Semantic Segmentation on MUAD Dataset: Calibration, Uncertainty, and OOD Detection

Team: Yulong MA, Boyuan ZHANG, Huanshan HUANG

Abstract

This report summarizes experiments on semantic segmentation. We use the Multiple Uncertainties for Autonomous Driving (MUAD) dataset. We train a UNet model and a Deep Ensemble. These models segment urban driving scenes. We investigate model calibration via Expected Calibration Error (ECE). We estimate uncertainty through Monte Carlo (MC) Dropout. We perform Out-of-Distribution (OOD) detection using predictive entropy. Calibration improves from a single model to an ensemble. A progressive reduction in ECE block gaps shows this improvement. We discuss overfitting behavior in training curves. We explain the relationship between Neural Collapse (NC) and OOD detection.

1 Experiment Background

The dataset is the MUAD dataset. It is a small version. It is a synthetic benchmark. It simulates diverse driving conditions. It provides pixel-wise ground truth. It has 19 semantic classes. These classes follow the Cityscapes label set. The dataset has train, val, and ood subsets. The OOD split contains scenes with distributional shifts. These shifts do not appear during training.

2 Technical Detail

2.1 Model Architecture

The segmentation model is a UNet architecture. It is a fully convolutional network. It has a symmetric encoder-decoder structure. It uses skip connections.

- Encoder: It captures semantic context. It uses successive downsampling convolutional blocks.
- Decoder: It enables precise spatial localization. It uses upsampling. Skip connections concatenate encoder features at each resolution.
- Dropout: The model applies a 0.1 dropout rate. This provides regularization. The model reuses this rate during inference for MC Dropout sampling.

2.2 Training Setup

- Optimizer: We use Stochastic Gradient Descent (SGD). The momentum is 0.9. The weight decay is 0.0001.
- Loss Function: We use Cross Entropy Loss.
- Learning Rate Scheduler: We use StepLR. The step size is 10 epochs. The decay factor is 0.1. The initial learning rate is 0.01.
- Epochs: We train each model for 100 epochs. The ensemble uses 5 models.
- Metric: We use Mean Intersection over Union (mIoU). We evaluate it on train and val splits.

3 Results Analysis

3.1 Training Curves and Overfitting

Figure 1 shows the loss and mIoU evolution. It covers 100 epochs for a single UNet model.

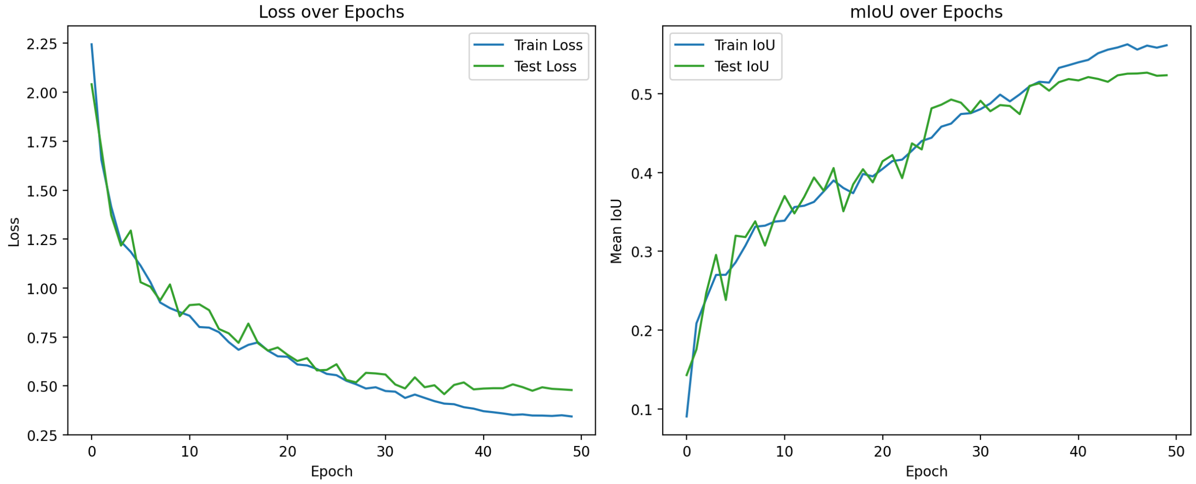

Figure 1: The left side shows training and validation loss over 100 epochs. The right side shows mIoU. The growing gap between train and val mIoU indicates overfitting.

