Uncertainty-aware Cross-Entropy for Semantic Segmentation

Steven Landgraf *, Markus Hillelmann, Kira Wursthorn, Markus Ulrich

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

Keywords: Semantic Segmentation, Uncertainty Quantification, Monte Carlo Dropout, Vision Metrology

Abstract

Deep neural networks have shown exceptional performance in various tasks, but their lack of robustness, reliability, and tendency to be overconfident pose challenges for their deployment in safety-critical applications like autonomous driving. In this regard, quantifying the uncertainty inherent to a model’s prediction is a promising endeavour to address these shortcomings. In this work, we present a novel Uncertainty-aware Cross-Entropy loss (U-CE) that incorporates dynamic predictive uncertainties into the training process by pixel-wise weighting of the well-known cross-entropy loss (CE). Through extensive experimentation, we demonstrate the superiority of U-CE over regular CE training on two benchmark datasets, Cityscapes and ACDC, using two common backbone architectures, ResNet-18 and ResNet-101. With U-CE, we manage to train models that not only improve their segmentation performance but also provide meaningful uncertainties after training. Consequently, we contribute to the development of more robust and reliable segmentation models, ultimately advancing the state-of-the-art in safety-critical applications and beyond. *

1. Introduction

Humans often make poor decisions and reach erroneous conclusions while overestimating their abilities, a phenomenon known as the Dunning-Kruger effect (Kruger and Dunning, 1999). Although deep neural networks are highly effective at solving semantic segmentation problems (Minaee et al., 2022), they also suffer from overconfidence (Guo et al., 2017). Additionally, neural networks lack interpretability (Gawlikowski et al., 2022) and struggle to distinguish between in-domain and out-of-domain samples (Lee et al., 2018). These flaws are particularly relevant in safety-critical applications, such as autonomous driving (McAllister et al., 2017) and medical imaging (Leibig et al., 2017), as well as in computer vision tasks that have high demands on reliability, like industrial inspection (Steger et al., 2018, Heizmann et al., 2022) and automation (Landgraf et al., 2023a, Ulrich and Hillelmann, 2021), where robust predictions are crucial. Misclassifying pixels in these contexts can lead to severe consequences, emphasizing the need for robust and trustworthy segmentation models.

Previous work suggests that quantifying the uncertainty inherent to a model’s prediction is a promising endeavour to enhance the safety and reliability of such applications (Landgraf et al., 2023b, Leibig et al., 2017, Lee et al., 2018, Mukhoti and Gal, 2018, Mukhoti et al., 2023). These uncertainties provide additional insights beyond the common softmax probabilities, revealing regions where the model is indecisive and likely to make errors. Surprisingly, the utilization of these uncertainties during the training of segmentation models has not been thoroughly explored.

In this work, we present a novel Uncertainty-aware Cross-Entropy loss, referred to as U-CE, that addresses this gap by incorporating dynamic uncertainty estimates into the training process as shown in Figure 1. Through pixel-wise uncertainty weighting of the well-known cross-entropy loss (CE), we harness the valuable insights provided by the uncertainties for more effective training. With U-CE, we manage to train models that are naturally capable of predicting meaningful uncertainties after training while simultaneously improving their segmentation performance.

Our contributions can be summarized as follows: Firstly, we propose the U-CE loss function, which utilizes uncertainty estimates to guide the optimization process, emphasizing regions with high uncertainties. Secondly, we conduct extensive experiments on two benchmark datasets, Cityscapes (Cords et al., 2016) and ACDC (Sakaridis et al., 2021), using two common backbones, ResNet-18 and ResNet-101 (He et al., 2016), demonstrating the superiority of U-CE over regular CE training. Lastly, we present additional insights, limitations, and potential improvements for U-CE through multiple ablation studies and a thorough discussion.

2. Related Work

Hereinafter, we briefly review the related work on uncertainty quantification and uncertainty-aware segmentation.

