Exploiting Pseudo Labels in a Self-Supervised Learning Framework for Improved Monocular Depth Estimation

Supplementary Material

Overview. We propose a self-distillation based self-supervised depth estimation learning framework (SD-SSMDE) and achieve state-of-the-art results on the KITTI and Cityscapes datasets. In this supplementary material, we provide more experimental results.

A. Inference Time

In Table 8 we measure the inference time on an NVIDIA Tesla V100 GPU and generate the number of multiply-add computations (MACs) with the PyTorch 1.7.1 framework and the THOP library[1]. We compare with the baseline Monodepth2 [18] and FSRE-Depth [28]. Our model is more accurate than Monodepth2, but is slightly more computational intensive. On the other hand, our model is much more efficient than FSRE-Depth which provides similar absolute relative error. FSRE-Depth is slower because it employs a semantic segmentation network for cross-task feature refinement.

Model	Backbone	Resolution	MACs	Time (ms)	AbsRel ↓
Monodepth2 [18]	ResNet-18	192 × 640	**8.0**	**11**	0.115
FSRE-Depth [28]	ResNet-18	192 × 640	20.4	18	**0.105**
SD-SSMDE	ResNet-18	192 × 640	10.8	12	0.106
Monodepth2 [18]	ResNet-18	320 × 1024	**21.4**	**12**	0.115
FSRE-Depth [28]	ResNet-18	320 × 1024	54.5	29	0.102
SD-SSMDE	ResNet-18	320 × 1024	28.8	15	**0.101**
Monodepth2 [18]	ResNet-50	192 × 640	**16.6**	**15**	0.110
FSRE-Depth [28]	ResNet-50	192 × 640	32.0	29	0.102
SD-SSMDE	ResNet-50	192 × 640	18.6	17	**0.100**
SD-SSMDE	ResNet-50	320 × 1024	44.3	22	0.098

Table 8. **Inference time and MACs.**

B. Results on improved KITTI

Table 9 shows our results on the KITTI Eigen set with improved ground truth [51]. We experiment with a fixed scaling factor and ground truth median scaling. We outperform the Monodepth2 baseline [18], with both scaling methods, with ResNet-18 and ResNet-50 backbones and on both medium and high resolution. We observe the largest performance difference when using a fixed scale factor. Monodepth2 does not enforce scale consistency between depth predictions, therefore it has a large scale variance across frames. As a result, the error increases when using a single scale for all predictions. On the other hand, our model performs better due to the fact that it learns from scale-consistent pseudo-labels and in consequence, depth predictions are inter-frame scale-consistent. This allows us to use a single scale factor for all depth predictions with minimal loss in accuracy compared to ground truth median scaling.

[1] https://github.com/Lyken17/pytorch-OpCounter

C. Experiments on Cityscapes

In this section we give more details about the training and evaluation schemes used on the Cityscapes dataset. During evaluation, we center crop the original image of size 1024 × 2048 to 512 × 1664 following [2, 54]. During the training of the self-supervised teacher network, we center crop the images to 768 × 2048. We generate high-resolution pseudo-labels of size 768 × 2048. In the second stage of training, we train the student network and follow the same cropping scheme used for evaluation, but we scale the images to 128 × 416. In Table 10 we present the full set of metrics for the Cityscapes experiments as an extension of Table 7 from the main paper.

D. Qualitative Results on KITTI

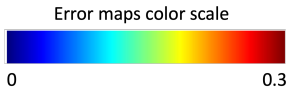

Figure 5. **Color scale used to generate depth error maps.**

In Figure 5 we present the color scale used to generate absolute error maps for our depth predictions in Figures 1, 4, 6. Blue indicates low error and red high error. In Figure 6 we present a qualitative comparison between [18,22,29] and our results obtained with our self-supervised teacher network, but also the final results from the pseudo-supervised student network. We observe that both our teacher and student networks yield more accurate depth maps.

