ECCV'20 ONLINE

23-28 AUGUST 2020

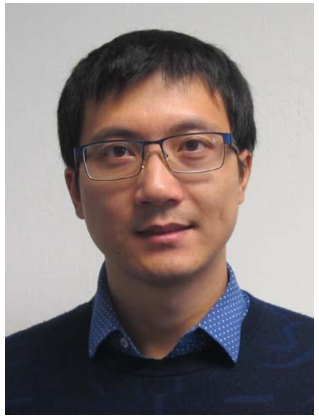
16TH EUROPEAN CONFERENCE ON

COMPUTER VISION

WWW.ECCV2020.EU

ECCV'20 ONLINE 23-28 AUGUST 2020

Pseudo RGB-D for Self-Improving Monocular SLAM and Depth Prediction



Lokender Tiwari | Pan Ji | Quoc-Huy Tran | Bingbing Zhuang | Saket Anand | Manmohan Chandraker

Presenter: Lokender Tiwari, Ph.D. Candidate at IIIT-Delhi

Project Page: https://lokender.github.io/self-improving-SLAM.html

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

NEC

NEC LABORATORIES AMERICA, INC.

Relentless passion for innovation

UC San Diego

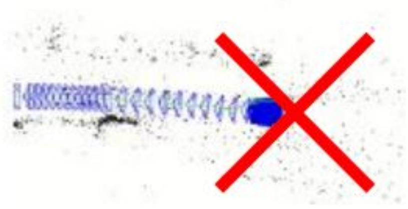
Outline

- Motivation
- Proposed Self-Improving Framework
- Experiments
- Analysis of Self-Improving Framework
- Conclusion

Motivation - Self-Improving Pseudo RGB-D SLAM

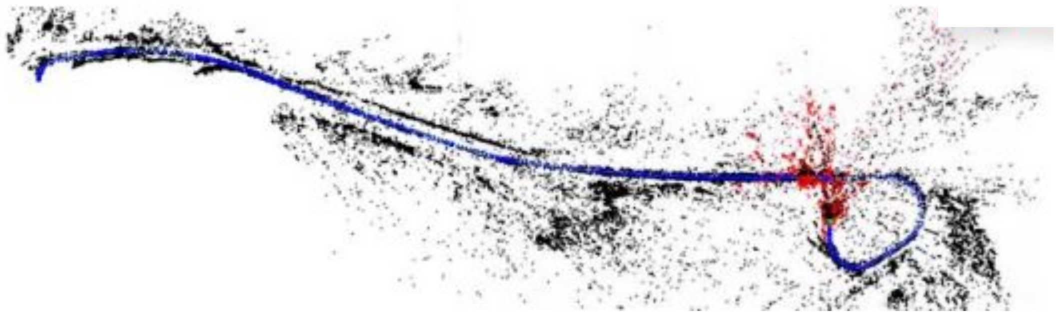
Geometric
Monocular
RGB SLAM

Tracking failure



RGB ORB-SLAM2 [1]
(KITTI Odometry Sequence 01)

Geometric
Monocular
RGB-D SLAM



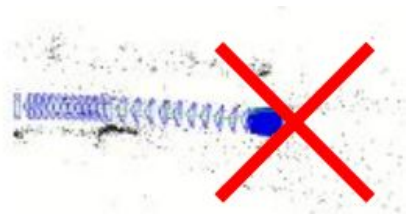
RGB-D ORB-SLAM2 [1]
(KITTI Odometry Sequence 01)

[1] Mur-Artal wt al."ORBSLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Transactions on Robotics* 2017

Motivation - Self-Improving Pseudo RGB-D SLAM

Geometric
Monocular
RGB SLAM

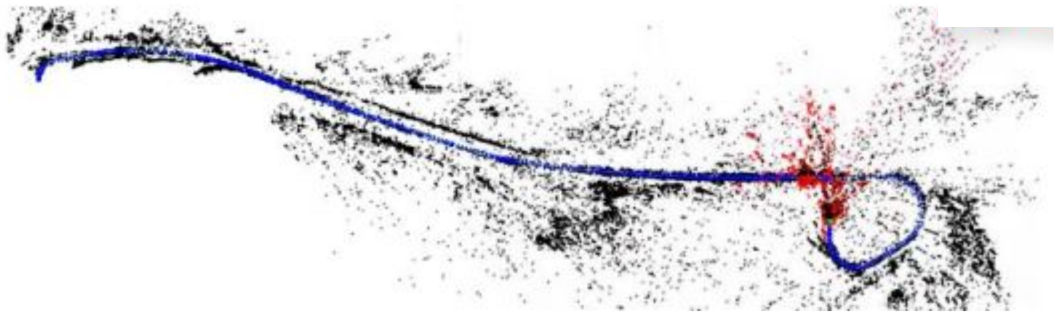
Tracking failure



RGB ORB-SLAM2 [1]
(KITTI Odometry Sequence 01)

Geometric
Monocular
RGB-D SLAM

Depth (D) from Active depth sensor (e.g LiDAR)



RGB-D ORB-SLAM2 [1]
(KITTI Odometry Sequence 01)

[1] Mur-Artal wt al."ORBSLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Transactions on Robotics* 2017

Motivation - Self-Improving Pseudo RGB-D SLAM

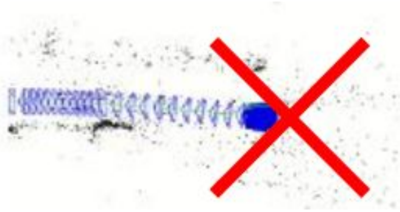
Geometric
Monocular
RGB SLAM

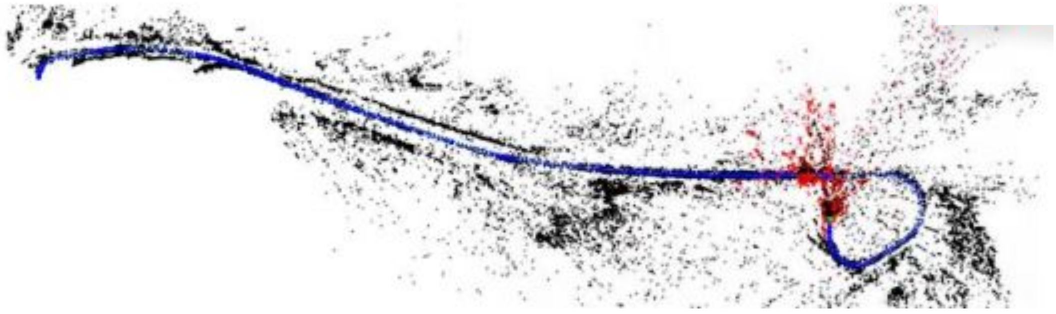
Pseudo Depth Sensor
Pseudo RGB-D SLAM

Geometric
Monocular
RGB-D SLAM

Depth (D) from
Active depth
sensor
(e.g LiDAR)

Tracking
failure





RGB ORB-SLAM2 [1]
(KITTI Odometry Sequence 01)

RGB-D ORB-SLAM2 [1]
(KITTI Odometry Sequence 01)

[1] Mur-Artal wt al."ORBSLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Transactions on Robotics* 2017

Motivation - Self-Improving Monocular Depth Prediction

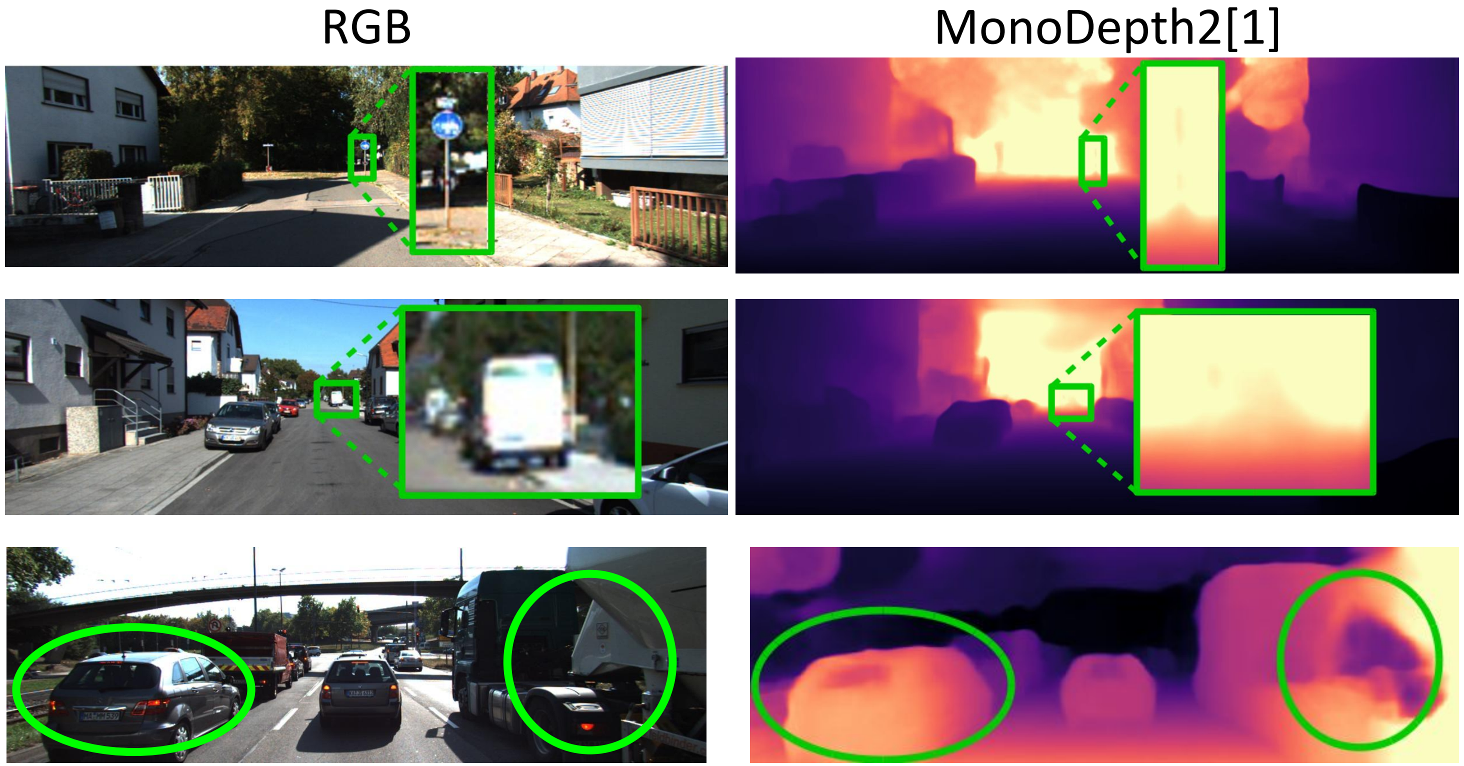
Unsupervised
CNN-Based
Monocular Depth
Prediction

Does not model:

- Photo changes
- Wide-baseline constraints (beyond 3-5 frames)
- ....

Motivation - Self-Improving Monocular Depth Prediction

Unsupervised CNN-Based Monocular Depth Prediction



Does not model:

- Photo changes
- Wide-baseline constraints (beyond 3-5 frames)
- ....

- Fails to predict accurate depths (especially for farther points)

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*

Motivation

Geometric
Monocular
RGB-SLAM

Unsupervised
CNN-Based
Monocular Depth
Prediction

Motivation

Geometric
Monocular
RGB-SLAM

Suffers from:

- Pure Rotational Motion
- Scale ambiguity/drift
- ...

Unsupervised
CNN-Based
Monocular Depth
Prediction

Does not model:

- Photo changes
- Wide-baseline constraints (beyond 3-5 frames)
- ....

Motivation

Geometric Monocular RGB-SLAM

Suffers from:

- Pure Rotational Motion
- Scale ambiguity/drift
- ...