*Corresponding author
°Code is available at: https://github.com/StevenLandgraf/U-CE
2.1 Uncertainty Quantification

Deep neural networks, with their millions of model parameters and non-linearities, have proven effective in solving complex tasks in natural language processing (Otter et al., 2020) and computer vision, like semantic segmentation (Minaee et al., 2022). Unfortunately, due to their complexity, the computation of the exact posterior probability distribution of the network’s output is infeasible (Blundell et al., 2015, Loquercio et al., 2020). Consequently, approximate uncertainty quantification methods are employed to offer a practical solution to tackle the intractability of the exact posterior distribution. The most prominent methods include Bayesian Neural Networks (MacKay, 1992), Monte Carlo Dropout (Gal and Ghahramani, 2016), and Deep Ensembles (Lakshminarayanan et al., 2017). We will refer to these methods as traditional uncertainty quantification techniques throughout the following.

A mathematically grounded, though computationally complex, approach to uncertainty quantification is provided by Bayesian Neural Networks, which transform a deterministic network into a stochastic one using probabilistic distributions placed over the activations or the weights (Jospin et al., 2022). For instance, Bayes by Backprob (Blundell et al., 2015) employs variational inference to learn approximate distributions over the weights. These can be used to create an ensemble of models with differently sampled weights to approximate the posterior distribution of the predictions.

Gal and Ghahramani simplify this approximation process by using Monte Carlo Dropout (Gal and Ghahramani, 2016). While dropout is usually applied as a regularization technique (Srivastava et al., 2014), Monte Carlo Dropout uses this concept to sample from the posterior distribution of a network’s prediction at test time. In its original form, Monte Carlo Dropout only captures the epistemic uncertainty inherent to the model. To obtain a more comprehensive measure of uncertainty that includes the aleatoric uncertainty, which captures the noise inherent in the observations, Monte Carlo Dropout can be combined with learned uncertainty predictions and assumed density filtering (Gast and Roth, 2018, Kendall and Gal, 2017, Loquercio et al., 2020).

The current state-of-the-art uncertainty quantification method are Deep Ensembles, which consist of an ensemble of trained models that generate diverse predictions at test time (Lakshminarayanan et al., 2017). Due to the introduction of randomness through random weight initialization or different data augmentations across ensemble members (Fort et al., 2020), Deep Ensembles are well-calibrated (Lakshminarayanan et al., 2017). Multiple studies demonstrated that Deep Ensembles generally outperform other uncertainty quantification methods across varying tasks (Ovadia et al., 2019, Wursthorn et al., 2022, Gustafsson et al., 2020). However, this performance gain is associated with high computational cost.

In addition to the aforementioned approximate uncertainty quantification methods, there has been a growing interest in deterministic single forward-pass approaches, which offer advantages in terms of memory usage and inference time. For example, Van Amersfoort et al. (Van Amersfoort et al., 2020) and Liu et al. (Liu et al., 2020) explore the concept of distance-aware output layers. While these methods demonstrate good performance, they are not competitive with the current state-of-the-art and require significant modifications to the training process (Mukhoti et al., 2023). Another approach, proposed by Mukhoti et al. (Mukhoti et al., 2023), simplifies the two previous methods by employing Gaussian Discriminant Analysis for feature-space density estimation after training. Although they perform on par with Deep Ensembles in some settings, their approach still necessitates a more sophisticated training approach. Additionally, fitting the feature-space density estimator is only possible after training, which is not suitable for U-CE where meaningful uncertainties are required during training.

Overall, uncertainty quantification remains an active and evolving field of research, with various approaches offering their own advantages and disadvantages. For our specific case, Monte Carlo Dropout emerges as the preferred option due to its ease of use, minimal impact on the training process, and computational efficiency compared to Deep Ensembles. Through Monte Carlo Dropout sampling, we can compute the predictive uncertainty to apply pixel-wise weighting of the well-known cross-entropy loss. With predictive uncertainties, we refer to the standard deviation of the softmax probabilities of the predicted class provided by Monte Carlo Dropout sampling.

