PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science (2024) 92:101–114
https://doi.org/10.1007/s41064-024-00280-4

ORIGINAL ARTICLE

DUDES: Deep Uncertainty Distillation using Ensembles for Semantic Segmentation

Steven Landgraf[1] · Kira Wursthorn[1] · Markus Hillemann[1] · Markus Ulrich[1]

Received: 27 November 2023 / Accepted: 1 February 2024 / Published online: 25 March 2024

Abstract

The intersection of deep learning and photogrammetry unveils a critical need for balancing the power of deep neural networks with interpretability and trustworthiness, especially for safety-critical application like autonomous driving, medical imaging, or machine vision tasks with high demands on reliability. Quantifying the predictive uncertainty is a promising endeavour to open up the use of deep neural networks for such applications. Unfortunately, most current available methods are computationally expensive. In this work, we present a novel approach for efficient and reliable uncertainty estimation for semantic segmentation, which we call **D**eep **U**ncertainty **D**istillation using **E**nsembles for **S**egmentation (DUDES). DUDES applies student-teacher distillation with a Deep Ensemble to accurately approximate predictive uncertainties with a single forward pass while maintaining simplicity and adaptability. Experimentally, DUDES accurately captures predictive uncertainties without sacrificing performance on the segmentation task and indicates impressive capabilities of highlighting wrongly classified pixels and out-of-domain samples through high uncertainties on the Cityscapes and Pascal VOC 2012 dataset. With DUDES, we manage to simultaneously simplify and outperform previous work on Deep-Ensemble-based Uncertainty Distillation.

Keywords Deep Learning · Semantic Segmentation · Uncertainty Quantification · Deep Ensemble · Knowledge Distillation

1 Introduction

In recent years, approaches based on deep neural networks have become the most popular and successful solution for semantic segmentation problems (Minaee et al. 2022). Despite their unrivaled performance on established benchmark datasets like Cityscapes (Cordts et al. 2016) or PASCAL VOC (Everingham et al. 2010), neural networks lack interpretability (Gawlikowski et al. 2022), are unable to distinguish between in-domain and out-of-domain samples (Lee et al. 2018), and tend to be overconfident (Guo et al. 2017). These shortcomings are especially severe for safety-critical applications like autonomous driving (McAllister et al. 2017) and the analysis of medical imaging (Leibig et al. 2017) or computer vision tasks that have high demands on reliability like industrial inspection (Steger et al. 2018) and automation (Ulrich and Hillemann 2023).

Quantifying the predictive uncertainty is a promising endeavour to make such applications safer and more reliable, e.g., by preemptively making risk-averse predictions or by providing feedback to a human operator when predictions are uncertain. Some of the most relevant methods include Bayesian Neural Networks (MacKay 1992), Monte Carlo Dropout (Gal and Ghahramani 2016), and Deep Ensembles (Lakshminarayanan et al. 2017). Unfortunately, most methods require a computationally expensive estimation of a distribution of outputs by sampling from a stochastic process. Recently, the concept of knowledge distillation has been introduced as a potential solution (Shen et al. 2021; Besnier et al. 2021; Holder and Shafique 2021; Simpson et al. 2022). Knowledge distillation is a technique for transferring the knowledge embodied in a complex model, re-

✉ Steven Landgraf
steven.landgraf@kit.edu

Kira Wursthorn
kira.wursthorn@kit.edu

Markus Hillemann
markus.hillemann@kit.edu

Markus Ulrich
markus.ulrich@kit.edu

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

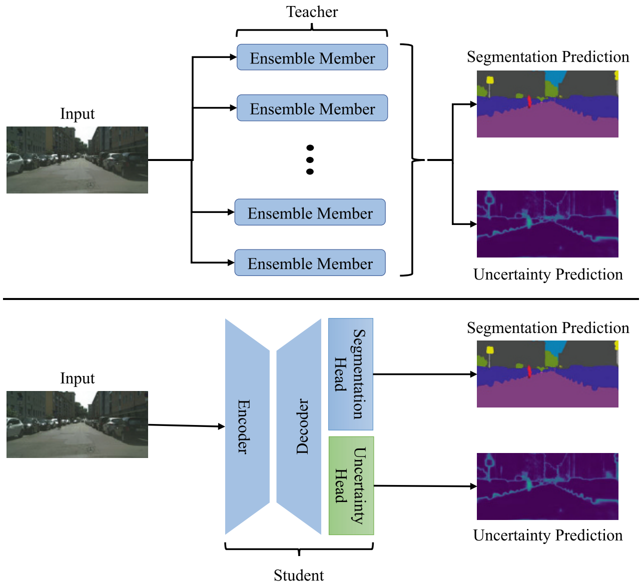

Fig. 1 DUDES applies student-teacher distillation with a Deep Ensemble (DE) to accurately approximate predictive uncertainties with a single forward pass while maintaining simplicity and adaptability

ferred to as the teacher, to a smaller model, referred to as the student. By incorporating the knowledge learned by a more complex model, the student's performance can be enhanced (Hinton et al. 2015; Romero et al. 2015; Malinin et al. 2019).

In this work, we present a novel approach for efficient and reliable uncertainty quantification, which we call **D**eep **U**ncertainty **D**istillation using **E**nsembles for **S**egmentation (DUDES) as shown in Fig. 1. DUDES applies student-teacher distillation with a Deep Ensemble to accurately approximate predictive uncertainties while maintaining simplicity and adaptability. In comparison to the Deep Ensemble teacher, the student only needs a single forward pass to obtain predictive uncertainties, which massively reduces the inference time and eliminates the computational overhead that is associated with having to deal with multiple models and forward passes. DUDES simultaneously simplifies and outperforms previous work on Deep-Ensemble-based uncertainty distillation, which we experimentally evaluate on the Cityscapes and Pascal VOC 2012 dataset.

After an overview of the related work on uncertainty quantification and knowledge distillation in Sect. 2, the methodology of DUDES is described in Sect. 3. In Sect. 4, we demonstrate the ability of DUDES through quantitative and qualitative analysis and investigate the potential for identifying wrongly classified pixels or out-of-domain samples with the help of the predictive uncertainties qualitatively. Thereafter, we provide an extended set of experiments with a Vision-Transformer-based architecture, a different uncertainty quantification method for the teacher, and a different dataset to demonstrate the generalizability of DUDES in Sect. 5. Sect. 6 discusses the experimental results and their potential impact on future research. Sect. 7 concludes the paper.

2 Related Work

In this section, we summarize the related work on uncertainty quantification and knowledge distillation.

2.1 Uncertainty Quantification

Deep neural networks consist of a large number of model parameters and include non-linearities, which generally makes the exact posterior probability distribution of a network's output prediction intractable (Blundell et al. 2015; Loquercio et al. 2020). This leads to approximate uncertainty quantification approaches including softmax probability, Bayesian techniques like Bayesian Neural Networks (MacKay 1992), Monte Carlo Dropout (Gal and Ghahramani 2016), and Deep Ensembles (Lakshminarayanan et al. 2017).

While the softmax predictions are easy to implement, they tend to be overconfident and need to be calibrated in order to produce reliable confidence predictions where the predicted probability and the actual likelihood are in agreement (Guo et al. 2017). Additionally, softmax predictions are often erroneously interpreted as model confidence (Gal and Ghahramani 2016). A mathematically sound approach based on Bayesian inference is provided by Bayesian Neural Networks, which transform a deterministic network into a stochastic one. This is done by placing probability distributions over the activations and/or weights (Jospin et al. 2022). For example, Bayes by Backprop (Blundell et al. 2015) uses variational inference to learn approximate distributions over the weights. At test time, weights are sampled from the learned distributions, resulting in an ensemble of models that is used to sample from the posterior distribution over the predictions. To overcome the high computational cost of Bayesian Neural Networks, Gal and Ghahramani (2016) propose Monte Carlo Dropout as an approximation of a stochastic Gaussian process using a common regularization method. While dropout regularization (Srivastava et al. 2014) is usually only used during training, Monte Carlo Dropout (Gal and Ghahramani 2016) applies this technique to sample from the posterior distribution of the predictions at test time.

