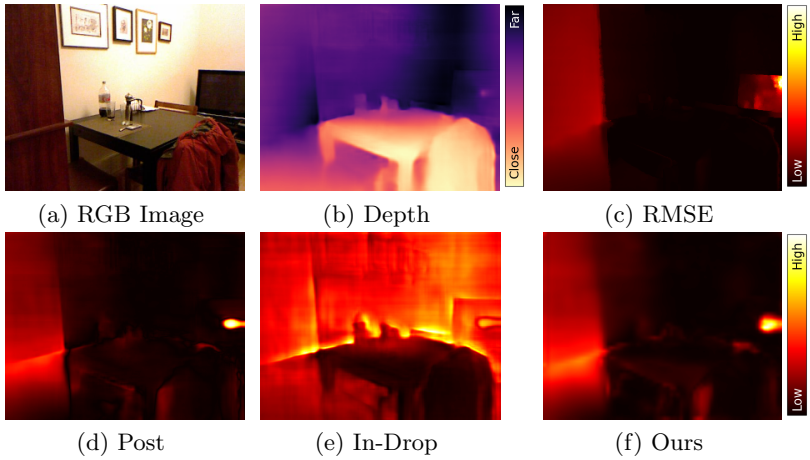
Supplementary Material
Gradient-based Uncertainty for Monocular Depth Estimation

Julia Hornauer[1] and Vasileios Belagiannis[2⋆]

[1] Institute of Measurement, Control and Microtechnology, Ulm University, Germany
julia.hornauer@uni-ulm.de
[2] Department of Simulation and Graphics, Otto von Guericke University Magdeburg, Germany

1 Further Visual Results

In Fig. 1 to Fig. 4, further visual results are provided. The examples show the RGB images, the depth prediction, the true error in terms of RMSE as well as the uncertainty obtained from the base model, inference dropout (*In-Drop*) and the gradients extracted from the neural network (Ours).

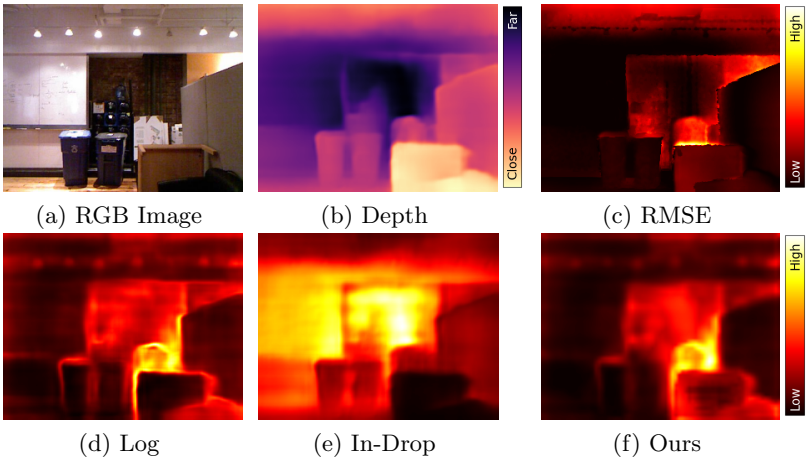


(a) RGB Image (b) Depth (c) RMSE

(d) Post (e) In-Drop (f) Ours

Fig. 1. Uncertainty estimation example from Monodepth2 [2] trained on NYU Depth V2 [3]. In (a), the input image is shown, while (b) and (c) display the depth prediction and the true error in terms of RMSE, respectively. The second row demonstrates the uncertainty estimated by post-processing (*Post*) (d), inference dropout (*In-Drop*) (e) and gradients (Ours) (f).

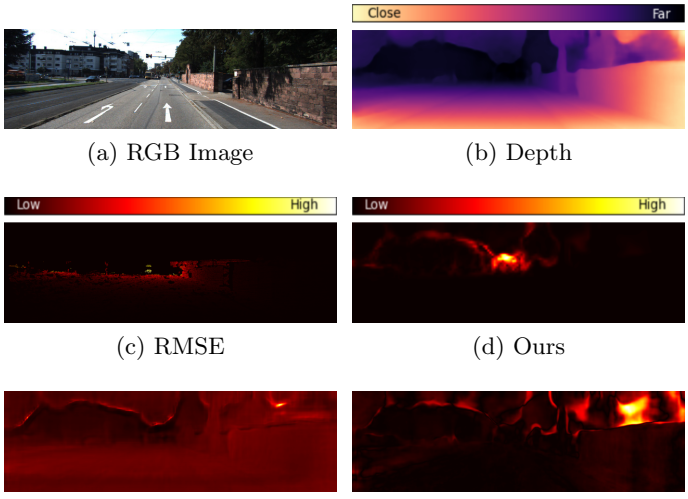
⋆ Most of this work was done while Vasileios Belagiannis was with Ulm University.



