- Otherwise, keep stereo:

$$D_{\text{fused}}(u) = D_{\text{stereo}}(u)$$

Post-Processing

$$\mathcal{D}_{\text{smooth}}(\mathbf{u}) = \frac{1}{W(\mathbf{u})} \sum_{\mathbf{v} \in \mathcal{N}(\mathbf{u})} w_s(\mathbf{u}, \mathbf{v}) \cdot w_r(I(\mathbf{u}), I(\mathbf{v})) \cdot \mathcal{D}_{\text{fused}}(\mathbf{v})$$

Methodology

Monocular Scale Alignment via LiDAR

Problem: Monocular depth D_{mono} lacks absolute scale

Solution: Use sparse LiDAR to compute global scale $\hat{s}$

- **Valid set:**

$$V = \{u \in \Omega \mid D_{\text{lidar}}(u) > 0 \wedge D_{\text{mono}}(u) > 0\}$$

- **Scale factor:**

$$\hat{s} = \text{median} u \in V\left(\frac{D\text{lidar}(u)}{D_{\text{mono}}(u)}\right)$$

- **Aligned depth:**

$$D_{\text{aligned}}(u) = \hat{s} \cdot D_{\text{mono}}(u)$$

Methodology

Monocular Scale Alignment via LiDAR

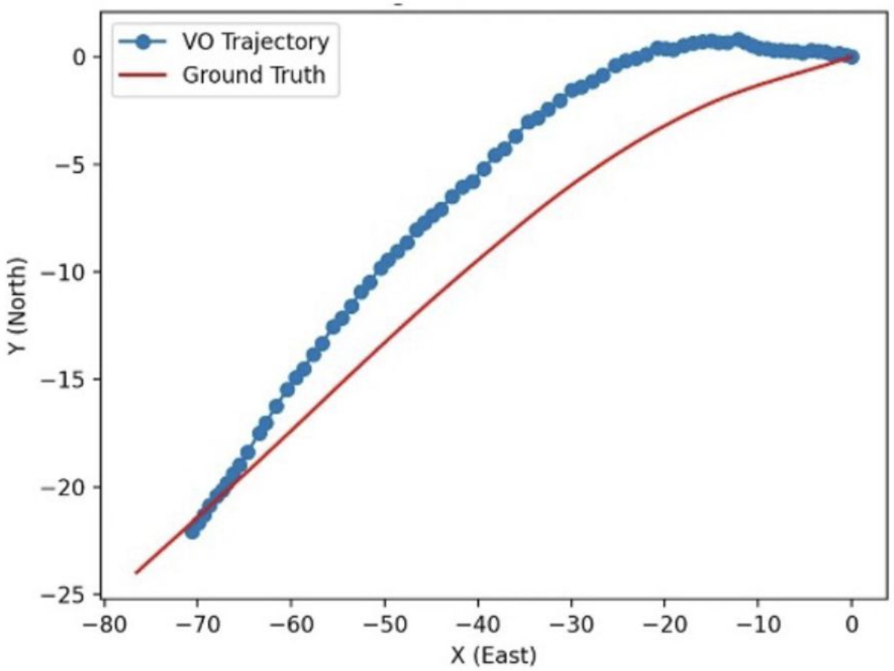
While global scaling suffices for many applications, local inconsistencies may still remain, especially near object boundaries or in textureless regions. We optionally adopt sparse LiDAR corrections at selected pixels, following a similar blend or replacement rule as in last slides, using the aligned monocular depth as a base and using the post-processing smooth as well.

Post-Processing

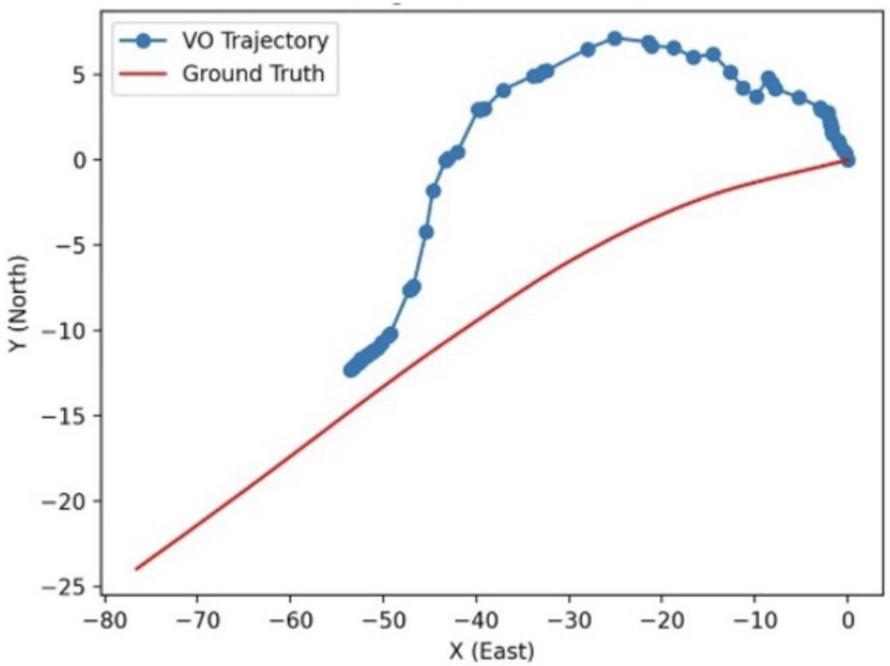
$$\mathcal{D}_{\text{smooth}}(\mathbf{u}) = \frac{1}{W(\mathbf{u})} \sum_{\mathbf{v} \in \mathcal{N}(\mathbf{u})} w_s(\mathbf{u}, \mathbf{v}) \cdot w_r(I(\mathbf{u}), I(\mathbf{v})) \cdot \mathcal{D}_{\text{fused}}(\mathbf{v})$$

