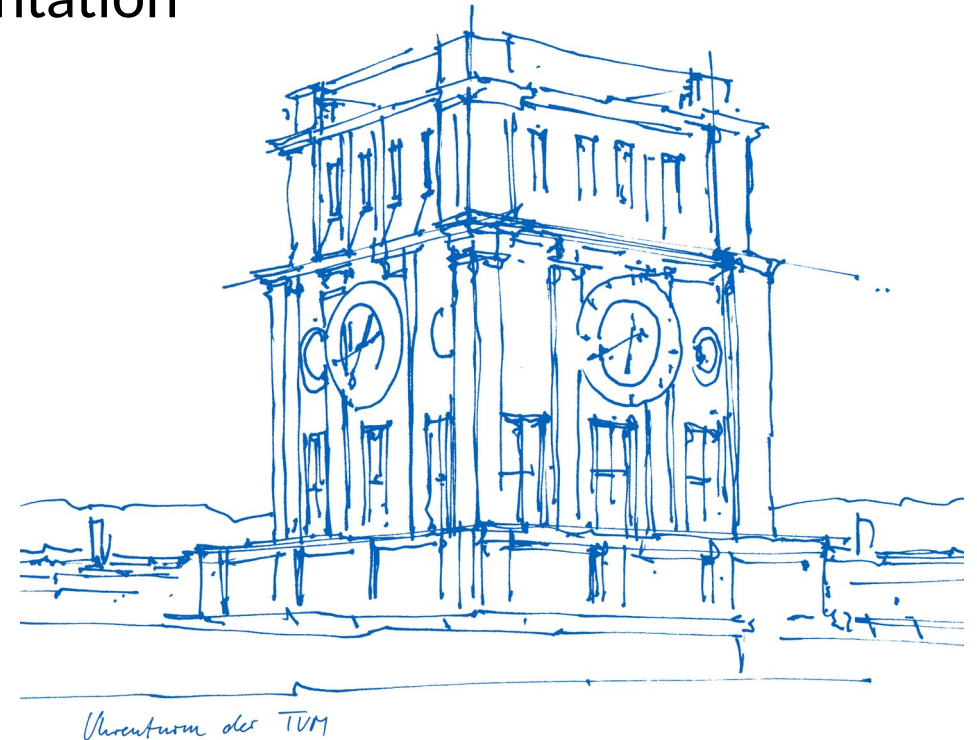


Deep Learning-based Keypoint Matching for Visual-Inertial Odometry

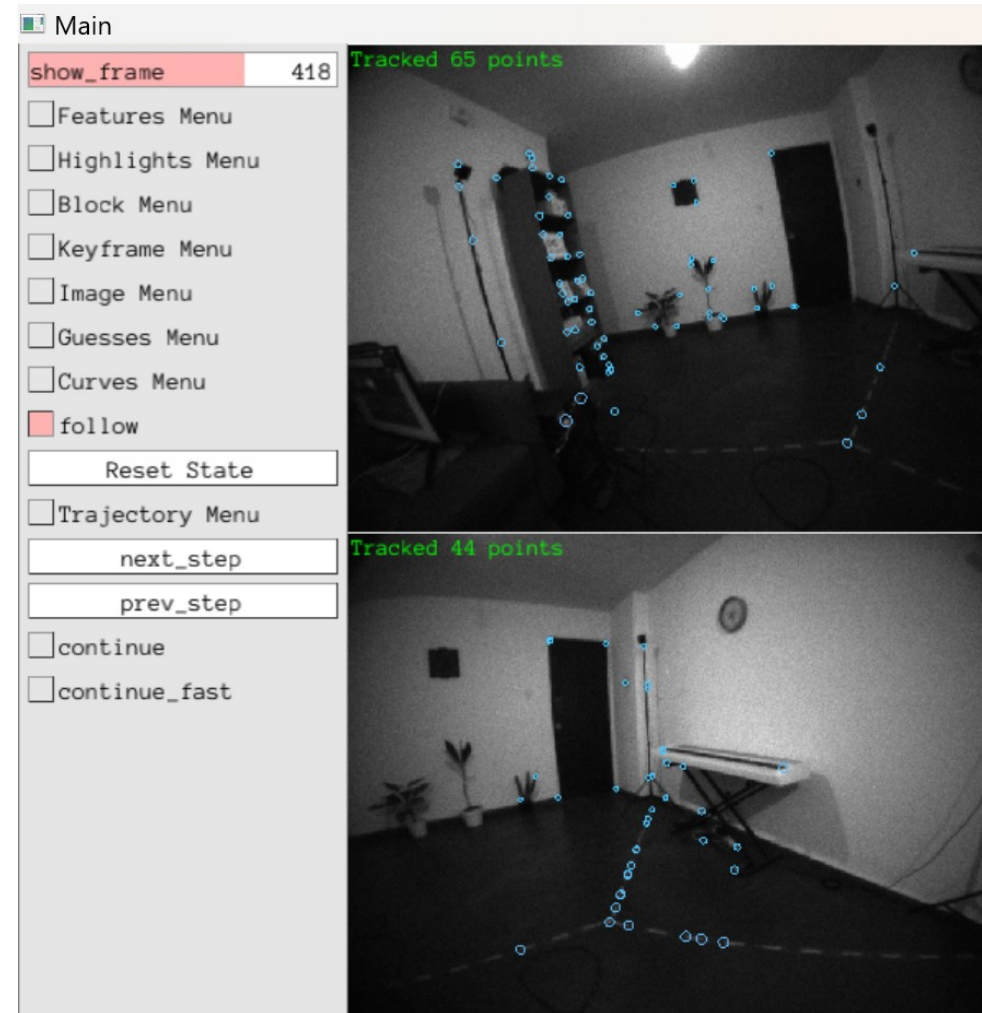
Master's Thesis Final Presentation

Pranav Athrapullikal

