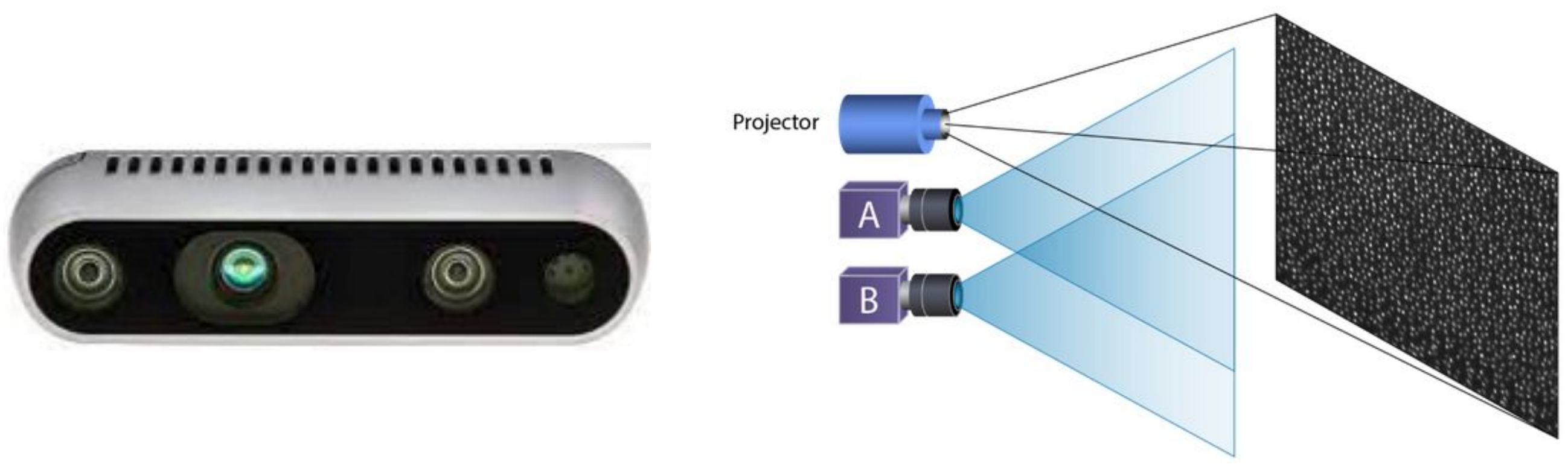
Self-Supervised Depth Completion for Active Stereo

[1]SLAMcore, London UK [2]Technical University of Denmark

Frederik Warburg[2], Daniel Hernandez-Juarez[1], Juan Tarrio[1], Alexander Vakhitov[1], Ujwal Bonde[1], Pablo F. Alcantarilla[1]

