On Calibrating Semantic Segmentation Models: Analyses and An Algorithm

Dongdong Wang
University of Central Florida
daniel.wang@knights.ucf.edu

Boqing Gong
Google Research
bgong@google.com

Liqiang Wang
University of Central Florida
liqiang.wang@ucf.edu

Abstract

We study the problem of semantic segmentation calibration. Lots of solutions have been proposed to approach model miscalibration of confidence in image classification. However, to date, confidence calibration research on semantic segmentation is still limited. We provide a systematic study on the calibration of semantic segmentation models and propose a simple yet effective approach. First, we find that model capacity, crop size, multi-scale testing, and prediction correctness have impact on calibration. Among them, prediction correctness, especially misprediction, is more important to miscalibration due to over-confidence. Next, we propose a simple, unifying, and effective approach, namely selective scaling, by separating correct/incorrect prediction for scaling and more focusing on misprediction logit smoothing. Then, we study popular existing calibration methods and compare them with selective scaling on semantic segmentation calibration. We conduct extensive experiments with a variety of benchmarks on both in-domain and domain-shift calibration and show that selective scaling consistently outperforms other methods.

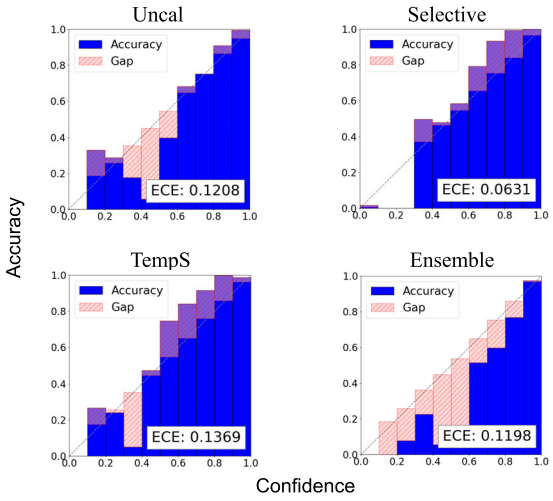


Figure 1. Reliability diagrams [5] of visualized ECE for Segmenter [40]. We randomly select and evaluate fifty validation images from COCO-164K [1] and compare uncalibrated model (Uncal), selective scaling (Selective), temperature scaling (TempS), and ensembling (Ensemble). Smaller gap implies less ECE and better calibration. Ensembling yields comparable accuracy. Despite limited improvement by TempS and Ensemble, Selective exhibits better segmentation model calibration.

1. Introduction

Deep neural networks (DNNs) have become the "go-to" models in various computer vision tasks, such as image classification, object detection, semantic segmentation. However, recent work found that DNNs are often overconfident when they make mistakes [11], misleading downstream applications. To calibrate DNNs' confidence in prediction, researchers have developed a rich line of works on image classification using regularized training [30, 31, 41, 46], post-hoc processing [11, 19, 25, 36], and Bayesian modeling [15, 18, 21, 47], to name a few.

However, the extensive pursuit of image classification has left it unclear how to calibrate DNNs for other computer vision tasks and how well existing calibration methods generalize to the tasks beyond image classification. In this paper, we conduct a comprehensive study of the calibration of deep semantic segmentation models.

Semantic segmentation tags a semantic label to every pixel in an image. Over the past years, we have witnessed increasingly accurate DNN models [2, 7, 40, 49, 51, 53] for semantic segmentation over various benchmark datasets [1, 14, 38, 44, 50, 54]. The progress has benefited many downstream applications, such as medical imaging and diagnostics, autonomous driving, and robotics. While accuracy is essential for the applications, the segmentation models' uncertainties also provide crucial signals, especially for safety-critical applications — by a segmentation model's *uncertainty*, we refer to its confidence in the label it assigns to a pixel. For example, an autonomous driving system can use the uncertainty of a semantic segmentation model (e.g., about drivable areas) to make informed decisions.

Code and models: https://github.com/dwang181/selectivecal

Existing work on the semantic segmentation uncertainty focuses on the medical image domain except [6]. Deep ensemble [8, 21, 27], MC-dropout [16], and stochastic processes [29] are introduced from image classification to calibrate a medical image segmentation model's uncertainty. These methods are applied to the training stage and some could reduce a segmentation model's accuracy. We instead focus on the post-hoc calibration methods that do not change the model's prediction, and our study covers various domains.

Ding et al.'s work [6] is the most related to ours. While they proposed and validated a novel calibration method for semantic segmentation, our objective is two-fold. We first investigate the uncertainty in semantic segmentation from multiple perspectives, hoping to reveal the major challenges with the segmentation task's uncertainty. This comprehensive investigation and insights are necessary for future research, given the limited existing work in this area. We then propose a simple and effective post-hoc algorithm that outperforms and also can apply to almost all existing calibration methods initially designed for image classification, significantly improving their performance in semantic segmentation.

To fill the gap of current research on semantic segmentation calibration, we conduct a systematic study. We analyze five state-of-the-art models [24, 40, 49, 51] with six existing calibration methods [6, 11, 19, 25, 27] over seven latest diverse benchmark datasets [1, 14, 38, 39, 44, 50, 54]. First, we investigate different factors to study how they affect miscalibration of semantic segmentation models through experiment. Given the observation, we find that model size, crop size, multi-scaling testing, and misprediction can affect calibration. In particular, segmentation model calibration error is more relevant to misprediction when model size is fixed. Next, we study different existing post-hoc calibrators and introduce our proposed selective scaling algorithm based upon the observation of misprediction. Our selective scaling optimizes ECE by reducing confidence of misprediction, which is reflected by more samples falling into lower confidence region (the far left bin) in Figure 1. Then, we carry out extensive experiments across different models, calibrators, and benchmark datasets. The experiments examine both in-domain calibration and domain-shift calibration to reveal calibration variability across different tasks.

