

Uncertainty Aware Training to Improve Uncertainty Active Learning for Semantic Segmentation

Moritz Thoma^{1,3*} Tobias Preintner^{3,4} Emad Aghajanzadeh^{1,3} Shambhavi Balamuthu Sampath^{1,3}
 Pierpaolo Mori³ Nael Fafous³ Manoj-Rohit Vemparala³ Alexander Frickenstein³
 Daniel Mueller-Gritschneider² Ulf Schlichtmann¹

¹Technical University of Munich ²TU Wien ³BMW Group ⁴Leiden University

Abstract

Active learning (AL) on deployed devices involves collecting data to improve the performance of machine learning models beyond their initial version. In this context, the collected data must be uploaded to a remote server for training the next version of the model. Data collection on deployed devices can pose one of two challenges: (1) if done naively, large quantities of uninformative data are uploaded or (2) if complex data selection algorithms are used, careful curation of data introduces a compute overhead that is unrelated to the main task. In this paper, we introduce a novel Uncertainty Aware Active Learning (UAAL) method for semantic segmentation that integrates uncertainty estimates into the training process, thereby conditioning the model for an AL setup before deployment time. This enhances the quality of the collected data for an existing selection technique without adding more complexity to the selection algorithm. Essentially, the model outputs provide a clearer indication to the selection algorithm about its own uncertainty. We integrate UAAL with five uncertainty AL methods and demonstrate its efficacy on the CityScapes and ADE20K datasets using three DeepLabv3+ variants. On the CityScapes dataset, with a 10% data budget, UAAL achieves an average mIoU increase of +1.74 p.p., peaking at +3.49 p.p. for MobileNetV3 combined with MC-Dropout. Similarly, on ADE20K, UAAL boosts ResNet-50 with MC-Dropout by +2.22 p.p. at the same data budget, all while not adding any complexity to the data selection process.

1. Introduction

Modern deep neural network (DNN) architectures [4, 7, 8] can achieve impressive prediction capabilities. However, to unlock the full potential of these models, a substantial amount of labeled data is required. This is often costly and labor-intensive to obtain, particularly for complex pixel-

wise tasks such as semantic segmentation [10, 29]. Active learning (AL) on deployed devices involves selection algorithms to find and curate the informative data required to improve the networks beyond their initial deployment [35]. Beginning with a small, labeled data set, AL starts a continuous loop of cycles that train a model, deploy it, curate data during operation, upload it to the central server and finally ask an oracle to label it before starting the next cycle as depicted on the left side of Fig. 1.

However, collecting data using deployed devices that often come with constrained memory, computation and bandwidth resources poses one of two challenges: (1) naive data collection has little to no computational overhead but results in the upload of large data quantities to the central server. (2) complex data selection algorithms can significantly reduce upload quantities but come at the expense of increased computational overhead. Striking a balance between the competing interests of minimizing bandwidth requirements and minimizing computational complexity while improving model performance as much as possible is a non-trivial challenge and the optimal solution depends on many factors including dataset, model and hardware capabilities.

To identify informative data many AL approaches have proposed data *selection algorithms*. Generally, for identifying informative data three categories exist: Reconstruction, Uncertainty, and Diversity. **Reconstruction** approaches use an encoder-decoder structure to compress and reconstruct an image and compare the reconstruction with the original one to measure differences [12, 48]. **Uncertainty**-based methods evaluate uncertainty in the model prediction for a given sample and use it as a proxy for informativeness [16, 18, 24, 28, 49, 50, 52]. **Diversity**-based methods select samples based on their representativeness for the entire, or subsets of the dataset [1, 2, 41]. **Hybrid** approaches combine the two former categories [15, 27].

While all the proposed AL algorithms [16, 18, 24, 28, 49, 50, 52] are proficient at identifying informative images, adding additional or more complex criteria for improved data informativeness always comes with an increased com-

*Corresponding author moritz.thoma@tum.de

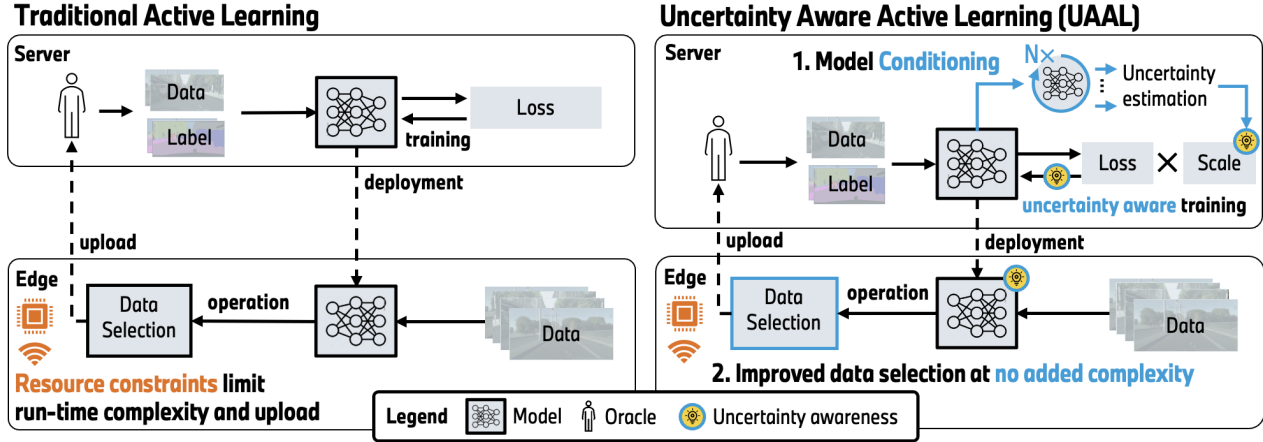


Figure 1. Traditional active learning (left) and our Uncertainty-Aware Active Learning (UAAL) (right). Our method conditions the model with uncertainty-based loss scaling during training (light blue) to improve existing data selection without adding complexity at runtime.

putational overhead [50, 52], which is undesirable when targeting data collection on deployed devices where data collection is usually not the main task purpose.

