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

Hang Zhou
hang.zhou@uea.ac.uk
David Greenwood
david.greenwood@uea.ac.uk
Sarah Taylor
s.l.taylor@uea.ac.uk
Michal Mackiewicz
M.Mackiewicz@uea.ac.uk

School of Computing Sciences
University of East Anglia
Norwich, UK

Abstract

For self-supervised monocular depth estimation (SDE), recent works have introduced additional learning objectives, for example semantic segmentation, into the training pipeline and have demonstrated improved performance. However, such multi-task learning frameworks require extra ground truth labels, neutralising the biggest advantage of self-supervision. In this paper, we propose **SUB-Depth** to overcome these limitations. Our main contribution is that we design an auxiliary self-distillation scheme and incorporate it into the standard SDE framework, to take advantage of multi-task learning without labelling cost. Then, instead of using a simple weighted sum of the multiple objectives, we employ generative task-dependent uncertainty to weight each task in our proposed training framework. We present extensive evaluations on KITTI to demonstrate the improvements achieved by training a range of existing networks using the proposed framework, and we achieve state-of-the-art performance on this task.

1 Introduction

Depth perception plays an important role in real-world applications including autonomous driving, augmented reality, 3D reconstruction and other high-level computer vision tasks. Although physical sensors such as LiDAR have been deployed widely, estimating depth from pixels is appealing due to the lower cost and compatibility where a camera is available. Supervised depth estimation approaches [1, 5, 9, 11, 31, 35, 38] can predict dense depth maps but require costly ground truth depth labels. In contrast, self-supervised approaches require no labelled data [13, 14, 23, 30, 34, 40, 41, 48, 49, 50, 51] and are performing competitively. At a high level, self-supervised depth estimation uses a depth network's output as an intermediate representation for a stereo matching problem or an image reconstruction task. For the latter, the models are trained within a standard self-supervised monocular depth estimation (SDE) framework [14, 34, 44, 49, 50, 51]. To build such a system, a pose network is introduced to predict the camera pose change between two consecutive frames. Hence, in a standard SDE training framework, a depth network and a pose network are trained simultaneously for **an image reconstruction** task by optimising the photometric loss. Previous works [6, 32, 33, 41] have shown that training a single depth model benefits from multiple regression or classification objectives. Inspired by prior works that train a student depth network with a trained teacher network [34, 37], we extend the single-task SDE framework to a multi-task setting by introducing a self-distillation scheme associated with a regression objective. Compared with other multi-task settings which introduce supervised tasks such as semantic segmentation, one of the advantages of self-distillation is that the framework remains a self-supervised regime.

Performance of multi-task systems is dependent on the relative loss weighting for each task. Instead of manually tuning weights of loss terms, inspired by Kendall [24], we propose two uncertainty modelling strategies to calculate uncertainty for the self-distillation task and the image reconstruction task respectively. Specifically, the self-distillation uncertainty down-weights the regression loss when a teacher network outputs noisy depth values, and the photometric uncertainty outputs higher confidence where input frames satisfy the image reconstruction tasks' assumptions, that is, static world and ego-motion. We call our system **SUB-Depth**, and summarise its following key contributions:

- We propose a novel multi-task training framework for self-supervised monocular depth estimation.
- Instead of manually tuning loss terms' weights, we utilize the task-dependent uncertainty idea, and experiment with several ways of uncertainty modeling.
- We conduct exhaustive experiments to show that the proposed training framework is able to boost existing models' performance significantly.

2 Related literature

In this section, we review works relating to self-supervised monocular depth estimation, multi-task learning and predictive uncertainty modelling.

