# Uncertainty Aware Training to Improve Uncertainty Active Learning for Semantic Segmentation

Supplementary Material

## 6. Robust Uncertainty Scaling for UAAL

Robustness is a key design goal in our Uncertainty-Aware Active Learning (UAAL) method. In the main paper, we demonstrated robustness across various active learning (AL) methods and datasets. In this supplementary paper, we further investigate the robustness of hyperparameter selection for our uncertainty scaling (Eq. 6). We dive into the impact of using different exponents $\mathcal{E}$ in the scaling and examine the robustness of our scaling against its hyperparameter $\eta$. We also present the detailed scheduling functions employed in our method. If not stated otherwise we follow the same training setup as described in Sec. 4.1.

### 6.1. Exponent Effect on Scaling Factors

Our scaling $\mathcal{S}$ relies on a power factor $\mathcal{P}$ to give more weight to higher uncertainty values (refer to Sec. 3.2 for details). Fig. 6 illustrates the result of different exponent values $\mathcal{P}$ on the scaling, highlighting the increased emphasis on higher uncertainties for increasing exponent values. As can be seen in those images, larger exponent scaling reduces the focus to few regions with the highest uncertainty (right), while lower exponent values keep a good general focus on diverse areas (left). To extend beyond qualitative observations, we engage in further quantitative analyses to evaluate the benefits of emphasizing higher uncertainty values through exponents and to determine the optimal degree of emphasis. In this context, we perform a series of experiments that combine Entropy-based AL with our UAAL scaling on the CityScapes [10] dataset. These experiments are conducted using four distinct power values. The findings, which are collated in Tab. 5, reveal that an exponent of 2 (i.e. square operation) strikes an ideal balance by accentuating high uncertainty values without excessive amplification.

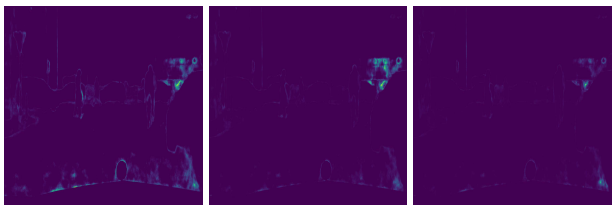

Figure 6. Uncertainty scaling maps for low, medium and high exponent values (left to right), showing increased emphasis on fewer regions with increasing power. Images resized to 1:1 aspect ratio.

| **Budget $\mathcal{B}$** | **Exponent $\mathcal{E}$ for $\mathcal{S}$** | | | |
|---|---|---|---|---|
| | 1 | **2** | 3 | 5 |
| **10 %** | 67.53±0.51 | **67.95±1.81** | 66.45±1.48 | 65.70±0.97 |
| **20 %** | 71.86±0.20 | **72.03±0.55** | 71.31±0.85 | 70.41±0.38 |

Table 5. Prediction quality (mIoU) of a ResNet-50 [21] DeepLabv3+ [7] with Entropy + UAAL using different exponent values $\mathcal{E}$ for the scaling on two budgets of the CityScapes [10] dataset. An exponent of 2 performs best.

### 6.2. Noise Level Robustness

Robustness against hyperparameters is crucial for active learning methods to find informative data even in previously unknown conditions. This section discusses both the theoretical motivation for scaling using noise magnitude and evaluates the impact of different noise magnitudes with the scaling.

#### 6.2.1. Theoretical Foundation

Starting from Eq. 1, and using again the normalized noise $\Delta\theta_k = \eta\|\theta\|\bar{n}$ the variance after the perturbation can be defined as:

$$\sigma^2 = \text{Var}(f(\theta + \Delta\theta, x)) \tag{9}$$

The perturbation is expected to only cause a minor change in output if chosen sufficiently small. Hence, the output change can be expressed as a first order Taylor expansion around the original parameters:

$$f(\theta + \Delta\theta, x) \approx f(\theta, x) + \nabla_\theta f(\theta, x) \cdot \Delta\theta \tag{10}$$

$$\text{Var}(f(\theta + \Delta\theta, x)) \approx \text{Var}(f(\theta, x) + \nabla_\theta f(\theta, x) \cdot \Delta\theta) \tag{11}$$

Now, the variance of the output is the variance of this perturbation:

$$\text{Var}(f(\theta + \Delta\theta, x)) \approx \text{Var}(\nabla_\theta f(\theta, x) \cdot \Delta\theta) \tag{12}$$

Replacing $\Delta\theta$ with its components and extracting $\eta$, we get

$$\text{Var}(f(\theta + \Delta\theta, x)) = \sigma^2 \approx \eta^2 \text{Var}(\nabla_\theta f(\theta, x) \cdot \|\theta\|\bar{n}) \tag{13}$$

