DualRefine: Self-Supervised Depth and Pose Estimation Through Iterative Epipolar Sampling and Refinement Toward Equilibrium Supplementary Material

Antyanta Bangunharcana[1], Ahmed Magd[2], Kyung-Soo Kim[1]
[1]Mechatronics, Systems, and Control Laboratory, [2]Vehicular Systems Design and Control Lab
Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
{antabangun, a.magd, kyungsoo}@kaist.ac.kr

1. DEQ Framework

We adhere to the general framework of DEQ [2, 3] and employ a quasi-Newton solver to accelerate convergence. In our experiments, we utilize the Anderson solver [1]. A DEQ model computes $A = I - \frac{\partial \mathrm{U}}{\partial z^*}$ at the fixed point z^* to obtain the gradient. This is typically achieved by performing another fixed-point iteration. However, in line with [2, 5, 8], we approximate $A = I$ and utilize the inexact gradient for training.

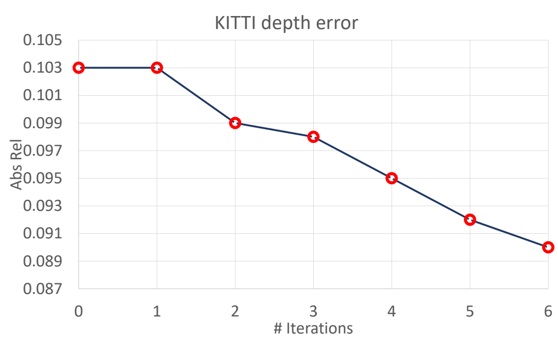

Figure 1. The progression of Abs Rel errors in each DualRefine iteration for KITTI depth.

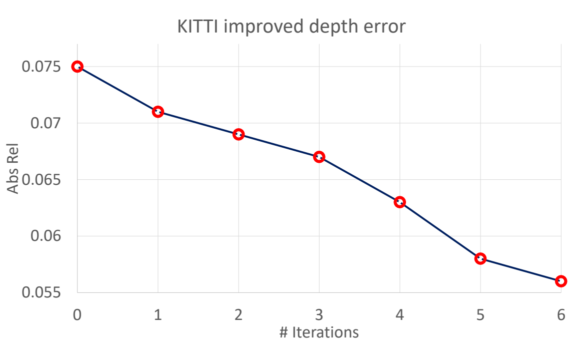

Figure 2. The progression of Abs Rel errors in each DualRefine iteration for KITTI improved depth.

2. Training Loss Combinations

Determining the optimal pairings to calculate the self-supervision losses at the refined fixed point is not straightforward. For each refined estimate (D^* and T^*), we can calculate the self-supervision loss using either the detached initial estimates ($[D^* \leftrightarrow$ detached $T_0]$ pair and $[T^* \leftrightarrow$ detached $D_0]$ pair) or the corresponding refined estimate ($[D^* \leftrightarrow T^*]$ pair and $[D^* \leftrightarrow T^*]$).

We observe a worse accuracy when both final estimates are paired with the corresponding initial estimates. We infer that, by pairing the final estimates with the initial ones, we impose a strong constraint on the model, limiting the scope of the output. We observe the best results when at least one of the final estimates is paired with the corresponding initial estimate. One example is when the depth loss is computed using the $[D^* \leftrightarrow$ detached $T_0]$ pair, while the pose loss is computed using the $[T^* \leftrightarrow D^*]$ pair. From this experiment, pairing the refined estimates with each other seems to display the best accuracy. However, to ensure scale consistency with the teacher networks, we follow the third loss pairing in the table.

3. Additional results on KITTI Depth

3.1. KITTI improved depth

In Tab. 1 we present evaluation results on the improved dense ground truth [19] of the KITTI [7] eigen split [4]. We perform garg cropping [6] and report the error for distances up to 80m. Our refinement module improves the initial estimates and outperforms most previous models while still being competitive with the Transformer [20]-based Depth-Former [12] model.