2.1 Self-supervised monocular depth estimation

Different from supervised learning based approaches that are trained for a regression task with ground truth depth, self-supervised monocular depth estimation (SDE) methods are trained for an image reconstruction task with a photometric loss. Inspired by Structure from Motion (SfM), the seminal work of Zhou [51] proposed a fundamental framework consisting of a depth network and a pose network which are trained simultaneously with sequential video frames. Many works have further advanced this framework in several different ways. Monodepth2 [14] introduced minimum-reprojection, structural similarity and multi-scale reconstruction strategies, and is now the most widely used baseline and builds the standard SDE framework. More recently, a series of works proposed improved network architectures within this framework [16, 34, 50]. An image distortion handling model was proposed for UnRectDepthNet [28], and to guide depth feature learning, Guizilini [17] exploited a semantic segmentation network. Instead of fixed camera parameters, learnable camera parameters have been used [3, 15]. By introducing self-attention and a discrete disparity volume, Johnston and Carneiro further improved Monodepth2 [21]. To boost the single-frame SDE

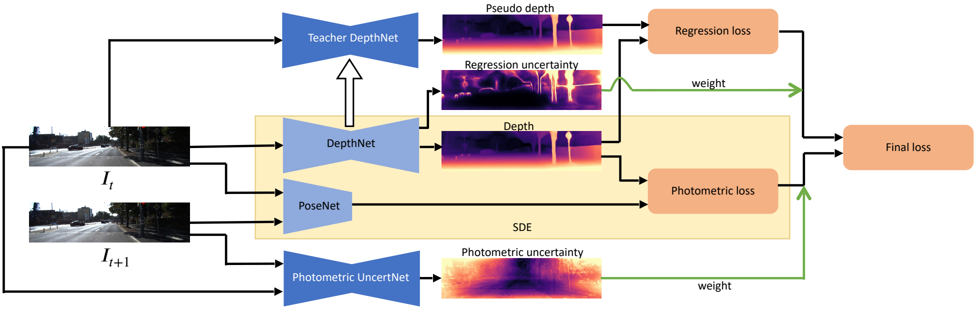

Figure 1: An overview of the SUB-Depth framework. SUB-Depth extends the standard existing self-supervised monocular depth estimation framework (SDE) (highlighted) using self-distillation and uncertainty modelling. The teacher DepthNet outputs a supervisory signal for training the DepthNet, and enables computation of a regression loss. Both regression and photometric uncertainty maps are learned and used to weight the respective losses. The teacher DepthNet is pretrained with the highlighted SDE framework by optimising the photometric loss.

frameworks, Senoh [40] proposed a multi-frame training and testing framework. Poggi [37] introduced uncertainty modelling into SDE approaches, and showed how different strategies impact depth and uncertainty estimation.

To our best knowledge, we first time propose and incorporate self-distillation into SDE to build a multi-objective learning framework. We demonstrate the performance improvements of our proposed training framework using three existing architectures as underlying models, Monodepth2 [14], HR-depth [34] and DIFFNet [50], which represent base-level, middle-level and high-level performance methods respectively.

2.2 Multi-task learning and optimisation

Visual scene understanding models, trained with multi-task learning systems, always outperform counterparts trained individually [24]. Recognising similarities with the semantic segmentation task, a series of supervised depth estimation methods [18, 24, 45] and self-supervised methods [2, 4, 26, 29] introduced additional segmentation networks. These methods boosted performance by using a shared representation learning encoder for both tasks or by distilling knowledge across tasks. However, building such a multi-task system requires extra semantic annotations, thus losing the most important advantage of self-supervised learning. Differing from prior works, we exploit knowledge distillation [19] as an auxiliary task, in which a student network can learn from a teacher network. In our proposed scheme, as the teacher and student have the same network architecture, we name the additional task *self-distillation*. The teacher network is trained with the standard SDE framework, so this scheme can be implemented without extra labeling cost.

