A Critical Synthesis of Uncertainty Quantification and Foundation Models for Semantic Segmentation

Steven Landgraf *, Joceline Hinz*, Markus Ulrich

Institute of Photogrammetry and Remote Sensing (IPF),
Karlsruhe Institute of Technology (KIT), Germany,
(steven.landgraf, joceline.hinz, markus.ulrich)@kit.edu

Keywords: Uncertainty Quantification, Reliability, Robustness, Foundation Models, Semantic Segmentation

Abstract

Foundation models are increasingly breaking what seemed to be impossible not long ago by enabling unprecedented accuracy and cross-domain generalization. Yet their lack of interpretability, tendency to be overconfident, and sensitivity to real-world domain shifts pose critical challenges for safety- and mission-critical applications. Uncertainty quantification (UQ) offers a principled way to address these issues, but its integration into segmentation foundation models has yet to be explored. In this paper we present the first systematic evaluation of UQ methods applied to a foundation model for semantic segmentation. We fine-tune a lightweight DPT decoder on top of the pretrained SAM2 encoder to establish a simple yet competitive baseline and benchmark four representative UQ approaches – Monte Carlo Dropout, Deep Sub-Ensemble, Test-Time Augmentation, and Evidential Deep Learning – across Cityscapes, NYUv2, and two challenging out-of-domain settings. Our analysis compares segmentation accuracy, calibration, uncertainty quality, and inference time, revealing clear trade-offs between predictive performance, reliability, and computational cost. These results highlight both the promise and the current limitations of uncertainty-aware foundation models, pointing to the need for future work that jointly optimizes accuracy, robustness, and efficiency for real-world deployment.

1. Introduction

Semantic segmentation is a foundational machine vision task involving assigning class labels to every pixel in an image (Mo et al., 2022). Recently, segmentation models have evolved rapidly from convolutional neural networks (Li et al., 2021) and vision transformers (Han et al., 2022) to foundation models trained on internet-scale data (Zhou et al., 2024). These models promise unprecedented accuracy and generalization, enabling zero-shot capabilities across diverse domains and marking a paradigm shift in semantic segmentation.

However, even the most potent neural networks remain subject to critical limitations (Awais et al., 2025). Foundation models in particular can act as black boxes, lacking interpretability (Gawlikowski et al., 2023), failing to recognize out-of-domain inputs (Ovadia et al., 2019), producing systematically overconfident outputs (Wilson and Izmailov, 2020), or being sensitive to adversarial perturbations (Rawat et al., 2017). It goes without saying that these issues are particularly harmful in safety- and mission-critical applications, where segmentation errors can propagate to downstream decisions with severe consequences.

Incorporating uncertainty quantification (UQ) into deep neural networks directly addresses many of the challenges outlined above. Reliable uncertainty estimates not only mitigate overconfidence and sensitivity to domain shifts but also enhance interpretability by indicating the model's confidence and highlighting areas of potential error, thereby improving the deployability deep learning-based systems. For instance, in safety-critical applications like autonomous driving (McAllister et al., 2017) or medical imaging (Nair et al., 2020), uncertainty-aware models can flag unreliable predictions to trigger human intervention or additional verification steps, reducing the risk of catastrophic failures.

Despite these advantages of UQ, the integration of UQ into foundation models for semantic segmentation remains underexplored. By systematically evaluating how different UQ methods interact with these powerful models, we offer not only improved reliability and explainability but also bridge the gap between state-of-the-art research and real-world deployment. Our contributions can be summarized as follows:

1. We present the first systematic study of UQ methods applied to foundation models for semantic segmentation, evaluating Monte Carlo Dropout (MCD), Deep Sub-Ensemble (DSE), Test-Time Augmentation (TTA), and Evidential Deep Learning (EDL).
2. We establish a simple yet competitive baseline by fine-tuning a DPT head on top of the pretrained SAM2 encoder, and benchmark its in-domain and out-of-domain performance across multiple datasets.
3. We provide a comprehensive comparison between segmentation performance, calibration, uncertainty quality, and inference time for all methods, highlighting trade-offs between accuracy, reliability, and computational cost.

2. Related Work

Our work lies at the intersection of semantic segmentation and UQ, focusing on adapting foundation models like the Segment Anything Model (SAM) (Kirillov et al., 2023). We first review advancements in segmentation, from CNNs to vision transformers and foundation models. We then discuss key deep learning UQ methods, including recent efforts to tailor them for foundation models. Finally, we highlight the research gap addressed by this paper.

* Corresponding Authors

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-673-2026

2.1 Semantic Segmentation

Convolutional Neural Networks (CNNs). The foundational work in deep learning-based segmentation began with fully convolutional networks, which enable end-to-end pixel-wise classification (Long et al., 2015). Subsequent CNN methods introduced richer context and multi-scale features. For instance, DeepLab established dilated convolutions and atrous spatial pyramid pooling to capture image context at multiple scales (Chen et al., 2017). Similarly, encoder-decoder networks like U-Net used symmetric contracting and expanding paths with skip connections to recover even fine details (Ronneberger et al., 2015). Together, these CNN-based approaches represented the state of the art for many years and have only recently been challenged by attention-based architectures.