3.2. DEQ results

In Tab. 2 we present the error for the output of our model in each DEQ iteration. Iteration 0 corresponds to the depth

	Method	Test frames	$W \times H$	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Low & mid res	Ranjan [17]	1	832×256	0.123	0.881	4.834	0.181	0.860	0.959	0.985
	EPC++ [15]	1	832×256	0.120	0.789	4.755	0.177	0.856	0.961	0.987
	Johnston [13] *et al.*	1	640×192	0.081	0.484	3.716	0.126	0.927	0.985	0.996
	Monodepth2 [9]	1	640×192	0.090	0.545	3.942	0.137	0.914	0.983	0.995
	PackNet-SFM [11]	1	640×192	0.078	0.420	3.485	0.121	0.931	0.986	0.996
	DualRefine-initial (D_0)	1	640×192	0.075	0.379	3.490	0.117	0.936	0.989	0.997
	Patil *et al.* [16]	N†	640×192	0.087	0.495	3.775	0.133	0.917	0.983	0.995
	Wang *et al.* [21]	2 (-1, 0)	640×192	0.082	0.462	3.739	0.127	0.923	0.984	0.996
	ManyDepth [22]	2 (-1, 0)	640×192	0.064	0.320	3.187	0.104	0.946	0.990	0.995
	DepthFormer [12]	2 (-1, 0)	640×192	**0.055**	**0.271**	**2.917**	**0.095**	0.955	0.991	**0.998**
	DualRefine-refined (D^*)	2 (-1, 0)	640×192	0.056	0.281	3.040	**0.095**	**0.960**	**0.992**	**0.998**
High res	DRO [10]	2 (-1, 0)	960×320	0.057	0.342	3.201	0.123	0.952	0.989	0.996
	ManyDepth (HR ResNet50) [22]	2 (-1, 0)	1024×320	0.062	0.343	3.139	0.102	0.953	0.991	0.997
	DualRefine-refined (D^*)	2 (-1, 0)	960×288	**0.052**	**0.282**	**2.880**	**0.090**	**0.966**	**0.993**	**0.998**

Table 1. Results and comparison with other state-of-the-arts models on the KITTI [7] Eigen split [4] with improved depth maps [19]. **Bold**: Best, Underscore: Second best. † : evaluated on whole sequences

# iters	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
0	0.103	0.726	4.497	0.181	0.893	0.965	0.983
1	0.103	0.702	4.360	0.179	0.900	0.967	0.984
2	0.099	0.700	4.321	0.174	0.906	0.968	0.984
3	0.098	0.695	4.318	0.175	0.905	0.968	0.984
4	0.095	0.690	4.308	0.174	0.908	0.967	0.984
5	0.092	0.673	4.264	0.172	0.911	0.968	0.984
6	0.090	0.658	4.237	0.171	0.912	0.967	0.984
7	0.089	0.653	4.23	0.172	0.912	0.967	0.983
8	0.090	0.653	4.234	0.173	0.910	0.967	0.983
9	0.091	0.655	4.251	0.174	0.909	0.966	0.983

Table 2. The progression of the errors on the KITTI [7] Eigen split in each DualRefine iteration.

# iters	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
0	0.075	0.379	3.490	0.117	0.936	0.989	0.997
1	0.071	0.329	3.186	0.108	0.950	0.991	0.997
2	0.069	0.324	3.143	0.105	0.953	0.991	0.997
3	0.067	0.319	3.135	0.104	0.953	0.991	0.997
4	0.063	0.307	3.098	0.101	0.956	0.992	0.998
5	0.058	0.291	3.050	0.097	0.959	0.992	0.998
6	0.056	0.281	3.040	0.095	0.960	0.992	0.998
7	0.055	0.278	3.041	0.095	0.960	0.992	0.998
8	0.055	0.279	3.056	0.096	0.958	0.992	0.998
9	0.057	0.283	3.091	0.097	0.957	0.992	0.998

Table 3. The progression of the errors on the KITTI [7] Eigen split [4] with improved depth maps [19] in each DualRefine iteration.

	Loss pairs D^*	Loss pairs T^*	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1 ↑
1	T_0	D_0	0.99	0.765	4.449	0.898
2	T^*	D_0	0.093	0.698	4.342	0.907
3	T_0	D^*	0.092	0.657	4.34	0.908
4	T^*	D^*	0.089	0.632	4.305	0.907

Table 4. Ablation experiment for the effect of pose updates and self-supervision pairings on the KITTI [7] Eigen split. **Bold**: Best.

estimates produced by the initial depth estimator. We can see that our model converges around the 6^{th} iteration. We also plot the Abs Rel error on Fig. 1

3.3. KITTI improved depth DEQ results

We also present detailed DEQ errors in Tab. 3 and plot the Abs Rel error in each iteration on Fig. 2 for the KITTI improved depth ground truth. Similarly, our model converges around the 6^{th} iteration.