A significant challenge for multi-task learning is how to simultaneously optimise multiple objectives. Sener at el. [39] developed a Frank-Wolfe optimiser to find a Pareto optimal solutions. To minimise the negative conflicts with other gradients during back-propagation, Yu at el. [47] proposed a form of gradient surgery that projects each task's gradient onto the normal plane of the gradient of any other task and modifies the gradients for each task. In SUB-Depth, we utilise the uncertainty-based weighting approach proposed by Kendall at el. [24] for jointly learning, which down-weights the task loss contribution where the task-dependent (homoscedastic) uncertainty is high.

3 SUB-Depth training framework

In this section, we first introduce the standard SDE framework, then the proposed self-distillation, and two task-dependent homoscedastic uncertainty formulations. The final system overview is shown in Figure 1.

3.1 Self-supervised monocular depth estimation

An SDE framework (highlighted by the yellow box in Figure 1) trains a DepthNet Θ_{depth} and a PoseNet Θ_{pose} simultaneously for an image reconstruction task with a triplet of sequential RGB frames $I_t \in \mathbb{R}^{H\times W\times 3}, t \in \{-1,0,1\}$. At training time Θ_{depth} takes a target frame I_0 as input and predicts a depth map $d = \Theta_{\text{depth}}(I_0)$, while the PoseNet estimates a relative pose change between the target frame and a source frame, $T_{0\to t^{'}} = \Theta_{\text{pose}}(I_0, I_{t^{'}}), t^{'} \in \{-1,1\}$.

Based on the assumption that the world is static and the view change is only caused by a moving camera, a reconstructed counterpart to target frame I_0 can be generated using only pixels from the source frames $I_{t^{'}}, t^{'} \in \{-1,1\}$:

$$I_{t^{'}\to 0} = I_{t^{'}}[\text{proj}(\text{reproj}(I_0, d, T_{0\to t^{'}}), K)] \tag{1}$$

where K are known camera intrinsics, [ ] is the sampling operator, $reproj$ returns a 3D point cloud of camera $t^{'}$, and $proj$ outputs the 2D coordinates when projecting the point cloud onto $I_{t^{'}}$.

Using the predicted depth map d, the generated view $I_{t^{'}\to 0}$ and the corresponding target frame I_0, we build a supervisory signal consisting of two items:

Photometric Loss, ℓ_p, is an appearance matching loss which calculates the difference between I_0 and $I_{t^{'}\to 0}$. Following [13, 14], the similarity between a synthesised frame and a target frame is computed using a Structural Similarity term (SSIM) [43]. Then, combining with the L1 norm, the final photometric loss function is defined:

$$\ell_p(I_0, I_{t^{'}\to 0}) = \alpha \frac{1 - \text{SSIM}(I_0, I_{t^{'}\to 0})}{2} + (1-\alpha)|I_0 - I_{t^{'}\to 0}| \tag{2}$$

Edge-aware Smoothness [13], ℓ_s, regularises the depth in low gradient regions:

$$\ell_s(d) = |\frac{\nabla d}{\partial x}| e^{-|\frac{\nabla I_0}{\partial x}|} + |\frac{\nabla d}{\partial y}| e^{-|\frac{\nabla I_0}{\partial y}|} \tag{3}$$

We also employ the minimum photometric error, auto-masking and multi-scale depth loss techniques which were introduced in [14]. The final self-supervised photometric objective is defined:

$$\ell_{photometric} = \min(\ell_p(I_0, I_{t^{'}\to 0})) + \beta \ell_s(d), t^{'} \in \{-1,1\} \tag{4}$$

where β is a weighting coefficient between the photometric loss ℓ_p and depth smoothness ℓ_s. The objective loss is averaged per pixel, pyramid scale and image batch.

Table 1: Comparison between manually tuned objective weights, evaluated on the KITTI [12] Eigen split. We experiment with several combinations of ω_{pho} and ω_{reg}. The best weighting pairs are in red. The best (Rel Abs and δ_1) scores are **bold and underlined**. Error and accuracy metrics' definitions are given in 5.1.