2.2 Uncertainty-aware Segmentation

In the domain of uncertainty-aware segmentation, researchers have explored various techniques to incorporate uncertainty measures into the training process. While traditional uncertainty quantification methods have successfully been employed in tasks such as visual bias mitigation in classification (Stone et al., 2022), these techniques have been largely overlooked or underutilized in the field of semantic segmentation. We provide an overview of notable works that leverage uncertainty-aware techniques for segmentation tasks in various domains. Additionally, we discuss how U-CE addresses the gap towards full utilization of traditional uncertainty quantification methods during training.

Some of the earlier work on more effective training has originally been designed for object detection. For example, Lin et al. (Lin et al., 2017) introduced the Focal Loss (FL) that down-weights the contribution of easy examples to shift the focus more towards hard examples. Another closely related technique is online hard example mining by Srivastava et al. (Srivastava et al., 2016). They propose to automatically select hard examples to only learn from them and completely ignore the easy examples. By now, both methods have been successfully adapted for semantic segmentation (Jadon, 2020, Wang et al., 2022).

Another line of work focuses on the identification and compensation of ambiguities and label noise. Kaiser et al. (Kaiser et al., 2023) propose adding a learned bias to a network’s logits and introducing a novel uncertainty branch to induce the compensation bias only to relevant regions. However, unlike U-CE, their approach does not utilize uncertainties to make training more robust, rather they aim to avoid new noise during data annotation.

More closely related to our work, Bischke et al. (Bischke et al., 2018) and Bressan et al. (Bressan et al., 2022) propose to leverage uncertainties to improve training on imbalanced aerial image datasets. The former use the per-class uncertainty of the model together with the median frequency to balance training (Bischke et al., 2018). We argue that dynamically weighting each pixel individually during training, which is what U-CE does, is even more valuable. The latter utilize pixel-wise weights, but only consider the class and labeling uncertainty.
3. Methodology

In the following, we provide an overview of U-CE, explain our novel uncertainty-aware cross-entropy loss and outline the implementation details.

3.1 Overview

The central idea of U-CE is to incorporate predictive uncertainties into the training process to enhance segmentation performance. As depicted in Figure 2, we propose two simple yet highly effective adaptations to the regular training process:

1. During training, we sample from the posterior distribution with Monte Carlo Dropout to obtain predictive uncertainties alongside the regular segmentation prediction.
2. We apply pixel-wise weighting to the regular cross-entropy loss based on the collected uncertainties.

To compute predictive uncertainties during training, we choose Monte Carlo Dropout. It is straightforward to implement, requires minimal tuning, and is computationally more efficient than Deep Ensembles. However, it is worth noting that other uncertainty quantification methods could also be utilized for U-CE. Exploring these alternatives is an interesting avenue for future work, which we will discuss in Section 5.

3.2 Uncertainty-aware Cross-Entropy

Segmentation Sampling. In contrast to typical usage of Monte Carlo Dropout, U-CE incorporates the sampling process from the posterior distribution not only at test time but also during training. To compute the necessary uncertainties for our uncertainty-aware cross-entropy loss, we perform $\beta$ sampling iterations at each training step. This generates $\beta$ segmentation samples in addition to the regular segmentation prediction. Notably, gradient computation is disabled during the sampling process as it is unnecessary for backward propagation, which relies solely on the regular segmentation prediction. By disabling gradient computation during sampling, we reduce the additional computational overhead of U-CE in terms of training time and GPU memory usage.

Uncertainty-aware Cross-Entropy Loss. The final objective function of U-CE builds upon the well-known categorical cross-entropy loss and can be defined as:

$$L_{u-ce} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} y_{n,c} \cdot \log(p_{n,c}),$$

where $L_{u-ce}$ is the uncertainty-aware cross-entropy 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 ground truth label, $p_{n,c}$ is the respective predicted softmax probability, and $w_n$ represents the pixel-wise uncertainty weight. It is worth noting that Equation 1 simplifies to the regular cross-entropy loss by setting $w_n$ to one for all pixels.