The uncertainties produced by Monte Carlo Dropout are not calibrated (Gal and Ghahramani 2016), which is a major drawback that is overcome by Deep Ensembles (Lakshminarayanan et al. 2017) where an ensemble of trained models produces samples of predictions at test time. Random weight initialization and diverse data augmentations across ensemble members introduce randomness in Deep Ensembles, enabling exploration of diverse modes in function space (Fort et al. 2020). This characteristic contributes to their reputation for being well-calibrated (Lakshminarayanan et al. 2017) and establishes them as the state-of-the-art in uncertainty quantification (Ovadia et al. 2019; Gustafsson et al. 2020; Wursthorn et al. 2022). Besides their outstanding ability to quantify high-quality uncertainties, Deep Ensembles are also a popular method to improve the prediction quality itself (Marmanis et al. 2016; Nigam et al. 2018; Kang and Gwak 2019; Thanh et al. 2020; Lumini et al. 2021; Nanni et al. 2023).

Next to these approximate uncertainty quantification methods, there has also been an increasing interest in using deterministic single forward-pass methods, which need less memory and have a lower inference time. For instance, Van Amersfoort et al. (2020) and Liu et al. (2020) build on the idea of a well-regularized feature space in which they quantify the uncertainty through distance-aware output layers. Although these methods perform well, they are not quite competitive with Deep Ensembles and require a substantial adaptation of the training process. Mukhoti et al. (2023) propose to simplify the beforementioned approaches by using Gaussian Discriminant Analysis post-training for feature-space density estimation. With their approach, they manage to perform on par with a Deep Ensemble in some settings but still require a more sophisticated training approach. In general, these deterministic single-forward pass methods are a worthwhile alternative to the traditional uncertainty quantification methods MacKay (1992); Gal and Ghahramani (2016); Lakshminarayanan et al. (2017), yet they all introduce conceptual complexity that require changes in the architecture, the training process, and introduce additional hyperparameters.

2.2 Knowledge Distillation

Knowledge distillation is a technique for transferring the knowledge embodied in a complex model, referred to as the teacher, to a usually smaller model, referred to as the student. The teacher can be a model with a large number of parameters or even a Deep Ensemble. The student is trained to imitate the predictions of the teacher on a given dataset, with the goal of minimizing the difference of the student's outputs and the teacher's outputs. By incorporating the knowledge learned by a more complex model, the student's performance can be enhanced. Usually, this results in a more compact student model that achieves similar performance compared to the teacher model (Hinton et al. 2015; Romero et al. 2015; Malinin et al. 2019).

Recently, the concept of knowledge distillation has attracted increasing interest in the context of efficient uncertainty quantification to enable real-time uncertainty estimation (Shen et al. 2021; Besnier et al. 2021; Holder and Shafique 2021; Simpson et al. 2022). For instance, Shen et al. (2021) have used student-teacher distillation for real-time uncertainty quantification based on Monte Carlo Dropout (Gal and Ghahramani 2016). Holder and Shafique (2021) are the first to use Deep Ensembles, generally regarded as the most powerful uncertainty quantification method (cf. Sect. 2.1), as the teacher for efficient uncertainty quantification and out-of-domain detection for semantic segmentation. While this is a logical step towards higher quality uncertainties in real-time environments, their approach requires a custom segmentation and uncertainty head, and they introduce two additional losses with three new hyperparameters that determine the smoothness of the softmax probability distribution and the loss weights. This makes their method rather difficult to implement and especially costly to adapt to new applications since these hyperparameters might need to be tuned. Besides, their student suffers a severe degradation in terms of segmentation performance in comparison to the teacher and systematically underestimates uncertainties for classes with high uncertainties and vice versa.

We believe that the process of ensemble-based uncertainty distillation can be improved upon by simplification. Instead of distilling the entire uncertainty map, which is what Holder and Shafique (2021) propose, we only regard the uncertainty of the respective predicted class, i.e. the predictive uncertainty. This basic, yet highly effective, simplification ensures that the student's segmentation performance is not degraded and the corresponding uncertainties can be learned more easily. As a result, we manage to train a student model that achieves similar or better segmentation performance than the Deep Ensemble teacher and does not suffer from any systematic shortcomings with regard to the uncertainty quantification. Additionally, DUDES does not rely on custom segmentation or uncertainty head architectures and introduces only a single uncertainty loss without hyperparameters. Thereby, we provide a distinct improvement over all of the shortcomings and complexities of previous work.

3 Methodology

In the following, we provide an overview of DUDES, explain the methodology behind our uncertainty distillation approach, and lay out the implementation details.

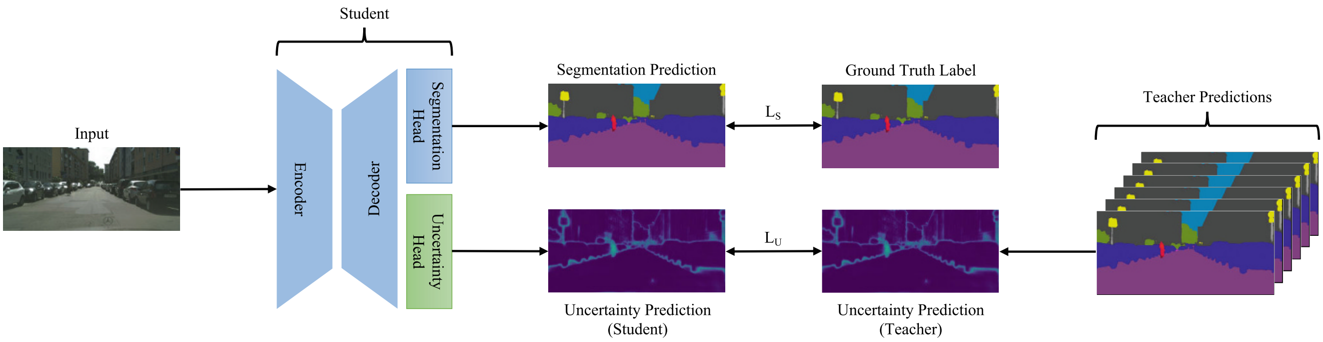

Fig. 2 A schematic overview of the training process of the student model of DUDES. DUDES is an easy-to-adapt framework for efficiently estimating predictive uncertainty through student-teacher distillation. The student model simultaneously outputs a segmentation prediction alongside a corresponding uncertainty prediction. Training the student involves a regular segmentation loss with the ground truth labels and an additional uncertainty loss. As ground truth uncertainties, we compute the predictive uncertainty of a Deep Ensemble, thereby acting as the teacher

3.1 Overview

DUDES is an easy-to-adapt framework for efficient and reliable uncertainty quantification through student-teacher distillation. The overall goal is to train a student model that can simultaneously output a segmentation prediction and a corresponding predictive uncertainty in the form of standard deviations that correlate with wrongly classified or out-of-domain pixels with a single forward pass as shown in Fig. 1. Although the student and the teacher could be trained jointly, in principle, we propose a two-step framework for the sake of simplicity and computational constraints:

1. Training the teacher with the ground truth labels
2. Training the student with the ground truth labels and the teacher's uncertainty predictions

As shown in Fig. 2, the training of the student model consists of two loss components. The first component L_S assesses the dissimilarity between the student's segmentation prediction and the ground truth labels, while the second component L_U evaluates the disparity between the student's uncertainty prediction and the output of one of the uncertainty quantification methods described in Sect. 2.1. As mentioned before, we propose to use a Deep Ensemble as the teacher for the concrete implementation of DUDES. Deep Ensembles are simple to implement, easily parallelizable, require little tuning, and represent the current state-of-the-art uncertainty quantification method (Ovadia et al. 2019; Gustafsson et al. 2020; Wursthorn et al. 2022). Nevertheless, since DUDES is flexible with regards to the chosen uncertainty quantification method, the Deep Ensemble can simply be replaced by any other uncertainty quantification method as long as the resulting uncertainty measure is limited between 0 and 1, which we will show in Sect. 5.

