On the uncertainty of self-supervised monocular depth estimation

Matteo Poggi, Filippo Aleotti, Fabio Tosi, Stefano Mattoccia

University of Bologna- Department of Computer Science and Engineering

{m.poggi, filippo.aleotti2, fabio.tosi5, stefano.mattoccia}@unibo.it

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

1. Problem Definition

Goal: joint depth prediction and uncertainty estimation from a single input image.

Key Contributions:

- A comprehensive evaluation of uncertainty estimation approaches tailored for **self-supervised** monocular depth estimation.
- An in-depth study of how different self-supervised strategies impacts on both uncertainty and depth.
- A new **Self-Teaching** paradigm to model uncertainty, crucial in the case of unknown camera poses during training.

2. Self-supervised monocular depth estimation

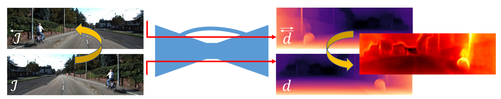
Supervision from hard to source ground truth labels can be replaced by an **image reconstruction loss** $\mathcal{L}_{ss}$ [1]. Given an input image $\mathcal{I}$ and a second frame $\mathcal{I}^{\dagger}$, a warped image $\tilde{\mathcal{I}}$ is obtained as a function π of intrinsics K, $K^{\dagger}$, relative pose $(R|t)$, $\mathcal{I}^{\dagger}$ and estimated depth d.

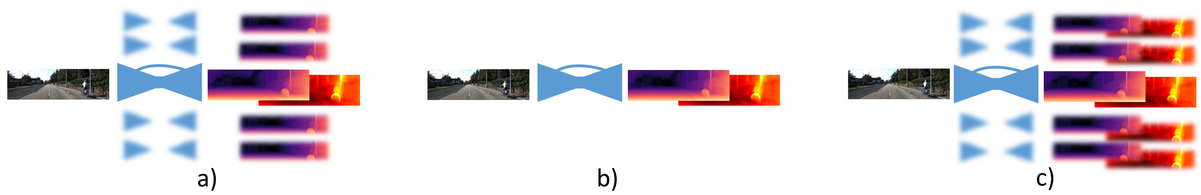
$$\mathcal{L}_{ss} = \mathcal{F}(\tilde{\mathcal{I}}, \mathcal{I}) = \alpha \cdot \frac{1 - \text{SSIM}(\tilde{\mathcal{I}}, \mathcal{I})}{2} + (1-\alpha) \cdot |\tilde{\mathcal{I}} - \mathcal{I}| \quad \text{with} \quad \tilde{\mathcal{I}} = \pi(\mathcal{I}^{\dagger}, K^{\dagger}, R|t, K, d)$$

Supervision can be sourced from K frames $\tilde{\mathcal{I}}$ selecting the minimum for each pixel q as $\mathcal{L}_{ss}(q) = \min_{i \in [0..K]} \mathcal{F}(\tilde{\mathcal{I}}_i(q), \mathcal{I}(q))$. The K frames can be acquired together with $\mathcal{I}$ by means of a single moving camera (M), a stereo rig (S) or both (MS). If $\mathcal{I}, \tilde{\mathcal{I}}$ are taken by a single camera, relative pose $(R|t)$ is unknown and thus estimated together with d.

3. Uncertainty estimation for self-supervised monocular estimation

Uncertainty from flipping. By estimating two depth maps d and $\overleftarrow{d}$ for image $\mathcal{I}$ and its horizontally flipped counterpart $\overleftarrow{\mathcal{I}}$, a naive uncertainty can be obtained as $|d - \overleftrightarrow{d}|$ by flipping back $\overleftarrow{d}$, similarly to the post-processing proposed in [2] (**Post**)





Uncertainty estimation strategies can be classified as follows [3]:

a) Empirical estimation. These methods sample a subset of all possible networks and estimate uncertainty as the **variance** $\sigma^2(d)$ over their predictions. Examples: MC Dropout (**Drop**), Bootstrapped Ensemble (**Boot**), Snapshots Ensemble (**Snap**).

b) Predictive estimation. This category aims at encoding uncertainty as a function of network parameters and the input image by means of a **predictive** model. Examples: Learned Reprojection (**Reproj**), Log-likelihood maximization (**Log**). This latter, traditionally used when ground truth is available, can be adapted to the self-supervised case as follows [4]