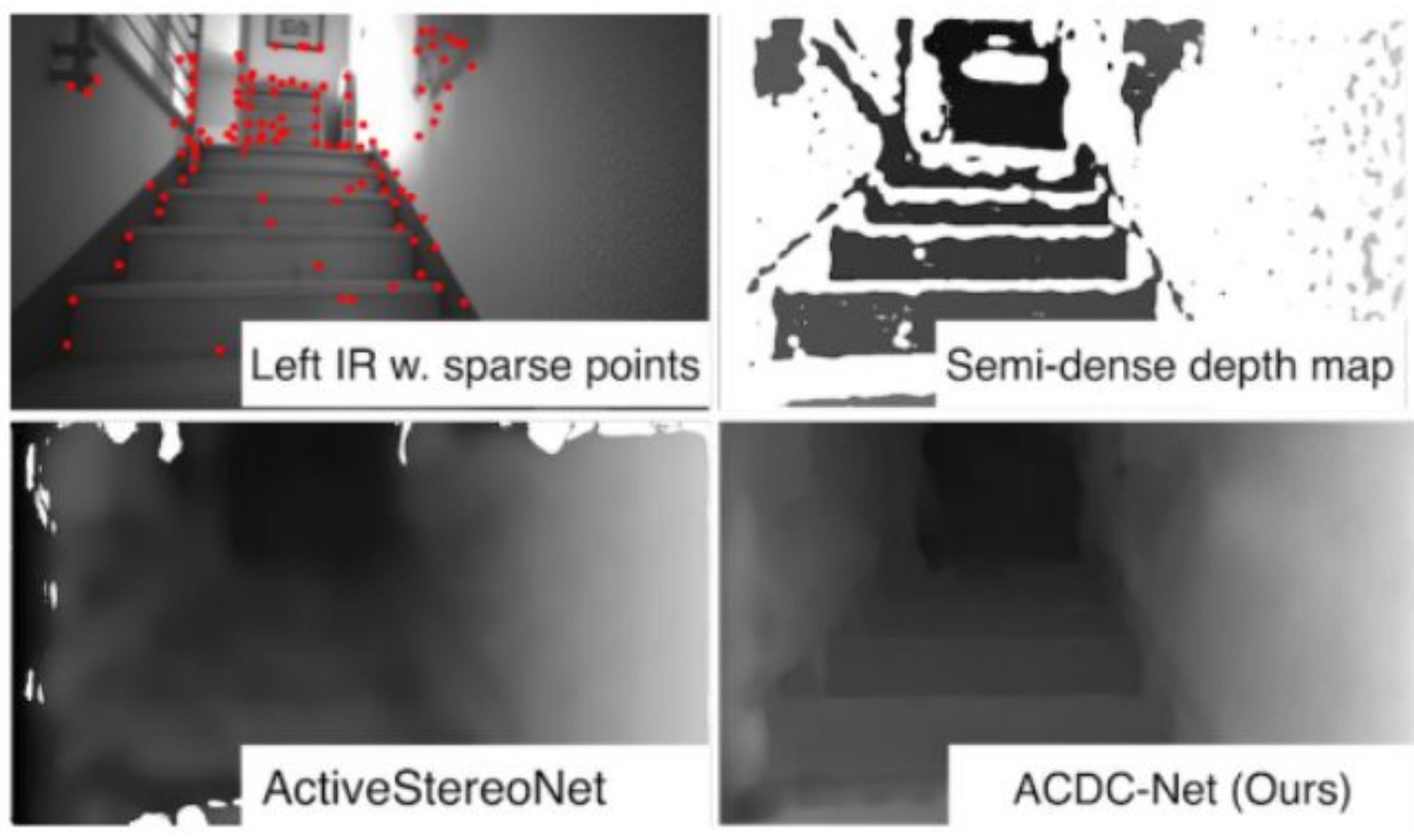
Introduction



- Active Stereo (AS) consists of a stereo pair of cameras that actively employs a light to simplify the problem
- AS suffers from artifacts and do not provide dense estimates

Self-supervised methods for active stereo	3D landmarks	Depth Sensor, e.g. LiDAR, RealSense	Images
Active Stereo Net [Y. Zhang et al, ECCV18]			✓
S2D [F. Ma et al, ICRA19] … and many more (especially supervised methods)		✓	✓
VOICED [Wong et al. ICRA20]	✓		✓
ACDC-net (ours)	✓	✓	✓

Goal



- Use depth completion with self-supervised learning to improve our depth estimates for robotics tasks (e.g. mapping and navigation)
- Self-supervised learning leverages all available data without labelling
- We leverage visual SLAM trajectory (loss) and keypoints (input and loss)