Unsupervised CNN-Based Monocular Depth Prediction

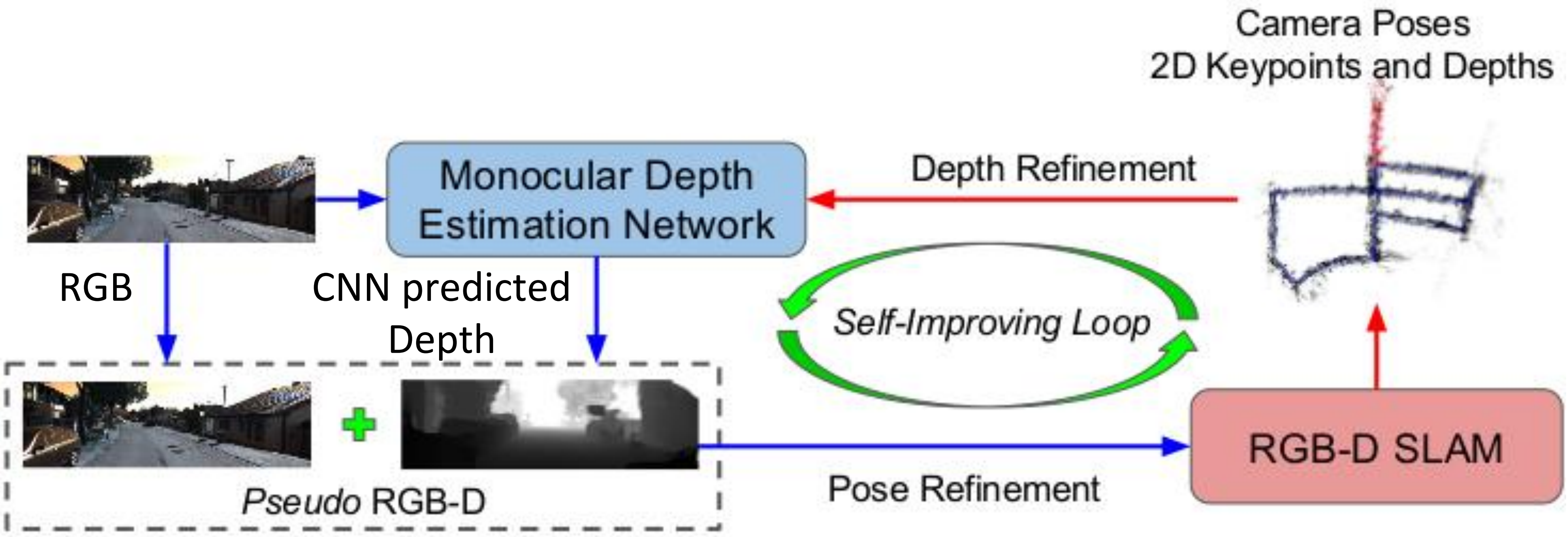
Does not model:

- Photo changes
- Wide-baseline constraints (beyond 3-5 frames)
- ....

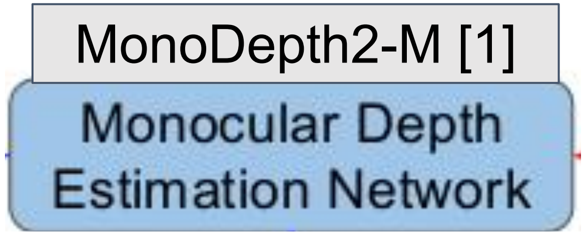
geometric-CNN framework

We propose a Self-Supervised, Self-Improving framework.

A Self-Supervised, Self-Improving Framework



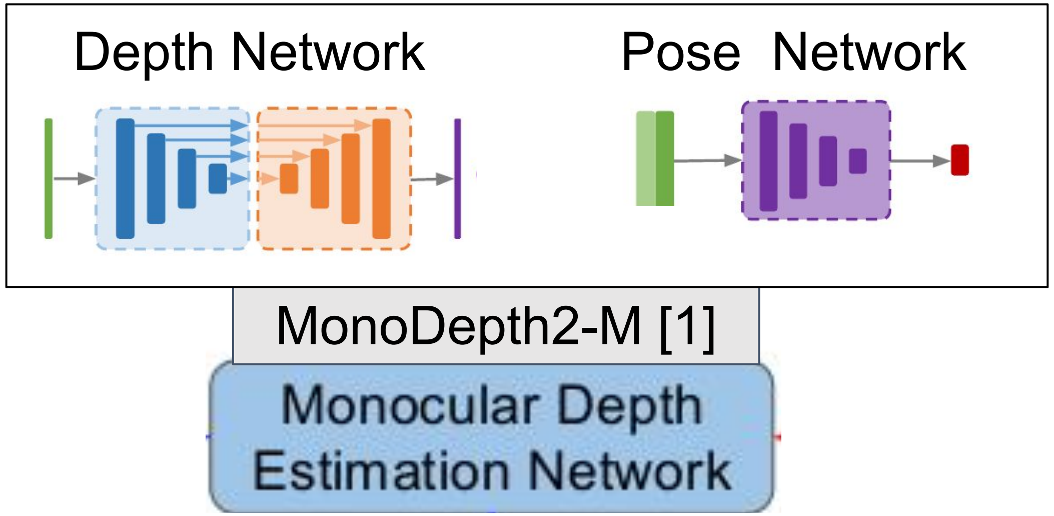
A Self-Supervised, Self-Improving Framework



- Base Unsupervised Monocular Depth Network: **MonoDepth2-M [1]**

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*

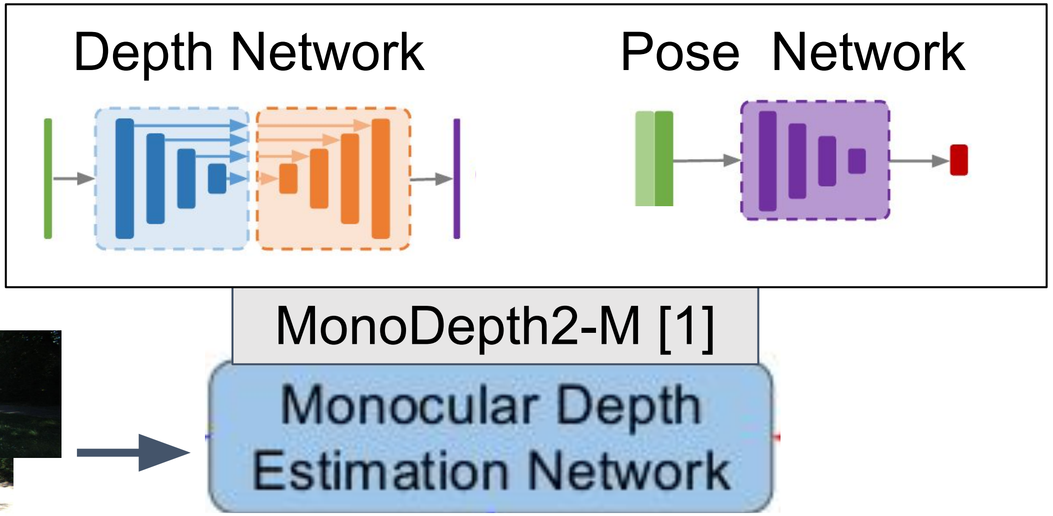
A Self-Supervised, Self-Improving Framework



- Base Unsupervised Monocular Depth Network: **MonoDepth2-M [1]**

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*

A Self-Supervised, Self-Improving Framework

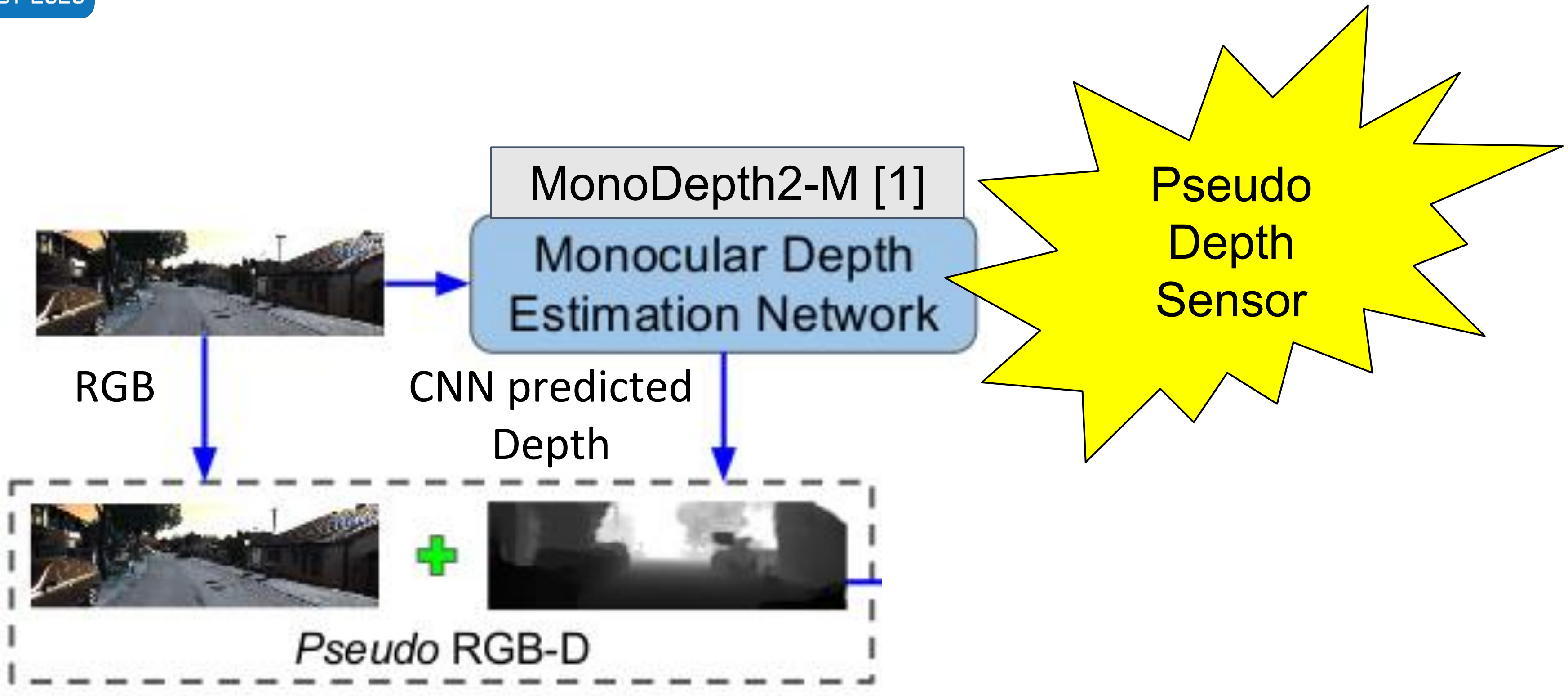


Monocular
Video/Images

- Base Unsupervised Monocular Depth Network: **MonoDepth2-M [1]**
- Train MonoDepth2-M using monocular videos in a complete unsupervised manner.

