Category Differences Matter: A Broad Analysis of Inter-Category Error in Semantic Segmentation

Jingxing Zhou
Porsche Engineering Group GmbH
jingxing.zhou@porsche-engineering.de

Jürgen Beyerer
Fraunhofer IOSB & Karlsruhe Institute of Technology
juergen.beyerer@iosb.fraunhofer.de

Abstract

In current evaluation schemes of semantic segmentation, metrics are calculated in such a way that all predicted classes should equally be identical to their ground truth, paying less attention to the various manifestations of the false predictions within the object category. In this work, we propose the Critical Error Rate (CER) as a supplement to the current evaluation metrics, focusing on the error rate, which reflects predictions that fall outside of the category from the ground truth. We conduct a series of experiments evaluating the behavior of different network architectures in various evaluation setups, including domain shift, the introduction of novel classes, and a mixture of these. We demonstrate the essential criteria for network generalization with those experiments. Furthermore, we ablate the impact of utilizing various class taxonomies for the evaluation of out-of-category error.

1. Introduction

Deep convolutional neural networks (CNN) [15, 16, 19, 31, 33] and more recently transformer-based neural networks [9,32,60] pushed the boundary of computer vision and pattern recognition forward in the past decade, which also enabled data-driven automated driving in certain scenarios. Meanwhile, the robustness of the neural networks has become a hurdle to overcome for the serial introduction of automated vehicles in public traffic, and it is considered as a crucial criterion of driving functions during validation and certification [21, 22]. Recent work from academia and industry has emphasized the robustness concern, including anomaly detection [8, 53], out-of-distribution detection [10, 36] and open-set classification [23, 34, 68]. Since the real-world driving environment is complicated and constantly changing, ranging from weather and light conditions to traffic behavior, it is essential to ensure that effective features are learned during network training so that the networks can be generalized to a wide range of possible driving conditions.

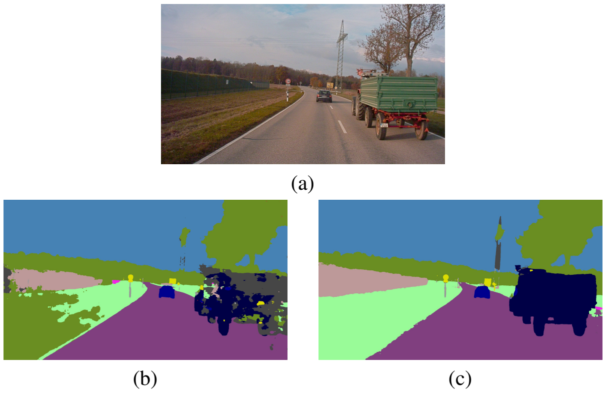

Figure 1. Motivation for category-related semantic segmentation evaluation: Given the input image (a) from A2D2 dataset [12]. The tractor is a previously unknown object as the networks are trained with Cityscapes [7]. While output (b) considers part of the tractor as *building*, (c) depicts the object as *truck*, which is part of the category *vehicle*. In this work, we quantitatively evaluate category-relevant prediction errors as a complement to the IoU metric.

Semantic segmentation plays an essential role in scene interpretation for automated driving, aiming to assign a class label to each pixel of the input image from a predefined class set. However, in the current evaluation scheme, the difference between whether an occurring classification error assigns a comparable class (w.r.t. the given use case) or a critically different class is disregarded. In the real-world scenario, erroneously assigned class may represent divergent levels of criticality: Labeling a *rider* as a *pedestrian*, for example, would typically be considered a less critical and more understandable confusion than mistaking the *rider* for a static *wall* – while established metrics for the performance of neural networks would assign equal scores to both results as they represent false predictions, as shown in Fig. 1.

To the best of our knowledge, previous work of the semantic segmentation evaluation metric mainly concentrated on region-based agreement of the predictions with ground truth [66], while this type of evaluation is not sufficiently addressed, which distinguishes between errors within comparable classes, based on a given hierarchy, and errors across critically distinct classes. Therefore, we aim to introduce a novel evaluation metric for semantic segmentation in the context of automated driving, emphasizing prediction errors based on an equivalence relationship that can be derived from the class taxonomies of the datasets. Beside that, with the new evaluation metric, multi-dataset semantic segmentation can be evaluated with higher flexibility since label remapping is considered to be intricate, time-consuming, and introduces ambiguities according to their class definitions [27].

In this work, we differentiate between the definitions of *class* and *category*, where we consider *class* as the leaf element of the class hierarchy, while *category* indicates their parent nodes, or superclasses. We evaluate some frequently-used network architectures with three different setups to evaluate the properties of the neural networks. We perform our experiments on five automotive-centric datasets that are frequently used in the research, unveiling the properties and differences of the models trained by certain datasets. Our main contributions are as follows:

- **Novel evaluation metric:** We propose a novel evaluation metric for semantic segmentation that specifically distinguishes in-category and out-of-category errors, which also facilitates the evaluation of domain shift and domain generalization as the proposed metric reduces the effort needed for a unified label space.
- **Variety of experiments:** We conduct the experiments not only in the source domain but also under domain shift, and with novel-classes, covering a range of architectures from classical neural networks to recent transformers. In addition, we also investigate the impact of exploiting different class taxonomies to address safety concerns in the automated driving domain.
- **In-depth analysis:** We analyze the behavior of various neural networks, providing new insights from the neural networks when they encounter underrepresented or previously unknown classes, which is essential for safety-critical applications like automated driving.

2. Related Work

2.1. Robustness Evaluation of the Neural Networks

Previous work has shown the challenges of learning effective features from the training datasets and keeping high variance for data perturbations [11, 51, 52]. If the input distribution shifts from the original training set, which can be frequently observed in real-world applications, abnormally high softmax confidences may appear, leading to unreliable predictions [39]. For the evaluation of robustness, models are frequently trained on a clean source dataset and tested on another datasets [17, 18], which include corrupted images or out-of-distribution samples, in a zero-shot fashion without any adaptation. Most evaluations are based on metrics like class-wise IoU, accuracy and Dice, which demand predicted labels matching the ground truth, even when class hierarchy is explicitly utilized for training regularization [29]. Beside that, certain architectural characteristics may harm network robustness, making predictions only trustworthy when input images are devoid of noise and blur [24]. To enhance against such disturbances, methods like advanced data augmentation strategies [55, 61, 62, 64], large-scale supervised and unsupervised pretraining [5, 45, 48, 70] can be utilized.