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	**<u>0.110</u>**	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	**<u>0.884</u>**	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

3.2 Self-distillation scheme

Most related works focus on integrating other supervised learning based tasks into an SDE framework. Typically, when introducing a segmentation task, the segmentation network shares the encoder in the SDE depth network, and all components are trained jointly with the sum of the photometric loss and the segmentation loss. Although depth models trained with such a multi-task system can improve their performance, it neutralises the advantage of SDE frameworks.

Unlike existing multi-task strategies, self-distillation avoids introducing extra manual annotations. Instead, we use an SDE trained teacher depth network Θ_{teacher} to output pseudo depth ground truth $d_{pseudo} = \Theta_{\text{teacher}}(I_0)$. We then let the depth map from the DepthNet $d = \Theta_{\text{depth}}(I_0)$ regress the d_{pseudo}. The objective can be formulated as an L1 **regression loss**:

$$\ell_{regression} = |\Theta_{\text{depth}}(I_0) - \Theta_{\text{teacher}}(I_0)| \tag{5}$$

As Θ_{teacher} and Θ_{depth} have the same network architecture, we name this task **self-distillation**.

By simply introducing a Θ_{teacher}, we retrain depth networks using following weighted loss function:

$$\ell = \omega_{pho} \times \ell_{photometric} + \omega_{reg} \times \ell_{regression} \tag{6}$$

Where ω_{pho} and ω_{reg} are weights for $\ell_{photometric}$ and $\ell_{regression}$ respectively. We train and evaluate models using different weighting settings, shown in Table 1. From the table, we observe that this naive multi-task learning framework can output a Θ_{depth} which outperforms the Θ_{teacher} trained with standard SDE framework, no matter what the ratio is. However, when we set $\omega_{pho} = 0.2$ and $\omega_{reg} = 0.8$, models gain best performance for Rel Abs, while they are improved significantly for δ_1 when $\omega_{pho} = 0.6$ and $\omega_{reg} = 0.4$. As it is hard to get an optimal weight settings, we utilize uncertainty based methods to automatically balance loss terms.

3.3 Task-dependent uncertainty formulation

Following [22], given a dataset (x, y), we let the network output the mean $\hat{y}$ and the variance σ of a posterior probability distribution $p(y|\hat{y}, \sigma)$ over ground truth y, which can be modelled as Laplacian or Gaussian. If Laplace's distribution:

$$p(y|\hat{y},\sigma) = \frac{1}{2\sigma}\exp\frac{-|\hat{y}-y|}{\sigma} \tag{7}$$

is used, then the network can be trained by minimising the loss [27]:

$$loss = \frac{|\hat{y}-y|}{\sigma} + \log(\sigma) \tag{8}$$

where the variance σ increases when the ground truth y is unreliable. As a result, we can treat σ as task-dependent uncertainty, and the penalty term $\log(\sigma)$, avoids the degenerate solution $\sigma = +\infty$.

We introduce uncertainty modelling for each sub-task in the framework:

Uncertainty for image reconstruction. Intuitively, as photometric loss is a measurement of the difference between two images, it is natural to estimate its uncertainty with a model that takes two images as input. While prior works [37, 46] use the DepthNet Θ_{depth} for modelling the photometric uncertainty, we propose a separate Photometric UncertNet Θ_{pho} to estimate the uncertainty. As for the input of the proposed uncertainty network, we experiment with different settings: 1). feeding the target frame I_t, 2). feeding the target I_t and aligned $I_{t'}$ (see in supplementary material for more details). Finally, we let UncertNet take the target frame and the source frame as inputs and output the photometric uncertainty map σ_{pho}, as shown in Figure 1. Then the uncertainty weighted photometric loss, with the penalty term $\log(\sigma_{pho})$, for the image reconstruction task is given by:

$$\ell_{reconstruction} = \frac{\ell_{photometric}}{\sigma_{pho}} + \log(\sigma_{pho}) \tag{9}$$

