International Journal of Computer Vision (2026) 134:142
https://doi.org/10.1007/s11263-026-02751-0

EMUFormer: Efficient Multi-task Uncertainties for Reliable Joint Semantic Segmentation and Monocular Depth Estimation

Steven Landgraf[1] · **Markus Hillemann**[1] · **Theodor Kapler**[1] · **Markus Ulrich**[1]

Received: 7 February 2025 / Accepted: 4 January 2026

Abstract

Quantifying the predictive uncertainty emerged as a possible solution to common challenges like overconfidence, lack of explainability, and robustness of deep neural networks, albeit one that is often computationally expensive. Many real-world applications are multi-modal in nature and hence benefit from multi-task learning. In autonomous driving or robotics, for example, the joint solution of semantic segmentation and monocular depth estimation has proven to be valuable. To this end, we introduce EMUFormer, a novel student-teacher distillation approach for efficient multi-task uncertainties in the context of joint semantic segmentation and monocular depth estimation. By leveraging the predictive uncertainties of the teacher, EMUFormer achieves new state-of-the-art results on Cityscapes and NYUv2 and additionally estimates reliable predictive uncertainties for both tasks that are comparable or superior to a Deep Ensemble despite being an order of magnitude more efficient to compute. These findings even extend to out-of-domain and domain adaptation scenarios, highlighting EMUFormer's remarkable reliability.

Keywords Uncertainty Quantification · Semantic Segmentation · Monocular Depth Estimation · Knowledge Distillation · Out-of-Domain · Domain adaptation.

1 Introduction

Due to their unparalleled performance in fundamental perception tasks like semantic segmentation (Minaee et al., 2022) or monocular depth estimation (Dong et al., 2022), deep neural networks are increasingly being deployed in real-time and safety-critical applications such as autonomous driving (McAllister et al., 2017), industrial inspection (Heizmann et al., 2022; Steger et al., 2018), and automation (Landgraf et al., 2023). However, they often suffer from overconfidence (Guo et al., 2017), lack explainability (Gawlikowski et al., 2022), and struggle to distinguish between in-domain and out-of-domain samples (Lee et al., 2018), which is of paramount importance for applications where prediction reliability is crucial. Since incorrect predictions can lead to severe consequences, previous work suggests that quantifying the uncertainty inherent in a model's prediction is a promising endeavor to make such applications safer (Landgraf et al., 2024a, b; Lee et al., 2018; Leibig et al., 2017; Loquercio et al., 2020; Mukhoti & Gal, 2018; Mukhoti et al., 2023). In autonomous driving, for instance, the car could provide feedback to the driver when it is uncertain or preemptively make risk-averse predictions based on the uncertainty.

In recent years, a number of promising uncertainty quantification methods for deep neural networks have been proposed (Amini et al., 2020; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Liu et al., 2020; MacKay, 1992; Mukhoti et al., 2023; Valdenegro-Toro, 2023; Van Amersfoort et al., 2020). Unfortunately, these methods either introduce technical complexity or require computationally expensive sampling from a stochastic process to estimate the uncertainty of a prediction. Additionally, they do not exploit that many real-world applications in robotics (Nekrasov et al., 2019) or autonomous driving (Chen et al., 2018) are

Communicated by Zorah Laehner.

✉ Steven Landgraf
steven.landgraf@kit.edu

Markus Hillemann
markus.hillemann@kit.edu

Theodor Kapler
theodor.kapler@student.kit.edu

Markus Ulrich
markus.ulrich@kit.edu

[1] Institute of Photogrammetry and Remote Sensing (IPF), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Published online: 06 March 2026

Table 1 Overview of the segmentation (Seg.), depth estimation (Depth) and uncertainty quantification (Pred. Unc.) capabilities as well as the respective number of parameters, FLOPs and FPS for different single-task and multi-task models and their respective Deep Ensemble (DE) versions with 10 members. SegFormer (Xie et al., 2021) and DepthFormer represent single-task models, whereas SegDepthFormer and EMUFormer depict multi-task models. B2 represents the medium-sized encoder of SegFormer, which was used for all models. Results are based on single-scale inference conducted on the NYUv2 (Silberman et al., 2012) dataset using an NVIDIA A100 GPU

	Seg.	Pred. Unc.	Depth	Pred. Unc.	Parameters	FLOPs	FPS
a) SegFormer-B2 Xie et al. (2021)	✓	×	×	×	27.3M	72.6G	55.3
b) DepthFormer-B2	×	×	✓	×	27.3M	72.1G	57.1
c) SegDepthFormer-B2	✓	×	✓	×	30.5M	120.1G	44.8
DE of a)	✓	✓	×	×	273.6M	726.4G	5.6
DE of b)	×	×	✓	✓	273.5M	720.8G	7.2
DE of c)	✓	✓	✓	✓	305.1M	1201.1G	4.9
EMUFormer-B2 (Ours)	✓	✓	✓	✓	30.5M	120.1G	44.8

multi-modal in nature and, hence, have the potential to benefit from multi-task learning, especially within the context of semantic segmentation and monocular depth estimation (Chen et al., 2018; Nekrasov et al., 2019). Although there have been successful attempts to make uncertainty quantification methods more efficient through knowledge distillation (Besnier et al., 2021; Holder & Shafique, 2021; Landgraf et al., 2024b; Shen et al., 2021; Simpson et al., 2022), they have either focused on semantic segmentation (Landgraf et al., 2024b; Besnier et al., 2021; Holder & Shafique, 2021; Shen et al., 2021) or monocular depth estimation (Shen et al., 2021; Simpson et al., 2022). This represents a notable research gap in the current literature.
Our contributions can be summarized as follows:

- We propose a novel student-teacher distillation approach for **E**fficient **M**ulti-task **U**ncertainties for joint semantic segmentation and monocular depth estimation with a modern Vision Trans**Former**, which we call **EMUFormer**.
- We highlight the generalizability and reliability of EMUFormer on two out-of-domain scenarios: Foggy Cityscapes (Sakaridis et al., 2018) and Rainy Cityscapes (Hu et al., 2019).
- We assess the domain adaptation capabilities of EMUFormer. This work introduces a novel perspective for evaluating efficient uncertainty quantification methods. It emphasizes their capacity to adapt efficiently to domain shifts without requiring extensive re-training or re-distillation efforts.
- We show that by leveraging the predictive uncertainties during training through the use of the Gaussian Negative Log-Likelihood loss, EMUFormer achieves state-of-the-art results on Cityscapes and NYUv2, while providing high-quality uncertainties with a single forward pass.

As Table 1 demonstrates, EMUFormer is able to estimate predictive uncertainties for both tasks that are comparable to the Deep Ensemble teacher despite being an order of magnitude more efficient to compute.

Please note that only the second and third contributions are novel to this work, while the others have been previously published in Landgraf et al. (2024).

2 Related Work

In this section, we summarize the related work on joint semantic segmentation and monocular depth estimation, uncertainty quantification, and knowledge distillation.

2.1 Joint Semantic Segmentation and Monocular Depth Estimation

Semantic segmentation and monocular depth estimation are both fundamental problems in image understanding that involve pixel-wise predictions based on a single input image. Motivated by the strong correlation and complementary properties of the two tasks, multiple previous works have focused on solving both tasks jointly (Bruggemann et al., 2020; Brüggemann et al., 2021; Gao et al., 2022; He et al., 2021; Ji et al., 2023; Jiao et al., 2018; Kendall et al., 2018; Lin et al., 2019; Liu et al., 2018, 2019; Mousavian et al., 2016; Nekrasov et al., 2019; Wang et al., 2015; Xu et al., 2018, 2022). General multi-task approaches with joint representation sharing (Zhang & Yang, 2021) or methods that leverage the depth map to improve the semantic segmentation prediction (Hu et al., 2018; Wang et al., 2021) are not relevant for our work and, therefore, are not covered by this review.

Wang et al. (2015) and Liu et al. (2018) propose frameworks for combining semantic segmentation and monocular

depth estimation using conditional random fields. In contrast, Mousavian et al. (2016) train parts of the model for each task separately and then fine-tune the full model on both tasks with a single loss function. Multiple previous works introduce attention mechanisms to improve the results (Brüggemann et al., 2021; Gao et al., 2022; Jiao et al., 2018; Liu et al., 2019). Gao et al. (2022) and Kendall et al. (2018) introduce confidences to weight the individual losses accordingly. Xu et al. (2018) propose a multi-task prediction-and-distillation network, where the predictions of intermediate auxiliary tasks are the multi-modal input for the final task – a concept that others (Nekrasov et al., 2019; Vandenhende et al., 2020) followed as well. Finally, there are multiple works (He et al., 2021; Ji et al., 2023; Lin et al., 2019) that propose specialized architectures, focusing on task-relevant feature separation, geometric constraints, and dynamic loss balancing, respectively.

Remarkably, most of the discussed approaches use older CNN-based architectures and require complex adaptations to either the model, the training process, or both. To advance the state of the art, we adapt a relatively recent Vision-Transformer-based architecture, similar to Xu et al. (2022), which offers strong accuracy–efficiency trade-offs. Also, to maintain methodological simplicity and transparency of the results, we do not use cross-task attention mechanisms, contrastive self-supervised learning algorithms, or a demanding loss weighting strategy like that of Kendall et al. (2018), and nevertheless achieve superior results. However, these strategies could be applied to our method as well, potentially further improving the results.

2.2 Uncertainty Quantification

A variety of uncertainty quantification methods (Amini et al., 2020; Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Liu et al., 2020; MacKay, 1992; Mukhoti et al., 2023; Valdenegro-Toro, 2023; Van Amersfoort et al., 2020) address the shortcomings of deep neural networks. Predictive uncertainty can be decomposed into aleatoric and epistemic uncertainty (Gal, 2016), which is essential for applications like active learning and detecting out-of-distribution samples (Gal et al., 2017; Schwaiger et al., 2020). Aleatoric uncertainty captures the irreducible data uncertainty, such as image noise or noisy labels from imprecise measurements. Epistemic uncertainty accounts for the model uncertainty and can be reduced with more or better training data (Gal, 2016; Kendall & Gal, 2017).

Most well-known uncertainty quantification methods require multiple forward passes at test time, making them computationally expensive. For instance, Gal and Ghahramani (2016) propose Monte Carlo Dropout (MCD) as an approximation of a stochastic Gaussian process. While dropout is usually only used for regularization during training (Srivastava et al., 2014), MCD applies this technique during test time to sample from the posterior distribution of the predictions at test time. Although MCD is easy to implement and thus very popular, Deep Ensembles (Lakshminarayanan et al., 2017) are commonly regarded as the state-of-the-art approach for uncertainty quantification across varying tasks (Gustafsson et al., 2020; Ovadia et al., 2019; Wursthorn et al., 2022, 2024). They consist of an ensemble of trained models that generate diverse predictions due to the introduction of randomness through random weight initialization or different data augmentations during training (Fort et al., 2020).