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*

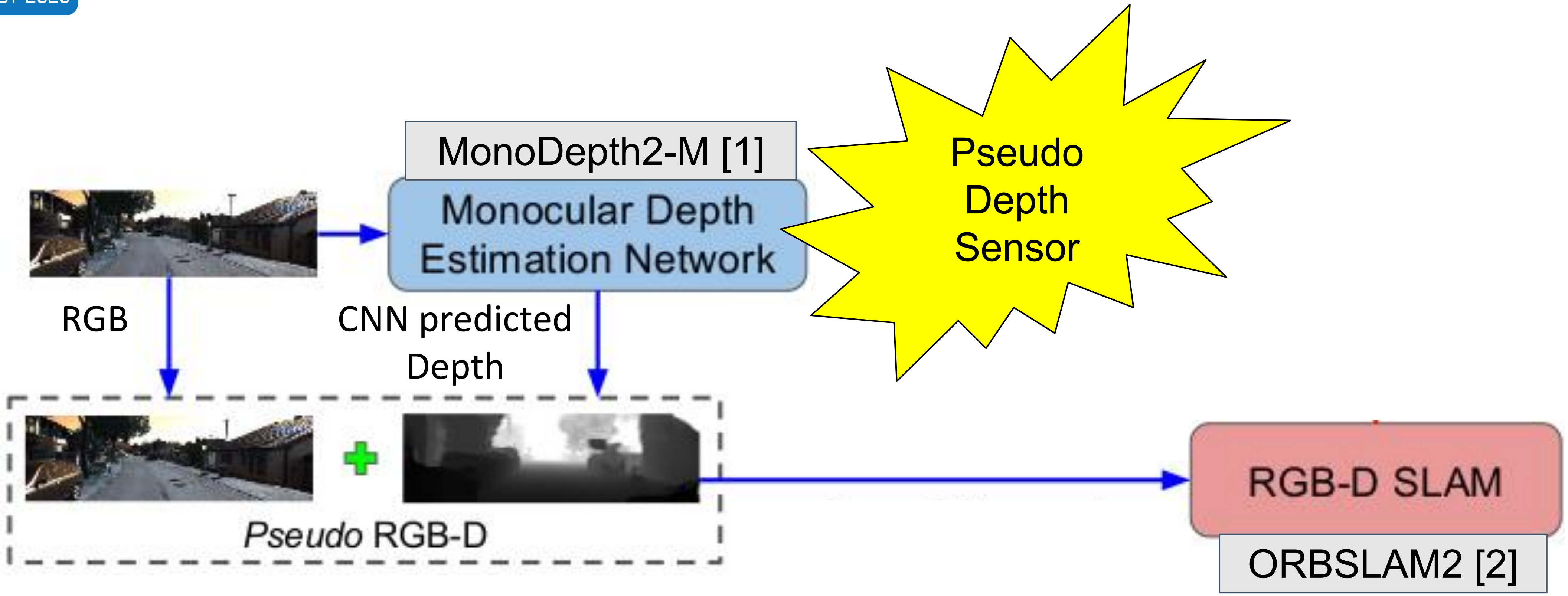
A Self-Supervised, Self-Improving Framework



- Prepare **Pseudo RGB-D** data

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*

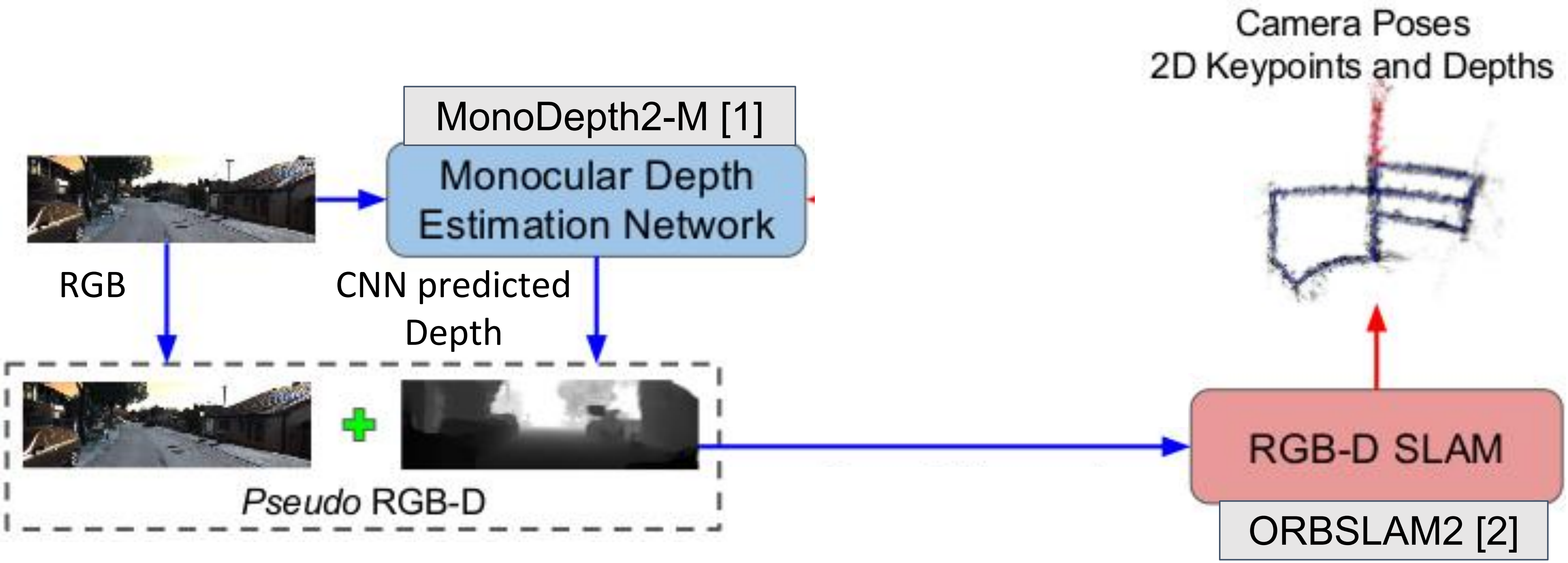
A Self-Supervised, Self-Improving Framework



- Prepare **Pseudo RGB-D** data
- Run RGB-D SLAM on **Pseudo RGB-D** pairs. We use RGB-D version of **ORB-SLAM2 [2]** as base RGB-D SLAM

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*
[2] Mur-Artal wt al."ORBSLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Transactions on Robotics* 2017

A Self-Supervised, Self-Improving Framework

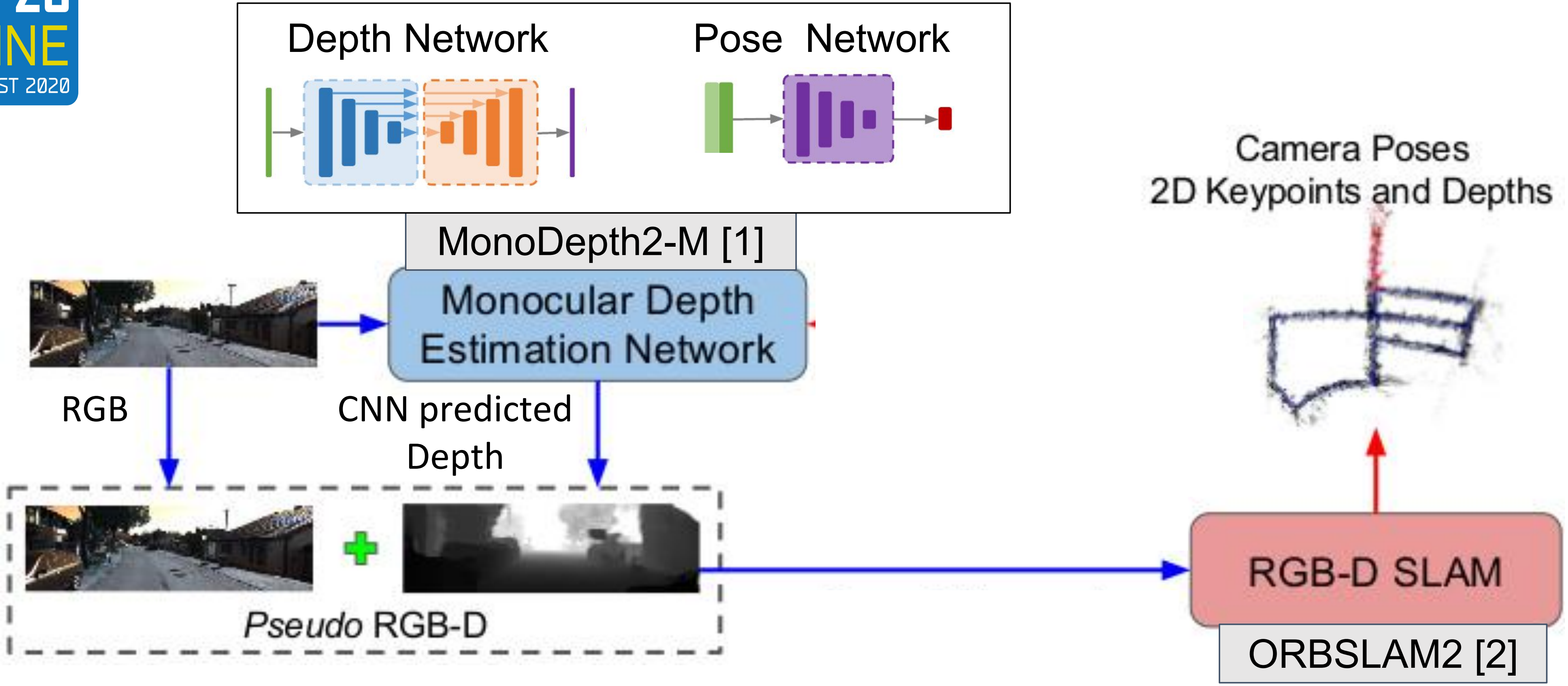


- Prepare **Pseudo RGB-D** data
- Run RGB-D SLAM on **Pseudo RGB-D** pairs. We use RGB-D version of **ORB-SLAM2 [2]** as base RGB-D SLAM
- Save Pseudo RGB-D SLAM outputs (Camera poses, keyframes, tracked keypoints and their depth values)

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*

[2] Mur-Artal wt al."ORBSLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Transactions on Robotics* 2017

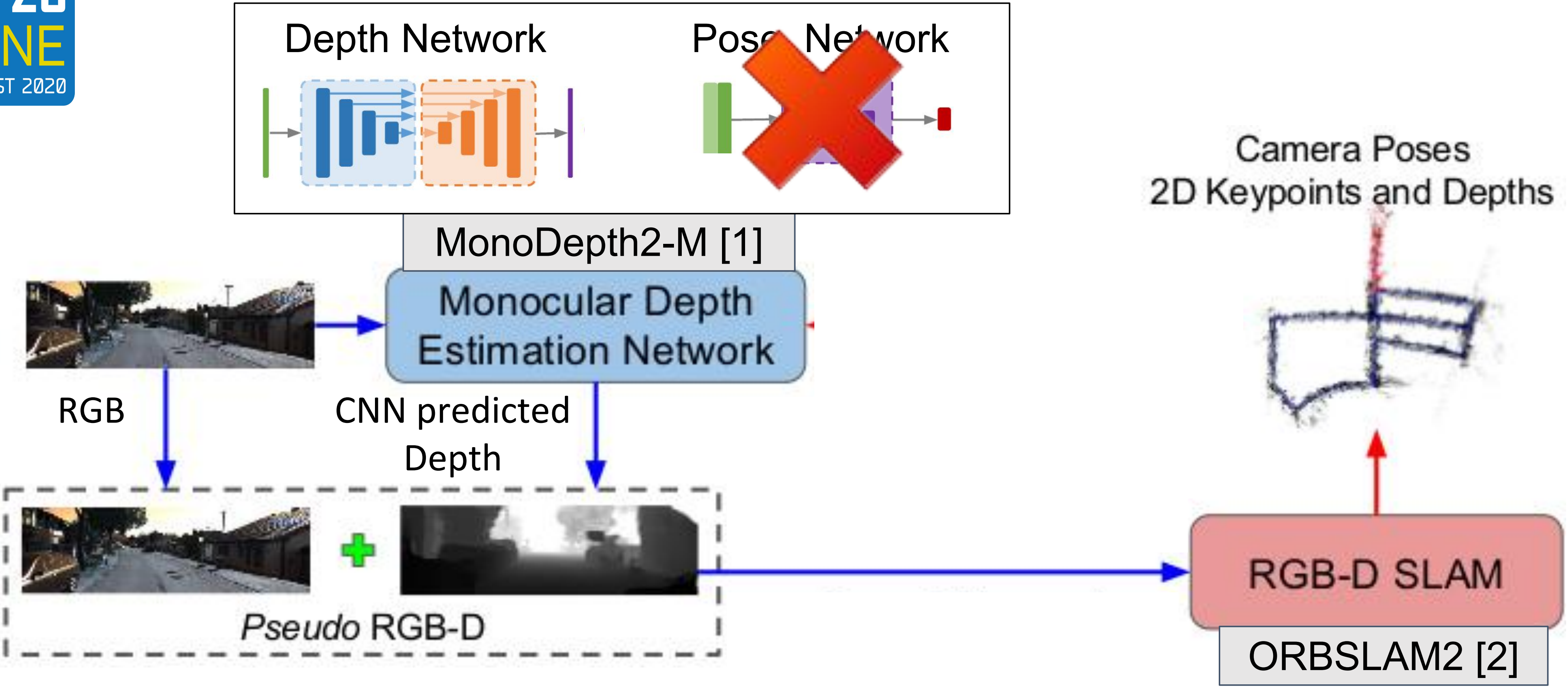
A Self-Supervised, Self-Improving Framework



- Depth Refinement

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*
[2] Mur-Artal wt al."ORBSLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Transactions on Robotics* 2017

