Are Synthetic Corruptions A Reliable Proxy For Real-World Corruptions?

Anonymous CVPR submission

Paper ID 0005

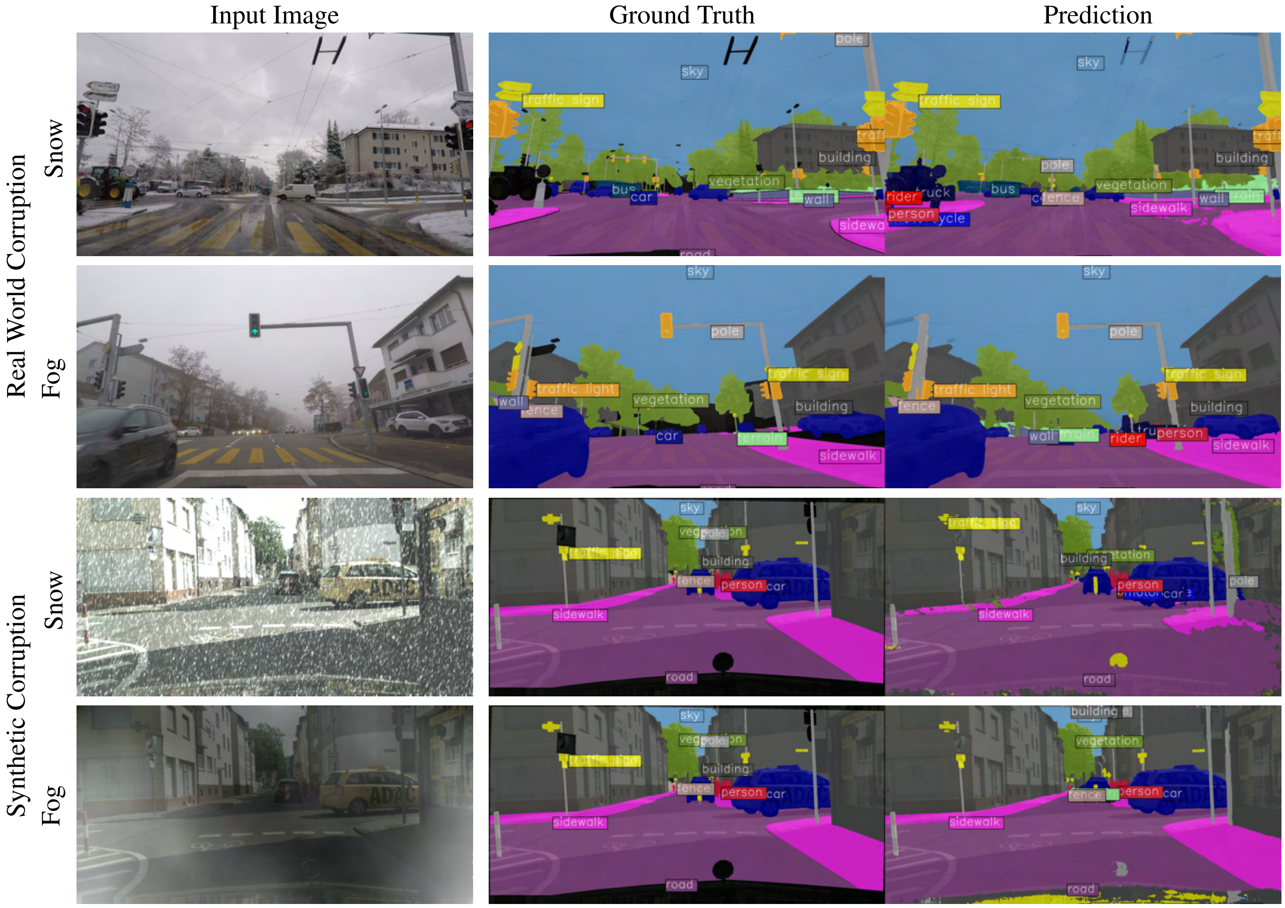

Figure 1. Comparing images with weather corruptions captured in the wild (ACDC [31]) and images corrupted using synthetic corruptions [19] and the predictions using a Mask2Former [7] with a Swin-Base [26] backbone trained on the Cityscapes [9] dataset.

