Evaluation of multi-task uncertainties in joint semantic segmentation and monocular depth estimation

Steven Landgraf, Markus Hilleman, Theodor Kapler, and Markus Ulrich

Karlsruhe Institute of Technology (KIT),
Institute of Photogrammetry and Remote Sensing (IPF),
Karlsruhe, Germany.

Abstract Deep neural networks achieve outstanding results in perception tasks such as semantic segmentation and monocular depth estimation, making them indispensable in safety-critical applications like autonomous driving and industrial inspection. However, they often suffer from overconfidence and poor explainability, especially for out-of-domain data. While uncertainty quantification has emerged as a promising solution to these challenges, multi-task settings still need to be investigated in this regard. In an effort to shed light on this, we evaluate Monte Carlo Dropout, Deep Sub-Ensembles, and Deep Ensembles for joint semantic segmentation and monocular depth estimation. Thereby, we reveal that Deep Ensembles stand out as the preferred choice and show the potential benefit of multi-task learning with regard to the uncertainty quality in comparison to solving both tasks separately.

Keywords Deep learning, uncertainty quantification, multi-task learning, semantic segmentation, monocular depth estimation

1 Introduction

Deep neural networks are increasingly being used in real-time and safety-critical applications like autonomous driving [1], industrial inspection [2], and automation [3]. Although they achieve incomparable

DOI: 10.58895/ksp/1000174496-13 erschienen in:
Forum Bildverarbeitung 2024
DOI: 10.58895/ksp/1000174496 | https://www.ksp.kit.edu/site/books/m/10.58895/ksp/1000174496/

performance in fundamental perception tasks like semantic segmentation [4] or monocular depth estimation [5], they still suffer from problems like overconfidence [6], lack explainability [7], and struggle to distinguish between in-domain and out-of-domain samples [8].

In order to tackle these critical challenges and prevailing shortcomings of deep neural networks, a number of promising uncertainty quantification methods [9–12] have been proposed. Surprisingly, however, quantifying predictive uncertainties in the context of joint semantic segmentation and monocular depth estimation has not been thoroughly explored yet [13]. Since many real-world applications are multi-modal in nature and, hence, have the potential to benefit from multi-task learning, this is a substantial gap in current literature.

To this end, we conduct a comprehensive series of experiments to study how multi-task learning influences the quality of uncertainty estimates in comparison to solving both tasks separately. Our contributions can be summarized as follows:

- We combine three different uncertainty quantification methods - Monte Carlo Dropout (MCD), Deep Sub-Ensembles (DSE), and Deep Ensembles (DE) - with joint semantic segmentation and monocular depth estimation and evaluate how they perform in comparison to each other.
- In addition, we reveal the potential benefit of multi-task learning with regard to the uncertainty quality compared to solving semantic segmentation and monocular depth estimation separately.

2 Related Work

2.1 Joint Semantic Segmentation and Monocular Depth Estimation

Semantic segmentation and monocular depth estimation are both essential tasks in image understanding, requiring pixel-wise predictions from a single input image. Due to the strong correlation and complementary nature of these tasks, several previous works have focused on addressing them jointly [14–18].

Notably, almost all previous works employ out-of-date architectures and require complex adaptations to either the model, the training process, or both. Instead of following this trend, we adapt a modern

Vision-Transformer-based architecture similar to Xu et al. [18], achieving competitive predictive performance while maintaining simplicity and transparency of the results.

2.2 Uncertainty Quantification

In order to address the shortcomings of deep neural networks, a variety of uncertainty quantification methods [9–12] and studies [19–21] have been proposed. The predictive uncertainty can be decomposed into aleatoric and epistemic uncertainty [22], which can be an essential for applications like active learning and detecting out-of-distribution samples [23]. The aleatoric component captures the irreducible data uncertainty, such as image noise or noisy labels from imprecise measurements. The epistemic uncertainty accounts for the model uncertainty and can be reduced with more or higher quality training data [22, 24].