Vision Transformers (ViTs). Transformer architecture have been the de-facto standard for natural language processing for a long time (Vaswani et al., 2017). More recently, they have also been brought into the vision domain (Dosovitskiy et al., 2020). In the context of semantic segmentation, SegFormer stands out with its hierarchical transformer encoder and a lightweight multilayer perceptron decoder, achieving impressive results and high efficiency (Xie et al., 2021). Mask2Former further unified semantic, instance and panoptic segmentation with a mask-classification transformer, setting new state-of-the-art results (Cheng et al., 2022). Beyond these task-specific advances, the field has shifted toward foundation models trained on internet-scale data. For example, SAM was trained with over 1B masks on 11M images (Kirillov et al., 2023). As noted in a recent study (Zhou et al., 2024), adapting these foundation models can yield superior segmentation performance and entirely new capabilities in terms of zero/few-shot segmentation, interactive prompting, or cross-domain generalization that were unseen until now.

2.2 Uncertainty Quantification

Overview. A variety of techniques have been proposed to capture predictive uncertainty in deep learning models (MacKay, 1992; Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017; Valdenegro-Toro, 2023; Van Amersfoort et al., 2020; Liu et al., 2020; Mukhoti et al., 2023; Amini et al., 2020). The most popular methods remain sampling-based due to their ease of use and effectiveness: for instance, MCD (Gal and Ghahramani, 2016) treats dropout (Srivastava et al., 2014) as a Bayesian approximation. Similarly, Deep Ensembles (Lakshminarayanan et al., 2017) consist of multiple, individually trained models to obtain state-of-the-art uncertainty results. Unfortunately, these approaches require multiple forward passes and sometimes more training time, which incurs high computational cost. To mitigate this, a number of deterministic, more efficient, methods have been proposed (Van Amersfoort et al., 2020; Liu et al., 2020; Mukhoti et al., 2023; Landgraf et al., 2024, 2025b). Likewise, EDL enables efficient uncertainty estimation by training the model to infer the parameters of a probability distribution with a single prediction (Amini et al., 2020).

Uncertainty-aware Foundation Models. Being able to estimate the uncertainty of the output of large foundation models is an emerging research field. Recently, Landgraf et al. (2025c) studied the synthesis of UQ methods and foundation models in monocular depth estimation and explicitly note that extending this work to other tasks, such as semantic segmentation, is an open opportunity. A few early methods have begun to address UQ for SAM in medical imaging (Jiang et al., 2024; Deng et al., 2023; Zhang et al., 2023). Beyond, USAM proposes a post-hoc training method for additional multilayer perceptrons to estimate the expected uncertainty (Kaiser et al., 2025). Liu et al. (2024) introduce SUM that quantifies uncertainty in SAM-generated pseudo-labels and uses it to enable uncertainty-aware fine-tuning of the model.

Research Gap. While all of these works have made progress toward uncertainty-aware foundation models in semantic segmentation with tailor-made approaches, they also reveal a critical gap: There has yet to be a systematic study of existing UQ methods in combination with SAM, despite its status as a de-facto foundation model for large-scale segmentation.

3. Methodology

In this section, we describe how we derive our baseline model from the Segment Anything Model 2 (SAM2) foundation model and how we combine four UQ techniques with it.

3.1 Baseline Model

We employ a hybrid architecture that combines the image encoder of the SAM2 (Ravi et al., 2024) with the DPT decoder (Ranftl et al., 2021) for semantic segmentation (see Fig. 1). The motivation for this design is twofold: First, by leveraging the SAM2 encoder, which builds on the hierarchical Vision Transformer Hiera (Ryali et al., 2023) pretrained with masked autoencoding (He et al., 2022), we directly harness state-of-the-art foundation model knowledge in the form of robust, multi-scale image representations. Second, coupling this encoder with the DPT decoder allows us to rely on a simple and widely adopted architecture for dense prediction, ensuring interpretability, comparability, and computational efficiency. The SAM2 encoder partitions the image into patches, projects them into tokens with positional embeddings, and processes them through four hierarchical transformer stages, producing multi-scale feature maps. These are fused by the DPT decoder through RefineNet-based (Lin et al., 2017) fusion modules that progressively upsample and align resolution, before passing them to a lightweight segmentation head. The head simply applies a 3x3-convolution, injects a dropout layer, followed by a linear projection, and lastly uses bilinear upsampling to restore the original image resolution. Training the model is equally simple as its architectural layout: we simply fine-tune the model with the regular cross-entropy loss

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

where $\mathcal{L}_{\text{CE}}$ represents the loss for a single image, N is the number of pixels, C is the number of classes, $y_{n,c}$ is the one-hot encoded ground truth label, and $p(\hat{y}_{n,c})$ is the predicted softmax pseudo-probability. Overall, our intent was to maximize the benefits of foundation model pretraining while keeping the rest of the architecture as simple as possible.

