Reliability in Semantic Segmentation: Are We on the Right Track?

Pau de Jorge
University of Oxford
NAVER LABS Europe*

Riccardo Volpi
NAVER LABS Europe

Philip H. S. Torr
University of Oxford

Grégory Rogez
NAVER LABS Europe

Abstract

Motivated by the increasing popularity of transformers in computer vision, in recent times there has been a rapid development of novel architectures. While in-domain performance follows a constant, upward trend, properties like robustness or uncertainty estimation are less explored—leaving doubts about advances in model reliability. Studies along these axes exist, but they are mainly limited to classification models. In contrast, we carry out a study on semantic segmentation, a relevant task for many real-world applications where model reliability is paramount. We analyze a broad variety of models, spanning from older ResNet-based architectures to novel transformers and assess their reliability based on four metrics: robustness, calibration, misclassification detection and out-of-distribution (OOD) detection. We find that while recent models are significantly more robust, they are not overall more reliable in terms of uncertainty estimation. We further explore methods that can come to the rescue and show that improving calibration can also help with other uncertainty metrics such as misclassification or OOD detection. This is the first study on modern segmentation models focused on both robustness and uncertainty estimation and we hope it will help practitioners and researchers interested in this fundamental vision task[1].

1. Introduction

Humans tend to overestimate their abilities, a cognitive bias known as Dunning-Kruger effect [27]. Unfortunately, so do deep neural networks. Despite impressive performance on a wide range of tasks, deep learning models tend to be overconfident—that is, they predict with high-confidence even when they are wrong [19]. This effect is even more severe under domain shifts, where models tend to underperform in general [23, 40, 45].

While these vulnerabilities affect deep models in general, they are often studied for classification models and are comparably less explored for semantic segmentation, a fundamental task in computer vision that is key to many critical applications such as autonomous driving and AI-assisted medical imaging. In those applications, domain shifts are more the rule than the exception (*e.g.*, changes in weather for a self-driving car or differences across patients for a medical imaging system). Therefore, brittle performance and overconfidence under domain shifts are two important and challenging problems to address for a safe deployment of artificial intelligence systems in the real world.

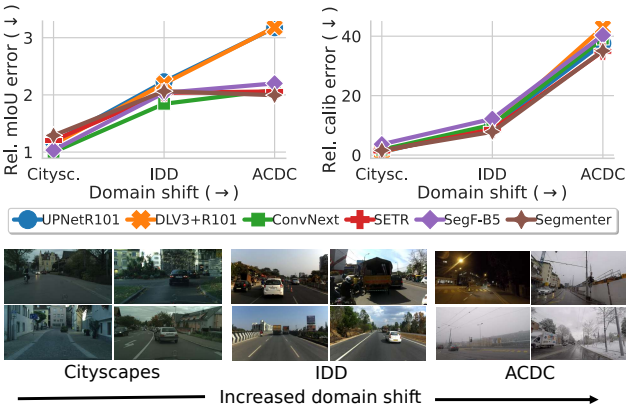

Figure 1. **Top: mIoU and ECE *vs*. domain shift.** Errors are normalized with respect to the lowest error on the training distribution (Cityscapes). We compare recent segmentation models, both transformer-based (SETR [58], SegFormer [55] and Segmenter [48]) and convolution-based (ConvNext [30]) with ResNet baselines (UPerNet [54] and DLV3+ [6]). All recent models (both transformers and CNNs) are remarkably more robust than ResNet baselines (whose lines in mIoU overlap), however, ECE increases sharply for all methods. **Bottom:** Sample images for each dataset.

With that in mind, we argue that a *reliable* model should *i)* be robust to domain shifts and *ii)* provide good uncertainty estimates. The core goal of this study is providing an answer to the following, crucial question: **are state-of-the-art semantic segmentation models improving in terms of robustness *and* uncertainty estimation?**

To shed light on this, we evaluate a large body of segmentation models, assessing their in-domain (ID) *vs*. out-

*https://europe.naverlabs.com
[1]Code available at https://github.com/naver/relis

of-domain (OOD) prediction quality (**robustness**) together with their calibration, misclassification detection and OOD detection (**uncertainty estimation**).

We argue that a study of this kind is crucial to understand whether research on semantic segmentation is moving in the right direction. Following the rise of transformer architectures in computer vision [4, 15, 29, 50], several studies have compared recent self-attention and CNN-based *classification* models in terms of robustness [2, 3, 30, 32, 36, 43] and predictive uncertainty [33, 44]. Yet, when it comes to *semantic segmentation*, prior studies [55, 59] only focused on robustness, using synthetic corruptions as domain shifts (*e.g.*, blur, noise) [25]. In contrast, we consider natural, realistic domain shifts and study segmentation models both in terms of robustness and uncertainty, leveraging datasets captured in different conditions—see Fig. 1 (bottom).

Task-specific studies are important, since task-specific architectures and learning algorithms may carry different behaviors and some observations made for classification might not hold true when switching to segmentation. For instance, contrary to Minderer *et al*. [33], we observe that improvements in calibration are far behind those in robustness, see Fig. 1 (top). Furthermore, previous analyses only consider simple calibration approaches [19] while assessing model reliability; in contrast, we make a step forward and explore content-dependent calibration strategies [14, 17], which show promise to improve reliability out of domain.

