Multi-Session SLAM with Differentiable Wide-Baseline Pose Optimization

Lahav Lipson and Jia Deng

PRINCETON VISION & LEARNING LAB

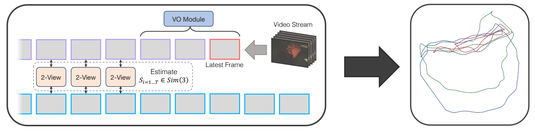
CVPR JUNE 17-21, 2024 SEATTLE, WA

Multi-Session SLAM

We introduce a new system for Monocular Multi-Session SLAM, which tracks camera motion across multiple disjoint videos under a single global reference. Our approach couples the prediction of optical flow with optimization layers to estimate camera pose.

Simultaneous Localization and Mapping (SLAM) is the task of estimating camera motion and a 3D map from video. Video data in the wild often consists of not a single continuous stream, but rather multiple disjoint sessions, either deliberately such as in collaborative mapping when multiple robots perform joint rapid 3D reconstruction, or inadvertently due to visual discontinuities in the video stream which can result from camera failures, extreme parallax, rapid turns, auto-exposure lag, dark areas, or occlusions.

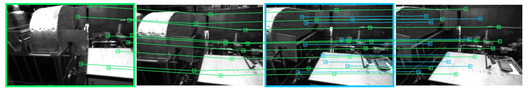
Key Challenges: Real-time camera pose + scale/location/position ambiguities between monocular videos



Example: Multi-Session SLAM

Visual SLAM with Sparse Optical Flow

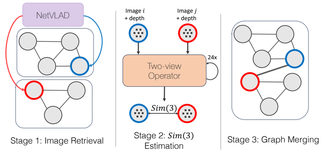
Our approach uses *sparse optical flow* to track keypoints between frames. This is as opposed to keypoint-matching approaches like Superglue and ORB. Using sparse flow has both advantages, and disadvantages.
Benefit: Robust in low-texture environments. No requirement for keypoint detector.
Drawback: Existing approaches to MS-SLAM are incompatible, e.g., ORB-SLAM3



Connecting Disjoint Trajectories

The goal in Multi-Session SLAM is to estimate camera motion for all monocular video streams under a single global reference.

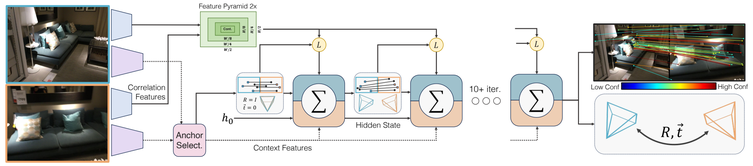
Our approach attempts **(1)** to estimate camera motion from video streams *individually* (like a typical SLAM system) and **(2)** to connect disjoint sequences by aligning their respective coordinate systems. To perform **(2)**, we identify co-visible image pairs using image retrieval and then run our two-view pose estimator to predict a 7DOF alignment between sequences.



Our approach to connecting disjoint sequences

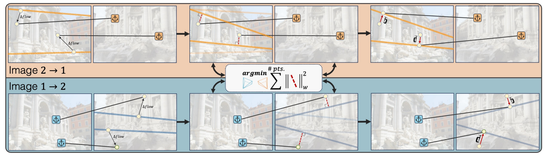
Recurrent Sparse Optical Flow for 2-view Camera Pose

A core subsystem of our Multi-Session SLAM method is a new approach to wide-baseline, 2-view relative camera pose. Given two views as input, we alternate between estimating sparse optical flow residuals using a weight-tied network, and updating the relative pose estimate with an optimization layer. Our method also implicitly learns a confidence measure for each predicted flow vector.



Updating Camera Poses and Optical Flow

Each update iteration follows three steps: **(1)** We predict an update to sparse optical flow. **(2)** We then update the camera pose estimates to align with the predicted flow. **(3)** Finally, we clamp the optical flow to the newly-induced epipolar lines.



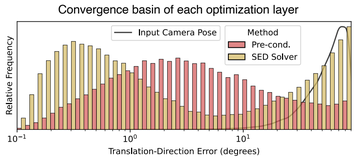
The optimizer in **(2)** parameterizes the epipolar lines as a function of the camera poses, and seeks to minimize Symmetric Epipolar Distance (SED) between the predicted matches and the epipolar lines.