Our contributions can be summarized as follows:

- We conduct a systematic study on the calibration of semantic segmentation models and provide insights to segmentation model calibration. We find that larger models tend to be less calibrated. When model architecture is fixed, larger crop size and multi-scale testing help calibrate models. Moreover, misprediction is important to miscalibration.

- We compare different existing popular calibration methods on state-of-the-art semantic segmentation models, and propose a simple but effective approach, selective scaling, which carries out separate scaling on correct and incorrect predictions.

- We conduct extensive experiments and justify the effectiveness of selective scaling, as shown in Figure 1. We also extend experiments from in-domain to domain-shift data and show that selective scaling consistently outperforms.

- We provide useful calibration observations as a comprehensive reference for further research on segmentation model calibration.

2. Related Work

DNN calibration. The methods of calibrating DNNs can be categorized into three groups, regularized training, uncertainty estimation, and post-hoc calibration [9]. Regularized training focuses on calibrating DNN over training, such as [30, 31, 41, 46]. Some methods estimate uncertainty using Bayesian methods [15, 18, 47] and deep ensemble [21, 26]. Post-hoc calibration conducts post-processing on the output of DNN to obtain well-calibrated results. While some post-hoc approaches exchange hurting accuracy for better calibration, like histogram binning [11, 36], some retain original accuracy of DNNs and only add calibration map to the last layer of DNNs to adjust probability distribution pattern. Several existing binary calibration maps are extended to multiclass calibration. For example, [11] extended temperature scaling and vector scaling aka Logistic scaling [11], to multiclasss calibration for DNN, and revealed temperature scaling is simple but more effective. To enrich calibration map, Dirichlet scaling [19] was proposed by extending Beta scaling [20] to multiclass DNN calibration. Recently, Meta-cal [25] integrated bipartite-ranking model with selective classification to improve calibration map for better calibrated model. Despite the success of these existing approaches in image classification calibration, there is no work to discuss their performance on semantic segmentation.

Calibration beyond image classification. Several extensions based upon existing calibrators have been proposed to solve calibration beyond image classification. For example, [29] modeled predictive uncertainty with spatial correlation matrix in stochastic process and incorporated this information into model training for more calibrated segmentation model. [6] proposed local temperature scaling that extends temperature scaling to pixels, and thus, better captures spatial heterogeneity of temperature for a better calibrated segmentation model. [27] conducted model ensembling to improve calibration and evaluate the improvement by both conventional and the proposed segment-level uncertainty metrics. [17] addressed semantic segmentation calibration through adversarial training a stochastic network which enables more accurate uncertainty estimation and robust prediction. [8] proposed a multi-headed Variational U-Net by

combining ensemble modeling with variational inference and improves model training with better in-distribution calibration and out-of-distribution detection. [32] addressed visual object detector calibration with plug-and-play training-time calibration loss and improved uncertainty optimization.

Domain-shift calibration. Domain generalization is an important challenge to DNN since model generalization to unseen data could be unpredictable when training and test data distributions differ [43]. In terms of model calibration, performance degradation can also be critical [13, 34]. To alleviate this problem, a variety of approaches have been proposed. [35] corrected covariant shift by importance weighting to achieve better calibrated models. [42] proposed a simple yet effective scheme by perturbing image for classification model calibration improvement. [45] proposed TransCal to improve image classification calibration degradation under the scenario of natural distribution shift. [10] leveraged multiple calibration domain to reduce disparity between source and target domain, thus improving model calibration on shift target domain. Despite some progress in domain-shift calibration, there is still no domain-shift study on semantic segmentation calibration.

Semantic segmentation. Semantic segmentation is an important visual understanding task. Different from image classification, semantic segmentation focuses on pixel-level classification within an image. Due to this difference, semantic segmentation models are developed in consideration of spatial correlation between pixels. DeepLab family, as an important dilated convolution model, integrates multiscale spatial context into segmentation modeling [2]. SegFormer incorporates hierarchical architectures to encode multiscale features and performs efficient semantic segmentation through ViT structure [49]. Segmenter is a straightforward ViT extension to semantic segmentation by decoding global context of the encoded embedding features from ViT outputs [40]. K-net is a unifying framework for all segmentation tasks [51]. ConvNeXt is proposed as an alternative to ViT and shows accurate segmentation [24].

3. Preliminaries

Similar to image classification, semantic segmentation can be formulated to a multiclass classification problem with a deep neural network. Let $x \in X$ and $y \in Y$ denote an input and its label, respectively. A deep neural network $h(x) = (\hat{y}, \hat{p})$ yields $\hat{y}$ as the predicted label with inference confidence $\hat{p}$. We expect a well-calibrated model to provide accurate prediction when its confidence is high.

ECE. There are several metrics to measure a model's calibration, and one of the most popular and accepted metrics is *expected calibration error (ECE)* [33], which reflects the gap between predictive confidence and accuracy. The formal definition with continuous variable is as follows.

$$ECE = \mathbb{E}_{\hat{p}}\left[|\mathbb{P}(\hat{y} = y|\hat{p} = p) - p|\right], \tag{1}$$

where $\hat{y}$ is the predicted label, y is the true label, p is expected confidence, and $\mathbb{P}(\hat{y} = y|\hat{p} = p)$ is model predictive accuracy. The expectation is about the discrepancy between accuracy and confidence. A perfectly calibrated model has zero ECE.