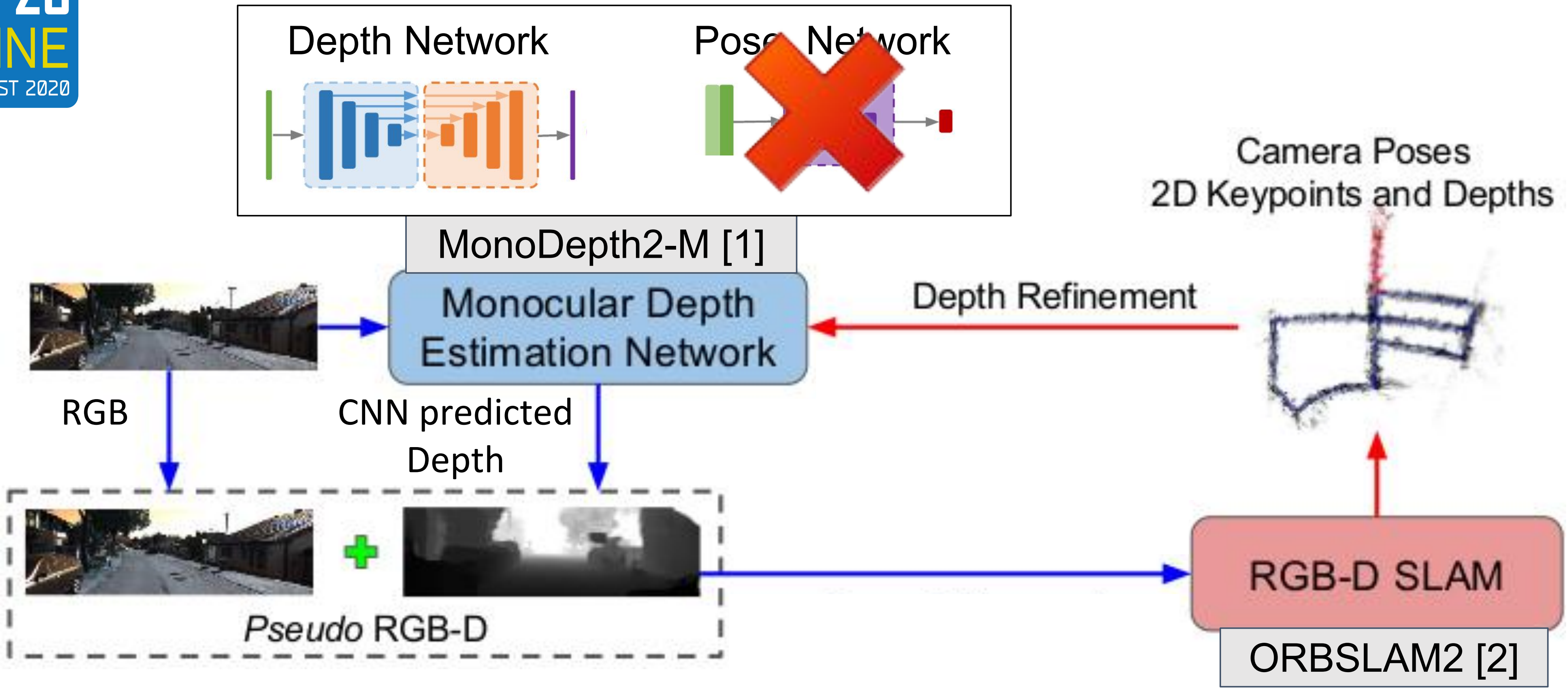
A Self-Supervised, Self-Improving Framework



- **Depth Refinement**
  - Disable MonoDepth2 pose network
  - Use camera poses obtained from Pseudo RGB-D SLAM

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*
[2] Mur-Artal wt al."ORBSLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Transactions on Robotics* 2017

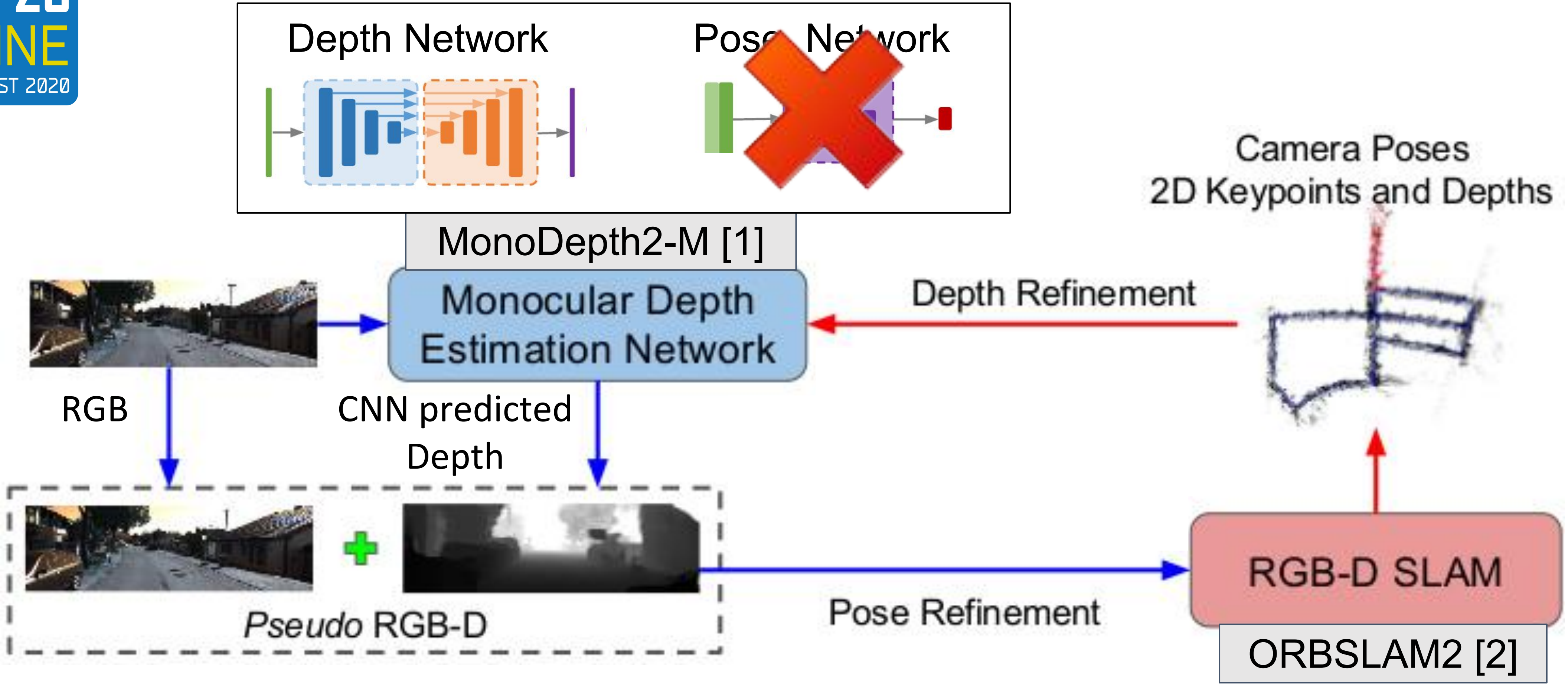
A Self-Supervised, Self-Improving Framework



- **Depth Refinement**
  - Disable MonoDepth2 pose network
  - Use camera poses obtained from Pseudo RGB-D SLAM

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*
[2] Mur-Artal wt al."ORBSLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Transactions on Robotics* 2017

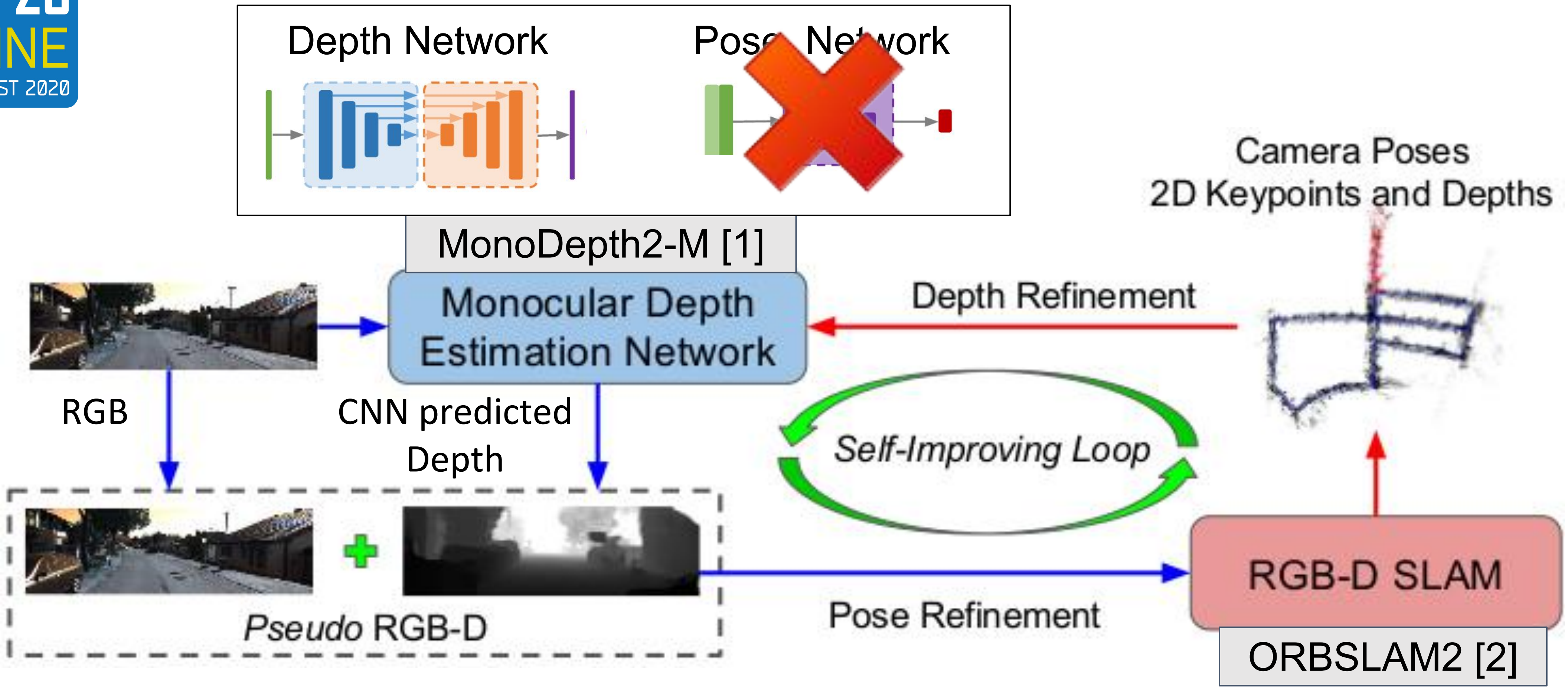
A Self-Supervised, Self-Improving Framework



- **Pose Refinement**
  - Use the refined depth model to prepare Pseudo RGB-D data
  - Re-run Pseudo RGBD-D SLAM and get refined camera poses, keypoints amd their updated locations

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*
[2] Mur-Artal wt al."ORBSLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Transactions on Robotics* 2017

A Self-Supervised, Self-Improving Framework



- **Self-Improving Loop**
  - Run until we see no improvement in depth and/or pose

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*
[2] Mur-Artal wt al."ORBSLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras." *IEEE Transactions on Robotics* 2017

A Self-Supervised, Self-Improving Framework

Pose Refinement

- Cannot use Pseudo RGB-D data directly to run RGB-D SLAM
- Pseudo Depth Sensor
  - CNN predict depth values at different scales compared to real active sensors e.g LiDAR

ECCV'20 ONLINE 23-28 AUGUST 2020

A Self-Supervised, Self-Improving Framework

Pose Refinement

- Cannot use Pseudo RGB-D data directly to run RGB-D SLAM
- Pseudo Depth Sensor
  - CNN predict depth values at different scales compared to real active sensors e.g LiDAR
- Adaptive Baseline (b)
  - Mimic the setup of KITTI dataset [1]

$$b = \frac{b^{KITTI}}{d^{KITTI}} * d_{max}$$

b^{KITTI} 0.54 meters

d^{KITTI} 80 meters

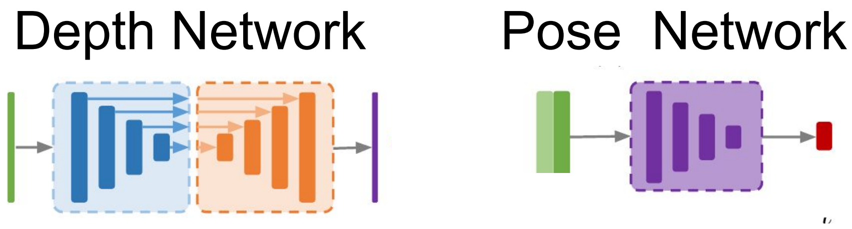
d_{max} Max CNN-predicted depth of the input sequence

[1] Geiger, Andreas, et al. "Vision meets robotics: The kitti dataset." *The International Journal of Robotics Research* 32.11 (2013): 1231-1237.