Uncertainty for self-distillation. We let the DepthNet Θ_{depth} encode and output depth regression uncertainty σ_{reg}. Besides, we explore using a standalone regression uncertainty network to estimate depth uncertainty (see in supplementary material for more details). Then the uncertainty weighted regression loss with the penalty term $\log(\sigma_{reg})$ for the self-distillation task can be computed as:

$$\ell_{distillation} = \frac{\ell_{regression}}{\sigma_{reg}} + \log(\sigma_{reg}) \tag{10}$$

3.4 Multi-task learning with uncertainty

Finally we combine the uncertainty weighted photometric loss ($\ell_{reconstruction}$) and regression loss ($\ell_{distillation}$) to build **SUB-Depth**:

$$\ell_{final} = \ell_{reconstruction} + \ell_{distillation} \tag{11}$$

The result is a multi-task learning system, which trains Θ_{depth} for an image reconstruction task and a self-distillation task using the sum of task-dependent uncertainty weighted losses.

The difference during training between the naive unweighted sum of losses and uncertainty weighted losses is shown in Figure 2. On the left plot, Θ_{depth} is trained with self-distillation as a prime task. In this graph we observe that, although the unweighted regression loss is lower than the unweighted photometric loss throughout most of the training, after applying the task-dependent uncertainty weighting the self-distillation task contributes more to

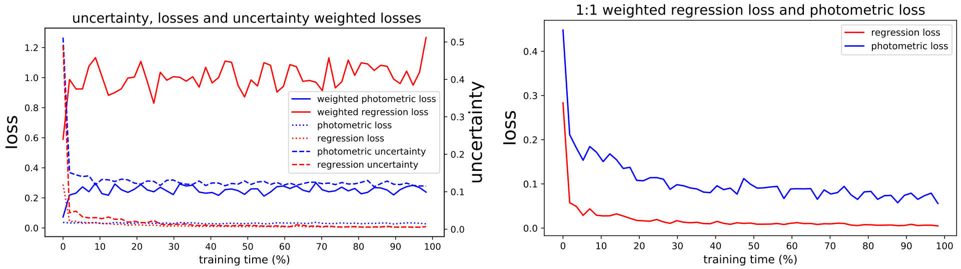

Figure 2: Left: The task-dependent losses, uncertainty weighted losses and uncertainty estimates during SUB-Depth training. Right: The corresponding task-dependent losses of the same system trained with no uncertainty modelling. Uncertainty modelling increases the contribution of the regression loss, and down-weights photometric loss.

the ℓ_{final} than the reconstruction loss. This change is due to the regression uncertainty σ_{reg} being lower than the photometric uncertainty σ_{pho}, and indicates that pseudo-labels from the teacher DepthNet provide a more reliable supervisory signal than the pixel-level metrics used in the photometric loss. For comparison, the right plot in Figure 2 shows the naive 1:1 weighted multi-task training framework without uncertainty modelling. In this case, the photometric task dominates the loss throughout training following similar curves to the respective unweighted losses on the left plot.

4 Implementation

Our models are trained and tested on a single NVIDIA RTX 6000 GPU using PyTorch [36]. A depth network and a pose network are trained for 20 epochs using the Adam optimiser [25] with the default betas 0.9 and 0.999. They were trained with a batch size of 8 and an input and output resolution of 640 × 192. We set the initial learning rate as 10^{-4} for the first 14 epochs and then 10^{-5} for fine-tuning the remainder. In the objective function ℓ_{final} (Equation 11), we set the SSIM weight $\alpha = 0.85$ and the edge-aware smoothness weight $\beta = 1 \times 10^{-3}$.