Quantitative results

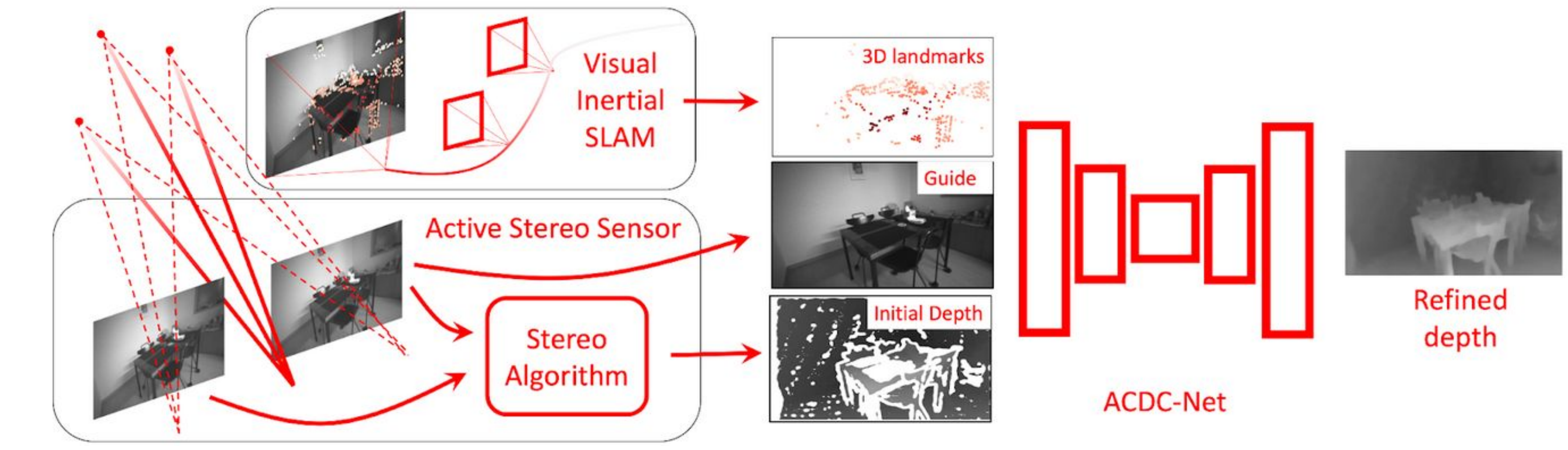
Method	Sup.	Result on whole image						With initial depth		W/O initial depth		time ↓ (ms)
		Rel. ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	%*val* ↑	Rel. ↓	RMSE ↓	Rel. ↓	RMSE ↓	
SGM [18]	✗	0.176	1.549	0.773	0.789	0.808	77.7	**0.023**	0.505	-	-	-
ELAS (Robotics) [14]	✗	0.120	1.483	0.861	0.875	0.885	77.7	0.070	1.086	0.337	2.759	-
Bilateral Solver [3]	✗	0.190	**0.568**	0.905	0.952	0.969	**100**	0.062	**0.393**	0.326	**0.880**	-
ActiveStereoNet [47]	✗	0.158	1.377	0.810	0.853	0.879	86.4	0.110	1.182	0.296	2.093	31
ACDC-Net-R18 (**ours**)	✗	0.130	1.049	0.875	0.954	0.977	**100**	0.096	0.875	0.215	1.616	**29**
ACDC-Net-R50 (**ours**)	✗	**0.087**	0.805	**0.932**	**0.964**	**0.979**	**100**	0.037	0.558	**0.174**	1.416	135
DMNet [35]	✓	0.110	1.217	0.846	0.933	0.968	100	0.103	1.170	0.144	1.532	38

We compare ACDC-Net with state of the art methods on both regions w./w.o initial depth estimates in Active TartanAir sequences