3.2 Uncertainty Quantification

Our central research question is how to make segmentation foundation models not only accurate but also reliable in real-world deployment where well-calibrated uncertainties are essential. Following (Gawlikowski et al., 2023), existing UQ approaches can be broadly grouped into test-time augmentation,

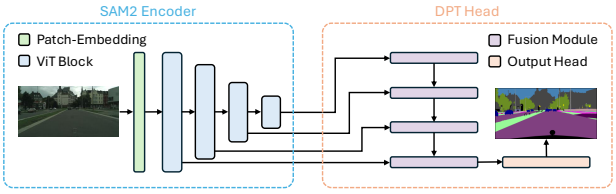

Figure 1. Schematic illustration of the used model architecture.

Bayesian methods, ensembles, and deterministic methods. To cover this wide methodological spectrum, we evaluate one representative technique for each group: MCD, DSE, EDL, and TTA. While Fig. 2 provides a schematic overview of how we fused our model architecture with all four UQ approaches, the following will provide a more detailed description of each.

Monte Carlo Dropout. MCD estimates predictive uncertainty through stochastic sampling with dropout layers. Following Gal and Ghahramani (2016), dropout can be interpreted as a variational approximation to Bayesian neural networks, where the posterior distribution over weights is intractable and thus approximated by a Bernoulli distribution. By keeping dropout active during inference, each forward pass yields a slightly different softmax output $p(\hat{y}_t)$.

To compute the predictive mean $\hat{\mu}$, we average across all T stochastic samples:

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} p(\hat{y}_t) \ . \tag{2}$$

The corresponding uncertainty can be quantified by the variance as

$$\hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^{T}(p(\hat{y}_t) - \hat{\mu})^2 \ . \tag{3}$$

Besides the standard deviation $\hat{\sigma}$, the predictive entropy can also be computed as a complete measure of the predictive uncertainty, which is composed of a combination of the aleatoric and epistemic uncertainty (Wolf et al., 2025):

$$H(\hat{\mu}) = -\sum_{c=1}^{C} \hat{\mu}_c \log(\hat{\mu}_c) \ . \tag{4}$$

Deep Sub-Ensemble. Deep Ensembles (Lakshminarayanan et al., 2017) are considered a gold standard for UQ (Ovadia et al., 2019; Gustafsson et al., 2020; Landgraf et al., 2025a), as they capture variability across independently trained models. However, their computational cost scales linearly with the number of ensemble members, limiting their practicality in large-scale vision tasks. DSEs approximate the benefits of DEs at reduced cost by sharing most of the architecture and varying only a subset of layers close to the output head (Valdenegro-Toro, 2023).

In our setup, the SAM2 encoder is shared across all ensemble members, while multiple DPT decoders are initialized independently. During training, only one decoder head is optimized per mini-batch, while the others are frozen. This cycling strategy increases diversity across decoders while keeping training efficient. At inference, each decoder in general produces a different softmax prediction $p(\hat{y}_t)$ based on the same encoder features. As in MCD, the predictive mean and predictive uncertainty are computed across all T decoder heads (see Eqs. 2 - 4).

Evidential Deep Learning. EDL directly models predictive uncertainty by parameterizing a Dirichlet distribution over class probabilities (Amini et al., 2020). Instead of outputting a single categorical distribution via softmax, the network produces non-negative evidence values e_c for each class c. These are mapped to the Dirichlet parameters

$$\alpha_c = e_c + 1 \ . \tag{5}$$

The Dirichlet strength is then defined as

$$S = \sum_{c=1}^{C} \alpha_c \ , \tag{6}$$

where C is the number of classes. From here, the belief masses b_c and the total uncertainty u can be obtained as

$$b_c = \frac{e_c}{S} \ , \quad u = \frac{C}{S} \ . \tag{7}$$

The expected class probabilities follow as

$$p(\hat{y}_c) = \frac{\alpha_c}{S} \ . \tag{8}$$

This representation enables the model to express both confident assignments – with large α_c values for one class – and uncertain states – with low, evenly distributed α_c values across classes.

Training is performed with an evidential loss function based on the mean squared error, which, for a single sample i, can be calculated as

$$\mathcal{L}_i(\Theta) = \sum_{c=1}^{C} (y_{ik} - p(\hat{y}_{ic}))^2 + \frac{p(\hat{y}_{ic})(1-p(\hat{y}_{ic})}{S_i + 1} \ , \tag{9}$$

where the first term penalizes prediction errors and the second term incorporates the variance of the predictive distribution, encouraging the model to calibrate its uncertainty appropriately. To prevent overconfidence on misclassified samples, a Kullback-Leibler (KL) divergence regularization is added to obtain the final loss function

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \mathcal{L}_i(\Theta) + \lambda_t \sum_{i=1}^{N} \mathrm{KL}[\mathrm{D}(p(\hat{y}_i)|\tilde{a}_i)||\mathrm{D}(p(\hat{y}_i)|1)] \ , \tag{10}$$

where λ_t is an annealing coefficient, t is the current training epoch, $\mathrm{D}(p(\hat{y}_i)|1)$ is the uniform Dirichlet distribution, and lastly $\tilde{\alpha}_i = y_i + (1 - y_i) \odot \alpha_i$ is the Dirichlet parameters after removal of the non-misleading evidence from predicted parameters α_i. More details on this can be found in Sensoy et al. (2018).

