Self-distillation and Uncertainty Boosting Self-supervised Monocular Depth Estimation

Hang Zhou, Sarah Taylor, David Greenwood, Michal Mackiewicz
{hang.zhou, s.l.taylor, david.greenwood, m.mackiewicz}@uea.ac.uk
University of East Anglia, Norwich, UK

Code and model available at https://github.com/brandleyzhou/SUB-Depth

INTRODUCTION

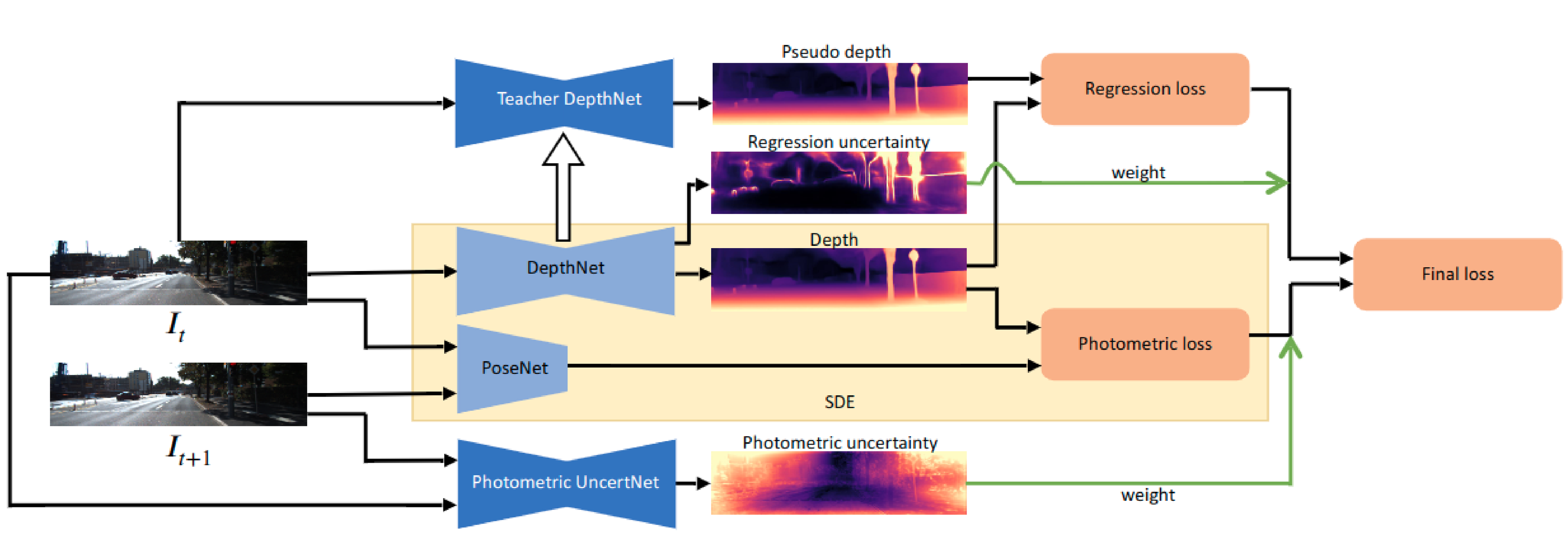
Goal:

Develop a two-stage training scheme for self-supervised monocular depth estimation approaches.

Contributions:

- Introducing an auxiliary teacher-student objective for SDE training
- Utilizing heteroscedastic uncertainty modelling to select optimal settings.
- Conducting extensive experiments to show the generalization ability to existing SOTA models.

4. Overview of SUB-Depth training



METHODS

1. Self-supervised monocular depth estimation (SDE)

Avoiding acquisition of depth ground truth, SDE trains a depth network and a pose network simultaneously for an image reconstruction object. Given an intrinsic matrix K, it uses estimated depth d and camera pose T change to warp a source frame I_s to a target frame I_t.

