Semantic segmentation and uncertainty quantification with vision transformers for industrial applications

Kira Wursthorn[1], Lili Gao[2], Steven Landgraf[1], and Markus Ulrich[1]

[1] Karlsruhe Institute of Technology (KIT), Institute of Photogrammetry and Remote Sensing (IPF), Englerstr. 7, 76131 Karlsruhe
[2] Torc Robotics, Augsburger Str. 540, 70327 Stuttgart

Abstract Vision Transformers (ViTs) have recently achieved state-of-the-art performance in semantic segmentation tasks. However, their deployment in critical applications necessitates reliable uncertainty quantification to assess model confidence. To tackle this challenge, we combine a state-of-the-art ViT with the popular uncertainty quantification method Monte Carlo Dropout (MCD) to predict both segmentation and uncertainty maps. We focus on an industrial machine vision setting and carry out the experiments on the T-LESS dataset. The evaluation is carried out with regard to both the segmentation accuracy and the predicted uncertainties using appropriate metrics.

Keywords Semantic segmentation, uncertainty quantification, vision transformers

1 Introduction

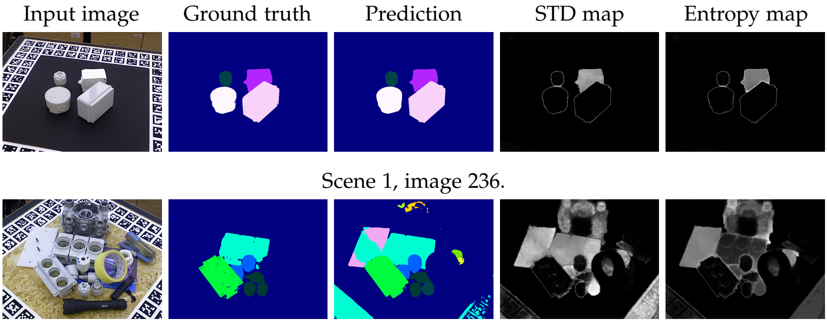
In computer vision, deep-learning-based approaches like convolutional neural networks (CNNs) have proven their success at solving the fundamental task of semantic segmentation of (RGB) images. Recently, Vision Transformers (ViTs) have been applied to this task and have gained much attention. The prediction of pixel-wise class labels in images is relevant for applications such as autonomous driving, and quality assurance in industry. These applications involve safety-critical

DOI: 10.58895/ksp/1000174496-12 erschienen in:
Forum Bildverarbeitung 2024
DOI: 10.58895/ksp/1000174496 | https://www.ksp.kit.edu/site/books/m/10.58895/ksp/1000174496/

and high-risk scenarios. Therefore, it is important to not only predict the class labels correctly but also to determine the prediction's reliability [1–4]. Estimating uncertainty of predictions allows to make informed decisions and to identify potentially inaccurate predictions.

Most classification and segmentation tasks use softmax to estimate class-wise pseudo probabilities to quantify the confidence in the predictions. It is well-known that softmax predictions tend to be overconfident, especially in cases where the input data of the model is out-of-domain [5, 6]. One popular method to quantify uncertainty in deep learning is Monte-Carlo Dropout (MCD) [7] that uses dropout at inference time. Multiple forward passes are used to sample from the posterior distribution of the predictions and approximate it, e.g., with a Gaussian distribution. The final segmentation map is determined by assigning each pixel the class with the highest average softmax output across all classes. The corresponding uncertainty map is either its standard deviation (STD) over the samples or the entropy of the mean values over the classes.

