# Supplementary

## A Experimental Setup

In this part, we want to provide some additional details regarding our experimental setup, which allow a deeper understanding into our experimental setup.

### A.1 Detailed Dataset Overview

In Table 4, we give an overview over the number of images in our used data (sub)sets. The Cityscapes dataset has 2,975 labeled training images on which we train the semantic segmentation part of our network. As we do not optimize our hyperparameters for the semantic segmentation and thereby do not use the validation set during training, our evaluation on this dataset is conducted on the official validation set containing 500 labeled images.

While we always train the segmentation part of our model on the Cityscapes dataset, the depth part of the network is trained on various splits of the KITTI dataset. The split of the KITTI dataset, which is most frequently used to compare depth estimation models, is the *Eigen split* [2], containing 697 images for testing. While the number of test images is constant throughout recent approaches, the number of training and validation images has been redefined by [12] to exclude static scenes. We also compare our method on the *Benchmark split* [10], which contains 500 test images with labels, which are only available on an evaluation server.

Finally, we train and evaluate on the *KITTI split* [4], whose test set are the official 200 training images from the KITTI 2015 Stereo dataset [8]. This test set has the advantage that it has available labels for both depth and semantic segmentation, which makes it suitable to observe the benefits of multi-task training for depth and semantic segmentation. While the Cityscapes validation set in principle also provides labels for both tasks, here the depth labels are obtained by a classical model-based algorithm, while the depth labels of the KITTI dataset are physical measurements from a LiDAR sensor and thereby better suited for evaluating a depth estimation model. Also, as our depth estimation training requires a preceding and a succeeding frame, the number of training images differs slightly from the original definition.

### A.2 Definition of the DC Object Classes

Also, we defined the DC object classes as all classes belonging to the human and vehicle categories inside the Cityscapes dataset [1] which contains in total 19 labeled classes. More specifically, that means that we consider the person, rider, car, truck, bus, train, motorcycle and bicycle class as DC object classes, as they are often observed as moving inside an image sequence. Opposed to that

**Table 4:** Overview over the used databases and available labels. Labels only available on a benchmark server are denoted by "(✓)".

| Dataset | Subset | # Images | Depth Labels | Segmentation Labels |
|---|---|---|---|---|
| Eigen split | train | 21,880 | ✓ | ✗ |
| | val | 4,187 | ✓ | ✗ |
| | test | 697 | ✓ | ✗ |
| Benchmark split | train | 36,040 | ✓ | ✗ |
| | val | 3,030 | ✓ | ✗ |
| | test | 500 | (✓) | ✗ |
| KITTI split | train | 28,937 | ✓ | ✗ |
| | val | 1,158 | ✓ | ✗ |
| | test | 200 | ✓ | ✓ |
| Cityscapes | train | 2,975 | ✓ | ✓ |
| | val | 500 | ✓ | ✓ |
| | test | 1,525 | ✓ | (✓) |

the classes road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain and sky are considered as static, as they are usually not in motion.

### A.3 Evaluation Metrics

In Section 4, we simply referred to previous approaches for the exact definition of the evaluation metrics. In this section, we provide the exact mathematical expressions which are used to evaluate the predicted depth maps $\boldsymbol{d}_t$ with available depth label $\overline{\boldsymbol{d}}_t$ as well as to evaluate the predicted segmentation maps $\boldsymbol{m}_t$ with regard to the ground truth label $\overline{\boldsymbol{m}}_t$. Note that the depth maps are evaluated using a sparse ground truth, where only the pixels with an available LiDAR measurement are considered during evaluation. Also, we apply median scaling to the predicted depth maps before evaluation to compensate the global scale ambiguity [12].

While the defintions of all metrics are equal for all data(sub)sets, there are two exceptions: On the Eigen test split we apply a crop defined by [2], which is in accordance with previous approaches, while on the Benchmark test split we cannot apply median scaling, as the depth labels from the evaluation server are not freely available. Therefore, we determine the median over all image-wise scale factors on the validation set and use this value as a global scale factor for our predictions on the test set. Also note that the subsequent metrics (12)-(21) are to be averaged over the respective test subset respectively.