#### 6.2.2. Empirical Evaluation

To test our approach's robustness against its hyperparameter kernel noise magnitude $\eta$ (see Eq. 6), we run a study on its impact on a DeepLabv3+ [7] with RN50 [21] backbone across various noise levels on both the ADE20K [53] and CityScapes [10] dataset. The results for both 10 % and 20 % data are summarized in Tab. 6. As can be seen there, neither

noise magnitude dominates. Despite changing $\eta$ more than an order of magnitude (up to $\times 50$ difference), the method does not result in failure of the training and still improves over the baseline, showcasing the robustness of our method towards the hyperparameter $\eta$.

| Budget $\mathcal{B}$ | CityScapes | | | |
|---|---|---|---|---|
| | $\eta = 0.001$ | $\eta = 0.005$ | $\eta = 0.01$ | $\eta = 0.05$ |
| **10 %** | 66.97±0.21 | 67.57±1.27 | **68.30±0.39** | 67.51±0.68 |
| **20 %** | 71.52±0.58 | 71.51±0.45 | 71.54±0.81 | **72.19±0.26** |
| | **ADE20K** | | | |
| | $\eta = 0.001$ | $\eta = 0.005$ | $\eta = 0.01$ | $\eta = 0.05$ |
| **10 %** | 25.20±0.52 | 24.90±0.05 | 24.94±0.16 | **25.25±0.28** |
| **20 %** | **30.79±0.25** | 30.66±0.73 | 30.70±0.23 | 30.41±0.28 |

Table 6. Entropy + UAAL prediction quality (mIoU) using different noise magnitudes $\eta$ for 10 % and 20 % budget on CityScapes [10] and ADE20K [53]. No $\eta$ dominates, choice can matter in 10 % data regime.

### 6.3. Scheduler Functions

This section provides an in-depth look at the uncertainty scheduling functions $\alpha$ employed in our experiments. In particular, we use three categories of scheduling strategies, each with an intuitive rationale for potentially enhancing model performance: increasing, decreasing and combined. The increasing strategies start with a lower influence of uncertainty early in training, gradually raising it as the model learns the target task.

$$\text{Lin. incr.:} \quad \alpha_{li}(\text{epoch}) = \frac{\text{epoch}}{\text{epoch}_{max}} \tag{A1}$$

$$\text{Sine incr.:} \quad \alpha_{si}(\text{epoch}) = \frac{1}{2} \cdot (\sin(\frac{\pi \cdot \text{epoch}}{\text{epoch}_{max}} - \frac{\pi}{2}) + 1) \tag{A2}$$

Conversely, decreasing strategies begin with a higher emphasis on uncertainty, which diminishes over time to focus on the target task, avoiding the neglect of less uncertain regions.

$$\text{Lin. decr.:} \quad \alpha_{ld}(\text{epoch}) = 1 - \frac{\text{epoch}}{\text{epoch}_{max}} \tag{A3}$$

$$\text{Cos. decr.:} \quad \alpha_{cd}(\text{epoch}) = \frac{1}{2} \cdot (\cos(\frac{\pi \cdot \text{epoch}}{\text{epoch}_{max}}) + 1) \tag{A4}$$

This final strategy combines the increasing and decreasing strategies, initially elevating and then reducing the influence of uncertainty, aiming for a balanced focus throughout training.

$$\text{Sin-Cos:} \quad \alpha_{sc}(\text{epoch}) = \frac{1}{2} \cdot (\sin(\frac{2\pi \cdot \text{epoch}}{\text{epoch}_{max}} - \frac{\pi}{2}) + 1) \tag{A5}$$

## 7. Detailed ADE20K Results

In Sec. 4.2 we presented in-depth results for our AL experiments for CityScapes data. This section provides the second half of the presented results for the ADE20K dataset. If not stated otherwise we follow the same training setup as described in Sec. 4.1. The results for the ADE20K dataset on the 10 % and 20 % budget $\mathcal{B}$ are summarized in Tab. 7. The overall pattern for ADE20K mirrors our findings from CityScapes, showing superior performance with our UAAL method in most scenarios. Given that ADE20K is a more challenging benchmark, it is harder for AL methods to exceed the random selection baseline. Despite these challenges, UAAL showcases its robustness and scalability by effectively managing the complexity and achieving performance gains. For example, the Xc65 model combined with MC-Dropout observed a 2.22 p.p. increase over the baseline. Moreover, when UAAL was paired with the Entropy method on the Xc65 model, it surpassed the random baseline - a feat that the standard Entropy AL method did not accomplish. These results highlight the effectiveness of UAAL in enhancing model performance, even in more demanding datasets.