Pixel-wise Uncertainty Weight. The pixel-wise uncertainty weight $w_n$ can be formulated as:

$$w_n = (1 + \sigma_n)^\alpha,$$

where $\sigma_n$ denotes the predictive uncertainty, and $\alpha$ controls the influence of the uncertainties in an exponential manner.

Figure 2. A schematic overview of the training process of U-CE. U-CE integrates the predictive uncertainties of a Monte Carlo Dropout (MC-Dropout) model into the training process to enhance segmentation performance. In comparison to most applications of Monte Carlo Dropout, U-CE utilizes the uncertainties not only at test time but also dynamically during training by applying pixel-wise weighting to the regular cross-entropy loss.
predictive uncertainty $\sigma$ represents the standard deviation of the softmax probabilities of the predicted class of the segmentation samples.

### 4. Experiments
In this section, we conduct an extensive range of experiments to demonstrate the value of incorporating predictive uncertainties into the training process. Firstly, we provide quantitative results comparing regular CE to U-CE under diverse settings. Secondly, we analyze qualitative examples. Lastly, we provide multiple ablation studies.

#### 4.1 Setup
**Architecture.** For all of our experiments, we employ DeeplabV3+ (Chen et al., 2018) as the decoder and either a ResNet-18 or ResNet-101 (He et al., 2016) as the encoder. Both backbones are commonly used for semantic segmentation (Maina et al., 2022, Zhang et al., 2020), making our work highly comparable and serving as an excellent baseline for future research.

**Monte Carlo Dropout.** In order to convert our architectures into Monte Carlo Dropout models, we add a dropout layer after each of the four residual block layers of the ResNets, inspired by Kendall et al. (Kendall et al., 2015) and Gustafsson et al. (Gustafsson et al., 2020).

**Training.** For all training processes, we use a Stochastic Gradient Descent (SGD) optimizer (Robbins and Monro, 1951) with a base learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001. Additionally, we multiply the learning rate of the decoder and segmentation head by ten. Finally, we employ polynomial learning rate scheduling to decay the initial learning rate during the training process, following the formula:

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

where $lr$ is the current learning rate, and $lr_{base}$ is the initial base learning rate. In all training processes, we use a batch size of 16 and train on four NVIDIA A100 GPUs with 40 GB of memory using mixed precision (Micikevicius et al., 2017).

**Datasets.** All of our experiments are based on either the Cityscapes dataset (Cordts et al., 2016) or the ACDC dataset (Sakaridis et al., 2021). Both datasets are publicly available street scene datasets aimed at advancing the current state-of-the-art in autonomous driving. The former consists of 2975 training images, 500 validation images, and 1525 test images. The latter contains 1600 training images, 406 validation images, and 2000 test images. Although both datasets share the same 19 evaluation classes and a void class, the ACDC dataset exclusively focuses on four adverse conditions: fog, nighttime, rain, and snow.

**Data Augmentations.** To prevent overfitting, we apply a common data augmentation strategy for all training procedures, regardless of the dataset or architecture used. The strategy includes the following steps:

1. Random scaling with a factor between 0.5 and 2.0.
2. Random cropping with a crop size of 768 $\times$ 768 pixels.
3. Random horizontal flipping with a flip chance of 50%.

#### 4.2 Quantitative Evaluation
Table 1 and 2 outline a quantitative comparison between regular CE and U-CE on the Cityscapes dataset (Cordts et al., 2016) for different dropout ratios. The provided numbers represent the mIoU ↑ in %. Best respective results are marked in bold.

**Evaluation.** Since both test splits are withheld for benchmarking purposes, we utilize the validation images for testing in all our experiments. Unless otherwise specified, we only report single forward pass results based on the original validation images without resizing or sampling for a fair comparison between all of the models. Also, we set the number of segmentation samples $\beta$ to ten by default.

**Metrics.** For quantitative evaluations, we primarily report the mean Intersection over Union (mIoU), also known as the Jaccard Index, to measure the segmentation performance. In addition to the mIoU, we also utilize the Expected Calibration Error (ECE) (Naeini et al., 2015) to evaluate the calibration as well as the mean class-wise predictive uncertainty (mUnc) to quantitatively compare the resulting uncertainties.