Abstract

Deep learning (DL) models are widely used in real-world applications but remain vulnerable to distribution shifts, especially due to weather and lighting changes. Collecting diverse real-world data for testing the robustness of DL models is resource-intensive, making synthetic corruptions an attractive alternative for robustness testing. However, are synthetic corruptions a reliable proxy for real-world corruptions? To answer this, we conduct the largest benchmarking study on semantic segmentation models, comparing performance on real-world corruptions and synthetic corruptions datasets. Our results reveal a strong correlation in mean performance, supporting the use of synthetic corruptions for robustness evaluation. We further analyze corruption-specific correlations, providing key insights to understand when synthetic corruptions succeed in representing real-world corruptions. The code and datasets will be released upon acceptance.

1. Introduction

Although very successful in benchmark scenarios, the reliability of deep-learning (DL)-based models for semantic segmentation in real-world scenarios remains a major concern. Potentially unseen variations in the data (a.k.a. distribution shifts), for example, due to changes in weather conditions (e.g., fog, rain, snow) and lighting (e.g., nighttime, glare), can heavily degrade model performance. Ensuring robustness to such shifts is critical for safe and reliable deployment, particularly in applications like autonomous driving [9, 27] or medical imaging [11, 30]. To evaluate model robustness, researchers often rely on synthetic corruptions, such as [19]. These perturbations — designed to mimic real-world conditions — offer a scalable and controlled way to assess model performance without the cost of real-world data collection.

Several previous works [4, 23, 31] have also attempted to draw focus towards threats posed in real-world applications when facing slight domain shifts, for example, through noise or simply through changing weather. Specific evaluations involve the study of Out-Of-Distribution (OOD) samples to mimic realistic domain shifts.

Despite their widespread use, the correlation between model performance on synthetic and real-world corruptions is not well understood. Figure 1 shows one such scenario with real-world corruptions (Snow and Fog) captured in the ACDC dataset [31] and similar synthetic corruptions added on in-domain images from the cityscapes validation dataset. We observe very similar trends in the lack of robustness of the model towards both real-world and synthetic corruptions. However, a fundamental question remains:

> “Are synthetic corruptions a reliable proxy for real-world corruptions?”

If a strong correlation exists, synthetic corruptions could serve as a cost-effective alternative for robustness evaluation. Conversely, if the correlation is weak, extensive tests on real-world settings remain necessary at all stages.

Here, we conduct a large benchmarking study, analyzing the correlation between model performance on real-world and synthetic corruptions for semantic segmentation. The main contributions of this work are as follows:

- We benchmark multiple DL-based semantic segmentation models on real-world corruptions from the ACDC dataset and synthetic corruptions from Cityscapes + 2D Common Corruptions.
- We provide an in-depth analysis of corruption-specific trends, identifying cases where synthetic corruptions succeed or fail as proxies.
- We provide benchmarking of semantic segmentation methods against synthetic corruptions on ADE20k [37] and PASCAL VOC 2012 [13] datasets.

Our findings reveal a high correlation in mean performance, suggesting that synthetic corruptions can indeed serve as a reliable proxy for real-world robustness evaluation. However, we also highlight key cases where synthetic corruptions fail to fully capture real-world effects, underscoring the need for more nuanced evaluation methods.

2. Related Work

The robustness of DL-based methods to distribution shifts is often used as a measure of their generalization ability [20, 21]. Common Corruptions [19] and 3D Common Corruptions [24] are tools proposed for benchmarking the robustness of image classification models, but they can be extended to other vision tasks as for example done in [23]. However, both are synthetic corruptions, and distribution shifts occurring in the real world might be slightly different. Conversely, Sakaridis et al. [31] proposed “ACDC: The Adverse Conditions Dataset with Correspondences for Robust Semantic Driving Scene Perception”. This dataset contains images captured in the wild in different conditions, such as during Night, Rain, Snow, and Fog. While ACDC does not cover many other possible conditions that can cause distribution shifts, it serves as a community-accepted tool for benchmarking real-world OOD robustness to a certain extent.

