Supplementary Material:
Two Stream Networks for Self-Supervised Ego-Motion Estimation

Rares Ambrus **Vitor Guizilini** **Jie Li** **Sudeep Pillai** **Adrien Gaidon**
Toyota Research Institute (TRI)
`firstname.lastname@tri.global`

1 Qualitative Results

We present qualitative results of our method on the training sequences 00-08 of the KITTI [1] odometry benchmark in Figure 1.

3rd Conference on Robot Learning (CoRL 2019), Osaka, Japan.

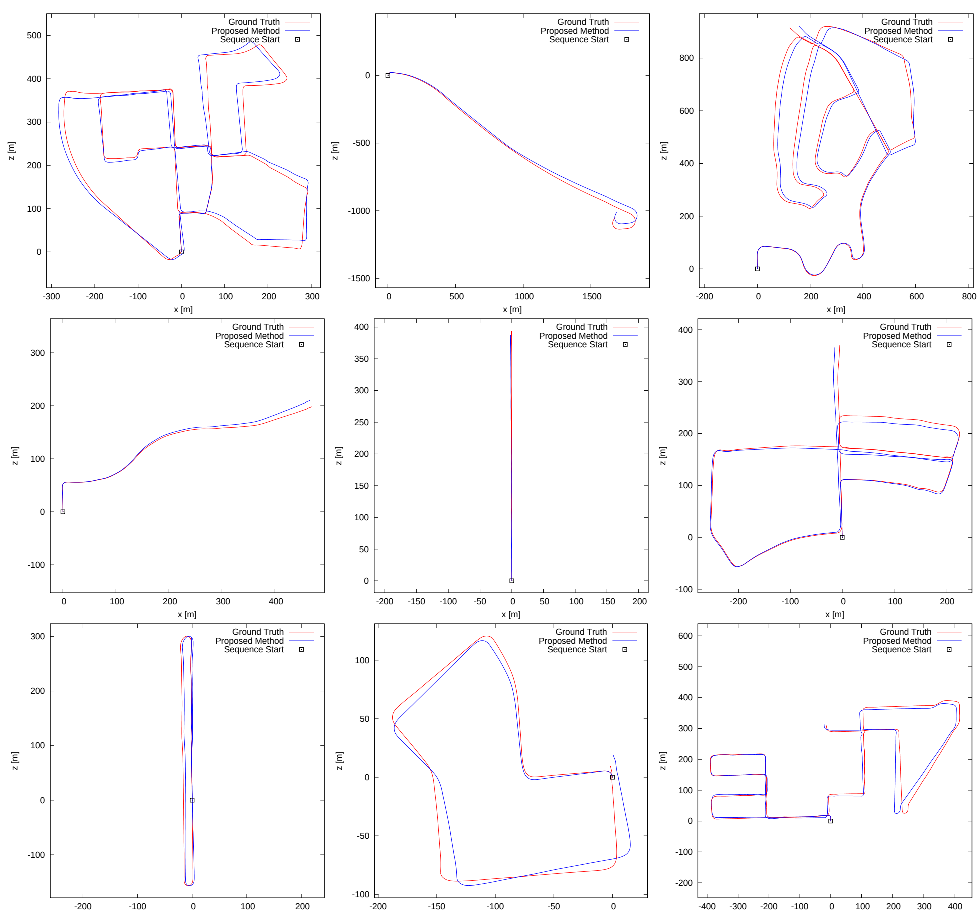

Figure 1: Qualitative trajectory results of the proposed method on train sequences 00-08 of the KITTI odometry benchmark.

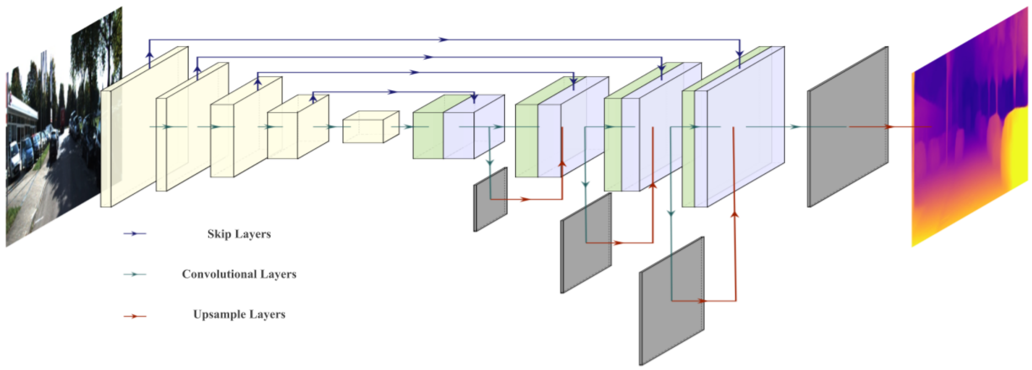

Figure 2: The architecture of our depth estimation network.

2 Depth estimation network architecture

We show in Fig. 2 the architecture of the depth network used. We base our architecture on [2] and follow [3] to add skip connections and output depth at 4 scales.

3 Structural Similarity (SSIM) loss component

As described in [4], the SSIM loss between two images is defined as:

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{1}$$

In all our experiments $C_1 = 1e^{-4}$ and $C_2 = 9e^{-4}$, and we use a $3x3$ block filter to compute μ_x and σ_x - the per-patch mean and standard deviation.

4 Discussion on scaling the monocular predictions

As described in the experimental setup section, in order to be consistent with the experimental protocol of [5, 6] we compute the scaling factor for each prediction by optimizing it over 5-frame long trajectories. However since our network is evaluated on pairs of frames, we have the option of computing the scaling factor using 2-frame trajectories (this is consistent with the way monocular depth methods are evaluated, with the scale being computed for each prediction). We present results when scaling using 2-frame trajectories and when scaling from 5-frame trajectories.

Method	**ATE Seq 09**	**ATE Seq 10**	t_{rel} **train**	t_{rel} **test**	r_{rel} **train**	r_{rel} **test**
Ours - scale from 5-frame trajectories	0.0096 ± 0.002	0.0089 ± 0.002	1.44	2.92	0.64	1.53
Ours - scale from 2-frame trajectories	0.0083 ± 0.002	0.0075 ± 0.002	1.38	2.92	0.64	1.53

Table 1: Results of our method when computing scale using 5-frame tranjectories versus 2-frame trajectories. Our method is trained on the KITTI odometry Sequences 00-08. We report ATE on the test sequences 09 and 10, as well as t_{rel} - average translational RMSE drift ($\%$) on trajectories of length 100-800m, and r_{rel} - average rotational RMSE drift ($^\circ/100m$) on trajectories of length 100-800m, averaged over the training and testing, respectively.

We summarize our analysis in Table 1. Interestingly, the ATE metric improves significantly, while the test t_{rel} metric suffers only minor variations. The r_{rel} metric is unaffected, as the scaling operation only affects the predicted translation between frames.

References

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[2] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016.

[3] S. Pillai, R. Ambrus, and A. Gaidon. SuperDepth: Self-Supervised, Super-Resolved Monocular Depth Estimation. In *ICRA*, 2019.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004.

[5] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017.

[6] C. Godard, O. Mac Aodha, and G. Brostow. Digging into self-supervised monocular depth estimation. *arXiv preprint arXiv:1806.01260*, 2018.