In this contribution, we combine a state-of-the-art ViT, the SegFormer [8], with MCD for semantic segmentation with uncertainty quantification (UQ). We choose SegFormer as our ViT baseline because of its efficient design and good performance, which both are relevant criteria in industry. Our goal is to quantify the quality and reliability of the SegFormer's predicted semantic segmentation maps as well as the corresponding uncertainty maps for industrial applications. Therefore, we train the model on the T-LESS [9] dataset that consists of various scenes of parts with characteristics that are typical for industry. As part of the Benchmark for 6D Object Pose Estimation (BOP) [10], the T-LESS training set can be augmented with physically-based rendered (PBR) synthetic training data. While the real training images show systematically captured and isolated views of each object respectively, the PBR subset consists of cluttered scenes with varying image acquisition conditions, scene backgrounds, and occlusions by both T-LESS objects and those of other BOP datasets. Figure 1 shows two examples of the T-LESS dataset from both a simple as well as a cluttered scene together with the corresponding segmentation and uncertainty maps that our trained uncertainty-aware SegFormer model predicted. We use the mean Intersection over Union (IoU) and the expected calibration error (ECE) [11] as metrics to measure the segmentation quality and model



Scene 1, image 236.

Scene 18, image 206.

Figure 1: Example predictions of segmentation and uncertainty maps for images from a simple (top row) and a complex scene (bottom row) of the T-LESS test dataset, using the MCD with a dropout rate of 30 % and 20 samples. In the uncertainty maps, brighter pixels represent higher uncertainty values.

calibration and the Patch Accuracy versus Patch Uncertainty (PAvPU), $p(\textbf{accurate}|\textbf{certain})$, and $p(\textbf{uncertain}|\textbf{inaccurate})$ [12] for the uncertainty evaluation.

After giving a short overview over the state-of-the-art approaches for semantic segmentation with ViTs and uncertainty quantification in Section 2, we explain our training and evaluation methodology in Section 3. In Section 4, we describe our experiments and present our results, which are discussed in Section 5. Section 6 concludes our paper.

2 Related Work

Due to the success of ViTs for image classification, many publications have been dedicated to applying the method to the task of semantic segmentation. Next to SegFormer, notable approaches include Segmenter [13], SETR [14], MaskFormer [15] and its successor Mask2Former [16] as well as general ViT approaches for dense predictions like Swin Transformer [17], DPT [18], and HRFormer [19].

Regarding UQ in RGB image-based semantic segmentation tasks, many works have successfully integrated MCD in their workflows, including applications like landcover prediction from remote sensing images [20], medical imaging [21], autonomous driving, and robotics [22–24]. To overcome the disadvantage of the additional runtime of sample-based UQ methods, knowledge distillation can be applied [25].

Recently, successful efforts have been made to combine SegFormer with UQ. While Chen et al. [26] propose their own UQ approach and compare its performance against MCD and ensembling using SegFormer, Landgraf et al. [27] add monocular depth estimation and and UQ with MCD to the SegFormer architecture. Both works conduct their experiments in the context of autonomous driving.

3 Methodology

Our methodology aims to achieve two main goals: i) Training and testing a SegFormer model to achieve the best possible segmentation performance on T-LESS, and ii) combining SegFormer with MCD for UQ. Both the segmentation and the uncertainty results are evaluated by their respective metrics (see below). The first goal provides a basic training setup, including suitable hyperparameters such as learning rate, model backbone, dataset settings, and data augmentations. This also leads to a baseline model without UQ. Next to testing the segmentation quality of the baseline model, it also includes the evaluation of the mean segmentation maps of the trained MCD models and the influence of performing dropout at inference time. For this, the mean IoU and the ECE metrics are used. The second goal that focuses on the UQ with SegFormer includes model training with different dropout rates for MCD and the evaluation of the predicted uncertainty maps with different sample sizes.

The uncertainty evaluation metrics proposed by Mukhoti and Gal (2018) [12] are computed based on the confusion matrix that includes four categories of pixel counts: accurate and certain (n_{ac}), accurate and uncertain (n_{au}), inaccurate and certain (n_{ic}), and inaccurate and uncertain (n_{iu}). To determine whether a prediction is certain or uncertain, an uncertainty threshold has to be defined. Here, we use the mean uncertainty over all pixels across the T-LESS test dataset. Based on the estimated counts, two metrics are computed that are defined as $p(\textbf{accurate}|\textbf{certain}) = n_{ac}/(n_{ac} + n_{ic})$ and $p(\textbf{uncertain}|\textbf{inaccurate}) = n_{iu}/(n_{ic} + n_{iu})$. The former returns higher values if predictions are accurate when the model is certain. The latter returns higher values if the model is uncertain when the predictions are inaccurate. Consequently, meaningful uncertainty values lead to large values for both metrics. Furthermore, the third metric PAvPU $= (n_{ac} + n_{iu})/(n_{ac} + n_{au} + n_{ic} + n_{iu})$ combines the first two metrics and, hence, presents an equivalent UQ metric to an overall accuracy. In the following, the metrics $p(\mathbf{accurate}|\mathbf{certain})$ and $p(\mathbf{uncertain}|\mathbf{inaccurate})$ are abbreviated as p_{ac} and p_{ui}.

