Fast Monocular Visual-Inertial Initialization Leveraging Learned Single-View Depth

Nathaniel Merrill, Patrick Geneva, Saimouli Katragadda, Chuchu Chen, and Guoquan Huang

Robot Perception and Navigation Group (RPNG)

University of Delaware, USA

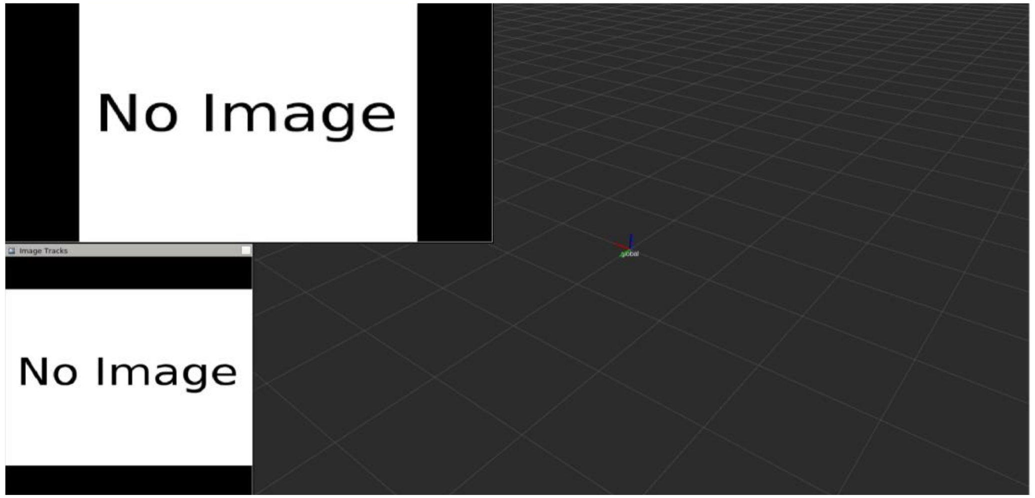
Introduction

- Visual-Inertial Odometry (VIO) requires **accurate initial conditions** to run
- State-of-the-art systems require 2sec, large parallax and many features to init

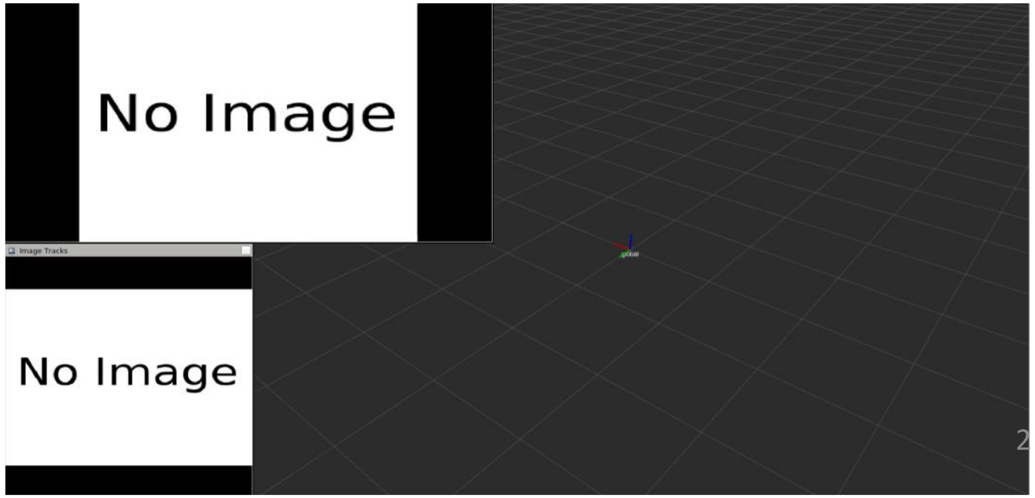
Introduction

- Visual-Inertial Odometry (VIO) requires **accurate initial conditions** to run
- State-of-the-art systems require 2sec, large parallax and many features to init
- **This work**
  - Propose a new initialization method for monocular VIO using **learned monocular depth**