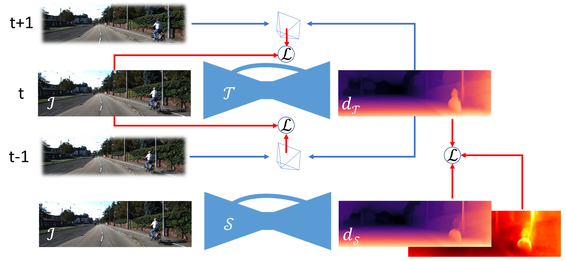
$$\mathcal{L}_{Log} = \frac{\min_{i \in [0..K]} \mathcal{F}(\tilde{\mathcal{I}}_i(q), \mathcal{I}(q))}{\sigma(d)} + \log \sigma(d)$$

c) Bayesian estimation. It models uncertainty by marginalizing over all possible network weights rather than choosing a point estimate. An approximation [5] is obtained combining **empirical** and **predictive** techniques (for instance, **Boot+Log**).

4. Self-teaching

In case of unknown pose (M, MS), predictive methods such as Log jointly model uncertainty for both estimated depth and pose. Indeed, $\mathcal{L}_{\text{Log}}$ is computed over $\mathcal{F}(\tilde{\mathcal{I}}, \mathcal{I})$, for which $\tilde{\mathcal{I}}$ is obtained as a function π of unknown variables d and $(R|t)$. Thus, we propose a **Self-Teaching** scheme (**Self**), training a student instance $\mathcal{S}$ to mimic the distribution sourced from a teacher model $\mathcal{T}$ trained with self-supervision

$$\mathcal{L}_{Self} = \frac{|\mu(d_{\mathcal{S}}) - d_{\mathcal{T}}|}{\sigma(d_{\mathcal{S}})} + \log \sigma(d_{\mathcal{S}})$$



5. Experiments & Results

- We train 10 new variants of MonoDepth2 [1] implementing different uncertainty modelling strategies.
- We train on 192×640 resized images, with batch size 12 for 20 epochs with M, S and MS supervision and evaluate on the Eigen test split with improved ground truth, defined in [6].
- For depth, we measure Abs Rel, RMSE (Lower is better) and $\delta < 1.25$ (Higher is better). For uncertainty, we measure the Area Under Sparsification Error (AUSE) and Area Under Random Gain (AURG).
- **Self-teaching** consistently improves depth estimation. When pose is unknown (M,MS) it outperform Log.

Depth evaluation:

Method	#Trn	#Par	#Fwd	Abs Rel	RMSE	$\delta < 1.25$	Abs Rel	RMSE	$\delta < 1.25$	Abs Rel	RMSE	$\delta < 1.25$
				(M)			(S)			(MS)		
Monodepth2 [1]	1×	1×	1×	0.090	3.942	0.914	0.085	3.942	0.912	0.084	3.739	0.918
Monodepth2-*Post* [1]	1×	1×	2×	0.088	3.841	0.917	0.084	3.777	**0.915**	**0.082**	**3.666**	**0.919**
Monodepth2-*Drop*	1×	1×	N×	0.101	4.146	0.892	0.129	4.908	0.819	0.172	5.885	0.679
Monodepth2-*Boot*	N×	N×	1×	0.092	3.821	0.911	0.085	**3.772**	0.914	0.086	3.787	0.910
Monodepth2-*Snap*	1×	N×	1×	0.091	3.921	0.912	0.085	3.849	0.912	0.085	3.806	0.914
Monodepth2-*Repr*	1×	1×	1×	0.092	3.936	0.912	0.085	3.873	0.913	0.084	3.828	0.913
Monodepth2-*Log*	1×	1×	1×	0.091	4.052	0.910	0.085	3.860	**0.915**	0.083	3.790	0.916
Monodepth2-*Self*	(1+1)×	1×	1×	**0.087**	3.826	**0.920**	0.084	3.835	**0.915**	0.083	3.682	**0.919**
Monodepth2-*Boot+Log*	N×	N×	1×	0.092	3.850	0.910	0.085	3.777	0.913	0.086	3.771	0.911
Monodepth2-*Boot+Self*	(1+N)×	N×	1×	0.088	**3.799**	0.918	0.085	3.793	0.914	0.085	3.704	0.915
Monodepth2-*Snap+Log*	1×	1×	1×	0.092	3.961	0.911	**0.083**	3.833	0.914	0.084	3.828	0.914
Monodepth2-*Snap+Self*	(1+1)×	1×	1×	0.088	3.832	0.919	0.086	3.859	0.912	0.085	3.715	0.916