Remarkably, quantifying uncertainties in joint semantic segmentation and monocular depth estimation has been largely overlooked [13]. Therefore, we compare multiple uncertainty quantification methods for this task and show how multi-task learning influences the quality of the uncertainty quality in comparison to solving both tasks separately.

3 Evaluation Strategy

3.1 Baseline Models.

To explore the impact of multi-task learning on the uncertainty quality, we conduct our evaluations with three models:

1. SegFormer [25] for the segmentation task,
2. DepthFormer for the depth estimation task,
3. SegDepthFormer for joint semantic segmentation and monocular depth estimation.

SegFormer. For solving the semantic segmentation task by itself, we use SegFormer [25], a modern Transformer-based architecture. Due to its high efficiency and performance, it is particularly suitable for real-time applications that might rely on uncertainty quantification. We

train all SegFormer models with the categorical Cross-Entropy loss

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\cdot\log(p(z)_{n,c}) \tag{1}$$

for a single image, where N is the number of pixels in the image, C is the number of classes, $y_{n,c}$ is the corresponding ground truth label, and $p(z)_{n,c}$ is the predicted softmax probability.

To obtain a measure for the aleatoric uncertainty [24] of the baseline model, we compute the predictive Entropy

$$H(p(z)) = -\sum_{c=1}^{C} p(z)_c\cdot\log(p(z)_c)\ . \tag{2}$$

DepthFormer. Highly inspired by the efficiency and performance of SegFormer [25], we propose DepthFormer for monocular depth estimation. We use the same hierarchical Transformer-based encoder and all-MLP decoder. In contrast to SegFormer, the output layer differs by having two output channnels: one for the predictive mean $\mu(z)$ and one for the predictive variance $s^2(z)$ [26]. The first output channel uses a ReLU output activation function, while the second output channel applies Softplus activation, which is a smooth approximation of the ReLU functon with the advantage of being differentiable at $z = 0$. We found Softplus to work better than ReLU for the predictive variance, following the work of Lakshminarayanan et al. [11].

For all DepthFormer models we follow Nix and Weigend [27] and treat the output of the model as a sample from a Gaussian distribution with the predictive mean $\mu(z)$ and a corresponding predictive variance $s^2(z)$. Based on this, we can minimize the Gaussian Negative Log-Likelihood (GNLL) loss

$$\mathcal{L}_{\text{GNLL}} = \frac{1}{2}\left(\frac{(y-\mu(z))^2}{s^2(z)} + \log(s^2(z))\right)\ , \tag{3}$$

where y is the the ground truth depth.

Through GNLL minimization, DepthFormer inherently learns corresponding variances, which can be interpreted as the aleatoric uncertainty [24,26].

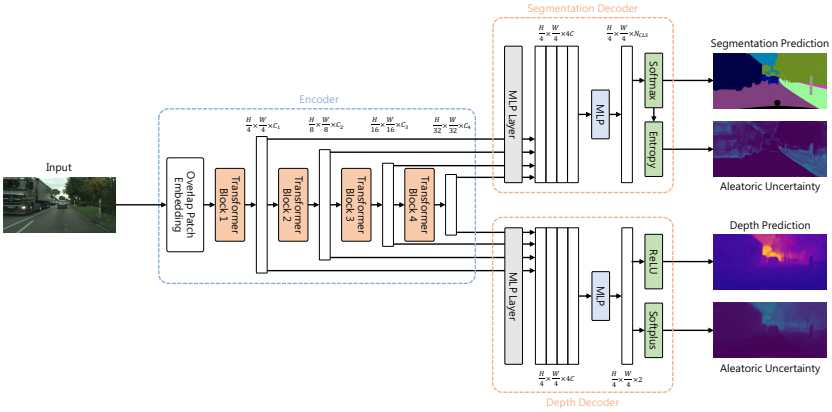

Figure 1: A schematic overview of the SegDepthFormer architecture. It combines the SegFormer [25] architecture with a lightweight all-MLP depth decoder.