Our analysis allows us individuating in which directions we are improving and in which we are lagging behind. This is the first work to systematically study robustness and uncertainty under domain shift for a large suite of segmentation models and we believe it can help practitioners and researchers working on semantic segmentation. We summarize our main observations in the following.

i) Remarkable improvements in robustness, but poor in calibration. Under domain shifts, recent segmentation models perform significantly better (in terms of mIoU)—with larger improvements for stronger shifts. Yet, OOD calibration error increases dramatically for all models.

ii) Content-dependent calibration [14] can improve OOD calibration, especially under strong domain shifts, where models are poorly calibrated.

iii) Misclassification detection shows different model ranking in and out of domain. When tested in domain, recent models underperform the ResNet baseline. As the domain shift increases, recent models take the lead.

iv) OOD detection is inversely correlated with performance. Indeed, a small ResNet-18 backbone performs best.

v) Content-dependent calibration [14] can improve OOD detection and misclassification out of domain. We observe a significant increase in misclassification detection under strong domain shifts after improving calibration. We also observe improvements for OOD detection, albeit milder.

	Sem. segm.	Robust performance	Uncertainty estimation	Natural shifts	OOD calib methods
Kamann *et al*. [25]	✓	✓			
Bhojanapalli *et al*. [3]		✓		✓	
Xie *et al*. [55]	✓	✓			
Naseer *et al*. [36]		✓			
Bai *et al*. [2]		✓		✓	
Minderer *et al*. [33]		✓	✓	✓	
Paul and Cheng [43]		✓		✓	
Mao *et al*. [32]		✓		✓	
Liu *et al*. [30]		✓		✓	
Zhou *et al*. [59]	✓	✓			
Pinto *et al*. [44]		✓	✓	✓	
Ours	✓	✓	✓	✓	✓

Table 1. **Studies of recent architectures**. While several prior works studied robustness and uncertainty of transformer- and CNN- based *classifiers*, studies on *segmentation* limited to robustness. This is the first study assessing robustness *and* uncertainty of modern segmentation models. Moreover, we consider natural domain shifts and are the only analysis to include content-dependent methods [14, 17] to improve calibration in OOD settings.

2. Related work

We study robustness and uncertainty in semantic segmentation. In doing so, we touch several fields, which we cover in the following. We further discuss related studies.

Segmentation models. Modern segmentation pipelines typically consist of encoder-decoder architectures [1, 13, 39, 46]. Decoders are usually designed *ad hoc* for segmentation, with DeepLab [5–7] and UPerNet [54] being two of the most prominent. On the other hand, the evolution of encoders has been closely related to that of classification models, with ResNet [21] being one of the most popular for years. The rise of transformers in computer vision [15] has led to a flurry of works leveraging self-attention for segmentation [48, 55, 58]. Novel convolutional architectures inspired by transformers have also risen [30]. We compare several recent segmentation models against ResNet baselines, in terms of both robustness and uncertainty.

Robustness. The brittleness of neural networks to changes in the input domain is a well-studied problem and many sub-formulations exist [49]. Robustness against synthetic shifts takes into account samples crafted by artificially altering images, for example injecting noise or blur (corruption robustness [23, 25]), or crafting imperceptible perturbations to induce model failure (adversarial robustness [18]). Robustness against *natural* shifts focuses on changes that may arise naturally, without human intervention [24, 45].

In this work we are interested in comparing the robustness of different off-the-shelf segmentation models under

natural domain shifts, since these are particularly relevant in real-world applications. In particular, we focus on semantic segmentation of urban scenes, hence, we evaluate models on samples from unseen geographical locations [51] and weather conditions [47]. Segmentation robustness against natural shifts has been studied before [52, 56], yet not in tandem with uncertainty and within a large-scale study taking into consideration several recent models.

Uncertainty. Guo *et al*. [19] have shown that deep models are overconfident. They have proposed a simple, yet effective solution known as temperature scaling (TS) where the output logits are divided by a temperature parameter before the softmax layer. Other calibration methods have been proposed (*e.g*., [20, 28, 35]), but TS is still very popular due to its simplicity and the fact that it does not alter predictions.

Calibration with TS is effective in ID settings; yet, Ovadia *et al*. [40] have shown that model calibration degrades significantly out of domain. Some methods have been proposed that address this problem [41, 42, 53], by assuming access to unlabeled OOD images beforehand. On the other hand, Gong *et al*. [17] have proposed methods that improve OOD calibration without any data from the target domain. They propose to cluster the calibration set in different "domains" and find a different temperature value for each. At test time, images are calibrated using the temperature from the closest cluster. *Ad hoc* for semantic segmentation, Ding *et al*. [14] propose a content-dependent calibration strategy that learns a small calibration network to predict a temperature for each pixel in an image.

In our study we test ID and OOD performance of several calibration methods, focusing on techniques that do not require access to OOD samples [14, 17, 19]—as generally robustness is evaluated on unseen domains [23, 25, 44, 49]

Previous analyses. In Tab. 1 we compare related studies on different aspects of reliability. Several works have suggested that transformer-based *classifiers* are more robust than CNNs [2, 3, 32, 36, 43]. Yet, the recent ConvNeXt [30] has challenged this result and later work have suggested that further investigation is needed [44]. Minderer *et al*. [33] have compared calibration of several classifiers, concluding that convolution-free models are more robust *and* better calibrated. In contrast, Pinto *et al*. [44] have compared recent transformers and CNNs, arguing there is "no clear winner". Some works have compared the robustness of transformers and CNNs for segmentation [55, 59]—but only against *synthetic* domain shifts. We broadly study robustness *and* uncertainty in segmentation under *natural* domain shifts. Similarly to [44], we do not observe a single model family which is better calibrated in all scenarios. In contrast with [33] though, we observe that robustness and calibration *do not* go hand in hand. This shows that not all trends observed in classification transfer to segmentation, confirming the importance of task-specific studies like ours.