Observations:

- Both losses decrease steadily. The validation loss stabilizes around 0.65. The training loss continues falling to 0.2.
- The training mIoU converges to 0.63. The validation mIoU plateaus near 0.50.
- A persistent gap of 0.13 exists. The model memorizes training patterns. These patterns do not generalize. Stronger data augmentation could mitigate this.

3.2 Ensemble Baseline: MCP Visualization

Figure 2 shows the Maximum Class Probability (MCP). It displays predicted segmentation, ground truth, and entropy maps. This uses a representative In-Distribution (ID) validation sample. High entropy concentrates at object boundaries.

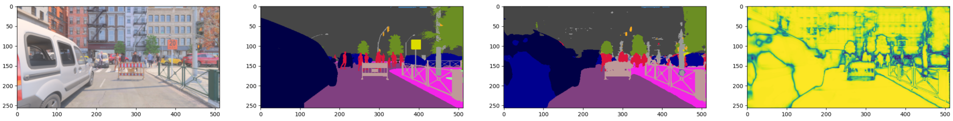

Figure 2: The images show input, ground truth, ensemble prediction, and entropy map from left to right. Bright regions indicate high entropy at boundaries.

The Deep Ensemble achieved a mIoU of 0.6558. It evaluated the ID validation set. It outperformed single models. It provided calibrated uncertainty estimates.

4 ECE Analysis: Calibration Across Configurations

ECE measures confidence alignment with actual accuracy. A calibrated model lies on the diagonal of a reliability diagram. The equation defines ECE:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

The size of each pink gap block reflects local miscalibration. The blocks progressively shrink. This happens as the model improves from a single model to an ensemble.

4.1 Before Calibration: Single Model

Figure 3 shows the reliability diagram. It evaluates a single UNet model.