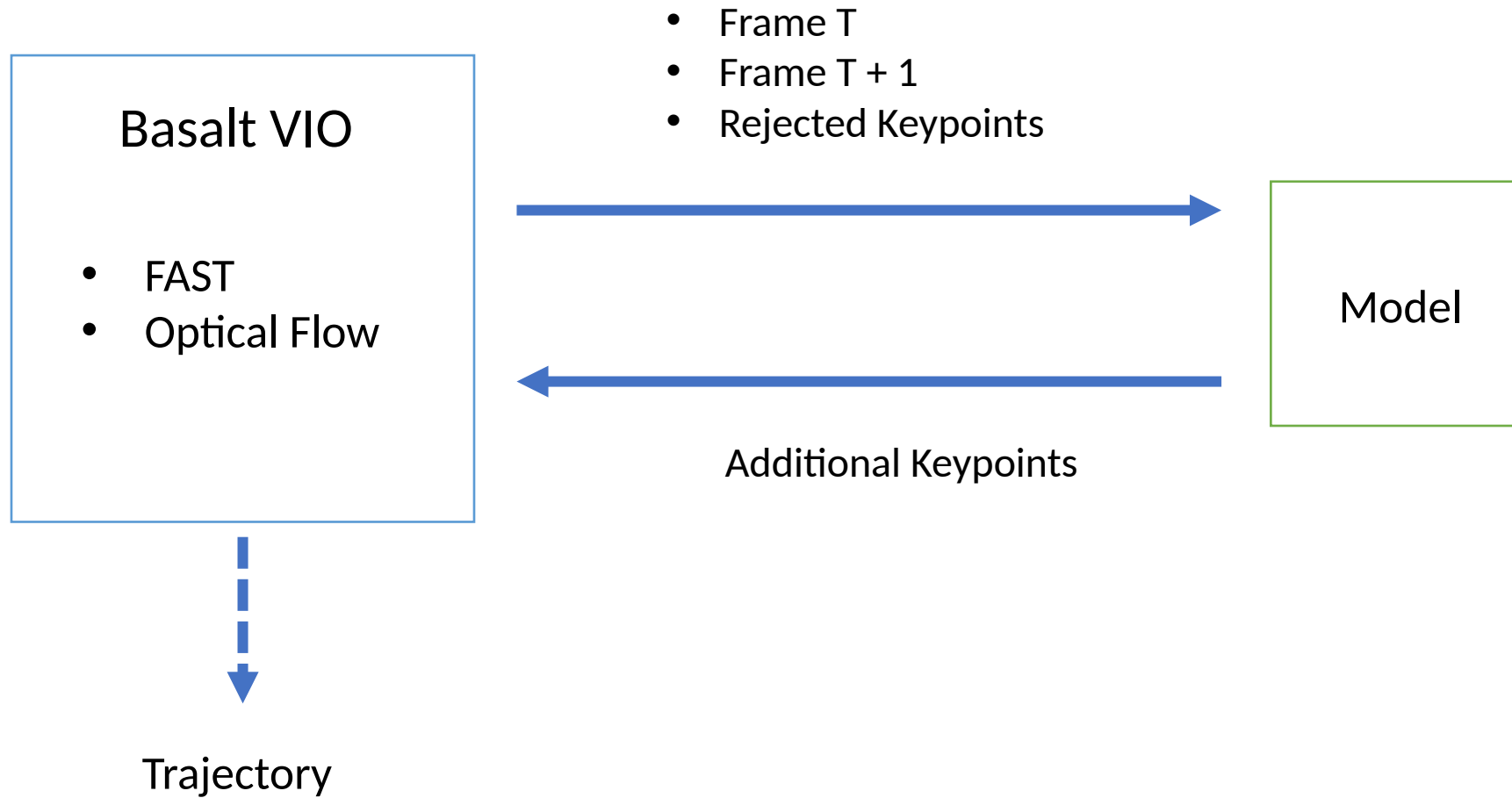


Purpose

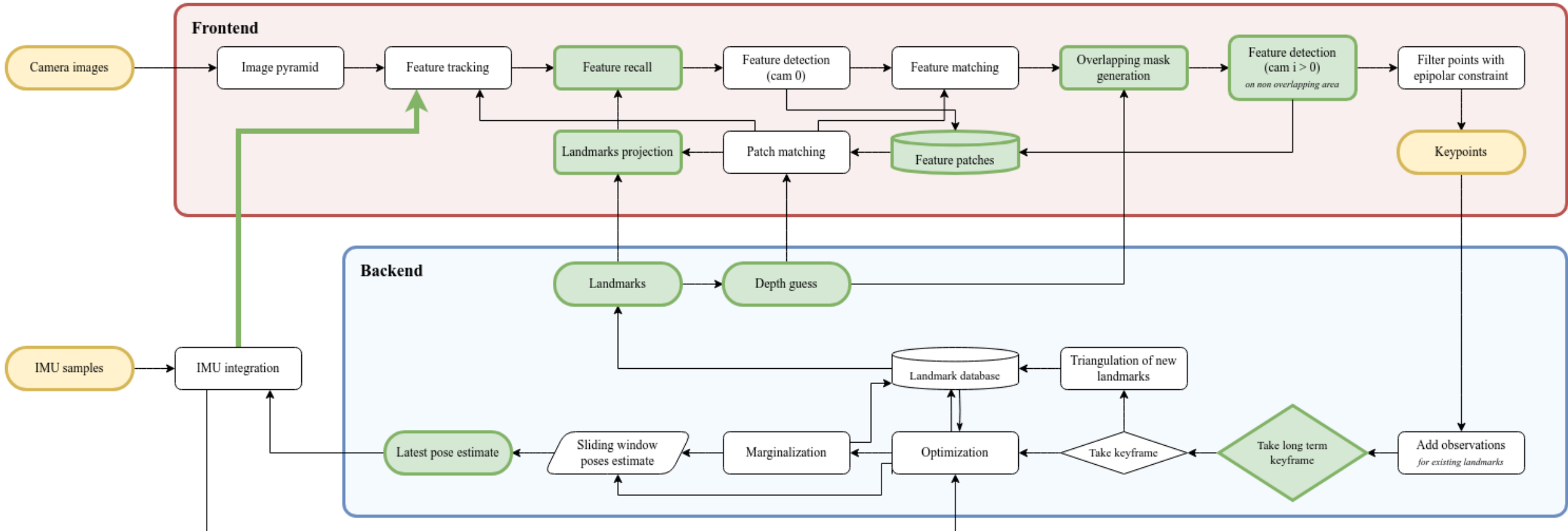
- Basalt
 - Visual-Inertial Odometry
 - Motion Estimation
- FAST / Optical Flow
- Rejected Keypoints
- ATE / RTE