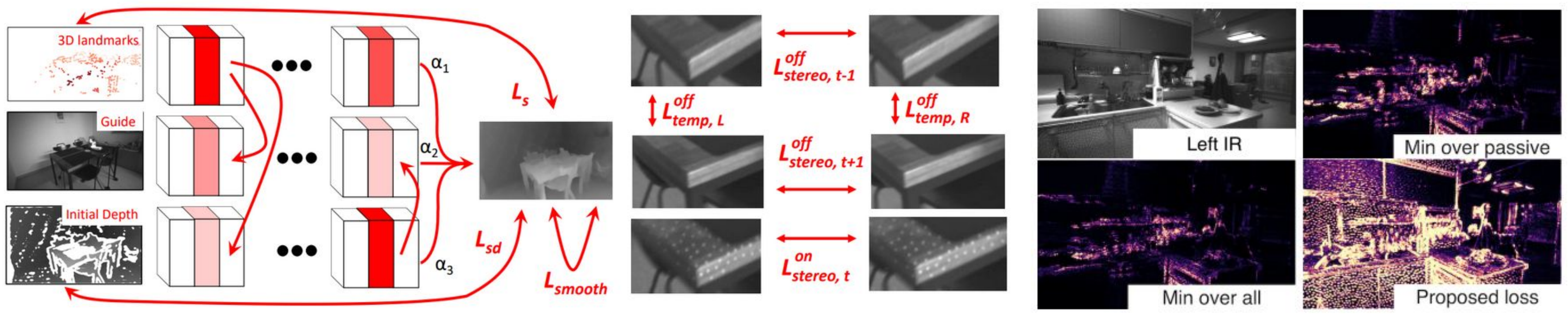
Method	S	C	A	Result on whole image						With initial depth		W/O initial depth	
				Rel. ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	%*val* ↑	Rel. ↓	RMSE ↓	Rel. ↓	RMSE ↓
SGM [18]	✓			0.236	0.957	0.719	0.736	0.748	57.8	0.051	0.297	-	-
ELAS (Robotics) [14]	✓			0.078	0.402	0.931	0.945	0.952	84.1	**0.045**	0.227	0.193	0.715
ActiveStereoNet [47]	✓		✓	0.123	0.538	0.903	0.957	0.973	96.4	0.066	0.261	0.308	0.997
Bilateral Solver [3]		✓		0.090	0.307	0.931	0.974	0.984	97.7	0.070	0.226	0.160	0.479
S2D-R34 [26]		✓	✓	0.383	1.008	0.326	0.507	0.667	**100**	0.360	0.982	0.407	1.074
Concat inputs (R50)		✓	✓	0.126	0.468	0.860	0.945	0.970	**100**	0.090	0.323	0.194	0.701
VOICED [44]		✓		0.194	0.569	0.737	0.874	0.934	98.7	0.179	0.482	0.239	0.761
ACDC-Net-R50 (**Mean**)		✓	✓	0.128	0.374	0.911	0.957	0.973	**100**	0.101	0.268	0.164	0.556
ACDC-Net-R18 (**ours**)		✓	✓	0.095	0.361	0.909	0.974	0.986	**100**	0.075	0.253	0.148	0.550
ACDC-Net-R50 (**ours**)		✓	✓	**0.075**	**0.290**	**0.945**	**0.981**	**0.988**	**100**	0.055	**0.184**	**0.117**	**0.460**

With a combination of inputs and complementary losses ACDC-Net-{R18,R50} outperforms competing methods on the D435i sequences. We benchmark against Stereo (S) and Completion (C) methods w./w.o. the Active pattern (A).