## 8. Additional Images for Qualitative Analysis

In this supplementary section, we provide additional examples to extend the qualitative analysis presented in Sec. 4.4 of the main content. These additional examples (Fig. 7) further illustrate the improvements brought by our UAAL approach in terms of uncertainty map clarity and provide a more crowded scene. Like in the main content, the comparisons between standard training and our UAAL approach consistently show enhanced clarity in the uncertainty maps for the majority of cases, particularly around edges and small details. Looking into the specific AL approaches, we observe superior results when using our UAAL method with Entropy [24] and MC-Dropout [16] and only fall slightly behind when applied to Noise Stability [28] in combination the large Xc65 as can be seen on the very right of the figure.

| Data Selection Method | Training Method | ADE20K ($\mathcal{B} = 10\,\%$) | | | ADE20K ($\mathcal{B} = 20\,\%$) | | |
|---|---|---|---|---|---|---|---|
| | | MobileNetV3 | ResNet-50 | Xception-65 | MobileNetV3 | ResNet-50 | Xception-65 |
| Random | standard | 16.87±0.20 | 25.55±0.19 | 26.02±0.17 | 21.03±0.19 | 30.35±0.59 | 31.09±0.22 |
| Entropy [24] | standard | 17.66±0.29 | 24.62±0.42 | 25.92±0.19 | 21.93±0.27 | 30.26±0.13 | 31.31±0.47 |
| | **UAAL** (ours) | **17.75±0.45** | **24.94±0.16** | **27.04±0.51** | **22.02±0.34** | **30.70±0.23** | **32.13±0.28** |
| MC-Dropout [16] | standard | 15.81±0.32 | 20.23±0.80 | 24.04±0.43 | 20.37±0.35 | 24.27±1.02 | 30.50±0.49 |
| | **UAAL** (ours) | 15.47±0.24 | **22.44±1.23** | 23.98±0.16 | **20.74±0.65** | **26.44±1.53** | 30.32±0.66 |
| Noise Stability [28] | standard | 15.51±0.38 | 21.00±0.71 | 23.82±0.57 | 20.21±0.05 | 27.98±0.13 | 29.67±0.36 |
| | **UAAL** (ours) | **15.62±0.15** | **21.74±0.47** | 23.45±0.44 | 20.20±0.24 | 27.96±0.37 | 29.58±0.36 |

Table 7. Evaluation of image-level uncertainty active learning methods at 10 % and 20 % data budget $\mathcal{B}$ for MobileNetV3, ResNet-50, and Xception-65 model variants on the ADE20K dataset with and without our UAAL training.

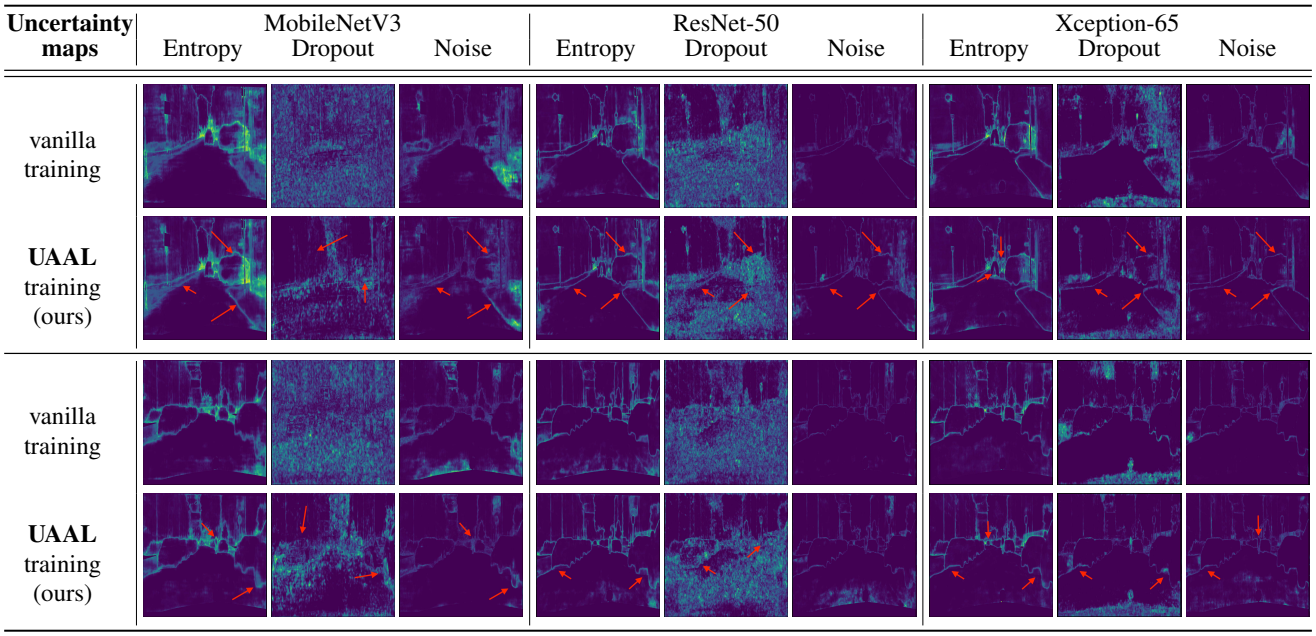


Figure 7. Qualitative comparison of uncertainty maps produced by MobileNetV3, ResNet-50 and Xception-65 DeepLabv3+ variants on two example CityScapes images using standard training (top row) and our UAAL enhanced training (bottom row), showing that the clarity of the uncertainty map improves with our method across all uncertainty methods. Some improvements are pointed out with red arrows. Only in one case our method fails to improve the quality. Images are reshaped to 1:1 aspect ratio for space efficiency.