3. Experimental settings and preliminaries

3.1. Datasets

As discussed in Section 2, we use different datasets for semantic segmentation of urban scenes to model natural domain shifts—inspired by prior art [10, 52, 56].

Cityscapes (CS) [12] contains images taken across 50 European cities at day time with overall good weather. Training and validation sets use sequences from disjoint sets of cities. Following this protocol, we further split validation cities into a calibration and a test set. Since CS is a mainstream benchmark in semantic segmentation, we use it as our training set (ID) to leverage available trained weights.

IDD [51] was captured in the cities of Hyderabad, Bangalore and their outskirts. Given the different geographical location, it poses a clear domain shift for CS models.

ACDC [47] contains images captured in adverse conditions (Fog, Rain, Snow and Night), which translate into strong domain shifts. Similarly to previous work [44, 55, 59], we focus on *covariate* shifts, *i.e*., changes in the input distribution—keeping the label set fixed. In practice, for OOD settings (IDD and ACDC) we consider the 19 classes from CS, ignoring the others. The only exception is one experiment in Appendix L, where we consider label shifts.

3.2. Architectures

We implement our models with MMsegmentation [11]. Following prior work [33, 44], we use the original training recipes for each model to compare them at their best. For completeness, we also explore the effects of pre-training dataset and number of training iterations in Appendix I.

SETR [58]. The first convolution-free segmentation model. It uses a ViT backbone [15] and different decoders (SETR-Naive, SETR-MLA and SETR-PUP). We use the ViT-Large backbone and analyze all three decoders.

Segmenter [48]. Similarly to SETR, it also uses a ViT backbone; yet, unlike the simpler SETR decoders, it carries a transformer-based one. We also test ViT-Large.

SegFormer [55]. This model incorporates an original self-attention mechanism and several architectural changes to be more efficient. We evaluate all models from this family (B0–B5), gradually increasing the number of parameters.

ConvNeXt [30]. Convolutional model with changes inspired by transformers. We use ConvNeXt-Large, comparable in size to ViT-Large, with an UPerNet decoder [54].

ResNet-based [22]. We use ResNet-V1c model, the default ResNet in MMsegmentation library [11]. Compared to vanilla ResNet [21] it uses a stem with three 3x3 convs (instead of a 7x7 conv). We use ResNet-18/50/101 models and two popular decoders: **DLV3+** [6] and **UPerNet** [54]

Additionally, in Appendix J we explore **Mask2Former**, an architecture for *universal image segmentation* [8] that does not follow the conventional *logits* + *softmax* paradigm.

3.3. Reliability metrics

We evaluate model reliability on four aspects: robustness, calibration, misclassification detection and OOD detection.

Robustness. We measure robustness by evaluating the standardized mean Intersection-over-Union (mIoU) performance in OOD settings, *i.e.*, on ACDC and IDD. We also provide ID performance, evaluating models on CS.

Calibration. A model is said to be calibrated when the predictive probabilities (*i.e.*, the logits after a softmax) correspond to the true probabilities. For instance, if we group all samples where the predicted probability is 90%, we would expect that 90% of those predictions are correct. The most common calibration metric is the *Expected Calibration Error (ECE)* [35], which looks at the expected difference between the predicted and actual probabilities. To estimate the ECE we quantize the predicted probabilities and compare the accuracy with the mean probability in each bin. Since the binning strategy can affect the results, we test ECE with equally spaced bins [35], equally populated bins [37, 38] and the Kolmogorov-Smirnov test [20], which gets rid of the binning strategy altogether. We report results with standard ECE, but find that all three aforementioned metrics yield similar conclusions (see Appendix A).

Given that in segmentation we have per-pixel predictions, the number of calibration samples explodes (a single CS image contains $2048 \times 1024 \simeq 2$M pixels). We ablate the different calibration metrics as we sub-sample the number of pixels per image and observe that 20k pixels per image in enough to estimate the ECE (see Appendix B).

Misclassification detection. A desiderata for a reliable model is to assign a larger confidence to correct outputs than incorrect ones[2]. In the ideal case, if we sorted all predictions from least to most confident, we would have all the incorrect predictions first and correct ones later. *Misclassification detection* measures how far away are we from such an ideal case. This can be measured with *Rejection-Accuracy curves* [16, 24]: we reject samples with low confidence and compute the accuracy *vs*. amount of rejected samples. However, these are biased in favor of better-performing models, since the base accuracy is higher in the first place. To avoid that, we follow Malinin *et al*. [31] and normalize the area under the curve by that of an oracle and subtracts a baseline score with randomly sorted samples. The resulting metric, known as the Prediction Rejection Ratio (PRR), will be positive if the confidence is better than the random baseline and will have a maximum score of 100% when the model matches the oracle.