4 Experiments

To address our first goal described in Section 3, we test different combinations of hyperparameters and training settings. We find that the best model performance in terms of mean IoU on the BOP test dataset of T-LESS is achieved by combining both real and PBR training data, a SegFormer-B5 backbone, and a learning rate of $6 \cdot 10^{-5}$. The combination of real and synthetic training data increases the mean IoU by roughly 50 %. Thus, we train all models in our experiments on both training data subsets. Similarly, a subsequent increase in the size of the backbone from B1 to B5 leads to increasing mean IoU scores and decreasing ECE values. For instance, replacing the smaller SegFormer-B1 architecture with the larger SegFormer-B5, which has the highest parameter count, leads to a 19.23 % increase in mean IoU and a 5.38 % reduction in ECE, as shown in Table 1. Thus, we select SegFormer-B5 for testing different subsets of data augmentation techniques of the AugSeg [28] framework. AugSeg includes geometric augmentations (random flip, random scale, and random crop) as well as a list of intensity-based augmentations (e.g., blurring, brightness and contrast modifications). The hyperparameter k denotes how many intensity-based augmentation techniques are randomly selected for each training instance. The results in terms of mean IoU and ECE are shown in Table 1.

We find that a combination of the geometric augmentations and the random intensity-based augmentations with a random selection parameter $k = 3$ works best, both in terms of highest mean IoU of 79.40 % and lowest ECE value of 4.29 %. We also test different learning rates where a learning rate of $1 \cdot 10^{-4}$ achieves the best results of a mean

IoU of 80.70 % and an ECE of 2.88 %. As learning rates higher than $2 \cdot 10^{-4}$ lead to model divergence during in our experiments at training time, we adopted the learning rate of $6 \cdot 10^{-5}$ of the original SegFromer publication to guarantee a stable training procedure. For a better comparison, all models are trained for 100 epochs on a NVIDIA H100 hardware.

Table 1: Ablation study using different model backbones and geometric and intensity-based data augmentation techniques with different values of the random selection parameter k from AugSeg [28].

	Augmentations		Metrics in %	
Model	Geometric	Intensity-based	Mean IoU ↑	ECE ↓
SegFormer-B1	-	-	44.69	9.92
SegFormer-B5	-	-	63.92	8.92
	✓	-	74.81	6.65
	✓	$k = 1$	76.35	6.07
	✓	$k = 3$	**79.40**	**4.29**
	✓	$k = 5$	79.22	5.30