Weights are optimized by the colour differences between warped $I_{s'}$, and I_t via photometric loss $\boldsymbol{L_P}$ and an edge-aware smoothness penalty term $\boldsymbol{L_S}$:

$$\boldsymbol{L_P} = \alpha \frac{1 - SSIM(I_{s'}, I_t)}{2} + \alpha |I_t - I_{s'}|$$

$$\boldsymbol{L_S} = \left|\frac{\nabla d}{\partial x}\right| e^{-\left|\frac{\nabla I_0}{\partial x}\right|} + \left|\frac{\nabla d}{\partial y}\right| e^{-\left|\frac{\nabla I_0}{\partial y}\right|}$$

The final loss for this image reconstruction task:

$$\boldsymbol{l_{photometric}} = \boldsymbol{L_P} + \beta\, \boldsymbol{L_S}$$

2. Self-distillation scheme:

We introduce a teacher depth model $\boldsymbol{T}$ and let d from a student depth network to regress $\boldsymbol{d_{pseudo}} = \boldsymbol{T(I_t)}$ using an L$\mathbf{1}$ loss:

$$\boldsymbol{l_{regression}} = |d - d_{pseudo}|$$

Then, we firstly combine $\boldsymbol{l_{regression}}$ with $\boldsymbol{l_{photometric}}$ using several manually-tuned settings:

$$\mathrm{l} = w_{pho} * \boldsymbol{l_{photometric}} + w_{reg} * \boldsymbol{l_{regression}}$$

And we find that it is hard to select the optimal weight setting, based on the table below.

Objective weights		Error metrics				Accuracy metrics		
ω_{pho}	ω_{reg}	Rel Abs	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
0	1	0.112	0.884	4.740	0.189	0.881	0.961	0.982
0.2	0.8	**0.110**	0.855	4.724	0.188	0.881	0.961	0.982
0.4	0.6	0.112	0.866	4.736	0.189	0.881	0.961	0.982
0.5	0.5	0.112	0.888	4.766	0.189	0.882	0.961	0.981
0.6	0.4	0.113	0.876	4.774	0.189	**0.884**	0.962	0.983
0.8	0.2	0.113	0.885	4.799	0.190	0.882	0.961	0.981
1	0	0.115	0.903	4.863	0.193	0.877	0.959	0.981

3. Task-dependent uncertainty formulation:

Following [1], we reformulate $\boldsymbol{l_{photometric}}$ and $\boldsymbol{l_{regression}}$ to $\boldsymbol{l_{reconstruction}}$ and $\boldsymbol{l_{distillation}}$ with their corresponding uncertainty:

$$\boldsymbol{l_{reconstruction}} = \frac{\boldsymbol{l_{photometric}}}{\sigma_{pho}} + \log(\sigma_{pho})$$

$$\boldsymbol{l_{distillation}} = \frac{\boldsymbol{l_{regrssion}}}{\sigma_{reg}} + \log(\sigma_{reg})$$

As a result, we use a combination of two losses above:

$$\boldsymbol{l_{final}} = \boldsymbol{l_{distillation}} + \boldsymbol{l_{reconstruction}}$$