Baseline



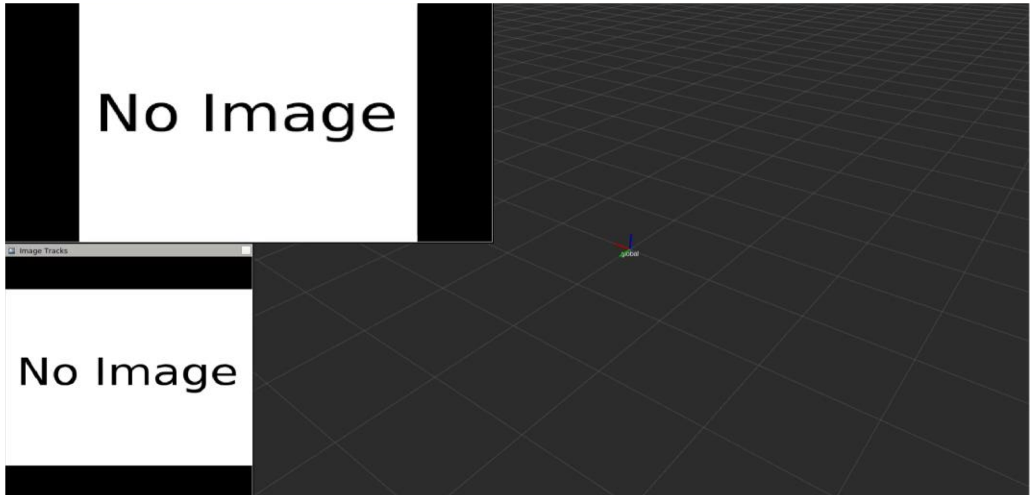
Ours



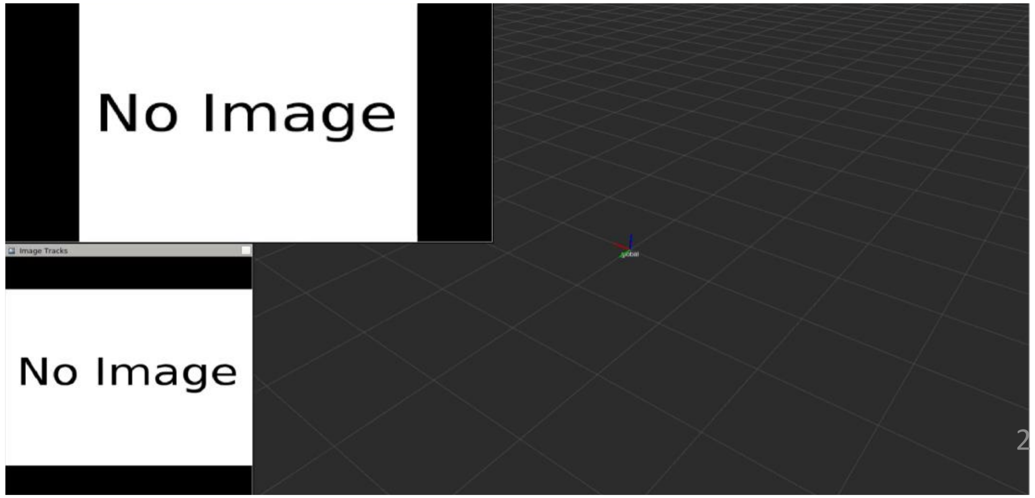
Introduction

- Visual-Inertial Odometry (VIO) requires **accurate initial conditions** to run
- State-of-the-art systems require 2sec, large parallax and many features to init
- **This work**
  - Propose a new initialization method for monocular VIO using **learned monocular depth**
  - Our method is shown to be faster, more accurate, and **more robust,** initializing in only **300ms** with low parallax and as low as **15 features**

Baseline

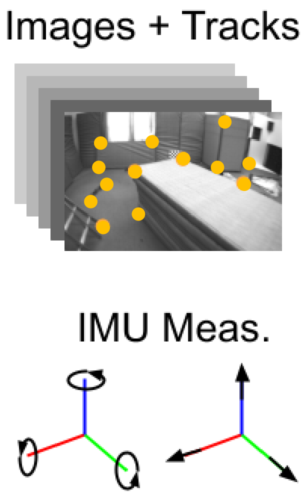


Ours



Baseline Monocular Visual-Inertial Initialization

- State-of-the-art monocular initialization methods [1] use image tracks and IMU measurements in a VI-SfM to solve for initial conditions
- Large number (M) features required to initialize

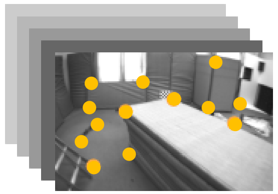


[1] T.-C. Dong-Si and A. I. Mourikis, "Estimator initialization in vision-aided inertial navigation with unknown camera-imu calibration," in IROS 2012

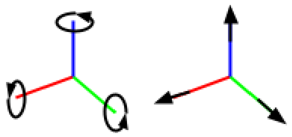
Baseline Monocular Visual-Inertial Initialization

- State-of-the-art monocular initialization methods [1] use image tracks and IMU measurements in a VI-SfM to solve for initial conditions
- Large number (M) features required to initialize

Images + Tracks



IMU Meas.