The four error depth metrics used for evaluation on the Eigen and KITTI split are defined as:

$$\text{Abs Rel} = \frac{1}{HW} \sum_{i \in \mathcal{I}} \frac{|d_i - \overline{d}_i|}{\overline{d}_i}, \tag{12}$$

$$\text{Sq Rel} = \frac{1}{HW} \sum_{i \in \mathcal{I}} \frac{\left(d_i - \overline{d}_i\right)^2}{\overline{d}_i}, \tag{13}$$

$$\text{RMSE} = \sqrt{\frac{1}{HW} \sum_{i \in \mathcal{I}} \left(d_i - \overline{d}_i\right)^2}, \tag{14}$$

$$\text{RMSE log} = \sqrt{\frac{1}{HW} \sum_{i \in \mathcal{I}} \left(\log d_i - \log \overline{d}_i\right)^2}, \tag{15}$$

with $\mathcal{I}$ being the set of all pixels and $H$ and $W$ being the width and height of the image, respectively. The accuracy metrics are defined as follows:

$$\text{The "}\delta < 1.25\text{" measure equals } \frac{1}{HW} \sum_{i \in \mathcal{I}} \left[ \max\left( \frac{d_i}{\overline{d}_i}, \frac{\overline{d}_i}{d_i} \right) < 1.25 \right], \tag{16}$$

$$\text{the "}\delta < 1.25^2\text{" measure equals } \frac{1}{HW} \sum_{i \in \mathcal{I}} \left[ \max\left( \frac{d_i}{\overline{d}_i}, \frac{\overline{d}_i}{d_i} \right) < 1.25^2 \right], \tag{17}$$

$$\text{the "}\delta < 1.25^3\text{" measure equals } \frac{1}{HW} \sum_{i \in \mathcal{I}} \left[ \max\left( \frac{d_i}{\overline{d}_i}, \frac{\overline{d}_i}{d_i} \right) < 1.25^3 \right], \tag{18}$$

$$\tag{19}$$

where $[\cdot]$ is defined as the Iverson bracket, which is 1 if the condition inside the bracket is true, and 0 if the condition is false. Furthermore on the benchmark split there are two more metrics, which are defined as

$$\text{SILog} = \frac{1}{HW} \sum_{i \in \mathcal{I}} \left(\log d_i - \log \overline{d}_i\right)^2 - \frac{1}{(HW)^2} \left( \sum_{i \in \mathcal{I}} \log d_i - \log \overline{d}_i \right)^2, \tag{20}$$

$$\text{iRMSE} = \sqrt{\frac{1}{HW} \sum_{i \in \mathcal{I}} \left( \frac{1}{d_i} - \frac{1}{\overline{d}_i} \right)^2}. \tag{21}$$

Finally, we evaluate our semantic segmentation using the mean intersection over union (mIoU) metric, which is defined as:

$$\text{mIoU} = \frac{1}{S} \sum_{s \in \mathcal{S}} \frac{\text{TP}_s}{\text{TP}_s + \text{FP}_s + \text{FN}_s}, \tag{22}$$

where $\mathcal{S} = \{1, 2, ..., S\}$ is the set of all classes defined in the Cityscapes dataset as described in Section A.2. Considering all labeled pixels for class $s \in \mathcal{S}$

in the predicted segmentation map $\boldsymbol{m}_t$, $\mathrm{TP}_s$ is the number of true positive predictions, $\mathrm{FP}_s$ is the number of false positive predictions, and $\mathrm{FN}_s$ is the number of false negative predictions. Note that while all depth metrics are computed image-wise and then averaged over all images inside the test set, for the mIoU calculation first $\mathrm{TP}_s$, $\mathrm{FP}_s$ and $\mathrm{FN}_s$ are summed up for all images of the test set and only afterwards the mIoU is calculated.

## B Evaluation

In this part, we give some additional examples of our proposed SGDepth method in comparison to several depth estimation baselines and also in comparison to our method trained without the semantic guidance (SGDepth only depth).

### B.1 Depth Comparison to Baselines

In this section, we provide additional examples of the proposed SGDepth method, which we compare to results of the baseline approaches. All models were trained and tested on the Eigen splits [2] of the KITTI dataset [3].