Out-of-domain detection. Another important aspect in reliability is that models are aware of their "domain of expertise" (*i.e.*, their training domain). When a sample differs significantly from the training samples, we would expect the model to be more uncertain of its prediction. Similarly to misclassification detection, in OOD detection we try to separate ID from OOD images based on the network confidence. This has broad applicability, *e.g.*, generating alerts for samples too far from the training domain or, connecting to active learning, gathering them for further review and annotation. As in the rest of our work, we consider a whole image to be out of domain if it presents a significant domain shift; that is, we consider CS in domain and IDD/ACDC out of domain. Since we define ID and OOD samples at the image level, we consider the average confidence of all pixels in a given image. We use the Area Under the Receiver Operating Characteristic curve (AUROC) [34], which goes from 0 to 1 (1 being the best score). Additionally, in Appendix L we consider OOD detection at a region level, considering classes in the IDD dataset not present in CS, *i.e.*, *rickshaw*, *billboard*, *guard rail*, *tunnel* and *bridge*.

[2]We use max softmax as confidence in the paper; in Appendix C we present similar results with negative entropy.

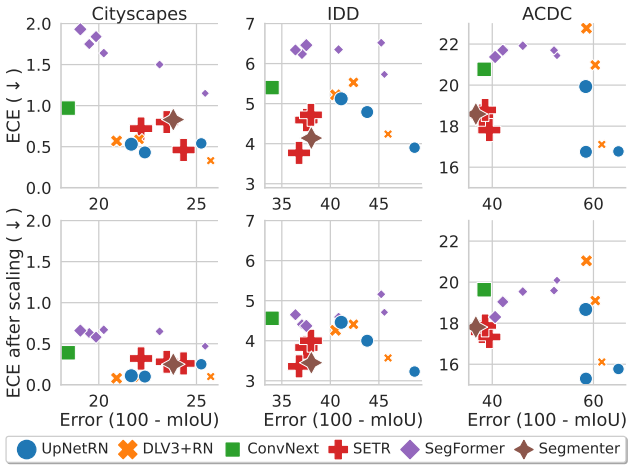

Figure 2. **Expected mIoU error** (↓) ***vs*. calibration error** (↓) before and after TS for different model families (top and bottom, respectively). All models trained and calibrated on CS. Marker-size proportional to number of parameters. Notice how TS yields marginal *relative* gains in OOD settings (IDD and ACDC).

4. Are modern segmentors more *reliable*?

In the following, we present the main findings of our study on the reliability of semantic segmentation models.

4.1. Robustness

Since the domain shifts we consider are natural and not synthetically induced, there is no straightforward way to evaluate their strength. To this end, we establish an ordering for the severity of the shift based on performance degradation of the ResNet baselines (DLV3+R101 and UPNetR101), which results in CS < IDD < ACDC. This aligns with a qualitative evaluation of the different datasets (see

Fig. 1, bottom, for a few samples). In Fig. 1 (top left) we present the mIoU error (*i.e.*, 100 − mIoU) for several models evaluated on the three datasets. To highlight the loss in performance as we increase the shift, we normalize all errors w.r.t. the best performing CS model (ConvNeXt). The trend is clear: the larger the domain shift, the larger the improvement brought by more recent segmentation models.

We expand on this in Fig. 2 (top), where we plot the mIoU error (on the x-axis) *vs*. the calibration error for all the models belonging to the different families (Sec. 3.2). The size of the markers is proportional to the number of parameters. Similarly to Fig. 1, we observe that the gap in mIoU between ResNet baselines and recent models grows larger as we increase the domain shift (*cf*. marker position on the x-axis). Interestingly, two of the models that perform best under ACDC's strong shift (SETR and Segmenter) are not significantly better than the ResNet baseline on the training domain (CS). This indicates that, in our setting, only assessing ID performance can hide the real value of newly crafted models and, hence, it is important to evaluate architectures out of domain in order to fully grasp their potential.

While there is no single model family that performs significantly better in all datasets, we can reach the clear conclusion that *all recent models are significantly more robust than well established baselines under natural shifts*.

4.2. Calibration error

Off-the-shelf calibration. In Fig. 1 (top-right) we present the ECE for different models as we increase the domain shift. Similarly to the mIoU error, ECE is normalized by the best CS model (in terms of calibration). Interestingly, despite the remarkable improvements in terms of robustness, recent models are not significantly better calibrated. When moving from CS to ACDC, the *relative* mIoU error increases by a factor ∼2× for recent models *vs*. ∼3× for ResNet baselines; yet, in terms of *relative* calibration error, all models increase by a ∼40× factor. This clearly highlights the need for further advances in model calibration.

In Fig. 2 (top) we show the ECE *vs*. mIoU error for all models and datasets. When it comes to calibration *vs*. robustness trade-off, there is no clear winner among the model families we consider. Moreover, we do not observe a clear trend between mIoU and ECE in any domain.

Calibration with TS. In Fig. 2 (bottom) we present the same results after applying TS [19], tuned on CS. Comparing top and bottom, we observe an overall improvement in calibration for all networks. In particular, SegFormer models (◆), which had the largest ECE on CS and IDD, seem to benefit the most from TS. Nevertheless, even after this improvment, OOD calibration error (IDD, ACDC) remains significantly larger than the ID one (CS) for all models. Regarding ECE *vs*. mIoU out of domain, a mild trend emerges after TS: For transformer models, better-calibrated models