Teacher. For the reasons stated above, we use a Deep Ensemble as the teacher. The Deep Ensemble consists of ten (cf. Sect. 4.5) regular semantic segmentation models that are not pre-trained, thus following prior work on Deep-Ensemble-based uncertainty quantification (Lakshminarayanan et al. 2017; Fort et al. 2020). By randomly initializing all the parameters before training, we aim to capture different aspects of the input data distribution for each ensemble member, boosting the teacher's overall performance, robustness, and uncertainty quantification capabilities. During inference, each ensemble member produces slightly different predictions, enabling the calculation of a mean segmentation prediction and an uncertainty prediction. In our case, we decided to use the softmax standard deviation as a measure of the respective uncertainty.

Student. As our student has to output a corresponding predictive uncertainty in addition to the segmentation prediction, we add a second head to the segmentation model's decoder. We propose to use an additional uncertainty head that is identical to the regular segmentation head of the segmentation model, except for the output layer. For the segmentation head, we use a softmax activation to obtain class-wise probabilities. Whereas for the uncertainty head, we use a sigmoid activation that limits the outputs between 0 and 1. Our uncertainty head only needs one output channel instead of the number of classes, as needed by the segmentation head. Since this is a key modification to improve upon previous work, we will discuss this simplification in detail in Sect. 6. In contrast to the randomly initialized ensemble members, the student's parameters are initialized with ImageNet pre-training (Deng et al. 2009) to drastically reduce the required training time as shown in Sect. 4.5.

3.2 Uncertainty Distillation

To efficiently estimate the predictive uncertainty of the Deep Ensemble with a single student model, we utilize student-teacher distillation as Fig. 2 shows.

Segmentation Loss. The main objective function that is being minimized for the segmentation task is the well-known categorical cross-entropy loss:

$$L_S = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_{n,c} \log(p_{n,c}(x)), \qquad (1)$$

where L_S is the segmentation 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 respective 1-hot encoded ground truth label, and $p_{n,c}(x)$ is the respective predicted probability based on the input image x. The categorical cross-entropy loss measures the dissimilarity between the ground truth probability distribution and the predicted probability distribution. By minimizing this loss during training, the model is encouraged to produce pixel-wise class predictions that are as close as possible to the ground truth classes.

Uncertainty Loss. To distill the predictive uncertainties of our teacher into the student, we introduce an additional uncertainty loss, which is formulated as the root mean squared logarithmic error (RMSLE)

$$L_U = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (\log(z_n(x) + 1) - \log(q_n(x) + 1))^2}, \qquad (2)$$

where L_U is the uncertainty loss for a single image, N is the number of pixels in the image, $z_n(x)$ is the teacher's predictive uncertainty for the n-th pixel as ground truth, and $q_n(x)$ is the corresponding student's uncertainty prediction based on the input image x. The teacher's predictive uncertainty z_i represents the standard deviation of the softmax probabilities of the predicted class in the segmentation map. By minimizing the RMSLE during training, the student is encouraged to produce uncertainty estimates that are as close as possible to the teacher's uncertainties. The natural logarithm provides special attention to the pixels where uncertainties are higher by penalizing underestimations more than overestimations.

Total Loss. The total loss is the sum of both individual losses:

$$L = L_S + L_U. \qquad (3)$$

For the sake of simplicity and because of the empirical results, which we will demonstrate in Sects. 4 and 5, we refrain from introducing additional hyperparameters to weight the individual losses. However, it is worth mentioning that, depending on the application, the introduction of weights for the individual loss terms could be valuable.

4 Evaluation of Performance

In this section, we describe a variety of experiments that demonstrate the advantages of DUDES. Firstly, we go over our experimental setup. Secondly, we compare the student and the teacher quantitatively. More specifically, we examine the class-wise segmentation performance as well as the class-wise uncertainties. In addition, we investigate the uncertainty quality and highlight the substantial difference in terms of inference time and trainable parameters between the teacher and the student model. Thirdly, we evaluate the student's predictions qualitatively. Fourthly, we assess how well our student model performs on out-of-domain (OOD) datasets in comparison to the teacher. Lastly, we provide two ablation studies, which explore the influence of the number of ensemble members and analyze the impact of pre-training.

4.1 Experimental Setup

Architecture. For our baseline semantic segmentation model, we use a DeepLabv3+ (Chen et al. 2018) as the decoder and a ResNet-18 (He et al. 2016) as the backbone because they both are very commonly used architectures for semantic segmentation. All ensemble members are trained with just the segmentation loss of Eq. (1) and generally follow the training procedure of the student with regards to data augmentations and hyperparameters.

Training. To prevent overfitting, we apply the following data augmentation strategy to all training procedures:

1. Random scaling with a scaling factor between 0.5 and 2.0,
2. Random cropping with the crop size of 768 × 768,
3. Random horizontal flipping with a flip chance of 50%.

Besides, we employ a Stochastic Gradient Descent (SGD) optimizer (Robbins and Monro 1951) with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005 as optimizer-specific hyperparameters. In all experiments, the decoder's learning rate is ten times higher than that of the backbone. Additionally, we use polynomial learning rate scheduling to decay the initial learning rate during the training process:

$$\mathrm{lr} = \mathrm{lr}_{\mathrm{initial}} \cdot \left(1 - \frac{\text{iteration}}{\text{total iterations}}\right)^{0.9}, \qquad (4)$$

where lr is the current learning rate, and $\mathrm{lr}_{\mathrm{initial}}$ is the initial learning rate. In all training processes, we train for 200 epochs with a batch size of 16 on a NVIDIA A100 GPU. We empirically found this to be sufficient for the models to converge and did not employ any early stopping techniques.

	Road	Sidewalk	Building	Wall	Fence	Pole	Tr. Light	Tr. Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorbike	Bicycle	Mean
Teacher IoU (Theirs)	96.1	79.4	91.4	43.2	56.3	58.4	62.0	73.0	91.7	59.6	93.7	78.2	55.3	93.5	66.8	79.3	67.7	53.4	74.3	72.3
Student IoU (Theirs)	96.4	77.2	90.0	42.6	54.7	47.6	51.1	65.2	90.4	56.4	91.9	73.9	49.8	92.1	61.7	72.3	62.5	49.3	68.9	68.1
Teacher IoU (Ours)	97.8	82.2	90.7	50.4	54.5	54.9	57.8	69.3	91.5	62.7	94.3	75.4	53.5	93.2	69.6	75.9	64.0	47.6	69.5	71.3
Student IoU (Ours)	98.0	83.5	91.4	46.7	55.7	59.1	63.3	73.3	91.8	63.1	94.2	79.0	57.9	93.9	74.7	83.8	69.4	50.1	73.6	73.8
Difference (Theirs) ↑	**0.3**	-2.2	-1.4	**-0.6**	-1.6	-10.8	-10.9	-7.8	-1.3	-3.2	-1.8	-4.3	-5.5	-1.4	-5.1	-7.0	-5.2	-4.1	-5.4	-4.2
Difference (Ours) ↑	0.2	**1.3**	**0.7**	-3.7	**1.2**	**4.2**	**5.5**	**4.0**	**0.3**	**0.4**	**-0.1**	**3.6**	**4.4**	**0.7**	**5.1**	**7.9**	**5.4**	**2.5**	**4.1**	**2.5**

Fig. 3 Quantitative comparison between the student's and the teacher's class-wise Intersection over Union (IoU). Higher IoU values denote better segmentation results, which are preferred. For the difference, the teacher's results are subtracted from the student's results. Results of Holder and Shafique (2021) are indicated by "Theirs"