In practical problems, statistical binning is used to quantize continuous variables and estimate eq. (1) by equally binning the probability interval,

$$\widehat{ECE} = \sum_{i=1}^{m} \frac{|B_i|}{n} |\text{acc}(B_i) - \text{conf}(B_i)|, \tag{2}$$

where n is the number of samples, m is the number of bins, B_i denotes a set of samples falling into the bin, and $\text{acc}(B_i)$ and $\text{conf}(B_i)$ are accuracy and confidence averaged over the samples in the bin. $\widehat{ECE}$ can be visualized with the gaps in reliability diagram [5] as Figure 1.

ECE in semantic segmentation. We extend ECE to semantic segmentation by considering each pixel as a sample. Instead of pooling all pixels of different images into a set, we calculate ECE over each image first before taking an average across images,

$$ECE = \frac{1}{N} \sum_{I=1}^{N} ECE_I, \tag{3}$$

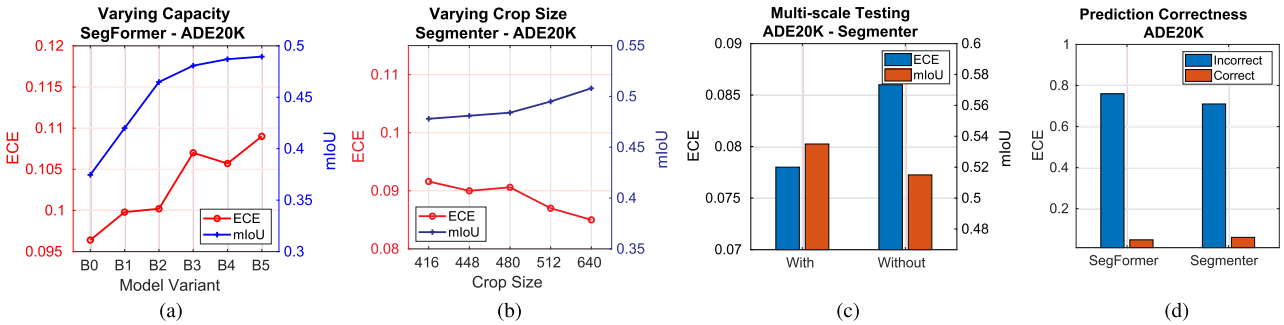
where I is an inference image, and N is the total number of images. Note that we adopt ECE_I because an image-wise metric is more popular in segmentation model evaluation. Also, ECE_I may significantly vary across images.

4. Uncertainty of Segmentation Models

We analyze the uncertainty of five segmentation models, including Segmenter [40], SegFormer [49], Knet-DeepLab [51], Knet-SWIN [51], and ConvNeXt [24], with the ADE20K [54] dataset, a popular semantic segmentation benchmark with 20,210/2,000 training/validation images. The models are state-of-the-art and released by OpenMMLab [3]. Due to limit of space, we focus on Segmenter [40] and SegFormer [49] in this section and report other models' results in the supplementary materials.

4.1. Larger models tending to be less calibrated

Among the models we analyze, we find that the model capacity is a primary factor determining their generalization accuracy. Large models often lead to high inference accuracy. Will it be valid for semantic segmentation models? To answer it, we conduct experiments on different model capacity by varying model depth and width. We select five variants of SegFormer [49], from B0 to B5 with a deeper and wider architecture, for model scaling-up study. From Figure 2a, we observe that the deeper and wider models,



(a) (b) (c) (d)

Figure 2. The effect of model capacity, image crop size, multi-scale testing, and prediction correctness on miscalibration. Prediction correction separately computes the ECE of incorrect/correct predictions. Image-based ECE in eq.(3) is used.

like B5, exhibit higher ECE (*i.e.*, worse miscalibration) despite better mIoU. It is consistent with the observation for image classification from Figure 2 in [11], which is caused by the over-confidence during the later stage of NLL optimization [11]. This implies that more accurate prediction leads to worse calibration when it comes to varying model size.

4.2. Larger crop size tending to be more calibrated

Image cropping is usually used to prepare input data for semantic segmentation models. Crop size determines model input resolution, *i.e.*, the number of pixels for modeling. Larger crop size indicates larger input image, more pixels fed to model, and wider modeling region. Although larger crop size shows higher predictive accuracy, its effect on model calibration is unknown. We conduct experiments to explore their interaction. We select Segmenter-L [40] as a representative model with five crop sizes of 416, 448, 480, 512, and 640 to show the crop size scaling effect. From Figure 2b, we observe that larger crop size yields lower ECE and better calibration. We associate it with modeling field. When crop size increases, larger modeling field provides richer global context information which helps more accurately and confidently locate object boundary and position. This implies that more accurate prediction leads to better calibration in terms of varying crop size.

4.3. Multi-scaling testing improving calibration

Multi-scale testing is an important technique to boost semantic segmentation model. It usually resizes the input images into multiple scales, feed these resized images into the model, and average all inferences for final prediction. Predictive accuracy is improved after multi-scale testing because multi-scale features help address scale variation between different objects [52]. However, its effectiveness on semantic segmentation calibration is unknown. We conduct experiments with Segmenter-L and cropping size of 640×640 of ADE20K to investigate the effect of multi-scale testing. The rescale ratios include 0.5, 0.75, 1.0, 1.25, 1.5, and 1.75. Given Figure 2c, with multi-scale testing, ECE is reduced from 0.86 to 0.78 while mIoU increases from 0.515 to 0.535. This suggests that multi-scale testing enables both alleviating miscalibration and improving accuracy. The intuition is that the aggregation of multi-scale features improves the object boundary delineation and predictive confidence in objects. Also, this ensembling helps correct mispredictions and reduce miscalibration over the over-confidence. This implies that more accurate prediction leads to better calibration concerning multi-scaling testing.