#### 4.3 Qualitative Evaluation
Tables 1 and 2 outline a quantitative comparison between FL (Lin et al., 2017), regular CE, and our proposed U-CE loss using two different $\alpha$ values for various dropout ratios and training lengths on the Cityscapes (Cordts et al., 2016) and ACDC (Sakaridis et al., 2021) datasets. For FL, we followed the original publication and set the focusing parameter $\gamma$ to 2.0 as this worked best in their experiments (Lin et al., 2017).

FL (Lin et al., 2017) performed the worst in all our experiments, possibly due to insufficient hyperparameter tuning. Re-
Table 3 provides additional information on the ECE and mUnc for CE and U-CE using a dropout ratio of 20%. In comparison to regular CE and U-CE\(_{\alpha=1}\), which exhibit similar results, U-CE\(_{\alpha=10}\) not only improves segmentation performance but also yields slightly better calibrated networks, as measured by the ECE. Moreover, the mUnc is also slightly lower for U-CE\(_{\alpha=10}\).

Overall, Tables 1, 2 and 3 provide strong evidence for the effectiveness of leveraging predictive uncertainties in the training process.

### 4.3 Qualitative Evaluation

In addition to the quantitative evaluation, we also provide qualitative examples in Figure 3 showing the original input image, the corresponding ground truth label, the model’s segmentation prediction, a binary accuracy map, and the student’s predictive uncertainty. The first three rows depict results from models with a ResNet-18 backbone and dropout ratio of 20%, trained for 200 epochs using a ResNet-18 with a dropout ratio of 40%. On average, U-CE\(_{\alpha=10}\) outperforms CE by 2.0% on Cityscapes (Cordts et al., 2016) and by 4.6% on ACDC (Sakaridis et al., 2021). Interestingly, U-CE\(_{\alpha=1}\) also matches or improves upon regular CE training in most cases. On average, U-CE\(_{\alpha=1}\) outperforms CE by 0.3% on Cityscapes and by 1.3% on ACDC.

Remarkably, U-CE\(_{\alpha=10}\) achieves the highest mIoU across all dropout ratios, even outperforming dropout-free baseline models in most cases. Notably, U-CE\(_{\alpha=10}\) achieves a maximum improvement of up to 9.3% over regular CE when training on ACDC (Sakaridis et al., 2021) for 200 epochs using a ResNet-18 with a dropout ratio of 40%. On average, U-CE\(_{\alpha=10}\) outperforms CE by 2.0% on Cityscapes (Cordts et al., 2016) and by 4.6% on ACDC (Sakaridis et al., 2021). Interestingly, U-CE\(_{\alpha=1}\) also matches or improves upon regular CE training in most cases. On average, U-CE\(_{\alpha=1}\) outperforms CE by 0.3% on Cityscapes and by 1.3% on ACDC.

Figure 3. Example images from the Cityscapes and ACDC validation set (a), corresponding ground truth labels (b), the model’s segmentation predictions (c), a binary accuracy map (d), and the predictive uncertainty (e). White pixels in the binary accuracy map are either incorrect predictions or void classes, which appear black in the ground truth label. For the uncertainty prediction, brighter pixels represent higher predictive uncertainties. The first three rows depict results from models with a ResNet-18 backbone and dropout ratio of 20%, trained for 200 epochs on Cityscapes (Cordts et al., 2016). The last three rows show examples from models using a ResNet-101 backbone and a dropout ratio of 20%, trained for 500 epochs on the ACDC dataset (Sakaridis et al., 2021). Generally, for large areas and well-represented classes like road, building, sky, and car, all models perform exceptionally well with minimal errors. Furthermore, there is a strong correlation between the binary accuracy map and the predictive uncertainty, indicating that all models provide meaningful uncertainties.