Hence, instead of adding more selection criteria to the data collection algorithm, we propose a novel **Uncertainty-Aware Active Learning (UAAL)** training scheme to enhance existing uncertainty-based selection algorithms *without* an increased runtime complexity. By utilizing uncertainty estimates from multiple forward passes and loss scaling *during* training, we condition the model for its data selection task at deployment time (as depicted in the right half of Fig. 1). This way our UAAL training can be applied to existing uncertainty-based AL methods, enhancing their performance without increasing their runtime complexity. Our key contributions are summarized as follows:

- We introduce a novel training scheme called **Uncertainty-Aware Active Learning (UAAL)**, which simultaneously optimizes model parameters and enhances data selection by integrating uncertainty estimates *during* training. This approach enables the development of high-accuracy models and is compatible with state-of-the-art uncertainty-based AL methods *without* increasing runtime complexity during deployment.
- Our UAAL method is the first to employ kernel noising to create robust, pixel-wise scaling factors, rendering the loss function uncertainty-aware. This innovation leads to significant improvements in the model’s ability to express uncertainty, thereby enhancing existing data selection strategies.
- We demonstrate the *effectiveness, scalability, and robustness* of UAAL by applying it to five widely-used active learning methods for curating training datasets on prominent semantic segmentation benchmarks, namely CityScapes and ADE20K.

2. Related Work

Active Learning for Semantic Segmentation. AL focuses on algorithms that select the most informative subset from a larger set of data for labeling [39]. This ability to identify highly informative data makes AL approaches compelling for selecting which data to upload from distributed edge devices with constrained resources.

For semantic segmentation AL, two bodies of works must be considered: Pixel-level data selection and image-level data selection. While the latter focuses on labels for the entire image, the former has proven to be highly effective in reducing label costs down to individual pixels or regions [5, 6, 25, 31, 44, 45, 51] and can also incorporate semi-supervised components to further reduce labeling expenses [19, 38]. However, all pixel and region-based approaches require access to unlabeled data. This requirement is apparent for semi-supervised methods, but also applies to pixel-based methods, as model predictions still depend on the full image for context, even if only individual pixels or regions are labeled. Using the context information in the unlabeled data, pixel- and region-based AL approaches can reduce label cost far beyond the capabilities of image-level approaches. However, when using distributed edge devices to curate and collect data, upload costs must be considered [40]. Hence, for that purpose, pixel and region-based AL approaches are prohibitively expensive as labeling few pixels or regions across many [5, 6, 31, 45] or even all images [25, 44, 51] incurs upload of large data quantities from distributed devices, even if only very few labels will be necessary. Differently, image-level methods select few images, that are labeled completely causing higher label cost but result in lower upload cost. Hence, in this work, we focus on image-level methods.

Within the image-level methods, given the scope of this work, our focus is on uncertainty-based AL methods. Within this branch we categorize single and multipass approaches that rely on predictions of 1 and N passes of data through the model, respectively. Entropy [24] AL evaluates the Shannon Entropy on a single model prediction. Generally, its values are high if no class is dominating in the prediction. LLAL [52] also only uses a single pass but adds a module to the model that is trained to estimate the model loss as an informativeness proxy for any given sample. Similarly, DEAL [50] adds a module to predict a difficulty score for every pixel for the task of semantic segmentation. More complex approaches require multiple passes of data through the model. Monte Carlo Dropout (MC-Dropout) AL [16] uses N forward passes while randomly dropping out a certain percentage of activations in the forward pass. Subsequently, it evaluates how much the predictions differ from each other to quantify the uncertainty. Recently, Noise Stability [28] also compares multiple passes, but instead of dropping out components, it adds a random noise relative to the kernel weights of each convolutional layer of the model. EquAL [18] and PCDAL [49] flip transform images and evaluate the consistency in predictions for driving and medical data, respectively. BALD [17] enhances model performance by estimating uncertainty in the model’s posterior distribution. While all of these methods use some form of uncertainty estimation to identify informative data, none of them enhances the selection capability of the model with uncertainty-aware training.

Uncertainty-Aware Training. Models tend to be overconfident in their predictions [20, 43]. This overconfidence often leads to a discrepancy between the model’s output certainty and the actual uncertainty in its predictions. To bridge this gap, researchers have explored various approaches to infuse uncertainty awareness into models during training, with a focus on improving confidence calibration [3, 13, 14, 26, 32, 43, 47] and boost out-of-distribution detection [23, 30, 42].

Among these works, two have targeted uncertainty-aware training for semantic segmentation tasks [13, 26]. In [13], a criterion is introduced that maximizes accuracy on high-confidence images while deprioritizing lower-confidence ones. Although prioritizing high-confidence images may seem counter-intuitive, in their medical use case, low-confidence images are typically reviewed by a physician, making it more important to ensure correctness on high-confidence predictions. The authors of [26] propose an uncertainty loss scaling technique that utilizes uncertainty estimates during training to improve calibration error and overall performance. By incorporating uncertainty into the loss, the model learns to better understand and quantify its own uncertainty, leading to more reliable predictions. However, the contribution of [13] is highly specific to medi-

Categories	[16–18, 24, 28, 49, 50, 52]	[13, 26]	[33, 37]	[This Work]
Active Learning	✓	✗	✓	✓
Uncertainty-Aware Training	✗	✓	✓	✓
Semantic Segmentation	✓	✓	✗	✓

Table 1. Classification of related works.

cal domain challenges while [26] rely on hyperparameter-intensive scaling to enhance prediction performance, making them ill-suited for a general, robust AL application under varying conditions.