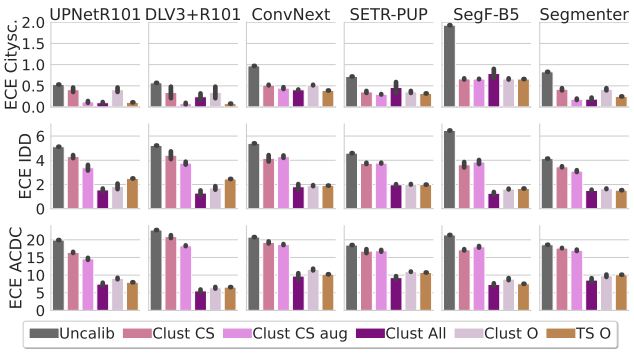

Figure 3. **ECE (↓) after clustering TS** for a selection of models (best mIoU on CS per family). If the calibration samples are representative of the test domains *Clust All* can indeed improve ECE, however, without access to OOD samples (*Clust CS* and *Clust CS aug.*) benefits of clustering are limited. Oracle baselines (*O*) always use calibration images from the test domain.

are also the most robust (before TS, SegFormer (◆) did not follow this trend). On the other hand, for ResNet baselines (×,●) the better-calibrated models are generally the most brittle. Regarding ConvNeXt (■), although it is one of the best performing models, it seems to be worse calibrated than other models with similar mIoU. Overall, *recent models are not significantly better calibrated than ResNet baselines neither before nor after TS*.

4.3. Can we improve out-of-domain calibration?

Calibration error on samples from the training domain in not alarming, especially after TS. Nonetheless, the sharp increase in ECE out of domain is concerning for many applications, especially since recent segmentation models do not show a clear improvement in this direction. This renders methods that seek to improve calibration out of domain all the more important, but yet this is a rather underexplored research area. As discussed in Sec. 2, to the best of our knowledge, only Gong *et al*. [17] tackle OOD calibration *without* additional information about the test domain. They suggest clustering the calibration set into multiple "domains" based on the image features extracted by the network. A different temperature per cluster is then selected and test-time predictions are scaled according to the cluster assigned to the images. This *adaptive* TS method was originally devised for classification, but we extend it to the segmentation task by scaling all the logits of a given image with the same temperature. Regarding the number of clusters, we find 16 to be a reasonable number (see Appendix D for this analysis).

Clustering on different calibration sets. In Fig. 3 we present the results of calibrating with clusters computed on different calibration sets. Since our training dataset is CS, we assume that only CS images are available for calibration. As a naive alternative to obtain more diverse clusters, we in-

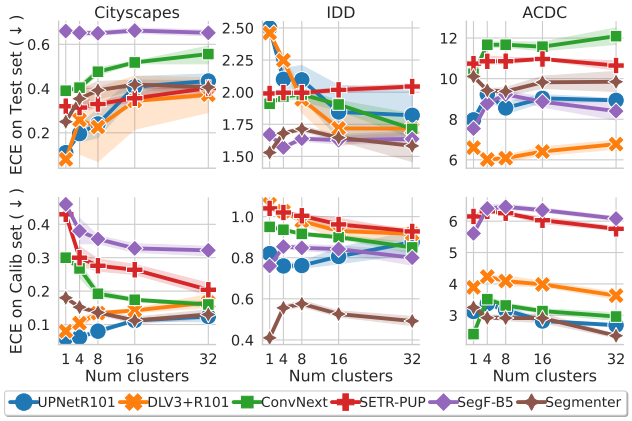

Figure 4. **ECE** (↓) ***vs*. number of clusters for oracles** (calibration images extracted from the test domain). Even when evaluating on the calibration set, where there can be no overfitting of the temperatures, ECE does not decrease monotonically as we increase the number of clusters (one cluster is equivalent to vanilla TS).

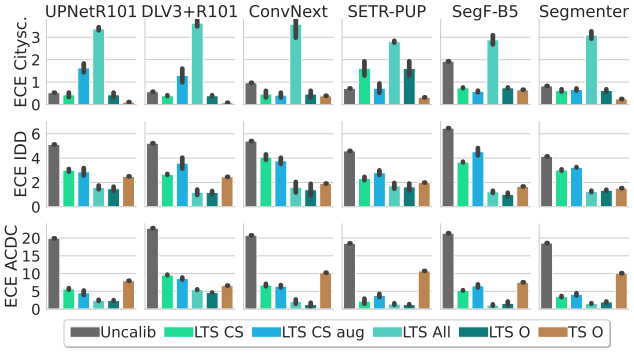

Figure 5. **ECE** (↓) **after local TS (LTS)** for a selection of models (best mIoU on CS per family). Even without access to OOD samples (*LTS CS*), LTS calibration improves ECE out of domain, especially under strong domain shifts (ACDC). With access to OOD samples (*LTS All*), ECE out of domains improves further, albeit it degrades in domain.

troduce a *CS aug.* dataset where some of the CS calibration images are randomly augmented with different transformations (*e.g.*, color scaling, changes in brightness, contrast, etc). The rationale is that more diverse clusters might generalize better to new domains.

To assess how beneficial OOD samples can be during calibration, we also add another calibration set which contains images from all the datasets (CS, IDD and ACDC) mixed together, we will refer to it as *All*. This serves as a sort of upper bound, since our main goal is still assessing robustness in unseen domains. Furthermore, we introduce two more oracle baselines (*O*), which use calibration images from the test domain. For instance, when evaluating on IDD, the oracle calibration set will consist of *only* IDD while *All* will contain images of IDD, ACDC and CS mixed. One oracle baseline uses clustering, while the other uses vanilla TS (*Clust O* and *TS O*, respectively).

As expected, calibrating on all datasets (*Clust All*) significantly improves ECE, with comparable performance to oracles in most settings. In contrast, without access to OOD samples (*Clust CS*), calibrating with the method by Gong *et al*. yields rather limited improvements. Moreover, increasing cluster diversity via data augmentation (*Clust CS aug.*) is not always beneficial.