Due to the impracticality of multiple forward passes in time-critical applications, there's interest in deterministic single forward-pass methods. Van Amersfoort et al. (2020) and Liu et al. (2020) consider distance-aware output layers in the form of radial basis functions or Gaussian processes for uncertainty quantification. Mukhoti et al. (2023) propose an alternative approach with Gaussian Discriminant Analysis post-training. Valdenegro-Toro (2023) proposes Deep Sub-Ensembles (DSE), where the ensemble covers only a subset of layers instead of the whole model. They enable a trade-off between uncertainty quality and computational cost (Valdenegro-Toro, 2023).

Overall, quantifying uncertainties in joint semantic segmentation and monocular depth estimation has been largely overlooked. To the best of our knowledge, Landgraf et al. (2024) are the only ones to consider this research question so far. They compare multiple uncertainty quantification methods for this task and show how multi-task learning has the potential to positively influence the quality of uncertainty estimates in comparison to solving both tasks separately.

2.3 Knowledge Distillation

Knowledge distillation, introduced by Hinton et al. (2015), involves transferring the knowledge from a complex model (teacher) to a typically smaller model (student). It aims to enhance the student's performance on a given task by imitating the predictions of the teacher (Hinton et al., 2015) or transferring knowledge from intermediate features (Romero et al., 2015). More recent work has adapted knowledge distillation to enable real-time uncertainty quantification. While some previous work employs MCD to estimate uncertainties for the student to learn (Besnier et al., 2021; Gurau et al., 2018; Shen et al., 2021), the majority proposes to use a Deep Ensemble (Deng et al., 2021; Holder & Shafique, 2021; Landgraf et al., 2024b; Malinin et al., 2019; Simpson et al., 2022). Among these, Deng et al. (2021) are the only ones to consider a multi-task problem in the field of emotion recognition.

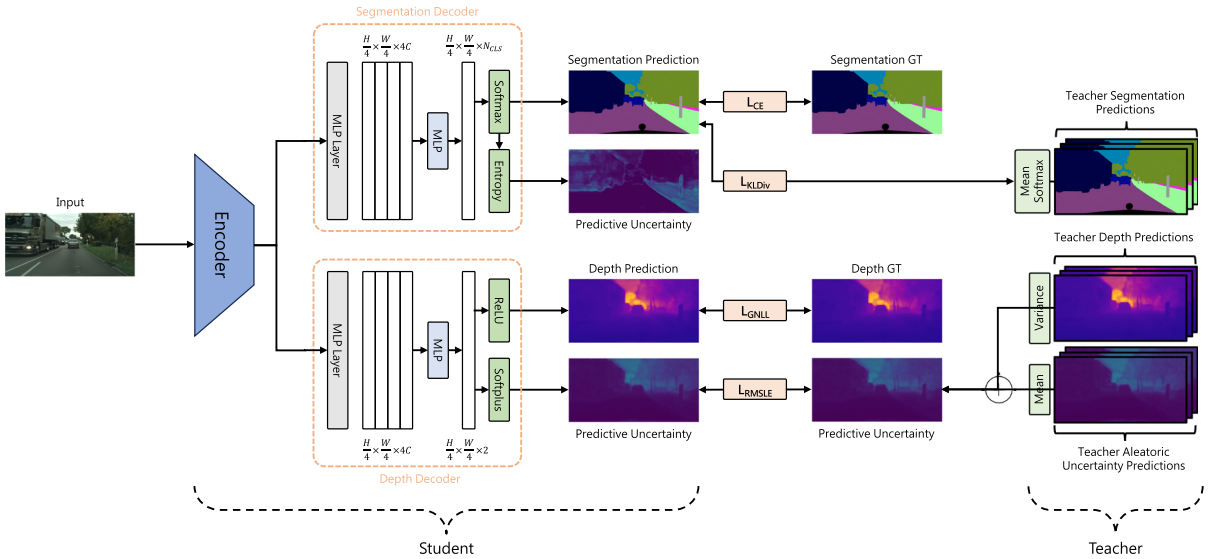

Fig. 1 A schematic overview of EMUFormer. In addition to the regular Cross-Entropy (CE) loss for the semantic segmentation task and the Gaussian Negative Log-Likelihood (GNLL) loss for the monocular depth estimation task, EMUFormer utilizes two additional losses that distill the predictive uncertainties of the teacher into the student model: the Kullback-Leibler (KL) divergence loss for segmentation uncertainty distillation and the root mean squared logarithmic error (RMSLE) for depth uncertainty distillation

3 EMUFormer

In the following, we explain our student-teacher distillation framework for efficient multi-task uncertainties, which we call EMUFormer. Our objective with EMUFormer is threefold:

1. Achieve state-of-the-art joint semantic segmentation and monocular depth estimation results
2. Estimate well-calibrated predictive uncertainties for both tasks
3. Avoid introducing additional computational overhead during inference

In order to achieve these goals, EMUFormer employs a two-step student-teacher distillation framework:

1. Training a teacher with ground truth labels
2. Training the student with ground truth labels while distilling the teacher's predictive uncertainties

In principle, any architecture capable of outputting a semantic segmentation mask along with a predictive mean and variance for monocular depth estimation is suitable for EMUFormer.

Student. To solve both predictive tasks simultaneously, we use a modified version of SegFormer (Xie et al. 2021), which we call SegDepthFormer. In addition to the efficient yet effective hierarchical Transformer-based encoder and entirely composed of multilayer perceptrons (MLP) decoder of SegFormer, we add an all-MLP depth decoder like shown in the left part of Fig. 1. Since the intricate details of SegDepthFormer are not part of the primary contributions of this paper, we provide details of the architecture and the training criterion in the Appendix (cf. Section A).

Teacher. Our framework is flexible with regard to the type of teacher. We select a Deep Ensemble that is known for producing high-quality estimates (Gustafsson et al., 2020; Ovadia et al., 2019; Wursthorn et al., 2022), which is also the case for multi-task uncertainty estimation (Landgraf et al., 2024).

Improving Uncertainty Distillation. To determine both predictive uncertainties for the uncertainty distillation, we compute multiple prediction samples from the teacher, as presented in Fig. 1. Since we use the same dataset for training and distillation, it is possible that the student underestimates the epistemic uncertainty of the teacher because of overfitting. Hence, we add color jittering as additional data augmentation to the teacher's input $\tilde{x}$, which was shown to be helpful by previous work on uncertainty distillation (Landgraf et al., 2024b; Shen et al., 2021). The color jitter causes the teacher's uncertainty distribution on the training dataset to be more closely aligned with the test-time distribution. In our experiments, we observed that removing this augmentation leads to noticeably less reliable uncertainty estimates, confirming that the diversity induced by color jittering plays a crucial role in preventing the student from overfitting to overly confident teacher predictions.

Training Criterion. As Fig. 1 shows, EMUFormer is trained to minimize the weighted sum of four objective func-

tions:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + w_1\mathcal{L}_{\text{GNLL}} + w_2\mathcal{L}_{\text{KL}} + w_3\mathcal{L}_{\text{RMSLE}}, \tag{1}$$

where

- $\mathcal{L}_{\text{CE}}$ is the categorical Cross-Entropy (CE) loss for the semantic segmentation task,
- $\mathcal{L}_{\text{GNLL}}$ is the Gaussian Negative Log-Likelihood loss for the monocular depth estimation task,
- $\mathcal{L}_{\text{KL}}$ is the Kullback-Leibler divergence loss for the segmentation uncertainty distillation,
- $\mathcal{L}_{\text{RMSLE}}$ is the root mean squared logarithmic (RMSLE) error for the depth uncertainty distillation,
- and the weighting factors are empirically set to $w_1 = w_3 = 1$ and $w_2 = 10$.

Segmentation Criterion. For the semantic segmentation task, we use the well-known categorical Cross-Entropy loss

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c} \cdot \log(p(z)_{n,c}) \ , \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ is the Cross-Entropy (CE) loss for a single image, 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.

Depth Criterion. For regression tasks, neural networks typically output only a predictive mean $\mu(z)$ and the parameters are, in the most straightforward approach, optimized by minimizing the mean squared error (MSE). However, the MSE does not cover uncertainty. Therefore, we follow the approach of Nix and Weigend (1994) instead: By treating the neural networks prediction as a sample from a Gaussian distribution with the predictive mean $\mu(z)$ and corresponding predictive variance $s^2(z)$, 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 ground truth depth.

Since this is a major insight of our work, we want to highlight that, usually, $s^2(z)$ is solely learned implicitly through the optimization of the predictive means based on the ground truth labels. In the case of EMUFormer, however, the network is also being trained to mimic the predictive uncertainty of the teacher in parallel. Consequently, the depth uncertainty does not need to be learned implicitly, rather it can be used explicitly to improve the depth estimation itself.

Segmentation Uncertainty Distillation. The segmentation uncertainty knowledge of the teacher model is transferred into the student model by using the Kullback-Leibler divergence loss:

$$\mathcal{L}_{\text{KL}} = \sum_{c=1}^{C} q_c(\tilde{z}) \cdot \log\left(\frac{q_c(\tilde{z})}{p_c(z)}\right) \ , \tag{4}$$

where $\tilde{z}$ are the logits based on the color jittered input image $\tilde{x}$, $q_c(\tilde{z})$ is the teacher's mean softmax probability map, and $p_c(z)$ is the student's softmax probability map. Minimizing this loss ensures that the student learns to match the well-calibrated softmax probabilities provided by the teacher, allowing the predictive entropy

$$H(p(z)) = -\sum_{c=1}^{C} p(z)_c \cdot \log(p(z)_c) \tag{5}$$

to capture the underlying predictive uncertainty.

Depth Uncertainty Distillation. Since it is not possible to match two distributions for the unbound uncertainties in the regression task, we introduce the root mean squared logarithmic error (RMSLE) for the depth uncertainty distillation:

$$\mathcal{L}_{\text{RMSLE}} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\log(\sigma_n^2(\tilde{z})+1) - \log(s_n^2(z)+1)\right)^2} \ , \tag{6}$$

where $\sigma_n^2(\tilde{z})$ is the teacher's predictive uncertainty based on the color jittered input image $\tilde{x}$ and $s_n^2(z)$ is the student's predictive uncertainty estimate. The natural logarithm penalizes underestimations more than overestimations, thereby providing special attention to the pixels with higher uncertainties. By minimizing the depth uncertainty loss, the student is trained to mimic the predictive uncertainty of the teacher.