VI Linear System

$$\mathbf{A}\left[\mathbf{p}_{f_1}^\top \; \cdots \; \mathbf{p}_{f_M}^\top \; \mathbf{v}^\top \; \mathbf{g}^\top\right]^\top = \mathbf{b}$$

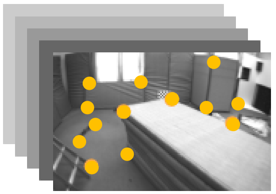
Landmarks, velocity, and gravity solved in linear system of dim 6+3M

[1] T.-C. Dong-Si and A. I. Mourikis, “Estimator initialization in vision-aided inertial navigation with unknown camera-imu calibration,” in IROS 2012

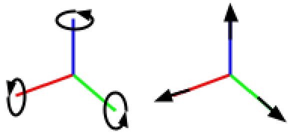
Baseline Monocular Visual-Inertial Initialization

- State-of-the-art monocular initialization methods [1] use image tracks and IMU measurements in a VI-SfM to solve for initial conditions
- Large number (M) features required to initialize

Images + Tracks



IMU Meas.



VI Linear System

$$\mathbf{A}\left[\mathbf{p}_{f_1}^\top \ \cdots \ \mathbf{p}_{f_M}^\top \ \mathbf{v}^\top \ \mathbf{g}^\top\right]^\top = \mathbf{b}$$

Landmarks, velocity, and gravity solved
in linear system of dim 6+3M

VI-BA

Initial vel, gravity,
landmarks

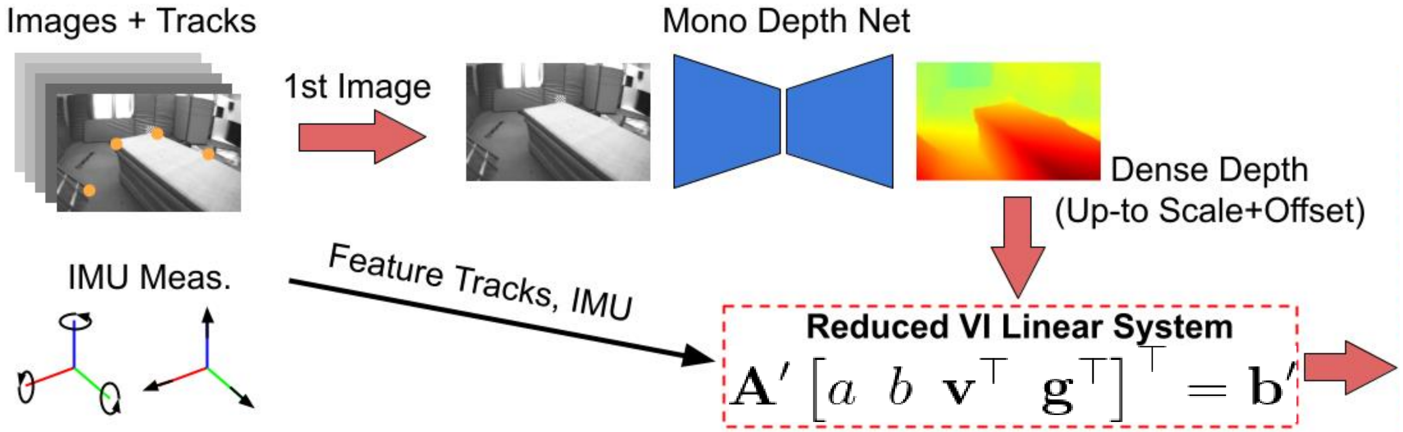
VIO

[1] T.-C. Dong-Si and A. I. Mourikis, “Estimator initialization in vision-aided inertial navigation with unknown camera-imu calibration,” in IROS 2012

Proposed Initialization Method

- **Key idea:** leverage learned monocular depth to *reduce* the linear system
  - Propose new model of 3D landmarks w.r.t. learned affine-invariant depth d_i

$$\mathbf{p}_{f_i} = z_i \, \boldsymbol{\theta}_{f_i}$$
$$= (a d_i + b) \, \boldsymbol{\theta}_{f_i}$$