A Self-Supervised, Self-Improving Framework

Depth Refinement

- **Pre-training:** Use MonoDepth2's pose network (*Once*).
- **Depth Refinement:** Use Pseudo RGB-D SLAM's output poses.



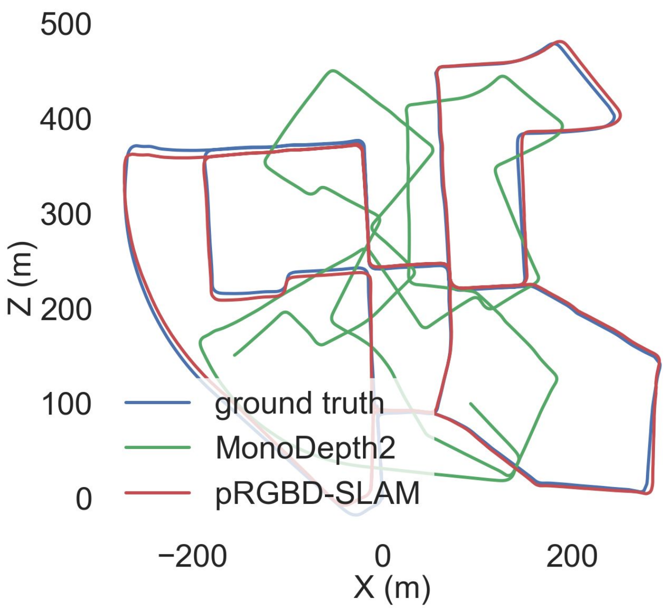
Pre-training Configuration

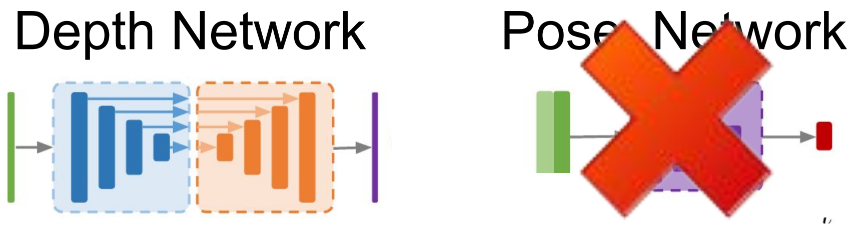
[1] Geiger, Andreas, et al. "Vision meets robotics: The kitti dataset." *The International Journal of Robotics Research* 32.11 (2013): 1231-1237.

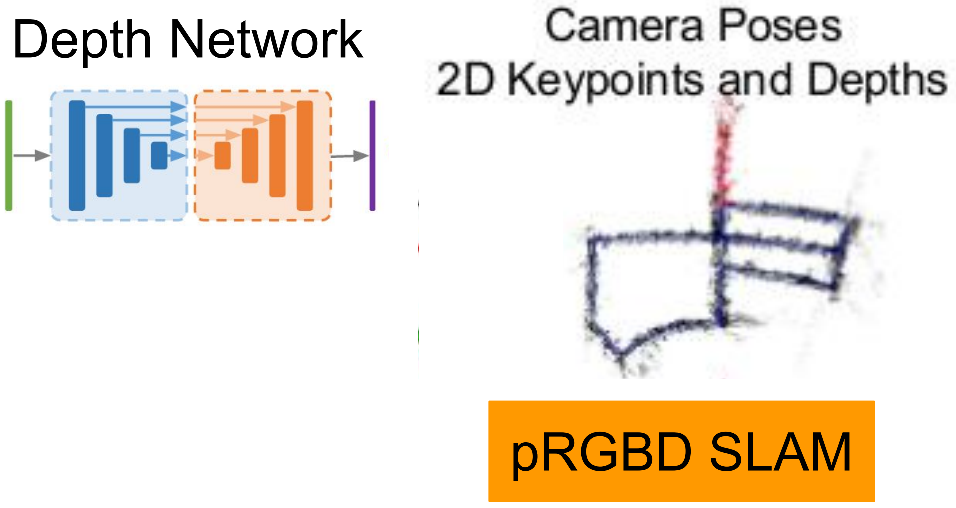
A Self-Supervised, Self-Improving Framework

Depth Refinement

- **Pre-training:** Use MonoDepth2’s pose network (*Once*).
- **Depth Refinement:** Use Pseudo RGB-D SLAM’s output poses.







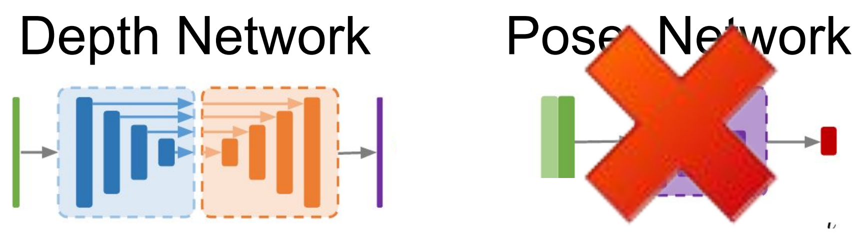
Refinement Configuration

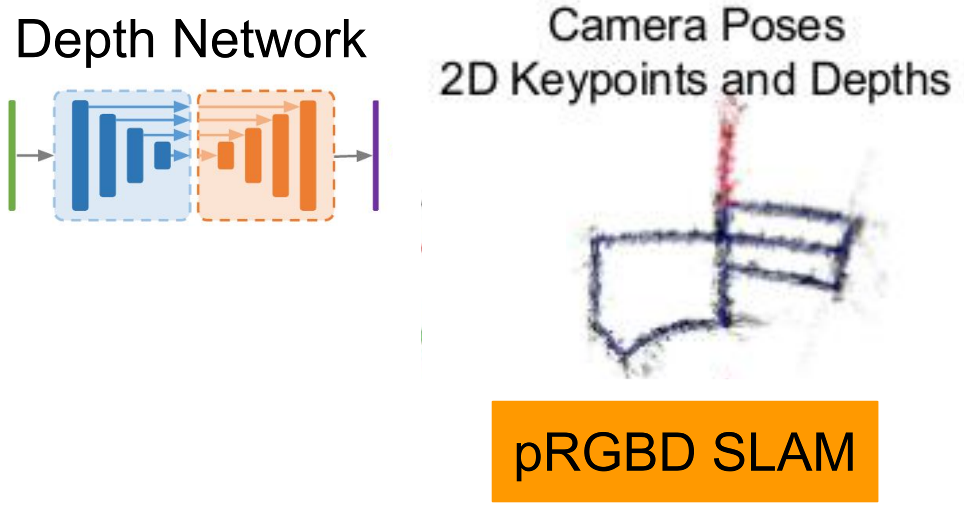
[1] Geiger, Andreas, et al. "Vision meets robotics: The kitti dataset." *The International Journal of Robotics Research* 32.11 (2013): 1231-1237.

A Self-Supervised, Self-Improving Framework

Depth Refinement

- **Pre-training:** Use MonoDepth2's pose network (*Once*).
- **Depth Refinement:** Use Pseudo RGB-D SLAM's output poses.
- True Camera Intrinsics
    - Instead of average camera intrinsics , we use true camera intrinsics during refinement.





Refinement Configuration

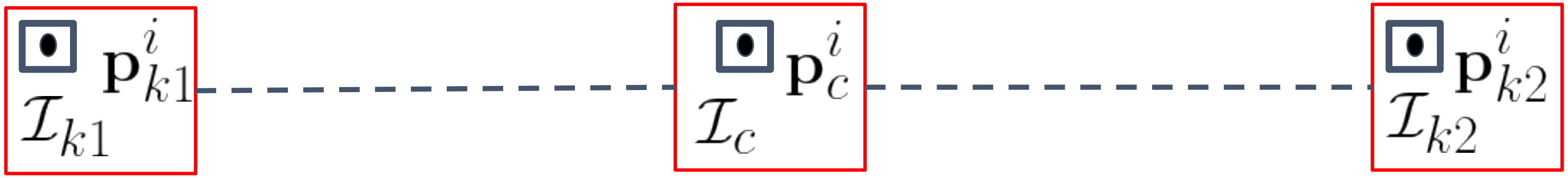
[1] Geiger, Andreas, et al. "Vision meets robotics: The kitti dataset." *The International Journal of Robotics Research* 32.11 (2013): 1231-1237.

A Self-Supervised, Self-Improving Framework

Depth Refinement

$$k1 < c < k2$$

$$\mathbf{p}_c^i = [\, p_c^{i1}, p_c^{i2} \,]$$



$\mathcal{X} = \{\mathbf{p}^i\}$ Set of keypoints visible in all the three **keyframes**

$d_c^i(\mathbf{w})$ Depth of ith keypoint in the **keyframe** $\mathcal{I}_c$ obtained from the **depth network**

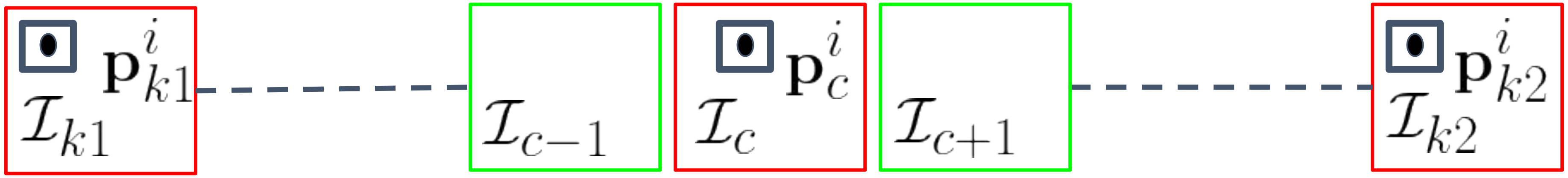
ECCV'20 ONLINE 23-28 AUGUST 2020

A Self-Supervised, Self-Improving Framework

Depth Refinement

$$k1 < c < k2$$

$$\mathbf{p}_c^i = [\, p_c^{i1}, p_c^{i2} \,]$$



$\mathcal{X} = \{\mathbf{p}^i\}$ Set of keypoints visible in all the three **keyframes**

$d_c^i(\mathbf{w})$ Depth of ith keypoint in the **keyframe** $\mathcal{I}_c$ obtained from the **depth network**

$\mathcal{I}_{c-1}$ $\mathcal{I}_{c+1}$ Temporally adjacent **frames** of the central **keyframe** $\mathcal{I}_c$

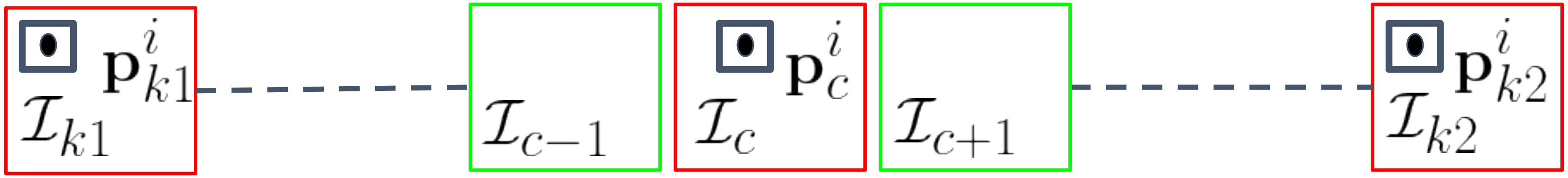
$\mathbf{T}_{c\rightarrow c-1}$ $\mathbf{T}_{c\rightarrow c+1}$ Relative camera poses between $\mathcal{I}_c$ and its temporally adjacent frames, obtained from **Pseudo RGB-D SLAM**

A Self-Supervised, Self-Improving Framework

Depth Refinement

$$k1 < c < k2$$

$$\mathbf{p}_c^i = [\, p_c^{i1}, p_c^{i2} \,]$$



$\mathcal{I}_{c-1}$ $\mathcal{I}_{c+1}$ Temporally adjacent **frames** of central **keyframe** $\mathcal{I}_c$