Overview



Basalt



Credit: Mateo de Mayo

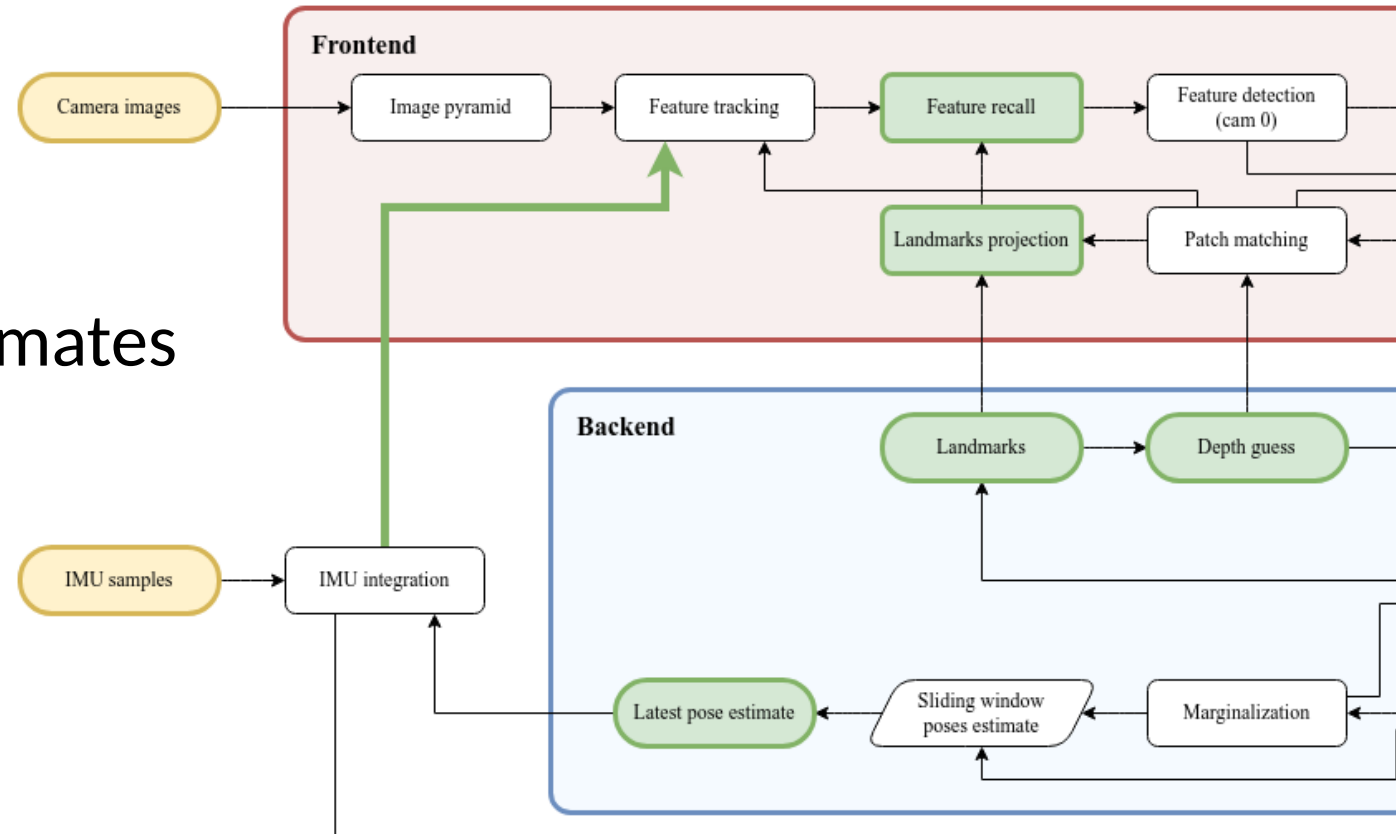
ATE / RTE

- Monado SLAM Dataset
 - Camera / IMU
- ATE
 - Evaluates global accuracy of trajectory w.r.t ground truth
- RTE
 - Evaluates local accuracy of trajectory w.r.t ground truth

	Optical Flow	
	ATE	RTE
MO001_hand_puncher_1	0.2753	0.0166
MO002_hand_puncher_2	0.2358	0.0154
MO003_hand_shooter_easy	0.1983	0.0142
MO004_hand_shooter_hard	0.0671	0.0098
MO005_inspect_easy	0.0187	0.0031
MO006_inspect_hard	0.0571	0.0084
MO 7 easy	0.0133	0.0034
MO 8 hard	0.0259	0.0111
MO 9 short up down	0.0037	0.0032
MO 10 short panorama	0.0100	0.0038
MO 11 short back forth	0.0186	0.0051
AVG	0.0840	0.0086
MEDIAN	0.0259	0.0084

Keypoint Estimates

- Feature Tracking
 - Optical Flow
 - Adjacent Frames T and $T + 1$
- Requires Good Keypoint Estimates in Frame $T + 1$
 - Uses IMU data after Frame T

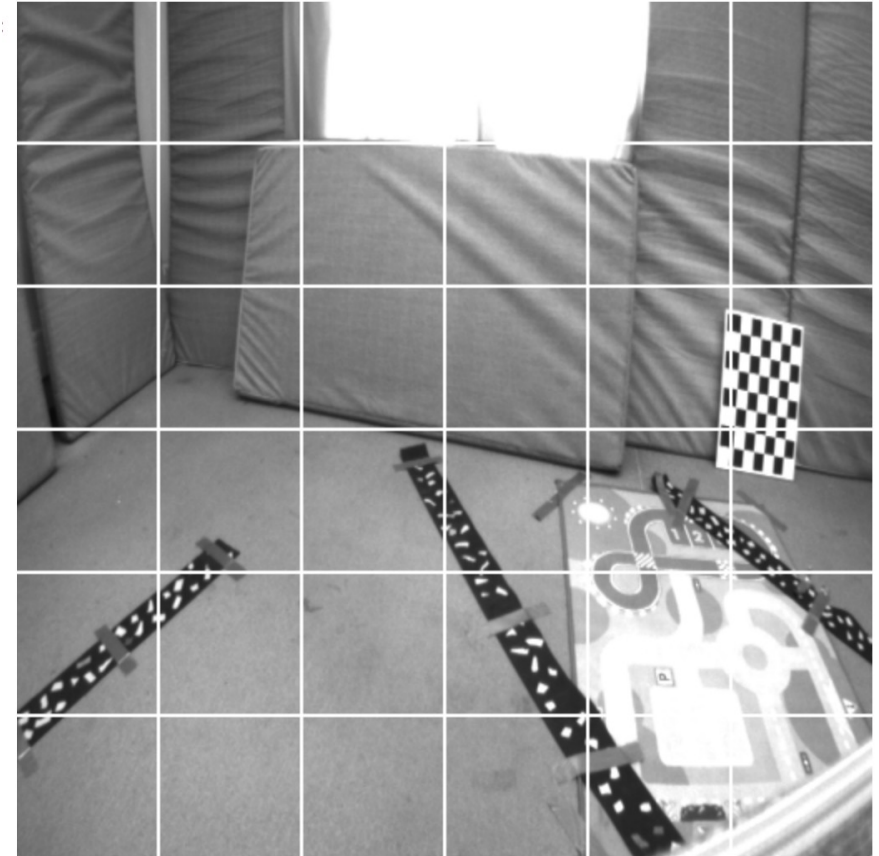


Using the Estimates as Predictions

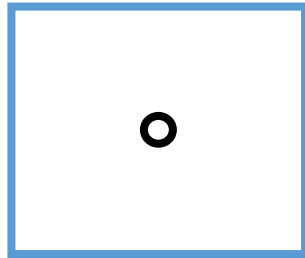
	Optical Flow	RoMa with p >= 0.8	Use Estimates as Coordinate Predictions	
			Filter by LT >= 3	Filter by p>=0.98 using Model R
		ATE		
MOO01_hand_puncher_1	0.2753	0.3930	-	0.24051
MOO02_hand_puncher_2	0.2358	0.3645	-	0.21628
MOO03_hand_shooter_easy	0.1983	0.2341	-	0.24046
MOO04_hand_shooter_hard	0.0671	0.1864	-	0.16991
MOO05_inspect_easy	0.0187	0.0808	-	0.01530
MOO06_inspect_hard	0.0571	0.0911	-	0.05286
MO 7 easy	0.0133	0.0293	-	0.01962
MO 8 hard	0.0259	0.0496	-	0.03636
MO 9 short up down	0.0037	0.0152	0.30405	0.00427
MO 10 short panorama	0.0100	0.0125	0.27350	0.00955
AVG	0.0905	0.1456		0.10051
MEDIAN	0.0415	0.0859		0.04461

Dataset Generation

- Monado SLAM Datasets
 - Images
 - IMU Data
- Basalt
 - Feature Detection Component
 - FAST
 - Cells / Grid
 - Store corners