Nonetheless, there are nuanced differences between the models. For example, in the first two rows of Figure 3, which represent models trained CE and U-CE\(_{\alpha=1}\), there are noticeable misclassifications on top of the human standing in front of the truck. Naturally, this area is also accompanied with high uncertainties. In contrast, the model trained with U-CE\(_{\alpha=10}\) exhibits significantly fewer difficulties, resulting in a better segmentation prediction and lower uncertainties.

A similar situation is observable in the last three rows, showing examples from the more challenging ACDC dataset (Sakaridis et al., 2021).
Table 4. Ablation study on the impact of $\alpha$. The provided numbers represent the mIoU $\uparrow$. Best results are marked in **bold**.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>RN18 (Cityscapes)</td>
<td>69.5</td>
<td>70.0</td>
<td>70.7</td>
<td>71.2</td>
<td>71.5</td>
<td><strong>71.8</strong></td>
<td>71.0</td>
<td>47.0</td>
<td>70.9</td>
</tr>
<tr>
<td>RN101 (Cityscapes)</td>
<td>74.8</td>
<td>75.2</td>
<td>75.6</td>
<td>76.1</td>
<td>76.4</td>
<td><strong>76.6</strong></td>
<td>76.3</td>
<td>75.8</td>
<td>72.6</td>
</tr>
<tr>
<td>RN18 (ACDC)</td>
<td>56.1</td>
<td>56.9</td>
<td>57.6</td>
<td>58.8</td>
<td>58.8</td>
<td><strong>60.5</strong></td>
<td>60.3</td>
<td>60.1</td>
<td>37.5</td>
</tr>
<tr>
<td>RN101 (ACDC)</td>
<td>65.0</td>
<td>65.0</td>
<td>65.7</td>
<td>65.5</td>
<td>66.0</td>
<td><strong>66.7</strong></td>
<td>64.5</td>
<td>64.5</td>
<td>19.9</td>
</tr>
</tbody>
</table>

In summary, the qualitative findings presented in Figure 3 concur with our quantitative evaluation, manifesting the efficacy of U-CE across different datasets and architectures.

4.4 Ablation Studies

In addition to the quantitative and qualitative evaluation, we also present multiple ablation studies. Unless otherwise noted, we confined all of the ablation studies to models that use a ResNet-18 as the backbone, have a dropout ratio of 20%, and were trained for 200 epochs.

**Impact of $\alpha$.** The most influential hyperparameter of U-CE is $\alpha$ as it exponentially controls the weighting of the CE loss. Table 4 demonstrates the impact of different $\alpha$ values on the mIoU for both backbones, ResNet-18 (RN18) and ResNet-101 (RN101), on both Cityscapes and ACDC. Evidently, the segmentation performance consistently improves as $\alpha$ increases until it reaches ten, which stands as the best value in three out of four cases across the two datasets and architectures. Thus, using ten as the default value for $\alpha$ seems to be a fair estimation to achieve the best results, not only for the mentioned cases but potentially for other applications as well. Further increasing $\alpha$ leads to a degradation in mIoU. Additionally, training becomes more unstable as models overly focus on uncertain pixels, resulting in some models failing to converge properly. Nonetheless, U-CE exhibits robustness against changes in $\alpha$, offering a wide range of valid hyperparameters that lead to improved segmentation results compared to regular CE training.

**Impact of $\beta$.** Table 5 exhibits another ablation study on the number of segmentation samples $\beta$. Interestingly, there is no clear benefit of sampling more often than six times, especially with regard to the training time. As indicated by the training times, U-CE$_{\beta=10}$ increases the necessary training time by approximately 10%, whereas U-CE$_{\beta=10}$ extends it by roughly 35%. For comparison, Gal and Ghahramani (Gal and Ghahramani, 2016) recommend sampling ten times to get a reasonable estimation of the predictive mean and uncertainty.