Only two works [33, 37] consider AL. The work in [33] introduces a loss function that stores and considers the frequency of correct predictions in the last epochs. The approach in [37] leverages the temporal continuity between video frames, to estimate uncertainty of predictions leading to improved performance in video object detection tasks. However, neither [33] nor [37] do apply to the task of semantic segmentation, as the scaling in [33] would require storing millions of predictions per image while [37] mandate the availability of temporal information, which cannot be assumed in semantic segmentation [53].

Summary. Existing semantic segmentation AL methods have largely overlooked uncertainty-aware training, resulting in untapped performance potential. The few studies [13, 26, 33, 37, 53] that do incorporate uncertainty for segmentation tasks are not applicable or do not scale to semantic segmentation AL tasks. UAAL addresses this gap by introducing a robust uncertainty-based scaling factor that enhances model expressiveness and data selection efficiency without increasing runtime complexity.

3. Methodology

The field of semantic segmentation AL has largely overlooked the potential of uncertainty-aware training, resulting in untapped performance (Tab. 1). Addressing this gap, our UAAL approach for semantic segmentation incorporates uncertainty estimates into the training process, effectively conditioning the model for an AL setup prior to deployment. With UAAL we enhance the data selection quality of uncertainty-based active learning methods *without* increasing their runtime complexity.

Our UAAL approach has three distinct components: (I.) kernel noising for generating uncertainty estimates during training, (II.) scaling calculation to translate the uncertainty into scaling factors, and (III.) scheduling to control the influence of uncertainty during the course of the training. The components of our approach are illustrated in Fig. 2.

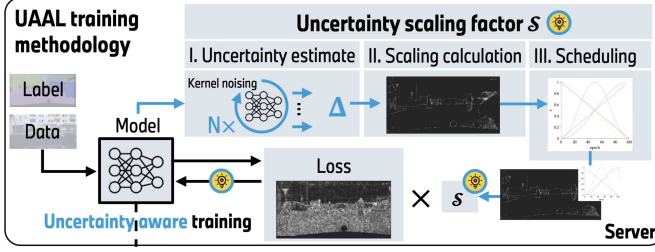


Figure 2. UAAL uncertainty-aware training on server before deployment (also shown in top right corner of overview in Fig. 1) consists of three modules: (I.) Uncertainty estimation using kernel noising with N forward passes, (II.) uncertainty scaling factor calculation and (III.) uncertainty scheduling.

3.1. Uncertainty Estimation

One key component of our UAAL methodology is uncertainty estimation of the predictions *during training*. Literature has explored various ways how to estimate uncertainty [16, 28] in model predictions to understand its uncertainty and use it for data selection *at inference time*. Borrowing from these approaches, we choose to adopt the idea of weight noising [28] over the more popular MC-Dropout [16] as it applies subtle modifications across a broad range of model parameters, instead of inducing large perturbation at few locations like MC-Dropout. This comes at the advantage of easy integration as the design choices collapse from *where* to place layers and *how much disturbance* for Monte-Carlo to only *how much disturbance* for weight noising.

For the weight noising-based uncertainty estimates during training we follow [28] and consider the k^{th} convolutional layer with weights θ_k , for which the noise is calculated by $\Delta\theta_k = \eta\|\theta_k\|\bar{n}_k$, with noise (i.e. disturbance) magnitude $\eta \in [0, 1]$ controlling the relative noise and a random direction $\bar{n}_k$ with unit length. The original weights are then perturbed to $\tilde{\theta}_k = \theta_k + \Delta\theta_k$. Using the set of perturbed weights $\tilde{\theta}$, we get an updated prediction $\tilde{p} = f(\tilde{\theta}, x)$ instead of our original prediction $p = f(\theta, x)$. Using N forward passes, we compute N perturbed predictions and use them to compute their variance as a quantification of the uncertainty in the predictions:

$$\sigma^2 = \text{Var}(f(\tilde{\theta}, x)) \quad (1)$$

For each forward pass $n \in [0, N)$ in every training step, we recompute the weight noise $\{\tilde{\theta}^{(1)}, \dots, \tilde{\theta}^{(N)}\}$ to account for the changing model weight magnitude during training.

3.2. Uncertainty Scaling

Building upon the uncertainty estimation framework established in the previous section, we introduce a novel uncertainty-aware loss scaling mechanism, denoted as $\mathcal{S}$,

which dynamically modulates the contribution of each pixel during the training process. Our approach is meticulously designed to fulfill three essential criteria for effective scaling: (1) dependence on uncertainty estimation, (2) amplification of pixels exhibiting higher uncertainties, and (3) robustness to variations in hyperparameters.

To achieve uncertainty-dependent loss scaling, we draw inspiration from the work of [26] and propose a scaling factor $\mathcal{S}_i$ for each pixel i within an image containing a total of P pixels. The foundation of the scaling factor is based on the variance σ observed during N stochastic forward passes of the network, as outlined in Eq. 1. To obtain pixel wise scaling, we compute the standard deviation σ_i from the softmax class-wise predictions for each pixel and class. Given that the softmax outputs p_i of pixel i are constrained within the interval $[0, 1]$, the maximum achievable standard deviation for pixels is given by:

$$\sigma_{i,\max} = \sqrt{\frac{1}{N} \sum_{n=1}^N (p_i^{(n)} - \bar{p}_i)^2} \leq 0.5 \quad (2)$$