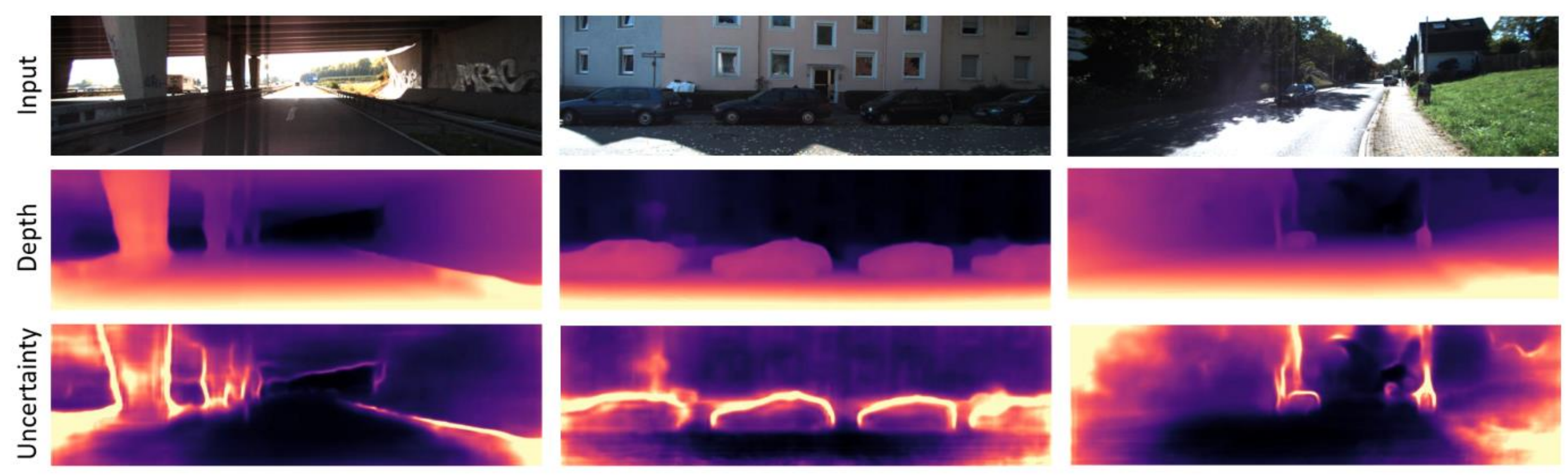
RESULTS

We validate SUB-Depth on three different SDE approaches: Monodepth2 [2], HR-depth [3] and DIFFNet [4] with KITTI benchmark.

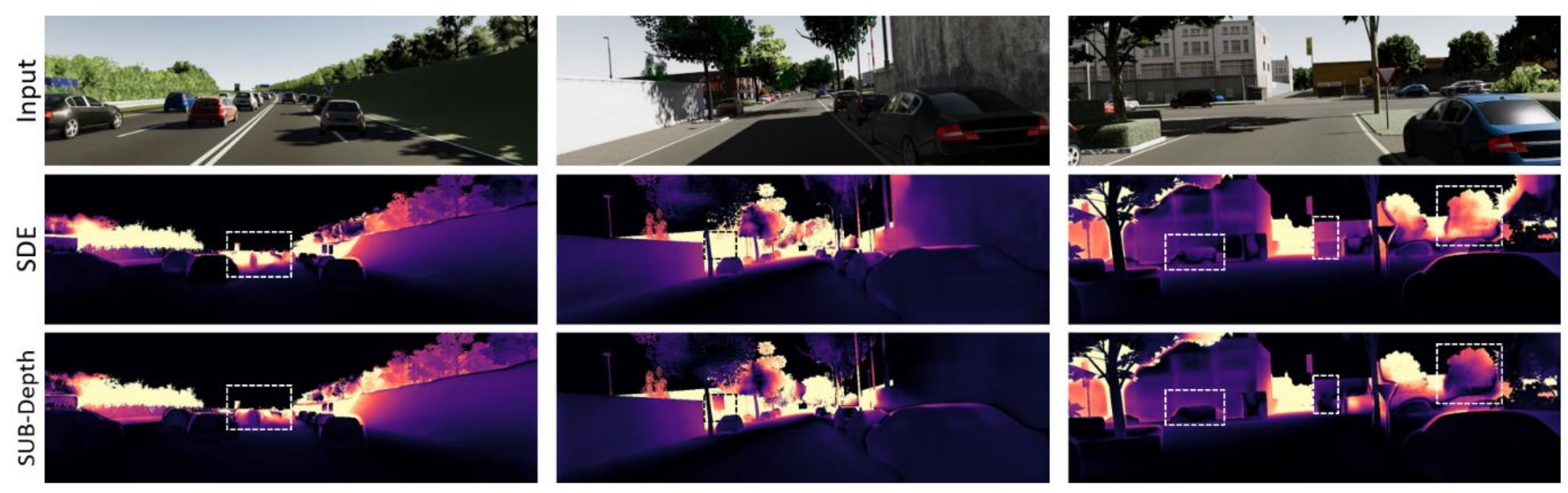
Quantitative comparison on KITTI Eigen split