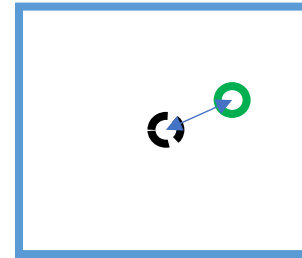


Dataset Generation



Reference Patch

- Reference keypoint from Basalt



Target Patch

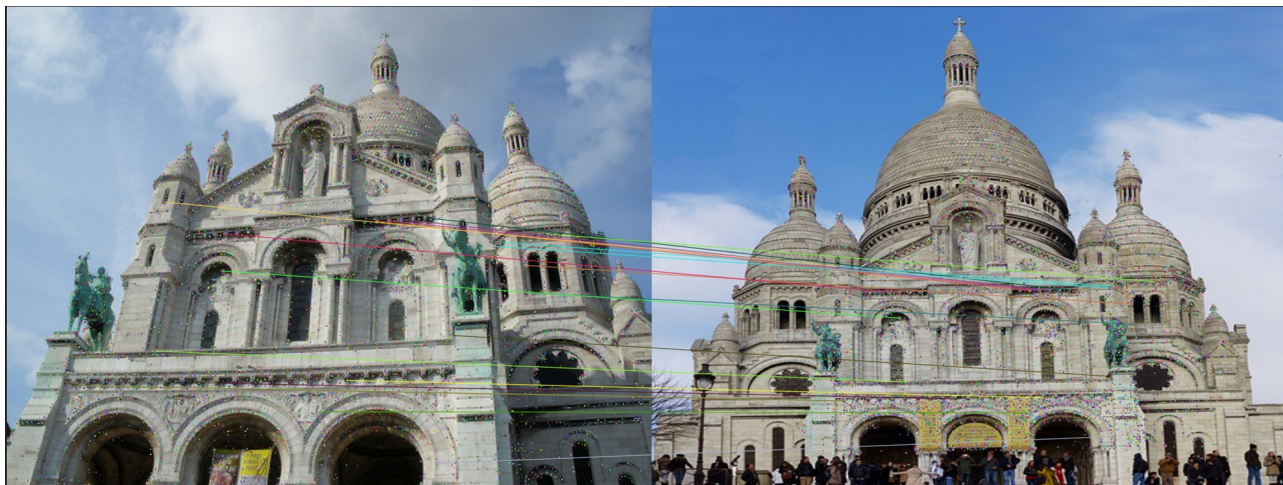
- Keypoint Estimate from Basalt
- True Keypoint Match from RoMa

Dataset Generation

- RoMa
 - Dense Matching
- Warp
 - Tensor $H \times W \times 4$
 - 2 x-y coordinates
- Confidence Map
 - Tensor $H \times W \times 1$
 - Confidence of each match



FAST + RoMa



- DeDoDe vs FAST
- Keypoint Distribution
- Cell Level Confidence Unreliable

Table 1. **SotA comparison on the Megadepth-1500-Unlimited benchmark.** We follow DeDoDe [12] and evaluate each detector with its optimal number of keypoints. In the case of SiLK and DeDoDe, we cap the number to 30k. Measured in AUC (higher is better).

Method ↓	AUC →	@5°	@10°	@20°
SuperPoint [9] CVPRW'18		31.7	46.8	60.1
DISK [30] NeurIPS'20		36.7	52.9	65.9
ALIKED [34] TIM'23		41.9	58.4	71.7
SiLK [14] ICCV'23		39.9	55.1	66.9
DeDoDe v1-L — v1-B [12] 3DV'24		49.4	65.5	77.7
DeDoDe v2-L — v1-B		52.5	67.4	78.7
DeDoDe C4-L — C4-B [7] CVPR'24		51	67	79
DeDoDe v2-L — C4-B		52.6	67.9	79.5
DeDoDe v1-L — v1-G [12] 3DV'24		52.8	69.7	82.0
DeDoDe v2-L — v1-G		54.6	70.7	82.4
DeDoDe v1-L — RoMa [13] CVPR'24		55.1	71.6	83.5
DeDoDe v2-L — RoMa		57.6	73.3	84.4

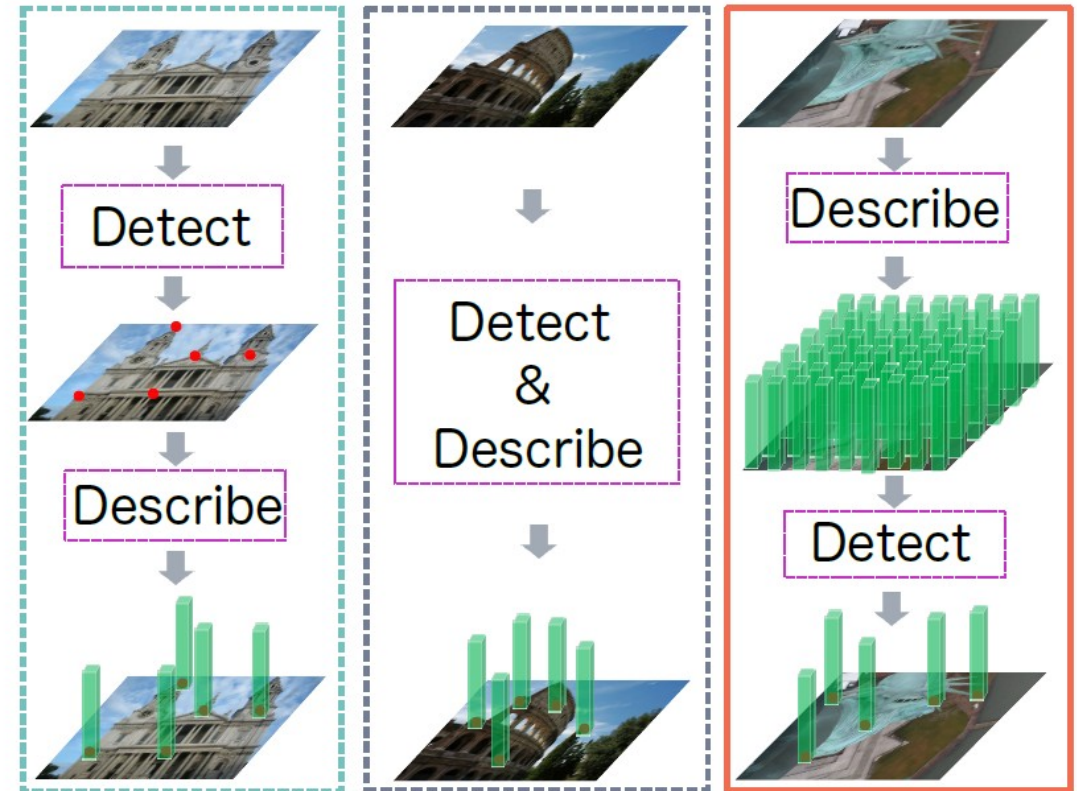
Feature Matchers

Is there a Detector?

- Detector-Based
 - Detect then Describe
 - Describe then Detect
 - Joint
- Detector Free