Results

Differences between monocular with and without alignment



(a)



(b)

(a): VO of the monocular SLAM with the alignment (b) VO of the monocular SLAM without the alignment

Results

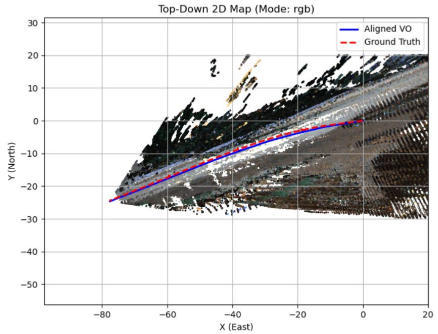
Metric differences between stereo depth estimation with and without alignment

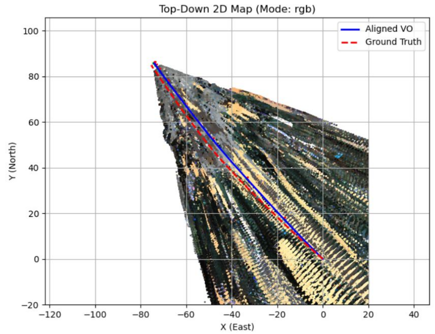
ATE RMSE without enhancement

Sequence	ORB BF_crosscheck	ORB BF_KNN	ORB FLANN_KNN	SIFT BF_crosscheck	SIFT BF_KNN	SIFT FLANN_KNN	AKAZE BF_crosscheck	AKAZE BF_KNN	AKAZE FLANN_KNN
0002	27.2579	12.0044	13.0190	9.5553	19.9516	2.9547	0.8046	10.0485	7.7121
0011	2.9519	50.7017	33.3848	1.9058	94.4574	95.4215	7.9700	92.5149	81.5799
0009	19.5827	187.8725	18.6645	16.3576	187.1956	16.4777*	18.3227	55.1842	22.8239

ATE RMSE with enhancement

Sequence	ORB BF_crosscheck	ORB BF_KNN	ORB FLANN_KNN	SIFT BF_crosscheck	SIFT BF_KNN	SIFT FLANN_KNN	AKAZE BF_crosscheck	AKAZE BF_KNN	AKAZE FLANN_KNN
0002	32.1204	11.9972	13.0965	2.3043	20.0736	2.9350	0.5432	10.6094	6.7473
0011	9.6483	51.4318	25.0061	4.6827	1.8504	97.8086	87.4956	92.3244	81.3993
0009	18.1985	187.3683	187.1941	16.4432	187.1988	187.1518	17.9459	54.5246	22.9818





Discussion

- **Monocular + LiDAR (:**
- ✅ **Significantly improves** trajectory accuracy
- ✅ Enables **metric scale** for downstream mapping and planning
- ✅ **Model-agnostic**: no retraining or architectural changes
- 🔁 Slight local inconsistencies may remain near object edges

- **Stereo + LiDAR (:**
- ✅ Reduces **local outliers** and improves mapping fidelity
- ⚠️ Only **modest gains** in overall pose estimation — stereo already metric
- 🧩 Acts more like a **robustness enhancer**, not a necessity

Conclusion

Sparse LiDAR provides **complementary benefits** to both monocular and stereo vision:

- For monocular SLAM, it solves a fundamental scale problem;
- For stereo SLAM, it enhances depth reliability in edge cases.

Future works

Future directions include:

- Learning-based depth–LiDAR fusion with uncertainty modeling
- Online calibration and self-supervised refinement
- Robustness to dynamic objects and environmental changes

Thank you!

GROUP 3
Phurithat Tangsripairoje, Meng Wu, Yikuan Xiong

EECS 568/ROB 530 - Mobile Robotics