Recent research also suggests that transformers exhibit greater resilience than CNNs when exposed to out-of-distribution samples [28, 65] or adversarial attacks [37, 42]. However, if a similar training setup is applied, transformers are not proven to be more robust than CNNs, although their operating mechanisms are vastly dissimilar with different focus [1, 37, 57]. Another aspect that is often disregarded is the availability of the data points during the training process [69], as recent work [26, 44, 58] demonstrated the potential of robustness when the model is trained with large-scale cross-domain composite datasets.

2.2. Domain Generalization

Domain Generalization (DG) focuses on learning general feature representations from the source domain so that the learned feature representations can handle arbitrary data distribution, which is not available during training time. Lots of current research concentrates on image classification [13, 50], while only a small proportion of work focuses on the semantic segmentation scenario [4, 6, 20, 40, 41, 43]. Pan *et al.* [40] integrated two different normalization layers to ensure that features are invariant to changes in appearance. The effectiveness of feature presentation normalization is further improved with the assistance of domain-variant frequency analysis [20]. Other approaches have attempted to normalize global features by removing style-specific information [6] or using other normalization layers [54]. Data augmentation methods [43] generate additional training samples with style transfer models to avoid domain-specific bias. Chen *et al.* [4] introduced a contrastive learning pipeline with knowledge distillation for better generalization.

In semantic segmentation, DG is frequently evaluated by a sim-to-real setup [14, 43] from synthetic datasets like SYNTHIA [47] or GTA5 [46] to real datasets like Cityscapes [7] and BDD100k [63]. Although the real-to-real setup seems to be more natural for real applications, one essential prerequisite for multi-dataset learning [30, 56] is to generate an unified label space, which was demonstrated by Mseg [27], Kim *et al.* [25] and proved to be non-trivial [2]. The lack of a unified label space also decelerates the interest of real-to-real DG evaluation, which also motivated us to propose a novel evaluation metric that can simplify the evaluation of real-to-real domain generalization evaluation.

3. CER as a Category-Aware Metric

In this section, we define a novel evaluation metric that specifically addresses inter-category errors. Current evaluation methodologies concentrate on evaluating overlapping areas between prediction and ground truth if and only if the predicted class exactly corresponds to the ground truth. Frequently, the metric *intersection over union* (IoU), also known as the Jaccard index, is used for the evaluation, indicating the portion of True Positive (TP) predictions compared to the sum of False Positives (FP) and False Negatives (FN) of one certain class c from a fixed set of predefined classes C:

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c} \tag{1}$$

After the IoU score is calculated for each individual class, an average value of the IoU scores across all possible classes is calculated to indicate the global segmentation quality, known as the mean IoU (mIoU). Similarly, IoU can be calculated in a category-wise way, however, in this setup, it is not possible to have insights to evaluate if one certain class within the category has a higher error rate in comparison to other classes in the same category when that class is insufficient represented in the dataset.

Inspired by this, we propose a novel evaluation metric that concentrates on the error rates of the predictions that are outside of the ground truth *category*, called *Critical Error Rate* (CER) with respect to class c:

$$\mathrm{CER}_c = \frac{\mathrm{FP}_{\mathrm{out},c} + \mathrm{FN}_{\mathrm{out},c}}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c} \tag{2}$$

$$\text{where } \mathrm{FP}_{\mathrm{out},c} = \sum_{c' \in C \setminus K_c} M_{c,c'} \text{ and } \mathrm{FN}_{\mathrm{out},c} = \sum_{c' \in C \setminus K_c} M_{c',c} \,,$$

given the confusion matrix M, which has a dimension of $|C| \times |C|$ (predictions × ground truth) and K_c, the set of classes in the same category as c. With CER, we swap the numerator from TP_c to two terms that represent the sum of false positive and false negative predictions that are not located in the ground truth category of the class c ($\mathrm{FP}_{\mathrm{out},c} + \mathrm{FN}_{\mathrm{out},c}$). Meanwhile, the denominators are kept identical to IoU metric to assure comparability with $\mathrm{CER_c} \in [0, 1]$ and $\mathrm{IoU_c} + \mathrm{CER_c} \leq 1, \forall c \in C$. CER facilitates class-wise out-of-category error analysis.

In the case of a novel class setup, where a class has never appeared before in the training set, the network does not give predictions for that novel class. CER can be simplified to another representation indicating the portion of the predictions from the network which are not from the correct object category with:

$$\mathrm{CER}_c^{\text{novel class}} = \frac{\mathrm{FN}_{\mathrm{out},c}}{\mathrm{FN}_c} = \frac{\sum_{c' \in C \setminus K_c} M_{c',c}}{\mathrm{FN}_c} \tag{3}$$

Qualitative comparisons utilizing CER can be found in Fig. 2 and Appendix C.

In comparison to the IoU metric, the error rate outside of the category indicates the portion of the error that is, by definition, considered to be more critical since the object category shifts. In addition to that, in the case of multi-dataset evaluation, the CER metric can be used as an indication demonstrating the ability to classify novel classes within the known class hierarchy, avoiding expensive dataset relabeling efforts. We discuss the inter-category confusion based on class taxonomies that are heavily appearance based given by the datasets in Sec. 4 and Sec. 5. For example, a *pedestrian* ↔ *rider* confusion is less critical than a confusion with a background classes such as *building*. But in certain scenarios, *e.g.*, for motion behavior prediction, the confusion *motorcycle* and *bicycle* is more critical as they represent varying motion pattern although they have similar semantic properties. Therefore, the class taxonomy that used for evaluation is not a natural constant, but rather a designed feature that should be adapted to a certain use-case. We ablate various class hierarchies in Sec. 6 as appearance hierarchy does not always represent safety aspects.

4. Experimental Setups

4.1. Datasets

Cityscapes [7] includes 5 000 fine-annotated images for semantic segmentation. The images are recorded in Europe with fair weather conditions. 19 annotated classes from 7 categories defined in the dataset are frequently used.

We use ACDC [49], BDD100k [63] and A2D2 [12] to address the problem of cross-domain adaptation and generalization since they share similar annotation strategies with Cityscapes. Those datasets include diverse driving conditions that are not included in Cityscapes including night, fog, rain, and snow around the world. We utilize the split for validation where applicable. We use the whole A2D2 dataset for evaluation in our work due to the absence of official split. For comparison, we also employ Mapillary Vistas [38] for training including 25 000 images from diverse training situations around the world. 66 classes are annotated in the dataset (v1.2) from 9 categories. To simplify the evaluation, we only utilize the top-level category and the leaf classes. The class hierarchies can be found in Appendix B.