To gain more intuition, we visualize the cluster assignments (see Appendix E). When using all datasets for calibration, test-time images are qualitatively close to the assigned clusters; yet, with *Clust CS* or *Clust CS aug.*, OOD images do not blend well with the calibration images of their corresponding clusters. We argue that one implicit assumption for clustering to work well is that test-time images are close to one of the clusters (domains) in the calibration set; therefore, under strong domain shifts, it is unlikely to bring much improvement.

On the bright side, if representative images from the deployment domains are available, clustering could be applied to allow a single model to be calibrated on multiple domains. Of course, with OOD annotated samples, additional fine-tuning or adaptation techniques could be applied, but this is out of the scope of this study, since our focus is *off-the-shelf* model robustness, without adaptation.

Clustering does not improve ECE in domain. Comparing oracles in Fig. 3, we can observe that clustering does not improve significanly over TS. In some settings, it is even worse. One possible explanation would be that this is due to an overfitting of the temperature parameters to the particular calibration clusters. However, in Fig. 4 we observe that even when evaluating the ECE in the calibration set, the error does not monotonically decrease with the number of clusters. Although somewhat surprising, this is in fact possible since decreasing the ECE for several disjoint subsets of images (clusters) independently does not guarantee that the ECE on the union set will decrease. We provide a formal theorem in Appendix F, in support of this claim. Note that we are not asserting that clustering will not improve ECE *in general* (we empirically observed it can, if provided with a representative calibration set) but rather that it is not guaranteed to do so.

To sum up, *clustering the calibration set does not bring significant improvements unless representative images of the test domain are present in the calibration set*. Moreover, *it is not better than TS for ID calibration.*

4.3.1 Adaptive temperature via calibration network

The partial failure of the clustering approach motivates us to investigate other methods that adjust the temperature adaptively w.r.t. the input, since we can expect this to help im-

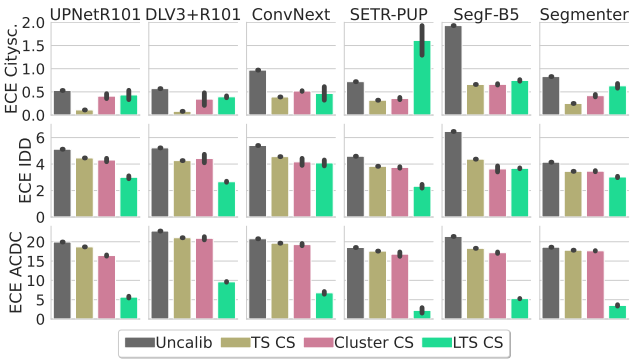

Figure 6. **Comparison of calibration methods.** ECE (↓) after calibration for a selection of models (best mIoU on CS per family). All models are calibrated on CS calibration set. LTS is markedly the best calibration method out of domain with remarkable improvements under strong domain shifts.

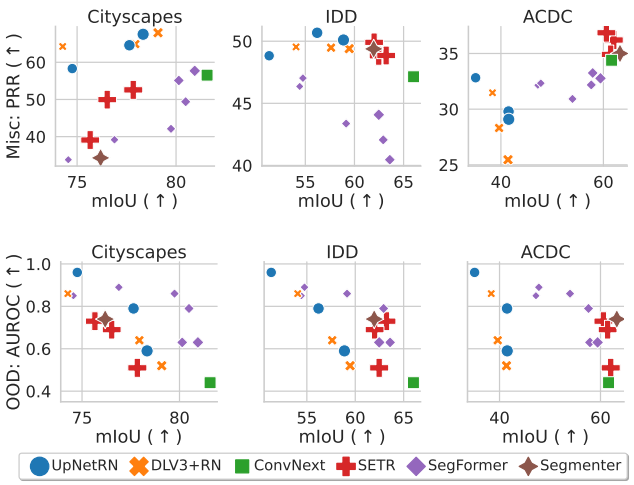

Figure 7. **Misclassification and OOD detection *vs*. Robustness** for different segmentation models and datasets. **(Top) PRR** (↑) ***vs*. mIoU** (↑)**:** ResNet-based models (●,×) outperform more recent models (other markers) in ID misclassification detection (CS, left), but the trend is opposite under strong domain shifts (ACDC, right). **(Bottom) AUROC** (↑) ***vs*. mIoU** (↑)**:** There is no free-lunch between robustness and OOD detection in any considered domain.

proving OOD calibration. In this regard, Ding *et al*. [14] suggest training a small calibration network that predicts the temperature values as a function of both the input image and the segmentation model logits. This Local Temperature Scaling (LTS) method is specific to segmentation so the output is not a single temperature per image but a "temperature map" with pixel-level temperature values. Despite not being designed for OOD conditions, our intuition is that a network providing sample-dependent temperatures can be beneficial under domain shifts.

LTS using different calibration sets. As in Sec. 4.3, we test LTS on multiple calibration sets (*CS*, *CS aug.*, *All*), which in this case are used to learn the calibration network. Also here, we compare against oracles, for which the calibration network is learned using images from the test domains. Results are shown in Fig. 5. Interestingly, we find that LTS using only CS images for calibration (*LTS CS*) leads to a noticeable improvement in OOD ECE. In particular, when testing on ACDC—where the domain shift is stronger—*LTS CS* outperforms even TS with access to images on the test domain (*TS O*) for some models. Also in these experiments, introducing naive data augmentations on CS (*LTS CS aug.*) does not yield substantial improvements.