3.4. Additional qualitative results

We illustrate through Figs. 3 and 4 additional results in the KITTI dataset. An interesting observation is how the model learns to give low confidence to vehicles and textureless image regions. We also show in Fig. 4 how the epipolar geometry differs between the initial estimates and the refined estimates, which may cause inaccurate photometric losses

Methods	$t_{err}(\%) \downarrow$	$r_{err}(^\circ/100m) \downarrow$	ATE $(m) \downarrow$
ORB-SLAM2 [61]	12.96	**0.7**	44.09
Monodepth2 [22]	12.28	3.1	99.36
Zou *et al*. [102]	7.28	1.4	71.63
DualRefine-initial (T_0)	12.50	4.04	118.29
DualRefine-refined (T^*)	**5.82**	1.51	**17.27**

Table 5. Additional results on KITTI odometry test split (Seq. $11 \sim 21$) using ORB-SLAM2 stereo as pseudo-GT. We provide a comparison with representative state-of-the-art self-supervised depth and odometry methods. ORB-SLAM2 is included as a representative non-learning based method.

as well as matching costs.

4. Additional results on KITTI odometry

We perform an additional evaluation on Seq. 11-21 of the KITTI odometry dataset, using the stereo version of ORB-SLAM2 as a pseudo-GT following Zou *et al*. [23] We present the average results in Tab. 5 The refinement greatly improves over the initial predictions and also displays better ATE even in comparison to ORB-SLAM2 with loop closure.

5. Conv-GRU Update Implementation

In our approach, we use the standard Conv-GRU block [18] to compute the updates as follows:

$$\begin{aligned} z_{k+1} &= \sigma(\mathrm{CNN}_z([h_k, x_k])) \\ r_{k+1} &= \sigma(\mathrm{CNN}_r([h_k, x_k])) \\ \tilde{h}_{k+1} &= \tanh(\mathrm{CNN}_{\tilde{h}}([r_{k+1} \odot h_k, x_k])) \\ h_{k+1} &= (1 - z_{k+1}) \odot h_k + z_{k+1} \odot \tilde{h}_{k+1} \end{aligned} \tag{1}$$

where σ represents the sigmoid activation function. Exploring other variants of the Conv-GRU block will be considered in the future.