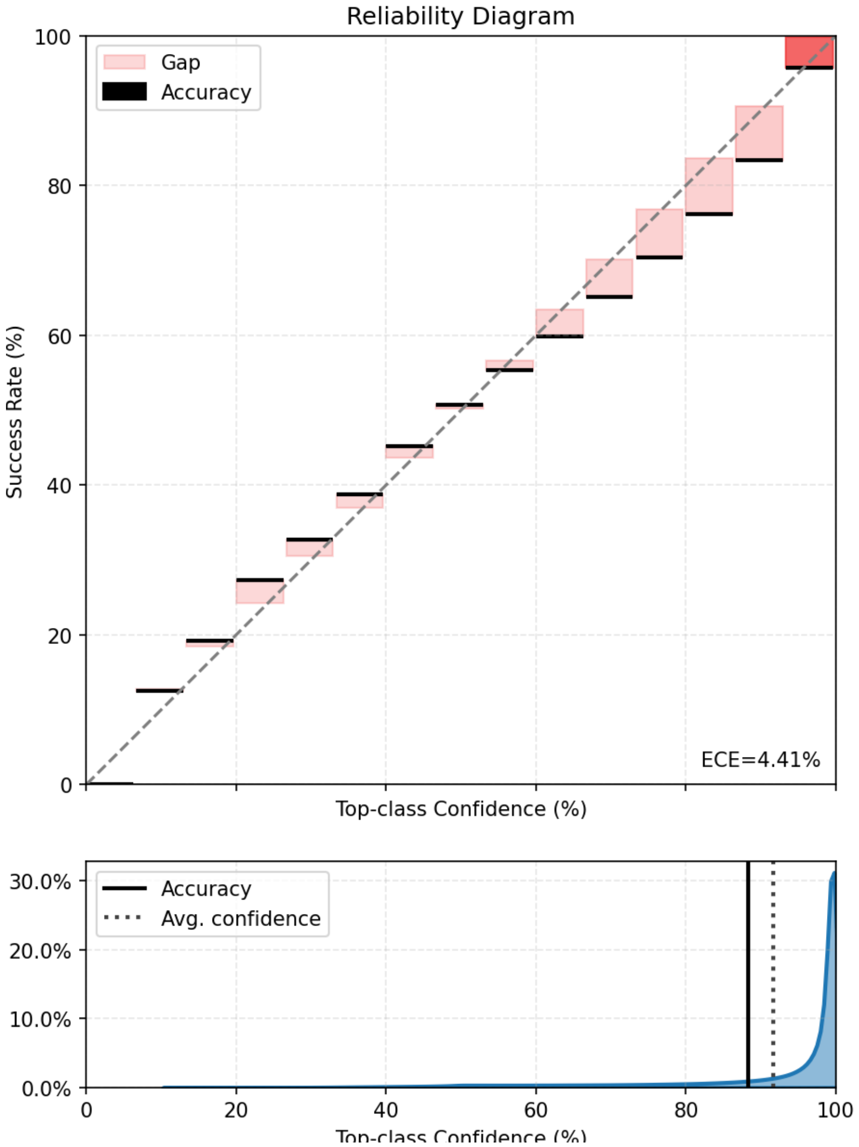

Figure 3: This is the reliability diagram before calibration. ECE is 4.41 percent. Large pink gaps indicate overconfidence in top bins.

Observations:

- The model is underconfident below 70 percent confidence. It becomes overconfident in top bins.
- The largest block sits in the top bin. It represents most predictions. It contributes heavily to the total ECE.
- ECE is 4.41 percent.

4.2 After Ensemble Averaging

Figure 4 shows the reliability diagram. It evaluates the Deep Ensemble. It averages softmax probabilities.