To increase the loss for pixels depending on the standard deviations, we establish a lower bound of one for the loss. Consequently, the initial scaling factor is defined as:

$$\mathcal{S}_i = 1 + \sigma_i \quad (3)$$

To further emphasize pixels characterized by high uncertainty, we introduce an exponent $\mathcal{E}$, which governs the degree of emphasis placed on pixels with significant standard deviations. This modification yields the following enhanced scaling equation:

$$\mathcal{S}_i = (1 + \sigma_i)^\mathcal{E} \quad (4)$$

While Eq. 4 aligns with our objective of integrating uncertainty into the scaling process, it is important to note that only substantial discrepancies between predictions will result in appreciable scaling effects. Moreover, with standard deviation approximately proportional to noise magnitude $\sigma \sim \eta$, noise magnitude has a strong effect for the scaling. Check the supplementary Sec. 6.2.1 for the theoretical foundation. To ensure the efficacy of our scaling approach across a diverse range of noise levels, we therefore adjust the influence of the standard deviation σ in relation to the inverse of the noise magnitude η . To highlight relative differences without runaway losses, we divide by the average image uncertainty score $\bar{\mathcal{U}}$ while setting a lower bound of one to prevent any pixel's loss from being diminished. Thus, the final scaling factor for pixel i in the image is expressed as:

$$\mathcal{U}_i = \left(1 + \frac{\sigma_i}{\eta}\right)^\mathcal{E} \quad (5)$$

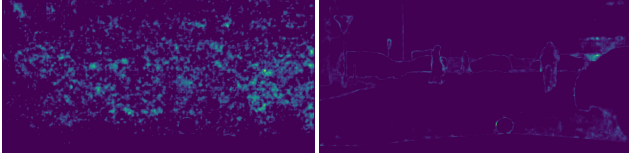


Figure 3. Qualitative comparison of uncertainty estimates for a CityScapes [10] image in the beginning (left), and middle (right) of training. Brighter pixels indicate higher uncertainty.

$$\mathcal{S}_i = \max\left[\frac{\mathcal{U}_i}{\bar{\mathcal{U}}}, 1\right] \quad \text{with } \bar{\mathcal{U}} = \frac{1}{P} \cdot \sum_{i=1}^P \mathcal{U}_i \quad (6)$$

Through this adaptive scaling $\mathcal{S}$, we ensure that our model is not only sensitive to its own uncertainty but also benefits from an enhanced focus on problematic regions. This approach leads to a more discriminative learning process supporting the model in identifying informative data points for its next deployment generations.

3.3. Uncertainty Scheduling

As training progresses, the model’s uncertainty estimates evolve. Initially, the model’s fit to the data is poor, leading to noisy and less reliable uncertainty estimates. However, as training continues, these estimates become more refined and indicative of the model’s confidence in its predictions. The transformation of uncertainty throughout the training is illustrated in Fig. 3, which contrasts early and mid-training uncertainty estimates. To adapt to the changing nature of uncertainty, we introduce a scheduling function $\alpha(\text{epoch})$ that modulates the influence of the scaling factor $\mathcal{S}$ over the course of the training:

$$\mathcal{S}_\alpha(\text{epoch}) = \mathcal{S} \cdot \alpha(\text{epoch}) \quad (7)$$

We explore different scheduling strategies, including a *constant* schedule that maintains a steady influence of uncertainty, as well as dynamic schedules that adjust the influence based on the training epoch. Specifically, we implement *linear* schedules for both increasing and decreasing influence, a *sine* schedule for a smooth increase, and a *cosine* schedule for a smooth decrease. Additionally, we propose a combined *sin-cos* schedule that smoothly shifts from increasing to decreasing. These varied scheduling strategies enable us to investigate the impact of changing uncertainties on the model’s training and performance. The linear increasing schedule $\alpha_{\text{lin, inc}}$ is given as an example in Eq. 8.

$$\alpha_{\text{lin, inc}}(\text{epoch}) = \frac{\text{epoch}}{\text{epoch}_{\text{max}}} \quad (8)$$

All other scheduling functions may be found in supplementary material Sec. 6.3.

4. Evaluation

4.1. Experimental Setup

Datasets. For an in-depth analysis of UAAL we evaluate the performance on the CityScapes [10] and the more challenging ADE20K [53] dataset. ADE20K comprises 20,210 training and 2,000 validation images of different shapes across 150 classes, while CityScapes is more homogeneous with more similar images and static image shapes within 2,975 training and 500 validation images with 19 classes.

Models. We evaluate the DeepLabv3+ architecture [7] with MobileNetV3-Small (MNV3) [22], ResNet-50 (RN50) [21], and Xception-65 (Xc65) [7, 9] backbones, all pre-trained on the ImageNet [11] dataset. All models are trained and evaluated with an output stride of 16.

Training Setup. All models are trained using SGD optimizer with a momentum of 0.9 and a weight decay of $4e-4$, excluding batch normalization layers. We use polynomial learning rate decay starting from 0.01. For CityScapes, we train the models for 200 epochs with a batch size of 8, random resize between 0.5 and 2, and crop size 512×1024 during training and full resolution validation. For ADE20K, we train for 64 epochs with a batch size of 16, resize of longest image dimension to 512 and then crop 512×512 . For validation, we pad outputs to 512×512 after resize instead of performing crop. For UAAL, we set kernel noise η to 0.01, set the exponent $\mathcal{E}$ to 2 in all cases and use a simple constant uncertainty schedule α if not stated otherwise.

Active Learning. For the AL experiments, we start with 2 % randomly selected data and perform 10 AL cycles, adding 2 % each time, ending with a cumulative budget $\mathcal{B}$ of 20 % (= 10 cycles · 2 %) following [28, 36, 52]. Since this setting is consistent, specifying both cycle and budget is redundant. Therefore, we may refer only to budgets: For example, a 10 % budget $\mathcal{B}$ refers to the data accumulated in 5 cycles of AL. Like [1, 41, 46] we train again from scratch after adding data in every cycle. All values are averages of three runs that start from the same random subset. We choose to evaluate our approach on image-level AL approaches, which, while less efficient in labeling than pixel and region-based methods, incur significantly lower upload costs, making them more suitable for distributed data collection use cases. Specifically, we use (1) Entropy [24], (2) BALD [17], (3) EquAL [18], (4) MC-Dropout [16], and (5) Noise stability [28] as they represent the category of image-level uncertainty-based AL well and do not require hard-to-generalize additional modules like LLAL [52] and DEAL [50]. We further use 50 % dropout rate for MC-Dropout AL and small noise $\eta = 0.01$ for Noise Stability following [34] and [28], respectively. We further add Random selection and the well-recognized diversity-based Core-set [41] as non-uncertainty-based reference points.