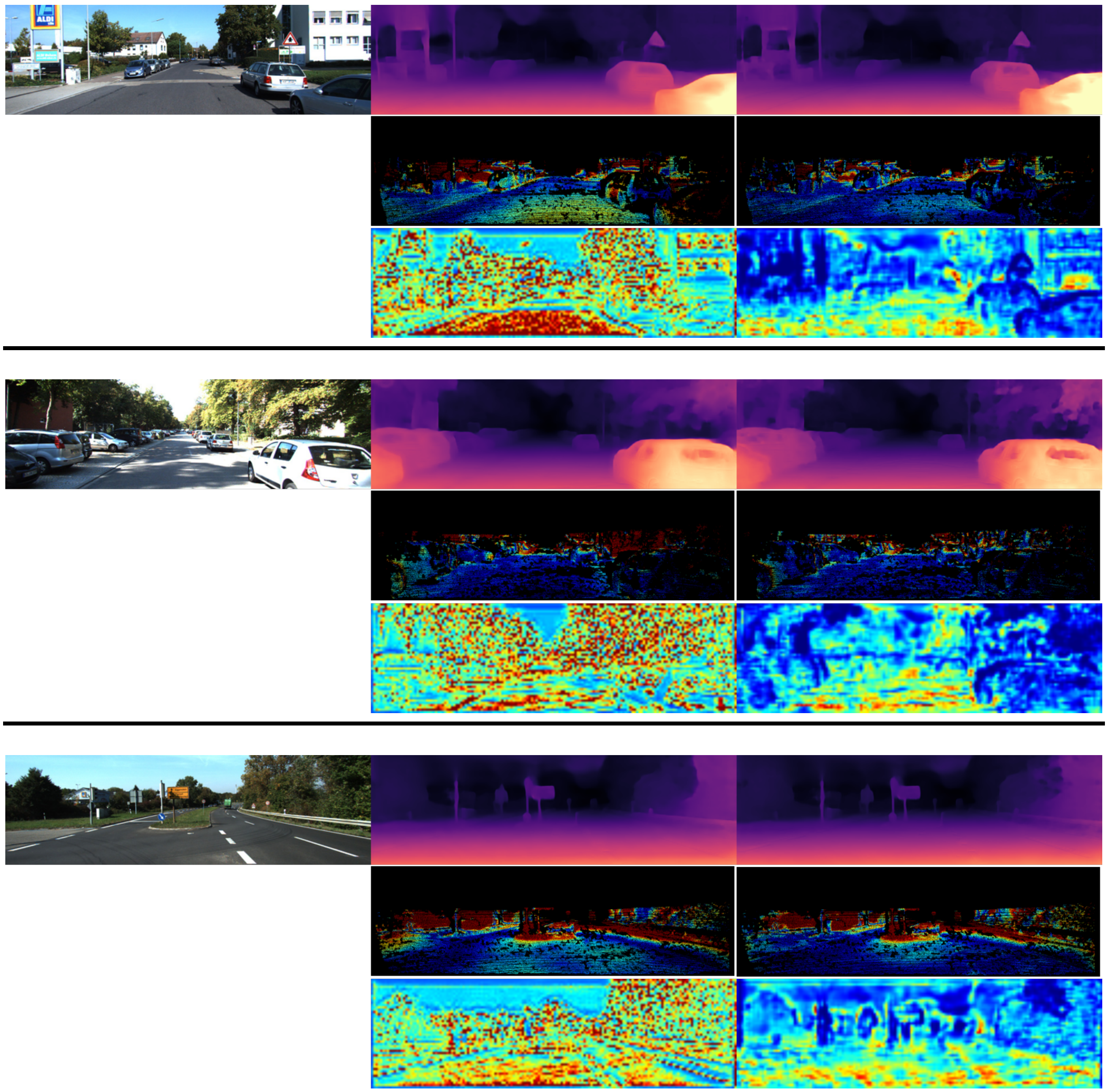

Figure 3. Qualitative results in the KITTI [7] dataset. top left: input image, top center: initial disparity, top right: refined disparity, middle center: initial error map, middle right: refined error map, bottom center: fixed confidence weights, bottom right: 6^{th} iter confidence weights.