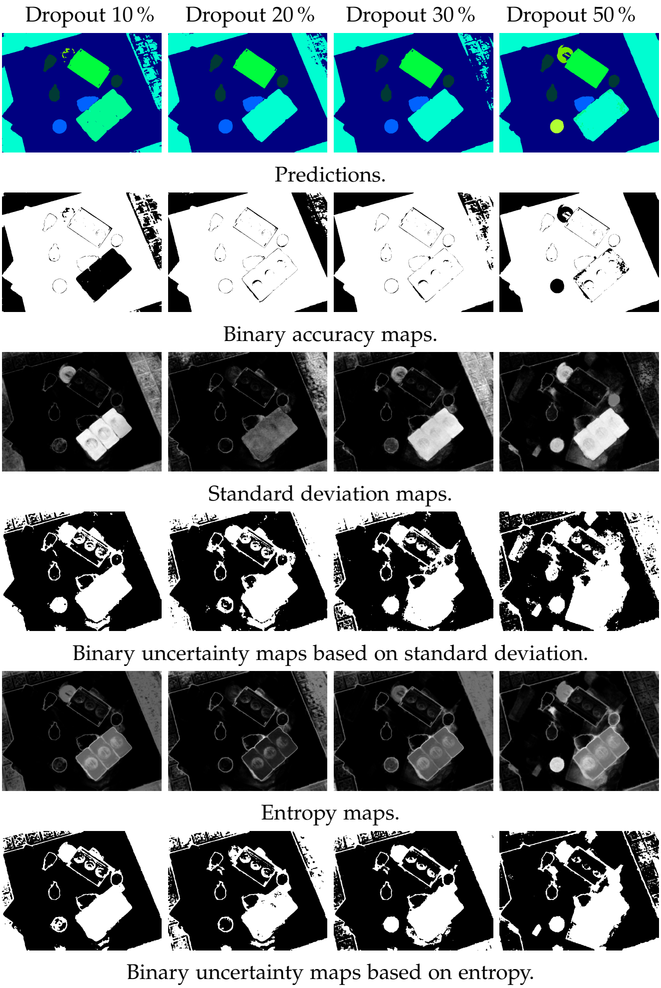
In order to incorporate MCD for the second goal of UQ, we activate the implemented but dormant dropout layers in the SegFormer architecture. We train the models with dropout rates of 10 %, 20 %, 30 %, and 50 % resulting in four different models. In contrast to dropout regularization, the dropout layers remain active for MCD at test time to obtain samples. We evaluate each model with sample sizes of $N = \{2, 5, 10, 20, 100\}$ respectively and compare them using both mean IoU and ECE for segmentation quality and p_{ac}, p_{ui}, and PAvPU for uncertainty quality. The results are summarized in Table 2.

Our evaluations show that smaller dropout rates lead to a higher mean IoU but not necessarily to lower ECE values. With regard to UQ metrics, all models achieve similar scores. Furthermore, increasing values for N result in increasing values in p_{ac} and p_{ui}, as expected. However, they surprisingly also result in slightly lower PAvPU scores. This is caused by decreasing counts of n_{ac} with increasing N. Nevertheless, these changes in PAvPU as well as in mean IoU and ECE are not substantial as they are all smaller than 3 %. In terms of required runtime, the minimum sample size of $N = 2$ takes around 89 ms while $N = 100$ results in 3711 ms runtime. Therefore, in time-critical applications, it should be possible to decrease N in order to speed-up the application without sacrificing too much predictive quality. For example, an uncertainty-aware prediction with $N = 20$ takes less than a second at 751 ms.

Table 2: Performance of SegFormer-B5 with MCD. Tested were different dropout rates and sample sizes N. The results were evaluated in terms of both the segmentation and uncertainty quality using the respective metrics described in Section 3. The subscript ″std″ indicates that the metrics are based on standard deviation, while the subscript ″en″ indicates that the metrics are based on entropy. All metrics are in %.