4 Experimental Setup

Baseline Models. For the sake of a more comprehensive comparison, we use two single-task models and one multi-task model as baselines:

1. SegFormer (Xie et al., 2021) for the semantic segmentation task
2. DepthFormer for the monocular depth estimation task
3. SegDepthFormer for joint semantic segmentation and monocular depth estimation

Considering the efficiency and performance of SegFormer (Xie et al., 2021), we derive the latter two models from it, making only minimal changes to adjust for the respective task. For DepthFormer, we only change the output layer to

produce two output channels: one for the predictive mean and one for the predictive variance, a common approach for uncertainty-aware monocular depth estimation (Lakshminarayanan et al., 2017; Loquercio et al., 2020). In order to jointly solve both tasks with SegDepthFormer, we extend SegFormer (Xie et al., 2021) by the decoder of DepthFormer and use one shared encoder.

More details on the baseline models can be found in the Appendix (cf. Sect. A).

Datasets. We conduct all experiments on Cityscapes (Cordts et al., 2016) and NYUv2 (Silberman et al., 2012). Cityscapes is a popular urban street scene benchmark dataset with 2975 training and 500 validation images. We focus on Cityscapes because it most closely reflects the real-world conditions where robust yet efficient uncertainty estimates are most needed. NYUv2 contains 795 training and 654 testing images of indoor scenes and serves as the main reference benchmark for previous joint semantic segmentation and monocular depth estimation models.

For the out-of-domain evaluation, we use the validation sets of Foggy Cityscapes (Sakaridis et al., 2018) and Rainy Cityscapes (Hu et al., 2019), which introduce progressively increasing perturbations to the original urban street scenes, simulating adverse weather conditions in the form of fog and rain. These synthetic variants enable a fine-grained robustness analysis under progressively increasing distribution shifts – an aspect that is difficult to achieve with real-world datasets such as ACDC (Sakaridis et al., 2021).

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. All training runs utilize the AdamW (Loshchilov & Hutter, 2017) optimizer with a base learning rate of 0.00006 and employ a polynomial learning rate scheduler defined as:

$$lr = lr_{\text{base}} \cdot \left(1 - \frac{\text{iteration}}{\text{total iterations}}\right)^{0.9} , \tag{7}$$

where lr is the current learning rate and lr_{base} is the initial base learning rate. Besides, we employ a batch size of 8 and train on four NVIDIA A100 GPUs with 40 GB of memory using mixed precision (Micikevicius et al., 2017). The encoders of the baseline models are initialized with weights pre-trained on ImageNet (Deng et al., 2009) and subsequently trained for 250 epochs on Cityscapes and for 100 epochs on NYUv2. In contrast, EMUFormer is initialized with the weights of a pre-trained SegDepthFormer and fine-tuned for only 100 epochs on both datasets. To maintain simplicity and transparency, we refrain from employing common techniques such as OHEM (Shrivastava et al., 2016), auxiliary losses, class imbalance compensation, or sliding window testing to boost performance.

Metrics. For the semantic segmentation task, we report the mean Intersection over Union (mIoU), also referred to as the Jaccard Index. Softmax probability calibration is evaluated using the Expected Calibration Error (ECE) (Naeini et al., 2015). For monocular depth estimation, we use the root mean squared error (RMSE). To assess uncertainty quality, we adopt the following metrics proposed by Mukhoti and Gal (2018):

$$\begin{aligned} p(\text{accurate}|\text{certain}) &= \frac{n_{ac}}{n_{ac} + n_{ic}} , \\ p(\text{uncertain}|\text{inaccurate}) &= \frac{n_{iu}}{n_{iu} + n_{ic}} , \\ \text{PAvPU} &= \frac{n_{ac} + n_{iu}}{n_{ac} + n_{au} + n_{ic} + n_{iu}} , \end{aligned} \tag{8}$$

where n_{ac} represents the number of pixels that are accurate and certain, n_{ic} the number of pixels that are inaccurate and certain, n_{iu} the number of pixels that are inaccurate and uncertain, and n_{au} is the number of pixels that are accurate and uncertain.

Although originally designed for semantic segmentation (Mukhoti & Gal, 2018), we extend these metrics to evaluate the depth regression uncertainties as well. To decide whether a depth prediction is accurate, we apply the following formula:

$$\max\left(\frac{\mu(z)}{y}, \frac{y}{\mu(z)}\right) = \delta_1 < 1.25 , \tag{9}$$

where $\mu(z)$ is the predicted depth value of a pixel and y is the corresponding ground truth depth (Ming et al., 2021). δ_1 serves as a standard metric for quantifying the accuracy of monocular depth estimation models, using 1.25 as the threshold that must not be exceeded by a depth prediction to be counted as accurate. In contrast, δ_2 and δ_3 are less strict, utilizing thresholds of 1.25^2 and 1.25^3, respectively.

5 Experiments

We conduct several experiments to demonstrate the efficiency and efficacy of EMUFormer. Firstly, we compare EMUFormer's performance with its Deep Ensemble teacher and the baseline models. Subsequently, we contrast our results with previous state-of-the-art approaches, followed by qualitative examples. Then, we evaluate the generalizability of EMUFormer on two out-of-domain datasets in comparison to its Deep Ensemble teacher and explore its capacity for domain adaptation. Thereafter, we study the impact of utilizing the distilled uncertainties in the GNLL loss. Lastly, we provide an ablation study on different backbone sizes.

Unless otherwise specified, we use SegFormer's B2 backbone as default for all experiments as a compromise between efficiency and performance. In addition, we report results for the smallest (B0) and largest (B5) backbones to analyze the effect of model capacity. We employ a SegDepthFormer Deep Ensemble with 10 members as the teacher for all experiments.

5.1 Quantitative Evaluation

Baseline vs. Teacher vs. Student. We present a comprehensive analysis in Table 2 by comparing the baseline models, their Deep Ensembles versions with 10 members, and EMUFormer. EMUFormer emerges as the standout performer, surpassing the baseline models across a large majority of the metrics on both datasets. Remarkably, this performance is achieved while maintaining an equivalent inference time. EMUFormer even outperforms the SegDepthFormer Deep Ensemble, which served as its teacher and has approximately 33 times higher inference time, in most cases. In terms of prediction performance, EMUFormer gives marginally worse segmentation results compared to the SegFormer Deep Ensemble on Cityscapes. However, it notably excels in the depth estimation task, especially on Cityscapes (Cordts et al., 2016), which is a phenomenon we observed across multiple experiments (cf. Tables 3 and 9). We primarily attribute this improvement to the utilization of the predictive uncertainties inside the Gaussian Negative Log-Likelihood loss, but investigate this more thoroughly in Sects. 5.5 and 6.

Comparison with SOTA.

Table 3 compares EMUFormer with previous state-of-the-art methods on Cityscapes (Cordts et al., 2016) and NYUv2 (Silberman et al., 2012). Across both datasets, EMUFormer consistently achieves competitive results, with EMUFormer-B5 outperforming all previous approaches in joint semantic segmentation and monocular depth estimation. For instance, on NYUv2, EMUFormer-B5 attains a 1.4% higher mIoU and a 0.007 lower RMSE than MTFormer (Xu et al., 2022), which also employs a Vision-Transformer-based architecture.

We note, however, that performance naturally scales with model capacity: larger backbones such as MIT-B5 offer higher representational power, while smaller variants provide stronger efficiency–accuracy trade-offs. To ensure transparency, we report results for multiple backbone sizes (B0, B2, and B5) and discuss EMUFormer-B5 as the reference configuration only because it establishes the upper performance bound of our approach. Importantly, even smaller EMUFormer variants remain competitive with methods of comparable or larger capacity, demonstrating the scalability and robustness of our design.

In contrast to prior work such as MTFormer (Xu et al., 2022), EMUFormer achieves these results without cross-task attention or self-supervised pre-training, keeping the architecture and approach lightweight. Moreover, EMUFormer provides high-quality uncertainty estimates without any additional computational overhead during inference.

5.2 Qualitative Evaluation

In Fig. 2, we provide qualitative examples of EMUFormer-B2 for Cityscapes (Cordts et al., 2016) and NYUv2 (Silberman et al., 2012).

Cityscapes. On Cityscapes, EMUFormer demonstrates good prediction performance for both tasks. In the segmentation task, its uncertainty prediction proves particularly insightful as highlighted by the red rectangles. For example, for the car hood, which is not part of the training labels (indicated by black pixels), the model exhibits high uncertainty values, indicating its ability to capture out-of-distribution information or epistemic uncertainty. Similarly, in noisy background areas, the model effectively captures the aleatoric uncertainty. Additionally, the model correctly predicts high uncertainties for challenging areas like the wall on the right of the image, demonstrating the benefit of uncertainties in identifying potential model errors. In the depth estimation task, analogously to the segmentation task, EMUFormer predicts high uncertainty on the car hood and the sky, which are both areas without ground truth information, i.e., areas of high epistemic uncertainty. Furthermore, the uncertainty is appropriately high at object boundaries, indicating sensitivity to significant depth discontinuities.

NYUv2. For the segmentation task, EMUFormer again outputs high uncertainties for pixels without ground truth information or that are misclassified, consistently providing useful predictive uncertainties. In the depth estimation task, the uncertainties correlate with the estimated depth, providing an intuitive and helpful indication. This alignment suggests that the model effectively captures the depth prediction quality, particularly as it relates to increasing distances.

In summary, the qualitative evaluation aligns with the quantitative findings of Sect. 5.1. It demonstrates the proficiency of EMUFormer in handling both the segmentation and the depth estimation tasks and its ability to generate meaningful predictive uncertainties that enable more thorough interpretations of the predictions.

5.3 Out-of-Domain Evaluation

For out-of-domain evaluation, we compare the SegDepthFormer baseline model, a SegDepthFormer Deep Ensemble with 10 members (teacher), and EMUFormer on two out-of-domain datasets: Foggy Cityscapes (Sakaridis et al., 2018) and Rain Cityscape (Hu et al., 2019). In order to evaluate the generalizability, we do not fine-tune any model.