SegDepthFormer. To jointly solve semantic segmentation and monocular depth estimation, we propose SegDepthFormer. The architecture, which is shown in Figure 1, combines SegFormer [25] and DepthFormer. It comprises three modules: a hierarchical Transformber-based encoder, an all-MLP segmentation decoder, and an all-MLP depth decoder. Both decoders fuse the multi-level features obtained through the shared encoder to solve the joint prediction task.

SegDepthFormer is trained to minimize the weighted sum of the two previously described objective functions: $\mathcal{L} = \mathcal{L}_{\text{CE}} + w_1 \mathcal{L}_{\text{GNLL}}$, where w_1 is a weighting factor, which we set to $w_1 = 1$ for the sake of simplicity and because both loss values are of similar magnitude.

The respective aleatoric uncertainty is obtained by computing the predictive entropy $H(p(z))$ for the segmentation task or by the predictive variance $s^2(z)$, which is learned implicitly through the optimization of $\mathcal{L}_{\text{GNLL}}$.

3.2 Uncertainty Quantification

We evaluate Monte Carlo Dropout (MCD) [10], Deep Ensembles (DEs) [11], and Deep Sub-Ensembles (DSEs) [12], motivated by their simplicity, ease of implementation, parallelizability, minimal tuning require-

ments, and state-of-the-art performance.

Monte Carlo Dropout. MCD depends on the number and placement of dropout layers and particularly the dropout rate. We adopt the original SegFormer [25] layer placement and consider two dropout rates , 20% and 50%. We sample ten times to obtain the prediction and predictive uncertainty [10,28].

Deep Ensemble. DEs achieve the best results if they are trained to explore diverse modes in function space, which we accomplish by randomly initializing all decoder heads, using random augmentations, and by applying random shuffling of the training data points [11,29]. We report results of a DE with ten members, following the suggestions of previous work [11,29,30].

Deep Sub-Ensemble. Consistent with DEs and MCD, we train the DSE with ten decoder heads for each task on top of a shared encoder [12]. During training, we only optimize a single decoder head per training batch and alternate between them. Thereby, we aim to introduce as much randomness as possible, analogous to the training of DEs. For inference, we utilize all decoder heads.

4 Experimental Setup

Predictions. Regardless of the uncertainty quantification method, we report the results of the mean prediction.

Uncertainty. For the segmentation task, we compute the predictive entropy based on the mean softmax probabilities as a measure for the predictive uncertainty [31]. For the depth estimation task, however, we calculate the predictive uncertainty based on the mean predictive variance and the variance of the depth predictions of the samples [26].

Datasets. We conduct all experiments on Cityscapes [32] and NYUv2 [33].

Data Augmentations. Regardless of the trained model, we apply random scaling with a factor between 0.5 and 2.0, random cropping with a crop size of 768×768 pixels on Cityscapes and 480×640 pixels on NYUv2, and random horizontal flipping with a flip chance of 50%.

Implementation Details. For all training processes, we use AdamW [34] optimizer with a base learning rate of $6 \cdot 10^{-5}$ and employ a polynomial rate scheduler. Besides, we use a batch size of 8 and train for

250 epochs on Cityscapes and for 100 epochs NYUv2, respectively.

Metrics. For semantic segmentation, we report mean Intersection over Union (mIoU) and Expected Calibration Error (ECE) [35]. For monocular depth estimation, we use root mean squared error (RMSE). The uncertainty is evaluated using the following metrics proposed by Mukhoti and Gal [31]:

1. $p(accurate|certain)$: The probability of accurate predictions given low uncertainty.
2. $p(uncertain|inaccurate)$: The probability of high uncertainty given inaccurate predictions.
3. *PAvPU*: The combination of both cases, i.e. accurate|certain and inaccurate|uncertain.

Although these metrics have originally been proposed for semantic segmentation [31], we also use them to evaluate the depth uncertainty. We use the following formula to determine whether a depth prediction is accurate:

$$\max\left(\frac{\mu(z)}{y}, \frac{y}{\mu(z)}\right) = \delta_1 < 1.25 \ , \tag{4}$$

where $\mu(z)$ is the predicted depth value of a pixel and y is the corresponding ground truth depth.