4.4. Mispredictions affecting calibration

Given eqs.(1) and (2), miscalibration results from discrepancy between predictive confidence and accuracy [33]. A perfectly calibrated model should yield predicted probability of 1, *i.e.*, 100% confidence, for correct prediction, and predicted probability of 0, *i.e.*, infinitesimal confidence, for misprediction. However, an optimal solution is hardly obtained in practice, which implies underconfidence in correct prediction and over-confidence in misprediction. How do these discrepancies affect the calibration property of a segmentation model? We conduct an experiment with SegFormer-B5 [49] and Segmenter-L [40] by grouping correct/incorrect predictions to separately compute their ECEs. This separation can more clearly reveal how much they contribute to final miscalibration. Figure 2d shows that the ECEs from incorrect predictions are significantly higher than those from correct predictions across both models. This implies that misprediction more contributes to miscalibration. This observation also justifies that over-confidence is a more critical problem for calibration [11].

5. Calibrating Semantic Segmentation Model

We first introduce our post-hoc scaling algorithm in this section. Next, we briefly review the popular existing calibration methods, including four image classification post-hoc calibrators, one semantic segmentation calibrator, and one Bayesian modeling calibrator. We select these calibrators because they exhibit state-of-the-art calibration performance in different groups.

Figure 3. Segmented examples from ADE20K, COCO-164K, BDD100K, DAVIS, SpaceNet-7, BraTs-2017, and SYNTHIA (left to right).

5.1. Selective Scaling

We propose selective scaling given that misprediction is more attributed to miscalibration (Figure 2d). Selective scaling is an extension of selective classification to model calibration. It introduces a binary classifier as a selector to categorize correct and incorrect predictions for separate scaling. For example, selected mispredictions' logits will be smoothed to alleviate the miscalibration from over-confidence.

$$\begin{cases} \hat{\boldsymbol{p}} = \sigma_{SM}(\boldsymbol{z}/T_1), & \text{if } \hat{y} \neq y \\ \hat{\boldsymbol{p}} = \sigma_{SM}(\boldsymbol{z}/T_2), & \text{otherwise} \end{cases} \tag{4}$$

where $\hat{\boldsymbol{p}}$ is calibrated probability vector, σ_{SM} is softmax activation function, $\boldsymbol{z}$ is the logit vector before activation, T_1 is a temperature to smooth logit distribution, and T_2 is a temperature to sharpen logit distribution, so $T_1 > T_2$. For over-confidence problem, we set T_2 to 1. It can be extended to under-confidence problem with T_2 less than 1. The selection of T_1 relies on the selector accuracy. When its accuracy is higher, T_1 is larger. Our selective scaling optimizes ECE by reducing confidence of misprediction, which is reflected by more samples falling into the lower confidence region (the far left bin) in reliability diagram (Figure 1).

Selective scheme was adopted for calibrator design in Meta-cal [25], but different from selective scaling. First, selective scaling does not degrade model accuracy while Meta-cal with miscoverage control does [25]. Different from Meta-cal, selective scaling is independent from entropy ranking model to determine classification threshold. Also, selective scaling obtains a selector without heuristic data resampling used in Meta-cal, which reduces hyperparameters. Selective scaling, to the best of our knowledge, is never used for semantic segmentation model calibration.

5.2. Existing calibration methods

Temperature Scaling is a simple but effective approach for multi-classification model calibration [11]. The calibration is carried out with a single temperature parameter to scale logits for overfitting problem resolution.

$$\hat{\boldsymbol{p}} = \sigma_{SM}(\boldsymbol{z}/T), \tag{5}$$

where T is a scaler of temperature to scale logit vector $\boldsymbol{z}$.

Logistic Scaling is an extension of temperature scaling a.k.a. vector scaling [11]. The scaling model is formulated with linear transformation for more complex calibration map.

$$\hat{\boldsymbol{p}} = \sigma_{SM}(w \odot \boldsymbol{z} + b), \tag{6}$$

where w and b are two vectors to scale the logit vector $\boldsymbol{z}$.

Dirichlet Scaling is the extension of logistic scaling and derived with probability output distribution [19], which enriches calibration map for better optimization with Dirichlet distribution. We adopt linear paramenterization [19] for Dirichlet scaling, which is formulated as follows.

$$\hat{\boldsymbol{p}} = \sigma_{SM}(W \cdot log(\sigma_{SM}(\boldsymbol{z})) + b), \tag{7}$$

where W is a matrix and b is a vector for linear parametrisation of the probability $\sigma_{SM}(\boldsymbol{z})$.

Meta-cal is the scaling strategy derived from bipartite-ranking model and selective classification [25]. With the entropy threshold by a ranking model, probability outputs are separately processed. The probability output with the entropy smaller than threshold is scaled by temperature scaling; otherwise, the output is changed to random prediction. Obviously, it severely degrades model accuracy. Although [25] proposed coverage accuracy control for image classification, it is impractical to semantic segmentation because mIoU is hard to be tuned with calibrator training set.

$$\begin{cases} \hat{\boldsymbol{p}} = \mathbf{1}/k, & \text{if } -\hat{p}log(\hat{p}) > \gamma, \\ \hat{\boldsymbol{p}} = \sigma_{SM}(\boldsymbol{z}/T), & \text{otherwise}, \end{cases} \tag{8}$$

where k is the number of classes, γ is the entropy threshold computed with a ranking model for calibration decision.

Local Temperature Scaling (LTS) is the extension of temperature scaling to semantic segmentation [6]. It provides pixel-level heterogeneous scaling temperature by combing input images with output logit through CNN. Although it considers spatial heterogeneity over an image for scaling, the calibrator training relies on strong correlation between image features and temperature heterogeneity and more training data for feature extraction [6].

$$\hat{\boldsymbol{p}} = \sigma_{SM}(\boldsymbol{z}/(\boldsymbol{T}(\boldsymbol{z}, I)), \tag{9}$$

where $\boldsymbol{T}(\boldsymbol{z}, I)$ is a pixel-wise temperature map by a convolution with segmented logit vector $\boldsymbol{z}$ and image I.