To stabilize training with our SAM2–DPT hybrid, we employ a two-head architecture, as shown by Fig. 2: One head predicts segmentation logits, while the second head outputs evidential uncertainty parameters. Empirically, this design significantly improved segmentation performance compared to a single shared head, where segmentation accuracy degraded drastically.

Test-Time Augmentation. TTA is applied only during inference to estimate the predictive uncertainty. Consequently, it is applied without modifying the model parameters or the training

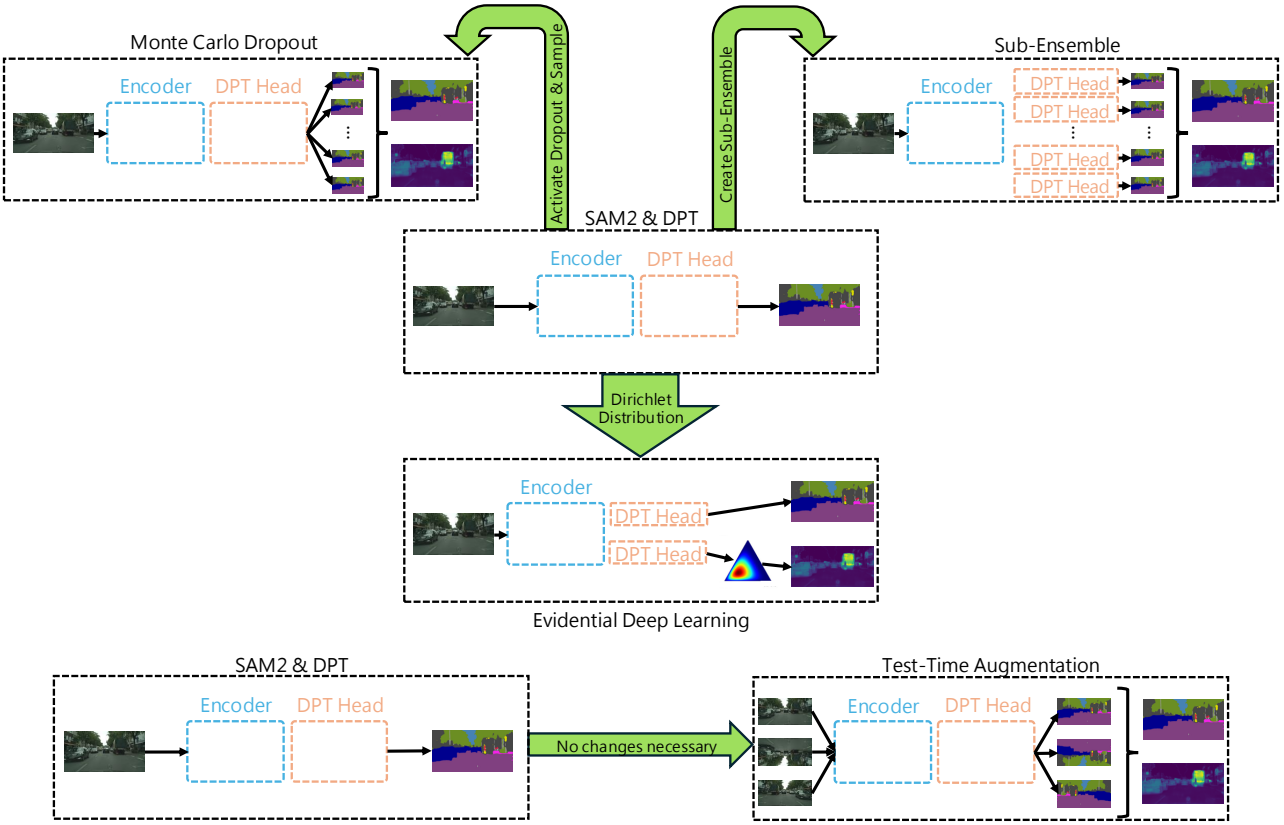

Figure 2. A schematic overview of how we combined the four different uncertainty quantification techniques with our model.

protocol. During inference, we generate T augmented variants of each input image and feed both the original and its augmented versions into the network for prediction. Akin to MCD and DSE, the final prediction is obtained by the mean and the corresponding uncertainty is quantified by the variance/standard deviation or predictive entropy (cf. Eqs. 2–4).

4. Experimental Setup

This section details the unified training configuration, datasets, and augmentations used in all experiments to ensure fair comparisons across UQ methods and robust evaluation under diverse conditions.