DepthNet and Teacher DepthNet. To verify the generalisation capability of SUB-Depth, we train three different architectures: Monodepth2 [14], HR-depth [34] and DIFFNet [50], which represent baseline-level, mid-level and state-of-the-art methods when trained with the standard SDE framework. DepthNet models are initialised on the Imagenet [8] pretrained weights. The teacher DepthNets are fixed models that are pretrained with the SDE framework. To generate the associated regression uncertainty, we modify output layers which originally produce one-channel depth maps to two-channel output.

PoseNet and Photometric UncertNet. For all training settings, we implement the architecture proposed in [14] for pose estimation, which is built on ResNet-18. The pose network takes the two adjacent frames as input and outputs the relative pose which is parameterised with a 6-DOF vector. The photometric uncertainty network uses an encoder-decoder with skip-connections. The encoder is based on the ResNet-18 architecture and the decoder follows the design of the Monodepth2 depthnet decoder [14]. The photometric uncertainty network takes adjacent frames as input and outputs photometric uncertainty maps.

5 Experiments and results

In this section we describe and evaluate our framework on the KITTI dataset. We explore the observed improvements in performance, and perform an ablation study to determine the contribution of each component of the SUB-Depth training framework.

5.1 Dataset and metrics

KITTI [12] is a dataset that contains stereo images and corresponding 3D laser scans of outdoor scenes captured by imaging equipment mounted on a moving vehicle [25]. The RGB images have a resolution of $\approx 1241 \times 376$ and the corresponding depth maps are sparse with a large amount of missing data. For training, we adopt the dataset split proposed by [9] and resize images to 640×192. After removing the static frames by a pre-processing step suggested by [51], this results in 39,810 monocular frame triplets for training and 4,424 frame triplets for validation. To simplify the training process, the camera intrinsic matrices are assumed identical for all the frames in different scenes. To obtain this "universal" intrinsic matrix, we offset the principal point of the camera to the image centre and reset the focal length as the average of all the focal lengths in KITTI.

Depth metrics described by Eigen [9] are the most common used metrics for evaluating depth estimation accuracy. They include four error metrics: the Absolute Relative Error (Abs Rel), Squared Relative Error (Sq Rel), Root Mean Squared Error (RMSE), and the log of RMSE; accuracy metric: δ_1, δ_2, δ_3. We report each of these measures for each setting in our evaluation.

Uncertainty metric. Although uncertainty modelling is not our main contribution, we validate and compare the uncertainty outputs with two selected methods from Poggi et al. [37]. When evaluating uncertainty, we treat the depth regression uncertainty from DepthNet Θ_{depth} as depth uncertainty. From Ilg et al. [20], we use the Area Under the Sparsification Error (AUSE), the lower the better, and the Area Under the Random Gain (AURG), the higher the better, to quantify the uncertainty modelling performance of three depth metrics: Abs Rel, RMSE and δ_1, respectively in Table 3.

5.2 Evaluation on KITTI

To evaluate the performance of SUB-Depth, we select and retrain three model architectures from prior work using our training framework: Monodepth2 [14], HR-depth [34] and DIFFNet [50]. In each case, when compared to the original model (teacher DepthNet), we see significant improvements in all metrics. Table 2 displays this quantitative comparison for all standard metrics for KITTI. We particularly draw attention to the improvement for DIFFNet, a recent state-of-the-art model, that still exhibits substantial improvement. DIFFNet trained using SUB-Depth establishes a new level of performance on the KITTI corpus. In Table 3, We evaluate the uncertainty modelling performance on three different depth metrics. With respect to AUSE, our proposed method outperforms other competitors from Poggi et al. [37], while, for AURG, there is a marginal gap between ours and the Self method.

To validate the performance improvements gained by SUB-Depth and evaluate the contribution of each design, we conduct an ablation study as shown in Table 5. Monodepth2 [14] is used as the underlying architecture for all results reported in this table. The first row $\ell_{photometric}$ is the result from the standard SDE framework, and performs the worst of all

Table 2: **Quantitative comparison of SUB-Depth to existing SDE framework trained models on KITTI [12] Eigen split**. The best results in each subsection are in **bold**. Models trained with SUB-Depth outperform the same models trained with SDE in every case.