$\mathcal{I}_{c-1}^{'}$ $\mathcal{I}_{c+1}^{'}$ Synthesized temporally adjacent **frames**

$\mathcal{P}_c = \mathrm{PE}(\mathcal{I}_{c-1}^{'}, \mathcal{I}_{c-1}) + \mathrm{PE}(\mathcal{I}_{c+1}^{'}, \mathcal{I}_{c+1})$ **Photometric error**

$\mathcal{S}_c$ **Smoothness loss**

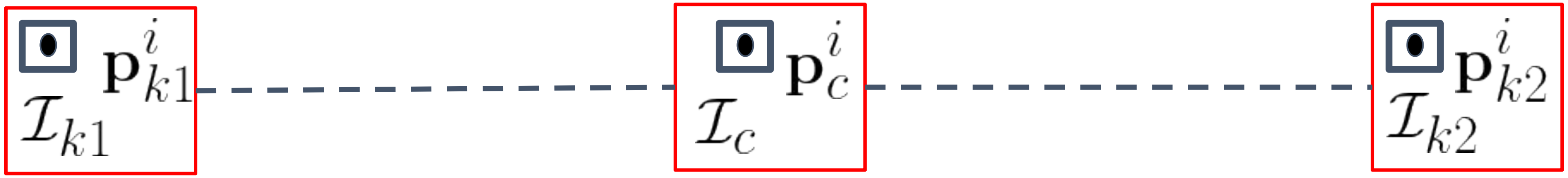
Narrow baseline losses

A Self-Supervised, Self-Improving Framework

Depth Refinement

$$k1 < c < k2$$

$$\mathbf{p}_c^i = [\ p_c^{i1}, p_c^{i2}\ ]$$



$d_c^i(\mathbf{w})$ Depth of ith keypoint in the **keyframe** $\mathcal{I}_c$ obtained from the **depth network**

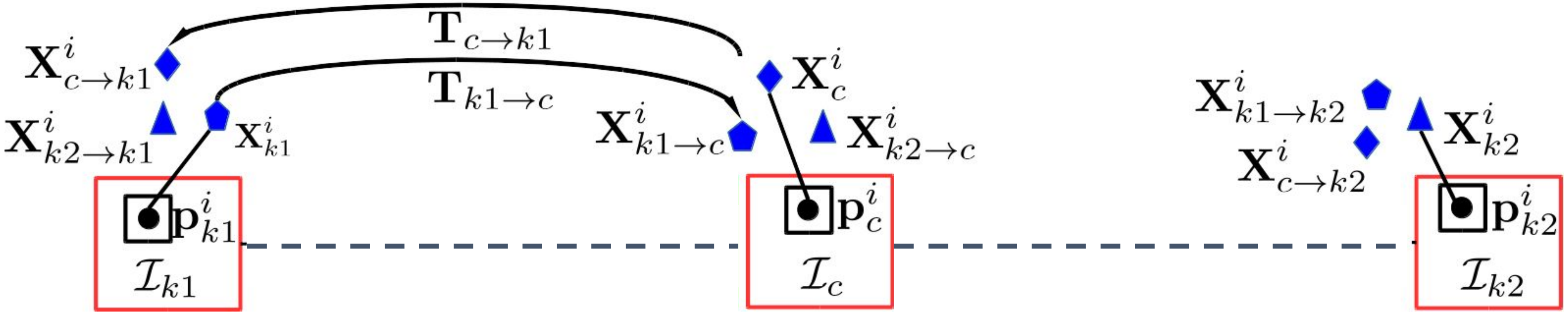
$$\mathbf{X}_c^i = \mathbf{K}^{-1}[\ \mathbf{p}_c^i,\ 1\ ]^T d_c^i(\mathbf{w})$$ **Backproject to 3D**

$$\mathbf{X}_{c\to k1}^i = \mathbf{T}_{c\to k1}\,\mathbf{X}_c^i = [x_{c\to k1}^i(\mathbf{w}),\ y_{c\to k1}^i(\mathbf{w}),\ d_{c\to k1}^i(\mathbf{w})]$$ **Depth transfer**

$\mathbf{T}_{c\to k1}$ Relative camera pose obtained from Pseudo RGB-D SLAM

A Self-Supervised, Self-Improving Framework

Depth Refinement



$$\mathbf{X}^i_{c\to k1} = \mathbf{T}_{c\to k1}\,\mathbf{X}^i_c = [x^i_{c\to k1}(\mathbf{w}),\ y^i_{c\to k1}(\mathbf{w}),\ d^i_{c\to k1}(\mathbf{w})]$$

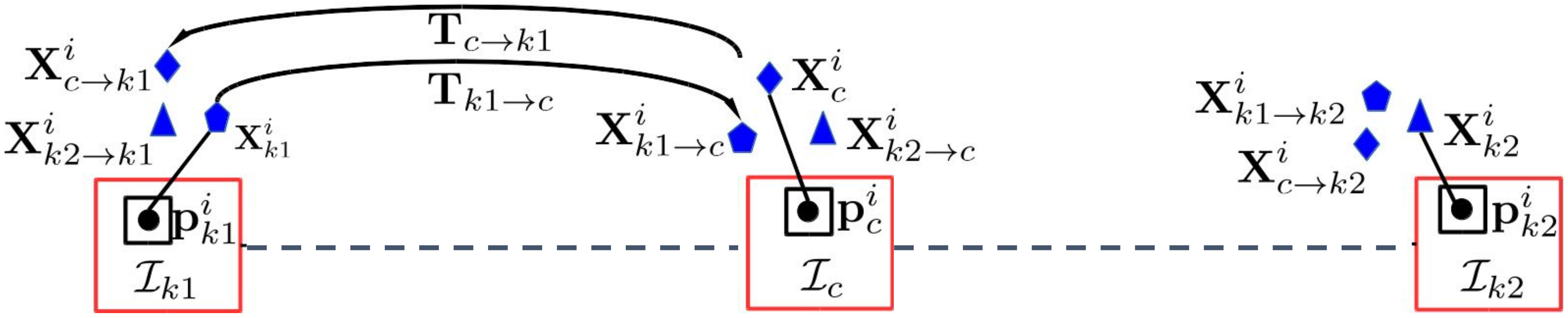
Depth transfer

$$|d^i_{c\to k1}(\mathbf{w})\text{-}d^i_{k1}(\mathbf{w})|$$

Depth Transfer loss

A Self-Supervised, Self-Improving Framework

Depth Refinement



$$\mathbf{X}^i_{c\to k1} = \mathbf{T}_{c\to k1}\,\mathbf{X}^i_c = [x^i_{c\to k1}(\mathbf{w}),\ y^i_{c\to k1}(\mathbf{w}),\ d^i_{c\to k1}(\mathbf{w})]$$

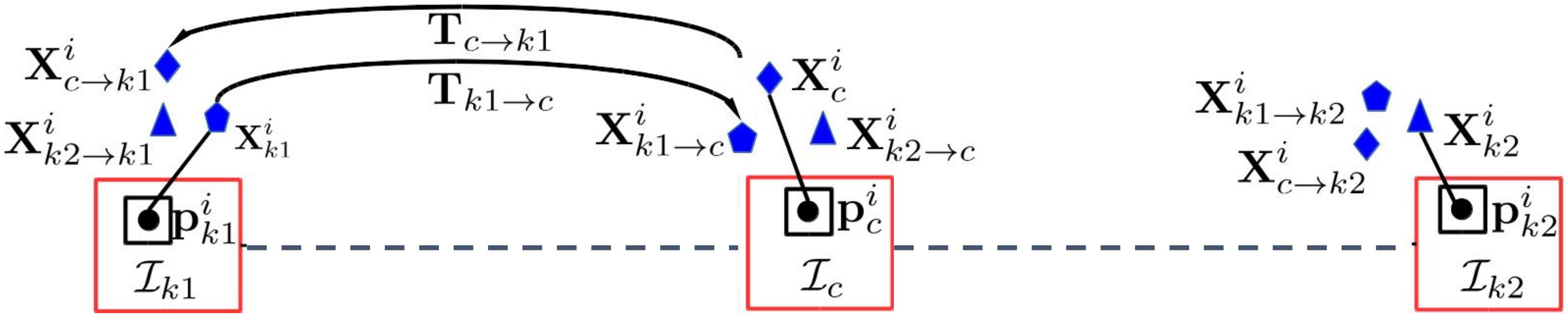
Depth transfer

$$|d^i_{c\to k1}(\mathbf{w})\text{-}d^i_{k1}(\mathbf{w})|+|d^i_{k1\to c}(\mathbf{w})\text{-}d^i_c(\mathbf{w})|$$

Depth Transfer loss Depth Transfer loss

A Self-Supervised, Self-Improving Framework

Depth Refinement



$$\mathbf{X}^i_{c\rightarrow k1} = \mathbf{T}_{c\rightarrow k1}\,\mathbf{X}^i_c = [x^i_{c\rightarrow k1}(\mathbf{w}),\ y^i_{c\rightarrow k1}(\mathbf{w}),\ d^i_{c\rightarrow k1}(\mathbf{w})]$$

Depth transfer

$$\mathcal{T}^i_{c\leftrightarrow k1}(\mathbf{w}) = |d^i_{c\rightarrow k1}(\mathbf{w}) - d^i_{k1}(\mathbf{w})| + |d^i_{k1\rightarrow c}(\mathbf{w}) - d^i_c(\mathbf{w})|$$

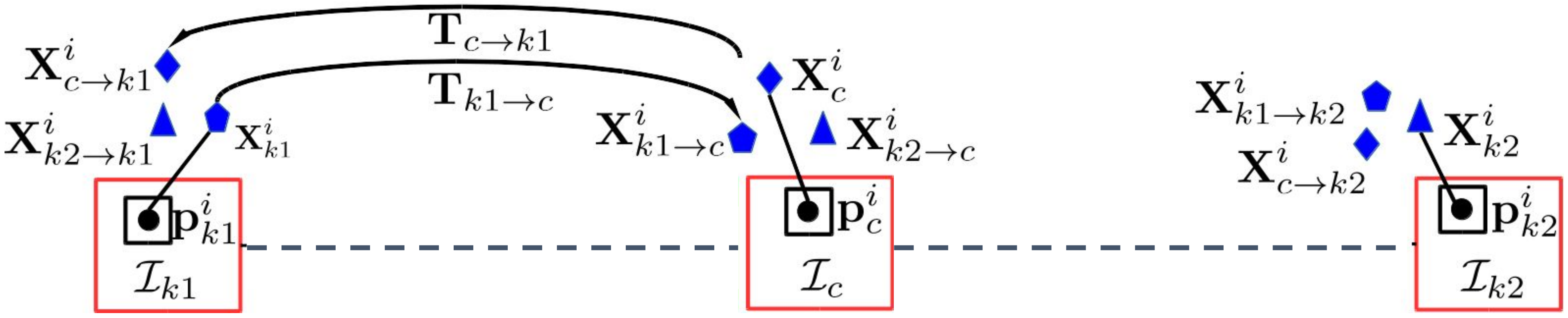
Symmetric Depth Transfer loss

Depth Transfer loss

Depth Transfer loss

A Self-Supervised, Self-Improving Framework

Depth Refinement



$$\mathbf{X}^i_{c\to k1} = \mathbf{T}_{c\to k1}\,\mathbf{X}^i_c = [x^i_{c\to k1}(\mathbf{w}),\ y^i_{c\to k1}(\mathbf{w}),\ d^i_{c\to k1}(\mathbf{w})]$$

Depth transfer

$$\mathcal{T}^i_{c\leftrightarrow k1}(\mathbf{w}) = |d^i_{c\to k1}(\mathbf{w}) - d^i_{k1}(\mathbf{w})| + |d^i_{k1\to c}(\mathbf{w}) - d^i_c(\mathbf{w})|$$