In this work, we use both Common Corruptions and ACDC to benchmark OOD robustness and thus measure the generalization ability of various semantic segmentation methods, including recently proposed SotA methods like Mask2Former [7] and InternImage [32], with the goal to investigate whether synthetic datasets that are easy to generate can serve as a proxy for a model’s real world OOD robustness.

[4] provides a new benchmark for robustness against anomalies, while relevant for real-world applications, we intend to focus this work on traditional OOD robustness.

In their work, Michaelis et al. [28] proposed datasets combining 2D Common Corruptions with datasets such as MS-COCO [25], PASCAL VOC 2007 [12], and Cityscapes. However, their evaluations were limited to 2D Common Corruptions and how different severities of the corruptions on the images impact the downstream task performance. We find correlations between performance against 2D Common Corruptions and real-world corruptions. We use their proposed Cityscapes-C (Cityscapes + 2D Common Corruptions) as our synthetic corruptions dataset.

3. Metrics For Analysis At Scale

This is the first work to analyze semantic segmentation methods, especially under the lens of reliability and generalization ability on such a large scale. The most commonly used metrics for reporting evaluations on semantic segmentation are mean Intersection over Union (mIoU), mean class Accuracy (mAcc), and mean Accuracy of all pixels (aAcc) [1, 2, 36]. We capture these metrics while evaluating models against both ACDC and the 15 2D Common Corruptions on the Cityscapes validation dataset. As per the commonly accepted practice of such OOD evaluations, all models are pre-trained on the Cityscapes training dataset.

Similar to [28], the 15 2D Common Corruptions [19] considered in this work are: 'gaussian noise', 'shot noise', 'impulse noise', 'defocus blur', 'frosted glass blur', 'motion blur', 'zoom blur', 'snow', 'frost', 'fog', 'brightness', 'contrast', 'elastic', 'pixelate', 'jpeg'. Similar to [19], Michaelis et al. [28] shows that synthetic corruptions with corruption severity=1 are too weak, and corruptions with corruption severity=5 are too strong for the downstream task. Thus, we use corruption severity=3 in our evaluations.

As discussed, multiple image classification works [10, 20, 21] and some semantic segmentation works [23, 28] use OOD Robustness of models for evaluating the generalization ability of the method. However, different image corruptions impact the performance of the semantic segmentation methods differently. As we are interested in the worst possible case, we define $\mathrm{Generalization\ Ability\ Measure}$ (GAM) as the worst mIoU across all image corruptions at a given severity level. That is, we ask the question "For a given dataset, what is the worst possible performance of a given method?". Answering this question tells us about the reliability and generalization ability of a method. We find the minimum of the mIoU of the segmentation masks predicted under image corruptions w.r.t. the ground truth masks for a given method, across all corruptions at a given severity and report this as the $\mathrm{GAM}_{severity\ level}$. For example, for severity=3, the measure would be denoted by GAM_3. The higher the GAM value, the better the generalization ability of the given semantic segmentation method. In Appendix A, we show that our observations are not limited to the mIoU metric and extend to other metrics as well.

4. Analysis And Key Findings

We analyze the correlation in mean performance to determine whether synthetic corruptions can serve as a reliable proxy for real-world corruptions. Additionally, we conduct an in-depth examination of corruption-specific trends, identifying cases where synthetic corruptions effectively mimic real-world effects and where they fall short.

4.1. Are Synthetic Corruptions Useful?

We attempt to study if synthetic corruption like that introduced by [19] does represent the distribution shifts in the real world. While this assumption has driven works such as [19, 23, 24], to the best of our knowledge, it has not yet been proven. Previous works on robustness [15] simply report performance on both, thus, to save compute in the future, we prove this assumption in Fig. 2.

For this analysis, we used methods trained on the training set of Cityscapes and evaluated them on 2D Common Corruptions [19] and the ACDC datasets. ACDC is the Adverse Conditions Dataset with Correspondences, consisting of images from similar regions and scenes as Cityscapes but captured under different conditions such as Day/Night, Fog, Rain, and Snow. These are corruptions in the real world, thus, we attempt to find correlations between performance against synthetic corruptions from 2D Common Corruptions (severity=3) and ACDC. We analyze each common corruption separately and also the mean performance across all 2D Common Corruptions.