For the sake of simplicity and to simulate real-world employment, we set the uncertainty threshold to the mean uncertainty of a given image for all evaluations.

5 Results

In this section, we describe the results of our joint uncertainty evaluation quantitatively. Tables 1 and 2 contain a detailed comparison, primarily focusing on the uncertainty quality.

Single-task vs. Multi-task. Looking at the differences between the single-task models, SegFormer and DepthFormer, and the multi-task model, SegDepthFormer, the single-task models generally deliver slightly better prediction performance. However, SegDepthFormer exhibits greater uncertainty quality for the semantic segmentation task in comparison to SegFormer. This is particularly evident for

Table 1: Quantitative comparison on the Cityscapes dataset [32] between the three baseline models paired with MCD, DSE, and DEs, respectively. Best results are marked in **bold**.

		Semantic Segmentation					Monocular Depth Estimation				
		mIoU ↑	ECE ↓	p(acc/cer) ↑	p(inacc/unc) ↑	PAvPU ↑	RMSE ↓	p(acc/cer) ↑	p(inacc/unc) ↑	PAvPU ↑	Inference Time [ms]
Baseline	SegFormer	0.772	0.033	0.882	0.395	0.797	-	-	-	-	17.90 ± 0.47
	DepthFormer	-	-	-	-	-	7.452	0.749	0.476	0.766	17.59 ± 0.82
	SegDepthFormer	0.738	0.028	0.913	0.592	0.826	7.536	0.745	0.472	0.762	22.04 ± 0.27
MCD (20%)	SegFormer	0.759	**0.007**	0.883	0.424	0.780	-	-	-	-	177.13 ± 0.64
	DepthFormer	-	-	-	-	-	7.956	0.749	0.555	0.739	139.32 ± 0.78
	SegDepthFormer	0.738	0.020	0.911	0.592	0.803	7.370	0.761	0.523	0.757	202.23 ± 0.39
MCD (50%)	SegFormer	0.662	0.028	0.883	0.485	0.760	-	-	-	-	176.98 ± 0.53
	DepthFormer	-	-	-	-	-	21.602	0.181	0.366	0.431	139.81 ± 1.20
	SegDepthFormer	0.640	0.021	0.906	0.616	0.782	8.316	0.733	**0.558**	0.723	203.82 ± 0.81
DSE	SegFormer	0.772	0.037	0.890	0.456	0.797	-	-	-	-	132.30 ± 3.16
	DepthFormer	-	-	-	-	-	**7.036**	0.762	0.467	0.772	91.82 ± 2.01
	SegDepthFormer	0.749	0.009	**0.931**	**0.696**	**0.844**	7.441	0.751	0.463	0.766	212.11 ± 8.44
DE	SegFormer	**0.784**	0.033	0.887	0.416	0.798	-	-	-	-	667.51 ± 2.89
	DepthFormer	-	-	-	-	-	7.222	0.759	0.486	0.771	626.79 ± 2.05
	SegDepthFormer	0.755	0.015	0.917	0.609	0.828	7.156	**0.763**	0.493	**0.773**	743.23 ± 32.95

Table 2: Quantitative comparison on the NYUv2 dataset [33] between the three baseline models paired with MCD, DSE, and DEs, respectively. Best results are marked in **bold**.