In the examples of Figure 7 two things can be observed. Firstly, the depth predictions of our full SGDepth method are sharpened at object boundaries. This effect can be observed especially for small objects such as traffic lights or traffic signs as, e.g., in rows 1 and 3 from the top and row 3 from the bottom. This effect is mainly observed due to the joint training approach of depth estimation and semantic segmentation as thereby the encoder better learns to extract object boundaries, provided by the semantic segmentation, which in return guides the depth estimation to predict sharper edges at these boundaries. We also suspect that this effect is not even fully considered by the numerical evaluation as the ground truth depth labels only cover about the bottom two thirds of the image and many traffic signs and traffic lights are above this zone.

Secondly, our approach also allows for better learned depth of DC objects, as, e.g., in rows 1, 2 and 6 from the bottom, where especially the pedestrians and cyclists are more sharply visible inside the depth map. This is most likely due to our semantic masking technique, where the depth of DC objects is mainly learned from frames containing rather non-moving DC objects.

### B.2 Benefits of Multi-Task Training

In this section we show additional examples of our SGDepth method for comparison with baselines trained only for the single tasks of depth estimation or semantic segmentation, respectively. The models were all trained and tested on the KITTI splits defined by [4].

The benefits of joint training for the depth estimation as discussed in the previous section also apply for our comparison to our own baseline, where one can clearly see the benefits of the semantic guidance for each single image inside Figure 8. However, also the semantic segmentation maps improve, compared to

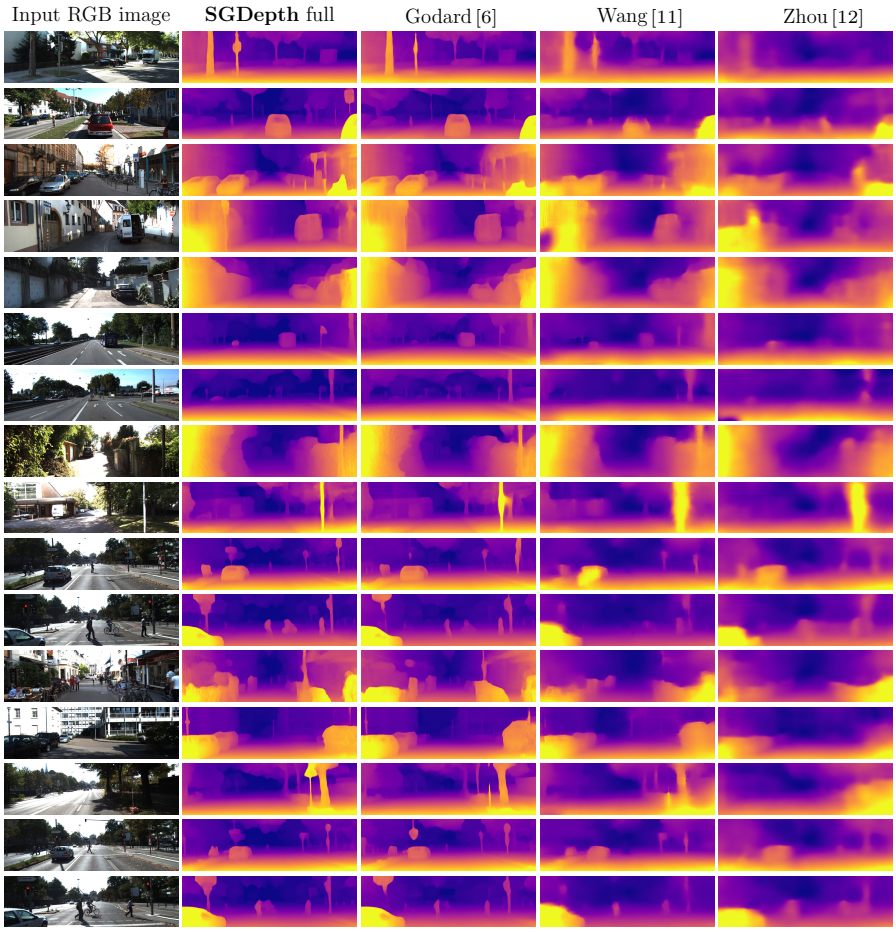


**Fig. 7:** Additional examples of our proposed full SGDepth method in comparison to baseline methods. The figure is best viewed on screen and in color.

a semantic segmentation baseline (SGDepth only seg.), which was solely trained for the task of semantic segmentation (on the Cityscapes dataset). As stated in Section 5.3, we believe that this improvement on the KITTI dataset is due to the fact that through the self-supervised depth estimation, suitable features for the KITTI dataset are extracted, which bridge the domain shift between the Cityscapes and the KITTI dataset. This claim is also supported by the qualitative results, which appear clearly improved compared to the baseline.