When using all the datasets for calibration (*LTS All*), OOD results improve even further; yet, there is a noticeable increase in calibration error on CS. Unlike clustering, where the temperature was optimized independently for each cluster, LTS trains the calibration network using all the samples at once and samples with large calibration error (like ACDC or IDD) may dominate the loss. We hypothesize that further improvements in the architecture and training schedule of the calibration network can lead to even better performance and are promising directions for future work.

Focusing on the oracle baselines, LTS outperforms TS on IDD and ACDC, but TS outperforms LTS on CS (*cf*. *LTS O* and *TS O*). This may be due to the fact that CS is a very homogeneous dataset if compared to the other two, hence, it can be reasonable that a simpler method may perform best.

Comparing all calibration methods. In Fig. 6 we compare *TS CS*, *Clust CS* and *LTS CS* (all calibrated on CS). LTS is markedly the best calibration method out of domain, especially under stronger domain shifts. In domain, TS works best, but it does not bring significant improvements out of domain. Since LTS predicts the temperature parameter at the pixel level, this motivates an ablation of clustering where we predict different temperatures per image; yet this does not improve results (see details in Appendix G). Additionally, we perform an ablation of LTS using only the image or the logits for calibration. Image information seems to be more important for OOD calibration, while the logits are more important in the ID setup (see Appendix H).

4.4. Misclassification detection

In Fig. 7 (top) we compare all models in terms of *misclassification detection vs. robustness* —PRR score (↑) *vs.* mIoU (↑). In domain, we observe a clear trend: within the same model family, better performing models tend to also show better PRR. However, when considering all models, higher mIoU does not generally imply higher PRR and ResNet-based backbones perform significantly better than more recent architectures. As we increase the domain shift, the trend changes: for ACDC, recent models perform best both in terms of mIoU and PRR. Moreover, out of domain, ResNet families show a negative correlation where better

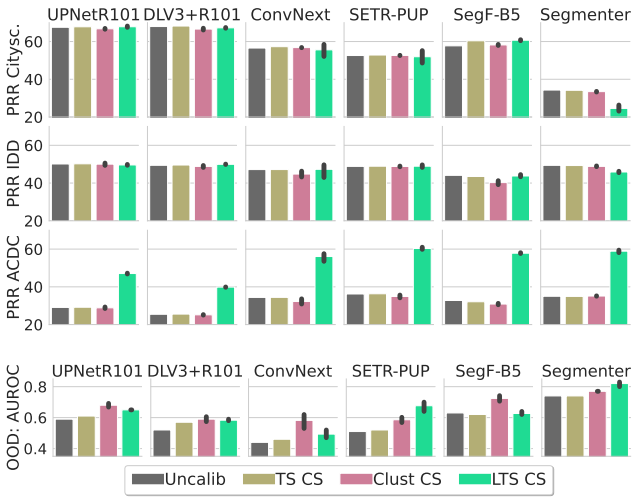

Figure 8. **Misclassification and OOD detection after calibration** for several models after applying different calibration techniques using CS samples. **Misclassification – PRR** (↑) **(first 3 rows)**: Under strong domain shifts (ACDC), LTS calibration significantly improves PRR. **OOD detection – AUROC** (↑) **(last row)**: Both clustering and LTS yield improvements in OOD detection.

mIoU leads to worse PRR. Overall, *recent models seem to improve misclassification detection under strong domain shifts, but underperform baselines in domain.*

4.5. Out-of-domain detection

In Fig. 7 (bottom) we compare *OOD detection vs. robustness* (AUROC score *vs*. mIoU) for all models. To measure OOD detection we separate the ID images (CS) from the OOD ones (IDD and ACDC). Therefore, the y-axis is the same in all three plots and only the mIoU changes. For OOD detection, there is a marked negative trend between CS mIoU and AUROC. When looking at IDD and ACDC, the negative trend continues but there seems to be a distinction between ResNet baselines and other recent models: the latter perform better in terms of mIoU, but at the same time show a drop in AUROC. In short, *there is no free-lunch between robustness and OOD detection. In terms of OOD detection, a small ResNet-18 (●) performs best.*

4.6. Can calibration improve misclassification and out-of-domain detection?

In Fig. 8 we show misclassification (top) and OOD (bottom) detection metrics for different models after calibration using only CS samples. For misclassification, we observe a sharp PRR improvement on ACDC after we calibrate models with LTS. This is encouraging as it indicates that the calibration network learned in LTS can help discern correct from incorrect predictions given its output temperature. We do not observe significant improvements in other datasets or with other methods. This is reasonable, since the largest calibration gain was observed with LTS on ACDC. (see Fig. 6).

Regarding OOD detection (Fig. 8 bottom) we observe that both clustering and LTS calibration can improve OOD detection. We find this interesting, since clustering on CS did not improve OOD calibration significantly. Although the clusters using only CS images are not representative enough to produce adequate temperatures for IDD or ACDC, OOD samples are assigned to clusters which have larger temperatures. This is enough to decrease the confidence for OOD samples compared to ID and leads to better OOD detection. Similarly for LTS, the calibration network assigns larger temperatures to OOD images.

In conclusion, *adaptive TS techniques are a promising avenue to improve OOD detection and misclassification detection under strong domain shifts.*

5. Conclusion

We have studied the reliability of recent segmentation models—in terms of **robustness** and **uncertainty estimation** under natural domain shifts. Overall, while no single model family is better in all scenarios, recent models are remarkably more robust to domain shifts than ResNet baselines. Yet, this does not translate into better *calibration*—severely degraded out of domain. Thus, it is crucial to find methods to improve model calibration in OOD settings. To this end, we have explored state-of-the-art methods and found that Local Temperature Scaling [14], although originally devised for ID settings, is a promising technique.