Table 2 Quantitative comparison on the Cityscapes (Cordts et al., 2016) and NYUv2 (Silberman et al., 2012) datasets between the baseline models, their Deep Ensemble (DE) versions with ten members, and EMUFormer. SegDepthFormer (DE) serves as the teacher

	Semantic Segmentation					Monocular Depth Estimation				Inference Time [ms]
	mIoU ↑	ECE ↓	p(acc/cer) ↑	p(inacc/unc) ↑	PAvPU ↑	RMSE ↓	p(acc/cer) ↑	p(inacc/unc) ↑	PAvPU ↑	
Cityscapes										
SegFormer Xie et al. (2021)	0.772	0.033	0.882	0.395	0.797	–	–	–	–	17.90 ± 0.47
SegFormer (DE)	**0.784**	0.033	0.887	0.416	0.798	–	–	–	–	667.51 ± 2.89
DepthFormer	–	–	–	–	–	7.452	0.749	0.476	0.766	17.59 ± 0.82
DepthFormer (DE)	–	–	–	–	–	7.222	0.759	0.486	0.771	626.79 ± 2.05
SegDepthFormer	0.738	0.028	0.913	0.592	0.826	7.536	0.745	0.472	0.762	22.04 ± 0.27
SegDepthFormer (DE)	0.755	0.015	0.917	0.609	**0.828**	7.156	0.763	**0.493**	0.773	743.23 ± 32.95
EMUFormer	0.752	**0.012**	**0.923**	**0.658**	0.811	**6.983**	**0.772**	0.491	**0.783**	22.04 ± 0.27
NYUv2										
SegFormer Xie et al. (2021)	0.470	0.159	0.768	0.651	0.734	–	–	–	–	18.09 ± 0.41
SegFormer (DE)	**0.486**	**0.125**	0.782	0.675	0.734	–	–	–	–	715.97 ± 7.55
DepthFormer	–	–	–	–	–	0.554	0.786	0.449	0.610	17.51 ± 0.87
DepthFormer (DE)	–	–	–	–	–	0.524	0.808	**0.475**	0.613	624.30 ± 2.07
SegDepthFormer	0.466	0.151	0.769	0.659	0.733	0.558	0.776	0.446	0.594	22.31 ± 0.23
SegDepthFormer (DE)	0.481	0.122	0.783	0.682	0.733	0.552	0.785	0.453	0.590	788.76 ± 2.00
EMUFormer	0.475	0.129	**0.787**	**0.692**	**0.737**	**0.514**	**0.810**	0.440	**0.633**	22.31 ± 0.23

Best results are highlighted in bold

Table 3 Comparison against state-of-the-art approaches for joint semantic segmentation and monocular depth estimation

	NYUv2		Cityscapes	
	mIoU ↑	RMSE ↓	mIoU ↑	RMSE ↓
HybridNet A2 Lin et al. (2019)	0.343	0.682	0.666	12.09
Mousavian et al. (2016)	0.392	0.816	–	–
C-DCNN Liu et al. (2018)	0.398	0.628	–	–
BMTAS Bruggemann et al. (2020)	0.411	0.543	–	–
Gao et al. (2022)	0.419	0.528	–	–
CI-Net Gao et al. (2022)	0.426	0.504	0.701	6.880
SOSD-Net He et al. (2021)	0.450	0.514	0.682	–
Wang et al. (2015)$_{\text{CVPR'15}}$	0.442	0.745	–	–
PAD-Net Xu et al. (2018)$_{\text{CVPR'18}}$	0.502	0.582	0.761	–
Nekrasov et al. (2019)$_{\text{ICRA'19}}$	0.420	0.565	–	–
MTI-Net Vandenhende et al. (2020)$_{\text{ECCV'20}}$	0.490	0.529	–	–
ATRC Brüggemann et al. (2021)$_{\text{ICCV'21}}$	0.463	0.536	–	–
MTFormer Xu et al. (2022)$_{\text{ECCV'22}}$	0.506	0.483	–	–
EMUFormer-B0 (Ours)	0.363	0.674	0.630	8.086
EMUFormer-B2 (Ours)	0.475	0.514	0.752	6.983
EMUFormer-B5 (Ours)	**0.520**	**0.476**	**0.771**	**6.157**

Best results are highlighted in bold

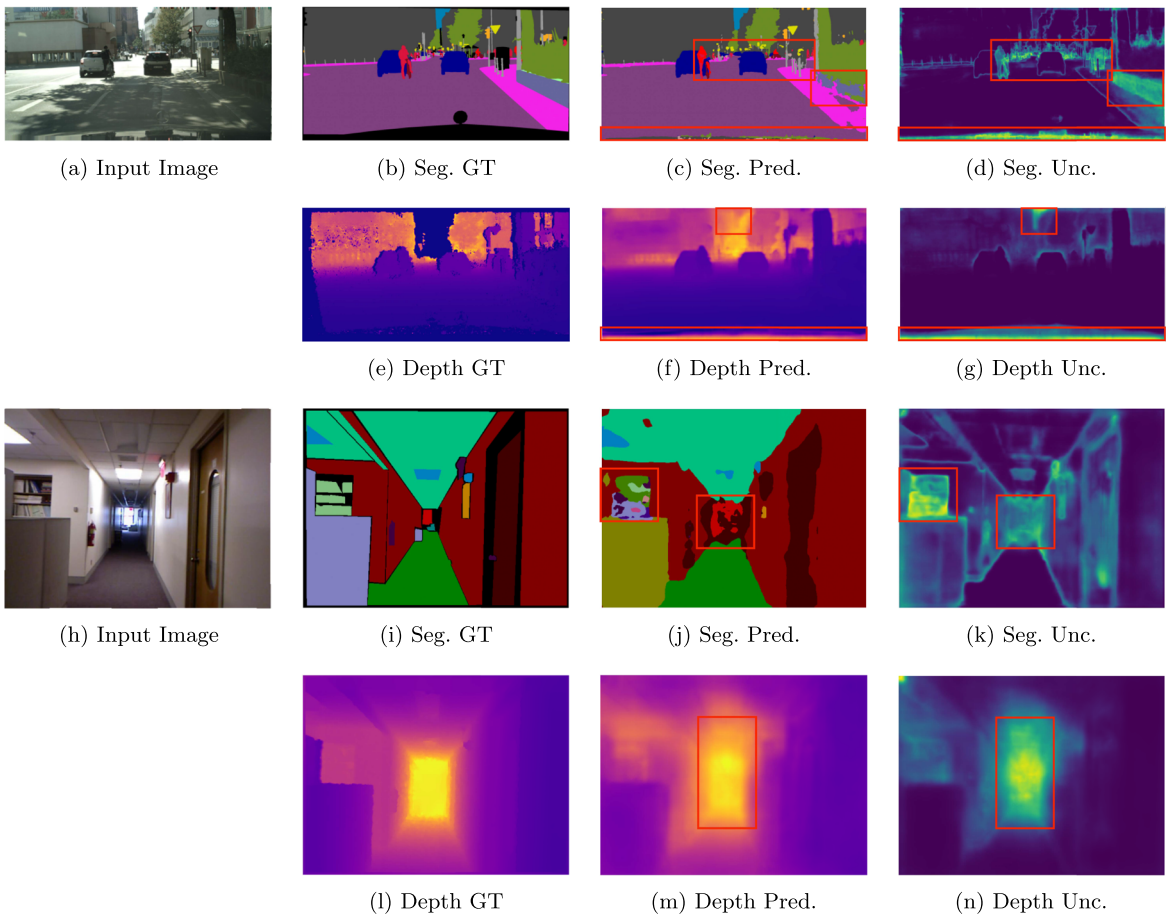

Fig. 2 Representative qualitative examples of EMUFormer-B2 on the Cityscapes (Cordts et al., 2016) (top) and NYUv2 (Silberman et al., 2012) (bottom) datasets. Red rectangles are added to highlight interesting areas. Best viewed in color. Brighter colors indicate higher uncertainty

Foggy Cityscapes. Compared to the original Cityscapes dataset, the Foggy Cityscapes dataset reveals significant performance degradation, as shown by Table 4. Both, the baseline and Deep Ensemble models, experience declines in predictive performance and calibration quality as the fog density increases, with the baseline showing more pronounced degradation. The Deep Ensemble demonstrates greater robustness, maintaining better performance and uncertainty quality even under severe fog conditions.

EMUFormer performs comparably to the Deep Ensemble in terms of predictive accuracy for the segmentation task while offering improved calibration, except for one case. Additionally, it delivers significantly better performance in terms of depth estimation. Regarding uncertainty quality, EMUFormer matches the Deep Ensemble in $p(\text{accurate}|\text{certain})$ and PAvPU for semantic segmentation but exhibits a notable improvement in $p(\text{uncertain}|\text{inaccurate})$. For depth estimation, EMUFormer provides equal or slightly better uncertainty quality across all evaluated metrics, further underscoring its effectiveness in handling challenging out-of-domain scenarios without the computational overhead of a Deep Ensemble.

Rainy Cityscapes. Table 5 highlights performance trends across varying levels of simulated rain. Both the baseline and Deep Ensemble models experience performance degradation as rain intensity increases, with the Deep Ensemble consistently demonstrating superior robustness in predictive performance, calibration quality, and uncertainty metrics.

Compared to the Deep Ensemble, EMUFormer shows mixed results under varying rain conditions. While it produces strong calibration on par with the Deep Ensemble, except for one case, its performance in predictive accuracy decreases with more intense rain. Under the most challenging conditions, EMUFormer performs worse than the Deep Ensemble and even slightly worse than the baseline in semantic segmentation accuracy, though it achieves significantly better results for depth estimation. Regarding segmentation uncertainty quality, EMUFormer is slightly worse than the Deep Ensemble for $p(\text{accurate}|\text{certain})$ and PAvPU, but performs on par for $p(\text{uncertain}|\text{inaccurate})$. For depth uncertainty, EMUFormer matches the Deep Ensemble in $p(\text{accurate}|\text{certain})$, performs slightly worse in $p(\text{uncertain}|\text{inaccurate})$, and slightly better in PAvPU. These results showcase that it can almost match the performance of a Deep Ensemble in out-of-domain scenarios while maintaining the computational efficiency of the baseline SegDepthFormer.