Method	Scaling	Backbone	Resolution	AbsRel ↓	SqRel ↓	RMS ↓	RMSlog ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Monodepth2 [18]	GT	ResNet-18	192 × 640	0.084	0.481	3.757	0.129	0.923	0.985	0.996
SD-SSMDE	GT	ResNet-18	192 × 640	0.079	0.399	3.442	0.121	0.929	0.986	0.997
SD-SSMDE	GT	ResNet-50	192 × 640	**0.072**	**0.347**	**3.219**	**0.111**	**0.941**	**0.990**	**0.998**
Monodepth2 [18]	Fixed	ResNet-18	192 × 640	0.104	0.561	3.961	0.147	0.882	0.980	0.995
SD-SSMDE	Fixed	ResNet-18	192 × 640	0.084	0.436	3.550	0.128	0.918	0.985	0.997
SD-SSMDE	Fixed	ResNet-50	192 × 640	**0.076**	**0.377**	**3.304**	**0.117**	**0.933**	**0.988**	**0.997**
Monodepth2 [18]	GT	ResNet-18	320 × 1024	0.085	0.450	3.542	0.126	0.925	0.987	0.996
SD-SSMDE	GT	ResNet-18	320 × 1024	0.072	0.344	3.255	0.112	0.940	0.989	0.998
SD-SSMDE	GT	ResNet-50	320 × 1024	**0.068**	**0.311**	**3.077**	**0.106**	**0.947**	**0.991**	**0.998**
Monodepth2 [18]	Fixed	ResNet-18	320 × 1024	0.096	0.504	3.691	0.137	0.903	0.984	0.996
SD-SSMDE	Fixed	ResNet-18	320 × 1024	0.077	0.370	3.338	0.118	0.931	0.988	0.997
SD-SSMDE	Fixed	ResNet-50	320 × 1024	**0.074**	**0.338**	**3.144**	**0.112**	**0.939**	**0.990**	**0.998**

Table 9. **Evaluation on KITTI Eigen set with improved ground truth [51].** During inference, we scale the depth predictions either using the ground truth median (GT) or a fixed scale factor. We evaluate Monodepth2 [18] with the authors' code.

Model	Train	Test	AbsRel ↓	SqRel ↓	RMS ↓	RMSlog ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Struct2Depth 2 [2]	C	C	0.145	1.737	7.280	0.205	0.813	0.942	0.976
Monodepth2 [18]	C	C	0.129	1.569	6.876	0.187	0.849	0.957	0.983
Videos in the Wild [20]	C	C	0.127	1.330	6.960	0.195	0.830	0.947	0.981
Li *et al.* [35]	C	C	0.119	1.290	6.980	0.190	0.846	0.952	0.982
Choi *et al.* [8]	C	C	0.115	1.125	6.584	0.195	0.857	0.963	0.986
SD-SSMDE (teacher - GT scaling)	C	C	0.117	1.090	6.468	0.176	0.856	0.964	0.990
SD-SSMDE (student - fixed scaling)	C	C	0.114	1.017	5.949	0.169	0.870	0.967	0.990
SD-SSMDE (student - GT scaling)	C	C	**0.110**	**0.988**	**5.953**	**0.165**	**0.876**	**0.970**	**0.991**
Monodepth2 [18]	K	C	0.153	1.785	8.590	0.234	0.774	0.926	0.976
SD-SSMDE (student - fixed scaling)	K	C	**0.143**	**1.635**	**8.441**	**0.221**	**0.789**	**0.931**	**0.980**

Table 10. **Full metrics on Cityscapes.** Evaluation of models on the Cityscapes dataset, trained on Cityscapes (C) or on KITTI (K).