Method

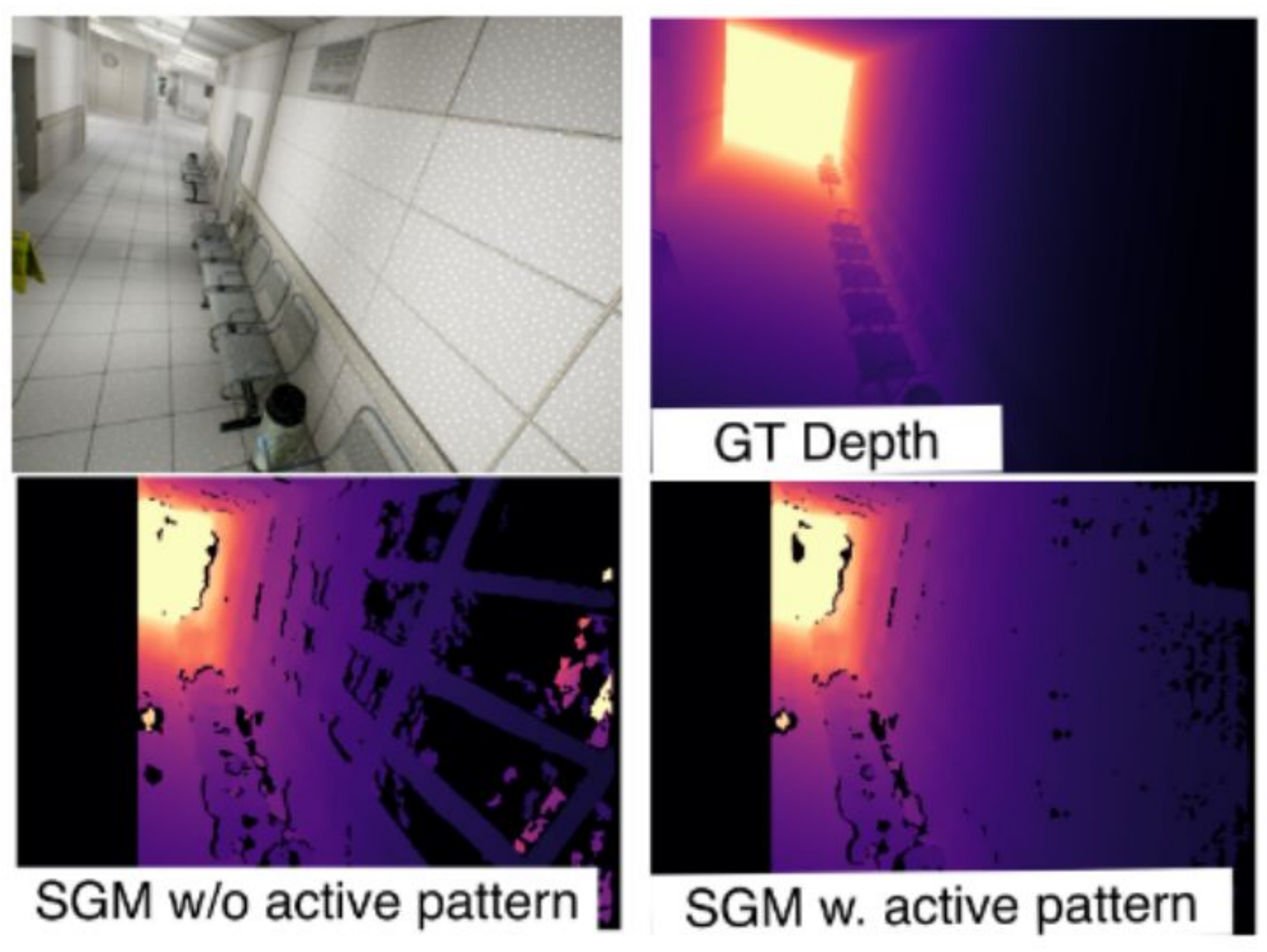


Adapt a channel exchanging arch. as backbone to fuse multi-modal input



Novel photometric loss for active and passive frames:
Minimum only over passive losses to remove occluded regions from temporal losses