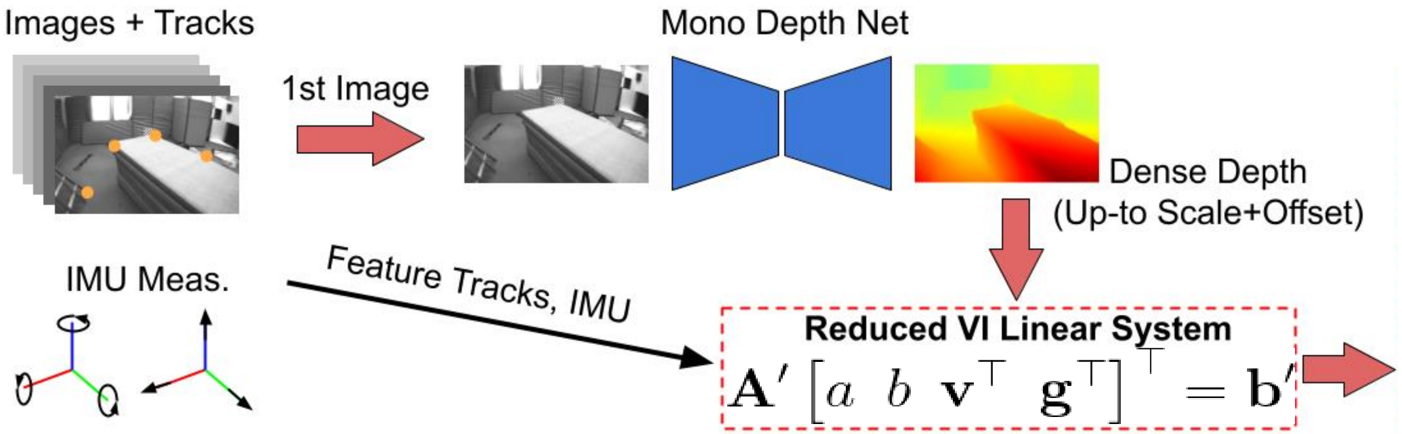
Scale, offset, velocity, and gravity solved in linear system of dim 8

Proposed Initialization Method

- **Key idea:** leverage learned monocular depth to *reduce* the linear system
  - Propose new model of 3D landmarks w.r.t. learned affine-invariant depth d_i

$$\mathbf{p}_{f_i} = z_i \, \boldsymbol{\theta}_{f_i}$$
$$\boxed{= (a d_i + b) \, \boldsymbol{\theta}_{f_i}}$$

Only estimate a, b to represent all landmarks



Scale, offset, velocity, and gravity
solved in linear system of dim 8

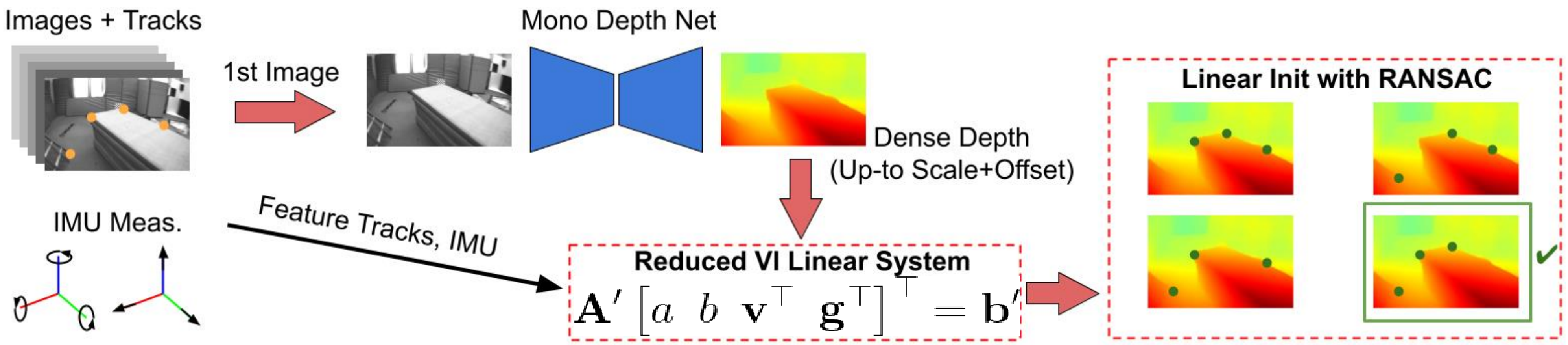
Proposed Initialization Method

- **Key idea:** leverage learned monocular depth to *reduce* the linear system
  - Propose new model of 3D landmarks w.r.t. learned affine-invariant depth d_i

$$\mathbf{p}_{f_i} = z_i \, \boldsymbol{\theta}_{f_i}$$
$$\boxed{= (a d_i + b) \, \boldsymbol{\theta}_{f_i}}$$