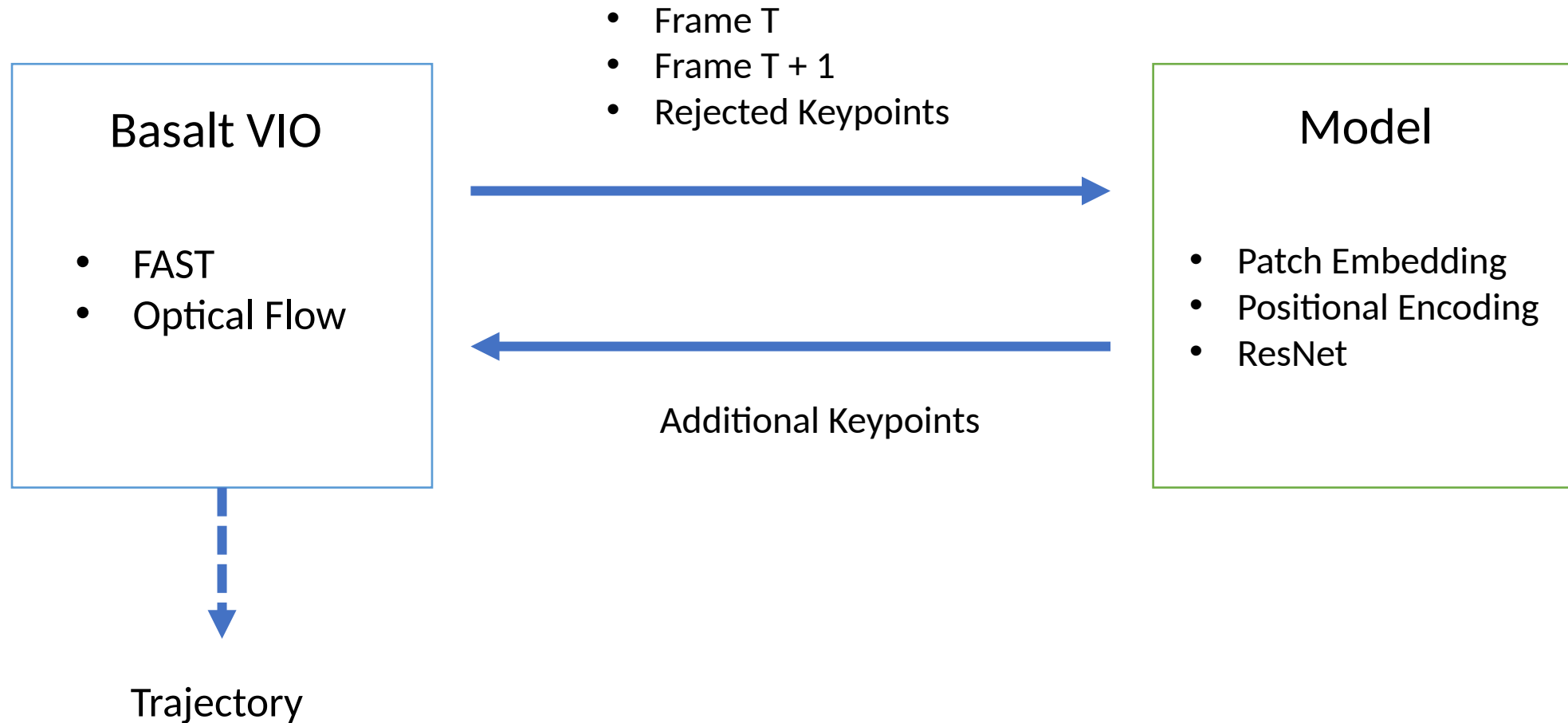
Few or Many Matches?

- Sparse Matching
- Dense Matching
- Semi-Dense Matching



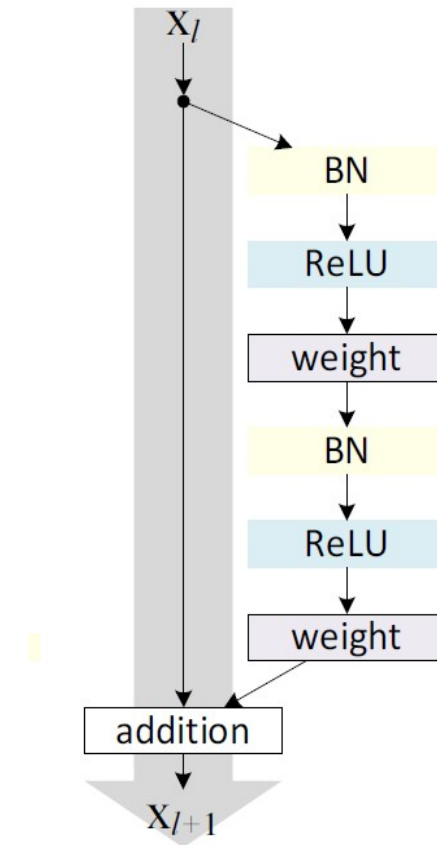
Source - Local Feature Matching Using Deep Learning: A Survey

Model



Model

- Outputs
 - Coordinate Offsets - $(\Delta x, \Delta y) \in [-1, 1]$ after tanh
 - Patch Level Confidence - $c \in [0, 1]$ after sigmoid
 - Point Level Confidence - $c \in [0, 1]$ after sigmoid
- ResNet
 - Basic Blocks
 - Pre Activation
 - Depth Separable Convolutions
- Patch Embedding using Convolution
- Positional Encoding



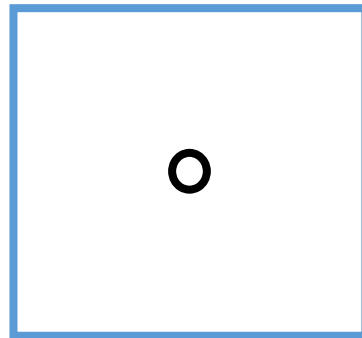
(e) full pre-activation

Training

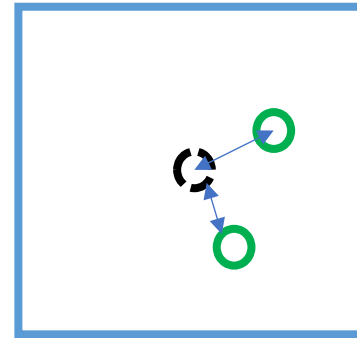
- Coordinate Offsets
 - x, y directions
 - Predicted and True Target patch keypoints
- Patch Level Confidence
 - BCE
 - Negative Samples
- Point Level Confidence
 - Use RoMa Confidence values ?
 - Image Context
 - Unreliable
 - Gaussian

Data Augmentation

- Blur
- Add Noise
- Perturb Target Patch



Reference Patch



Target Patch

Loss Function

	No Fallback		Fallback	
	Optical Flow	RoMa with p >= 0.8	Filter by LT >= 3 & p >=0.99	
			100	1000
	ATE			
MGO01_low_light	0.86800	0.85813	0.87422	0.871467
MGO02_hand_puncher	0.75280	0.51655	0.73844	0.733625
MGO03_hand_shooter_easy	0.17851	0.19091	0.18412	0.17792
MGO04_hand_shooter_hard	0.47272	0.74667	0.42358	0.43260
MGO05_inspect_easy	0.04564	0.09603	0.03720	0.03881
MGO06_inspect_hard	0.28920	0.27440	0.26541	0.27695
MGO07_mapping_easy	0.03981	0.04120	0.03021	0.034205
MGO08_mapping_hard	0.14307	0.13120	0.12440	0.153195
MGO09_short_1_updown	0.00952	0.00999	0.00959	0.009081
MGO10_short_2_panorama	0.01119	0.02034	0.01581	0.013251
MGO11_short_3_backandforth	0.01615	0.03209	0.01603	0.016062
AVG	0.25696	0.26523	0.24718	0.25065
MEDIAN	0.14307	0.13120	0.12440	0.15320