(a) RGB Image (b) Depth (c) RMSE

(d) Log (e) In-Drop (f) Ours

Fig. 2. Uncertainty estimation example from Monodepth2 [2] trained on NYU Depth V2 [3] with log-likelihood maximization (*Log*). In (a), the input image is shown, while (b) and (c) display the depth prediction and the true error in terms of RMSE, respectively. The second row demonstrates the uncertainty estimated by log-likelihood maximization (*Log*) (d), inference dropout (*In-Drop*) (e) and gradients (Ours) (f).

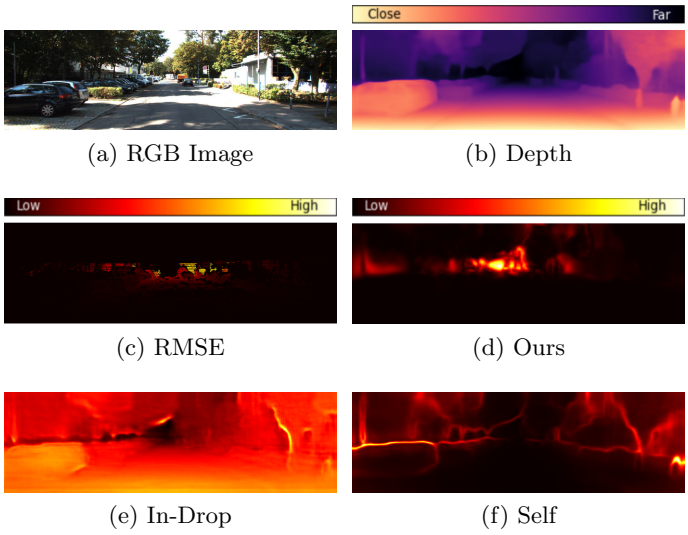


(a) RGB Image (b) Depth

(c) RMSE (d) Ours

(e) In-Drop (f) Post

Fig. 3. Uncertainty estimation example from Monodepth2 [2] trained on KITTI [1] with stereo pair supervision. In (a), the input image is shown, while (b) and (c) display the depth prediction and the true error in terms of RMSE, respectively. Note that ground truth depth is not available for all pixels. In (d), (e) and (f), the uncertainty estimated by gradients (Ours), inference dropout (*In-Drop*) and post-processing (*Post*) is demonstrated, respectively.



(a) RGB Image (b) Depth

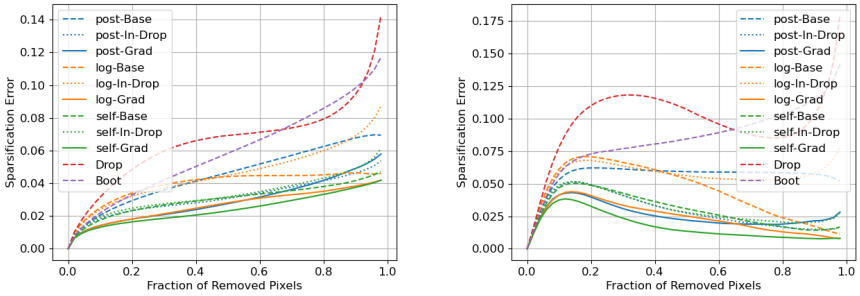
(c) RMSE (d) Ours

(e) In-Drop (f) Self

Fig. 4. Uncertainty estimation example from Monodepth2 [2] trained on KITTI [1] with stereo pair supervision and self-teaching. In (a), the input image is shown, while (b) and (c) display the depth prediction and the true error in terms of RMSE, respectively. Note that ground truth depth is not available for all pixels. In (d), (e) and (f), the uncertainty estimated by gradients (Ours), inference dropout (*In-Drop*) and self-teaching (*Self*) is demonstrated, respectively.