Only estimate a, b to represent all landmarks

- Because of the reduced system, RANSAC is practical
  - Minimal problem reduced from **6+3M** for M landmarks to **8**



Scale, offset, velocity, and gravity solved in linear system of dim 8

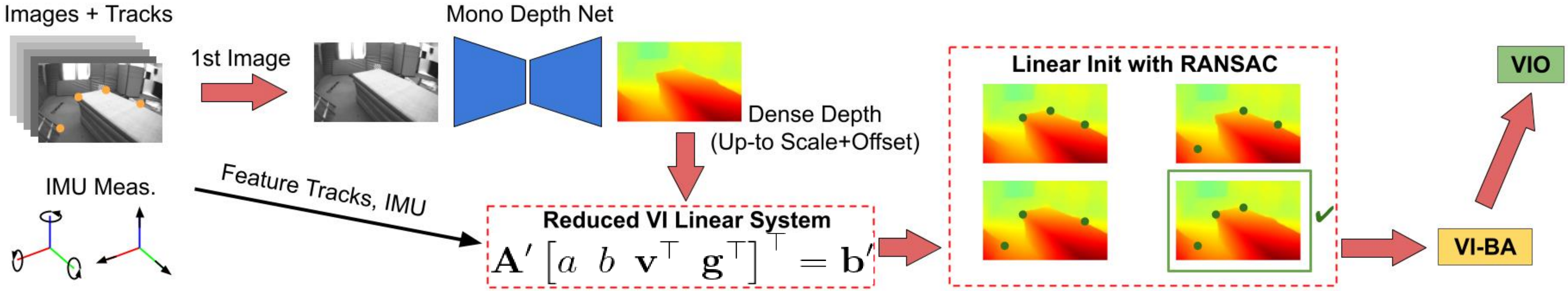
Proposed Initialization Method

- **Key idea:** leverage learned monocular depth to *reduce* the linear system
  - Propose new model of 3D landmarks w.r.t. learned affine-invariant depth d_i

$$\mathbf{p}_{f_i} = z_i \, \boldsymbol{\theta}_{f_i}$$
$$\boxed{= (a d_i + b) \, \boldsymbol{\theta}_{f_i}}$$

Only estimate a, b to represent all landmarks

- Because of the reduced system, RANSAC is practical
  - Minimal problem reduced from **6+3M** for M landmarks to **8**



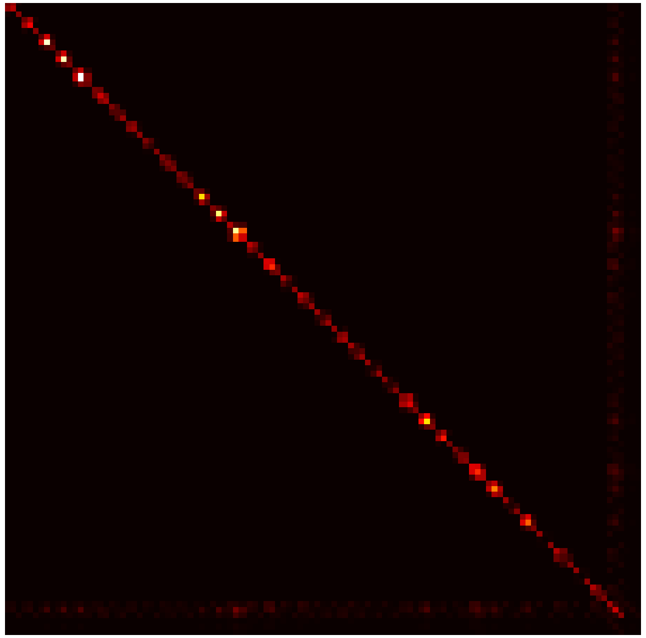
Scale, offset, velocity, and gravity solved in linear system of dim 8