4.2. Network Architectures and Training Setup

The objective of our experiment is to evaluate the impact of utilizing different image encoders and decoders. For backbones, we investigate three different sizes of the backbones from the classical ResNet family [16], which are well-established in previous work. From their counterparts of transformer models or heavily transformer-inspired models, we study ConvNeXt [33], Swin Transformer [32] and lately

		IoU% ↑									CER% ↓								
Backbone	Decoder	person	rider	car	truck	bus	train	m-bike	bicycle	mean	person	rider	car	truck	bus	train	m-bike	bicycle	mean
ResNet 18	FCN	80.1	57.3	93.5	46.4	72.7	46.3	53.9	76.2	70.7	16.4	22.7	4.3	22.5	8.9	25.4	23.2	20.4	16.3
	PSPNet	79.2	54.5	94.2	66.9	82.2	70.8	56.5	76.3	74.0	16.8	22.3	4.5	19.1	10.8	19.7	22.5	20.3	15.8
	ASPP	81.4	60.6	94.7	73.7	84.8	74.2	61.6	76.9	76.0	15.6	22.7	4.2	11.8	7.4	15.6	21.7	19.8	14.3
SegNeXt-T	HamHead	81.5	60.7	95	81.4	88.6	77.8	64.1	77.0	78.2	15.6	23.6	4.1	9.1	6.8	15.0	20.4	19.9	13.8
SegNeXt-S		83.1	64.1	95.4	82.5	91.4	82.4	69.0	78.7	80.0	14.4	22.0	3.8	7.8	5.8	14.0	18.7	18.6	12.9
ResNet 50	FCN	82.8	62.4	94.3	54.2	73.3	47.5	61.6	79.2	73.0	14.1	20.2	3.7	18.5	8.0	22.9	19.5	18.1	14.7
	PSPNet	82.5	61.8	95.3	75.7	87.6	80.9	66.7	79.0	77.5	14.4	20.3	3.9	14.8	8.2	14.7	19.2	18.3	13.9
	ASPP	84.0	66.0	95.7	79.9	88.9	80.6	69.3	79.5	79.3	13.6	20.0	3.5	10.8	5.7	12.6	18.3	17.8	12.7
ResNet 50	UperNet	82.6	61.2	95.3	78.7	87.7	76.8	67.8	78.9	77.9	14.1	20.8	3.8	12.5	7.6	16.1	18.2	18.3	13.5
Swin-T		82.9	62.5	95.4	80.7	88.8	79.3	68.2	78.7	79.0	14.2	21.6	3.7	9.4	6.1	13.8	18.4	18.3	12.9
ConvNeXt-T		84.0	64.7	95.8	86.0	91.8	83.5	70.3	80.3	80.8	13.2	19.9	3.5	7.0	5.3	12.2	17.0	17.0	11.9
SegNeXt-B	HamHead	84.6	67.2	95.9	88.1	93.0	86.8	72.7	80.3	81.9	13.2	20.3	3.4	5.6	4.9	10.5	16.2	17.2	11.6
ResNet 101	FCN	83.3	64.4	94.8	61.9	81.8	60.3	62.7	79.1	75.7	13.9	20.2	3.6	13.1	6.1	12.4	18.6	18.2	13.3
	PSPNet	83.6	64.5	95.6	80.9	89.6	79.5	69.3	79.6	78.8	13.7	20.2	3.7	11.3	5.9	12.9	17.8	17.8	13.0
	ASPP	84.4	67.0	95.8	79.4	89.8	82.8	70.3	80.2	79.9	13.2	19.6	3.5	9.1	5.4	12.4	17.4	17.4	12.3
ResNet 101	UperNet	83.3	63.5	95.4	77.7	88.2	79.4	67.8	79.2	78.6	13.9	21.1	3.7	11.4	7.3	15.7	18.9	18.1	13.3
Swin-S		84.2	65.0	95.9	83.4	91.6	85.8	71.1	80.1	81.0	13.2	20.1	3.4	9.6	5.2	10.9	17.2	17.3	12.0
ConvNeXt-S		84.8	66.6	95.9	84.3	92.2	86.3	72.5	81.1	81.8	12.6	18.9	3.3	5.8	5.1	11.6	16.0	16.4	11.2
SegNeXt-L	HamHead	85.2	68.1	96.1	88.5	93.4	88.4	73.6	80.8	82.5	12.6	19.5	3.3	5.5	4.7	10.0	15.6	16.8	11.2

Table 1. **Source (C)→Target (C):** Evaluation of semantic segmentation models trained on Cityscapes. We report the average class IoU and class CER over three runs.

introduced SegNeXt network [15] with corresponding network size. For image decoders, we compare the well-adapted ASPP head from DeepLabv3Plus(DLv3+) [3] and the UperNet decoder head [59], which is commonly combined with latest encoder architectures. Beside that, we ablate simple and also frequently used single and multiple scale decoder head FCN [35] and pyramid pooling module [67]. We use identical number of iterations and batch size across our experiments and datasets to assure comparability. The detailed training setup can be found in Appendix A.

5. Evaluation

In order to evaluate the ability of network generalization, we quantify the differences with 3 different setups. First, in Sec. 5.1 we only evaluate on source domain, so that we can show the inherent properties of intra- and inter-category confusion. In this setup, we train and evaluate our models separately utilizing Cityscapes or Mapillary datasets. We focus on IoU and CER for classes in the category *human* and *vehicle*. Beside that, we also calculate the mean IoU, mean CER across all the classes. Secondly, in Sec. 5.2 we evaluate the domain shift changes, where the visual appearance of the objects changes while the type of the objects remains unchanged. With that, we consider **C**ityscapes and **M**apillary as our source datasets and evaluate on the classes where they share similar class definition as **A2**D2, **AC**DC and **B**DD100k dataset. Finally, in Sec. 5.3 we consider the realistic open-world heterogeneous setup where the domain is changing and new classes are appearing at the same time. Similar to the domain shift setup, the networks are trained on both source datasets and evaluated on the **A2**D2 dataset with emphasis on those novel classes.

5.1. Source Domain

We first analyze the results from the source-to-source setup in the Cityscapes dataset. As shown in Tab. 1, there is a negative correlation between the IoU and CER, since it is trivial to derive that the absolute error that a certain network makes decreases when the segmentation performance increases. For the architectures with FCN decoder, which yield also worse segmentation performance measured by IoU due to the lack of multi-scale features, they have similar performance measured by CER comparing with other heads. This can be seen as an indication that out-of-category errors are more affected by the quality of feature representations that generated by the image encoder. On the other hand, although SegNeXt-S achieves similar performance measured by IoU to ResNet 101 backbone in combination with the ASPP decoder, SegNeXt-S model has more out of category errors: When there is a classification error, the model tends to classify the pixel as belonging to a different category, which usually poses a higher safety risk.