Data Selection Method	Training Method	CityScapes ($\mathcal{B} = 10\%$)			CityScapes ($\mathcal{B} = 20\%$)		
		MobileNetV3	ResNet-50	Xception-65	MobileNetV3	ResNet-50	Xception-65
Random selection	standard	48.06 $\pm$ 0.56	62.72 $\pm$ 1.30	62.72 $\pm$ 0.75	53.91 $\pm$ 0.46	68.52 $\pm$ 0.92	68.88 $\pm$ 0.99
Core-set [41]	standard	49.79 $\pm$ 0.38	65.82 $\pm$ 0.18	65.61 $\pm$ 0.26	55.50 $\pm$ 0.18	70.04 $\pm$ 0.35	71.52 $\pm$ 0.32
Entropy [24]	standard	52.10 $\pm$ 0.36	66.87 $\pm$ 0.61	67.12 $\pm$ 0.68	57.22 $\pm$ 0.15	71.17 $\pm$ 0.93	71.85 $\pm$ 0.71
	UAAL (ours)	53.34$\pm$0.31	68.60$\pm$0.30	68.48$\pm$0.17	58.91$\pm$0.29	72.27$\pm$0.33	72.76$\pm$0.41
BALD [17]	standard	49.30 $\pm$ 0.24	59.62 $\pm$ 0.44	61.39 $\pm$ 1.43	54.20 $\pm$ 0.29	66.22 $\pm$ 0.20	67.84 $\pm$ 0.26
	UAAL (ours)	51.90$\pm$0.21	60.97$\pm$0.59	62.02$\pm$0.22	57.67$\pm$0.34	66.33$\pm$0.22	66.39 $\pm$ 0.17
EquAL [18]	standard	51.82 $\pm$ 0.32	66.50 $\pm$ 0.34	67.06 $\pm$ 0.26	57.25 $\pm$ 0.06	71.12 $\pm$ 0.58	71.69 $\pm$ 0.53
	UAAL (ours)	54.37$\pm$0.62	67.88$\pm$0.23	67.64$\pm$0.77	59.52$\pm$0.52	71.34$\pm$0.73	72.65$\pm$0.31
MC-Dropout [16]	standard	48.19 $\pm$ 0.89	64.79 $\pm$ 0.63	67.13 $\pm$ 0.77	52.92 $\pm$ 0.55	69.96 $\pm$ 0.69	71.40 $\pm$ 0.53
	UAAL (ours)	51.68$\pm$0.51	66.23$\pm$0.59	69.35$\pm$0.54	56.51$\pm$0.22	70.92$\pm$0.73	72.76$\pm$0.66
Noise Stability [28]	standard	50.86 $\pm$ 0.06	66.05 $\pm$ 0.70	67.42 $\pm$ 0.92	56.77 $\pm$ 0.22	70.86 $\pm$ 0.43	71.78 $\pm$ 0.31
	UAAL (ours)	54.29$\pm$0.19	66.52$\pm$0.69	69.06$\pm$0.24	58.87$\pm$0.25	71.03$\pm$0.12	72.26$\pm$0.90

Table 2. Evaluation of image-level uncertainty AL methods at 10 % and 20 % data budget $\mathcal{B}$ for MNV3, RN50 and Xc65 model variants on the CityScapes dataset with and without our UAAL training.

4.2. Quantitative Analysis

In this section, we present a comprehensive analysis of the quantitative performance enhancements achieved by our UAAL approach. The results cover a variety of models, AL strategies, and datasets. In Tab. 2 we provide results for all AL methods on all models for CityScapes after 5 and 10 cycles, i.e. for 10 % and 20 % data budget $\mathcal{B}$, respectively. We further show results for all intermediate budgets $\mathcal{B}$, covering a broad range from 2 % to 20 % on both ADE20K and CityScapes for the RN50 model, as depicted in Fig. 4. Additional, detailed results for ADE20K are available in the supplementary material Sec. 7.

We evaluate our UAAL approach with respect to three core aspects: effectiveness, scalability, and robustness. Each of these aspects provides a unique perspective on our UAAL.

Effectiveness. To assess the effectiveness of UAAL, we evaluate it across all models and AL techniques for the 10 % budget $\mathcal{B}$ on the CityScapes dataset (Tab. 2, left). Starting with a general comparison, most uncertainty AL methods outperform the random selection reference, with a clear majority also outperforming the Core-set reference. Our UAAL approach expands on that, adding a noticeable mean intersection over union (mIoU) improvement to the standard training baselines. Specifically, UAAL consistently enhances existing methods by up to 3.49 percentage points (p.p.) for MNV3 and MC-Dropout, while averaging to 1.74 p.p. across all models and AL techniques. This improvement is especially noteworthy, given the number of models and AL methods put to the test, indicating that UAAL can *effectively* enhance the data selection capabilities without adding complexity at runtime.

Scalability. In this paragraph, we evaluate the scalability of our approach by analyzing its performance across varying budget ranges and its application to the challenging ADE20K dataset. We first analyze the results for RN50 on CityScapes across different budgets, as illustrated in Fig. 4a. The results demonstrate that UAAL (solid lines) consistently outperforms the baseline (dashed lines) across the entire budget $\mathcal{B}$ range from 2 % to 20 %, indicating its robustness across diverse scenarios. Specifically, the average improvements across all budgets are 0.94 p.p. for Entropy and 1.24 p.p. for MC-Dropout, highlighting the effectiveness of our approach. Focusing on the final 20 % budget $\mathcal{B}$ (Tab. 2), we observe a similar average improvement of 1.20 p.p. across all methods, with a peak enhancement of 3.59 p.p. for MNV3 when utilizing MC-Dropout data selection. This consistent superiority across the entire budget spectrum underscores UAAL’s capability to adapt and deliver performance gains, regardless of the available data budget. Next, we extend our evaluation to the ADE20K dataset (Fig. 4b). The superiority across all budgets observed for CityScapes is consistently replicated, showing noticeable improvements across all budgets $\mathcal{B}$. While UAAL yields a modest average improvement of 0.68 p.p. for Entropy across all budgets, the enhancements for MC-Dropout are more pronounced, averaging an increase of 1.68 p.p.. Detailed results for ADE20K are provided in the supplementary material, Sec. 7. Combining observations from both budget range scalability and the ADE20K dataset, we demonstrate that UAAL can effectively scale to handle both varying data budgets and complex datasets.