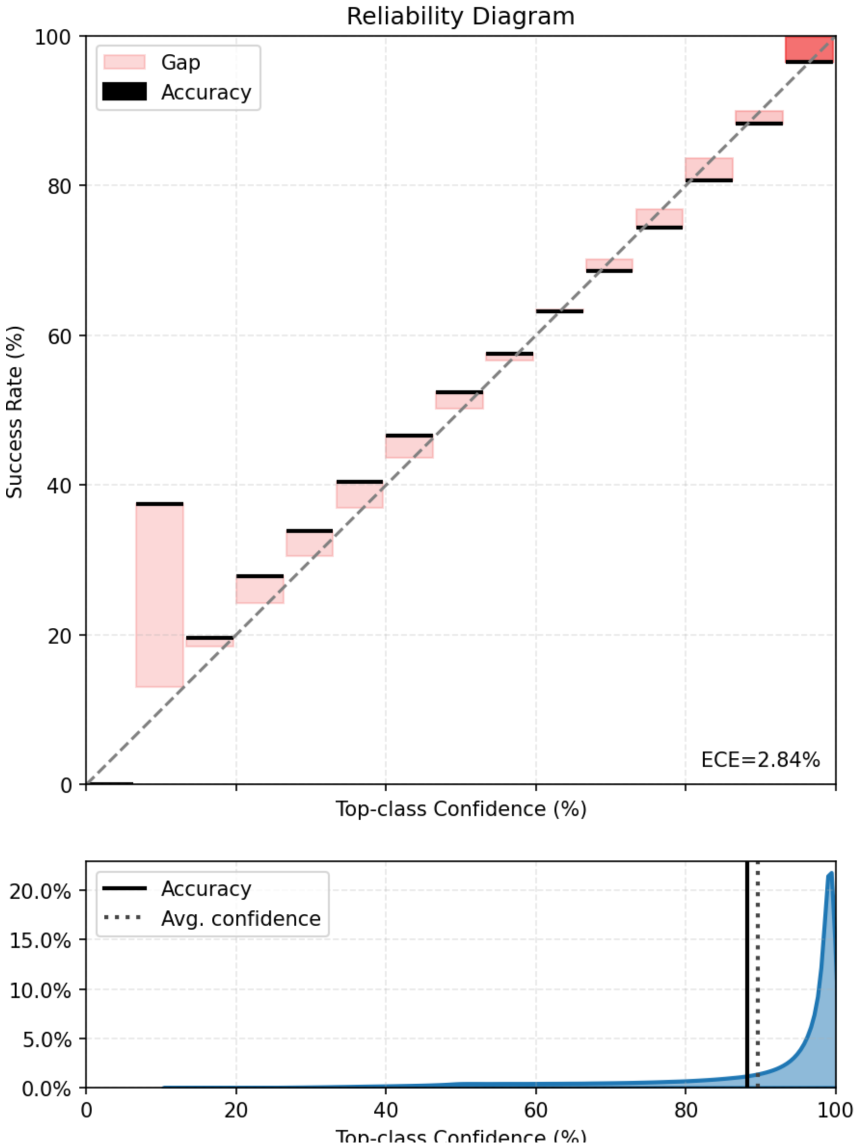

Figure 4: This is the reliability diagram after Deep Ensemble. ECE is 2.84 percent. The pink gap blocks are smaller.

Observations:

- Ensemble averaging forces predictions toward moderate confidence values. It breaks extreme overconfidence.
- The gap blocks shrink noticeably. Ensemble disagreement regularizes confidence.
- ECE improves to 2.84 percent.
- The confidence histogram becomes slightly more spread out.

4.3 After Temperature Scaling

Figure 5 shows the reliability diagram. It evaluates temperature scaling on ensemble outputs.

