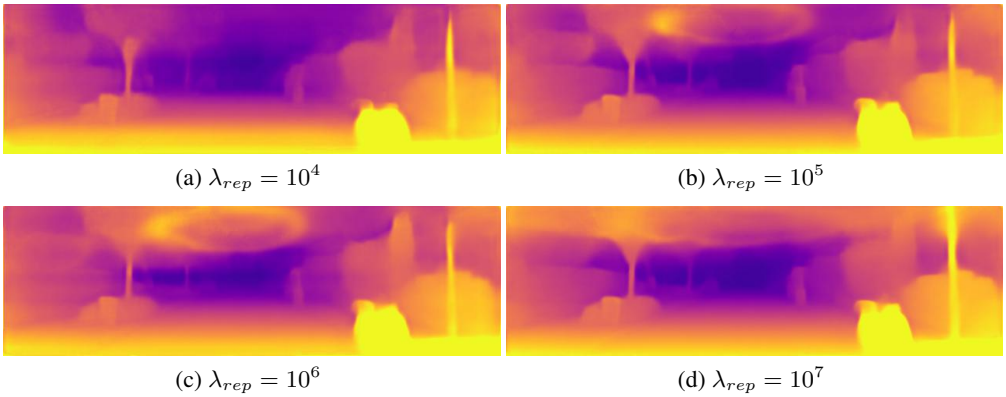
Supplementary Material: Robust Semi-Supervised Monocular Depth Estimation with Reprojected Distances

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

1 Self-Supervised and Supervised Losses Trade-Off

Our proposed semi-supervised loss (Eq. 1, main text) is composed of three individual terms: $\mathcal{L}_{photo}$, representing the self-supervised photometric loss, $\mathcal{L}_{smooth}$, representing a self-supervised smoothness depth regularizer, and $\mathcal{L}_{rep}$, representing the proposed supervised reprojected distance loss. Determining the correct balance between these terms is an important part of the training protocol, and in this section we discuss the effects that λ_{rep}, or the ratio between the self-supervised and supervised components of the loss, has in our overall results.

Interestingly, we did not notice any meaningful changes in numerical results when λ_{rep} varies, even if this variation is by a few orders of magnitude. However, there was a significant difference in how the resulting depth maps are visually represented, as depicted in Fig. 1. In particular, larger values for λ_{rep} promote a worse reconstruction of areas not observed by the LiDAR sensor. We suspect that this behavior is due to the supervised term of the loss overwhelming the self-supervised terms, which hinders the learning of denser, smoother depth maps via the photometric loss. This is supported by the fact that this is a typical behavior of purely supervised depth learning algorithms, where the loss is never calculated in areas where there are no valid depth values. When further lowering λ_{rep}, we started to see degradation in numerical results, indicating that the photometric loss was being over-represented in the loss and scale was not being learned properly, which led us to elect $\lambda_{rep} = 10^4$ as the optimal value for our proposed semi-supervised loss.



(a) $\lambda_{rep} = 10^4$ (b) $\lambda_{rep} = 10^5$

(c) $\lambda_{rep} = 10^6$ (d) $\lambda_{rep} = 10^7$

Figure 1: **Effects of varying the coefficient** λ_{rep} that weights the supervised loss term, for the KITTI dataset. Most noticeably, lower values of λ_{rep} produce a better reconstruction of areas not observed by the LiDAR sensor.

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

2 Degradation in the Number of Supervised Frames

In this section, we provide analysis of our model robustness to another type of degradation in supervision: the number of depth labels available. This is particularly useful as a way to combine large unlabeled datasets, produced without any sort of supervision, with a small amount of labeled images, obtained separately under more controlled circumstances. Our training schedule, on the KITTI dataset, consists of producing two separate splits:

- **Unlabeled** ($\mathcal{U}$): All available images (39810, following the pre-processing steps of [1]) are maintained, discarding all depth information.
- **Supervised** ($\mathcal{S}$): N images are randomly selected from the entire dataset and maintained, alongside their corresponding *Annotated* depth information.