### B.3 KITTI Eigen Split Ablation

We trained and optimized the parameters of our different model variants on the KITTI split [4] to observe the resulting performance on both tasks, depth and

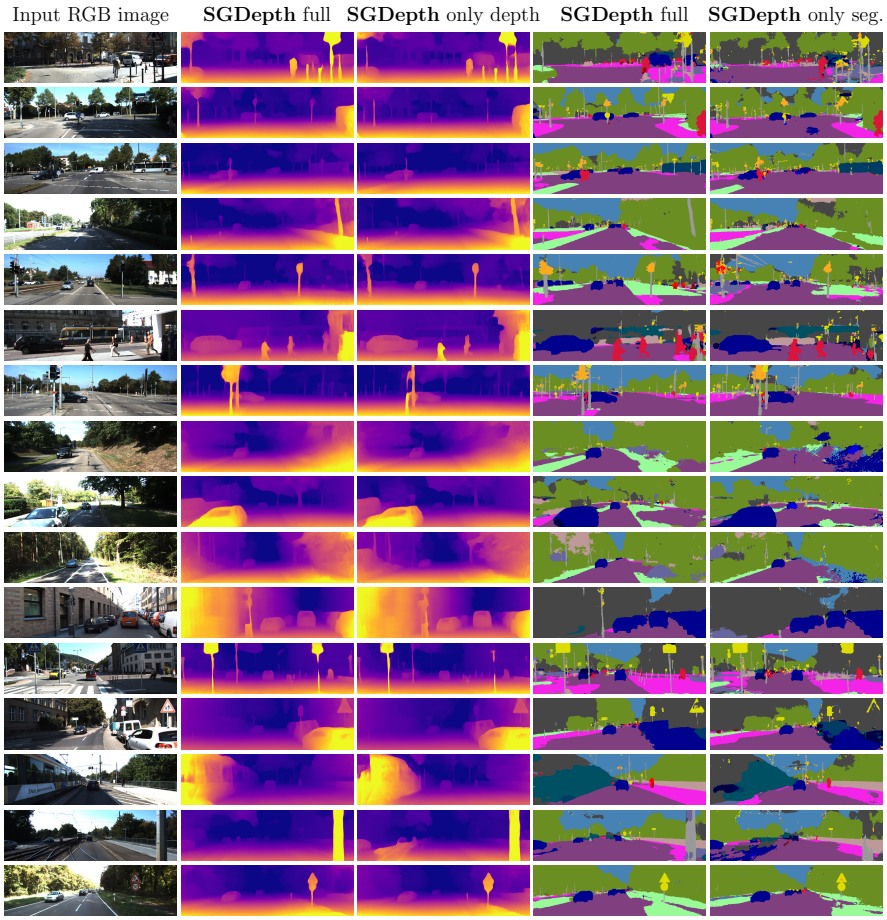


**Fig. 8:** Additional examples on how the **full** SGDepth model compares to the models trained only on the **single tasks** of depth estimation and semantic segmentation, respectively. The figure is best viewed on screen and in color.

semantic segmentation. In the end we only evaluated the final obtained models on the Eigen split benchmark. However, for completeness we also provide the same ablation experiments as executed on the KITTI split on the Eigen split in Table 5. We observe that all multi-task models outperform the single-task baseline (SGDepth only depth) and that our final model (SGDepth full) is best in the important metrics Abs. Rel. and $\delta < 1.25$, as has been observed on the KITTI split as well. *Thereby, our ablation on the Eigen split confirms our ablation experiments on the KITTI split.*

**Table 5: Ablation study** of different models on the **KITTI Eigen split**. CS indicates training of the depth estimation on Cityscapes, K training on the KITTI Eigen split, and (CS) training of the segmentation branch on Cityscapes. **Best results** at each resolution are written in **boldface**.