Tab. 2 shows the evaluation of three sets of models trained on the Mapillary dataset. We choose two mainstream architectures, DLv3+ and ConvNeXt, with the best performing model on Cityscapes for evaluation. The negative correlation between IoU and CER can still be observed. In Mapillary dataset, due to the multitude of classes that are available and

		IoU% ↑									CER% ↓								
Backbone	Decoder	person	bicyclist	o. rider	car	o. vehicle	trailer	truck	rails	mean	person	bicyclist	o. rider	car	o. vehicle	trailer	truck	rails	mean
ResNet-18	ASPP	70.2	35.6	0.0	90.3	0.9	0.0	64.3	0.0	36.9	26.0	27.3	46.8	6.8	29.2	7.5	14.9	30.7	33.4
ResNet-50		75.4	57.6	0.0	92.1	15.8	0.0	73.6	17.4	44.2	21.8	21.6	45.6	5.9	24.7	6.8	10.9	31.1	29.6
ResNet-101		76.3	57.2	0.0	92.5	21.5	0.0	73.7	39.1	45.5	21.1	21.1	44.1	5.6	24.6	7.6	11.1	24.2	29.0
ConvNeXt-T	UperNet	75.3	61.0	0.0	91.8	23.5	0.0	73.8	48.8	48.0	21.9	20.4	44.2	6.1	24.8	9.0	11.8	23.4	28.4
ConvNeXt-S		76.8	63.9	0.0	92.4	27.1	9.4	75.9	59.8	50.8	20.7	19.0	47.5	5.7	22.7	10.1	9.6	21.6	26.9
ConvNeXt-B		78.0	67.4	0.0	92.6	34.4	10.1	76.4	64.4	51.5	20.0	18.9	37.4	5.6	21.9	11.0	10.0	23.2	26.6
SegNeXt-S	HamHead	71.2	58.7	0.0	90.8	26.0	0.1	73.7	49.1	46.1	26.4	23.1	46.3	7.2	27.3	8.3	12.0	21.8	30.8
SegNeXt-B		73.6	63.0	0.0	91.6	31.6	9.6	77.2	61.8	49.5	24.2	21.5	43.6	6.8	23.6	14.3	11.0	22.5	28.2
SegNeXt-L		75.1	65.4	0.0	92.1	31.0	8.2	77.9	67.9	51.5	22.9	20.1	39.2	6.4	23.4	15.7	10.1	20.3	27.5

Table 2. **Source (M)→Target (M):** Evaluation of semantic segmentation models trained on Mapillary Vista dataset. We report the average class IoU and class CER over three runs.

the unevenly distributed class instances, *e.g.*, classes like *other rider*, *trailer* and *rails* have fewer training samples compared to others. Although the IoUs of the corresponding classes are low, they are still frequently classified into a class that is in the same class category. For instance, with the ASPP decoder, the class *trailer* has an IoU of 0 with all backbone configurations in our experiments, however, there is only a small portion of the predictions that are outside of the *vehicle* category. Similar effects can be observed with the class *trailer* and *rails*, although there is a significant improvement measured by IoU across the configurations, the CER is not negatively correlated to IoU. Despite the fact that underrepresented classes, such as *other rider*, cannot be classified correctly, the critical error rate decreases as the network's learning capacity grows. Our assumption is that the decision boundary of the class is not only shaped by the class itself but also by other classes that share common semantic or visual properties from the same object category.

Comparing the results from both datasets, it is confirmed that the IoU metric accurately reflects the performance of the networks when the instances of the classes in the dataset are well distributed and the number of available classes is limited. However, when the dataset addresses the long-tail distribution problem, CER can be utilized as a valuable supplement to evaluate the performance of the neural network.

5.2. Domain Shift

In the domain shift setup, we evaluate the generalization ability of the networks on three different target datasets: ACDC, BDD100k and A2D2. Since ACDC and BDD100k share the same label space with Cityscapes, we evaluate the networks that we trained on Cityscapes and calculate the corresponding metrics for each class. Due to the uncertainty of the network, we repeat the training three times with certain seed from ImageNet-pretrained backbones and report the standard deviation of the model performance.

Tab. 3 depicts zero-shot semantic domain generalization performance from eight architectures, including the well-adapted ResNet-based CNN, image transformer, and modern CNN models. Compared with their performance on the source domain, which can be found in Tab. 1, a significant performance drop can be observed. Beside the frequently-seen class *person*, we mainly focus on underrepresented classes in our analysis. On ACDC dataset, the state-of-the-art transformer model and modern CNN model, *e.g.*, ConvNeXt and Swin transformer, illustrate their strength at learning more generalized features by achieving better performance, as significant domain generalization differences can be observed between those models. Although ConvNext-Tiny and DLv3+ with ResNet 50 backbone achieve similar performance on source domain, the modern CNN model yields 12.2% absolute mIoU improvement on the target domain with lower standard deviation.

We observe similar results on BDD100k validation set. However, due to the extreme data distribution for the class *train* in the dataset, most of the models have poor performance on this class as judged by IoU. However, as the CER metrics indicate, a lot of predictions are still assigned to the category of *vehicle* instead of other categories. Similar effects can be seen in the class of *truck*: Although the IoU remains low, there are significantly more predictions in the category of *vehicle* compared to ACDC dataset, where the false predictions are to a large extent in other object categories. This may also due to the different appearance of certain classes and deviated labeling strategy. In addition to that, we notice the modern architectures with naive training organization have also achieved results that are similar to dedicated domain generalization methods reported by [6].

Since the A2D2 dataset does not share the same label space with other datasets, we choose a subset of classes that can be remapped to the Cityscapes evaluation scheme, as we evaluate the novel classes in the following subsection, Sec. 5.3. We consider the class *car* and *truck* from the A2D2 dataset in this subsection. The models that are used for evaluation are trained on Cityscapes or Mapillary. For a fair comparison they are trained with identical train-