Ensembling is proposed to solve medical image segmentation uncertainty prediction [27]. It carries out calibration improvement with simple average of ensemble model, a simplified version of Bayesian inference [47]. Also, deep ensemble is a strong baseline for image classification calibration [21, 47]. We compare its performance with other post-hoc approaches.

$$\hat{\boldsymbol{p}} = \frac{1}{N}\Sigma_{n=1}^{N}\hat{\boldsymbol{p}}^n, \tag{10}$$

where N is the number of ensemble members.

6. Experiments

Models. We consider five recent state-of-the-art segmentation models which cover CNN and ViT architectures.

1. **SegFormer** [49] is a ViT-based encoder-decoder model based upon lightweight multilayer perception (MLP) decoders with pyramid architecture.
2. **Segmenter** [40] is an encoder-decoder model based exclusively on Transformer.
3. **Knet** [51] is a unified segmentation decoder module which includes two important variants, Knet-DeepLab and Knet-SWIN. Knet-DeepLab consists of ResNet-50 [12] backbone and DeepLab-V3 [2] decoder. Knet-SWIN is contructed with SWIN [23] backbone and UperNet [48] decoder.
4. **ConvNeXt** [24] is a CNN backbone, but shaped to ViT-like architecture for better scaling. We select ConvNeXt backbone with UperNet [48] decoder.

For ADE20K, all models except Knet-DeepLab are trained with 640×640 crop size. For BDD100K and SYNTHIA, the crop size is 512×1024. For other benchmarks, the crop size is 512×512. The batch size is set to 8.

Dataset. We examine existing calibration methods across six important benchmarks from various applications. Figure 3 illustrates the example images from benchmarks.

1. **Scene and stuff segmentation.** We use ADE20K [54] and COCO-164K [1] as large-scale segmentation benchmark. ADE20K contains 150 object and stuff classes with 20,210/2,000 images in the training/validation set. COCO-164K, ∼164K images for 91 stuff and 80 thing classes, includes 118K/5K images in the training/validation set.
2. **Autonomous driving.** We choose BDD-100K [50], a latest benchmark for urban driving scene segmentation. BDD-100K contains 7K/1K 1280×720 images in 19 classes for training/validation set. We select the validation set of CityScapes [4] as test set from target domain for domain-shift calibration assessment. CityScapes has 2975/500 1024×2048 images for training/validation set. BDD-100K and CityScapes share the same label space.
3. **Video segmentation.** We adopt DAVIS2016 [37] to assess calibration on temporal domain shift problems. DAVIS2016 contains 50 object and 1 background classes. We select 480p resolution video sequences to prepare temporal domain shift dataset. The total of the frames in each sequence ranges from 24 to 103. Each frame is an image segmented with a single object and background.
4. **Remote sensing.** We select SPACENET-7 (SN-7) [44], a binary small object identification benchmark, to examine spatial and temporal domain shift calibration. SN-7 consists of 1,408 1024×1024 satellite images collected from monthly urban development across 101 cities. It contains 11,080,000 buildings over 41,000 km^2 observation area. Due to the development, images vary temporally; due to planning difference, the objects like the building styles differ by locations.
5. **Medical imaging.** We select BraTS2017 [14], a popular brain tumor segmentation dataset. BraTS2017 contains multimodal magnetic resonance imaging (MRI) scans from 1,210 patients, which consists of 3D voxels with 155 layers from top to bottom. Each layer is equivalent to a 2D semantic segmentation image with 3 tumor feature classes.
6. **Simulation domain transfer.** SYNTHIA [39] is a popular unban street synthetic image benchmark. We select its subset of RANDCITYSCAPES, which contains 9400 1280×760 images with 22 classes, as source domain for model training. We select CityScapes [4] as target domain for domain-shift calibration assessment.

In-domain calibration. We select the validation set of each benchmark for in-domain calibration study. For ADE20K, COCO-164K, and BDD100K, we select 100/100/1800 images, 250/250/4500 images, and 50/50/900 images for calibrator training/validation/testing.

Domain-shift calibration. We select the validation sets of COCO-164K and CityScapes as testing sets of ADE20K and BDD100K. For taxonomy incompatibility over datasets, like ADE20K-COCO164K, we follow the convention in [22] to merge and split label classes. For DAVIS, we merge all training and validation sequences and resplit them in half by frame. The first half of frames are used for training with 1731 images and the second half are split for validation/testing with 100/1624 images. For SN-7, all training data are split into 1133/50/225 images for training/validation/testing set by location, and 698/50/660 images for training/validation/testing set by time. For BraTS, we select 800/50 images of 75^{th} layer scan for training/validation and 360 images of 100^{th} layer scan for testing. For SYNTHIA, we split training set into 9000/400 images for training/validation, and CityScapes' validation set for testing.

Calibration methods. We select aforementioned six existing calibration methods to compare selective scaling. For selective scaling, selective model is simplified with three-layer MLP due to training data restriction. The inputs are pixel-wise predictive probabilities and the labels are binary which indicate correct/incorrect prediction. Moreover, Meta-cal with miscoverage control yields severe model degradation. We implement our extension by replacing random prediction with large temperature scaling to retain accuracy. For ensembling, we achieve a three-member ensemble with comparable accuracy for fair comparison.

Table 1. Segmentation accuracy (mIoU) and calibration error (ECE) on different benchmarks. TempS, LogS, DirS, LTS, Ens., and Selective denote temperature, logistic, Dirichlet, local temperature, ensembling, and selective scaling, respectively. Ensembling is carried out by three models with reduced size for comparable mIoU. Meta-Cal* is implemented by our extension with large temperature.