**Impact of Data Augmentations.** The impact of various data augmentation strategies on CE and U-CE is demonstrated in Table 6. The results show that incorporating additional data augmentations on top of the baseline strategy of random cropping with a crop size of 768 $\times$ 768 pixels improves the mIoU across the board. More importantly, this ablation study confirms that U-CE consistently outperforms CE across different data augmentation strategies, indicating its effectiveness in improving segmentation performance.

**Impact of $lr_{\text{base}}$.** Table 7 shows the ablation study on the base learning rate $lr_{\text{base}}$. The most notable comparison is between regular CE and U-CE$_{\alpha=1}$, which demonstrates that U-CE is not limited to specific learning rates. U-CE$_{\alpha=1}$ consistently outperforms regular CE for all examined base learning rates, despite increasing the training loss by approximately 9% as indicated by the mUnc in Table 3. Moreover, U-CE$_{\alpha=10}$ exceeds the results of CE and U-CE$_{\alpha=1}$ for all base learning rates except $10^{-5}$, which caused divergence. Overall, this ablation study confirms the value of leveraging predictive uncertainties during training, irrespective of the learning rate, which is arguably the single most important hyperparameter in deep learning (Bengio, 2012).

5. Discussion

In contrast to previous approaches, U-CE fully leverages predictive uncertainties obtained by Monte Carlo Dropout during training. As a result, we manage to train models that not only improve their segmentation performance but are also naturally capable of predicting meaningful uncertainties after training as well.

While U-CE appears to have no apparent shortcomings, except for a minor increase in training time, we acknowledge the need for a transparent discussion about its potential limitations. Our aim is to effectively guide future work in pushing the boundaries of state-of-the-art techniques, especially in safety-critical applications like autonomous driving.

**Limitations.** One limitation of U-CE arises in the absence of densely annotated ground truth labels. If most pixels are either
labeled as background or designated to be ignored while training, U-CE will likely offer next to no benefit, except for a higher loss around object boundaries. Additionally, U-CE may not contribute to improved segmentation performance if the network is already overfitting the training data. Having said that, the impact of U-CE on generalization needs further examination.

**Future Work.** With regards to future work, we have multiple suggestions that might be worth investigating. Potentially, the results of U-CE could be further improved if the quality of the uncertainty estimates would be better. Therefore, it would be interesting to integrate Deep Ensembles (Lakshminarayanan et al., 2017), the state-of-the-art uncertainty quantification method (Ovadia et al., 2019, Wursthorn et al., 2022, Gustafsson et al., 2020), with U-CE, which we could not realize because of computational restraints. On a similar note, it could be worth employing warmup epochs, which we omitted to refrain from introducing another hyperparameter. Additionally, we would like to see α removed from U-CE by incorporating statistical hypothesis testing. This would be beneficial in two ways: Firstly, it would remove the most influential hyperparameter of U-CE. Secondly, and maybe more importantly, it would leverage all of the available uncertainties and not just the predictive uncertainty. Finally, we encourage other researchers to incorporate U-CE into state-of-the-art semantic segmentation approaches and to explore its usefulness in other computer vision tasks that rely on pixel-wise predictions, such as depth estimation.

Overall, we believe that U-CE presents a promising paradigm in semantic segmentation by dynamically leveraging uncertainties to create more robust and reliable models. Despite a minor increase in training time and room for further improvement, we see no reason not to employ U-CE in comparison to regular CE.

### 6. Conclusion

In this paper, we introduced U-CE, a novel uncertainty-aware cross-entropy loss for semantic segmentation. U-CE incorporates predictive uncertainties, based on Monte Carlo Dropout, into the training process through pixel-wise weighting of the regular cross-entropy loss. As a result, we manage to train models that are naturally capable of predicting meaningful uncertainties after training while simultaneously improving their segmentation performance. Through extensive experimentation on the Cityscapes and ACDC datasets using ResNet-18 and ResNet-101 architectures, we demonstrated the superiority of U-CE over regular cross-entropy training.

We hope that U-CE and our thorough discussion of potential limitations and future work contribute to the development of more robust and trustworthy segmentation models, ultimately advancing the state-of-the-art in safety-critical applications and beyond.

### References