Afterwards, training is performed as instructed, however at each step half the batch size is sampled from $\mathcal{U}$ and half from $\mathcal{S}$, with the former not contributing for the proposed reprojected distance loss $\mathcal{L}_{rep}$ during loss calculation. Note that $\mathcal{S}$ is sampled with replacement, so the same labeled images can be processed multiple times in the same epoch, that is considered finished when all images from $\mathcal{U}$ are processed once. This is done to avoid data imbalance, as the number of training frames from $\mathcal{S}$ decrease relatively to $\mathcal{U}$.

Results obtained using this training schedule are shown in Table 1, indicating that our proposed method statistically did not degrade when observing only 10000 images, roughly 25% of the total of annotated depth maps. Additionally, when observing only 1000 images, or 2.5% the total number of annotated depth maps, our proposed methods achieved performance comparable to Amiri et al. [2] and Luo et al. [3], considered the current state-of-the-art for semi-supervised monocular depth estimation. As we further decrease the number of supervised frames, performance starts to degrade more steeply, however these are mostly due to the model's inability to learn proper scale with such sparse (and possibly biased) information.

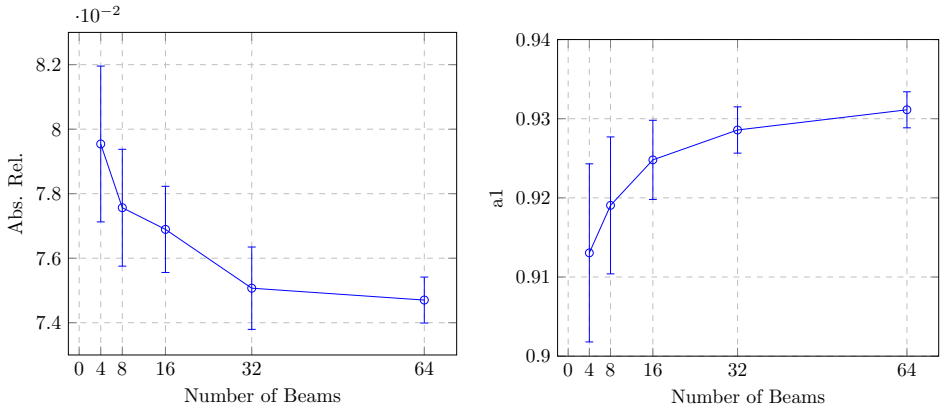
# Sup. Frames	Abs.Rel	Sq.Rel	RMSE	RMSE$_{log}$	$\delta < 1.25$
39810 (all)	0.073 ± 0.001	0.344 ± 0.004	3.273 ± 0.008	0.117 ± 0.001	0.932 ± 0.002
10000	0.074 ± 0.002	0.346 ± 0.006	3.298 ± 0.021	0.118 ± 0.002	0.934 ± 0.002
1000	0.080 ± 0.003	0.388 ± 0.010	3.550 ± 0.038	0.125 ± 0.005	0.923 ± 0.004
100	0.101 ± 0.007	0.532 ± 0.023	4.230 ± 0.078	0.155 ± 0.018	0.886 ± 0.013
10	0.249 ± 0.031	2.832 ± 0.081	10.412 ± 0.380	0.439 ± 0.059	0.561 ± 0.047

Table 1: **Quantitative results** showing how our proposed semi-supervised methodology behaves with a decreasing number of supervised frames at training time, for the KITTI dataset. For each row, statistical intervals were calculated based on 10 independent models trained using different random subsets from $\mathcal{S}$. For **all**, the entire $\mathcal{S}$ was used in all 10 sessions, with the statistical intervals being indicative of the noise inherent to stochastic training and random data augmentation.

3 Effects of Beam Selection for Sparse Depth Labels

In this section we explore how sensitive our semi-supervised depth estimates are to the selection of beams at training time, particularly as depth labels become sparser. In other words, we would like to investigate how the distribution of valid depth pixels throughout annotated labels impact overall results. In our original experiments, beam sparsification was achieved by keeping only those at equally spaced intervals, and by increasing these intervals the number of beams decreases. Naturally, when all 64 beams are used there is no interval, when 32 are used every second beam is kept, when 16 are used every fourth beam is kept, and so forth. It is important to note that not all beams are necessarily used by the reprojected depth map, since their point of contact might not be visible by the camera. In fact, we noticed that most of the information contained in beams below the 45^{th} is discarded, which makes the task of sparse semi-supervision even more challenging.