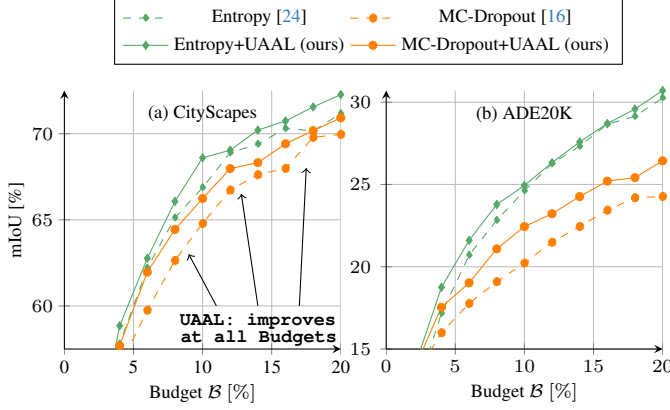


Figure 4. Performance of RN50 DeepLabv3+ on CityScapes (left) and ADE20K (right) for both Entropy and MC-Dropout AL and the respective UAAL variant. UAAL outperforms its vanilla counterpart for all budgets B in both methods and datasets.

Budget B	Increase		Decrease		Sin-Cos	Constant
	Sine	Linear	Cosine	Linear		
10 %	67.49	68.15	67.87	68.07	67.58	68.60
20 %	71.69	71.79	72.14	72.00	71.55	72.27

Table 3. Prediction capability (mIoU) of a RN50 variant of DeepLabv3+ when using our uncertainty scheduling functions α to train on different budgets B of the CityScapes [10] dataset, selected using Entropy [24].

Robustness. Finally, we assess the robustness of UAAL. For that, we analyze the impact of different scheduler functions α , which modulate our uncertainty scaling factor S_α during training. Our results for the RN50 DeepLabv3+ model on the CityScapes dataset with 10 % and 20 % data budgets B and all scheduler functions are summarized in Tab. 3. The results indicate that decreasing strategies generally outperform increasing ones, suggesting that uncertainty scores are more influential early in the training process. Notably, a constant schedule, without any scaling adjustments throughout the training, achieves the best mIoU performance. This observation indicates that our approach is inherently capable of *self-regulation*, thereby eliminating the need for manual scaling adjustments. With self-regulating nature of UAAL the training process simplifies, which is particularly advantageous in real-world applications with varying conditions, as it ensures reliable performance without constant hyperparameter tuning. For an extended analysis of hyperparameter robustness, focusing on parameters of our scaling factor S i.e. noise magnitude η as well as investigations on the exponent $\mathcal{E}$ we refer to the supplementary material Sec. 6.

Method		CityScapes		
		MobileNetV3	ResNet-50	Xception-65
10 % budget $\mathcal{B}$	standard: all data	68.27	77.62	78.05
	+UAAL (<i>baseline</i>)	69.34 (+1.07)	77.71 (+0.08)	78.57 (+0.52)
	Entropy	52.10	66.87	67.12
	+UAAL	53.34 (+1.25)	68.60 (+1.73)	68.48 (+1.36)
	BALD	49.30	59.62	61.39
	+UAAL	51.90 (+2.60)	60.97 (+1.35)	62.02 (+0.63)
	EquAL	51.82	66.50	67.06
	+UAAL	54.37 (+2.55)	67.88 (+1.38)	67.64 (+0.58)
	MC-Dropout	48.19	64.79	67.13
	+UAAL	51.68 (+3.49)	66.23 (+1.45)	69.35 (+2.22)
Noise Stability	50.86	66.05	67.42	
+UAAL	54.29 (+3.43)	66.52 (+0.47)	69.06 (+1.64)	

Table 4. Prediction quality (mIoU) of trainings with and without our UAAL method to assess the contribution of improved data selection quality and improved training from our UAAL for MNV3 [22], RN50 [21] and Xc65 [7, 9] DeepLabv3+ [7]. UAAL contributes to performance uplift through both training and improved data selection.

Overall, UAAL achieves remarkable effectiveness, scalability, and robustness across diverse datasets, models, active learning methods, and data budgets B , all *without adding runtime complexity* to the data selection process. It consistently provides significant gains in mIoU performance without hyperparameter adjustments. This positions UAAL as a highly appealing approach for enhancing existing uncertainty-based data selection algorithms, particularly in online environments.

4.3. Improved Data Selection vs Improved Training

While UAAL has improved AL methods with high consistency, it is interesting to quantify how much the enhanced training methodology contributes to the overall performance gains compared to the contribution of the improved data selection. To disentangle the impact of UAAL’s training enhancements from the improved data selection, we isolate the training contribution by training on the entire dataset. This allows us to directly measure the benefits conferred by UAAL’s training process without the impact of data selection and compare them with the advantages observed during AL experiments. The results of these experiments are summarized in Tab. 4. The table is organized as follows: the first row displays the results obtained from the full dataset, both with and without UAAL training, isolating the contribution to the training. Afterwards, the results of the different AL methods for 10 % data budget B with and without UAAL are given. For each paired comparison (baseline versus UAAL-enhanced), we provide the absolute performance delta. In the full data reference scenario outlined in the first row, UAAL yields improvements in mIoU

between 1.07 p.p. for MNV3 and 0.08 p.p. for RN50, establishing UAAL’s training contribution. When analyzing the improvements brought by UAAL across each AL method and model, we observe that the absolute performance gains (noted in parentheses) in AL consistently exceed the improvements seen in our reference scenario. Hence, while a modest enhancement can be traced back to UAAL’s training methodology, substantial gains of up to 2.36 p.p. in the case of Noise Stability with UAAL can be attributed to the improved data selection facilitated by our UAAL training approach. This data solidifies that UAAL improves both training *and* data selection, showing its overall efficacy in AL paradigms.