		ACDC					BDD100k				
	Architecture	person	truck	bus	train	mean	person	truck	bus	train	mean
IoU%	DLv3+ R50	38.1 ± 4.4	4.9 ± 1.5	31.4 ± 11.9	30.0 ± 8.4	37.6 ± 1.7	45.2 ± 21.3	21.8 ± 2.9	15.8 ± 4.7	0.2 ± 0.2	42.9 ± 0.9
	UperNet R50	43.2 ± 1.8	2.5 ± 2.1	40.3 ± 7.7	32.2 ± 1.8	35.1 ± 0.2	56.3 ± 1.6	16.4 ± 2.0	15.7 ± 6.6	0.0 ± 0.0	42.1 ± 1.2
	DLv3+ R101	47.7 ± 5.7	11.6 ± 9.6	38.5 ± 25.6	28.9 ± 6.5	40.3 ± 2.9	59.1 ± 0.5	28.2 ± 1.3	24.5 ± 6.4	0.2 ± 0.2	45.9 ± 0.9
	UperNet R101	46.4 ± 0.2	2.3 ± 2.1	33.6 ± 4.0	24.8 ± 15.3	35.8 ± 2.8	57.4 ± 1.0	22.6 ± 1.0	22.1 ± 2.7	0.1 ± 0.1	45.2 ± 0.7
	UperNet Swin-T	39.6 ± 3.9	18.5 ± 5.2	44.4 ± 1.7	36.7 ± 3.7	39.4 ± 0.9	54.8 ± 2.2	26.7 ± 1.6	31.8 ± 5.1	0.0 ± 0.0	43.7 ± 0.8
	UperNet Conv-T	52.1 ± 0.9	41.4 ± 4.0	44.0 ± 2.1	57.7 ± 1.7	49.8 ± 0.4	63.1 ± 0.4	38.6 ± 0.6	31.0 ± 3.4	0.1 ± 0.1	51.2 ± 0.3
	UperNet Swin-S	45.6 ± 2.8	41.5 ± 7.2	60.9 ± 8.2	40.5 ± 1.9	47.1 ± 1.3	59.9 ± 0.8	30.5 ± 2.4	34.5 ± 9.7	0.2 ± 0.2	48.0 ± 0.9
	UperNet Conv-S	60.3 ± 0.5	52.1 ± 11.0	52.3 ± 3.2	57.3 ± 5.5	54.5 ± 1.1	66.7 ± 0.4	39.6 ± 1.0	38.6 ± 3.6	0.4 ± 0.1	55.4 ± 0.3
CER%	DLv3+ R50	60.4 ± 5.3	90.8 ± 2.0	49.3 ± 12.4	48.8 ± 11.8	47.4 ± 1.2	53.6 ± 22.1	46.7 ± 8.7	22.2 ± 5.9	30.7 ± 13.7	36.2 ± 1.0
	UperNet R50	56.1 ± 1.7	93.6 ± 4.7	42.1 ± 12.3	33.6 ± 14.6	49.3 ± 0.6	43.2 ± 1.6	50.9 ± 3.2	23.0 ± 6.4	14.8 ± 2.4	34.9 ± 1.3
	DLv3+ R101	51.1 ± 6.0	79.1 ± 14.7	46.3 ± 31.1	43.9 ± 3.4	44.9 ± 2.4	40.3 ± 0.5	32.9 ± 3.4	25.1 ± 12.4	18.9 ± 8.1	33.0 ± 0.6
	UperNet R101	52.9 ± 0.2	94.0 ± 4.8	58.1 ± 3.9	51.9 ± 32.1	50.4 ± 3.7	41.9 ± 1.2	42.5 ± 2.4	28.2 ± 5.6	27.4 ± 13.4	34.2 ± 1.5
	UperNet Swin-T	58.7 ± 4.3	73.4 ± 6.6	41.4 ± 5.1	47.1 ± 4.6	45.6 ± 0.5	41.0 ± 0.5	40.4 ± 3.2	16.4 ± 3.0	27.3 ± 6.4	32.4 ± 1.3
	UperNet Conv-T	45.8 ± 1.4	43.6 ± 11.2	47.3 ± 5.2	34.1 ± 2.8	36.8 ± 1.0	35.4 ± 0.3	22.0 ± 0.3	10.3 ± 0.5	19.2 ± 4.8	27.1 ± 0.4
	UperNet Swin-S	53.1 ± 2.1	45.2 ± 4.7	30.8 ± 10.3	49.7 ± 4.5	38.6 ± 1.1	38.9 ± 0.3	30.1 ± 5.9	14.2 ± 6.5	25.5 ± 13.3	30.0 ± 1.3
	UperNet Conv-S	38.5 ± 0.4	32.0 ± 11.0	30.7 ± 3.9	33.8 ± 6.9	31.5 ± 0.9	32.1 ± 0.3	22.1 ± 2.3	7.9 ± 1.6	11.9 ± 1.7	23.5 ± 0.2

Table 3. **Source (C)→Target (AC, B):** Zero-shot evaluation of models trained on Cityscapes, evaluated on ACDC and BDD100k. We apply color gradient visualizing IoU shifts within each class across the two datasets. We report the values over three runs.

ing setup, i.e., the same number of iterations with the same batch size. Tab. 4 shows the evaluation of the generalization ability of networks trained on both datasets. It can be observed that the networks with the ResNet-18 image encoder trained on Mapillary dataset outperform models that have three times more parameters (ResNet-101) that were trained on Cityscapes before. We notice similar effects with the ConvNeXt-Tiny backbone, where models trained on Mapillary achieve a lower error rate, lower deviations, and higher IoUs than the more sophisticated ConvNeXt-Base model trained on Cityscapes, which corresponds to the recent work by Piva *et al*. [44]. Beside that, the better generalization ability of the modern CNN models is affirmed again [1, 57], as ConvNeXt-based models surpass the classical ResNet-based models regardless of the dataset that used for training.

		Car		Truck	
Setup		IoU%	CER%	IoU%	CER%
Cityscapes	R18	88.8 ± 1.3	5.0 ± 0.1	49.8 ± 8.1	24.1 ± 7.0
	R50	89.0 ± 1.1	5.7 ± 0.7	54.2 ± 9.9	24.5 ± 6.0
	R101	88.8 ± 2.0	5.0 ± 1.3	53.1 ± 9.6	17.4 ± 4.1
	Conv-T	91.8 ± 0.4	4.0 ± 0.2	62.5 ± 1.0	14.9 ± 1.1
	Conv-S	92.5 ± 0.4	3.9 ± 0.3	69.3 ± 0.3	9.9 ± 0.7
	Conv-B	93.6 ± 0.1	3.6 ± 0.2	69.4 ± 1.8	10.6 ± 1.8
Mapillary	R18	92.5 ± 0.3	4.1 ± 0.2	62.8 ± 2.0	14.0 ± 2.8
	R50	93.9 ± 0.2	3.5 ± 0.1	72.0 ± 1.7	7.3 ± 0.8
	R101	94.3 ± 0.2	3.4 ± 0.2	73.7 ± 0.3	6.2 ± 0.6
	Conv-T	94.2 ± 0.0	3.4 ± 0.1	74.0 ± 0.3	5.3 ± 0.3
	Conv-S	94.7 ± 0.1	3.3 ± 0.1	76.0 ± 0.3	5.0 ± 0.6
	Conv-B	94.6 ± 0.1	3.2 ± 0.1	75.1 ± 0.2	4.6 ± 0.3

Table 4. **Source (C/M)→Target (A2):** Evaluation of models trained on Cityscapes and Mapillary, evaluated on two classes from A2D2 dataset that share the same label space. We report the average IoU and CER over three runs.