Linear System Structure

- Our linear system is considerably smaller than the baseline one
- Replacing M landmark positions with a, b reduces the size to 8

Baseline $\mathbf{A}^\top\mathbf{A}$

6+3M



6+3M

Our $\mathbf{A}'^\top\mathbf{A}'$

8

8

EuRoC Results

- Tested on EuRoC Vicon room datasets with 5 KFs over a 0.5sec window (avg over hundreds of initializations), and using MiDaS [1] network
- **Comparisons**
  - **DS 3D:** Baseline initialization (Dong-Si [2]) with 3D landmarks
  - **DS + DP:** Our reimplementation of [3] (mono depth priors in VI-BA)

Table: Scale error (%)

Algorithm	Avg.
DS 3D	57.6
DS + DP	58.8
Ours w/o RANSAC	17.3
Ours	**5.8**

Table: ATE (deg / m)

Algorithm	Average
DS 3D	1.592 / 0.028
DS + DP	1.523 / 0.027
Zhou [3]	- / 0.024
Ours w/o RANSAC	1.467 / 0.026
Ours	**1.419 / 0.022**

[1] R. Ranftl et. al, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," in TPAMI 2022.

[2] T.-C. Dong-Si and A. I. Mourikis, "Estimator initial-ization in vision-aided inertial navigation with unknown camera-imu calibration," in IROS 2012

[3] Y. Zhou et. al, "Learned Monocular Depth Priors in Visual-Inertial Initialization," in ECCV 2022.

EuRoC Results

- Tested on EuRoC Vicon room datasets with 5 KFs over a 0.5sec window (avg over hundreds of initializations), and using MiDaS [1] network
- **Comparisons**
  - **DS 3D:** Baseline initialization (Dong-Si [2]) with 3D landmarks
  - **DS + DP:** Our reimplementation of [3] (mono depth priors in VI-BA)

Table: Scale error (%)

Algorithm	Avg.
DS 3D	57.6
DS + DP	58.8
Ours w/o RANSAC	17.3
Ours	**5.8**

Table: ATE (deg / m)

Algorithm	Average
DS 3D	1.592 / 0.028
DS + DP	1.523 / 0.027
Zhou [3]	- / 0.024
Ours w/o RANSAC	1.467 / 0.026
Ours	**1.419 / 0.022**

[1] R. Ranftl et. al, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” in TPAMI 2022.

[2] T.-C. Dong-Si and A. I. Mourikis, “Estimator initial-ization in vision-aided inertial navigation with unknown camera-imu calibration,” in IROS 2012

[3] Y. Zhou et. al, “Learned Monocular Depth Priors in Visual-Inertial Initialization,” in ECCV 2022.

TUM-VI Results

- On the TUM-VI dataset, we tested initialization with 5 KFs and only a **300ms window**
- Found reasonable performance of MiDaS [1] network on fisheye images despite being trained on rectified

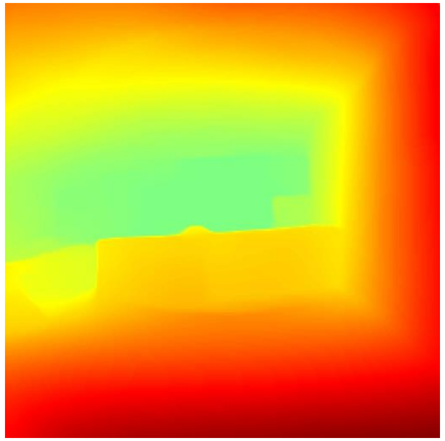
Table: Init window ATE and scale error (deg / m (%))

Algorithm	Average
DS 3D	**1.243** / 0.018 (9.20)
DS + DP	1.276 / 0.020 (8.73)
Ours	1.274 / **0.011** **(6.47)**

Table: VIO ATE using init conditions (deg / m)

Algorithm	Average
DS 3D	1.381 / 0.133
DS + DP	1.384 / 0.122
Ours	**1.214** / **0.059**

Example input and depth



[1] R. Ranftl et. al, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” in TPAMI 2022.