Summary. Overall, these results demonstrate that EMUFormer generalizes effectively to out-of-domain scenarios without fine-tuning, achieving competitive performance compared to the Deep Ensemble while maintaining the computational efficiency of the baseline SegDepthFormer. EMUFormer matches or exceeds the Deep Ensemble in cal-

Table 4 Quantitative comparison between the SegDepthFormer baseline model, a SegDepthFormer Deep Ensemble (DE) with 10 members (teacher), and EMUFormer (student) on the Foggy Cityscapes validation dataset (Sakaridis et al., 2018) without fine-tuning. β denotes the attenuation coefficient and controls the thickness of the fog. Higher β values result in thicker fog. The original Cityscapes and the Foggy Cityscapes datasets share the same validation images, enabling a fair comparison between in-domain and out-of-domain results

		Semantic Segmentation					Monocular Depth Estimation			
		mIoU ↑	ECE ↓	p(acc|cer) ↑	p(unc|inacc) ↑	PAvPU ↑	RMSE ↓	p(acc|cer) ↑	p(unc|inacc) ↑	PAvPU ↑
Cityscapes	SegDepthFormer	0.738	0.028	0.913	0.592	0.826	7.536	0.745	0.472	0.762
	SegDepthFormer (DE)	0.755	0.015	0.917	0.609	0.828	7.156	0.763	0.493	0.773
	EMUFormer	0.752	0.012	0.923	0.658	0.811	6.983	0.772	0.491	0.783
Foggy$_{\beta=0.005}$	SegDepthFormer	0.707	0.035	0.906	0.602	0.818	8.061	0.731	0.481	0.751
	SegDepthFormer (DE)	0.727	0.028	0.914	0.627	0.822	7.487	0.758	0.509	0.765
	EMUFormer	0.721	0.040	0.919	0.678	0.803	7.182	0.769	0.500	0.780
Foggy$_{\beta=0.01}$	SegDepthFormer	0.674	0.054	0.899	0.606	0.814	8.628	0.715	0.475	0.741
	SegDepthFormer (DE)	0.699	0.056	0.910	0.637	0.817	7.971	0.750	0.511	0.761
	EMUFormer	0.691	0.027	0.915	0.694	0.794	7.635	0.764	0.506	0.778
Foggy$_{\beta=0.02}$	SegDepthFormer	0.609	0.078	0.875	0.593	0.798	9.844	0.697	0.467	0.730
	SegDepthFormer (DE)	0.639	0.045	0.895	0.644	0.803	9.213	0.738	0.517	0.760
	EMUFormer	0.629	0.015	0.904	0.714	0.778	8.927	0.750	0.518	0.772

ibration and uncertainty quality for depth estimation and delivers robust segmentation performance under foggy conditions. Although performance declines slightly under heavy rain, particularly in semantic segmentation, EMUFormer still offers strong depth estimation and uncertainty calibration, highlighting its ability to handle domain shifts efficiently without the need for ensemble-based methods.

5.4 Domain adaptation

Domain adaptation is essential for achieving robust model performance across diverse environmental conditions by facilitating knowledge transfer to new domains. While some previous uncertainty quantification distillation approaches (Deng et al., 2021; Holder & Shafique, 2021; Landgraf et al., 2024b; Shen et al., 2021) have explored out-of-domain performance, they have largely overlooked domain adaptation. This gap is particularly significant for applications such as autonomous driving, where re-training a teacher model – often implemented as a Deep Ensemble – and repeating the distillation process for every new domain is prohibitively expensive and operationally impractical. To address this, we propose a novel perspective for evaluating uncertainty quantification distillation methods, emphasizing their capacity to adapt efficiently to domain shifts without requiring extensive re-training or re-distillation efforts.

More specifically, we evaluate the domain adaptation capabilities of EMUFormer by fine-tuning it on Foggy Cityscapes (Sakaridis et al., 2018) and Rainy Cityscapes (Hu et al., 2019). This setting aligns with the simplest homogeneous domain adaptation paradigm as defined by Wang and Deng (2018), where the source and target domains share identical feature spaces (semantically and dimensionally) but differ in input data distributions.

We follow the distillation process described in Sect. 3, with the exception of omitting the additional color jitter augmentation, as the training and distillation datasets are no longer identical. EMUFormer, initially trained on Cityscapes, is fine-tuned using the ground truth labels from Foggy Cityscapes or Rainy Cityscapes in conjunction with the outputs of the teacher Deep Ensemble trained on Cityscapes. The fine-tuning process is deliberately constrained to a single NVIDIA A100 GPU, with a maximum training duration of approximately 2.5 hours. This setup is designed to evaluate the domain adaptation capabilities of our approach while taking computational efficiency into account.

Quantitative Evaluation. Tables 6 and 7 present a quantitative comparison between the out-of-domain and fine-tuning results of EMUFormer on Foggy Cityscapes and Rainy Cityscapes, respectively. The results demonstrate significant benefits from domain adaptation across both semantic segmentation and depth estimation task.

Table 5 Quantitative comparison between the SegDepthFormer baseline model, a SegDepthFormer Deep Ensemble (DE) with 10 members (teacher), and EMUFormer (student) on the Rainy Cityscapes validation dataset (Hu et al., 2019) without fine-tuning. We evaluate on three sets of parameters, where Rain_1 uses [0.01, 0.005, 0.01], Rain_2 uses [0.02, 0.01, 0.005], and Rain_3 uses [0.03, 0.015, 0.002] as attenuation coefficients α and β and the raindrop radius a. α and β determine the degree of simulated rain and fog in the images

		Semantic Segmentation					Monocular Depth Estimation			
		mIoU ↑	ECE ↓	p(acc|cer) ↑	p(unc|inacc) ↑	PAvPU ↑	RMSE ↓	p(acc|cer) ↑	p(unc|inacc) ↑	PAvPU ↑
Rain_1	SegDepthFormer	0.608	0.020	0.936	0.658	0.810	7.187	0.792	0.558	0.767
	SegDepthFormer (DE)	0.673	0.004	0.954	0.741	0.813	6.740	0.804	0.559	0.767
	EMUFormer	0.647	0.006	0.943	0.739	0.784	6.538	0.805	0.534	0.774
Rain_2	SegDepthFormer	0.611	0.031	0.928	0.670	0.802	8.043	0.771	0.543	0.756
	SegDepthFormer (DE)	0.645	0.012	0.948	0.750	0.806	7.516	0.785	0.544	0.759
	EMUFormer	0.611	0.021	0.934	0.745	0.776	7.294	0.787	0.516	0.765
Rain_3	SegDepthFormer	0.582	0.045	0.917	0.671	0.795	8.848	0.751	0.534	0.749
	SegDepthFormer (DE)	0.612	0.023	0.943	0.756	0.799	8.294	0.767	0.535	0.755
	EMUFormer	0.576	0.026	0.928	0.751	0.767	8.033	0.772	0.510	0.761

Table 6 Quantitative comparison of out-of-domain (OOD) and fine-tuning (FT) results of EMUFormer (student) on the Foggy Cityscapes validation dataset (Sakaridis et al., 2018). The parameter β, representing the attenuation coefficient, determines the fog density, with higher β values corresponding to denser fog

		Semantic Segmentation					Monocular Depth Estimation			
		mIoU ↑	ECE ↓	p(acc|cer) ↑	p(unc|inacc) ↑	PAvPU ↑	RMSE ↓	p(acc|cer) ↑	p(unc|inacc) ↑	PAvPU ↑
$\text{Foggy}_{\beta=0.005}$	EMUFormer (OOD)	0.721	0.040	0.919	0.678	0.803	7.182	0.769	0.500	0.780
	EMUFormer (FT)	0.749 (+0.028)	0.011 (+0.029)	0.920 (+0.001)	0.631 (-0.047)	0.821 (+0.018)	6.314 (-0.868)	0.796 (+0.030)	0.522 (+0.022)	0.794 (+0.014)
$\text{Foggy}_{\beta=0.01}$	EMUFormer (OOD)	0.691	0.027	0.915	0.694	0.794	7.635	0.764	0.506	0.778
	EMUFormer (FT)	0.747 (+0.056)	0.019 (+0.008)	0.917 (+0.002)	0.635 (-0.059)	0.811 (+0.017)	5.631 (-2.004)	0.822 (+0.058)	0.547 (+0.041)	0.807 (+0.029)
$\text{Foggy}_{\beta=0.02}$	EMUFormer (OOD)	0.629	0.015	0.904	0.714	0.778	8.927	0.750	0.518	0.772
	EMUFormer (FT)	0.730 (+0.101)	0.004 (+0.011)	0.918 (+0.014)	0.662 (-0.052)	0.792 (+0.014)	5.463 (-3.464)	0.828 (+0.078)	0.563 (+0.045)	0.802 (+0.030)

Values in parentheses denote the difference relative to the OOD baseline after fine-tuning. Green text indicates an improvement, while red text indicates a degradation

Table 7 Quantitative comparison of out-of-domain (OOD) and fine-tuning (FT) results of EMUFormer (student) on the Rainy Cityscapes validation dataset (Hu et al., 2019). We evaluate on three sets of parameters, where Rain_1 uses [0.01, 0.005, 0.01], Rain_2 uses [0.02, 0.01, 0.005], and Rain_3 uses [0.03, 0.015, 0.002] for attenuation coefficients α and β and the raindrop radius a. α and β determine the degree of simulated rain and fog in the images

		Semantic Segmentation					Monocular Depth Estimation			
		mIoU ↑	ECE ↓	p(acc|cer) ↑	p(unc|inacc) ↑	PAvPU ↑	RMSE ↓	p(acc|cer) ↑	p(unc|inacc) ↑	PAvPU ↑
Rain_1	EMUFormer (OOD)	0.647	0.006	0.943	0.739	0.784	6.538	0.805	0.534	0.774
	EMUFormer (FT)	0.727 (+0.080)	0.010 (-0.004)	0.958 (+0.015)	0.705 (-0.034)	0.822 (+0.038)	4.730 (-1.808)	0.866 (+0.061)	0.582 (+0.048)	0.798 (+0.024)
Rain_2	EMUFormer (OOD)	0.611	0.021	0.934	0.745	0.776	7.294	0.787	0.516	0.765
	EMUFormer (FT)	0.680 (+0.069)	0.018 (+0.003)	0.957 (+0.023)	0.713 (-0.032)	0.800 (+0.024)	4.941 (-2.353)	0.873 (+0.096)	0.629 (+0.113)	0.787 (+0.022)
Rain_3	EMUFormer (OOD)	0.576	0.026	0.928	0.751	0.767	8.033	0.772	0.510	0.761
	EMUFormer (FT)	0.715 (+0.139)	0.025 (+0.001)	0.960 (+0.052)	0.733 (-0.018)	0.793 (+0.026)	4.566 (-3.467)	0.876 (+0.104)	0.606 (+0.096)	0.799 (+0.038)

Values in parentheses denote the difference relative to the OOD baseline after fine-tuning. Green text indicates an improvement, while red text indicates a degradation

For semantic segmentation, fine-tuning EMUFormer leads to improvements in mIoU of at least 2.8% and up to 13.9% compared to the out-of-domain baseline and an improved softmax calibration, as measured by ECE, in 5 out of 6 cases. Segmentation uncertainty quality also improves in terms of p(accurate|certain) and PAvPU. However, p(uncertain|inaccurate) shows a slight degradation, which can be attributed to the surprising strength of EMUFormer in terms of out-of-domain performance, outperforming its Deep Ensemble teacher on this specific metric, as shown by Table 4 in the previous Sect. 5.3. In the depth estimation task, fine-tuning yields substantial performance gains, with reductions in RMSE ranging from 0.868 to 3.467. Depth uncertainty quality also improves consistently across all evaluated metrics, highlighting the robustness and strong performance of EMUFormer, particularly in terms of monocular depth estimation, even while adapting to domain-specific conditions.