4.1 Training Configuration

All models are trained for 100 epochs, which was empirically found to be sufficient for convergence. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a base learning rate of 3×10^{-5} for both datasets and a polynomial decay scheduler. As the SAM2 encoder is pre-trained while the decoder is newly initialized, we apply a learning-rate multiplier of 10 to the decoder. The loss function is the standard cross-entropy loss (cf. Eq. 1). Due to GPU memory constraints, we set the batch size to 8 for all experiments. The initial weight decay is set to 1×10^{-2} for Cityscapes and 1×10^{-4} for NYUv2. All experiments are conducted on a single NVIDIA A100 GPU with 40 GB VRAM.

4.2 Datasets

We evaluate on Cityscapes (Cordts et al., 2016) and NYUv2 (Silberman et al., 2012). Cityscapes contains high-resolution street scenes (2048×1024 px). Following standard practice, we use the 2,875 training and 500 validation images with 19 semantic classes and 1 void label, which is ignored during training. Additionally, we employ the most challenging options of the synthetic Rainy-Cityscapes (Hu et al., 2019) and Foggy-Cityscapes (Sakaridis et al., 2018) variants for out-of-domain evaluation.

NYUv2 comprises 1,449 RGB-D indoor images captured with a Microsoft Kinect in 464 different rooms. We use the RGB images and the corresponding 40 semantic classes (grouped into 13 super-classes for visualization) split into 795 training and 654 test images at 640×480 px resolution.

The four datasets jointly allow testing under both outdoor and indoor conditions as well as in-domain and out-of-domain settings.

4.3 Data Augmentations

Regardless of the model, we apply random scaling with a factor between 0.5 and 2.0, random cropping with a crop size of 768×768 px on Cityscapes and 480×640 px on NYUv2, and random horizontal flipping with a flip chance of 50 % during training as data augmentations.

4.4 Uncertainty Quantification

Based on rigorous evaluations, we report results based on the following UQ configurations. For TTA, we apply vertical and horizontal flipping as well as scaling during inference to generate samples to compute our predictive mean and uncertainty. We note that vertical flipping can alter spatial priors and may therefore lead to an overestimation of uncertainty in some cases – we nevertheless found the inclusion of vertical flipping to

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-673-2026 | © Author(s) 2026. CC BY 4.0 License.

Cityscapes	mIoU ↑
DeepLabV3+ (Chen et al., 2018)	79.6%
U-Net++ (Ronneberger et al., 2015)	75.5%
SegFormer-B0 (Xie et al., 2021)	76.2%
SegFormer-B5 (Xie et al., 2021)	82.4%
Mask2Former (Cheng et al., 2022)	83.3%
SAM2-DPT [*tiny*] (ours)	76.6%
SAM2-DPT [*large*] (ours)	82.4%

Table 1. Comparison against previous approaches on Cityscapes.

NYUv2	mIoU ↑
TokenFusion (Wang et al., 2022)	54.2%
Omnivore (Girdhar et al., 2022)	54.0%
EMSANet (Seichter et al., 2022)	53.3%
SGNet (Chen et al., 2021)	51.0%
AsymFormer (Du et al., 2024)	54.1%
SAM2-DPT [*tiny*] (ours)	43.6%
SAM2-DPT [*large*] (ours)	55.5%

Table 2. Comparison against previous approaches on NYUv2.

work best. MCD uses the already existing dropout layers in the model architecture with a dropout rate of 20% and we sample ten times during test-time. Similarly, we employ ten DPT heads for DSE.

As the predictive entropy delivered slightly better results in our evaluations than the predictive variance, we only report uncertainty metrics based on the former for all models, including the baseline.

4.5 Metrics

In terms of metrics, we report the mean Intersection over Union (mIoU), the Expected Calibration Error (ECE) (Naeini et al., 2015; Wolf et al., 2025), and two uncertainty quality metrics from Mukhoti and Gal (2018):

1. **p(acc.|cer.):** The probability that the model is accurate on its output given that the uncertainty is below a specific threshold.
2. **p(unc.|ina.):** The probability that the uncertainty of the model exceeds a specified threshold given that the prediction is inaccurate.

Both metrics are expected to return high values for high-quality uncertainties. Based on our own empiric results and prior work from Landgraf et al. (2025a), we opt for the median uncertainty of any given image as the uncertainty threshold.

5. Experiments

The following section lays out numerous quantitative as well as qualitative results, encompassing not only in-domain datasets but also out-of-domain settings for a comprehensive set of experiments.

5.1 Quantitative Results

Comparison with SOTA. Tables 1 and 2 report segmentation performance on Cityscapes and NYUv2, respectively. Without additional architectural changes or complex training strategies, our simple SAM2-DPT baseline achieves results comparable to recent state-of-the-art methods, demonstrating that fine-tuning a foundation model with a lightweight DPT head can already yield competitive performance across both datasets. For all subsequent experiments, we report results using the SAM2-Tiny backbone due to computational constraints; however, we empirically verified that the observed trends are consistent with larger backbone variants.

In-Domain Evaluation. Table 3 shows in-domain results on Cityscapes and NYUv2. On Cityscapes, the baseline model achieves competitive mIoU, strong calibration, and offers the fastest inference time. At the same time, it offers the second worst results in terms of uncertainty quality – only undermined by EDL, which performs worst across all metrics. DSE attains the best calibration and uncertainty quality but at the cost of a slight mIoU reduction and substantially slower inference.