4.4. Qualitative Analysis

In this section, we provide a qualitative analysis to illustrate the improvements brought by our UAAL approach in terms of uncertainty map clarity corroborating the quantitative improvements found in the previous chapter. For the analysis, we create uncertainty maps for various images from the CityScapes validation set. The uncertainty maps are generated using the uncertainty estimation of Entropy, MC-Dropout and Noise Stability AL methods. The uncertainty is estimated on all model variants (MNV3, RN50, Xc65) after being trained on 10 % budget $\mathcal{B}$ i.e. after 5 cycles of AL for both the standard training and our novel UAAL training approach. Fig. 5 depicts a comparison between uncertainty maps generated after using standard training methods and those produced when using our UAAL training for one image. Examples for MNV3, Xc65 and a second image are provided in the supplementary material in Sec. 8.

As can be seen, the different uncertainty estimation methods produce distinct appearances for the uncertainty maps: MC-Dropout results in blurry maps due to the high dropout rate employed. In contrast, both Entropy and Noise Stability produce much sharper lines. Between the two, Noise Stability generates more homogeneous uncertainties, while Entropy displays many bright green spots indicating a wider range of uncertainties. When applying our method, these differences persist, but with notable improvements. Comparing the uncertainty maps after our UAAL training to those after vanilla training, our method significantly enhances the clarity of the maps. We showcase three key regions with improved uncertainty with red arrows. In general, the improvements are most noticeable around edges such as sidewalks, cars, and small signs, particularly in MNV3 (Supplementary Sec. 8). Overall, the qualitative analysis further highlights the effectiveness of UAAL: by improving certainty around edges, UAAL enhances the quality of the uncertainty maps, ensuring a more precise and reliable data acquisition process, without adding overhead to the selection algorithm.

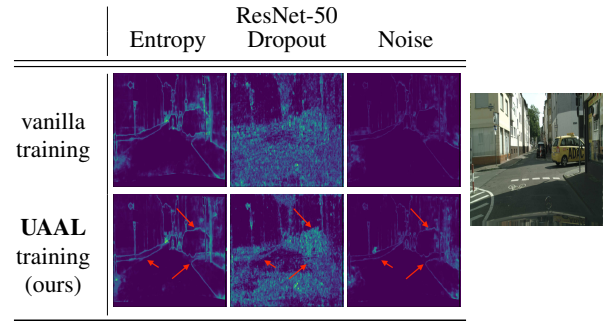


Figure 5. Qualitative comparison of uncertainty maps produced by RN50 DeepLabv3+ on an example CityScapes image using standard training (top row) and our UAAL enhanced training (bottom row). UAAL improves the clarity of the uncertainty maps for all uncertainty methods. Some improvements are pointed out with red arrows. Images are reshaped to 1:1 aspect ratio.

5. Conclusion

In this work, we introduced UAAL, a novel method that employs an uncertainty loss scaling factor $\mathcal{S}$ during training, addressing a significant gap in existing semantic segmentation AL methods, which have largely overlooked uncertainty-aware training. By harnessing kernel noise-based uncertainty estimates during training, our UAAL method enhances model expressiveness and improves data selection efficiency *without increasing runtime complexity*, making it an attractive solution for data collection in distributed, deployed devices. Our extensive evaluation across multiple AL methods, models and datasets establishes UAAL as a highly effective, scalable, and robust approach for improving data selection quality in uncertainty-based AL. By enhancing the selection quality of AL algorithms, UAAL facilitates more accurate data collection, ultimately advancing the capabilities of real-world AI applications.

Future work. Future research could explore integrating UAAL with various model architectures, including vision transformers, and applying it to a broader range of tasks. Additionally, while UAAL introduces no runtime complexity, enhancing its training speed presents a compelling avenue for further exploration. Future efforts may focus on adapting UAAL for post-training fine-tuning, which could reduce training time while maintaining its effectiveness.

References

- [1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 137–153. Springer, 2020. 1, 5
- [2] Zolán Bodó, Zsolt Minier, and Lehel Csató. Active learning