N	$p_{ac,\text{std}} \uparrow$	$p_{ui,\text{std}} \uparrow$	$\text{PAvPU}_{\text{std}} \uparrow$	$p_{ac,\text{en}} \uparrow$	$p_{ui,\text{en}} \uparrow$	$\text{PAvPU}_{\text{en}} \uparrow$	Mean IoU ↑	ECE ↓
dropout rate = 10 %								
2	98.88 ± 0.01	74.09 ± 0.15	92.41 ± 0.03	99.49 ± 0.01	88.14 ± 0.08	90.48 ± 0.03	76.93 ± 0.08	5.75 ± 0.13
5	99.28 ± 0.01	81.76 ± 0.11	91.74 ± 0.02	99.53 ± 0.01	88.60 ± 0.07	90.25 ± 0.02	77.13 ± 0.06	5.68 ± 0.11
10	99.34 ± 0.01	83.00 ± 0.07	91.42 ± 0.02	99.55 ± 0.01	88.83 ± 0.05	90.14 ± 0.02	77.24 ± 0.06	5.64 ± 0.10
20	99.38 ± 0.01	83.64 ± 0.07	91.17 ± 0.02	99.56 ± 0.01	88.96 ± 0.04	90.07 ± 0.02	77.25 ± 0.05	5.62 ± 0.09
100	99.43 ± 0.00	84.44 ± 0.03	90.82 ± 0.01	99.57 ± 0.00	89.12 ± 0.03	90.01 ± 0.01	77.25 ± 0.03	5.63 ± 0.02
dropout rate = 20 %								
2	98.25 ± 0.02	73.16 ± 0.20	91.57 ± 0.05	99.09 ± 0.01	87.94 ± 0.11	89.68 ± 0.04	75.51 ± 0.12	6.09 ± 0.13
5	98.78 ± 0.01	81.70 ± 0.15	90.78 ± 0.05	99.19 ± 0.01	88.77 ± 0.09	89.35 ± 0.04	75.95 ± 0.09	6.12 ± 0.14
10	98.93 ± 0.01	83.27 ± 0.11	90.37 ± 0.03	99.24 ± 0.01	89.11 ± 0.08	89.17 ± 0.03	76.10 ± 0.07	6.10 ± 0.10
20	99.01 ± 0.01	84.14 ± 0.08	90.02 ± 0.02	99.27 ± 0.00	89.32 ± 0.05	89.03 ± 0.02	76.16 ± 0.07	6.06 ± 0.07
100	99.12 ± 0.01	85.23 ± 0.03	89.50 ± 0.02	99.30 ± 0.00	89.55 ± 0.02	88.93 ± 0.01	76.27 ± 0.04	6.00 ± 0.04
dropout rate = 30 %								
2	98.49 ± 0.02	74.39 ± 0.28	91.54 ± 0.04	99.28 ± 0.01	89.02 ± 0.13	89.58 ± 0.04	74.52 ± 0.17	5.57 ± 0.20
5	98.04 ± 0.01	83.35 ± 0.17	90.56 ± 0.04	99.39 ± 0.01	89.95 ± 0.09	89.15 ± 0.04	75.00 ± 0.15	5.51 ± 0.17
10	98.20 ± 0.01	85.09 ± 0.09	90.03 ± 0.03	99.45 ± 0.01	90.42 ± 0.08	88.90 ± 0.02	75.24 ± 0.11	5.45 ± 0.13
20	99.28 ± 0.01	86.09 ± 0.10	89.62 ± 0.03	99.48 ± 0.01	90.70 ± 0.07	88.75 ± 0.03	75.34 ± 0.09	5.42 ± 0.14
100	99.39 ± 0.01	87.36 ± 0.07	88.96 ± 0.02	99.51 ± 0.00	90.99 ± 0.04	88.61 ± 0.02	75.41 ± 0.04	5.43 ± 0.06
dropout rate = 50 %								
2	95.84 ± 0.03	66.93 ± 0.33	88.12 ± 0.04	97.48 ± 0.03	83.34 ± 0.13	86.93 ± 0.03	68.66 ± 0.22	7.78 ± 0.16
5	96.86 ± 0.05	77.47 ± 0.36	87.17 ± 0.08	97.70 ± 0.03	85.02 ± 0.18	86.40 ± 0.07	69.54 ± 0.12	8.05 ± 0.18
10	97.17 ± 0.03	80.04 ± 0.18	86.53 ± 0.07	97.81 ± 0.02	85.80 ± 0.13	86.08 ± 0.06	69.84 ± 0.12	8.17 ± 0.17
20	97.37 ± 0.02	81.65 ± 0.09	85.99 ± 0.04	97.89 ± 0.01	86.32 ± 0.07	85.87 ± 0.04	70.06 ± 0.11	8.15 ± 0.10
100	97.67 ± 0.02	83.99 ± 0.11	85.14 ± 0.04	97.96 ± 0.01	86.85 ± 0.06	85.77 ± 0.03	70.24 ± 0.04	8.17 ± 0.06

Figure 2 shows some qualitative results for different dropout rates and with $N = 20$ on an example image of a complex scene in the T-LESS test dataset. Next to the predicted segmentation and uncertainty maps, the accuracy and the binary uncertainty maps are shown. For the binary uncertainty maps, we applied the same mean uncertainty threshold mentioned in Section 3 that is used for the estimation of the UQ metrics. Overall, it shows that accurate pixel predictions correspond to low uncertainty patches and vice versa. Increasing dropout rates lead to higher uncertainty values, which can be seen in the binary uncertainty maps. In case of the 10 % dropout model, the falsely segmented object in the lower right part of the image and background pixels exhibit high uncertainties.