Sample Predictions

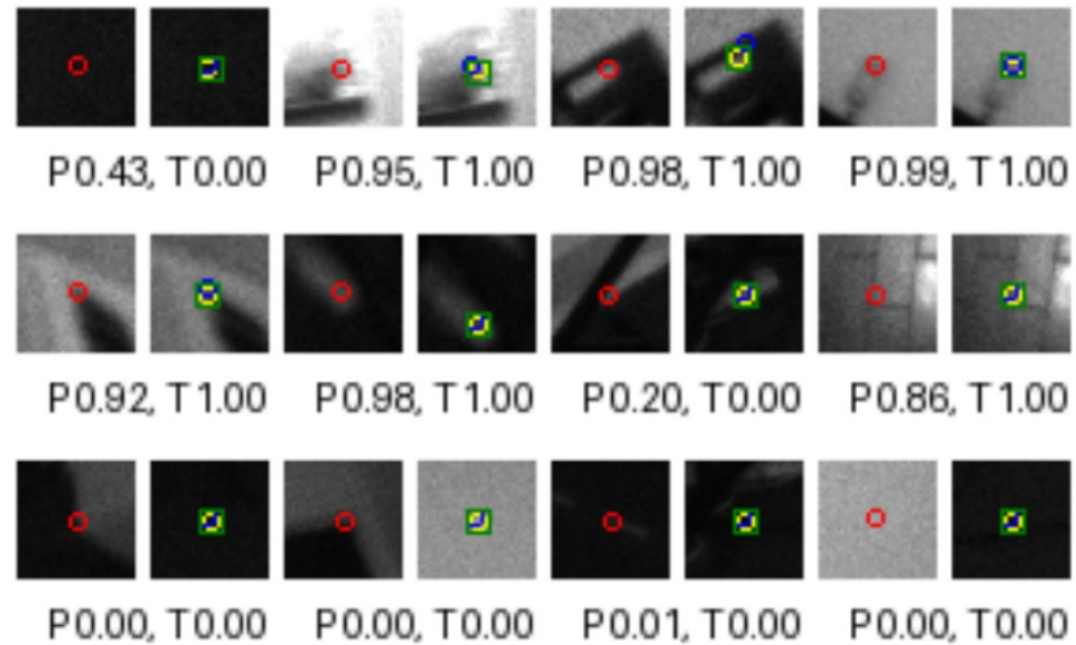


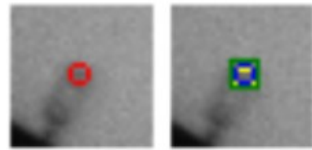
Figure 3.2: Prediction Samples from Validation Set

Thresholds

	No Fallback		Fallback			
	Optical Flow	RoMa [p >=0.8]	Filter by p >= 0.98	Filter by LT >= 3 & p >= 0.9	Filter by LT >= 3 & p >= 0.94	Filter by LT >= 3 & p >= 0.98
	ATE					
MGO01_low_light	0.86800	0.85813	0.87909	1.00282	0.97311	0.90153
MGO02_hand_puncher	0.75280	0.51655	0.71161	0.73204	0.73271	0.72167
MGO03_hand_shooter_easy	0.17851	0.19091	0.20051	0.23283	0.22849	0.20005
MGO04_hand_shooter_hard	0.47272	0.74667	0.42051	0.53661	0.47207	0.44496
MGO05_inspect_easy	0.04564	0.09603	0.04774	0.08998	0.08185	0.04925
MGO06_inspect_hard	0.28920	0.27440	0.23313	0.28955	0.26284	0.23256
MGO07_mapping_easy	0.03981	0.04120	0.02945	0.04905	0.03970	0.02827
MGO08_mapping_hard	0.14307	0.13120	0.11622	0.08803	0.09649	0.10859
MGO09_short_1_updown	0.00952	0.00999	0.00925	0.00805	0.00864	0.00893
MGO10_short_2_panorama	0.01119	0.02034	0.01596	0.02255	0.01393	0.01224
MGO11_short_3_backandforth	0.01615	0.03209	0.01587	0.02727	0.02314	0.01704
AVG	0.25696	0.26523	0.24358	0.27989	0.26663	0.24774
MEDIAN	0.14307	0.13120	0.11622	0.08998	0.09649	0.10859