Furthermore, we have explored *misclassification* and *OOD detection*—two other important tasks regarding uncertainty estimation. We have shown that recent and more robust models tend to perform better at misclassification under strong domain shifts, but yet they underperform ResNet baselines ID. On the other hand, OOD detection under domain shifts is negatively correlated with the mIoU, which translates into a trade-off between robustness and uncertainty. Finally, we find that adaptive temperature scaling techniques can help beyond calibration and improve OOD detection and misclassification in some settings.

All in all, although we appear to be *on the right track* for what concerns robustness, our findings motivate the need to improve reliability of segmentation models in other dimensions, where results are not equally postive. In that regard, we identify several promising directions which we hope may encourage future research on this important topic.

Acknowledgements We thank Francesco Pinto and Gabriela Csurka for helpful discussions. Prof. Philip Torr is supported by the UKRI grant: Turing AI Fellowship EP/W002981/1 and EPSRC/MURI grant: EP/N019474/1. We would also like to thank the Royal Academy of Engineering and FiveAI.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: a Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(12):2481–2495, 2017. 2

[2] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than CNNs? In *NeurIPS*, 2021. 2, 3

[3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *ICCV*, 2021. 2, 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 22

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(4):834–848, 2017. 2

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv:1706.05587, 2017. 1, 2, 3

[7] 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. 2

[8] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *CVPR*, 2022. 3, 22

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 22

[10] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne Kim, Seungryong Kim, and Jaegul Choo. RobustNet: Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening. In *CVPR*, 2021. 3

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

[12] 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 *CVPR*, 2016. 3

[13] Gabriela Csurka, Riccardo Volpi, Boris Chidlovskii, et al. Semantic image segmentation: Two decades of research. *Foundations and Trends® in Computer Graphics and Vision*, 14(1-2):1–162, 2022. 2

[14] Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local temperature scaling for probability calibration. In *ICCV*, 2021. 2, 3, 7, 8, 19, 20, 28

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 2, 3

[16] Giorgio Fumera and Fabio Roli. Support vector machines with embedded reject option. In *International Workshop on Support Vector Machines*, 2002. 4

[17] Yunye Gong, Xiao Lin, Yi Yao, Thomas G Dietterich, Ajay Divakaran, and Melinda Gervasio. Confidence calibration for domain generalization under covariate shift. In *ICCV*, 2021. 2, 3, 5, 13, 14, 15, 16, 19, 26

[18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *ICLR*, 2015. 2

[19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 1, 2, 3, 5

[20] Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. arXiv:2006.12800, 2020. 3, 4

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3

[22] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *CVPR*, 2019. 3

[23] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *ICLR*, 2019. 1, 2, 3

[24] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 2, 4

[25] Christoph Kamann and Carsten Rother. Benchmarking the Robustness of Semantic Segmentation Models. In *CVPR*, 2020. 2, 3

[26] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 22

[27] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999. 1

[28] 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. In *NeurIPS*, 2019. 3

[29] 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 *ICCV*, 2021. 2, 22

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022. 1, 2, 3, 22

[31] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. arXiv:1905.00076, 2019. 4

[32] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *CVPR*, 2022. 2, 3

[33] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *NeurIPS*, 2021. 2, 3, 22

[34] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 4

[35] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, 2015. 3, 4

[36] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *NeurIPS*, 2021. 2, 3

[37] Khanh Nguyen and Brendan O'Connor. Posterior calibration and exploratory analysis for natural language processing models. arXiv:1508.05154, 2015. 4

[38] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, 2019. 4

[39] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning Deconvolution Network for Semantic Segmentation. In *ICCV*, 2015. 2

[40] 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. *NeurIPS*, 2019. 1, 3

[41] Anusri Pampari and Stefano Ermon. Unsupervised calibration under covariate shift. arXiv:2006.16405, 2020. 3

[42] Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *AISTATS*, 2020. 3

[43] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *AAAI*, 2022. 2, 3

[44] Francesco Pinto, Philip HS Torr, and Puneet K Dokania. An impartial take to the cnn vs transformer robustness contest. In *ECCV*, 2022. 2, 3, 22

[45] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 1, 2

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 2

[47] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding. In *ICCV*, 2021. 3

[48] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for Semantic Segmentation. In *ICCV*, 2021. 1, 2, 3

[49] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, 2020. 2, 3

[50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2

[51] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and C.V. Jawahar. IDD: A Dataset for Exploring Problems of Autonomous Navigation in Unconstrained Environments. In *WACV*, 2019. 3

[52] Riccardo Volpi, Pau De Jorge, Diane Larlus, and Gabriela Csurka. On the Road to Online Adaptation for Semantic Image Segmentation. In *CVPR*, 2022. 3

[53] Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable calibration with lower bias and variance in domain adaptation. *NeurIPS*, 2020. 3

[54] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified Perceptual Parsing for Scene Understanding. In *ECCV*, 2018. 1, 2, 3

[55] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *NeurIPS*, 2021. 1, 2, 3

[56] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization without Accessing Target Domain Data. In *ICCV*, 2019. 3

[57] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 22

[58] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In *CVPR*, 2021. 1, 2, 3

[59] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *ICML*, 2022. 2, 3