Uncertainty evaluation:

Method	Abs Rel		RMSE		$\delta \geq 1.25$		Abs Rel		RMSE		$\delta \geq 1.25$		Abs Rel		RMSE		$\delta \geq 1.25$	
	AUSE	AURG	AUSE	AURG	AUSE	AURG	AUSE	AURG	AUSE	AURG	AUSE	AURG	AUSE	AURG	AUSE	AURG	AUSE	AURG
	(M)						(S)						(MS)					
Monodepth2-*Post*	0.044	0.012	2.864	0.412	0.056	0.022	0.036	0.020	2.523	0.736	0.044	0.034	0.036	0.018	2.498	0.655	0.044	0.031
Monodepth2-*Drop*	0.065	0.000	2.568	0.944	0.097	0.002	0.103	-0.029	6.163	-2.169	0.231	-0.080	0.103	-0.027	7.114	-2.580	0.303	-0.081
Monodepth2-*Boot*	0.058	0.001	3.982	-0.743	0.084	-0.001	0.028	0.029	2.291	0.964	0.031	0.048	0.028	0.030	2.269	0.985	0.034	0.049
Monodepth2-*Snap*	0.059	-0.001	3.979	-0.639	0.083	-0.002	0.028	0.029	2.252	1.077	0.030	0.051	0.029	0.028	2.245	1.029	0.033	0.047
Monodepth2-*Repr*	0.051	0.008	2.972	0.381	0.069	0.013	0.040	0.017	2.275	1.074	0.050	0.030	0.046	0.010	2.662	0.635	0.062	0.018
Monodepth2-*Log*	0.039	0.020	2.562	0.916	0.044	0.038	0.022	0.036	0.938	2.402	**0.018**	0.061	0.028	0.029	1.714	**1.562**	0.028	0.050
Monodepth2-*Self*	0.030	0.026	2.009	1.266	0.030	0.045	0.022	0.035	1.679	1.642	0.022	0.056	**0.022**	0.033	**1.654**	1.515	**0.023**	0.052
Monodepth2-*Boot+Log*	0.038	0.021	2.449	0.820	0.046	0.037	**0.020**	**0.038**	**0.807**	**2.455**	**0.018**	**0.063**	0.030	0.028	1.962	1.282	0.032	0.051
Monodepth2-*Boot+Self*	**0.029**	**0.028**	**1.924**	**1.316**	**0.028**	**0.049**	0.023	0.035	1.646	1.628	0.021	0.058	0.023	0.033	1.688	1.494	**0.023**	**0.056**
Monodepth2-*Snap+Log*	0.038	0.022	2.385	1.001	0.043	0.039	0.021	0.037	0.891	2.426	**0.018**	0.061	0.030	0.027	2.032	1.272	0.032	0.048
Monodepth2-*Snap+Self*	0.031	0.026	2.043	1.230	0.030	0.045	0.023	0.035	1.710	1.623	0.023	0.058	0.023	**0.034**	1.684	1.510	**0.023**	0.055

References

[1] C. Godard et al., "Digging into self-supervised monocular depth estimation" (ICCV 2019)

[2] C. Godard et al., "Unsupervised Monocular Depth Estimation with Left-Right Consistency" (CVPR 2017)

[3] E. Ilg et al., "Uncertainty Estimates and Multi-Hypotheses Networks for Optical Flow" (ECCV 2018)

[4] M. Klodt and A. Vedaldi, "Supervising the new with the old: learning SFM from SFM" (ECCV 2018)

[5] R. M. Neal, "Bayesian learning for neural networks" (Springer Science & Business Media)

[6] F. Aleotti et al., "Generative adversarial networks for unsupervised monocular depth prediction" (ECCVW 2018)

Links

Paper

GitHub Code

Youtube Video

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.