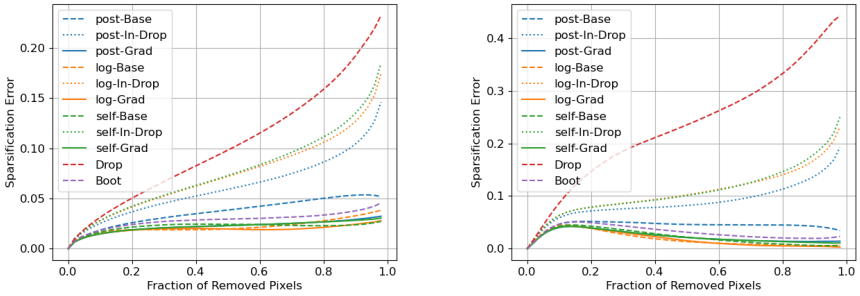
2 Additional Sparsification Error Plots

In Fig. 5 and Fig. 6, the sparsification error plots in terms of Abs Rel and $\delta \geq 1.25$ are displayed for Monodepth2 [2] trained on KITTI [1] with monocular sequences and stereo pairs, respectively. In Fig. 7, the sparsification error curves in terms of Abs Rel, $\delta \geq 1.25$ and RMSE for Monodepth2 [2] trained on NYU Depth V2 [3] are shown. The curves are averaged over the respective test sets.



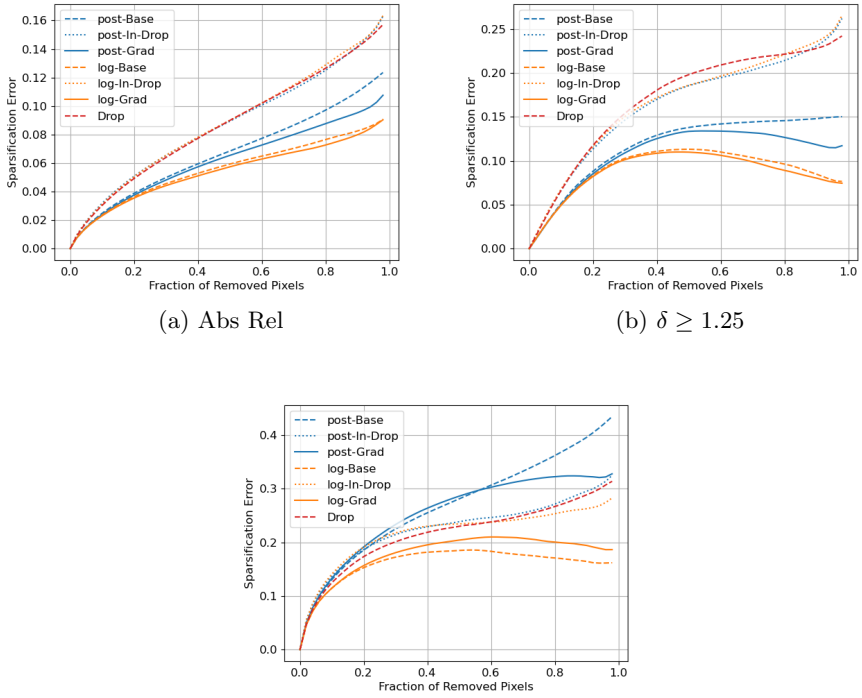
(a) Abs Rel (b) $\delta \geq 1.25$

Fig. 5. The sparsification error in terms of Abs Rel (a) and $\delta \geq 1.25$ (b) over the fraction of removed pixels is illustrated for Monodepth2 [2] trained on KITTI [1] with monocular sequences.



(a) Abs Rel (b) $\delta \geq 1.25$

Fig. 6. The sparsification error in terms of Abs Rel (a) and $\delta \geq 1.25$ (b) over the fraction of removed pixels is illustrated for Monodepth2 [2] trained on KITTI [1] with stereo pairs.



(a) Abs Rel

(b) $\delta \geq 1.25$

(c) RMSE

Fig. 7. The sparsification error in terms of Abs Rel (a), $\delta \geq 1.25$ (b) and RMSE (c) over the fraction of removed pixels is illustrated for Monodepth2 [2] trained on NYU Depth V2 [3].

3 Additional Ablation Studies

Importance Loss Components In Tab. 1, we evaluate the importance of the individual components of the loss function for the Bayesian models. Therefore, we compare the use of the entire loss function $\mathcal{L}_b$, the loss $\mathcal{L}_r$ without the variance and solely the variance σ^2 for the gradient generation. We conduct the experiments on the *Log* models.

Layer Selection In Tab. 2, we demonstrate the performance of our gradient-based uncertainty estimation for the Monodepth2 [2] *Post* model trained on NYU Depth V2 [3] with gradients extracted at different decoder layers. We consider gradients extracted from the 6th to the 9th decoder layer. The layers are counted starting from the first to the last decoder layer.