Method	Abs Rel	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
Monodepth2 [14]	0.115	0.903	4.863	0.193	0.877	0.959	0.981
+ SUB-Depth	**0.110**	**0.821**	**4.648**	**0.185**	**0.884**	**0.962**	**0.983**
Improvement	0.005	0.082	0.115	0.008	0.007	0.003	0.002
HR-depth [34]	0.109	0.792	4.632	0.185	0.884	0.962	0.983
+ SUB-Depth	**0.106**	**0.770**	**4.545**	**0.182**	**0.888**	**0.963**	**0.983**
Improvement	0.003	0.022	0.087	0.003	0.004	0.001	0
DIFFNet [49]	0.102	0.764	4.483	0.180	0.896	0.965	0.983
+ SUB-Depth	**0.099**	**0.695**	**4.326**	**0.175**	**0.900**	**0.966**	**0.984**
Improvement	0.003	0.059	0.157	0.005	0.004	0.001	0.001

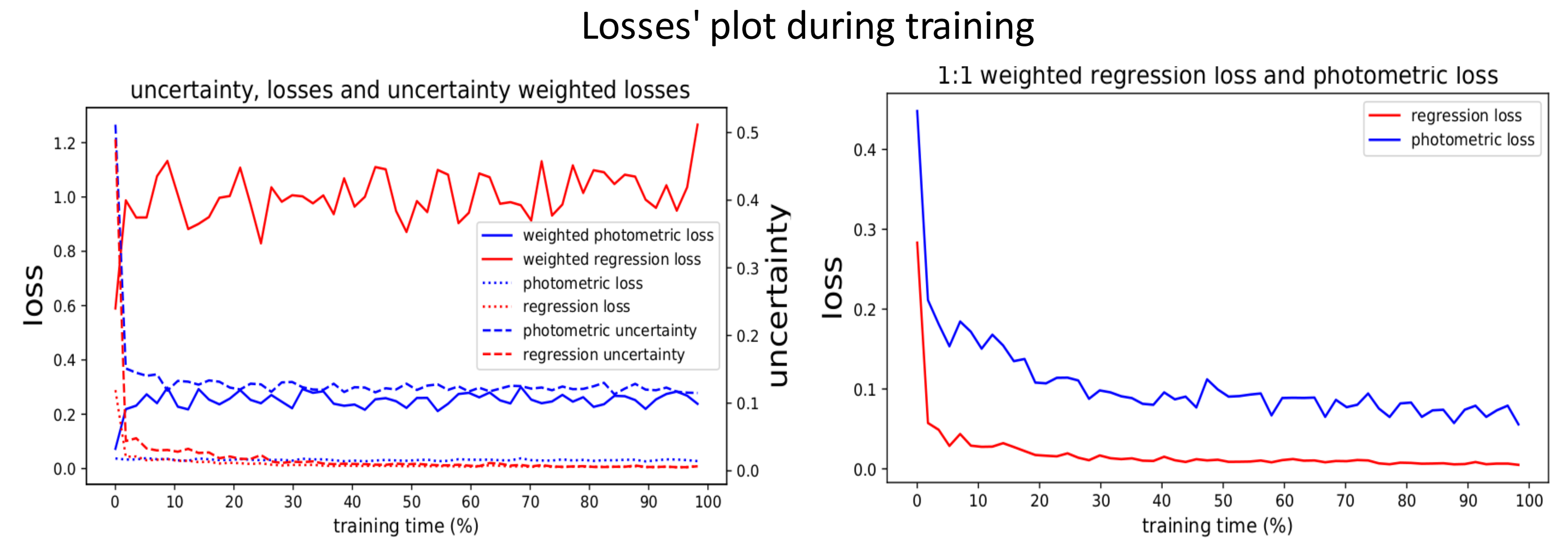
Output Visualizations



Error Visualizations



Losses' plot during training



References

[1] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. CVPR, 2018.

[2] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. ICCV, 2019.

[3]Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. AAAI, 2021.

[4] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. BMVC, 2021.