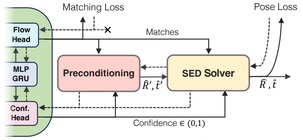
Differentiable Camera Pose Optimizer(s)

Unfortunately, the SED optimizer will converge to local minima if initialized far from the true optimum. To remedy this, we adopt a pre-conditioning stage which uses a weighted version of the 8-point algorithm. The combined optimizers are fully differentiable, meaning we can supervise on the final pose output.

Weighted 8-pt-algorithm (used for initialization)
- No local minimum
- Less accurate

SED Optimizer (used for refinement)
- Local minimum / non-convex
- More accurate





Our optimizer architecture, and the backward gradients (dotted lines)

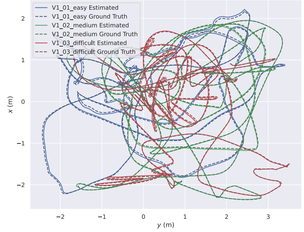
Quantitative Results

We evaluate our full system on the EuRoC and ETH-3D datasets in which all the ground truth trajectories are in a unified coordinate system. Compared to existing methods, our approach is significantly more robust and accurate. All reported methods run in real-time (camera hz = 20 FPS)

Scene name	MH01-03	MH01-05	V101-103	V201-203	Cam.
# Disjoint Trajectories	3	5	3	3	hz
Ours Mono-Visual	**0.022**	**0.036**	**0.031**	**0.024**	20
CCM-SLAM [36] Mono-Visual	0.077	-	-	-	**38**
ORB-SLAM3 [4] Mono-Visual	0.030	0.058	0.058	0.284	20
VINS [28] Mono-Inertial	-	0.210	-	-	-
ORB-SLAM3 Mono-Inertial	0.037	0.065	0.040	0.048	20

Multi-Session SLAM on EuRoC

Our prediction for sequence group V101-V103 of EuRoC



Scene name	Sofa	Table	Plant Scene	Einstein	Planar
# Disjoint Trajectories	4	2	3	2	2
ORB-SLAM3 [4] Mono-Visual	FAIL (no init)	0.018	FAIL (init→lost)	FAIL (init→lost)	**0.010**
Ours Mono-Visual	**0.010**	**0.010**	**0.021**	**0.032**	0.047

Multi-Session SLAM on ETH-3D

Ablation: Two-view Relative Pose on Scannet

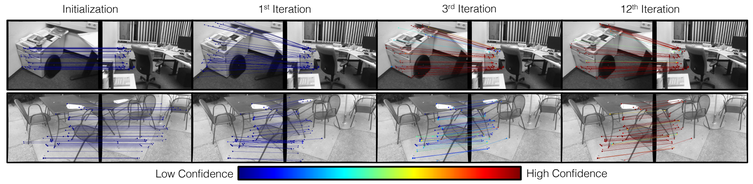
We evaluate our two-view relative pose subsystem in isolation. We outperforms existing methods on Scannet. Existing approaches perform matching as a pre-processing step, whereas ours is a weight-tied network alternating between optimization and matching.

	Overall Approach	Pose error AUC [%] ↑ @5°	@10°	@20°
Superglue [34]	Matching ↓ RANSAC [15] (LO-RANSAC [18])	16.2(17.7)	33.8(35.6)	51.8(54.2)
LightGlue [23]		16.5(19.4)	33.4(36.9)	50.1(53.5)
LoFTR [41]		22.1(25.7)	40.1(45.0)	47.6(61.4)
MatchFormer [47]		24.3(27.3)	43.9(47.6)	61.4(64.9)
ASpanFormer [5]		25.6(28.4)	46.0(48.8)	63.3(65.8)
Roessle&Nießner [31]	Matching→WSPA	20.7	41.6	61.7
Roessle&Nießner [31]	Matching→WSPA→BA	25.7	47.2	66.4
Ours	Optical-Flow ↘ Clamp +- Solver ↗	**30.5**	**50.9**	**67.5**

Quantitative two-view results on Scannet

Qualitative Two-View Matching Results

Qualitative results on Scannet. Our two-view subsystem estimates accurate relative poses across wide camera baselines. It initializes all matches with uniform depth and identity relative pose. Progressive applications of our update operator lead to more accurate matches and higher predicted confidence.



Qualitative results on Scannet