Dropout 10 % Dropout 20 % Dropout 30 % Dropout 50 %

Predictions.

Binary accuracy maps.

Standard deviation maps.

Binary uncertainty maps based on standard deviation.

Entropy maps.

Binary uncertainty maps based on entropy.

Figure 2: Comparison of uncertainty maps for the image from a complex scene (Scene 17, image 50) across different dropout rates. Predictions and uncertainties are generated with 20 samples. In uncertainty maps, brighter pixels represent higher uncertainty. In accuracy/uncertainty binary maps, white pixels represent accurate/uncertain pixels.

5 Discussion

Our experiments demonstrate that increasing the sample size generally improves the segmentation accuracy and calibration in terms of mean IoU and ECE, while also enhancing the reliability of uncertainty estimation, as indicated by higher p_{ac} and p_{ui} scores. However, PAvPU decreases with larger sample sizes due to an increase in accurately classified but uncertain pixels, n_{au}, suggesting a more cautious model that flags more pixels as uncertain. It has to be noted that the UQ metrics depend on the chosen uncertainty threshold used to generate the underlying confusion matrix as described in Section 3 and may therefore vary with different thresholds.

Lower dropout rates result in better segmentation accuracy and model calibration, with the best performance observed when dropout is deactivated. However, a 30 % dropout rate optimizes p_{ui}, which is critical for detecting potentially incorrect predictions while reducing the calibration and segmentation quality only by 1.19 % ECE and 4.30 % mean IoU on average compared to the baseline model of our first goal. Thus, a 30 % dropout rate balances accurate segmentation and effective uncertainty estimation, making it optimal for practical applications.

Entropy is identified as a more suitable uncertainty metric than standard deviation, as it provides higher p_{ui}, indicating a better capacity to flag incorrect predictions. Although entropy-based metrics slightly reduce PAvPU, the trade-off is justified by a significant improvement in detecting uncertain inaccuracies.

Overall, the results suggest that using 20 samples, a 30 % dropout rate, and entropy as the uncertainty metric provides an optimal configuration for balancing segmentation accuracy, calibration, and uncertainty quantification quality in the SegFormer model with MCD.

6 Conclusion

In this contribution, we successfully trained SegFormer, a ViT variant, on the T-LESS dataset for the task of semantic segmentation with UQ in an industrial application. In combination with MCD, SegFormer is able to effectively handle challenging objects in varying complex scenes while producing meaningful uncertainty estimates. In future work, we want to extend the methodology for instance segmentation, which allows the integration of an ViT model in a deep-learning-based 6D object pose estimation pipeline. In the evaluation, we want to include additional UQ metrics like UCS [29,30]. While MCD is easy to implement, it does not capture the full uncertainty in the predictions [23]. Therefore, in future work, we aim to combine SegFormer with other state-of-the-art UQ methods like the recently proposed Deep Deterministic Uncertainty (DDU) [31] approach to produce robust uncertainty estimates even under data shift.

Acknowledgments

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

References

1. B. Ghoshal, A. Tucker, B. Sanghera, and W. Lup Wong, “Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection,” *Computational Intelligence*, vol. 37, no. 2, pp. 701–734, 2021.
2. M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier, “Well-calibrated regression uncertainty in medical imaging with deep learning,” in *MIDL*, vol. 121, 2020, pp. 393–412.
3. D. Feng, L. Rosenbaum, and K. Dietmayer, “Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection,” in *ITSC*, 2018, pp. 3266–3273.
4. S. Shafaei, S. Kugele, M. H. Osman, and A. Knoll, “Uncertainty in machine learning: A safety perspective on autonomous driving,” in *SAFECOMP 2018 Workshops*, 2018, pp. 458–464.
5. D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” *arXiv e-prints*, vol. arXiv:1610.02136, 2016.
6. Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, “Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift,” in *NeurIPS*, vol. 32, 2019.

7. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *ICML*, vol. 48, 2016, pp. 1050–1059.
8. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *NeurIPS*, vol. 34, 2021, pp. 12 077–12 090.
9. T. Hodaň, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects," in *IEEE WACV*, 2017, pp. 880–888.
10. T. Hodaň, Y. Labbé, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, C. Sahin, F. Manhardt, F. Tombari, T.-K. Kim, J. Matas, and C. Rother, "BOP: Benchmark for 6D Object Pose Estimation," in *ECCV*, 2018.
11. M. P. Naeini, G. Cooper, and M. Hausknecht, "Obtaining well calibrated probabilities using bayesian binning," in *AAAI*, vol. 29, 2015, pp. 2901–2907.
12. J. Mukhoti and Y. Gal, "Evaluating bayesian deep learning methods for semantic segmentation," *arXiv e-prints*, vol. arXiv:1811.12709, 2018.
13. R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *IEEE/CVF ICCV*, 2021, pp. 7262–7272.
14. S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *IEEE/CVF CVPR*, 2021, pp. 6881–6890.
15. B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *NeurIPS*, vol. 34, pp. 17 864–17 875, 2021.
16. B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *IEEE/CVF CVPR*, 2022, pp. 1290–1299.
17. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF IICCV*, 2021, pp. 10 012–10 022.
18. R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *IEEE/CVF ICCV*, 2021, pp. 12 179–12 188.
19. Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution transformer for dense prediction," *arXiv e-prints*, vol. arXiv:2110.09408, 2021.

20. C. Dechesne, P. Lassalle, and S. Lefèvre, "Bayesian deep learning with monte carlo dropout for qualification of semantic segmentation," in *IEEE IGARSS*, 2021, pp. 2536–2539.
21. M. Abdar, S. Salari, S. Qahremani, H.-K. Lam, F. Karray, S. Hussain, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "Uncertaintyfusenet: robust uncertainty-aware hierarchical feature fusion model with ensemble monte carlo dropout for covid-19 detection," *Information Fusion*, vol. 90, pp. 364–381, 2023.
22. A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv e-prints*, vol. arXiv:1511.02680, 2015.
23. A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *NeurIPS 2017*, vol. 30, 2017.
24. S. Landgraf, M. Hillemann, K. Wursthorn, and M. Ulrich, "Uncertainty-aware cross-entropy for semantic segmentation," in *ISPRS Annals*, vol. X-2-2024, 2024, pp. 129–136.
25. 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.
26. B. Chen, W. Peng, X. Cao, and J. Röning, "Hyperbolic uncertainty aware semantic segmentation," *IEEE T-ITS*, vol. 25, no. 2, pp. 1275–1290, 2024.
27. S. Landgraf, M. Hillemann, T. Kapler, and M. Ulrich, "Efficient multi-task uncertainties for joint semantic segmentation and monocular depth estimation," in *GCPR*, 2024.
28. Z. Zhao, L. Yang, S. Long, J. Pi, L. Zhou, and J. Wang, "Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation," in *IEEE/CVF CVPR*, 2023, pp. 11 350–11 359.
29. K. Wursthorn, M. Hillemann, and M. Ulrich, "Uncertainty quantification with deep ensembles for 6d object pose estimation," in *ISPRS Annals*, vol. X-2-2024, 2024, pp. 223–230.
30. D. W. Wolf, P. Balaji, A. Braun, and M. Ulrich, "Decoupling of neural network calibration measures," in *GCPR*, 2024.
31. J. Mukhoti, A. Kirsch, J. van Amersfoort, P. H. Torr, and Y. Gal, "Deep Deterministic Uncertainty: A New Simple Baseline," in *IEEE/CVF CVPR*, 2023, pp. 24 384–24 394.