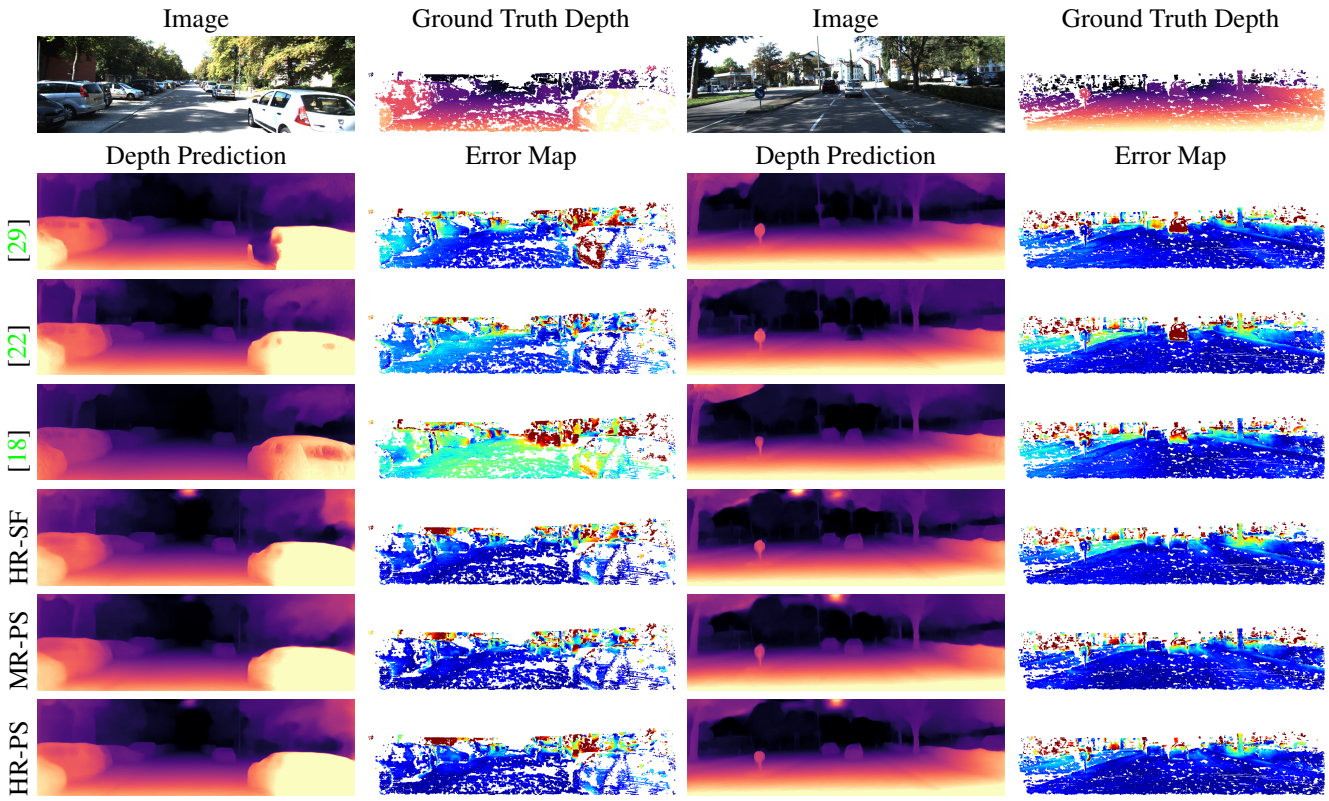

Figure 6. **Qualitative results on the KITTI Eigen test set with improved ground truth.** We compare our results with [18, 22, 29]. Both [29] and [22] employ external data, such as semantic segmentation and use medium resolution images. Monodepth2 [18] with ResNet-50 is our baseline on high resolution images (HR). HR-SF is the output from the first stage of training, $i.e.$ self-supervised teacher network, on high resolution images (320×1024), this network is used for generating pseudo-labels. And finally, MR-PS and HR-PS represent the output from the second-stage training of the student network, supervised with pseudo-labels on medium and high resolution images, respectively. Our HR-PS network provides the most qualitative depth maps as reflected in Table 6.