Qualitative Evaluation. Figure 3 presents qualitative examples of our domain-adapted, i.e., fine-tuned, EMUFormer-B2 on the most difficult versions of the Foggy Cityscapes (Sakaridis et al., 2018) and Rainy Cityscapes (Hu et al., 2019) validation datasets. EMUFormer demonstrates strong performance across both tasks and datasets, with uncertainty estimates effectively highlighting challenging regions. More specifically, on Foggy Cityscapes, segmentation uncertainty aligns with objects absent from the training data, such as an elderly person's walker in the foreground, as well as misclassified or noisy areas, exemplified by the region marked with a red rectangle in the right part of the image. Depth uncertainty is notably high for sky regions, where depth estimation is inherently ill-defined. Similarly, on Rainy Cityscapes, the model assigns high uncertainty to out-of-distribution objects, like a dumpster, and to distant regions obscured by rain and fog, as indicated in the central part of the image. Depth uncertainty remains elevated for sky pixels and distant, occluded regions, reflecting the model's sensitivity to visually ambiguous or uninformative cues.

Summary. Overall, these findings align with the quantitative evaluations, demonstrating that EMUFormer is capable of efficiently adapting to domain shifts without requiring extensive re-training or re-distillation efforts, while maintaining strong performance and reliable uncertainty estimates across both tasks.

5.5 Impact of Uncertainty Utilization

As described in Sect. 3 and shown by Equation 3, GNLL treats every prediction as a sample from a Gaussian distribution with a predictive mean and a corresponding predictive variance. Typically, these variances are learned implicitly through optimizing predictive means based on ground truth labels. However, with EMUFormer, the network is optimized to mimic the teacher's predictive uncertainty. This allows the depth uncertainty to be used explicitly to improve depth estimation. In order to explore this more thoroughly, we study the impact of the uncertainty utilization by replacing the GNLL loss with the Mean Squared Error (MSE) loss and the Huber loss (Huber, 1992), respectively, which do not account for the available predictive uncertainty.

Table 8 shows a quantitative comparison of the impact of the respective depth loss for EMUFormer-B2 on the Cityscapes and NYUv2 datasets. On Cityscapes, training with GNLL loss leads to the best performance across the board, especially with regard to the RMSE for monocular depth estimation. GNLL loss results in a RMSE of 6.983 in comparison to 7.217 and 7.340 for MSE and Huber loss (Huber, 1992), respectively. Similarly, on NYUv2, training with GNLL loss yields the best RMSE with 0.514 versus 0.527 and 0.533 for MSE and Huber loss (Huber, 1992), although at the cost of a very slight deterioration of 0.006 in mIoU. GNLL loss leads to the highest depth uncertainty quality for both datasets.

5.6 Ablation Studies

Backbone Size. Table 9 displays a comprehensive assessment of the influence of the backbone size on Cityscapes (Cordts et al., 2016) and NYUv2 (Silberman et al., 2012). In this context, we decided to evaluate the three baseline models as a Deep Ensemble with ten members each in comparison to EMUFormer for the smallest, B0, and the biggest, B5, backbone of SegFormer (Xie et al., 2021), respectively.

More specifically, EMUFormer emerges as the top performer on all segmentation metrics, except for the mIoU where the SegFormer Deep Ensemble gives slightly better results. On the Cityscapes dataset, EMUFormer stands out by delivering the best results for all depth metrics across both backbones. Notably, it achieves this superior performance while maintaining a 20 to 30 times faster inference time compared to the Deep Ensembles. On NYUv2, the DepthFormer Deep Ensemble performs marginally better on the depth metrics, although EMUFormer remains highly competitive, especially if inference time is considered.

6 Conclusion

EMUFormer employs student-teacher distillation to achieve state-of-the-art results in joint semantic segmentation and monocular depth estimation on Cityscapes (Cordts et al., 2016) and NYUv2 (Silberman et al., 2012). Simultaneously, it estimates well-calibrated predictive uncertainties for both tasks. This is achieved without introducing any additional computational overhead during inference, making EMUFormer usable for time-critical applications. EMUFormer

Fig. 3 Representative qualitative examples of our domain-adapted EMUFormer-B2 on the Foggy Cityscapes (Sakaridis et al., 2018) (top) and Rainy Cityscapes (Hu et al., 2019) (bottom) datasets. Red rectangles are added to highlight interesting areas. Best viewed in color. Brighter colors indicate higher uncertainty

Table 8 Impact of the depth loss on the results of EMUFormer-B2 on Cityscapes (Cordts et al., 2016) and NYUv2 (Silberman et al., 2012)

	Semantic Segmentation					Monocular Depth Estimation			
	mIoU ↑	ECE ↓	p(acc/cer) ↑	p(inacc/unc) ↑	PAvPU ↑	RMSE ↓	p(acc/cer) ↑	p(inacc/unc) ↑	PAvPU ↑
Cityscapes									
MSE	0.749	0.014	0.922	**0.659**	0.810	7.217	0.742	0.446	0.761
Huber (1992)	0.748	0.013	**0.923**	0.657	0.809	7.340	0.743	0.446	0.760
GNLL	**0.752**	**0.012**	**0.923**	0.658	**0.811**	**6.983**	**0.772**	**0.491**	**0.783**
NYUv2									
MSE	**0.481**	**0.127**	**0.788**	0.690	**0.737**	0.527	0.788	0.431	0.587
Huber (1992)	**0.481**	**0.127**	**0.788**	0.689	**0.737**	0.533	0.786	0.431	0.587
GNLL	0.475	0.129	0.787	**0.692**	**0.737**	**0.514**	**0.810**	**0.440**	**0.633**

Best results are highlighted in bold

even surpasses the performance of its Deep Ensemble teacher in certain cases, despite the latter having ten times the parameters and approximately 30 times higher inference time. Most interestingly, EMUFormer achieves particularly outstanding performance in the depth estimation task in comparison to the teacher. These findings remarkably extend to out-of-domain scenarios, where EMUFormer reliably matches the Deep Ensemble teacher's overall performance, while consistently providing superior depth estimates. This success can primarily be attributed to the use of the Gaussian Negative Log-Likelihood loss (cf. Sect. 5.5), which is commonly employed to implicitly learn corresponding variances in addition to the predictive means. In the case of EMUFormer, however, the teacher model already provides high-quality variances through distillation, allowing for a more accurate approximation of the predictive means and their associated uncertainties. Overall, these findings go along nicely with previous work (Landgraf et al., 2024a; Kendall et al., 2018) on leveraging uncertainties during training, making it an interesting venue for future work. Moreover, EMUFormer

Table 9 Quantitative comparison on the Cityscapes (Cordts et al., 2016) and NYUv2 (Silberman et al., 2012) datasets between the three baseline models as Deep Ensembles (DEs) and EMUFormer with SegFormer's B0 and B5 backbone (Xie et al., 2021). The respective SegDepthFormer DE served as the teacher for the corresponding EMUFormer

		Semantic Segmentation					Monocular Depth Estimation				Inference Time [ms]
		mIoU ↑	ECE ↓	p(acc/cer) ↑	p(inacc/unc) ↑	PAvPU ↑	RMSE ↓	p(acc/cer) ↑	p(inacc/unc) ↑	PAvPU ↑	
lityscapes											
B0	SegFormer (DE)	**0.689**	0.037	0.888	0.486	0.779	–	–	–	–	273.20 ± 1.38
	DepthFormer (DE)	–	–	–	–	–	8.452	0.692	0.414	0.719	236.13 ± 0.70
	SegDepthFormer (DE)	0.651	0.045	0.912	0.634	**0.803**	8.495	0.692	0.425	0.718	317.47 ± 15.64
	EMUFormer	0.630	**0.023**	**0.924**	**0.714**	0.791	**8.086**	**0.717**	**0.473**	**0.732**	**9.58 ± 0.07**
B5	SegFormer (DE)	**0.809**	0.032	0.896	0.435	0.819	–	–	–	–	1931.01 ± 12.77
	DepthFormer (DE)	–	–	–	–	–	6.588	0.782	0.487	0.791	1892.47 ± 9.24
	SegDepthFormer (DE)	0.789	0.037	0.928	0.657	**0.852**	6.664	0.785	0.502	0.792	2018.04 ± 32.31
	EMUFormer	0.771	**0.014**	**0.934**	**0.703**	0.845	**6.157**	**0.804**	**0.536**	**0.799**	**50.72 ± 0.45**
NYUv2											
B0	SegFormer (DE)	**0.376**	0.105	0.743	0.701	0.718	–	–	–	–	315.42 ± 2.41
	DepthFormer (DE)	–	–	–	–	–	**0.642**	**0.720**	0.476	**0.566**	227.92 ± 2.39
	SegDepthFormer (DE)	0.375	0.097	**0.744**	0.703	0.718	0.678	0.693	0.466	0.553	346.21 ± 2.72
	EMUFormer	0.363	**0.090**	0.743	**0.713**	**0.720**	0.674	0.705	**0.498**	0.558	**10.04 ± 0.06**
B5	SegFormer (DE)	**0.534**	0.138	0.792	0.653	**0.744**	-	-	-	-	1958.46 ± 36.71
	DepthFormer (DE)	–	–	–	–	–	0.468	**0.852**	**0.505**	**0.647**	1875.53 ± 12.83
	SegDepthFormer (DE)	0.526	**0.133**	0.794	0.665	0.743	**0.451**	0.838	0.478	0.619	2038.26 ±13.06
	EMUFormer	0.520	0.134	**0.798**	**0.688**	**0.744**	0.476	0.846	0.467	**0.647**	**52.27 ± 1.40**

Best results are highlighted in bold

shows promising results for domain adaptation, enabling substantial performance gains with minimal fine-tuning effort. Exploring domain adaptation in the context of uncertainty quantification fills a critical gap in current research and represents an exciting opportunity to further advance this field.

Appendix A Baseline Models

Hereinafter, we present the three baseline models SegFormer (Xie et al., 2021), DepthFormer, and SegDepthFormer in more detail. For each model, we briefly describe its architecture, training criterion, and how we obtain a measurement for the uncertainty. While these models are capable of estimating the aleatoric uncertainty (Kendall & Gal, 2017; Lakshminarayanan et al., 2017), they are not able to quantify the more complete predictive uncertainty, which includes the epistemic uncertainty. For this, one of the aforementioned uncertainty quantification methods (cf. Sect. 2 of the main part of the paper) has to be used.