Datasets

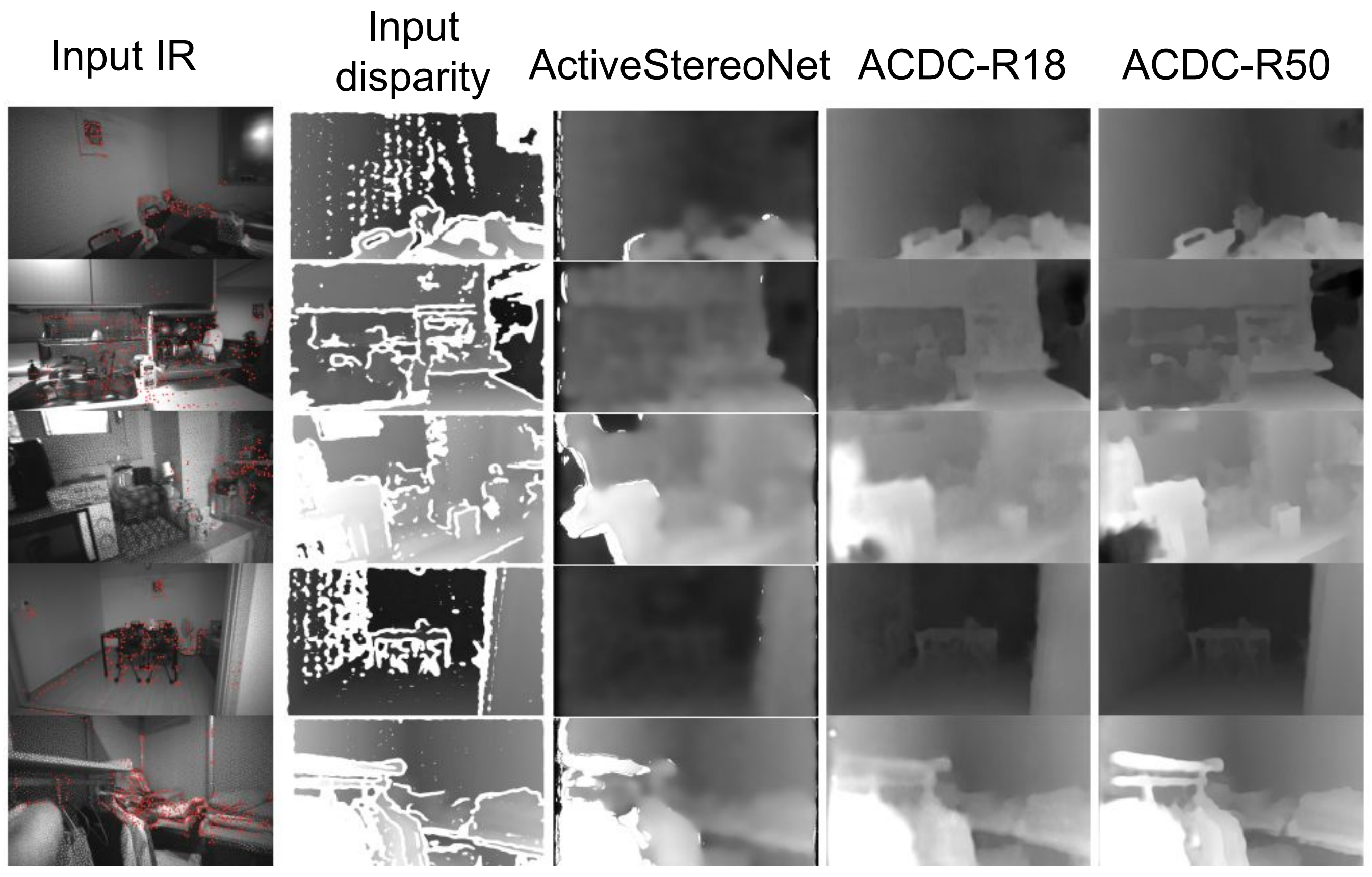


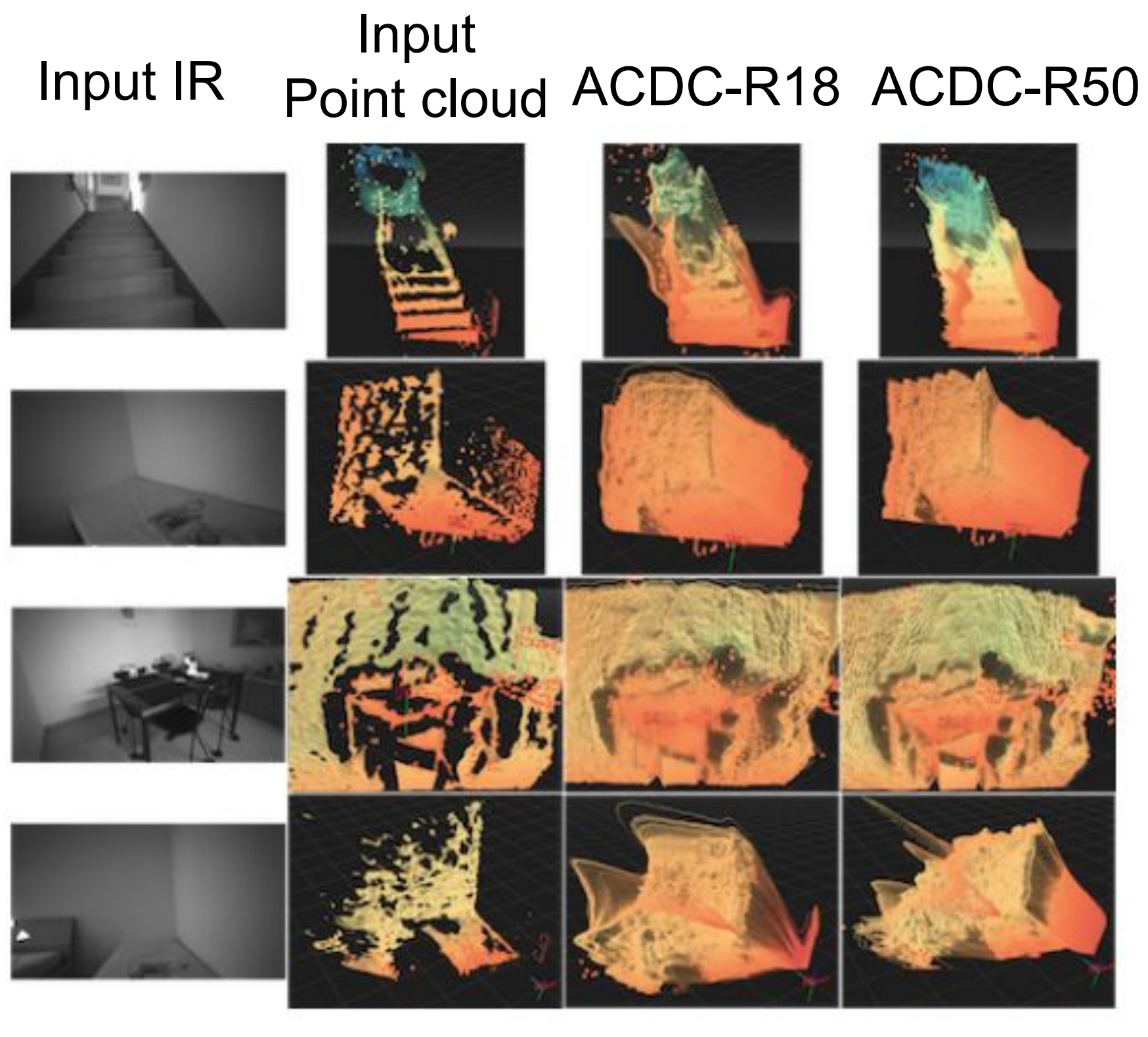
As there are no available active stereo datasets in the community, we release:

1) **RealSense** dataset: with initial stereo depth and infrared images

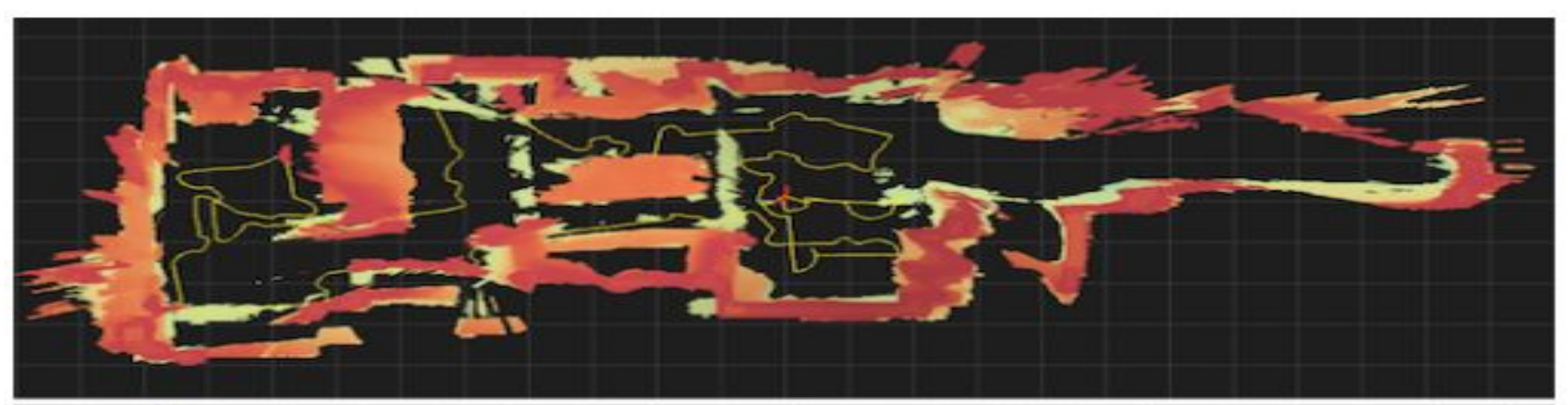
2) **Active TartanAir** dataset: we simulate the projected light and predicted SGM depth

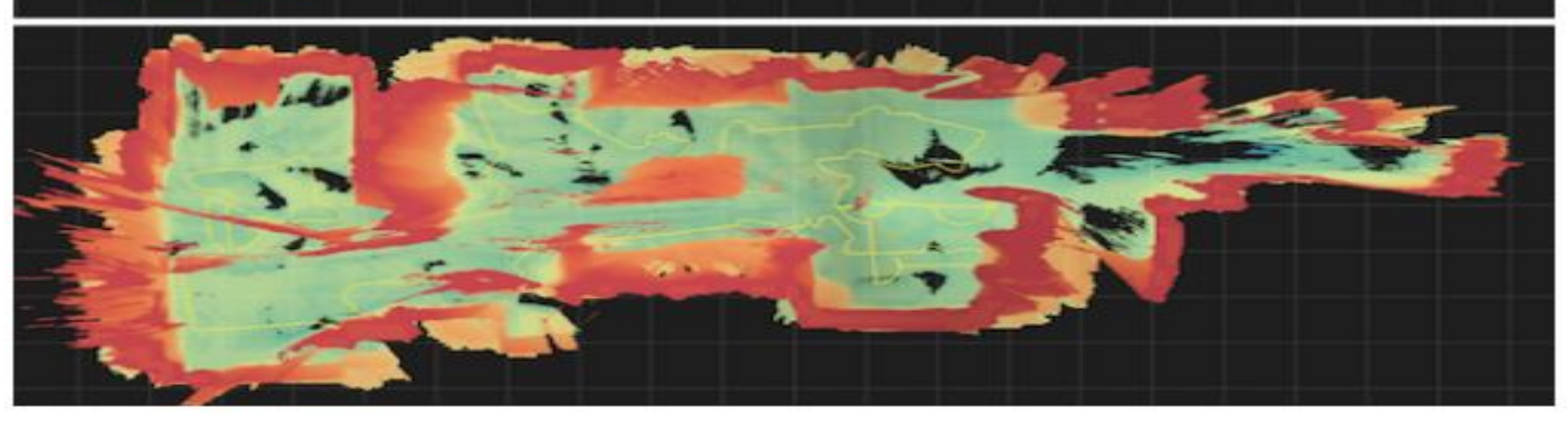
Qualitative results





3D reconstructions from raw D435i





3D reconstructions from completed ACDC-R50

Code & Datasets

SLAMCORE

ICRA 2022
IEEE International Conference on Robotics and Automation