5.3. Novel Class

We further extend our experiments to evaluate on previously unknown classes. The motivation is to assess whether the model tends to infer the correct class category, when an unknown object appears which is previously not included in the dataset but shares the same object category. We perform zero-shot evaluation on the A2D2 dataset and Mapillary dataset for the novel class setup with the neural networks previously trained on Cityscapes or Mapillary dataset. We report the CER of the classes *utility vehicles*, *tractor* from A2D2 dataset as well as the classes *other rider*, *caravan*, *other vehicle* and *wheeled slow* from Mapillary. All of the previously introduced classes are part of the category *vehicle* in A2D2 and Mapillary except *other rider* which is in the category of *human*. For reference, we note their mean IoUs as performance indications on the source domain. We report the average CER of the corresponding classes and the standard deviation to demonstrate the fluctuation of various seeds for network training. Utilizing the CER metric avoids potential ambiguity, which may cause unfair scores in the evaluation caused by varying labeling policies across different datasets. In order to evaluate the differences of the models trained with varying datasets, we also evaluate networks that trained with Mapillary dataset on A2D2.

The comparison can be seen in Tab. 5. Similar to the domain shift setup, although some of the models have similar performance on the source domain, they showed different levels of generalization. From a dataset point of view, a large, diverse dataset like Mapillary still shows impressive generalization ability, as a DLv3+ model with ResNet-50 backbone achieves similar performance in the class *utility vehicle* as its ConvNeXt counterparts, although there is still disparity in the *tractor* class. Beside that, the modern CNN models and transformer models outperform ResNet-based architectures with a lower CER and simultaneously lower performance deviation. With increasing learning capacity, similar to the observations in the last subsections, the networks can implicitly differentiate objects from various class categories, even if they are not explicitly given as learning objectives during training, as the sophisticated backbones

			Source	A2D2 CER% ↓		Mapillary CER% ↓			
Backbone	Decoder	Dataset	mIoU%	Util. Vehicles	Tractor	Other Rider	Caravan	Other Vehicle	Wheeled Slow
ResNet 18	FCN	CS	70.7	62.3 ± 3.6	63.7 ± 12.0	30.6 ± 2.9	38.8 ± 8.9	35.2 ± 2.2	63.1 ± 2.6
	ASPP	CS	76.0	56.7 ± 5.6	56.7 ± 8.7	33.2 ± 2.3	20.4 ± 7.4	31.5 ± 2.8	66.5 ± 5.6
		MV	36.9	48.4 ± 3.5	18.7 ± 6.6	-	-	-	-
ResNet 50	FCN	CS	73.0	71.7 ± 3.1	83.8 ± 5.1	32.3 ± 0.9	40.0 ± 5.3	36.4 ± 2.7	66.2 ± 0.7
	ASPP	CS	79.3	53.1 ± 3.2	76.7 ± 2.3	34.5 ± 3.2	34.5 ± 30.2	31.6 ± 2.3	61.2 ± 1.7
		MV	44.2	35.8 ± 2.1	9.5 ± 0.0	-	-	-	-
ResNet 50	UperNet	CS	77.9	57.3 ± 4.1	59.0 ± 19.2	44.8 ± 3.5	52.0 ± 10.1	32.1 ± 1.5	67.1 ± 4.7
Swin-T		CS	79.0	51.6 ± 9.8	73.5 ± 6.9	29.9 ± 0.5	70.6 ± 16.5	28.3 ± 3.6	63.7 ± 2.0
ConvNeXt-T		CS	80.8	35.6 ± 6.1	45.3 ± 4.2	23.0 ± 3.5	59.1 ± 13.3	28.7 ± 1.5	65.2 ± 1.6
		MV	48.0	24.8 ± 2.1	7.4 ± 0.3	-	-	-	-
SegNeXt-B	HamHead	CS	81.9	31.1 ± 1.7	28.4 ± 3.5	30.5 ± 3.5	27.0 ± 14.0	34.3 ± 1.3	64.8 ± 2.5
ResNet 101	FCN	CS	75.7	55.3 ± 4.1	65.5 ± 11.5	27.9 ± 2.0	30.9 ± 23.2	32.3 ± 1.8	67.1 ± 4.9
	ASPP	CS	79.9	46.3 ± 2.7	58.4 ± 18.2	31.4 ± 7.6	32.7 ± 34.1	33.6 ± 1.2	68.2 ± 5.4
ResNet 101	UperNet	CS	78.6	52.5 ± 17.1	57.5 ± 28.2	32.7 ± 5.2	54.7 ± 30.0	35.4 ± 6.1	73.4 ± 3.0
Swin-S		CS	81.0	47.4 ± 4.5	59.4 ± 3.3	21.7 ± 2.6	78.5 ± 0.5	29.7 ± 3.5	67.1 ± 2.7
ConvNeXt-S		CS	81.8	27.8 ± 3.1	25.0 ± 4.0	25.7 ± 1.3	45.0 ± 17.3	27.9 ± 0.8	70.8 ± 1.7
		MV	50.8	20.7 ± 3.3	6.8 ± 0.1	-	-	-	-
SegNeXt-L	HamHead	CS	82.5	33.5 ± 4.1	35.3 ± 7.0	29.8 ± 6.5	31.5 ± 34.4	30.5 ± 0.9	65.4 ± 1.9

Table 5. **Source (C/M)→Target (A2/M):** Zero-shot evaluation on six novel classes in an open-world semantic segmentation setup, we report class CER values and their standard deviations over three runs. As reference, we additionally provide their mIoU in source domain.

tend to classify those previously unknown objects into their superclasses. Multi-scale features prove to be beneficial for the generalization, while the feature pyramid network from UperNet in combination with ResNet may cause higher variation in the performance when the size of the model grows as the first layers of the encoder are trained to adapt the source domain. We observe that this property has less impact on the modern backbone. We assume that the introduction of new types of normalization layers may alleviate the negative effects. ConvNeXt and Swin Transformer show deviating behavior when encountering new objects. Swin models have a significantly higher error rate in the near *car* and *truck* classes compared with their ConvNeXt counterparts, although the two neural networks are trained with the same training configuration, *e.g.*, loss function and training scheduler. The new SegNext family indeed achieves similar generalization ability to ConvNeXt models, but the state-of-the-art segmentation performance cannot be directly transferred to a random, previously unknown domain. Combined with the results from Sec. 5.2, we suggest that CER can also be used as a quality indication for the datasets reflecting data variety or similarity. In Fig. 2, we show some qualitative results from A2D2 and Mapillary datasets, where there are previously unknown objects according to the training classes in the images. We also note the CER scores of the novel class in the images.