On NYUv2, the baseline achieves the highest segmentation performance and, unexpectedly, the best uncertainty quality, but shows the poorest calibration – though these results should be interpreted cautiously given the overall low segmentation scores. EDL again provides the second-fastest inference time but performs poorly across all other metrics. TTA delivers the best calibration, while DSE follows the baseline closely in uncertainty quality.

Out-of-Domain Evaluation. Table 4 shows out-of-domain results on Rainy- and Foggy-Cityscapes. On both datasets, segmentation performance drops markedly compared to the in-domain experiments, highlighting the vulnerability of deep learning models to domain shifts.

On Rainy-Cityscapes, TTA achieves the highest mIoU, the best calibration, and also improves the uncertainty quality compared to the baseline. MCD yields slightly higher mIoU than TTA but a substantially worse calibration. DSE performs comparably to the baseline with worse calibration but higher p(unc.|ina.). EDL again shows the weakest results across all metrics.

For Foggy-Cityscapes, TTA gives the best mIoU at the cost of notably worse calibration and uncertainty quality, whereas DSE attains the best uncertainty scores, especially p(unc.|ina.). MCD remains close to the baseline with moderate improvements in mIoU but worse calibration. EDL again performs worst on all metrics.

5.2 Qualitative Results

Fig. 3 shows in-domain qualitative results on Cityscapes. The baseline, MCD, and DSE produce the most accurate segmentation predictions in terms of mIoU, whereas TTA performs slightly worse and EDL exhibits the worst results. The corresponding binary accuracy maps reveal largely similar falsely segmented regions across all methods. Regarding the uncertainty, TTA and EDL show a higher and spatially more widespread uncertainty, while the baseline, MCD and DSE display more localized and lower levels of uncertainty. Notably, DSE captures the erroneous regions most effectively, as its uncertainties align well with the binary error maps, indicating a strong correspondence between prediction errors and uncertainty estimates.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-673-2026 | © Author(s) 2026. CC BY 4.0 License.

in-Domain	mIoU [%] ↑	ECE [%] ↓	p(acc.|cer.) [%] ↑	p(unc.|ina.) [%] ↑	Inference Time [ms] ↓
Cityscapes					
Baseline	76.55	0.95	92.61	76.18	**58.57**
TTA	**76.84**	1.76	94.29	81.32	202.50
MCD	76.10	1.30	93.52	80.18	656.07
DSE	74.44	**0.65**	**96.91**	**91.74**	315.46
EDL	70.84	4.97	84.40	51.14	73.40
NYUv2					
Baseline	**43.62**	16.80	**82.08**	**81.26**	**9.49**
TTA	43.61	**3.01**	79.53	77.09	34.52
MCD	38.15	13.15	78.12	77.26	124.75
DSE	39.10	11.21	80.36	80.00	55.40
EDL	37.94	7.13	76.01	74.50	11.15

Table 3. Quantitative in-domain evaluation on Cityscapes and NYUv2 datasets using SAM2-DPT.

Out-of-Domain	mIoU [%] ↑	ECE [%] ↓	p(acc.|cer.) [%] ↑	p(unc.|ina.) [%] ↑
Rainy-Cityscapes				
Baseline	46.72	5.09	92.66	80.92
TTA	48.99	**0.84**	**93.40**	**83.85**
MCD	**49.37**	5.33	92.29	78.28
DSE	45.71	7.07	92.65	83.21
EDL	40.91	12.24	89.68	75.12
Foggy-Cityscapes				
Baseline	53.96	**3.09**	89.54	80.11
TTA	**56.53**	7.46	88.48	77.43
MCD	54.49	5.67	90.39	81.44
DSE	51.14	4.82	**93.40**	**89.49**
EDL	50.97	14.26	83.64	67.09

Table 4. Quantitative out-of-domain evaluation on Rainy- and Foggy-Cityscapes using SAM2-DPT.