TUM-VI Results

- On the TUM-VI dataset, we tested initialization with 5 KFs and only a **300ms window**
- Found reasonable performance of MiDaS [1] network on fisheye images despite being trained on rectified

Example input and depth

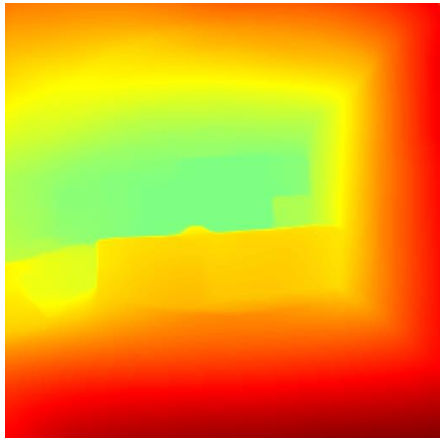


Table: Init window ATE and scale error (deg / m (%))

Algorithm	Average
DS 3D	**1.243** / 0.018 (9.20)
DS + DP	1.276 / 0.020 (8.73)
Ours	1.274 / **0.011** **(6.47)**

Table: VIO ATE using init conditions (deg / m)

Algorithm	Average
DS 3D	1.381 / 0.133
DS + DP	1.384 / 0.122
Ours	**1.214** / **0.059**

[1] R. Ranftl et. al, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” in TPAMI 2022.

TUM-VI Results

- On the TUM-VI dataset, we tested initialization with 5 KFs and only a **300ms window**
- Found reasonable performance of MiDaS [1] network on fisheye images despite being trained on rectified

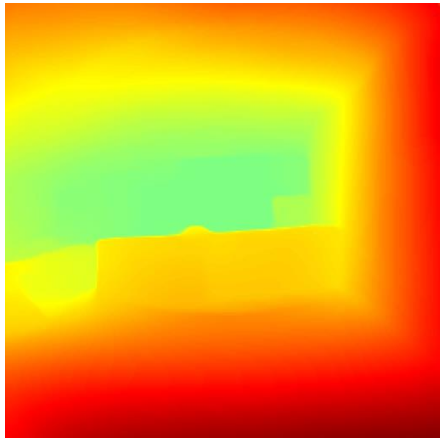
Table: Init window ATE and scale error (deg / m (%))

Algorithm	Average
DS 3D	**1.243** / 0.018 (9.20)
DS + DP	1.276 / 0.020 (8.73)
Ours	1.274 / **0.011 (6.47)**

Table: VIO ATE using init conditions (deg / m)

Algorithm	Average
DS 3D	1.381 / 0.133
DS + DP	1.384 / 0.122
Ours	**1.214 / 0.059**

Example input and depth



[1] R. Ranftl et. al, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” in TPAMI 2022.

Robustness to Outliers

- We tested our method's robustness to outlier measurements
- Added large random noise to the measurements for different outlier percentages

Table: Init window ATE (deg / m)

Outliers	Algorithm	Average
5%	DS 3D	1.257 / 0.017
	Ours w/o RANSAC	1.242 / **0.014**
	Ours	**1.047 / 0.014**
10%	DS 3D	1.280 / 0.016
	Ours w/o RANSAC	1.474 / 0.012
	Ours	**0.957 / 0.011**
25%	DS 3D	1.995 / 0.021
	Ours w/o RANSAC	2.413 / 0.025
	Ours	**1.409 / 0.014**
45%	DS 3D	2.929 / 0.035
	Ours w/o RANSAC	4.035 / 0.039
	Ours	**2.663 / 0.030**

Robustness to Outliers

- We tested our method's robustness to outlier measurements
- Added large random noise to the measurements for different outlier percentages
- Our method is **more robust to outliers** than the baseline

Table: Init window ATE (deg / m)

Outliers	Algorithm	Average
5%	DS 3D	1.257 / 0.017
	Ours w/o RANSAC	1.242 / **0.014**
	Ours	**1.047 / 0.014**
10%	DS 3D	1.280 / 0.016
	Ours w/o RANSAC	1.474 / 0.012
	Ours	**0.957 / 0.011**
25%	DS 3D	1.995 / 0.021
	Ours w/o RANSAC	2.413 / 0.025
	Ours	**1.409 / 0.014**
45%	DS 3D	2.929 / 0.035
	Ours w/o RANSAC	4.035 / 0.039
	Ours	**2.663 / 0.030**

Robustness to Tracking Failure

- We simulated tracking failure by reducing the number of features available to the VI-SfM

Table: % of successful initializations

Algorithm	60 feats	45 feats	30 feats	15 feats
DS 3D	81.25	17.50	33.75	2.50
DS 3D + DP	78.75	16.25	32.50	2.50
Ours w/o RANSAC	**100.00**	**98.75**	**97.50**	**55.00**
Ours	**100.00**	95.00	96.25	47.50