6. Ablation Study

Current class hierarchies are most commonly based on semantic differences between the classes. However, to address safety potentials, we propose two different concepts of class hierarchies as an ablation study to investigate the impact of utilizing different taxonomies: According to their possible motion behavior in the near future (as the velocity of *rider* may differ from that of a *car*) or, on the other hand, such that dynamic objects are distinguished into whether the class is considered a vulnerable road user (VRU), or if they have notable impact protection when a crash happens. To simplify our experiment, we only consider the classes that are under category *human* and *vehicle*. We ablate in this section if a flexible class taxonomy can provide more insight for the evaluation of the networks. We consider the following two alternative class taxonomies in the range of classes previously annotated as *vehicle* and *human*:

Behavior-based class taxonomy: We first consider a behavior-based class hierarchy, which depends on the possible moving speed of the class, to evaluate the use of segmentation results as input for behavior predictions. We divide the classes into two categories: one with low velocity in the scenes, like *person* and *wheeled slow*, while motorized objects and their riders are treated as objects that are considered to have more displacement in a certain time period.

VRU-based class taxonomy: We further divide the classes into two categories: VRU classes for objects that have little protection against crash forces, and non-VRU classes. We study the class ***M****otorcyclist*, *other* ***R****ider* and ***W****heeled slow* under various class taxonomies in Tab. 6. From the class *motorcyclist*, we observe a significant difference in reported CER when the class taxonomy changes. The evaluation under the VRU scheme indicates that the model generally has a lower amount of confusion due to the fact that we consider motorcycle and motorcyclist both as part of the VRU category, which can also be discovered in the other two evaluated classes. This is an indication that there is frequent confusion between various VRU classes due to their appearance. The motion behavior scheme presented by CER poses a challenge for the trained neural network. The reason for that is

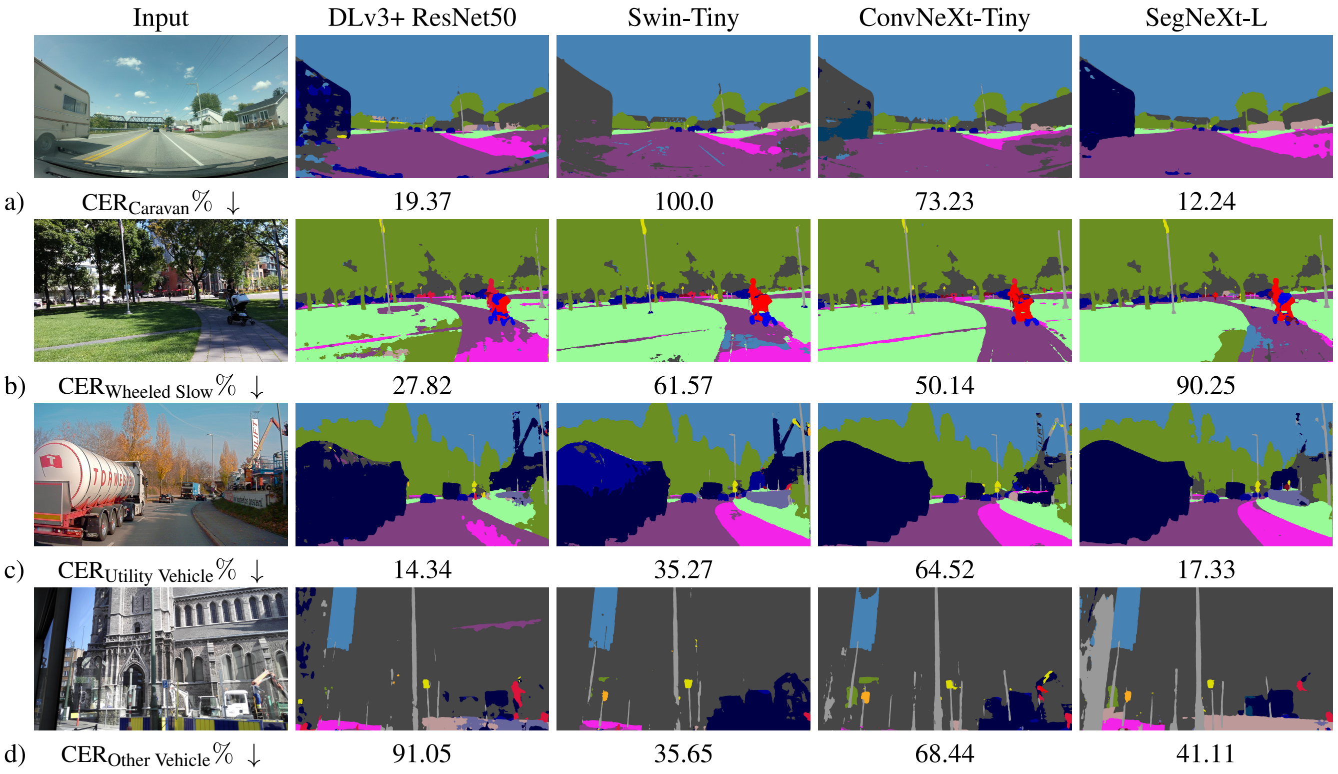
Input
DLv3+ ResNet50
Swin-Tiny
ConvNeXt-Tiny
SegNeXt-L
a) $\mathrm{CER}_{\mathrm{Caravan}}\% \downarrow$ 19.37 100.0 73.23 12.24
b) $\mathrm{CER}_{\mathrm{Wheeled\ Slow}}\% \downarrow$ 27.82 61.57 50.14 90.25
c) $\mathrm{CER}_{\mathrm{Utility\ Vehicle}}\% \downarrow$ 14.34 35.27 64.52 17.33
d) $\mathrm{CER}_{\mathrm{Other\ Vehicle}}\% \downarrow$ 91.05 35.65 68.44 41.11

Figure 2. Qualitative results on A2D2 dataset and Mapillary dataset. The neural networks are trained on Cityscapes dataset and evaluate in a zero-shot fashion facing previously unknown objects like a) caravan, b) stroller, c) and d) construction machines. Best viewed in color.

that visual differences between the classes are dominantly utilized in optimization during training. During evaluation, the metric treats visually similar classes into separate diverse semantic categories, *e.g.*, when *bicyclist* and *motorcyclist* are not considered in the same object category due to their motion difference, which is not included in the learning objectives and thus needs to be sufficiently addressed during training in order to achieve promising performance. Detailed class taxonomies can be found in Appendix B.