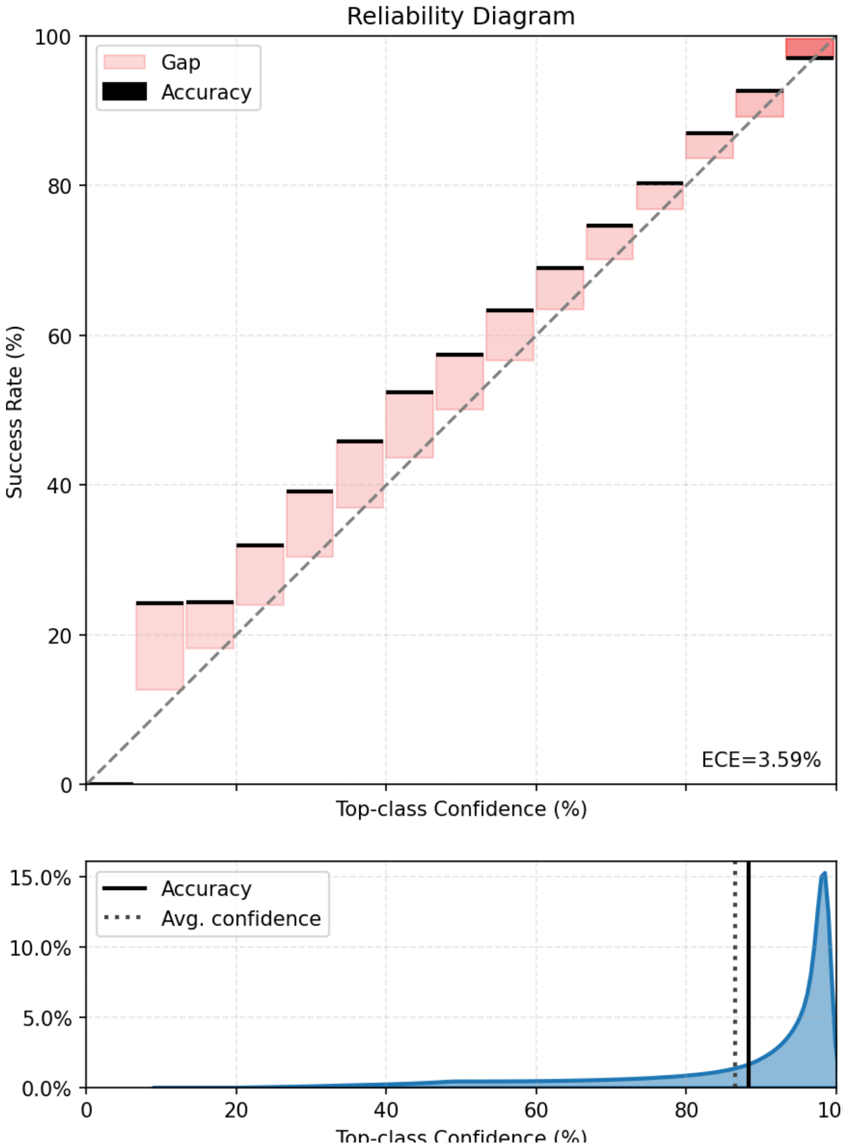

Figure 5: This is the reliability diagram after temperature scaling. ECE is 3.59 percent. The blocks are uniform but slightly larger.

Observations:

- Temperature scaling softens the logits. It increases entropy. It shifts predictions away from extremes.
- ECE becomes 3.59 percent. The imperfect temperature parameter undoes some ensemble benefits.
- The gap blocks become uniform. This helps threshold-sensitive applications.

4.4 Summary: ECE Block Evolution

Configuration	ECE Percent	Block Size Trend
Single Model	4.41	Large uneven gaps
Deep Ensemble	2.84	Smaller reduced bins
Ensemble and Scaling	3.59	More uniform gaps

Table 1: This compares ECE across configurations. Ensemble diversity smooths overconfident predictions.

Ensemble diversity acts as an implicit calibrator. Individual models show overconfidence in different regions. Averaging their probabilities corrects this issue.

5 OOD Detection via MC Dropout

MC Dropout estimates uncertainty at inference. It performs multiple forward passes. Predictive entropy serves as the uncertainty score. The equation defines predictive entropy:

$$H = -\sum_{c=1}^{C} \bar{p}_c \log \bar{p}_c$$

We compare sample counts for ID and OOD images.

5.1 In-Distribution Samples

Figures 6 and 7 show ID validation images.

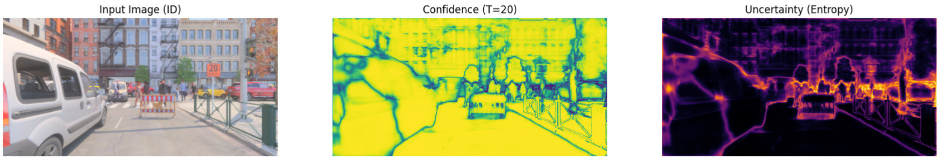

Figure 6: This shows ID sample 1 with 20 MC samples. The left is input. The center is confidence. The right is entropy. Uncertainty concentrates at boundaries.

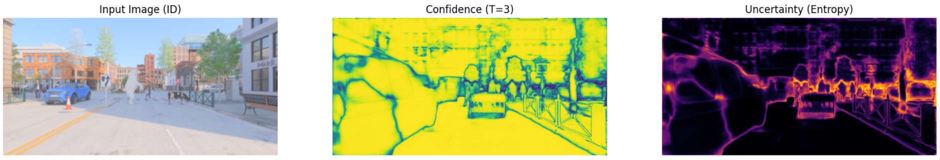

Figure 7: This shows ID sample 2 with 3 MC samples. Road regions have low entropy. Object boundaries show elevated uncertainty.