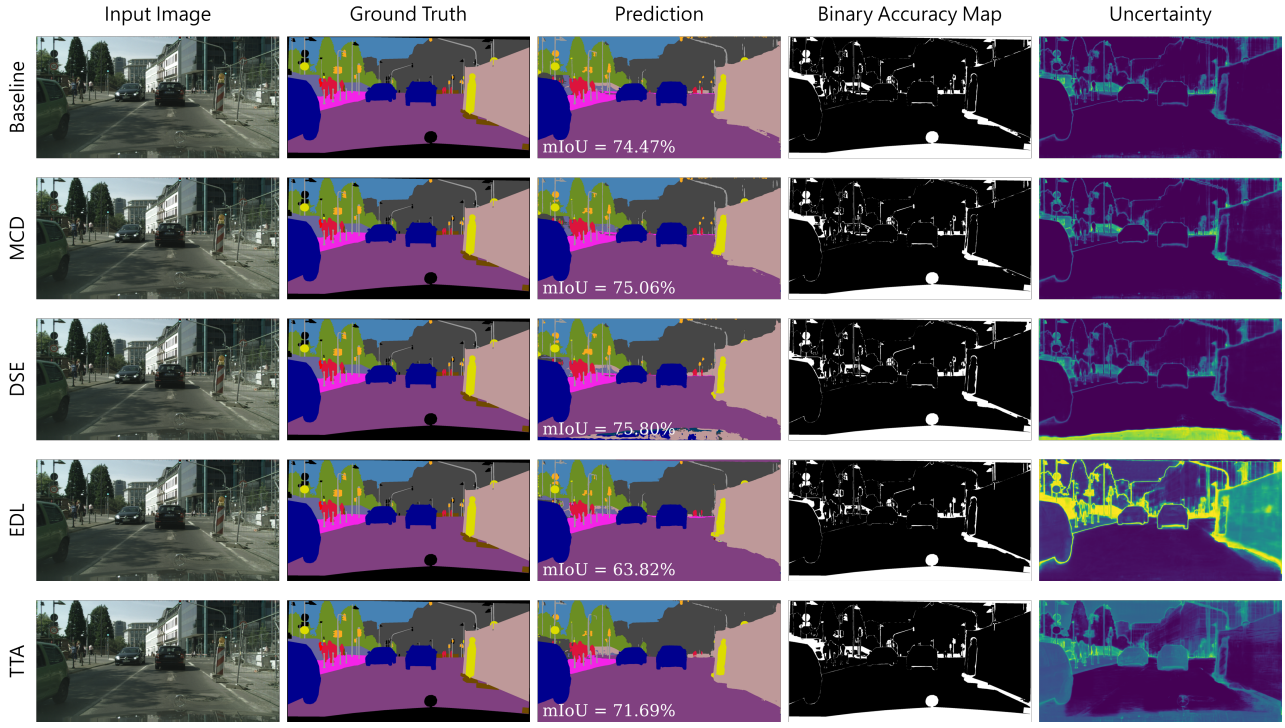

Figure 3. Qualitative in-domain evaluation on Cityscapes dataset using SAM2-DPT.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-673-2026 | © Author(s) 2026. CC BY 4.0 License.

5.3 Discussion

Our results reveal several important patterns regarding the critical synthesis of UQ and foundation models for segmentation. Although the SAM2–DPT baseline achieves competitive segmentation accuracy with minimal architectural changes, its calibration and uncertainty quality are insufficient and inconsistent across domains. This confirms that high predictive performance does not automatically translate to high reliability.

Additionally, we found clear trade-offs w.r.t. the UQ strategies. While DSE offers high uncertainty quality, it comes at a significant cost of inference speed. TTA improves upon the baseline in most cases but also suffers from high computation cost during inference. MCD, despite the highest inference times, often underperforms even the baseline, and EDL yields the least promising results. These observations highlight that there is no single best method and that the choice of UQ strategy highly depends on the deployment context.

6. Conclusion

We presented the first systematic evaluation of multiple UQ methods applied to a SAM2-based foundation model for semantic segmentation. Our experiments across Cityscapes, NYUv2 and two out-of-domain settings show that competitive segmentation accuracy is easily attainable with a lightweight decoder, but reliability and efficiency differ markedly between UQ strategies. It is clear that there is no one-size-fits-all solution yet. This underscores the importance of future work not only focusing on predictive performance but also reliability and efficiency.

References

Amini, A., Schwarting, W., Soleimany, A., Rus, D., 2020. Deep evidential regression. *Advances in neural information processing systems*, 33, 14927–14937.

Awais, M., Naseer, M., Khan, S., Anwer, R. M., Cholakkal, H., Shah, M., Yang, M.-H., Khan, F. S., 2025. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834–848.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, L.-Z., Lin, Z., Wang, Z., Yang, Y.-L., Cheng, M.-M., 2021. Spatial information guided convolution for real-time RGBD semantic segmentation. *IEEE Transactions on Image Processing*, 30, 2313–2324.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Deng, G., Zou, K., Ren, K., Wang, M., Yuan, X., Ying, S., Fu, H., 2023. Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 368–377.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, S., Wang, W., Guo, R., Wang, R., Tang, S., 2024. Asymformer: Asymmetrical cross-modal representation learning for mobile platform real-time rgb-d semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7608–7615.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *international conference on machine learning*, PMLR, 1050–1059.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X. X., 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1), 1513–1589.

Girdhar, R., Singh, M., Ravi, N., Van Der Maaten, L., Joulin, A., Misra, I., 2022. Omnivore: A single model for many visual modalities. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16102–16112.

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

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y. et al., 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1), 87–110.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hu, X., Fu, C.-W., Zhu, L., Heng, P.-A., 2019. Depth-attentional features for single-image rain removal. *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 8022–8031.

Jiang, M., Zhou, J., Wu, J., Wang, T., Jin, Y., Xu, M., 2024. Uncertainty-Aware Adapter: Adapting Segment Anything Model (SAM) for Ambiguous Medical Image Segmentation. *arXiv preprint arXiv:2403.10931*.