	Road	Sidewalk	Building	Wall	Fence	Pole	Tr. Light	Tr. Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorbike	Bicycle	Mean
Teacher Unc. (Theirs)	0.029	0.097	0.055	0.210	0.147	0.100	0.128	0.108	0.028	0.129	0.030	0.068	0.125	0.030	0.176	0.155	0.257	0.165	0.082	0.111
Student Unc. (Theirs)	0.032	0.086	0.077	0.155	0.141	0.133	0.135	0.111	0.055	0.127	0.046	0.097	0.127	0.046	0.108	0.100	0.127	0.135	0.130	0.104
Teacher Unc. (Ours)	0.024	0.064	0.038	0.150	0.165	0.100	0.142	0.102	0.025	0.101	0.031	0.109	0.120	0.043	0.175	0.163	0.195	0.158	0.108	0.106
Student Unc. (Ours)	0.018	0.065	0.035	0.144	0.160	0.128	0.112	0.097	0.027	0.126	0.025	0.117	0.105	0.038	0.200	0.144	0.171	0.190	0.150	0.108
Difference (Theirs) ↓	0.003	-0.011	0.022	-0.055	-0.006	0.033	0.007	0.003	0.027	0.002	0.016	0.029	0.002	0.016	-0.068	-0.055	-0.130	-0.030	0.048	-0.007
Difference (Ours) ↓	-0.006	0.001	-0.003	-0.006	-0.005	0.028	-0.03	-0.005	0.002	0.025	-0.006	0.008	-0.015	-0.005	0.025	-0.019	-0.024	0.032	0.042	0.002

Fig. 4 Quantitative comparison between the student's and the teacher's class-wise predictive uncertainties. In this case, a smaller difference is preferred as the student is trained to predict the same uncertainties as the teacher. The differences are calculated by subtracting the teacher's results from the student's results. They are highlighted based on the absolute differences being: ≤ 0.01, ≤ 0.02, ≤ 0.03, ≤ 0.04, ≤ 0.05, ≤ 0.06, ≥ 0.06. Results of Holder and Shafique (2021) are indicated by "Theirs"

Dataset. Our experiments are based on the Cityscapes dataset (Cordts et al. 2016), a freely available urban street scene dataset. It consists of 2975 training images, 500 validation images, and 1525 test images. Since the test images are not publicly available, we use the validation images for testing in all of our experiments. Each RGB image is 2048×1024 pixels in size, with each pixel assigned to one of 19 class labels or a void label. The void ground truth pixels are excluded during training and evaluation in the segmentation task, but they are used to qualitatively evaluate the uncertainty outputs as they indicate the model's ability to distinguish between in-domain and out-of-domain samples. Additionally, we test our student model and teacher ensemble on Foggy Cityscapes (Sakaridis et al. 2018) and Rain Cityscapes (Hu et al. 2019) to investigate the potential of DUDES for out-of-domain detection.

Metrics. For quantitative evaluations, we primarily report the mean Intersection over Union (mIoU), also known as the Jaccard Index to measure the quality of the segmentation prediction. In addition, we use the Expected Calibration Error (ECE) (Naeini et al. 2015) to evaluate the calibration of the softmax probabilities. Lastly, we report the mean class-wise predictive uncertainty (mUnc) (Holder and Shafique 2021) to compare the student's uncertainty with that of the teacher .

4.2 Quantitative Evaluation

Figs. 3 and 4 outline a quantitative comparison between the student's and the teacher's Intersection over Union (IoU) as well as their predictive uncertainties. The results of Holder and Shafique (2021) have been included as they are the most relevant previous work on Deep-Ensemble-based student-teacher distillation for efficient uncertainty quantification. Their teacher is based on 25 DeepLabv3+ models with a MobileNet backbone (Howard et al. 2017), whereas our teacher consists of 10 DeepLabv3+ models with a ResNet-18 backbone. The MobileNet backbone and our ResNet-18 backbone have been shown to have very similar performance (Bianco et al. 2018). Both students are initialized with ImageNet pre-training (Deng et al. 2009) and evaluated on the Cityscapes validation dataset (Cordts et al. 2016).

Segmentation Prediction. As shown in Fig. 3, our student network outperforms the teacher on the segmentation task for all classes except for *wall* and *sky*, with an average improvement of 2.5% in mIoU. We attribute this improvement to the student's ImageNet pre-training as compared to the randomly initialized ensemble members of the teacher. In comparison, the student by Holder and Shafique (2021) showed a mIoU deterioration of 4.2%.

Uncertainty Prediction. Fig. 4 shows that our student approximates the teacher's uncertainties very accurately: In 10 out of the 19 classes our student's class-wise uncertainties deviate by less than 0.01 compared to that of the

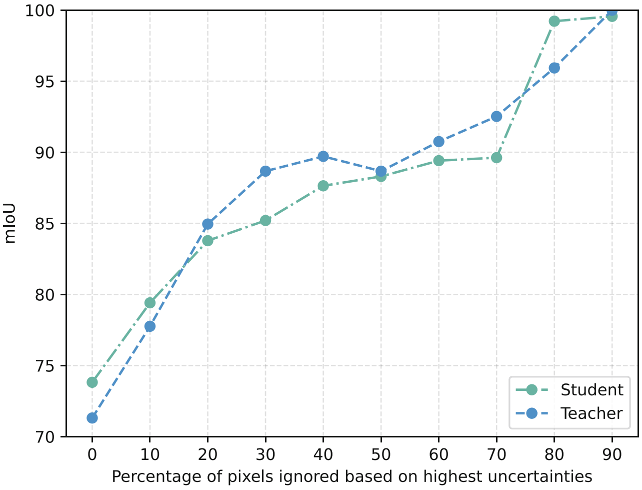

Fig. 5 Comparison between the student's and the teacher's mean Intersection over Union (mIoU). We progressively ignore an increasing percentage of pixels in the segmentation prediction and simultaneously re-evaluated the mIoU. The pixels are sorted based on their predictive uncertainty in descending order, thus removing the most uncertain segmentation predictions first

Table 1 Comparison of the inference time for a single image in milliseconds and the number of trainable parameters between the baseline, the teacher, and the student model. The inference time and corresponding standard deviation are based on 25 forward passes

	Inference time [ms]	Trainable Parameters
Baseline	18.3 ± 0.4	12,333,923
Teacher	217.1 ± 0.8	123,339,230
Student	18.5 ± 0.4	12,334,180

teacher. Our student manages to deviate by less than 0.03 in 17 out of the 19 classes, with a maximum deviation of 0.042 for the *bicycle* class. On the other hand, the student by Holder and Shafique (2021) deviates by less than 0.01 in 5 out of the 19 classes and by less than 0.03 in only 13 out of the 19 classes. Their student's maximum difference is 0.130 for the *train* class. On average across all classes, the uncertainties of both students deviate only slightly from those of the teachers, with our student model deviating by 0.002 and the student by Holder and Shafique (2021) deviating by -0.007. Generally speaking, both students struggle with accurately approximating the teacher's uncertainties for the last five classes: *Truck*, *bus*, *train*, *motorbike*, and *bicycle*. For these classes, our student has an average absolute deviation of 0.028, while their student (Holder and Shafique 2021) deviates by 0.066.

Fig. 5 displays another comparison between the student's and the teacher's ability to approximate reliable uncertainties: For this analysis, we progressively ignored an increasing percentage of uncertain pixels in the segmentation prediction and simultaneously re-evaluated the mIoU. For this, the pixels were sorted based on their predictive uncertainty in descending order. This initially removes the pixels with the most uncertain segmentation predictions from the evaluation until only the pixels with the most certain predictions are left. Consequently, meaningful uncertainties should result in a monotonically increasing function.