Dataset	Model	mIoU	Uncal	TempS	LogS	DirS	Meta-Cal*	LTS	Ens.	Selective
ADE20K	SegFormer-B5 [49]	49.13	0.111	0.109	0.110	0.110	0.103	0.105	0.109	**0.086**
ADE20K	Segmenter-L [40]	51.65	0.087	0.086	0.086	0.087	0.081	0.094	0.086	**0.069**
ADE20K	Knet-DeepLab [51]	45.06	0.111	0.105	0.107	0.106	0.102	0.118	0.110	**0.095**
ADE20K	Knet-SWIN-L [51]	52.46	0.098	0.094	0.093	0.097	0.089	0.134	0.096	**0.078**
ADE20K	ConvNeXt-L [24]	53.16	0.097	0.092	0.094	0.091	0.088	0.133	0.094	**0.082**
COCO-164K	SegFormer-B5 [49]	45.78	0.151	0.149	0.141	0.151	0.132	0.151	0.149	**0.113**
COCO-164K	Segmenter-L [40]	47.09	0.152	0.149	0.149	0.151	0.130	0.155	0.150	**0.109**
COCO-164K	Knet-DeepLab [51]	37.24	0.170	0.170	0.168	0.171	0.149	0.172	0.169	**0.093**
COCO-164K	Knet-SWIN-L [51]	46.49	0.161	0.159	0.161	0.160	0.142	0.162	0.160	**0.102**
COCO-164K	ConvNeXt-L [24]	46.48	0.160	0.157	0.158	0.159	0.141	0.162	0.159	**0.108**
BDD100K	SegFormer-B5 [49]	65.08	0.064	0.055	0.054	0.053	0.049	0.069	0.059	**0.040**
BDD100K	Segmenter-L [40]	61.33	0.055	0.045	0.043	0.042	0.037	0.071	0.052	**0.031**
BDD100K	Knet-DeepLab [51]	62.89	0.060	0.049	0.047	0.048	0.041	0.063	0.057	**0.035**
BDD100K	Knet-SWIN-L [51]	67.59	0.065	0.055	0.054	0.054	0.049	0.067	0.063	**0.040**
BDD100K	ConvNeXt-L [24]	67.26	0.064	0.054	0.053	0.056	0.049	0.065	0.063	**0.038**

6.1. Results

6.1.1 In-domain calibration evaluation

We present ECEs for five models and compare selective scaling with six existing calibrators on three-run average. Table 1 shows that majority of calibrators yield very limited improvement in model calibration. However, surprisingly, selective scaling consistently outperforms these scaling approaches. This calibration gain is attributed to more scaling on mispredictions. Moreover, this improvement is more significant for scene and object benchmarks such as ADE20K and COCO-164K than street view datasets like BDD-100K.

Furthermore, we find that BDD100K yields more calibrated models than other benchmarks. We associate it with larger crop size (512×1024) and higher model accuracy. We also find that ensembling exhibits weaker calibration. We connect it with model accuracy since we restrain the ensemble accuracy for fair comparison, which limits its calibration performance. Moreover, across models, Segmenter exhibits better calibration. Since it is a Transformer-exclusive model, we link it to different spatial inductive bias which is speculated in [28].

6.1.2 Domain-shift calibration evaluation

We present ECEs for five models on seven sets of domain-shift experiments and compare selective scaling with six existing calibrators on three-run average. Table 2 shows that selective scaling still consistently outperforms by a large margin. It implies that selective scaling can be generalized to domain-shift problems. In particular, Meta-cal exhibits performance variation. We speculate that the hard threshold from entropy calculation limits separating shift data. In contrast, selective scaling is more consistent because binary classifier has better generalization to separating shift data.

Across benchmarks, we find that label space shift causes more significant miscalibration. For example, ADE20K to COCO-164K and SYNTHIA to CityScapes show worse ECE than other domain-shift experiments. We link it to poorer domain adaptation and model generalization degradation. Also, we observe that small object segmentation like SN-7, as a challenged task, can cause worse miscalibration. Except for SN-7, more accurate models are more calibrated by benchmarks. For SN-7, background pixels engage in training but are excluded from calibration calculation, which causes larger ECE. Within these models, we find that Segmenter is still more calibrated.

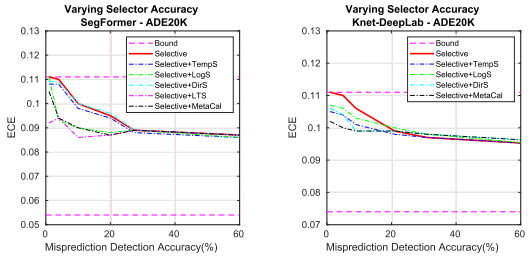


Figure 4. The effect of selector accuracy on calibration. Misprediction detection accuracy denotes the percentage of misprediction detected by selector. The lower bound (optimal case) is 100% misprediction detected for scaling. The upper bound (worst case) is 0%, *i.e.*, none misprediction detected is used for calibration.

6.1.3 Ablation study

To analyze the effect of selector on selective scaling, we conduct ablation study by varying misprediction detection accuracy, integrating with different existing calibrators, and examining calibration error across regions.

Selective scaling considering selection accuracy. Figure 4 shows that with the increase in selection accuracy, ECE decreases. This implies that when more mispredictions are detected and scaled, model calibration becomes better. It justifies the effectiveness of selective scaling.

Selective scaling with existing calibrators. We also

Table 2. Segmentation model accuracy (mIoU) and calibration error (ECE) on different benchmarks for domain-shift calibration assessment. InD and SD denotes in-domain and domain-shift test data. Meta-Cal* is our extension with large temperature scaling.