Confidence Outputs

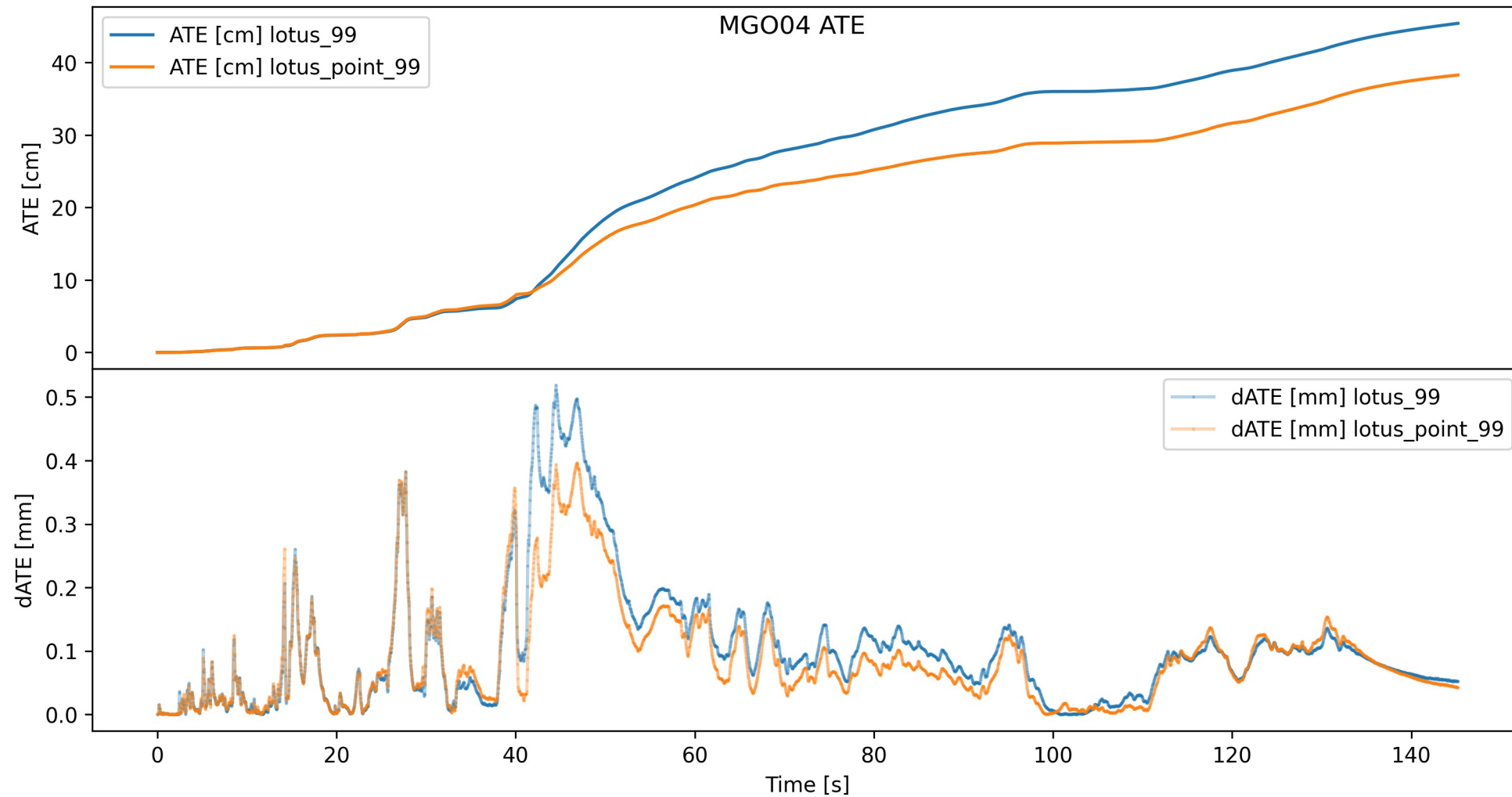
- Patch Level
- Point Level



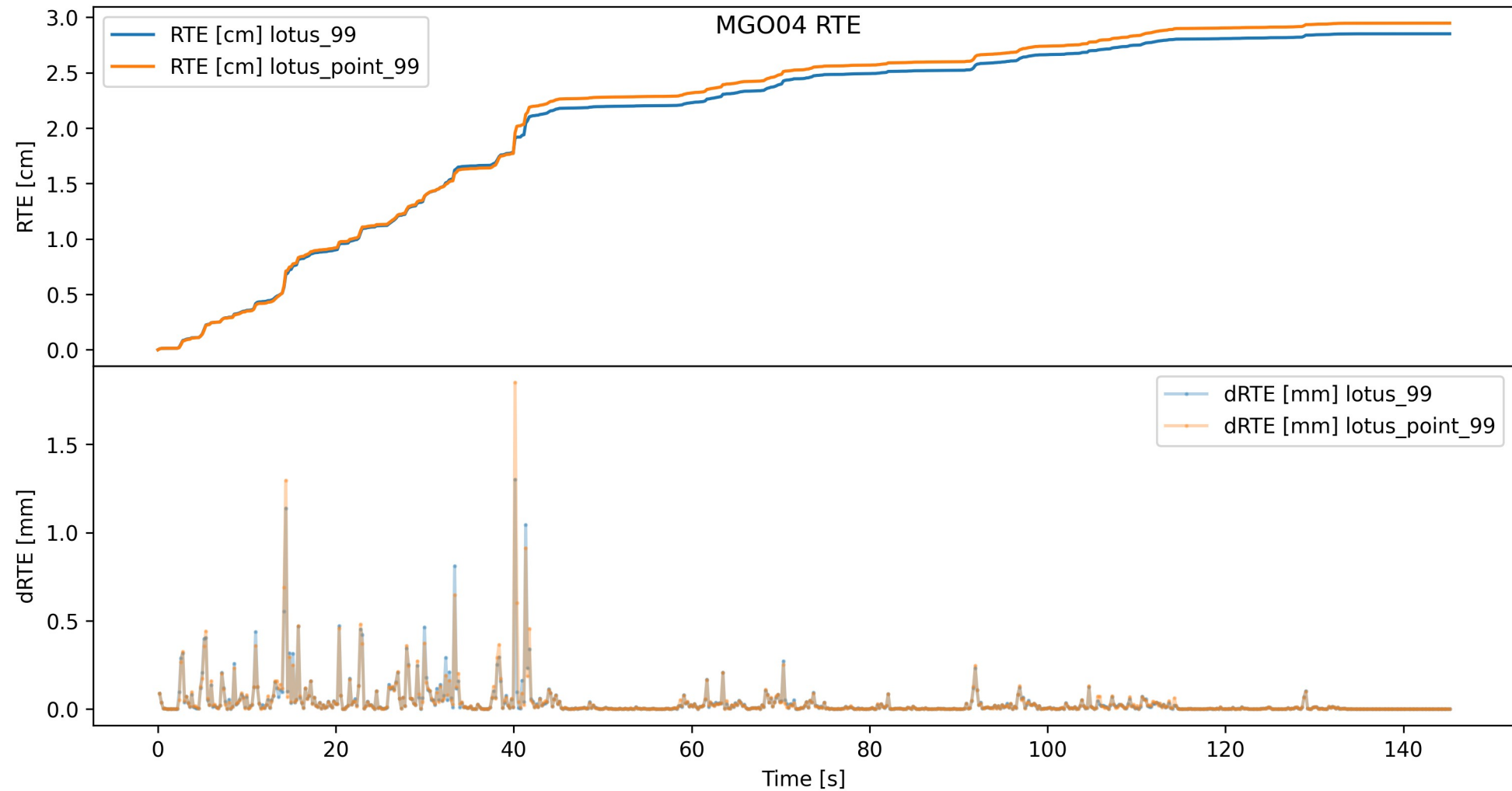
P0.99, T 1.00

	No Fallback		Fallback	
	Optical Flow	RoMa with p >= 0.8	Filter by LT >= 3 & p >=0.99	
			1 Confidence	2 Confidences
	ATE			
MGO01_low_light	0.86800	0.85813	0.87282	0.86800
MGO02_hand_puncher	0.75280	0.51655	0.67579	0.75305
MGO03_hand_shooter_easy	0.17851	0.19091	0.19613	0.16960
MGO04_hand_shooter_hard	0.47272	0.74667	0.38283	0.45425
MGO05_inspect_easy	0.04564	0.09603	0.04012	0.03889
MGO06_inspect_hard	0.28920	0.27440	0.25794	0.28956
MGO07_mapping_easy	0.03981	0.04120	0.03199	0.03851
MGO08_mapping_hard	0.14307	0.13120	0.13434	0.14141
MGO09_short_1_updown	0.00952	0.00999	0.009345	0.009693
MGO10_short_2_panorama	0.01119	0.02034	0.01623	0.01254
MGO11_short_3_backandforth	0.01615	0.03209	0.01598	0.01612
AVG	0.25696	0.26523	0.23941	0.25378
MEDIAN	0.14307	0.13120	0.13434	0.14141

ATE Graph



RTE Graph



Parameter Count

	No Fallback		Fallback	
	Optical Flow	RoMa with p >= 0.8	Filter by LT >= 3 & p >= 0.98	
			10 mil	2mil - DS Conv
	ATE			
MGO01_low_light	0.86800	0.85813	0.90153	0.88381
MGO02_hand_puncher	0.75280	0.51655	0.72167	0.74828
MGO03_hand_shooter_easy	0.17851	0.19091	0.20005	0.21067
MGO04_hand_shooter_hard	0.47272	0.74667	0.44496	0.444338
MGO05_inspect_easy	0.04564	0.09603	0.04925	0.046010
MGO06_inspect_hard	0.28920	0.27440	0.23256	0.225537
MGO07_mapping_easy	0.03981	0.04120	0.02827	0.031768
MGO08_mapping_hard	0.14307	0.13120	0.10859	0.117768
MGO09_short_1_updown	0.00952	0.00999	0.00893	0.009320
MGO10_short_2_panorama	0.01119	0.02034	0.01224	0.015800
MGO11_short_3_backandforth	0.01615	0.03209	0.01704	0.020490
AVG	0.25696	0.26523	0.24774	0.25035
MEDIAN	0.14307	0.13120	0.10859	0.11777

Positional Encoding

- None
- Sinusoidal
 - Split PE vector into 2 for height & width
 - Addition / Multiplication
- RoPE
 - Split image tensor into 2 for height & width
 - Apply rotations

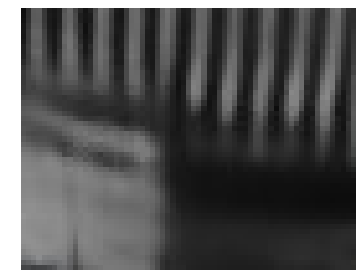
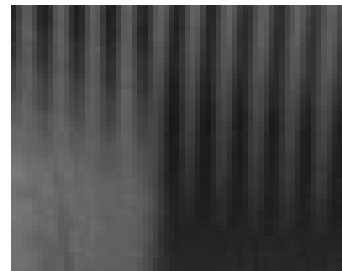
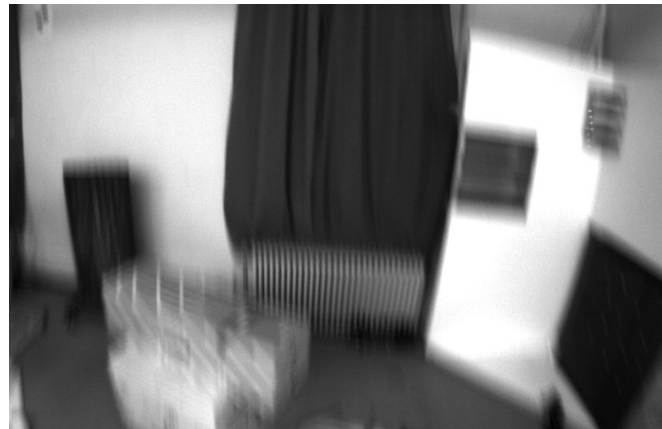
Positional Encoding