As Fig. 5 shows, the student as well as the teacher experience an almost linear rise in mIoU from 73.8% and 71.3%, respectively, to almost 100% after removing 90% of the most uncertain pixels. Both models attain a similar relative increase in mIoU by disregarding the first 10% of the most uncertain pixels. Up until ignoring 70% of the pixels, the teacher reaches a mIoU of 92.5%, while the student only attains 89.6%. Beyond this point, the student's mIoU surpasses that of the teacher, with the student achieving 99.2% after ignoring 80% of the pixels with the highest uncertainties, while the teacher only reaches 95.9%. This analysis yields two key findings: Firstly, predictive uncertainties prove to be related to the correctness of the prediction and hence provide an effective approach of identifying misclassified pixels. Secondly, our student's predictive uncertainties deviate only slightly from the teacher's uncertainties, revealing that they are equally meaningful.

Inference Time. Table 1 compares the inference time for a single image and the number of trainable parameters between the baseline, the teacher, and the student model. The experiment was conducted on a common NVIDIA GeForce RTX 3090 GPU with 24GB of memory. Obviously, there is only an insignificant difference of 0.2 milliseconds in inference time between the baseline and the student, despite the student's ability to output an additional predictive uncertainty. Furthermore, the student's inference is roughly 11.7 times faster than that of the teacher. The number of trainable parameters shows the efficiency of the student network. The additional uncertainty head of the student network only adds 257 parameters to the baseline model.

4.3 Qualitative Evaluation

Fig. 6 displays four example images from the Cityscapes validation set and their corresponding ground truth labels, our student's segmentation prediction, a binary accuracy map, and the student's uncertainty prediction. The binary accuracy map visualizes incorrectly predicted pixels and void classes in white and correctly predicted pixels in black.

Visually, for large areas and well-represented classes like road, sidewalk, building, sky, and car the student's segmentation is almost free of errors. This supports the quantitative evaluation described in Fig. 4. Like most segmentation models, our student struggles with class transitions, areas with lots of inherent noise, or areas that belong to the void class, which is visualized by the binary accuracy map.

A comparison of the binary accuracy map and our student's uncertainty prediction adds to the observations laid

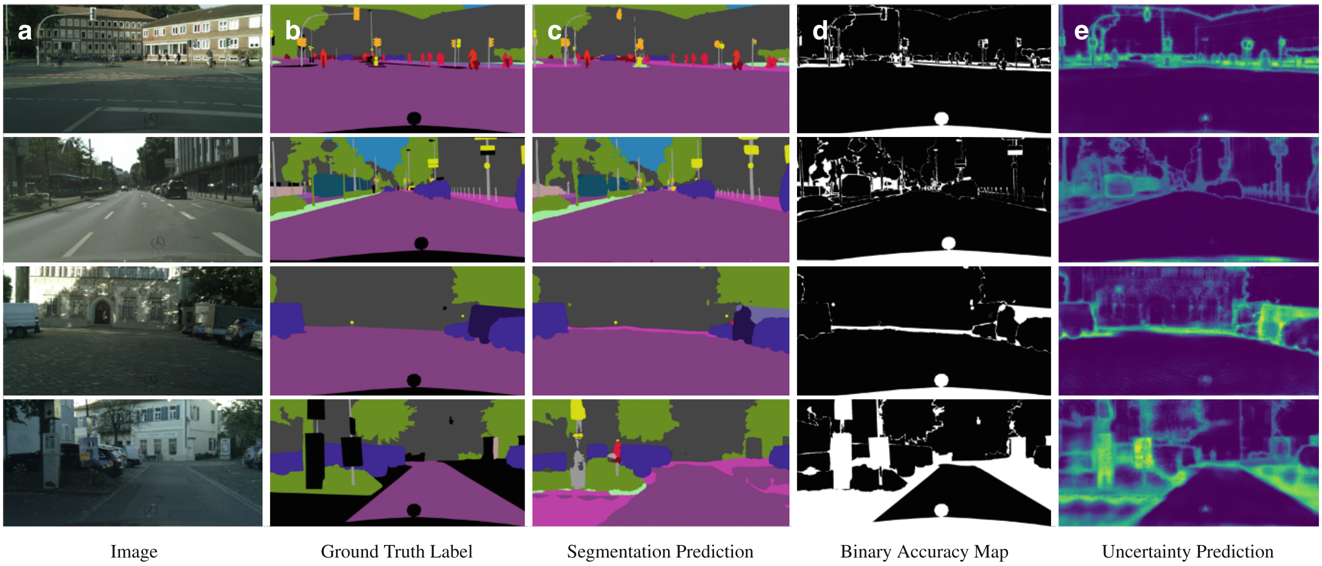

Fig. 6 Example images from the Cityscapes validation set **(a)** with corresponding ground truth labels **(b)**, our student's segmentation predictions **(c)**, a binary accuracy map **(d)**, and the student's uncertainty prediction **(e)**. White pixels in the binary accuracy map are either incorrect predictions or void classes. Latter appear black in the ground truth labels. For the uncertainty prediction, brighter pixels represent higher predictive uncertainties

out in Fig. 4 and Fig. 5: The uncertainty prediction reliably returns high uncertainties for wrongly classified pixels and out-of-domain samples, which both are visualized as white pixels in the binary accuracy map. For example, in the first image of Fig. 6, our student correctly predicts high uncertainties in the noisy parts of the background and for fine geometric structures like traffic lights. Conversely, the student predicts very low uncertainties for the road, buildings, sky, and vegetation. The second example image confirms this observation and adds two valuable insights about the quality of the student's uncertainty predictions. Firstly, although the train in the left part of the image is predicted correctly for the most part, the student still predicts high uncertainties. This is intuitively comprehensible and desired because the train class is underrepresented in the dataset (Cordts et al. 2016) and therefore potentially more difficult to detect reliably. Secondly, the student predicts high uncertainties in the bottom part of the image where reflections on the hood of the car cause incoherent segmentation predictions. The third image exemplifies another quality of the student's predictive uncertainty. In this case, the student struggles to correctly segment the truck in the right part of the image. Simultaneously, the student predicts high uncertainties for the entire truck, thus indicating the wrong segmentation prediction. The fourth image demonstrates the student's potential capability to identify out-of-domain samples: for areas that belong to the void class, high uncertainties are predicted.

4.4 Potential for Out-of-Domain Detection

To investigate the potential of DUDES for OOD detection, we evaluate our student model and the teacher ensemble on Foggy Cityscapes (Sakaridis et al. 2018) and Rain Cityscapes (Hu et al. 2019) without re-training them.

Quantitative Evaluation. As Tables. 2 and 3 show, our student model compares quite well with the teacher. Across all six validation datasets with varying amounts of simulated fog and rain, our student performs better on the segmentation task. It also manages to output similar predictive uncertainties, although underestimating them with increasing intensity of fog and rain in comparison to the teacher. Potentially, this gap can be closed by incorporating hold-out samples or additional data augmentations during the distillation process to improve the student's ability to gen-

Table 2 Comparison between the student's and the teacher's mean Intersection over Union (mIoU) and mean class-wise predictive uncertainty (mUnc) on the validation set of the Foggy Cityscapes dataset (Sakaridis et al. 2018). β denotes the attenuation coefficient and controls the thickness of the fog. Higher β values result in thicker fog

	$\text{Fog}_{\beta=0.005}$		$\text{Fog}_{\beta=0.01}$		$\text{Fog}_{\beta=0.02}$	
	mIoU ↑	mUnc	mIoU ↑	mUnc	mIoU ↑	mUnc
Teacher	64.2	12.3	57.6	14.1	46.7	16.2
Student	67.4	12.0	60.5	12.8	49.3	14.3

Table 3 Comparison between the student's and the teacher's mean Intersection over Union (mIoU) and mean class-wise predictive uncertainty (mUnc) on the validation set of the Rain Cityscapes 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

	Rain_1		Rain_2		Rain_3	
	mIoU ↑	mUnc	mIoU ↑	mUnc	mIoU ↑	mUnc
Teacher	47.7	13.2	40.6	14.9	34.9	16.2
Student	48.3	12.3	42.2	13.4	36.1	14.0