InD	SD	Model	mIoU	Uncal	TempS	LogS	DirS	Meta-Cal*	LTS	Ens.	Selective
ADE20K	COCO-164K	SegFormer-B5 [49]	8.20	0.467	0.466	0.468	0.469	0.415	0.420	0.465	**0.331**
ADE20K	COCO-164K	Segmenter-L [40]	9.60	0.395	0.394	0.395	0.394	0.322	0.331	0.393	**0.259**
ADE20K	COCO-164K	Knet-DeepLab [51]	7.02	0.447	0.445	0.446	0.446	0.403	0.439	0.444	**0.351**
ADE20K	COCO-164K	Knet-SWIN-L [51]	9.06	0.464	0.462	0.463	0.463	0.379	0.465	0.462	**0.327**
ADE20K	COCO-164K	ConvNeXt-L [24]	8.99	0.461	0.458	0.459	0.460	0.377	0.460	0.459	**0.320**
BDD100K	CityScapes	SegFormer-B5 [49]	63.05	0.087	0.083	0.082	0.084	0.076	0.085	0.087	**0.069**
BDD100K	CityScapes	Segmenter-L [40]	60.26	0.062	0.058	0.059	0.057	0.055	0.067	0.060	**0.048**
BDD100K	CityScapes	Knet-DeepLab [51]	61.27	0.073	0.070	0.069	0.071	0.063	0.073	0.071	**0.056**
BDD100K	CityScapes	Knet-SWIN-L [51]	67.46	0.080	0.078	0.077	0.077	0.071	0.079	0.079	**0.063**
BDD100K	CityScapes	ConvNeXt-L [24]	67.18	0.081	0.079	0.079	0.077	0.070	0.080	0.079	**0.065**
DAVIS-train	DAVIS-test	SegFormer-B4 [49]	89.33	0.033	0.031	0.031	0.031	0.055	0.098	0.032	**0.024**
DAVIS-train	DAVIS-test	Segmenter-B [40]	81.35	0.080	0.077	0.076	0.075	0.103	0.121	0.078	**0.049**
DAVIS-train	DAVIS-test	Knet-DeepLab [51]	83.12	0.076	0.072	0.073	0.073	0.098	0.107	0.075	**0.047**
DAVIS-train	DAVIS-test	Knet-SWIN-B [51]	89.22	0.045	0.041	0.041	0.040	0.065	0.101	0.044	**0.029**
DAVIS-train	DAVIS-test	ConvNeXt-B [24]	89.00	0.044	0.040	0.041	0.042	0.067	0.100	0.044	**0.031**
SN-7-SP-train	SN-7-SP-test	SegFormer-B4 [49]	57.23	0.655	0.636	0.625	0.634	0.711	0.701	0.652	**0.532**
SN-7-SP-train	SN-7-SP-test	Segmenter-B [40]	51.98	0.766	0.751	0.748	0.740	0.796	0.789	0.763	**0.687**
SN-7-SP-train	SN-7-SP-test	Knet-DeepLab [51]	55.58	0.704	0.688	0.679	0.681	0.768	0.720	0.696	**0.655**
SN-7-SP-train	SN-7-SP-test	Knet-SWIN-B [51]	59.42	0.609	0.576	0.559	0.563	0.631	0.618	0.605	**0.504**
SN-7-SP-train	SN-7-SP-test	ConvNeXt-B [24]	58.69	0.611	0.570	0.564	0.566	0.629	0.633	0.608	**0.517**
SN-7-TS-train	SN-7-TS-test	SegFormer-B4 [49]	54.39	0.700	0.673	0.669	0.673	0.712	0.723	0.695	**0.582**
SN-7-TS-train	SN-7-TS-test	Segmenter-B [40]	50.13	0.769	0.745	0.740	0.736	0.785	0.777	0.755	**0.624**
SN-7-TS-train	SN-7-TS-test	Knet-DeepLab [51]	56.73	0.651	0.622	0.621	0.619	0.689	0.675	0.647	**0.576**
SN-7-TS-train	SN-7-TS-test	Knet-SWIN-B [51]	62.42	0.619	0.587	0.579	0.584	0.630	0.621	0.609	**0.551**
SN-7-TS-train	SN-7-TS-test	ConvNeXt-B [24]	62.14	0.623	0.578	0.571	0.573	0.647	0.638	0.620	**0.549**
BraTS-train	BraTS-test	SegFormer-B4 [49]	45.81	0.155	0.137	0.135	0.135	0.129	0.150	0.151	**0.120**
BraTS-train	BraTS-test	Segmenter-B [40]	44.57	0.154	0.149	0.146	0.147	0.135	0.142	0.150	**0.124**
BraTS-train	BraTS-test	Knet-DeepLab [51]	46.05	0.188	0.179	0.176	0.176	0.158	0.187	0.185	**0.148**
BraTS-train	BraTS-test	Knet-SWIN-B [51]	47.99	0.155	0.146	0.143	0.144	0.132	0.157	0.152	**0.123**
BraTS-train	BraTS-test	ConvNeXt-B [24]	48.40	0.169	0.159	0.154	0.155	0.136	0.168	0.165	**0.121**
SYNTHIA	CityScapes	SegFormer-B5 [49]	33.12	0.286	0.264	0.262	0.263	0.230	0.288	0.284	**0.218**
SYNTHIA	CityScapes	Segmenter-L [40]	32.33	0.254	0.235	0.232	0.230	0.217	0.280	0.252	**0.203**
SYNTHIA	CityScapes	Knet-DeepLab [51]	30.19	0.311	0.298	0.296	0.296	0.272	0.305	0.309	**0.241**
SYNTHIA	CityScapes	Knet-SWIN-L [51]	34.31	0.281	0.273	0.270	0.270	0.239	0.288	0.279	**0.226**
SYNTHIA	CityScapes	ConvNeXt-L [24]	34.79	0.287	0.267	0.263	0.264	0.230	0.299	0.284	**0.211**

find that selective scaling can boost existing calibrators, vice versa. Figure 4 shows that the integration with selective scaling helps existing calibrators to improve model calibration. When detection accuracy is lower, like $< 20\%$, existing calibrators can help improve selective scaling.