5.2 Effect of Number of MC Samples

Figure 8 compares confidence maps. It evaluates 3 samples against 20 samples.

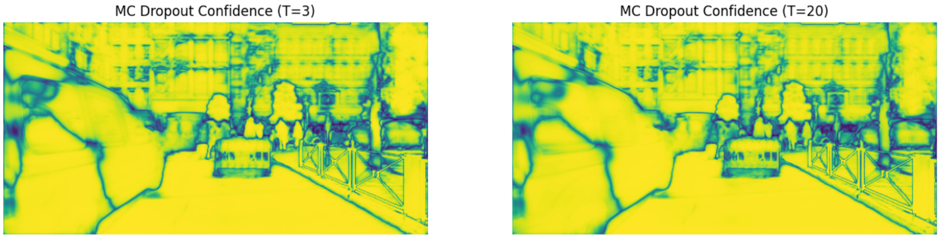

Figure 8: This compares 3 samples and 20 samples. Increasing samples yields smoother estimates. Road regions consolidate to uniform high confidence.

Low sample counts produce noisy maps. High sample counts reduce variance. High confidence regions become uniform. Ten samples provide reliable estimates.

5.3 OOD Samples

Figures 9 and 10 show an OOD image.

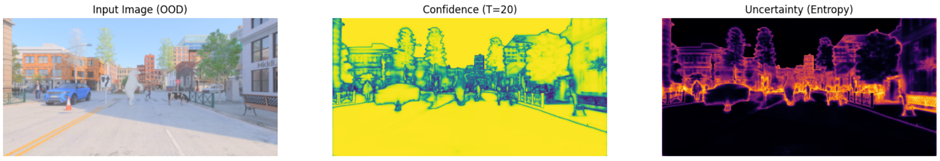

Figure 9: This shows an OOD sample with 20 MC samples. The right map shows elevated uncertainty across the entire scene.

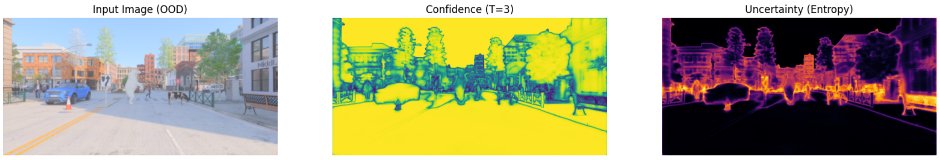

Figure 10: This shows an OOD sample with 3 MC samples. The upper half shows high entropy compared to the bottom region.

OOD entropy maps differ from ID maps. ID images localize entropy at boundaries. OOD images spread entropy across the scene. This global elevation enables threshold-based detection. A system flags a sample as OOD if the mean entropy exceeds a threshold.

6 Neural Collapse

Feature representations exhibit NC near the end of training. Class-conditional feature means converge. Within-class variability vanishes. The training mIoU reached 0.63. The validation mIoU stalled at 0.50. We hypothesize the encoder features collapsed toward class prototypes.

This affects OOD detection. Collapsed ID features produce high confidence. OOD features project to arbitrary positions. They produce higher entropy. The class prototypes might overfit. This makes the entropy threshold less reliable for similar scenes.

7 Conclusion

This work demonstrated semantic segmentation. We trained UNet on the MUAD dataset. We examined model calibration and uncertainty estimation. The key findings include several points.

1. Overfitting: A gap exists between training and validation mIoU. The model requires stronger regularization.

2. ECE reduction: Moving to a Deep Ensemble improves calibration. Ensemble diversity acts as an implicit calibrator.

3. Temperature scaling: It redistributes calibration error evenly. It did not improve ECE further.

4. MC Dropout: Predictive entropy distinguishes ID from OOD scenes. OOD images exhibit globally elevated entropy.

5. NC: The model overfits to class prototypes. This limits OOD detection reliability on borderline samples.

Future work includes stronger regularization. It includes joint optimization of temperature. It includes direct feature-space analysis of NC.