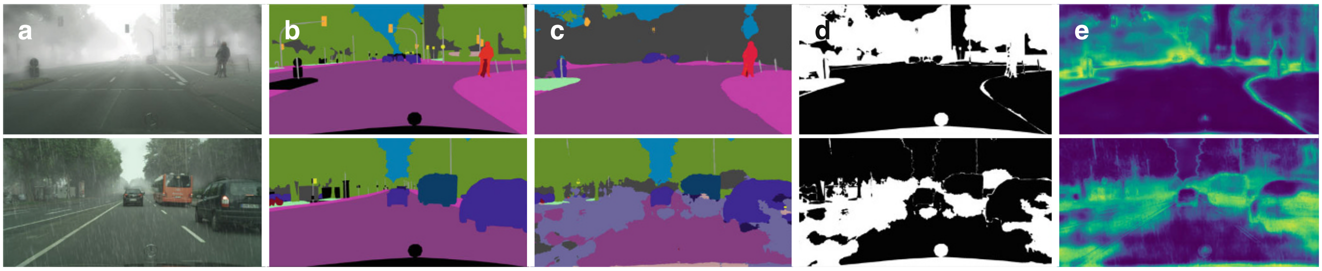

Image Ground Truth Label Segmentation Prediction Binary Accuracy Map Uncertainty Prediction

Fig. 7 Example images from the Foggy Cityscapes (top) and Rain Cityscapes (bottom) validation set **(a)** with corresponding ground truth labels **(b)**, our student's segmentation predictions **(c)**, a binary accuracy map **(d)**, and the student's uncertainty prediction **(e)**. White pixels in the binary accuracy map are either incorrect predictions or void classes. Latter appear black in the ground truth labels. For the uncertainty prediction, brighter pixels represent higher predictive uncertainties

eralize on OOD tasks. This certainly remains an interesting research question for future work.

Qualitative Evaluation. Fig. 7 supports the quantitative findings with qualitative examples. As expected, the simulated fog and rain degrade the quality of the segmentation prediction considerably. Nevertheless, the student model exhibits valuable predictive uncertainty estimations, particularly in regions with numerous incorrect classifications. Generally, this adds to the observations of Fig. 6: high uncertainties of the student correlate with wrongly classified pixels and out-of-domain samples.

4.5 Ablation Studies

Number of Ensemble Members. An essential part of DUDES is the quality of the teacher's uncertainty prediction because it represents an upper bound for the uncertainty quality that can be expected from the student. Fig. 8 shows the impact of the number of ensemble members on the mIoU and mean Uncertainty (mUnc). Naturally, adding more ensemble members improves the segmentation results. The mIoU increases from 70.6% when using just two ensemble members to a maximum 71.4% for twelve members. More importantly for DUDES, the mUnc increases from 0.092 for just two ensemble members to a maximum of 0.107 for six members. Adding more ensemble members to the teacher does not change the uncertainty prediction substantially as the mUnc stays within 0.106 and 0.107 until all twenty members are included. Overall, using ten members appears to strike a balance between segmentation performance and computational efficiency. While the mIoU is only 0.1% lower compared to using twelve members, opting for ten members reduces the computational cost in terms of training time and memory footprint considerably. These findings go along with prior work on Deep-Ensemble-based uncertainty quantification by Fort et al. (2020) and Lakshminarayanan et al. (2017). Consequently, we propose to use ten ensemble members for DUDES, which should be sufficient for most applications.

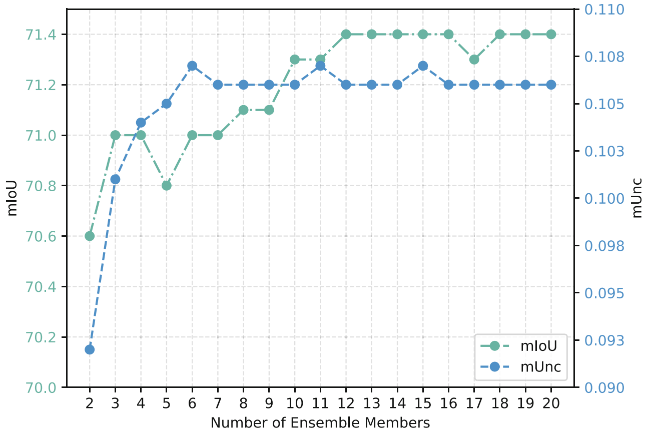

Fig. 8 Ablation study on the impact of the number of ensemble members on the mean Intersection over Union (mIoU) and mean class-wise predictive uncertainty (mUnc).

Table 4 Ablation study on the impact of ImageNet (Deng et al. 2009) pre-training on the mean Intersection over Union (mIoU), mean Uncertainty (mUnc), and Expected Calibration Error (ECE) (Naeini et al. 2015). We evaluate the ECE with the l_1 norm and a bin size of 10

	Training Epochs	mIoU ↑	ECE ↓	mUnc
$\text{Teacher}_{\text{Random},n=10}$	200	71.3	0.021	0.106
$\text{Baseline}_{\text{Random}}$	200	68.5	0.031	–
$\text{Student}_{\text{Random}}$	200	64.6	0.045	0.097
$\text{Baseline}_{\text{ImageNet}}$	200	73.7	0.019	–
$\text{Student}_{\text{ImageNet}}$	200	73.8	0.025	0.108
$\text{Student}_{\text{Random}}$	800	73.9	0.037	0.105

Impact of Pre-training. Table 4 shows the results of another ablation study on the impact of ImageNet (Deng et al. 2009) pre-training on the mIoU, mUnc, and expected calibration error (ECE). We comprehensively compare the baseline segmentation model with our student model and our teacher, which consists of ten randomly initialized baseline models. The study does not examine the impact of ImageNet pre-training on the ensemble members as this would lead to less reliable uncertainties compared to random initialization (Fort et al. 2020; Lakshminarayanan et al. 2017).

While training for 200 epochs and using random initialization, our student underperforms the baseline model by 3.9% and the teacher by 6.7% with a mIoU of 64.6% on the segmentation task. Our randomly initialized student also underestimates the teacher's uncertainties by 0.009 with a mUnc of 0.097. When using ImageNet pre-training for the baseline model and our student, both significantly improve their mIoU with 73.7% and 73.8% respectively. The student also manages to approximate the predictive uncertainties better with a mUnc of 0.108, which is close to the 0.106 of the teacher. It is worth noting that similar performance can also be achieved by randomly initializing our student when the number of training epochs is quadrupled to 800. This concurs with the findings of He et al. (2019). As a consequence, we suggest using ImageNet pre-training for the student to improve convergence speed. On top of that, using ImageNet pre-training leads to a lower ECE.

5 Evaluation of Generalizability

In this section, we provide more evidence for the simplicity and generalizability of DUDES by incorporating additional experiments with a modern Vision-Transformer-based architecture, Monte Carlo Dropout as the uncertainty quantification method, and a different dataset. Firstly, we lay out our adapted experimental setup. Secondly, we provide a quantitative as well as qualitative evaluation to finalize this section.

5.1 Experimental Setup

Architecture. In the following, we use a state-of-the-art Vision-Transformer-based architecture SegFormer-B5 (Xie et al. 2021) pre-trained on ImageNet (Deng et al. 2009) as the backbone of a U-Net (Ronneberger et al. 2015) decoder as the baseline model for the student and the teacher. In accordance with Sect. 3, we add a second uncertainty head to the U-Net's decoder, which is an exact copy of the segmentation head, except for the output layer.

Uncertainty Quantification Method. For uncertainty quantification, we apply Monte Carlo Dropout to our teacher, replacing the use of a Deep Ensemble. Since the SegFormer (Xie et al. 2021) already applies dropout layers throughout the entire network, we follow their work and consider two common dropout rates, 20% and 50% for the teacher model. In order to train the student model, we leave the teacher's dropout layers activated and sample ten times to obtain the predictive uncertainty for the uncertainty distillation (Gal and Ghahramani 2016; Shen et al. 2021; Gustafsson et al. 2020).