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 [50]	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

Table 3: **Quantitative comparison of uncertainty modelling.** We evaluate two uncertainty metrics for each selected depth metric and compare with two uncertainty modelling methods (Log and Self) in [37]. AUSE is lower the better, and AURG is higher the better.

	Abs Rel		RMSE		δ_1	
Method	AUSE	AURG	AUSE	AURG	AUSE	AURG
Poggi-Log [37]	0.051	0.027	3.097	1.188	0.060	0.056
Poggi-Self [37]	0.036	0.038	2.292	1.779	0.037	0.072
SUB-Depth	0.035	0.037	2.196	1.770	0.034	0.072

settings. In second row $\ell_{regression}$, by simply using the trained DepthNet as a teacher DepthNet we achieve improved performance across all measures. In last two rows, performance improves further as $\ell_{photometric}$ and $\ell_{regression}$ are combined and weighted by corresponding uncertainty estimation.

We offer additional evaluation on top-10 challenging subset [50] of KITTI, qualitative

Table 4: **Quantitative comparison of uncertainty modelling on improved ground truth [42]**.

	Abs Rel		RMSE		δ_1	
Method	AUSE	AURG	AUSE	AURG	AUSE	AURG
Poggi-Log [37]	0.039	0.020	2.562	0.916	0.044	0.038
Poggi-Self [37]	0.030	0.026	2.009	1.266	0.030	0.045
SUB-Depth	0.029	0.026	1.950	1.245	0.028	0.045

Table 5: **Ablation Studies**. We observe increased performance as self-distillation is introduced, and further improvements with the addition of uncertainty modelling. The best results in each subsection are in **bold**.

Objective	Abs Rel	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
$\ell_{photometric}$(Baseline)	0.115	0.903	4.863	0.193	0.877	0.959	0.981
$\ell_{regression}$	0.112	0.884	4.740	0.189	0.881	0.961	0.982
Ours(1:1 weighted)	0.112	0.888	4.766	0.189	0.882	0.961	0.981
Ours(uncertainty weighted)	**0.110**	**0.821**	**4.648**	**0.185**	**0.884**	**0.962**	**0.983**

results on KITTI, generalisation results on Cityscapes [7] and visualisation of error maps on Virtual KITTI [10] that are reported in supplementary material.

6 Conclusion

We presented a multi-task training framework for self-supervised monocular depth estimation, SUB-Depth. SUB-Depth extends the existing standard depth estimation framework with the introduction of self-distillation and uncertainty modelling. We introduce a teacher network and let the depth network be trained, not only for an image reconstruction task but also for a self-distillation task. To find the optimal objective weights, we utilize task-dependent uncertainty to weight losses for each task. Through analysing losses and uncertainty during training, we discovered that, initially the image reconstruction task contributes more than the self-distillation task, but then self-distillation quickly becomes the primary task since the estimated regression uncertainty is much lower than photometric uncertainty. We retrained three representative approaches using SUB-Depth to validate the generalisation capability of our proposed framework, and all outperform their counterparts. Our SUB-Depth training framework exhibits substantial improvements over the current state-of-the-art model on the KITTI benchmark for all depth metrics.

Acknowledgement
The research presented in this paper was carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia. .

References

[1] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. Bidirectional attention network for monocular depth estimation. In *International Conference on Robotics and Automation (ICRA)*, 2021.

[2] Hong Cai, Janarbek Matai, Shubhankar Borse, Yizhe Zhang, Amin Ansari, and Fatih Porikli. X-distill: Improving self-supervised monocular depth via cross-task distillation. *British Machine Vision Conference (BMVC)*, 2021.

[3] Sai Shyam Chanduri, Zeeshan Khan Suri, Igor Vozniak, and Christian Müller. Camlessmonodepth: Monocular depth estimation with unknown camera parameters. *arXiv preprint arXiv:2110.14347*, 2021.