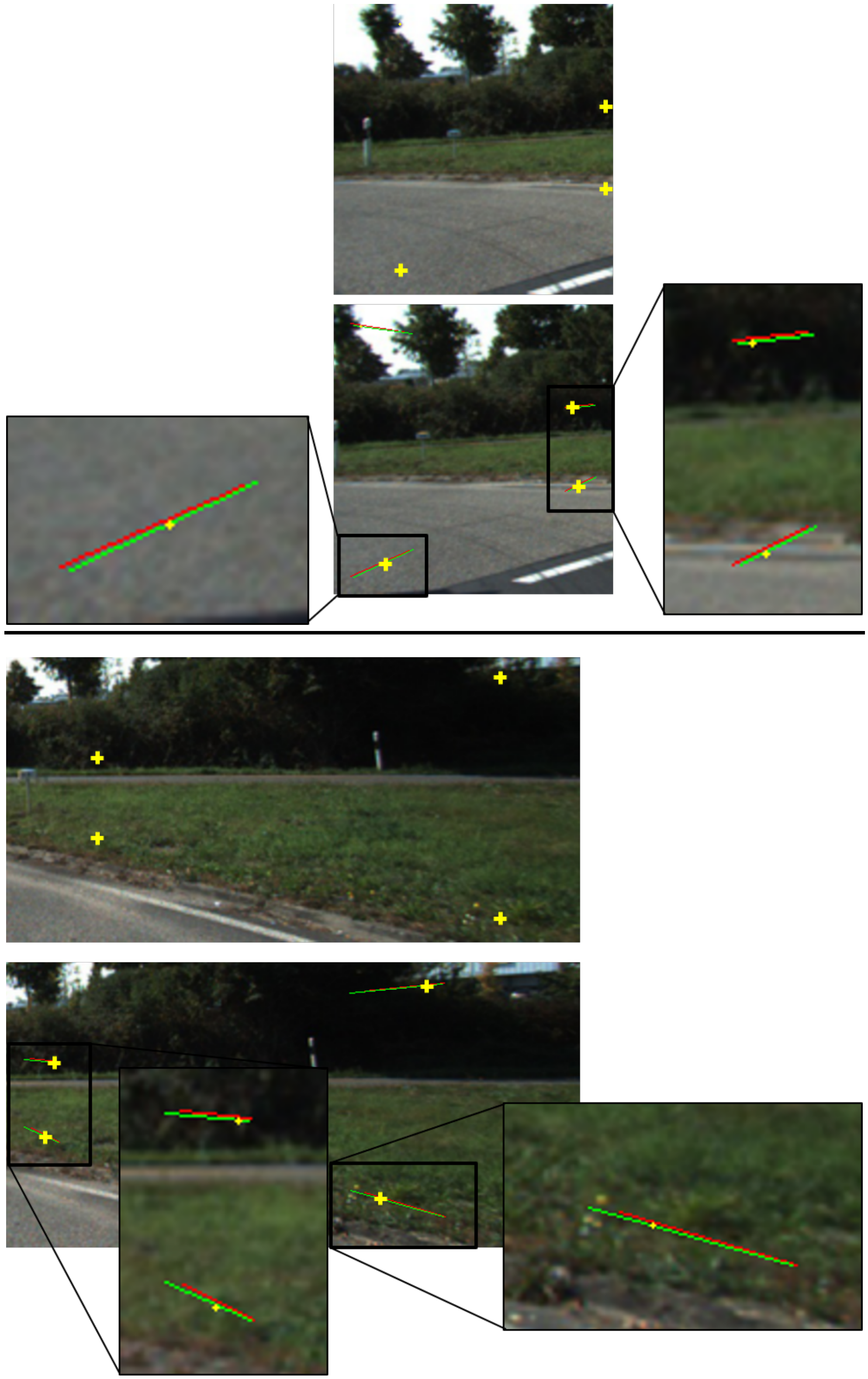

Figure 4. The epipolar line in the source image, calculated from yellow points in the target image, for the PoseNet [14] initial pose (red) and our refined pose (green). The yellow point in the source image is calculated based on our final depth and pose estimates.

References

[1] Donald G Anderson. Iterative procedures for nonlinear integral equations. *Journal of the ACM (JACM)*, 12(4):547–560, 1965. 1

[2] Shaojie Bai, Zhengyang Geng, Yash Savani, and J Zico Kolter. Deep equilibrium optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 620–630, 2022. 1

[3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[4] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 1, 2

[5] Samy Wu Fung, Howard Heaton, Qiuwei Li, Daniel McKenzie, Stanley J Osher, and Wotao Yin. Fixed point networks: Implicit depth models with jacobian-free backprop. 2021. 1

[6] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pages 740–756. Springer, 2016. 1

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1, 2, 4

[8] Zhengyang Geng, Meng-Hao Guo, Hongxu Chen, Xia Li, Ke Wei, and Zhouchen Lin. Is attention better than matrix decomposition? *arXiv preprint arXiv:2109.04553*, 2021. 1

[9] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 2

[10] Xiaodong Gu, Weihao Yuan, Zuozhuo Dai, Chengzhou Tang, Siyu Zhu, and Ping Tan. Dro: Deep recurrent optimizer for structure-from-motion. *arXiv preprint arXiv:2103.13201*, 2021. 2

[11] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020. 2

[12] Vitor Guizilini, Rareș Ambruș, Dian Chen, Sergey Zakharov, and Adrien Gaidon. Multi-frame self-supervised depth with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 160–170, 2022. 1, 2

[13] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 4756–4765, 2020. 2

[14] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 5

[15] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2624–2641, 2019. 2

[16] Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Don't forget the past: Recurrent depth estimation from monocular video. *IEEE Robotics and Automation Letters*, 5(4):6813–6820, 2020. 2

[17] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12240–12249, 2019. 2

[18] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 3

[19] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 1, 2

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[21] Jianrong Wang, Ge Zhang, Zhenyu Wu, XueWei Li, and Li Liu. Self-supervised joint learning framework of depth estimation via implicit cues. *arXiv preprint arXiv:2006.09876*, 2020. 2

[22] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021. 2

[23] Yuliang Zou, Pan Ji, Quoc-Huy Tran, Jia-Bin Huang, and Manmohan Chandraker. Learning monocular visual odometry via self-supervised long-term modeling. In *European Conference on Computer Vision*, pages 710–727. Springer, 2020. 3