- with clustering. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pages 127–139. JMLR Workshop and Conference Proceedings, 2011. 1
- [3] Patrik Olā Bressan, José Marcato Junior, José Augusto Correa Martins, Maximilian Jaderson de Melo, Diogo Nunes Gonçalves, Daniel Matte Freitas, Ana Paula Marques Ramos, Michelle Taís Garcia Furuya, Lucas Prado Osco, Jonathan de Andrade Silva, et al. Semantic segmentation with labeling uncertainty and class imbalance applied to vegetation mapping. *International Journal of Applied Earth Observation and Geoinformation*, 108:102690, 2022. 3
- [4] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17302–17313, 2023. 1
- [5] Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10988–10997, 2021. 2
- [6] Arantxa Casanova, Pedro O Pinheiro, Negar Rostamzadeh, and Christopher J Pal. Reinforced active learning for image segmentation. *International Conference on Learning Representations (ICLR)*, 2020. 2
- [7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 5, 7
- [8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1
- [9] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 5, 7
- [10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 5, 7, 2
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5
- [12] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16918–16927, 2021. 1
- [13] Yukun Ding, Jinglan Liu, Xiaowei Xu, Meiping Huang, Jian Zhuang, Jinjun Xiong, and Yiyu Shi. Uncertainty-aware training of neural networks for selective medical image segmentation. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, pages 156–173. PMLR, 2020. 3
- [14] Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in Neural Information Processing Systems*, 35:22380–22395, 2022. 3
- [15] Mélanie Gaillochet, Christian Desrosiers, and Hervé Lombaert. Active learning for medical image segmentation with stochastic batches. *Medical Image Analysis*, 90:102958, 2023. 1
- [16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 1, 3, 4, 5, 6, 7, 2
- [17] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 1183–1192. JMLR.org, 2017. 3, 5, 6
- [18] S Alireza Golestaneh and Kris M Kitani. Importance of self-consistency in active learning for semantic segmentation. *British Machine Vision Conference (BMVC)*, 2020. 1, 3, 5, 6
- [19] Licong Guan and Xue Yuan. Iterative loop method combining active and semi-supervised learning for domain adaptive semantic segmentation. *arXiv preprint arXiv:2301.13361*, 2023. 2
- [20] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 3
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 7, 1
- [22] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 5, 7
- [23] Ping Hu, Stan Sclaroff, and Kate Saenko. Uncertainty-aware learning for zero-shot semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 21713–21724. Curran Associates, Inc., 2020. 3
- [24] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2372–2379. IEEE, 2009. 1, 3, 5, 6, 7, 2
- [25] Hoyoung Kim, Minhyeon Oh, Sehyun Hwang, Suha Kwak, and Jungseul Ok. Adaptive superpixel for active learning in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 943–953, 2023. 2
- [26] Steven Landgraf, Markus Hillemann, Kira Wursthorn, and Markus Ulrich. U-ce: Uncertainty-aware cross-entropy for

- semantic segmentation. *arXiv preprint arXiv:2307.09947*, 2023. 3, 4
- [27] Xiaokang Li, Menghua Xia, Jing Jiao, Shichong Zhou, Cai Chang, Yuanyuan Wang, and Yi Guo. Hal-ia: A hybrid active learning framework using interactive annotation for medical image segmentation. *Medical Image Analysis*, 88:102862, 2023. 1
- [28] Xingjian Li, Pengkun Yang, Yangcheng Gu, Xueying Zhan, Tianyang Wang, Min Xu, and Chengzhong Xu. Deep active learning with noise stability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13655–13663, 2024. 1, 3, 4, 5, 6, 2
- [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1
- [30] Jeremiah Zhe Liu, Shreyas Padhy, Jie Ren, Zi Lin, Yeming Wen, Ghassen Jerfel, Zachary Nado, Jasper Snoek, Dustin Tran, and Balaji Lakshminarayanan. A simple approach to improve single-model deep uncertainty via distance-awareness. *Journal of Machine Learning Research*, 24(42): 1–63, 2023. 3
- [31] Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. Cereals-cost-effective region-based active learning for semantic segmentation. *British Machine Vision Conference (BMVC)*, 2018. 2
- [32] Pedro Mendes, Paolo Romano, and David Garlan. Error-driven uncertainty aware training, 2024. 3
- [33] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *international conference on machine learning*, pages 7034–7044. PMLR, 2020. 3
- [34] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 5
- [35] Enrique Nueve, Sean Shahkarami, Seongha Park, and Nicola Ferrier. Addressing the constraints of active learning on the edge. In *2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 845–849, 2021. 1
- [36] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton Van Den Hengel, and Javen Qin-feng Shi. Active learning by feature mixing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12237–12246, 2022. 5
- [37] Aayush J Rana and Yogesh S Rawat. Hybrid active learning via deep clustering for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18867–18877, 2023. 3
- [38] Aneesh Rangnekar, Christopher Kanan, and Matthew Hoffman. Semantic segmentation with active semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5966–5977, 2023. 2
- [39] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Comput. Surv.*, 54(9), 2021. 2
- [40] Sebastian Schmidt, Lukas Stappen, Leo Schwinn, and Stephan Günnemann. Generalized synchronized active learning for multi-agent-based data selection on mobile robotic systems. *IEEE Robotics and Automation Letters*, 2024. 2
- [41] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 1, 5, 6
- [42] Murat Sensoy, Lance Kaplan, Federico Cerutti, and Maryam Saleki. Uncertainty-aware deep classifiers using generative models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5620–5627, 2020. 3
- [43] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9030–9038, 2019. 3
- [44] Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: Semantic segmentation with pixelpick. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1687–1697, 2021. 2
- [45] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9433–9443, 2020. 2
- [46] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5972–5981, 2019. 5
- [47] Pegah Tabarisaadi, Abbas Khosravi, Saeid Nahavandi, Miadreza Shafie-Khah, and João P. S. Catalão. An optimized uncertainty-aware training framework for neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):6928–6935, 2024. 3
- [48] Tomas Vojir, Tomáš Šipka, Rahaf Aljundi, Nikolay Chumerin, Daniel Olmeda Reino, and Jiri Matas. Road anomaly detection by partial image reconstruction with segmentation coupling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15651–15660, 2021. 1
- [49] Tao Wang, Xinlin Zhang, Yuanbo Zhou, Junlin Lan, Tao Tan, Min Du, Qinquan Gao, and Tong Tong. Pcdal: A perturbation consistency-driven active learning approach for medical image segmentation and classification. *arXiv preprint arXiv:2306.16918*, 2023. 1, 3
- [50] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *Proceedings of the Asian conference on computer vision*, 2020. 1, 2, 3, 5
- [51] Amirsaeed Yazdani, Xuelu Li, and Vishal Monga. Maturity-aware active learning for semantic segmentation with hierarchically-adaptive sample assessment. In *34th British*

Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023. BMVA, 2023. [2](#)

- [52] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102, 2019. [1](#), [2](#), [3](#), [5](#)
- [53] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. [3](#), [5](#), [1](#), [2](#)