A.1 SegFormer

Architecture. For solely solving the semantic segmentation task, we use SegFormer (Xie et al., 2021), a modern Transformer-based architecture that stands out because of its high efficiency and performance. Thus, it is particularly suitable for real-time uncertainty quantification. As depicted in Fig. 4, SegFormer consists of two main modules: A hierarchical Transformer-based encoder that generates high-resolution coarse features and low-resolution fine features and a lightweight all-MLP segmentation decoder. The latter fuses the multi-level features of the encoder to produce a final segmentation prediction with the softmax activation function:

$$p(z) = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}, \tag{A1}$$

where $p(z)$ are the class probabilities of the softmax function that exponentiates each of the K elements of the input vector x, often referred to as logits, and then normalizes the results to obtain a probability distribution. Since SegFormer (Xie et al., 2021) only outputs logits at a $\frac{H}{4} \times \frac{W}{4}$ resolution given an input image of size $H \times W$, we use bilinear interpolation (Xie et al., 2021) before applying the softmax function on z to obtain the original resolution for the final segmentation prediction.

Training Criterion. For the objective function during training, we use the well-known categorical Cross-Entropy loss for a single image

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_{n,c} \cdot \log(p(z)_{n,c}) \ , \tag{A2}$$

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.

Aleatoric Uncertainty. We compute the predictive entropy

$$H(p(z)) = -\sum_{c=1}^{C} p(z)_c \cdot \log(p(z)_c) \ , \tag{A3}$$

which serves as the aleatoric uncertainty (Kendall & Gal, 2017).

A.2 DepthFormer

Architecture. Inspired by the efficiency and performance of SegFormer (Xie et al., 2021), we propose DepthFormer for monocular depth estimation. Fig. 4 shows that we use the same hierarchical Transformer-based encoder as SegFormer to generate high-level and low-level features. Similarly, these multi-level features are fused in an all-MLP decoder. In contrast to SegFormer, the output layer differs by having two output channels: one for the predictive mean $\mu(z)$ and one for the predictive variance $s^2(z)$ (Loquercio et al., 2020).

Predictive Mean. The first output channel uses a Rectified Linear Unit (ReLU) output activation function

$$\mu(z) = \max(0, z) \ , \tag{A4}$$

which serves as the predictive mean for monocular depth estimation.

Predictive Variance. The second output channel applies a Softplus activation

$$s^2(z) = \log(1 + e^z) \ , \tag{A5}$$

which is a smooth approximation of the ReLU function with the advantage of being differentiable, also at $z = 0$. Empirically, we found Softplus to work better than ReLU for the predictive variance, following the work by Lakshminarayanan et al. (2017).

Training Criterion. For regression tasks, neural networks typically output only a predictive mean $\mu(z)$ and the parameters are, in the most straightforward approach, optimized by minimizing the mean squared error (MSE). However, the MSE does not cover uncertainty. Therefore, we follow the approach of Nix and Weigend (1994) instead: By treating the neural networks prediction as a sample from a Gaussian distribution with the predictive mean $\mu(z)$ and corresponding predictive variance $s^2(z)$, we can minimize the Gaussian

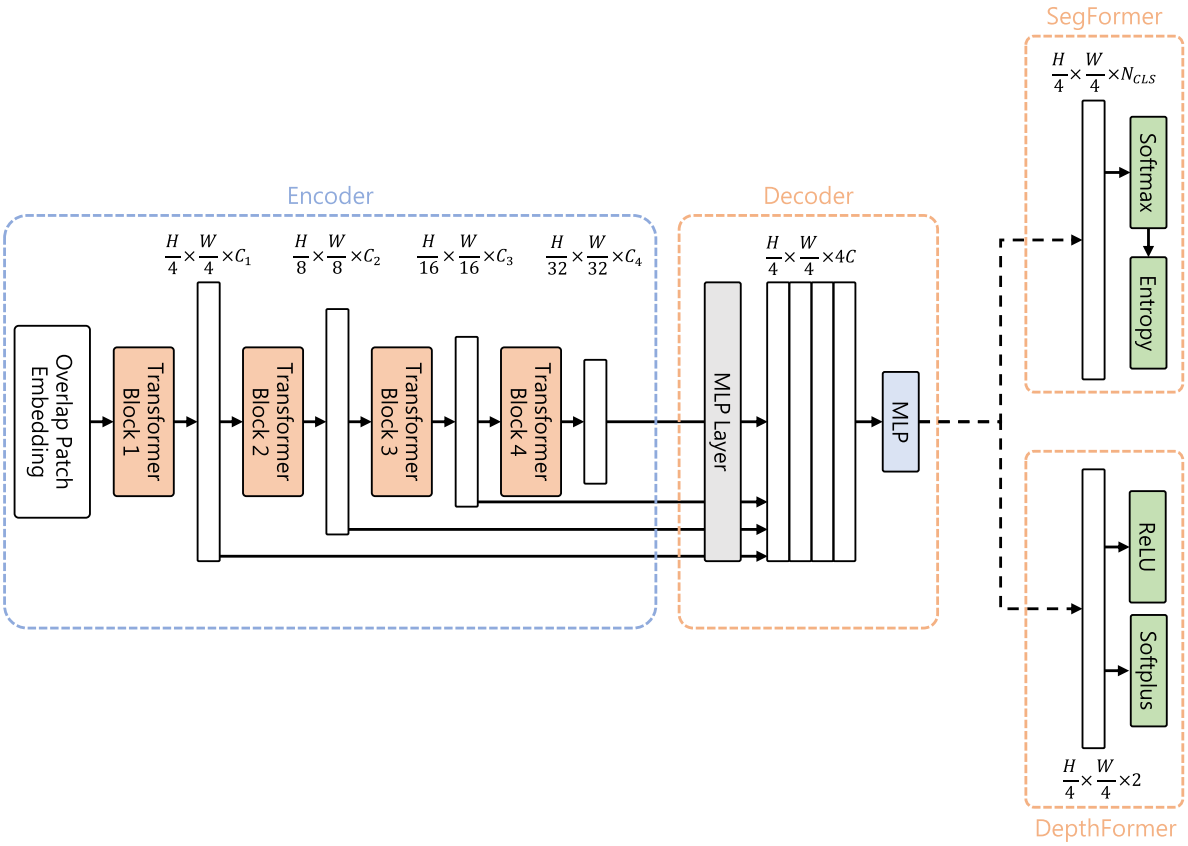

Fig. 4 A schematic overview of the SegFormer (Xie et al., 2021) and DepthFormer architectures. Both models share the same hierarchical Transformer-based encoder that generates high-resolution coarse features and low-resolution fine features, and a lightweight all-MLP segmentation decoder. They only differ in the number of output channels and in terms of output activations

Negative Log-Likelihood (GNLL) loss, which can be formulated as:

$$\mathcal{L}_{\text{GNLL}} = \frac{1}{2}\left(\frac{(y-\mu(z))^2}{s^2(z)} + \log(s^2(z))\right) , \tag{A6}$$

where y is the ground truth depth.

Aleatoric Uncertainty. Through GNLL minimization, DepthFormer does not only optimize the predictive means, but also inherently learns the corresponding variances, which can be interpreted as the aleatoric uncertainty (Kendall & Gal, 2017; Loquercio et al., 2020).

A.3 SegDepthFormer

Architecture. In order to jointly solve semantic segmentation and monocular depth estimation, we propose SegDepthFormer. The architecture shown in Fig. 5 comprises three modules: a hierarchical Transformer-based encoder, an all-MLP segmentation decoder, and an all-MLP depth decoder. The encoder and segmentation decoder are adapted from SegFormer (Xie et al., 2021) (Sect. A.1), while the depth decoder is from DepthFormer (Sect. A.2). Both decoders fuse the multi-level features obtained through the shared encoder to predict a final segmentation mask and a pixel-wise depth estimation, respectively.

Training Criterion. SegDepthFormer is trained to minimize the weighted sum of the two previously described objective functions:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + w_1\mathcal{L}_{\text{GNLL}} , \tag{A7}$$

where w_1 is a simple weighting factor. Since both loss values are of similar magnitude, we set $w_1 = 1$. However, tuning w_1 might slightly improve SegDepthFormer's performance.

Aleatoric Uncertainty. The respective aleatoric uncertainty is obtained by computing the predictive entropy $H(p(z))$ (see Eq. 5) for the segmentation task or by the predictive variance $s^2(z)$ (see Eq. A5), which is learned implicitly through the optimization of $\mathcal{L}_{\text{GNLL}}$.

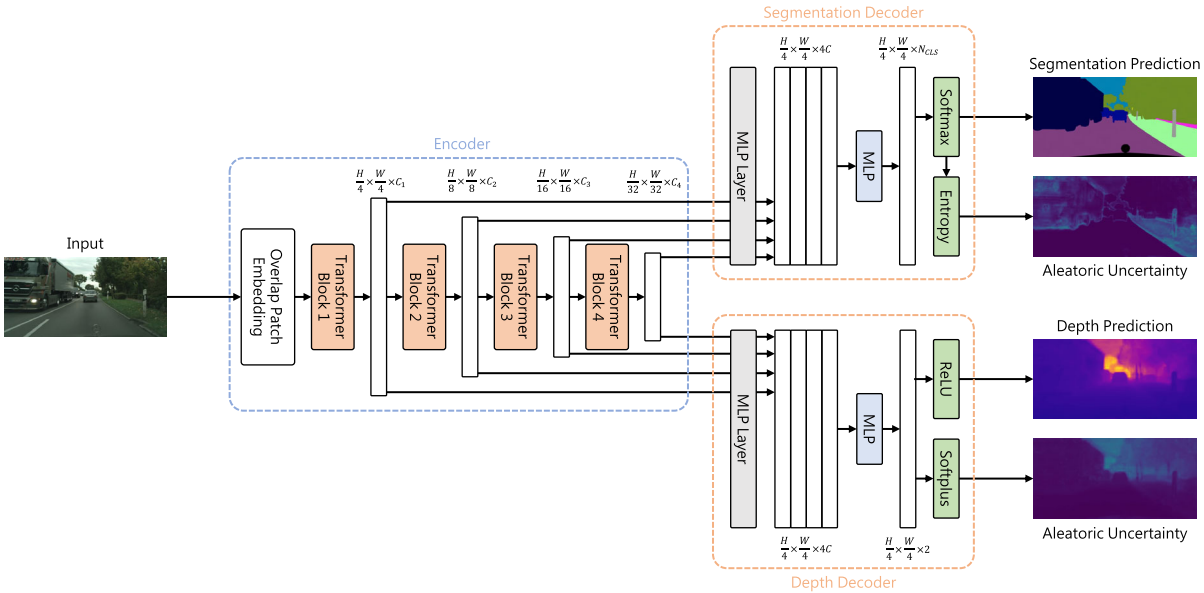

Fig. 5 A schematic overview of the SegDepthFormer architecture. The model combines the SegFormer Xie et al. (2021) architecture with a lightweight all-MLP depth decoder

Acknowledgements The authors acknowledge support by the state of Baden-Württemberg through bwHPC. This work is supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition.