Training. During the training processes, we make three changes compared to Sect. 3. Firstly, we decrease the initial learning rate to 0.001. Additionally, we apply color jittering in the distillation process to enhance the quality of the student's uncertainty estimates. Shen et al. (2021) showed that this is useful when the training dataset is used for training and distillation to prevent the student from underestimating the teacher's test-time uncertainty distribution. We followed their suggestion of random variation in the range of [-0.2, 0.2] in four aspects: brightness, contrast, hue, and saturation. Lastly, we change the crop size to 256 × 256.

Dataset. In addition to a different architecture and uncertainty quantification method, we also use the Pascal VOC 2012 (Everingham et al. 2010) dataset for evaluation. Unlike Cityscapes, Pascal VOC 2012 consists of only 1464 training images and 1449 validation images with varying resolutions, 20 semantic object classes, and 1 background class. Additionally, the dataset is less homogeneous than Cityscapes. These properties make it inherently difficult to achieve accurate segmentation results with corresponding uncertainty estimates.

Table 5 Quantitative comparison between the baseline's, the student's and the teacher's inference time, mean Intersection over Union (mIoU), Expected Calibration Error (ECE) (Naeini et al. 2015), and mean class-wise predictive uncertainty (mUnc) on the Pascal VOC 2012 dataset. Student$_A$ uses the uncertainties provided by Teacher$_A$ during training, whereas Student$_B$ uses the uncertainties provided by Teacher$_B$

	Dropout	Inference Time [ms]	mIoU ↑	ECE ↓	mUnc
Baseline	–	33.1 ± 0.7	78.8	0.027	–
Teacher$_A$	20%	355.5 ± 1.3	76.1	0.013	0.096
Teacher$_B$	50%		65.6	0.008	0.157
Student$_A$	–	34.5 ± 1.3	78.7	0.022	0.081
Student$_B$	–		78.4	0.024	0.135

5.2 Quantitative Evaluation

Table 5 shows a comparison between the baseline segmentation model, two teacher models with dropout rates of 20% and 50%, respectively, and two student models with an additional uncertainty head for uncertainty quantification.

Overall, the results align with the experimental findings on the Cityscapes dataset in Sect. 4. Our student models outperform their respective teacher models on the segmentation task while also capturing their predictive uncertainties. They only slightly underestimate their respective teacher's uncertainty, which we attribute to suboptimal hyperparameters and the inherently challenging properties of the Pascal VOC 2012 dataset. As a consequence, our student models match the performance of the baseline model on the segmentation task, while being slightly better calibrated in terms of ECE and they are able to output a meaningful predictive uncertainty, without significantly increasing the inference time.

5.3 Qualitative Evaluation

Fig. 9 corroborates the quantitative findings. The student model predicts high uncertainties for object boundaries, entirely wrong or missing classifications, and areas with fine-grained details that are challenging to classify. In contrast, easy-to-classify areas and background pixels exhibit low uncertainties.

For example, in the first image, our student segments the depicted eagle almost perfectly and accordingly only predicts high uncertainties for the object boundaries. Conversely, in the second and third image, our student either wrongly classifies pixels that should belong to the background or fails to classify parts of the object. Nonetheless, in both cases, high uncertainties are predicted for these areas, providing valuable information. In a similar way, the student predicts high uncertainties for both the bicycle and the human sitting on a bench in the fourth image as they are challenging to classify through all of the fine-grained details and noise.

6 Discussion

DUDES applies student-teacher distillation with a Deep Ensemble to accurately approximate predictive uncertainties with a single forward pass while maintaining simplicity and adaptability. Against the teacher, the needed inference time per image is reduced by an order of magnitude and the computational overhead in comparison to the baseline is neglectable. Additionally, the student exhibits impressive potential for identifying wrongly classified pixels and out-of-domain samples within an image by leveraging its uncertainties. Based on these observations, one could easily introduce an uncertainty-based threshold for OOD detection. However, it is essential to acknowledge that there remains a challenge in distinguishing between misclassified pixels and out-of-domain samples, as both may trigger the threshold. Therefore, careful consideration and further refinements may be necessary to address this challenge effectively.

DUDES represents a simple yet highly effective new approach for uncertainty quantification. In contrast to the work by Holder and Shafique (2021), DUDES requires no major changes to the student's architecture compared to the baseline and introduces only a single uncertainty loss without additional hyperparameters, yet delivers substantial improvements over their work. Firstly, our student model slightly outperforms its teacher in the segmentation task by 2.5% while their student suffers from a segmentation performance degradation in comparison to its teacher by 4.2%. Secondly, our student approximates its teacher's predictive uncertainties more closely than the student model by Holder and Shafique (2021). More precisely, their student tends to underestimate uncertainties for classes with high uncertainties and vice versa, whereas our student does not suffer from any systematic shortcomings.

A major factor of the effectiveness of DUDES lies in the simplification of what is distilled. Instead of distilling the teacher's uncertainty map, which is what Holder and Shafique (2021) proposed, we only use the teacher's predictive uncertainty. The teacher's uncertainty map is calculated by computing the standard deviation of the softmax probability maps of the individual models along the class dimen-