Robustness to Tracking Failure

- We simulated tracking failure by reducing the number of features available to the VI-SfM
- Our method is **more robust to tracking failure** than the baselines
  - Can initialize with *only 15 features*

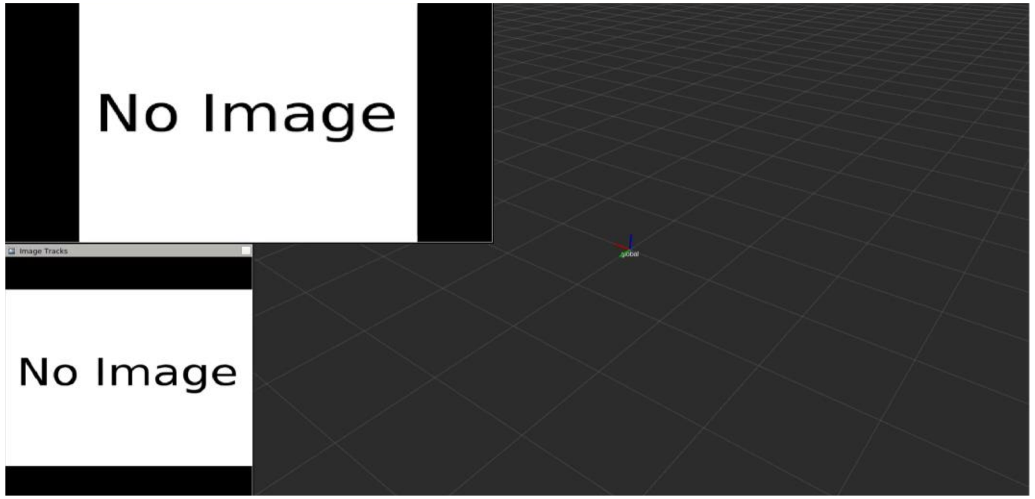
Table: % of successful initializations

Algorithm	60 feats	45 feats	30 feats	15 feats
DS 3D	81.25	17.50	33.75	2.50
DS 3D + DP	78.75	16.25	32.50	2.50
Ours w/o RANSAC	**100.00**	**98.75**	**97.50**	**55.00**
Ours	**100.00**	95.00	96.25	47.50

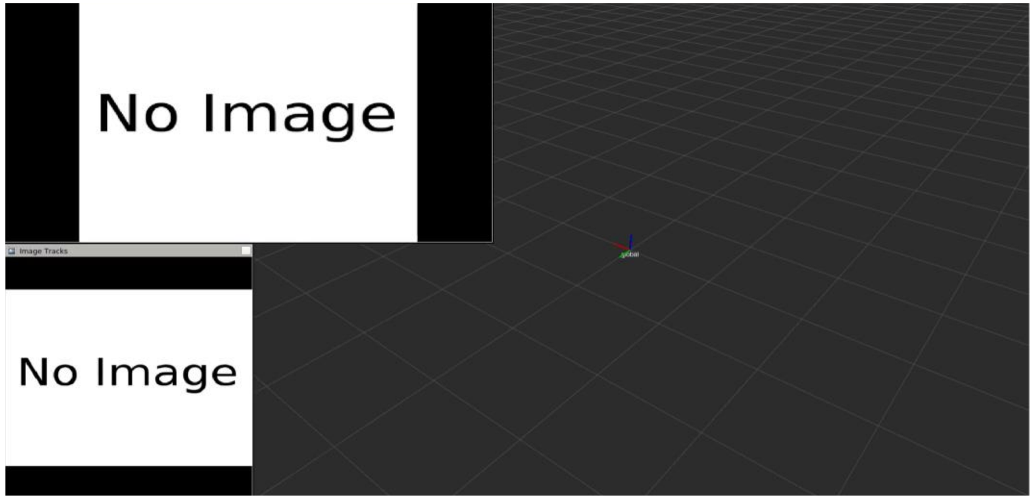
Robustness to Tracking Failure

- Our system can *quickly* initialize with only 15 features
- Baseline **fails** to initialize in reasonable amount of time

Baseline



Ours



Conclusion

- Proposed a new state-of-the-art mono visual-inertial initialization method
- Learned monocular depth is leveraged in linear init step
- Small linear system makes RANSAC practical
- Shown to outperform strong baselines

Nathaniel Merrill
nmerrill@udel.edu