In order to vary the position of depth information in the resulting sparse labels, while maintaining a proper distribution similar to what a real LiDAR sensor would provide, we opted for introducing an offset, determining where the top beam is located. Starting from 0, this offset increases until it



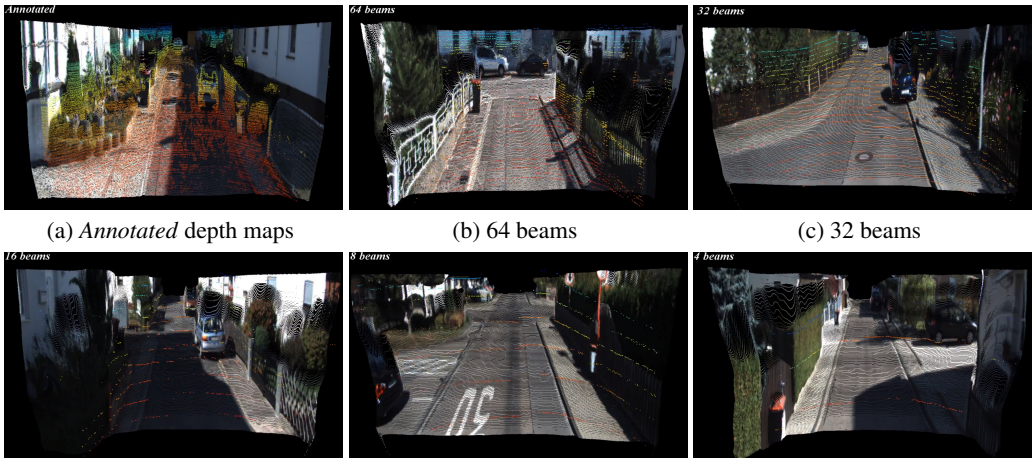
(a) Absolute Relative Error (abs_rel). (b) Threshold 1 ($\delta < 1.25$)

Figure 2: **Effects of beam selection** in monocular depth estimation performance, for different beam distributions. The error bars indicate the variation in depth estimates when different offset values for the top beam are considered. For 64 beams, since there is no variation, the error bars are indicative of the noise inherent to stochastic training and random data augmentation.

coincides with another beam that was selected when no offset is considered. Following this strategy, when 32 beams are considered there are 2 variations, when 16 beams are considered there are 4, and so forth. The results when using this strategy are depicted in Fig. 2, where we can see that sparser depth labels are more sensitive to the distribution of valid pixels, and there are indeed some configurations that lead to better results, however there was no configuration that resulted in catastrophic failures. Interestingly, as we further increased sparsity, considering only 2 or even 1 beam, some configurations failed to converge, showing that there is a limit to how much sparsity can be properly leveraged in our proposed semi-supervised learning framework, however a more thorough analysis is left for future work.

4 Additional Qualitative Results

Here we provide some more qualitative results of our proposed semi-supervised monocular depth estimation methodology, using the reprojected distance loss, on the KITTI dataset. Fig. 4 shows corresponding input RGB images and output depth maps, while Fig. 3 depicts reconstructed point-clouds from models trained using different numbers of LiDAR beams. More qualitative results can be found on the supplementary video attached.



(a) *Annotated* depth maps (b) 64 beams (c) 32 beams

(d) 16 beams (e) 8 beams (f) 4 beams

Figure 3: **Reconstructed point-clouds** from our proposed semi-supervised depth estimation methodology, with models trained using different numbers of LiDAR beams.

Figure 4: **Qualitative results** of our proposed semi-supervised monocular depth estimation methodology, showing input RGB images and output depth maps.

References

[1] 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.

[2] A. J. Amiri, S. Y. Loo, and H. Zhang. Semi-supervised monocular depth estimation with left-right consistency using deep neural network. *arXiv preprint arXiv:1905.07542v1*, 2019.

[3] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018.