	No Fallback		Fallback			
	Optical Flow	RoMa with p >= 0.8	Filter by LT >= 3 & p >= 0.98			
			None	Add	Multiply	RoPE
			ATE			
MGO01_low_light	0.86800	0.85813	0.93381	0.92014	0.93028	0.90153
MGO02_hand_puncher	0.75280	0.51655	0.70788	0.71675	0.76251	0.72167
MGO03_hand_shooter_easy	0.17851	0.19091	0.23136	0.20370	0.240980	0.20005
MGO04_hand_shooter_hard	0.47272	0.74667	0.43764	0.39882	0.525507	0.44496
MGO05_inspect_easy	0.04564	0.09603	0.08980	0.06261	0.058639	0.04925
MGO06_inspect_hard	0.28920	0.27440	0.25159	0.23845	0.211415	0.23256
MGO07_mapping_easy	0.03981	0.04120	0.05199	0.03397	0.040679	0.02827
MGO08_mapping_hard	0.14307	0.13120	0.09465	0.12839	0.109286	0.10859
MGO09_short_1_updown	0.00952	0.00999	0.00967	0.00866	0.008646	0.00893
MGO10_short_2_panorama	0.01119	0.02034	0.01409	0.01576	0.013633	0.01224
MGO11_short_3_backandforth	0.01615	0.03209	0.02046	0.01730	0.019893	0.01704
AVG	0.25696	0.26523	0.25845	0.24951	0.26559	0.24774
MEDIAN	0.14307	0.13120	0.09465	0.12839	0.10929	0.10859

Mixed Effect of Deblurring

- NAFNet
- Artifacts
- Datasets
 - EuRoC
 - Monado SLAM



Test Metrics

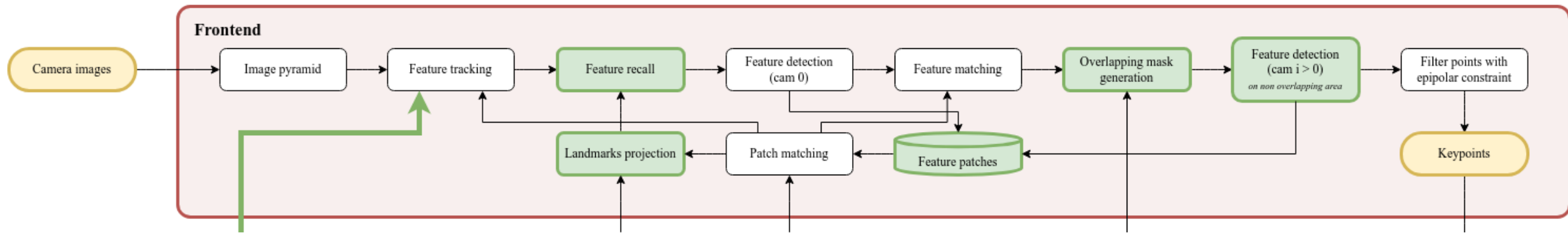
Table 5.1: Test Metrics Summary

Metric	Value
Confidence Accuracy	0.8499
Confidence F1 (0)	0.7326
Confidence F1 (1)	0.8813
Confidence Precision (0)	0.7920
Confidence Precision (1)	0.8655
Confidence Recall (0)	0.7119
Confidence Recall (1)	0.9067
Coordinates MAE	0.8618
PCK (0.5 pixels)	67.5367
PCK (1 pixels)	14.2504
PCK (1.25 pixels)	6.3636
PCK (1.5 pixels)	3.3172
PCK (2 pixels)	1.3175
PCK (3 pixels)	0.3812

	No Fallback	Fallback
	Optical Flow	Filter by LT ≥ 3 & $p \geq 0.99$ Model IV
	ATE	
AVG	0.20882	0.22068
MEDIAN	0.04900	0.04346
	RTE	
AVG	0.01423	0.01400
MEDIAN	0.00717	0.00654

Future Work

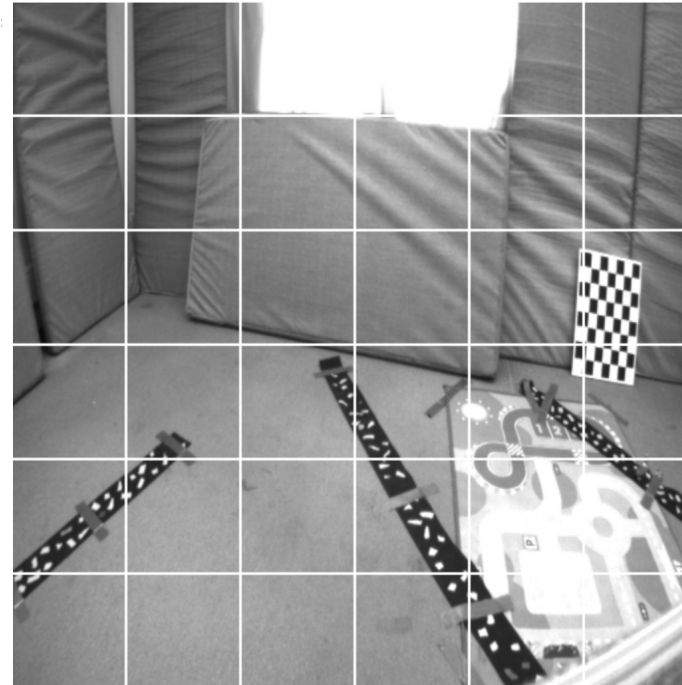
- Feature Matching Component



Credit: Mateo de Mayo

Future Work

- Dataset Generation
 - RoMa
 - MemFlow
 - ATE / RTE
- Model Size
- Offload to NPU



	Optical Flow		RoMa - $p \geq 0.8$		RoMa - Fallback	
	ATE	RTE	ATE	RTE	ATE	RTE
MO001_hand_puncher_1	0.2753	0.0166	0.3930	0.0181	0.25417	0.01658
MO002_hand_puncher_2	0.2358	0.0154	0.3645	0.0169	0.26431	0.01619
MO003_hand_shooter_easy	0.1983	0.0142	0.2341	0.0220	0.16271	0.01336
MO004_hand_shooter_hard	0.0671	0.0098	0.1864	0.0132	0.06595	0.00974
MO005_inspect_easy	0.0187	0.0031	0.0808	0.0068	0.01569	0.00315
MO006_inspect_hard	0.0571	0.0084	0.0911	0.0116	0.04604	0.00867
MO 7 easy	0.0133	0.0034	0.0293	0.0065	0.01571	0.00360
MO 8 hard	0.0259	0.0111	0.0496	0.0137	0.03045	0.01096
MO 9 short up down	0.0037	0.0032	0.0152	0.0067	0.00422	0.00354
MO 10 short panorama	0.0100	0.0038	0.0125	0.0067	0.01120	0.00443
MO 11 short back forth	0.0186	0.0051	0.0228	0.0080	0.01832	0.00525
AVG	0.0840	0.0086	0.1345	0.0118	0.08080	0.00868
MEDIAN	0.0259	0.0084	0.0808	0.0116	0.03045	0.00867

Conclusion

- Trained a CNN that uses depth-separable convolutions
- Operates on patches using Basalt's keypoint estimates
- Effect of positional encoding, deblurring and different confidence outputs
- Future Work – RoMa, Feature Matching Component