		Semantic Segmentation					Monocular Depth Estimation				
		mIoU ↑	ECE ↓	p(acc/cer) ↑	p(inacc/unc) ↑	PAvPU ↑	RMSE ↓	p(acc/cer) ↑	p(inacc/unc) ↑	PAvPU ↑	Inference Time [ms]
Baseline	SegFormer	0.470	0.159	0.768	0.651	**0.734**	-	-	-	-	18.09 ± 0.41
	DepthFormer	-	-	-	-	-	0.554	0.786	0.449	0.610	17.51 ± 0.87
	SegDepthFormer	0.466	0.151	0.769	0.659	0.733	0.558	0.776	0.446	0.594	22.31 ± 0.23
MCD (20%)	SegFormer	0.422	0.102	0.767	0.706	0.724	-	-	-	-	222.67 ± 0.61
	DepthFormer	-	-	-	-	-	0.605	0.741	0.478	0.568	139.58 ± 052
	SegDepthFormer	0.433	0.093	0.771	0.710	0.725	0.610	0.731	0.450	0.560	251.25 ± 0.81
MCD (50%)	SegFormer	0.273	0.083	0.705	**0.722**	0.713	-	-	-	-	223.25 ± 0.82
	DepthFormer	-	-	-	-	-	0.978	0.516	**0.492**	0.526	139.27 ± 0.69
	SegDepthFormer	0.272	0.084	0.702	0.721	0.711	0.837	0.576	0.473	0.525	251.98 ± 0.60
DSE	SegFormer	0.469	0.092	0.776	0.681	0.726	-	-	-	-	180.42 ± 3.93
	DepthFormer	-	-	-	-	-	0.547	0.782	0.423	0.596	91.66 ± 0.26
	SegDepthFormer	0.461	**0.077**	0.776	0.692	0.723	0.584	0.738	0.403	0.573	261.69 ± 5.10
DE	SegFormer	**0.486**	0.125	0.782	0.675	**0.734**	-	-	-	-	715.97 ± 7.55
	DepthFormer	-	-	-	-	-	**0.524**	**0.808**	0.475	**0.613**	624.30 ± 2.07
	SegDepthFormer	0.481	0.122	**0.783**	0.682	0.733	0.552	0.785	0.453	0.590	788.76 ± 2.00

$p(uncertain|inaccurate)$ on Cityscapes. For the depth estimation task, there is no significant difference in terms of uncertainty quality.

Baseline Models. As expected, the baseline models have the lowest inference times, being 5 to 30 times faster without using any uncertainty quantification method. While their prediction performance turns out to be quite competitive, only beaten by DEs, they show poor calibration and uncertainty quality for semantic segmentation. Surprisingly, the uncertainty quality for the depth estimation task is very

decent, often only surpassed by the DE.

Monte Carlo Dropout. MCD causes a significantly higher inference time compared to the respective baseline model. Additionally, leaving dropout activated during inference to sample from the posterior has a detrimental effect on the prediction performance, particularly with a 50% dropout ratio. Nevertheless, MCD outputs well-calibrated softmax probabilities and uncertainties, although the results should be interpreted with caution because of the deteriorated prediction quality.

Deep Sub-Ensemble. Across both datasets, DSEs show comparable prediction performance compared with the baseline models. Notably, DSEs consistently demonstrate a high uncertainty quality across all metrics, particularly in the segmentation task on Cityscapes.

Deep Ensemble. In accordance with previous work [28], DEs emerge as state-of-the-art, delivering the best prediction performance and mostly superior uncertainty quality. At the same time, DEs suffer from the highest computational cost.

6 Conclusion

By comparing uncertainty quantification methods in joint semantic segmentation and monocular depth estimation, we find Deep Ensembles offer the best performance and uncertainty quality, albeit at higher computational cost. Deep Sub-Ensembles provide an efficient alternative with minimal trade-offs. Additionally, we reveal the potential benefit of multi-task learning with regard to uncertainty quality of the semantic segmentation task compared to solving both tasks separately.

Acknowledgments

The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

References

1. R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, and A. Weller, “Concrete Problems for Autonomous Vehicle Safety: Advantages

of Bayesian Deep Learning," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 4745–4753.