Selective scaling across regions. Figure 5 shows segmentation calibration error statistics across regions. 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. It shows that: 1) boundary pixels are less calibrated than non-boundary pixels within an object bounding box; 2) selective scaling consistently yields lower mean ECE across both types of pixels.

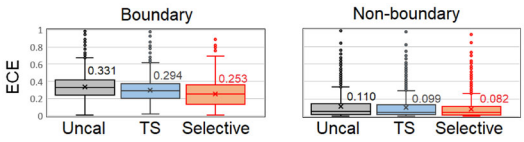


Figure 5. SegFormer-B5 ECE comparison in ADE20K between boundary and non-boundary pixels among uncalibration, temperature scaling, and selective scaling. Crosses and dots show means and outliers. The numbers are the means.

6.2. Limitation

Logit smoothing can cause entropy increase because post-hoc over-confidence calibration, such as temperature scaling, is equivalent to entropy maximization under certain logit constraints [11]. For example, original Meta-cal leads to significant entropy increase due to randomizing certain prediction outputs. Our method alleviates this issue, but still yields higher entropy. To resolve it, careful temperature tuning is a trade-off scheme for entropy improvement.

7. Conclusion

We analyzed semantic segmentation model calibration and proposed a simple but effective algorithm of selective scaling. We revealed the possible factors of segmentation model miscalibration. Given the analysis, we designed selective scaling to scale logits separately, and more focus on mispredicted logit smoothing. We performed extensive experiments on state-of-the-art models across various benchmarks. Both in-domain and domain-shift experiments justify the effectiveness of selective scaling. We also presented our observations for segmentation calibration research.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 1, 2, 6

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1, 3, 6

[3] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 3

[4] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[5] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. 1, 3

[6] Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local temperature scaling for probability calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6889–6899, 2021. 2, 5

[7] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9716–9725, 2021. 1

[8] Moritz Fuchs, Camila Gonzalez, and Anirban Mukhopadhyay. Practical uncertainty quantification for brain tumor segmentation. In *Medical Imaging with Deep Learning*, 2021. 2

[9] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021. 2

[10] Yunye Gong, Xiao Lin, Yi Yao, Thomas G Dietterich, Ajay Divakaran, and Melinda Gervasio. Confidence calibration for domain generalization under covariate shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8958–8967, 2021. 3

[11] 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. 1, 2, 4, 5, 8

[12] 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. 6

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 3

[14] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In *International MICCAI Brainlesion Workshop*, pages 287–297. Springer, 2017. 1, 2, 6

[15] Pavel Izmailov, Wesley J Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace inference for bayesian deep learning. In *Uncertainty in Artificial Intelligence*, pages 1169–1179. PMLR, 2020. 1, 2

[16] Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer, 2019. 2

[17] Elias Kassapis, Georgi Dikov, Deepak K Gupta, and Cedric Nugteren. Calibrated adversarial refinement for stochastic semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7057–7067, 2021. 2

[18] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 1, 2

[19] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019. 1, 2, 5

[20] Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, pages 623–631. PMLR, 2017. 2

[21] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 1, 2, 5

[22] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2879–2888, 2020. 6

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6

[24] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 6, 7, 8

[25] Xingchen Ma and Matthew B Blaschko. Meta-cal: Well-controlled post-hoc calibration by ranking. In *International Conference on Machine Learning*, pages 7235–7245. PMLR, 2021. 1, 2, 5

[26] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019. 2

[27] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020. 2, 5

[28] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021. 7

[29] Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in Neural Information Processing Systems*, 33:12756–12767, 2020. 2

[30] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020. 1, 2

[31] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019. 1, 2

[32] Muhammad Akhtar Munir, Muhammad Haris Khan, M Saquib Sarfraz, and Mohsen Ali. Towards improving calibration in object detection under domain shift. *arXiv preprint arXiv:2209.07601*, 2022. 3

[33] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 3, 4

[34] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019. 3

[35] Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 3219–3229. PMLR, 2020. 3

[36] Kanil Patel, William Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. Multi-class uncertainty calibration via mutual information maximization-based binning. *arXiv preprint arXiv:2006.13092*, 2020. 1, 2

[37] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 6

[38] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 1, 2

[39] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 2, 6

[40] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 1, 2, 3, 4, 6, 7, 8

[41] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[42] Christian Tomani, Sebastian Gruber, Muhammed Ebrar Erdem, Daniel Cremers, and Florian Buettner. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10132, 2021. 3

[43] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 3

[44] Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development spacenet dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2021. 1, 2, 6

[45] Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable calibration with lower bias and variance in domain adaptation. *Advances in Neural Information Processing Systems*, 33:19212–19223, 2020. 3

[46] Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. Energy-based open-world uncertainty modeling for confidence calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9302–9311, 2021. 1, 2

[47] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020. 1, 2, 5

[48] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 6

[49] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. 1, 2, 3, 4, 6, 7, 8

[50] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1, 2, 6

[51] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-Net: Towards unified image segmentation. In *NeurIPS*, 2021. 1, 2, 3, 6, 7, 8

[52] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 4

[53] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840*, 2020. 1

[54] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1, 2, 3, 6