Kaiser, T., Norrenbrock, T., Rosenhahn, B., 2025. Uncertainsam: Fast and efficient uncertainty quantification of the segment anything model. *arXiv preprint arXiv:2505.05049*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. et al., 2023. Segment anything. *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Landgraf, S., Hillemann, M., Kapler, T., Ulrich, M., 2025a. A comparative study on multi-task uncertainty quantification in semantic segmentation and monocular depth estimation. *tm-Technisches Messen*.

Landgraf, S., Hillemann, M., Kapler, T., Ulrich, M., 2025b. Efficient multi-task uncertainties for joint semantic segmentation and monocular depth estimation. D. Cremers, Z. Lähner, M. Moeller, M. Nießner, B. Ommer, R. Triebel (eds), *Pattern Recognition. DAGM GCPR 2024*, Lecture Notes in Computer Science, 157, Springer, 348–364.

Landgraf, S., Qin, R., Ulrich, M., 2025c. A critical synthesis of uncertainty quantification and foundation models in monocular depth estimation. *arXiv preprint arXiv:2501.08188*.

Landgraf, S., Wursthorn, K., Hillemann, M., Ulrich, M., 2024. Dudes: Deep uncertainty distillation using ensembles for semantic segmentation. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 92(2), 101–114.

Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J., 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12), 6999–7019.

Lin, G., Milan, A., Shen, C., Reid, I., 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1925–1934.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., Lakshminarayanan, B., 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33, 7498–7512.

Liu, K., Price, B., Kuen, J., Fan, Y., Wei, Z., Figueroa, L., Geras, K., Fernandez-Granda, C., 2024. Uncertainty-aware fine-tuning of segmentation foundation models. *Advances in Neural Information Processing Systems*, 37, 53317–53389.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization.

MacKay, D. J., 1992. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3), 448–472.

McAllister, R. T., Gal, Y., Kendall, A., Van Der Wilk, M., Shah, A., Cipolla, R., Weller, A., 2017. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. *International Joint Conferences on Artificial Intelligence, Inc.*

Mo, Y., Wu, Y., Yang, X., Liu, F., Liao, Y., 2022. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493, 626–646.

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

Mukhoti, J., Kirsch, A., Van Amersfoort, J., Torr, P. H., Gal, Y., 2023. Deep deterministic uncertainty: A new simple baseline. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24384–24394.

Naeini, M. P., Cooper, G., Hauskrecht, M., 2015. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI conference on artificial intelligence*, 29number 1.

Nair, T., Precup, D., Arnold, D. L., Arbel, T., 2020. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59, 101557.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J., 2019. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L. et al., 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Rawat, M., Wistuba, M., Nicolae, M.-I., 2017. Harnessing model uncertainty for detecting adversarial examples. *NIPS Workshop on Bayesian Deep Learning*.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, 234–241.

Ryali, C., Hu, Y.-T., Bolya, D., Wei, C., Fan, H., Huang, P.-Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J. et al., 2023. Hiera: A hierarchical vision transformer without the bells-and-whistles. *International conference on machine learning*, PMLR, 29441–29454.

Sakaridis, C., Dai, D., Van Gool, L., 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9), 973–992.

Seichter, D., Fischedick, S. B., Köhler, M., Groß, H.-M., 2022. Efficient multi-task rgb-d scene analysis for indoor environments. *2022 International joint conference on neural networks (IJCNN)*, IEEE, 1–10.

Sensoy, M., Kaplan, L., Kandemir, M., 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from rgbd images. *European conference on computer vision*, Springer, 746–760.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929–1958.

Valdenegro-Toro, M., 2023. Sub-ensembles for fast uncertainty estimation in neural networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4119–4127.

Van Amersfoort, J., Smith, L., Teh, Y. W., Gal, Y., 2020. Uncertainty estimation using a single deep deterministic neural network. *International conference on machine learning*, PMLR, 9690–9700.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y., Chen, X., Cao, L., Huang, W., Sun, F., Wang, Y., 2022. Multimodal token fusion for vision transformers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12186–12195.

Wilson, A. G., Izmailov, P., 2020. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in Neural Information Processing Systems*, 33, 4697–4708.

Wolf, D., Balaji, P., Braun, A., Ulrich, M., 2025. Decoupling of neural network calibration measures. D. Cremers, Z. Lähner, M. Moeller, M. Nießner, B. Ommer, R. Triebel (eds), *Pattern Recognition. DAGM GCPR 2024*, Lecture Notes in Computer Science, 157, Springer, 117–130.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34, 12077–12090.

Zhang, Y., Hu, S., Jiang, C., Cheng, Y., Qi, Y., 2023. Segment anything model with uncertainty rectification for auto-prompting medical image segmentation. *CoRR*.

Zhou, T., Xia, W., Zhang, F., Chang, B., Wang, W., Yuan, Y., Konukoglu, E., Cremers, D., 2024. Image segmentation in foundation model era: A survey. *arXiv preprint arXiv:2408.12957*.