| | | | Lower is better | | | | Higher is better | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | Resolution | Dataset | Abs Rel | Sq Rel | RMSE | RMSE log | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
| **SGDepth** only depth | 640 × 192 | K | 0.117 | 0.907 | 4.844 | 0.196 | 0.875 | 0.958 | 0.980 |
| **SGDepth** add multi-task training | 640 × 192 | (CS) + K | 0.117 | 0.918 | 4.777 | 0.193 | 0.872 | 0.960 | **0.982** |
| **SGDepth** add scaled gradients | 640 × 192 | (CS) + K | **0.113** | **0.817** | **4.671** | 0.191 | 0.877 | **0.961** | **0.982** |
| **SGDepth** add semantic mask | 640 × 192 | (CS) + K | 0.116 | 0.917 | 4.726 | **0.189** | 0.874 | **0.961** | **0.982** |
| **SGDepth** add threshold | 640 × 192 | (CS) + K | **0.113** | 0.861 | 4.724 | 0.191 | **0.879** | 0.960 | 0.981 |
| **SGDepth** full | 640 × 192 | (CS) + K | **0.113** | 0.835 | 4.693 | 0.191 | **0.879** | **0.961** | 0.981 |

**Table 6:** Pose estimation results on the KITTI odometry dataset sequences 9 and 10.

| Method | Sequence 9 | Sequence 10 | # frames |
|---|---|---|---|
| Zhou et al. [12] | $0.021 \pm 0.017$ | $0.020 \pm 0.015$ | 5 |
| Godard et al. [5] | $0.017 \pm 0.008$ | $0.015 \pm 0.010$ | 2 |
| Luo et al. [7] | $0.013 \pm 0.007$ | $\mathbf{0.012 \pm 0.008}$ | 3 |
| Ranjan et al. [9] | $\mathbf{0.012 \pm 0.007}$ | $\mathbf{0.012 \pm 0.008}$ | 5 |
| **SGDepth** only depth | $0.017 \pm 0.009$ | $0.014 \pm 0.010$ | 2 |
| **SGDepth** full | $0.019 \pm 0.010$ | $0.016 \pm 0.010$ | 2 |

### B.4 Pose Evaluation

Although the focus of our work is on depth estimation, we also provide results of our pose estimation network evaluated with the same strategy as introduced in [6, 12]. We trained on the sequences 0 to 8 of the KITTI odometry dataset and evaluated our models on the sequences 9 and 10 with the results compared to baselines shown in Table 6. Interestingly, the joint training of depth and semantic segmentation seems to have a negative effect on the pose estimation, whose optimization through multi-task learning could be subject to future works. Nevertheless we achieve competitive results compared to the baselines [7, 9], in particular when considering that most of them use more than 2 input images for pose estimation at test time.

## References


1. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: Proc. of CVPR. pp. 3213–3223. Las Vegas, NV, USA (Jun 2016)
2. Eigen, D., Puhrsch, C., Fergus, R.: Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. In: Proc. of NIPS. pp. 2366–2374. Montréal, QC, Canada (Dec 2014)
3. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision Meets Robotics: The KITTI Dataset. International Journal of Robotics Research (IJRR) **32**(11), 1231–1237 (Aug 2013)
4. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised Monocular Depth Estimation With Left-Right Consistency. In: Proc. of CVPR. pp. 270–279. Honolulu, HI, USA (Jul 2017)

5. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging Into Self-Supervised Monocular Depth Estimation. arXiv (1806.01260v4) (Jun 2018)
6. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging Into Self-Supervised Monocular Depth Estimation. In: Proc. of ICCV. pp. 3828–3838. Seoul, Korea (Oct 2019)
7. Luo, C., Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R., Yuille, A.: Every Pixel Counts ++: Joint Learning of Geometry and Motion with 3D Holistic Understanding. arXiv (1810.06125) (Jul 2019)
8. Menze, M., Geiger, A.: Object Scene Flow for Autonomous Vehicles. In: Proc. of CVPR. pp. 3061–3070. Boston, MA, USA (Jun 2015)
9. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation. In: Proc. of CVPR. pp. 12240–12249. Long Beach, CA, USA (Jun 2019)
10. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity Invariant CNNs. In: Proc. of 3DV. pp. 11–20. Verona, Italy (Oct 2017)
11. Wang, C., Miguel Buenaposada, J., Zhu, R., Lucey, S.: Learning Depth From Monocular Videos Using Direct Methods. In: Proc. of CVPR. pp. 2022–2030. Salt Lake City, UT, USA (Jun 2018)
12. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised Learning of Depth and Ego-Motion from Video. In: Proc. of CVPR. pp. 1851–1860. Honolulu, HI, USA (Jul 2017)