Symmetric Depth Transfer loss

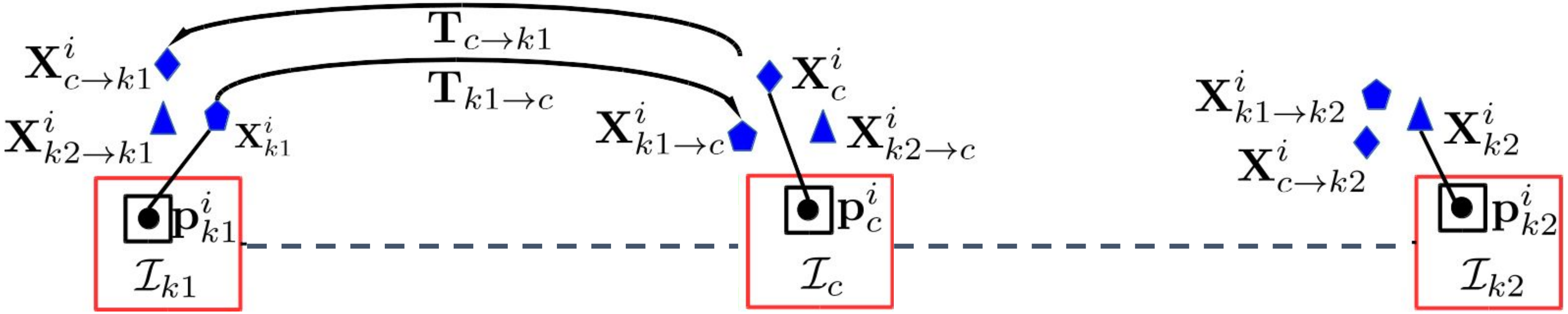
Depth Transfer loss

Depth Transfer loss

Similarly compute $\mathcal{T}^i_{c\leftrightarrow k2}$ $\mathcal{T}^i_{k1\leftrightarrow k2}$

A Self-Supervised, Self-Improving Framework

Depth Refinement



$$\mathbf{X}^i_{c\to k1} = \mathbf{T}_{c\to k1}\,\mathbf{X}^i_c = [x^i_{c\to k1}(\mathbf{w}),\; y^i_{c\to k1}(\mathbf{w}),\; d^i_{c\to k1}(\mathbf{w})]$$

Depth transfer

$$\mathcal{T}^i_{c\leftrightarrow k1}(\mathbf{w})=|d^i_{c\to k1}(\mathbf{w})\text{-}d^i_{k1}(\mathbf{w})|+|d^i_{k1\to c}(\mathbf{w})\text{-}d^i_c(\mathbf{w})|$$

Symmetric Depth Transfer loss

Depth Transfer loss

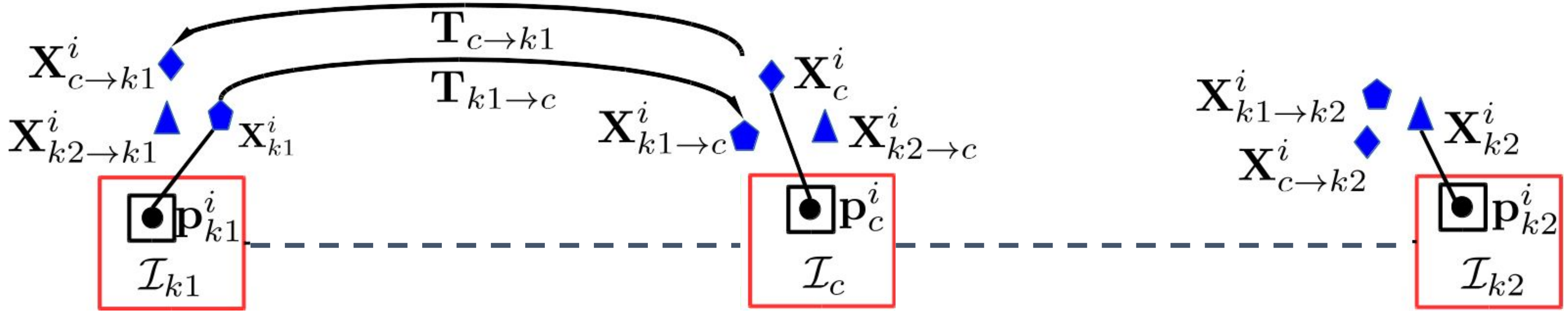
Depth Transfer loss

Wide baseline losses

Similarly compute $\mathcal{T}^i_{c\leftrightarrow k2}$ $\mathcal{T}^i_{k1\leftrightarrow k2}$

A Self-Supervised, Self-Improving Framework

Depth Refinement



$d^i_c(\mathbf{w})$ Depth of ith keypoint in the **keyframe** $\mathcal{I}_c$ obtained from the **depth network**

$d^i_c(\mathbf{SLAM})$ Depth of ith keypoint in the keyframe $\mathcal{I}_c$ obtained from Pseudo RGB-D SLAM

$$\mathcal{D}_c = \frac{\sum_{i\in\mathcal{X}} |d^i_c(\mathbf{w}) - d^i_c(\mathbf{SLAM})|}{|\mathcal{X}|}$$

Depth Consistency Loss

$$\mathcal{L} = \alpha\mathcal{P}_c + \beta\mathcal{S}_c + \gamma\mathcal{D}_c + \mu\left(\mathcal{T}^i_{c\leftrightarrow k1} + \mathcal{T}^i_{c\leftrightarrow k2} + \mathcal{T}^i_{k1\leftrightarrow k2}\right)$$

Total Loss

Experiments - Depth Refinement Evaluation

Quantitative Results

- Standard KITTI Eigen’s train-test split
- M : Monocular tranining
- S : Stereo training
- MS : Monocular and stereo training

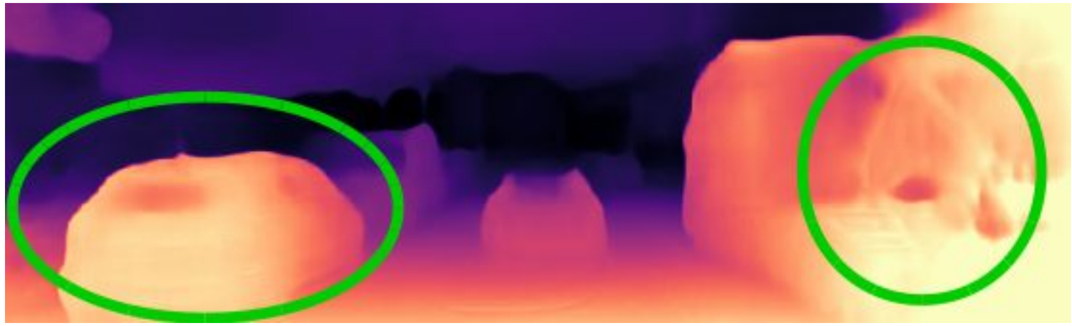
	Method	Train	Lower is better				Higher is better		
			Abs Rel	Sq Rel	RMSE	RMSE log	a1	a2	a3
self-supervised	Yang[55]	M	0.182	1.481	6.501	0.267	0.725	0.906	0.963
	Mahjourian[29]	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
	Klodt[22]	M	0.166	1.490	5.998	-	0.778	0.919	0.966
	DDVO[44]	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
	GeoNet[57]	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975
	DF-Net[64]	M	0.150	1.124	5.507	0.223	0.806	0.933	0.973
	Ranjan[35]	M	0.148	1.149	5.464	0.226	0.815	0.935	0.973
	EPC++[28]	M	0.141	1.029	5.350	0.216	0.816	0.941	0.976
	Struct2depth(M)[4]	M	0.141	1.026	5.291	0.215	0.816	0.945	0.979
	WBAF [59]	M	0.135	0.992	5.288	0.211	0.831	0.942	0.976
	MonoDepth2-M (re-train) [15]	M	0.117	0.941	4.889	0.194	0.873	0.957	0.980
	MonoDepth2-M (original) [15]	M	0.115	0.903	4.863	0.193	**0.877**	0.959	0.981
	pRGBD-Refined	M	**0.113**	**0.793**	**4.655**	**0.188**	0.874	**0.960**	**0.983**
	Garg[13]	S	0.152	1.226	5.849	0.246	0.784	0.921	0.967
	3Net (R50)[34]	S	0.129	0.996	5.281	0.223	0.831	0.939	0.974
	Monodepth2-S[15]	S	0.109	0.873	4.960	0.209	0.864	0.948	0.975
	SuperDepth [33]	S	0.112	0.875	4.958	0.207	0.852	0.947	0.977
	monoResMatch [43]	S	0.111	0.867	4.714	0.199	0.864	0.954	0.979
	DepthHints [49]	S	0.106	0.780	4.695	0.193	0.875	**0.958**	**0.980**
	DVSO[53]	S	**0.097**	**0.734**	**4.442**	**0.187**	**0.888**	**0.958**	**0.980**
	UnDeepVO [24]	MS	0.183	1.730	6.570	0.268	-	-	-
	EPC++ [28]	MS	0.128	0.935	5.011	0.209	0.831	0.945	**0.979**
	Monodepth2-MS[15]	MS	**0.106**	**0.818**	**4.750**	**0.196**	**0.874**	**0.957**	**0.979**
	Eigen[8]	D	0.203	1.548	6.307	0.282	0.702	0.890	0.890
	Liu[26]	D	0.201	1.584	6.471	0.273	0.680	0.898	0.967
	Kuznietsov[23]	DS	0.113	0.741	4.621	0.189	0.862	0.960	0.986
	SVSM FT[28]	DS	0.094	0.626	4.252	0.177	0.891	0.965	0.984
	Guo[19]	DS	0.096	0.641	4.095	0.168	0.892	0.967	0.986
	DORN[12]	D	**0.072**	**0.307**	**2.727**	**0.120**	**0.932**	**0.984**	**0.994**

Experiments - Depth Refinement Evaluation

RGB

MonoDepth2 [1]-Stereo Supervision

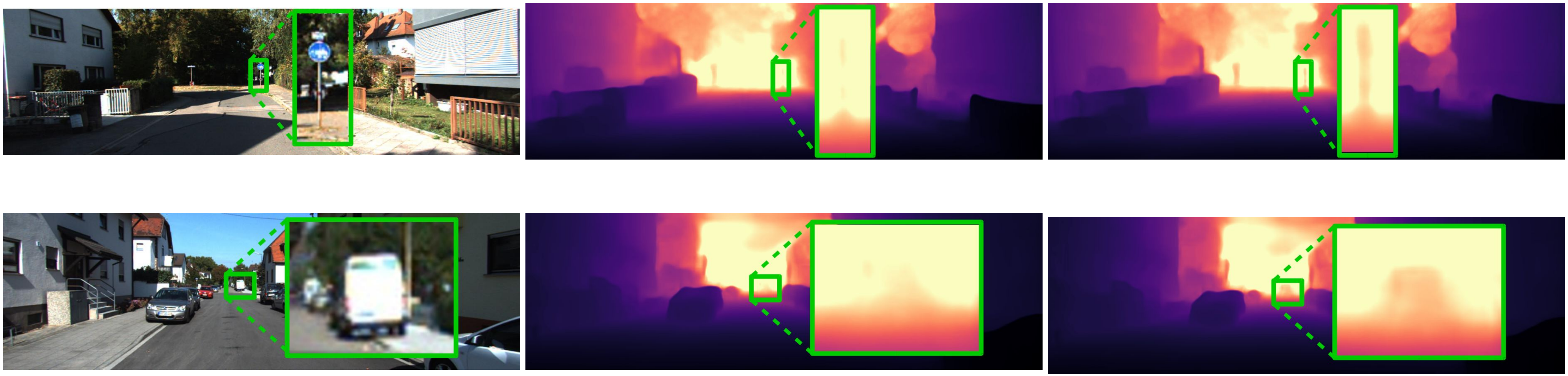
MonoDepth2 [1]-Monocular Supervision



pRGBD-Refined
(Proposed Method)

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." *in ICCV 2019*

Experiments - Depth Refinement Evaluation
Qualitative Results



RGB | MonoDepth2[1]-Monocular Supervision | pRGBD-Refined (Proposed Method)

- Visual improvements in the depth of farther points.

Experiments - Pose Refinement Evaluation

Quantitative Results

- KITTI Odometry Dataset
- Training Sequences: 00 - 08
- Testing Sequences: 09 and 10
- pRGBD-Initial: Pseudo RGB-D SLAM using pretrained CNN depths i.e 0th self-improving loop.

