Supplementary
On Calibrating Semantic Segmentation Models: Analyses and An Algorithm

1. Uncertainty of Semantic Segmentation

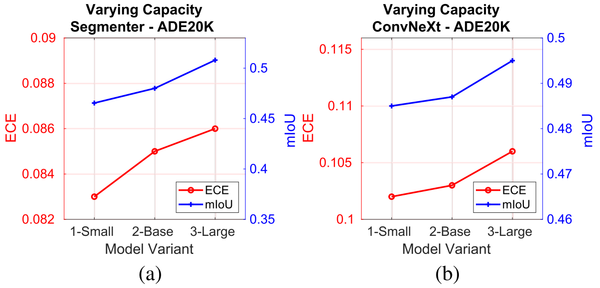

Figure 1. The effect of model capacity. Image-based ECE is employed to report model miscalibration. Model calibration error (ECE) tend to increase as model size increases given the observations from Segmenter and ConvNeXt.

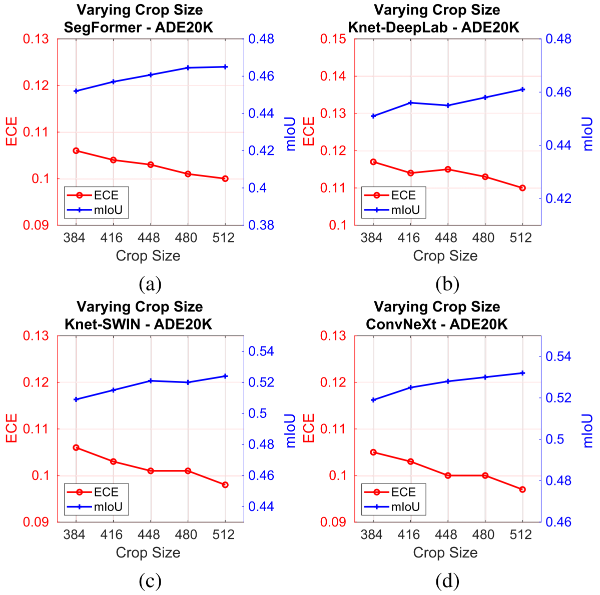

Figure 2. The effect of crop size. Image-based ECE is used to report miscalibration. Miscalibration tends to increase as crop size increases given the observations across four models.

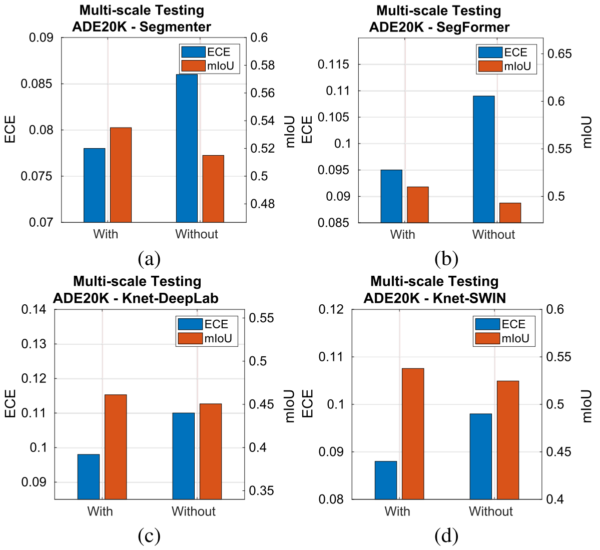

Figure 3. The effect of multi-scale testing. Image-based ECE is adopted to compare different testing strategies. From the observations, all four models show that mIoU increases, but ECE decreases when multi-testing is employed.

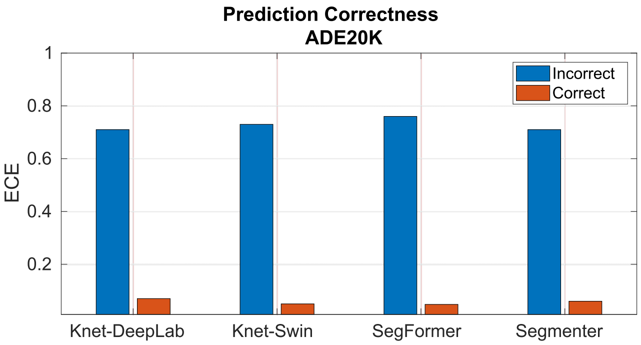

Figure 4. The effect of prediction correctness. Image-based ECE is adopted to assess miscalibration. Misprediction contributes more to miscalibration given the ECEs across four models.

2. Reliability Diagrams

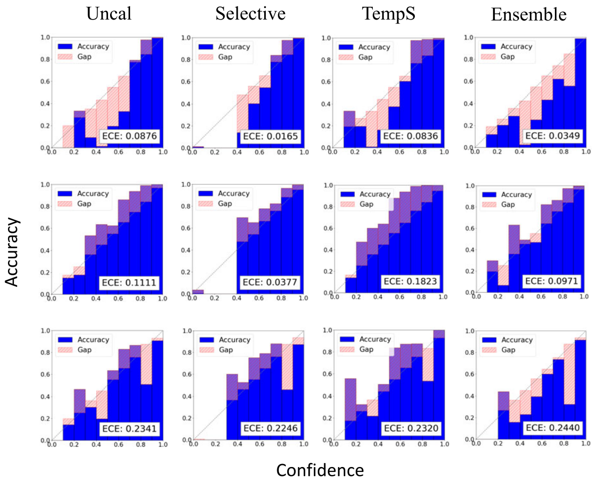

Figure 5. Reliability diagrams of visualized ECE for Segmenter-L on COCO-164K. Different calibration methods are compared across three randomly selected individual images. Selective scaling consistently outperforms temperature scaling and ensembling.

3. Spatial Statistics on Calibration Errors

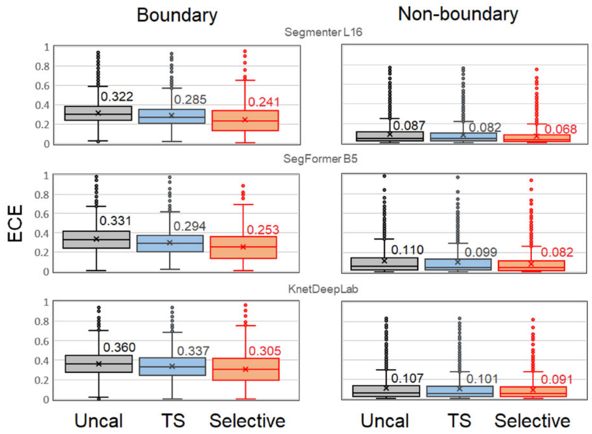

Figure 6. Calibration error (ECE) comparison in ADE20K between boundary and non-boundary pixels with boxplot of image-wise ECEs. Top and bottom bars denote maximum and minimum ECE while top, central, and bottom lines of the box indicate 25%, 50%, and 75% of ECE distribution. Crosses and dots show means and outliers. The numbers are the means. We conduct a comparative study on calibration error among uncalibration, temperature scaling, and selective scaling.

4. Selective Scaling Calibrator Training Setting

Table 1. Hyperparameter information.

BatchSize	20	Epoch	40	LR	0.001
Optimizer	AdamW	Decay	1e-6	Loss	CrossEntropy
Hyperparameter	1e10 when ACC$_{misprediction} > 50\%$				
T2	2 when ACC$_{misprediction} < 35\%$				