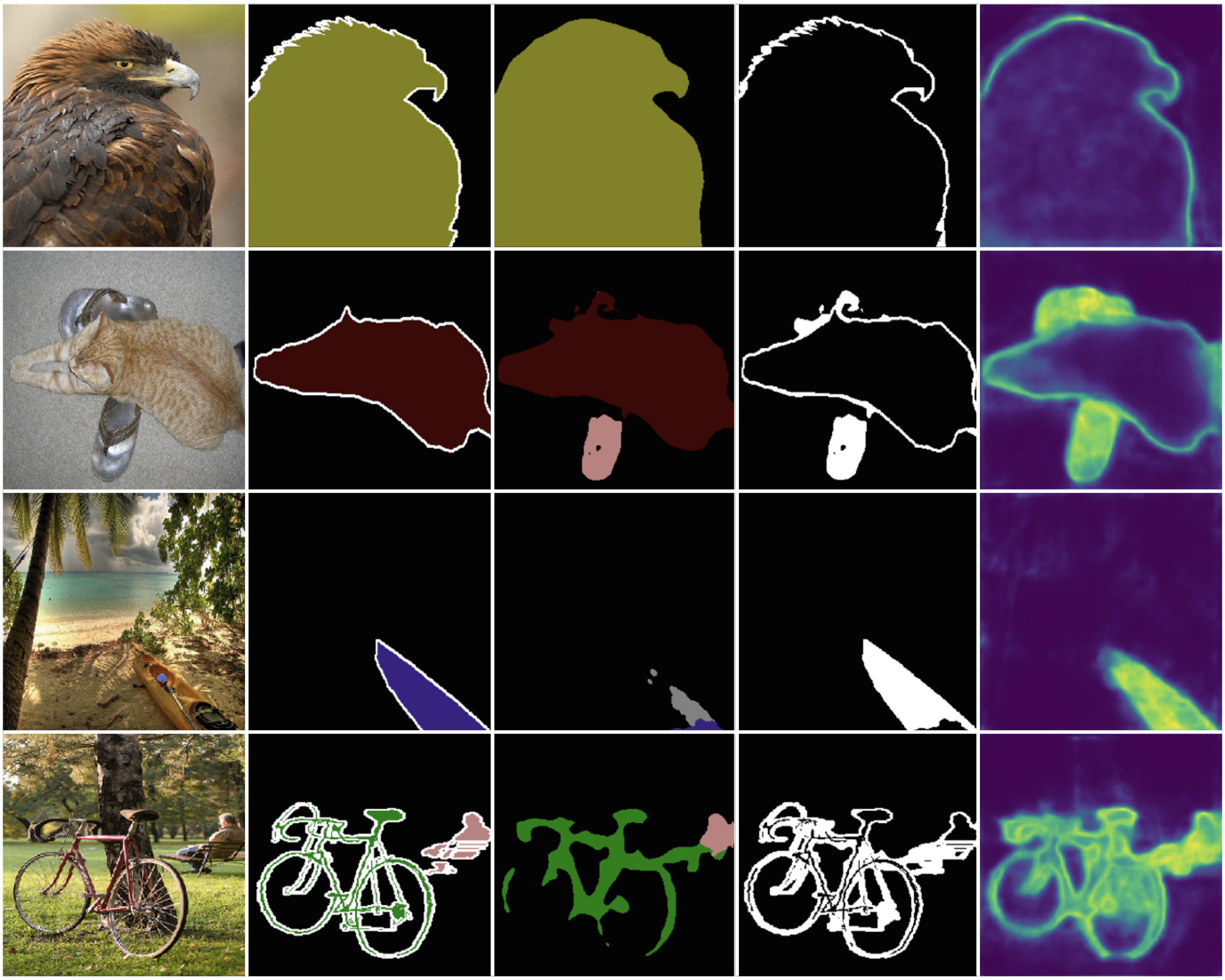

Fig. 9 Example images from the Pascal VOC 2012 validation set with corresponding ground truth labels, our student's segmentation predictions, a binary accuracy map, and the student's uncertainty prediction from left to right. White pixels in the binary accuracy map are either incorrect predictions or belong to the void class. Latter appears white in the ground truth label. For the uncertainty prediction, brighter pixels represent higher predictive uncertainties

sion. In the case of multi-class semantic segmentation, the resulting uncertainty map has dimensions of $C \times H \times W$, where C is the number of classes, H is the image height, and W is the image width. For DUDES, the class dimension is reduced to 1 by only considering the uncertainty of the predicted class in the segmentation map. Due to this simplification, the student's segmentation performance is not hindered and the predictive uncertainties can be learned more accurately.

We acknowledge the simplification in the uncertainty distillation to be a potential limitation of DUDES as the student is only capable of estimating the uncertainty of the predicted class. However, there are practically no negative implications of this limitation since the remaining uncertainties are usually discarded anyway. Hence, DUDES remains useful for efficiently estimating predictive uncertainties for a wide range of applications while being easy to adapt.

We believe that DUDES has the potential to provide a new promising paradigm in reliable uncertainty quantification by focusing on simplicity and efficiency. Except for the computational overhead during training, we found no apparent reason to not employ our proposed method in semantic segmentation applications where safety and reliability are critical.

7 Conclusion

In this work, we propose DUDES, an efficient and reliable uncertainty quantification method by applying student-teacher distillation that maintains simplicity and adaptability throughout the entire framework. We quantitatively demonstrated that DUDES accurately captures predictive uncertainties without sacrificing performance on the segmentation task. Additionally, qualitative results indicate impressive capabilities for the potential identification of wrongly classified pixels and out-of-domain samples through a simple uncertainty-based threshold. With DUDES, we managed to simultaneously simplify and outperform previous work on Deep-Ensemble-based uncertainty quantification.

We hope that DUDES encourages other researchers to incorporate uncertainties into state-of-the-art semantic segmentation approaches and to explore the usefulness of our proposed method for other tasks such as detection or depth estimation.

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.

Authors contributions Conceptualization: Steven Landgraf, Kira Wursthorn, Markus Hillemann, Markus Ulrich.
Methodology: Steven Landgraf, Kira Wursthorn.
Implementation and Experiments: Steven Landgraf.
Writing—original draft preparation: Steven Landgraf.
Writing—review and edition: Steven Landgraf, Kira Wursthorn, Markus Hillemann, Markus Ulrich.
Supervision: Markus Ulrich.

Funding Open Access funding enabled and organized by Projekt DEAL.

Availability of data and material All of the datasets used in this research are publicly available.

Code availability Code is available at: https://github.com/StevenLandgraf/DUDES

Conflict of interest The authors declare no conflics of interest.

References

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), IEEE, Anchorage, AK, USA, pp 3353–3357, https://doi.org/10.1109/ICIP42928.2021.9506719

Bianco S, Cadene R, Celona L, Napoletano P (2018) Benchmark Analysis of Representative Deep Neural Network Architectures. In: IEEE Access, vol 6, pp 64270–64277, https://doi.org/10.1109/ACCESS.2018.2877890

Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D (2015) Weight uncertainty in neural network. In: Bach F, Blei D (eds) Proceedings of the 32nd International Conference on Machine Learning, PMLR, Lille, France, vol 37, pp 1613–1622

Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV)

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

Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The PASCAL visual object classes (VOC) challenge. International journal of computer vision 88:303–338

Fort S, Hu H, Lakshminarayanan B (2020) Deep Ensembles: A Loss Landscape Perspective. arXiv:191202757 1912.02757

Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Balcan MF, Weinberger KQ (eds) Proceedings of The 33rd International Conference on Machine Learning, PMLR, New York, New York, USA, Proceedings of Machine Learning Research, vol 48, pp 1050–1059, https://proceedings.mlr.press/v48/gal16.html

Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, Kruspe A, Triebel R, Jung P, Roscher R, Shahzad M, Yang W, Bamler R, Zhu XX (2022) A Survey of Uncertainty in Deep Neural Networks. arXiv:210703342 2107.03342

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

Gustafsson FK, Danelljan M, Schon TB (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 K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

He K, Girshick R, Dollar P (2019) Rethinking ImageNet pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop, http://arxiv.org/abs/1503.02531

Holder CJ, Shafique M (2021) Efficient Uncertainty Estimation in Semantic Segmentation via Distillation. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), IEEE, Montreal, BC, Canada, pp 3080–3087, https://doi.org/10.1109/ICCVW54120.2021.00343

Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:170404861 1704.04861

Hu X, Fu CW, Zhu L, Heng PA (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

Jospin LV, Laga H, Boussaid F, Buntine W, Bennamoun M (2022) Hands-on Bayesian neural networks—a tutorial for deep learning users. IEEE Computational Intelligence Magazine 17(2):29–48, https://doi.org/10.1109/MCI.2022.3155327

Kang J, Gwak J (2019) Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. IEEE Access 7:26440–26447

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

Lee K, Lee H, Lee K, Shin J (2018) Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. arXiv:171109325 1711.09325

Leibig C, Allken V, Ayhan MS, 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

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

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

Lumini A, Nanni L, Maguolo G (2021) Deep ensembles based on stochastic activations for semantic segmentation. Signals 2(4):820–833

MacKay DJC (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:190500076 1905.00076

Marmanis D, Wegner JD, Galliani S, Schindler K, Datcu M, Stilla U (2016) Semantic segmentation of aerial images with an ensemble of cnns. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 3:473–480

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, International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, pp 4745–4753, https://doi.org/10.24963/ijcai.2017/661

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

Mukhoti J, Kirsch A, van Amersfoort J, Torr PH, 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 MP, Cooper G, Hauskrecht M (2015) Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI conference on artificial intelligence, vol 29

Nanni L, Fusaro D, Fantozzi C, Pretto A (2023) Improving existing segmentators performance with zero-shot segmentators. Entropy 25(11):1502

Nigam I, Huang C, Ramanan D (2018) Ensemble knowledge transfer for semantic segmentation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, pp 1499–1508

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: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32

Robbins H, Monro S (1951) A Stochastic Approximation Method. The Annals of Mathematical Statistics 22(3):400–407, https://doi.org/10.1214/aoms/1177729586

Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y (2015) FitNets: Hints for Thin Deep Nets. arXiv:14126550 1412.6550

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, Springer, pp 234–241

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

Shen Y, Zhang Z, Sabuncu MR, 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

Simpson IJA, Vicente S, Campbell NDF (2022) Learning structured gaussians to approximate deep ensembles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 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, 2nd Edition

Thanh NC, Long TQ, et al. (2020) Polyp segmentation in colonoscopy images using ensembles of u-nets with efficientnet and asymmetric similarity loss function. In: 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), IEEE, pp 1–6

Ulrich M, Hillemann M (2024) Uncertainty-aware hand–eye calibration. IEEE Transactions on Robotics 40:573–591, https://doi.org/10.1109/TRO.2023.3330609

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

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

Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P (2021) Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems 34:12077–12090