2. C. Steger, M. Ulrich, and C. Wiedemann, *Machine Vision Algorithms and Applications*. John Wiley & Sons, 2018.
3. S. Landgraf, M. Hillemann, M. Aberle, V. Jung, and M. Ulrich, "Segmentation of industrial burner flames: A comparative study from traditional image processing to machine and deep learning." *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 10, 2023.
4. S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, 2022.
5. X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, "Towards real-time monocular depth estimation for robotics: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 16 940–16 961, 2022.
6. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.
7. J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. X. Zhu, "A Survey of Uncertainty in Deep Neural Networks," *arXiv:2107.03342*, 2022.
8. K. Lee, H. Lee, K. Lee, and J. Shin, "Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples," *arXiv:1711.09325*, 2018.
9. D. J. C. MacKay, "A Practical Bayesian Framework for Backpropagation Networks," *Neural Computation*, vol. 4, no. 3, pp. 448–472, 1992.
10. Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 48. PMLR, 2016, pp. 1050–1059.
11. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.
12. M. Valdenegro-Toro, "Sub-ensembles for fast uncertainty estimation in neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4119–4127.

13. S. Landgraf, M. Hillemann, T. Kapler, and M. Ulrich, "Efficient multi-task uncertainties for joint semantic segmentation and monocular depth estimation," in *DAGM German Conference on Pattern Recognition*. Springer, 2024.
14. L. He, J. Lu, G. Wang, S. Song, and J. Zhou, "Sosd-net: Joint semantic object segmentation and depth estimation from monocular images," *Neurocomputing*, vol. 440, pp. 251–263, 2021.
15. T. Gao, W. Wei, Z. Cai, Z. Fan, S. Q. Xie, X. Wang, and Q. Yu, "Ci-net: A joint depth estimation and semantic segmentation network using contextual information," *Applied Intelligence*, vol. 52, no. 15, pp. 18 167–18 186, 2022.
16. N. Ji, H. Dong, F. Meng, and L. Pang, "Semantic segmentation and depth estimation based on residual attention mechanism," *Sensors*, vol. 23, no. 17, p. 7466, 2023.
17. D. Brüggemann, M. Kanakis, A. Obukhov, S. Georgoulis, and L. Van Gool, "Exploring relational context for multi-task dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 869–15 878.
18. X. Xu, H. Zhao, V. Vineet, S.-N. Lim, and A. Torralba, "Mtformer: Multi-task learning via transformer and cross-task reasoning," in *European Conference on Computer Vision*. Springer, 2022, pp. 304–321.
19. S. Landgraf, M. Hillemann, K. Wursthorn, and M. Ulrich, "Uncertainty-aware cross-entropy for semantic segmentation," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. X-2, 2024.
20. K. Wursthorn, M. Hillemann, and M. Ulrich, "Uncertainty quantification with deep ensembles for 6d object pose estimation," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. X-2, 2024.
21. D. W. Wolf, P. Balaji, A. Braun, and M. Ulrich, "Decoupling of neural network calibration measures," in *DAGM German Conference on Pattern Recognition*. Springer, 2024.
22. Y. Gal, "Uncertainty in deep learning," *Ph.D. thesis, University of Cambridge*, 2016.
23. Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *International conference on machine learning*. PMLR, 2017, pp. 1183–1192.
24. A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 5580–5590.

25. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

26. A. Loquercio, M. Segu, and D. Scaramuzza, "A general framework for uncertainty estimation in deep learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3153–3160, 2020.

27. D. A. Nix and A. S. Weigend, "Estimating the mean and variance of the target probability distribution," in *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, vol. 1. IEEE, 1994, pp. 55–60.

28. F. K. Gustafsson, M. Danelljan, and T. B. Schon, "Evaluating scalable bayesian deep learning methods for robust computer vision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 318–319.

29. S. Fort, H. Hu, and B. Lakshminarayanan, "Deep Ensembles: A Loss Landscape Perspective," *arXiv:1912.02757*, 2020.

30. S. Landgraf, K. Wursthorn, M. Hillemann, and M. Ulrich, "Dudes: Deep uncertainty distillation using ensembles for semantic segmentation," *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, vol. 92, no. 2, pp. 101–114, 2024.

31. J. Mukhoti and Y. Gal, "Evaluating bayesian deep learning methods for semantic segmentation," *arXiv preprint arXiv:1811.12709*, 2018.

32. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

33. N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, 2012, Proceedings, Part V 12*. Springer, 2012, pp. 746–760.

34. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

35. M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 29, no. 1, 2015.