Variance over Augmentations as Gradient Extraction Loss In Tab. 3, we evaluate the usage of the variance over test-time augmentations as loss func-

Table 1. Uncertainty estimation results for Monodepth2 [2] trained on NYU Depth V2 [3] or KITTI [1] with monocular (Mono) or stereo pair (Stereo) supervision comparing the importance of the single loss terms using the *Log* models. The estimated uncertainty is evaluated with the Area Under the Sparsification Error (AUSE) and the Area Under the Random Gain (AURG) in terms of absolute relative error (Abs Rel), root mean squared error (RMSE) and accuracy $\delta \geq 1.25$.

		Abs Rel		RMSE		$\delta \geq 1.25$	
Setup	Loss	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑
NYU	$\mathcal{L}_b$	0.053	0.032	0.176	0.193	0.086	0.069
	σ^2	0.054	0.031	0.163	0.205	0.085	0.070
	$\mathcal{L}_r$	0.059	0.026	0.236	0.133	0.100	0.055
KITTI Mono	$\mathcal{L}_b$	0.026	0.033	0.819	2.658	0.024	0.059
	σ^2	0.036	0.023	2.399	1.079	0.042	0.041
	$\mathcal{L}_r$	0.028	0.031	0.577	2.900	0.027	0.056
KITTI Stereo	$\mathcal{L}_b$	0.019	0.038	0.490	2.849	0.018	0.061
	σ^2	0.026	0.031	1.685	1.655	0.028	0.051
	$\mathcal{L}_r$	0.022	0.036	0.495	2.845	0.022	0.047

Table 2. Uncertainty estimation results for Monodepth2 [2] *Post* model trained on NYU Depth V2 [3] with gradients extracted from different decoder layers. The estimated uncertainty is evaluated with the Area Under the Sparsification Error (AUSE) and the Area Under the Random Gain (AURG) in terms of absolute relative error (Abs Rel), root mean squared error (RMSE) and accuracy $\delta \geq 1.25$.

	Abs Rel		RMSE		$\delta \geq 1.25$	
Layer	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑
5	0.063	0.023	0.237	0.135	0.111	0.044
6	0.061	0.025	0.252	0.120	0.106	0.048
7	0.063	0.023	0.259	0.113	0.110	0.045
8	0.064	0.022	0.284	0.088	0.114	0.041
9	0.066	0.021	0.298	0.073	0.118	0.037

tion for the gradients extraction in comparison to the proposed loss function, which is defined as the difference of the depth prediction on the input image and its flipped counterpart. We report the results for the gradient extraction on the Monodepth2 [2] *Post* model trained on NYU Depth V2 [3].

Reference Depth Augmentation In Tab. 4, the uncertainty estimation results of our gradient-based approach on Monodepth2 [2] trained with NYU Depth V2 [3] are reported for the *Log* model. We demonstrate different configurations to define the loss for the gradient generation. We consider the squared difference of the prediction to the ground truth depth (GT) and to transformed images by image flipping (Flip), gray-scale conversion (Gray) and additive Gaussian noise (Noise).

Table 3. Uncertainty estimation results for Monodepth2 [2] *Post* model trained on NYU Depth V2 [3] where the gradients are extracted with the variance over different test-time augmentations. The estimated uncertainty is evaluated with the Area Under the Sparsification Error (AUSE) and the Area Under the Random Gain (AURG) in terms of absolute relative error (Abs Rel), root mean squared error (RMSE) and accuracy $\delta \geq 1.25$.

	Abs Rel		RMSE		$\delta \geq 1.25$	
Layer	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑
Var-Grad	0.064	0.023	0.256	0.116	0.113	0.042
Ours	0.061	0.025	0.252	0.120	0.106	0.048

Table 4. Uncertainty estimation results for Monodepth2 [2] *Log* trained on NYU Depth V2 [3] when using different loss functions for the gradient generation. We compare the error of the prediction to the ground truth depth (GT) and depth predictions obtained by different image transformations. We consider image flipping (Flip), gray-scale conversion (Gray) and additive Gaussian noise (Noise). The estimated uncertainty is evaluated with the Area Under the Sparsification Error (AUSE) and the Area Under the Random Gain (AURG) in terms of absolute relative error (Abs Rel), root mean squared error (RMSE) and accuracy $\delta \geq 1.25$.

	Abs Rel		RMSE		$\delta \geq 1.25$	
Loss	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑	AUSE ↓	AURG ↑
GT	0.014	0.071	0.072	0.297	0.018	0.136
Flip	0.053	0.032	0.176	0.193	0.086	0.069
Gray	0.054	0.031	0.177	0.192	0.086	0.068
Noise	0.054	0.031	0.165	0.203	0.085	0.069

References

1. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
2. Godard, C., Aodha, O.M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3827–3837 (2019)
3. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)