	Method	Seq. 09 RMSE	Seq. 09 RelTr	Seq. 09 RelRot	Seq. 10 RMSE	Seq. 10 RelTr	Seq. 10 RelRot
Supervised	DeepVO[47]	-	-	-	-	8.11	0.088
	ESP-VO[48]	-	-	-	-	9.77	0.102
	GFS-VO[50]	-	-	-	-	6.32	0.023
	GFS-VO-RNN[50]	-	-	-	-	7.44	0.032
	BeyondTracking[51]	-	-	-	-	**3.94**	**0.017**
	DeepV2D[42]	79.06	8.71	0.037	48.49	12.81	0.083
Self-Supervised	SfMLearner [60]	**24.31**	8.28	0.031	20.87	12.20	**0.030**
	GeoNet[57]	158.45	28.72	0.098	43.04	23.90	0.090
	Depth-VO[58]	-	11.93	0.039	-	12.45	0.035
	vid2depth[29]	-	-	-	-	21.54	0.125
	UnDeepVO[24]	-	7.01	0.036	-	10.63	0.046
	Wang *et al.*[45]	-	9.88	0.034	-	12.24	0.052
	CC[35]	29.00	**6.92**	**0.018**	**13.77**	7.97	0.031
	DeepMatchVO[37]	27.08	9.91	0.038	24.44	12.18	0.059
	Li *et al.*[25]	-	8.10	0.028	-	12.90	0.032
	Monodepth2-M[15]	55.47	11.47	0.032	20.46	**7.73**	0.034
	SC-SfMLearer[2]	-	11.2	0.034	-	10.1	0.050
	RGB ORB-SLAM	18.34	7.42	**0.004**	8.90	5.85	**0.004**
	pRGBD-Initial	12.21	4.26	0.011	8.30	5.55	0.017
	pRGBD-Refined	**11.97**	**4.20**	0.010	**6.35**	**4.40**	0.016

[1] Geiger, Andreas, et al. "Vision meets robotics: The kitti dataset." *The International Journal of Robotics Research* 32.11 (2013): 1231-1237.

Experiments - Pose Refinement Evaluation
Quantitative Results

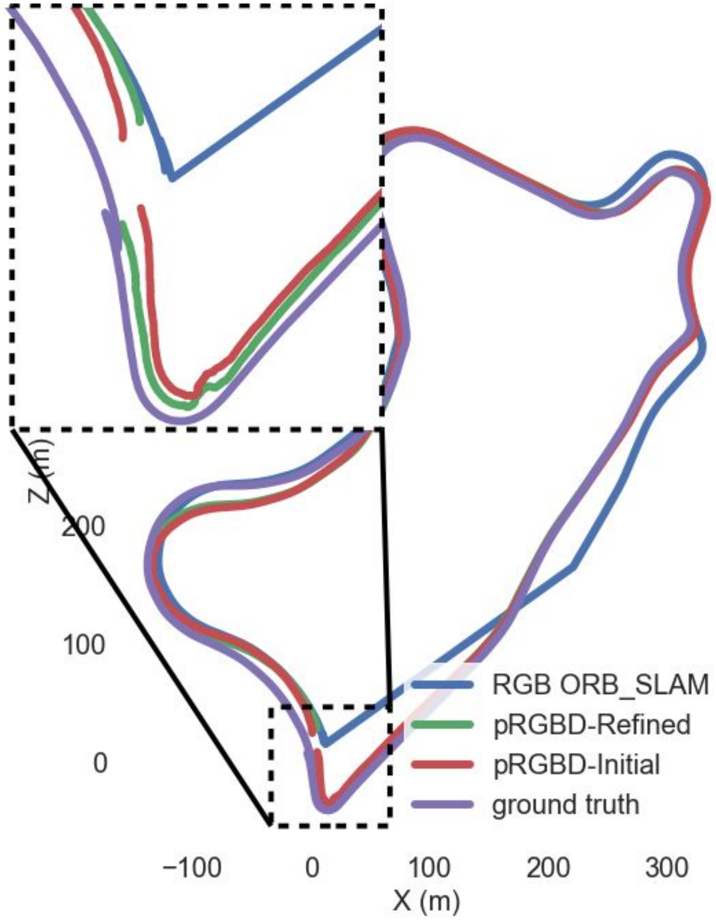
- KITTI Odometry Dataset
- Training Sequences: 00 - 08
- Testing Sequences: 11 - 21

Seq	RGB ORB-SLAM			pRGBD-Initial			pRGBD-Refined		
	RMSE	RelTr	RelRot	RMSE	RelTr	RelRot	RMSE	RelTr	RelRot
11	14.83	7.69	**0.003**	6.68	3.28	0.016	**3.64**	**2.96**	0.015
13	6.58	2.39	**0.006**	6.83	2.52	0.008	**6.43**	**2.31**	0.007
14	4.81	5.19	**0.004**	4.30	4.14	0.014	**2.15**	**3.06**	0.014
15	3.67	1.78	**0.004**	2.58	1.61	0.005	**2.07**	**1.33**	**0.004**
16	6.21	2.66	**0.002**	5.78	2.14	0.006	**4.65**	**1.90**	0.004
18	6.63	2.38	**0.002**	5.50	2.30	0.008	**4.37**	**2.21**	0.006
19	18.68	4.91	**0.002**	23.96	2.82	0.007	**13.85**	**2.52**	0.006
20	9.19	6.74	**0.016**	8.94	5.43	0.027	**7.03**	**4.50**	0.022
12	X	X	X	X	X	X	**94.2**	**32.94**	**0.026**
17	X	X	X	14.71	8.98	**0.011**	**12.23**	**7.23**	**0.011**
21	X	X	X	X	X	X	X	X	X

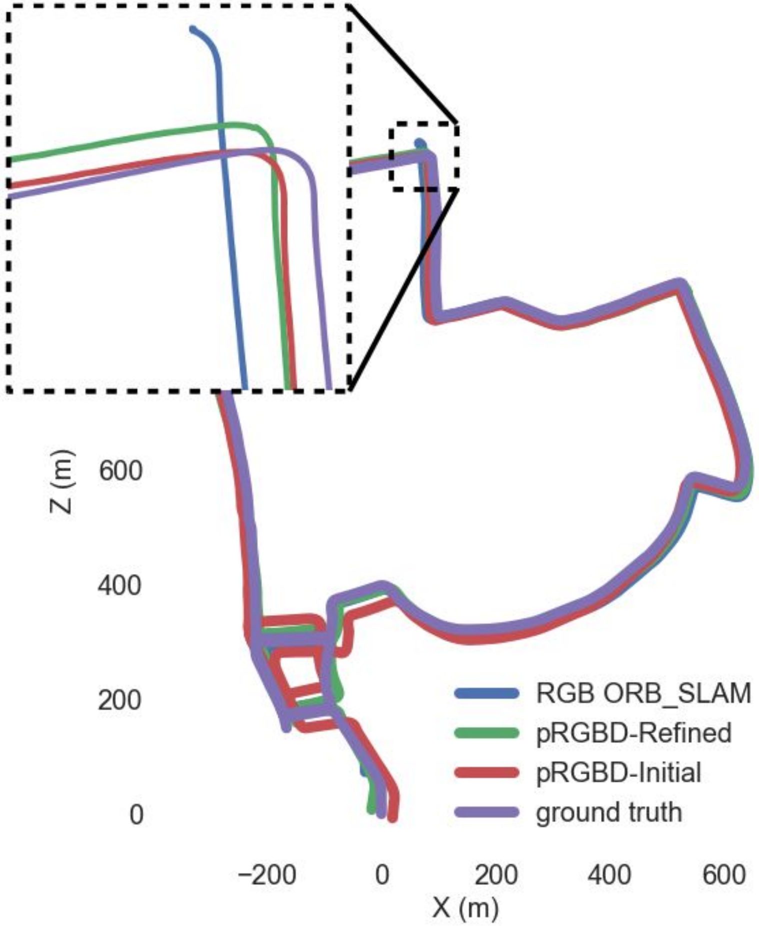
[1] Geiger, Andreas, et al. "Vision meets robotics: The kitti dataset." *The International Journal of Robotics Research* 32.11 (2013): 1231-1237.

Experiments - Pose Refinement Evaluation

Qualitative Results



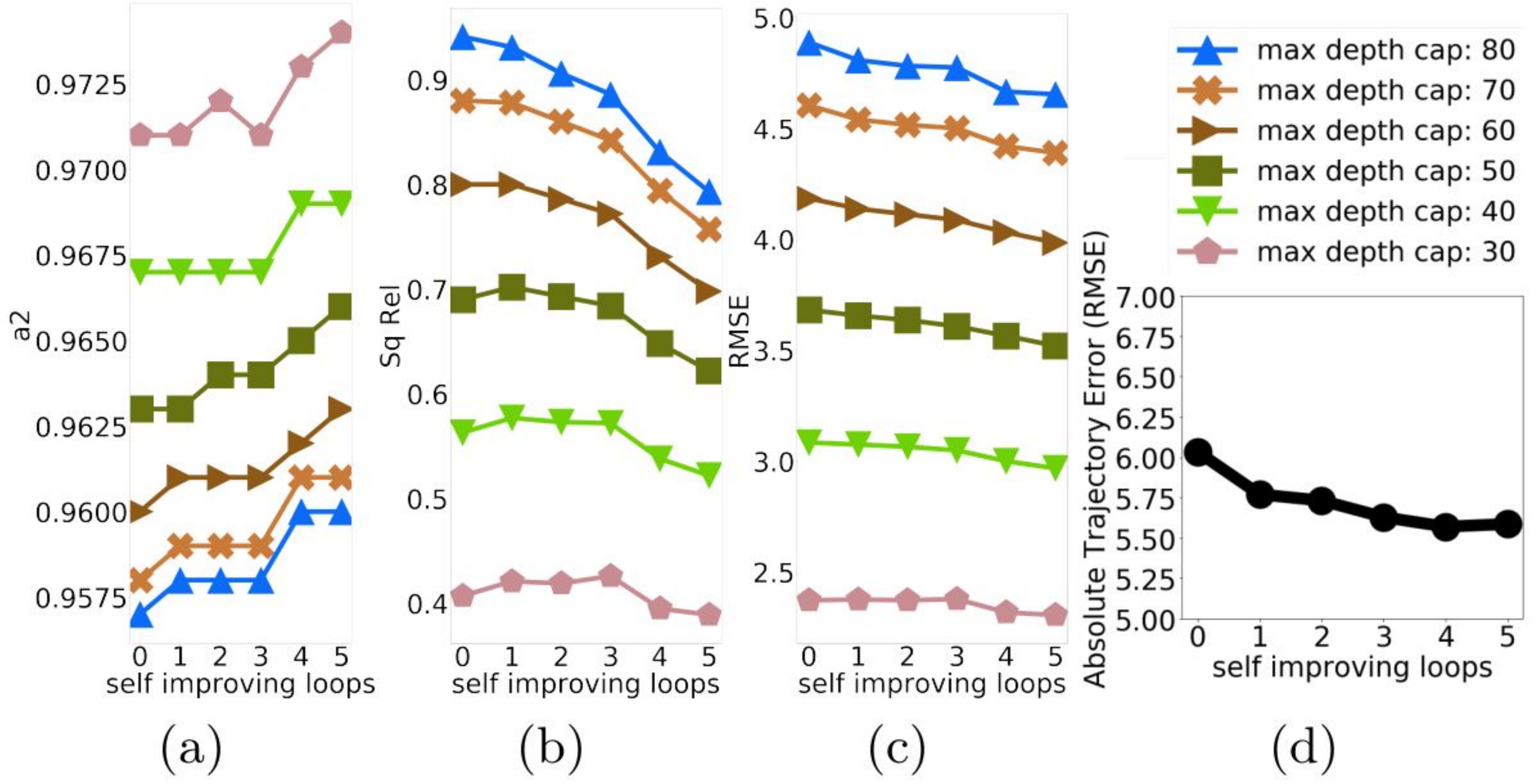
KITTI Odometry Sequence 09



KITTI Odometry Sequence 19

Analysis of Self-Improving Loops

- Improved depth predictions of both nearby and farther away points.
- Significant rate of reduction of errors.
- Pose refinement complements depth refinement.



Depth/Pose Evaluation metric w.r.t self-improving loops. Depth Evaluation metrics in (a-c) are computed at different max depth caps.

Conclusion

- Self-Improving framework to couple geometrical and learning based methods for 3D perception.
- Win-win situation achieved
- Both monocular SLAM and depth prediction are improved by a significant margin, without any active depth sensor and ground truth label

ECCV'20 ONLINE
23-28 AUGUST 2020

Thank you