[4] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Xiaotian Chen, Yuwang Wang, Xuejin Chen, and Wenjun Zeng. S2r-depthnet: Learning a generalizable depth-specific structural representation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] JaeHoon Choi, Dongki Jung, DongHwan Lee, and Changick Kim. Safenet: Self-supervised monocular depth estimation with semantic-aware feature extraction. In *Conference on Neural Information Processing Systems (NIPS)*, 2020.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Conference on Neural Information Processing Systems (NIPS)*, 2014.

[10] A Gaidon, Q Wang, Y Cabon, and E Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T. Barron. Learning single camera depth estimation using dual-pixels. In *International Conference on Computer Vision (ICCV)*, 2019.

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, 2013.

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. In *International Conference on Computer Vision (ICCV)*, 2019.

[15] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *International Conference on Computer Vision (ICCV)*, 2019.

[16] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *International Conference on Learning Representations (ICLR)*, 2020.

[18] Lei He, Jiwen Lu, Guanghui Wang, Shiyu Song, and Jie Zhou. Sosd-net: Joint semantic object segmentation and depth estimation from monocular images. *Neurocomputing*, 440:251–263, 2021.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[20] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *European Conference on Computer Vision (ECCV)*, September 2018.

[21] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[22] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Conference on Neural Information Processing Systems (NIPS)*, 2017.

[23] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[24] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[26] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision (ECCV)*, 2020.

[27] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning sfm from sfm. In *European Conference on Computer Vision (ECCV)*, pages 698–713, Munich, Germany, 2018. Springer.

[28] Varun Ravi Kumar, Senthil Yogamani, Markus Bach, Christian Witt, Stefan Milz, and Patrick Mader. Unrectdepthnet: Self-supervised monocular depth estimation using a generic framework for handling common camera distortion models. In *International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[29] Varun Ravi Kumar, Marvin Klingner, Senthil Yogamani, Stefan Milz, Tim Fingscheidt, and Patrick Mader. Syndistnet: Self-supervised monocular fisheye camera distance estimation synergized with semantic segmentation for autonomous driving. In *Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[30] Hyunmin Lee and Yongho Shin. Real-time stereo matching network with high accuracy. In *IEEE International Conference on Image Processing (ICIP)*, 2019.

[31] Sihaeng Lee, Janghyeon Lee, Byungju Kim, Eojindl Yi, and Junmo Kim. Patch-wise attention network for monocular depth estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[32] Lukas Liebel and Marco Körner. Multidepth: Single-image depth estimation via multi-task regression and classification. In *Intelligent Transportation Systems Conference (ITSC)*, 2019.

[33] Yawen Lu, Michel Sarkis, and Guoyu Lu. Multi-task learning for single image depth estimation and segmentation based on unsupervised network. In *International Conference on Robotics and Automation (ICRA)*, 2020.

[34] 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. *In AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[35] S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yağız Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

[37] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[38] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[39] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

[40] Takanori Senoh, Koki Wakunami, Hisayuki Sasaki, Ryutaro Oi, and Kenji Yamamoto. Fast depth estimation using non-iterative local optimization for super multi-view images. In *Global Conference on Signal and Information Processing*, 2015.

[41] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[42] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017.

[43] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Transactions on Image Processing (TIP)*, 2004.

[44] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[45] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[46] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[47] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

[48] Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, and Sean Fanello. Activestereonet: End-to-end self-supervised learning for active stereo systems. In *European Conference on Computer Vision (ECCV)*, 2018.

[49] Hang Zhou, David Greenwood, Sarah Taylor, and Han Gong. Constant velocity constraints for self-supervised monocular depth estimation. In *European Conference on Visual Media Production (CVMP)*, 2020.

[50] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. In *British Machine Vision Conference (BMVC)*, 2021.

[51] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.