	CER% ↓			$\mathrm{CER_B}$% ↓			$\mathrm{CER_V}$% ↓		
Setup	**M**	**R**	**W**	**M**	**R**	**W**	**M**	**R**	**W**
R18	20.9	33.2	66.5	86.2	32.9	48.1	11.4	24.6	34.4
R50	12.4	34.5	61.2	92.8	34.2	56.2	7.7	27.4	39.2
R101	13.4	31.4	68.2	92.3	30.3	49.8	8.9	26.3	39.2
C-T	11.7	23.0	65.2	93.3	21.3	44.6	7.9	19.3	36.2
C-S	11.5	25.7	70.8	92.6	25.3	45.0	7.5	24.3	37.7
C-B	11.6	21.0	72.3	92.9	20.0	42.2	8.5	18.9	33.8

Table 6. Ablation study on different class taxonomies, we show the CER with class hierarchy from Mapillary dataset, its variant $\mathrm{CER_B}$ based on behaviour of the classes and $\mathrm{CER_V}$ according to VRU property. We report CER values over three runs.

7. Conclusion and Outlook

In our work, we propose a novel evaluation metric for semantic segmentation that specifically addresses the distinction between intra- and inter-category errors. We first evaluate the out-of-category error behavior of current state-of-the-art models and classical CNN models without domain shifts, revealing the fact that the optimization process of underrepresented classes is also driven by appearance-similar classes. In a domain-shift setup, we confirmed the good generalization ability of modern architectures like ConvNeXt. Beside that, we showed the importance of data variety, which affects the performance of the neural networks more than learning capacity. We also evaluate the models' behavior when a novel object appears and show the differences between the models. Last but not least, we discuss the impact of utilizing class taxonomies that are independent from appearance for the category-related semantic segmentation evaluation. Future work could additionally conduct more exhaustive evaluations, considering other application domains and a wider variety of class hierarchy principles.

ACKNOWLEDGMENT

The research leading to these results is supported by the German Federal Ministry for Economic Affairs and Climate Action within the projects "KI Delta Learning" (www.ki-deltalearning.de) and "AVEAS" (www.aveas.org).

References

[1] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843, 2021. 2, 6

[2] Petra Bevandić, Marin Oršić, Ivan Grubišić, Josip Šarić, and Siniša Šegvić. Multi-domain semantic segmentation with overlapping labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2615–2624, 2022. 2

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 4

[4] Wuyang Chen, Zhiding Yu, Zhangyang Wang, and Animashree Anandkumar. Automated synthetic-to-real generalization. In *International Conference on Machine Learning*, pages 1746–1756. PMLR, 2020. 2

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[6] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 2, 5

[7] 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(CVPR)*, 2016. 1, 2, 3

[8] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16918–16927, 2021. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[10] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? *ECCV*, 2022. 1

[11] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *NeurIPS*, 31, 2018. 2

[12] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 1, 3

[13] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015. 2

[14] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2477–2486, 2019. 2

[15] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv preprint arXiv:2209.08575*, 2022. 1, 4

[16] 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. 1, 3

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 2

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 2

[19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1

[20] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 2

[21] ISO. ISO 26262: Road vehicles—Functional Safety, 2018. 1

[22] ISO. ISO 21448: R.V.—Safety of the Intended Functionality, 2022. 1

[23] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, 2021. 1

[24] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *CVPR*, 2020. 2

[25] Dongwan Kim, Yi-Hsuan Tsai, Yumin Suh, Masoud Faraki, Sparsh Garg, Manmohan Chandraker, and Bohyung Han. Learning semantic segmentation from multiple datasets with label shifts. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 20–36. Springer, 2022. 2

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[27] 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. 2

[28] Aodi Li, Liansheng Zhuang, Shuo Fan, and Shafei Wang. Learning common and specific visual prompts for domain generalization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 4260–4275, December 2022. 2

[29] Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1246–1257, 2022. 2

[30] Xiaodan Liang, Hongfei Zhou, and Eric Xing. Dynamic-structured semantic propagation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 752–761, 2018. 2

[31] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 1

[32] 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. 1, 3

[33] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1, 3

[34] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 1

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4

[36] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. Motion inspired unsupervised perception and prediction in autonomous driving. In *ECCV*. Springer, 2019. 1

[37] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021. 2

[38] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 3

[39] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015. 2

[40] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 2

[41] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1863–1871, 2019. 2

[42] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022. 2

[43] Duo Peng, Yinjie Lei, Lingqiao Liu, Pingping Zhang, and Jun Liu. Global and local texture randomization for synthetic-to-real semantic segmentation. *IEEE Transactions on Image Processing*, 30:6594–6608, 2021. 2

[44] Fabrizio J. Piva, Daan de Geus, and Gijs Dubbelman. Empirical generalization study: Unsupervised domain adaptation vs. domain generalization methods for semantic segmentation in the wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 499–508, January 2023. 2, 6

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[46] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 2

[47] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016. 2

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2

[49] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 3

[50] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9624–9633, 2021. 2

[51] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Processings of International Conference on Learning Representations (ICLR)*, 2014. 2

[52] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robust-

ness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:18583–18599, 2020. 2

[53] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 246–263. Springer, 2022. 1

[54] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6924–6932, 2017. 2

[55] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 2

[56] Li Wang, Dong Li, Han Liu, Jinzhang Peng, Lu Tian, and Yi Shan. Cross-dataset collaborative learning for semantic segmentation in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2487–2494, 2022. 2

[57] Zeyu Wang, Yutong Bai, Yuyin Zhou, and Cihang Xie. Can cnns be more robust than transformers? In *International Conference on Learning Representations*, 2023. 2, 6

[58] Junfei Xiao, Zhichao Xu, Shiyi Lan, Zhiding Yu, Alan Yuille, and Anima Anandkumar. 1st place solution of the robust vision challenge (rvc) 2022 semantic segmentation track. *arXiv preprint arXiv:2210.12852*, 2022. 2

[59] 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. 4

[60] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1

[61] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020. 2

[62] Suorong Yang, Weikang Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*, 2022. 2

[63] 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. 2, 3

[64] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2

[65] Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7277–7286, 2022. 2

[66] Yuxiang Zhang, Sachin Mehta, and Anat Caspi. Rethinking semantic segmentation evaluation for explainability and model selection. *arXiv preprint arXiv:2101.08418*, 2021. 1

[67] 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

[68] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *CVPR*, 2021. 1

[69] Jingxing Zhou and Jürgen Beyerer. Impacts of data anonymization on semantic segmentation. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 997–1004, 2022. 2

[70] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. 2