Author Contributions Conceptualization: Steven Landgraf, Markus Hillemann, Markus Ulrich. Methodology: Steven Landgraf. Implementation and Experiments: Steven Landgraf, Theodor Kapler. Writing (original draft): Steven Landgraf. Writing (review): Steven Landgraf, Markus Hillemann, Theodor Kapler, Markus Ulrich. Supervision: Markus Ulrich.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability All of the utilized datasets are publicly available and referenced.

References

Amini, A., Schwarting, W., Soleimany, A., & Rus, D. (2020). Deep evidential regression. *Advances in Neural Information Processing Systems, 33*, 14927–14937.

Besnier, V., Picard, D., & Briot, A. (2021). Learning uncertainty for safety-oriented semantic segmentation in autonomous driving. In *2021 IEEE International Conference on Image Processing (ICIP)*. (pp. 3353–3357). IEEE.

Bruggemann, D., Kanakis, M., Georgoulis, S., & Van Gool, L. (2020). Automated search for resource-efficient branched multi-task networks. arXiv preprint arXiv:2008.10292.

Brüggemann, D., Kanakis, M., Obukhov, A., Georgoulis, S., & Van Gool, L. (2021). Exploring relational context for multi-task dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. (pp. 15869–15878).

Chen, L., Yang, Z., Ma, J., & Luo, Z. (2018). Driving scene perception network: Real-time joint detection, depth estimation and semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. (pp. 1283–1291). https://doi.org/10.1109/WACV.2018.00145

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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, D., Wu, L., & Shi, B. E. (2021). Iterative distillation for better uncertainty estimates in multitask emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. (pp. 3557–3566).

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. (pp. 248–255). IEEE, Miami, FL https://doi.org/10.1109/CVPR.2009.5206848

Dong, X., Garratt, M. A., Anavatti, S. G., & Abbass, H. A. (2022). Towards real-time monocular depth estimation for robotics: A survey. *IEEE Transactions on Intelligent Transportation Systems, 23*(10), 16940–16961.

Fort, S., Hu, H., & Lakshminarayanan, B. (2020). Deep Ensembles: A Loss Landscape Perspective. arXiv:1912.02757.

Gal, Y. (2016). Uncertainty in deep learning. Ph.D. thesis, University of Cambridge.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In M.F. Balcan, & K.Q. Weinberger (eds.) *Proceedings of The 33rd*

International Conference on Machine Learning. Proceedings of Machine Learning Research, (vol. 48, pp. 1050–1059). PMLR, New York, New York, USA (20–22 Jun 2016), https://proceedings.mlr.press/v48/gal16.html

Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep bayesian active learning with image data. In *International conference on machine learning*. (pp. 1183–1192). PMLR

Gao, T., Wei, W., Cai, Z., Fan, Z., Xie, S. Q., Wang, X., & Yu, Q. (2022). Ci-net: A joint depth estimation and semantic segmentation network using contextual information. *Applied Intelligence, 52*(15), 18167–18186.

Gao, T., Wei, W., Wang, X., Yu, Q., & Fan, Z. (2022). Predictive uncertainties for multi-task learning network. In *International Conference on Advanced Algorithms and Neural Networks (AANN 2022)*. (vol. 12285, pp. 294–300). SPIE.

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. (2022). A Survey of Uncertainty in Deep Neural Networks. arXiv:2107.03342

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup, & Y.W. Teh (eds.) *Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research*, (vol. 70, pp. 1321–1330). PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/guo17a.html

Gurau, C., Bewley, A., & Posner, I. (2018). Dropout distillation for efficiently estimating model confidence. arXiv preprint arXiv:1809.10562.

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

He, L., Lu, J., Wang, G., Song, S., & Zhou, J. (2021). Sosd-net: Joint semantic object segmentation and depth estimation from monocular images. *Neurocomputing, 440*, 251–263.

Heizmann, M., Braun, A., Glitzner, M., Günther, M., Hasna, G., Klüver, C., Krooß, J., Marquardt, E., Overdick, M., & Ulrich, M. (2022). Implementing machine learning: chances and challenges. *at-Automatisierungstechnik, 70*(1), 90–101.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, arXiv:1503.02531

Holder, C. J., & Shafique, M. (2021). Efficient uncertainty estimation in semantic segmentation via distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. (pp. 3087–3094).

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

Hu, Y., Chen, Z., & Lin, W. (2018). Rgb-d semantic segmentation: a review. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. (pp. 1–6). IEEE.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*. (pp. 492–518). Springer.

Ji, N., Dong, H., Meng, F., & Pang, L. (2023). Semantic segmentation and depth estimation based on residual attention mechanism. *Sensors, 23*(17), 7466.

Jiao, J., Cao, Y., Song, Y., & Lau, R. (2018). Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European conference on computer vision (ECCV)*. (pp. 53–69).

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. (pp. 5580–5590). NIPS'17, Curran Associates Inc.

Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. (pp. 7482–7491).

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems.* (Vol. 30). Curran Associates Inc.

Landgraf, S., Hillemann, M., Aberle, M., Jung, V., & Ulrich, M. (2023). Segmentation of industrial burner flames: A comparative study from traditional image processing to machine and deep learning. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **X-1/W1-2023**, 953–960 https://doi.org/10.5194/isprs-annals-X-1-W1-2023-953-2023, https://isprs-annals.copernicus.org/articles/X-1-W1-2023/953/2023/

Landgraf, S., Hilleman, M., Kapler, T., & Ulrich, M. (2024). Evaluation of multi-task uncertainties in joint semantic segmentation and monocular depth estimation. In *Forum Bildverarbeitung 2024*. p. 147. KIT Scientific Publishing.

Landgraf, S., Hillemann, M., Kapler, T., & Ulrich, M. (2024). Efficient multi-task uncertainties for joint semantic segmentation and monocular depth estimation. In *DAGM German Conference on Pattern Recognition*. Springer.

Landgraf, S., Hillemann, M., Wursthorn, K., & Ulrich, M. (2024). Uncertainty-aware cross-entropy for semantic segmentation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 10*, 129–136.

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.

Lee, K., Lee, H., Lee, K., & Shin, J. (2018). Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. arXiv:1711.09325

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports, 7*(1), 17816. https://doi.org/10.1038/s41598-017-17876-z

Lin, X., Sánchez-Escobedo, D., Casas, J. R., & Pardàs, M. (2019). Depth estimation and semantic segmentation from a single rgb image using a hybrid convolutional neural network. *Sensors, 19*(8), 1795.

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, J., Wang, Y., Li, Y., Fu, J., Li, J., & Lu, H. (2018). Collaborative deconvolutional neural networks for joint depth estimation and semantic segmentation. *IEEE transactions on neural networks and learning systems, 29*(11), 5655–5666.

Liu, S., Johns, E., & Davison, A. J. (2019). End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (pp. 1871–1880).

Loquercio, A., Segu, M., & Scaramuzza, D. (2020). A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters, 5*(2), 3153–3160.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

MacKay, D. J. C. (1992). A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation, 4*(3), 448–472. https://doi.org/10.1162/neco.1992.4.3.448

Malinin, A., Mlodozeniec, B., & Gales, M. (2019). Ensemble Distribution Distillation. arXiv:1905.00076.

McAllister, R., 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. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. (pp. 4745–4753). International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia.https://doi.org/10.24963/ijcai.2017/661

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., & others. (2017). Mixed precision training. arXiv preprint arXiv:1710.03740.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968

Ming, Y., Meng, X., Fan, C., & Yu, H. (2021). Deep learning for monocular depth estimation: A review. *Neurocomputing, 438*, 14–33.

Mousavian, A., Pirsiavash, H., & Košecká, J. (2016). Joint semantic segmentation and depth estimation with deep convolutional networks. In *2016 Fourth International Conference on 3D Vision (3DV)*. (pp. 611–619). IEEE.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (pp. 24384–24394).

Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*. vol. 29.

Nekrasov, V., Dharmasiri, T., Spek, A., Drummond, T., Shen, C., & Reid, I. (2019). Real-time joint semantic segmentation and depth estimation using asymmetric annotations. In *2019 International Conference on Robotics and Automation (ICRA)*. (pp. 7101–7107). IEEE.

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

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. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems.* (Vol. 32). Curran Associates, Inc.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015). FitNets: Hints for Thin Deep Nets. arXiv:1412.6550.

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

Sakaridis, C., Dai, D., & Van Gool, L. (2021). Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*. (pp. 10765–10775).

Schwaiger, A., Sinhamahapatra, P., Gansloser, J., & Roscher, K. (2020). Is uncertainty quantification in deep learning sufficient for out-of-distribution detection? *Aisafety@ ijcai, 54*, 1–8.

Shen, Y., Zhang, Z., Sabuncu, M. R., & Sun, L. (2021). Real-time uncertainty estimation in computer vision via uncertainty-aware distribution distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. (pp. 707–716).

Shrivastava, A., Gupta, A., & Girshick, R. (2016). Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. (pp. 761–769).

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*. (pp. 746–760). Springer.

Simpson, I. J., Vicente, S., & Campbell, N. D. (2022). Learning structured gaussians to approximate deep ensembles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (pp. 366–374).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(56), 1929–1958.

Steger, C., Ulrich, M., & Wiedemann, C. (2018). *Machine Vision Algorithms and Applications*. John Wiley & Sons.

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

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

Vandenhende, S., Georgoulis, S., & Van Gool, L. (2020). Mti-net: Multi-scale task interaction networks for multi-task learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. (pp. 527–543). Springer.

Wang, C., Wang, C., Li, W., & Wang, H. (2021). A brief survey on rgb-d semantic segmentation using deep learning. *Displays, 70*, Article 102080.

Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing, 312*, 135–153.

Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., & Yuille, A. L. (2015). Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. (pp. 2800–2809).

Wursthorn, K., Hillemann, M., & Ulrich, M. (2022). Comparison of uncertainty quantification methods for CNN-based regression. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLIII-B2-2022**, 721–728.

Wursthorn, K., Hillemann, M., & Ulrich, M. (2024). Uncertainty quantification with deep ensembles for 6d object pose estimation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 10*, 223–230.

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.

Xu, D., Ouyang, W., Wang, X., & Sebe, N. (2018). Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (pp. 675–684).

Xu, X., Zhao, H., Vineet, V., Lim, S. N., & Torralba, A. (2022). Mtformer: Multi-task learning via transformer and cross-task reasoning. In *European Conference on Computer Vision*. (pp. 304–321). Springer.

Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering,34*(12), 5586–5609.