In Fig. 2, we observe a very strong positive correlation in performance against ACDC and mean performance across all 2D Common Corruptions. This novel finding helps the community significantly. It means that we do not need to go into the wild to capture images with distribution shifts, as synthetic corruptions serve as a reliable proxy for real-world conditions. Next, we look at the correlation between the worst-case scenario measure using GAM_3 and ACDC. Here, we observe a higher correlation than the previous case, indicating that the performance against worst-case corruption serves as a reliable proxy for real-world corruptions. Lastly, as a sanity check, we find the correlation between mean performance against all corruptions and performance against worse-case corruption to observe a very high correlation. Showing that the two can be used interchangeably.

4.2. When Do Synthetic Corruptions Succeed?

Since some synthetic corruptions attempt to directly mimic the real-world scenarios in ACDC, like changes in lighting due to Day/Night changes or changes in weather due to snowfall or fog, we analyze the correlation of relevant corruptions to ACDC. As discussed in Sec. 4.1, the mean performance correlation is high. However, we observe in Fig. 3 that individual corruptions exhibit varying levels of agreement between synthetic and real-world effects. We observe that the Snow corruption shows a very strong alignment (Pearson correlation 0.867), indicating that synthetic snow corruptions effectively mimic real-world snow-related degradation, despite the corrupted images looking different to a human observer (as shown in Fig. 1).

Brightness (Pearson correlation 0.270) and Fog (Pearson correlation 0.349) exhibit weak alignment, suggesting that synthetic versions of these corruptions fail to fully capture real-world complexities. Specifically, brightness corruptions struggle to model real-world nighttime conditions, while synthetic fog does not accurately represent atmospheric distortions seen in real-world data.

These findings highlight that while synthetic corruptions

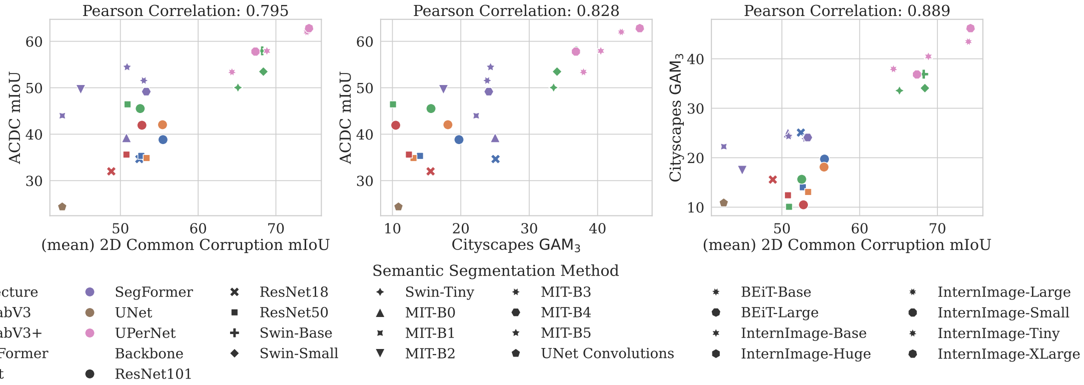

Figure 2. To empirically determine if synthetic common corruptions such as those proposed by [19] truly represent the distribution and domain shifts in the real world, we try to find correlations in evaluations on ACDC and 2D Common Corruptions. Each model is trained on the training dataset of the Cityscapes dataset. Left plot: The y-axis represents values from evaluations on the ACDC dataset, and the x-axis represents mean performance from evaluations on the Common Corruptions at severity=3. We observe a high positive correlation. Centre plot: The y-axis again represents values from evaluations on the ACDC dataset, while the x-axis represents GAM_3, which is the worst performance of the methods across all the Common Corruptions at severity=3. We observe a slightly higher positive correlation. Right plot: serves as a sanity check, where the y-axis represents GAM_3 and the x-axis represents mean performance from evaluations on the Common Corruptions at the same severity. We observe a very high correlation in performance. Thus, given the high positive correlations between performance on the ACDC and mean performance against all synthetic common corruption, we conclude for relative analysis that synthetic corruptions do serve as a reliable proxy for real-world corruptions.

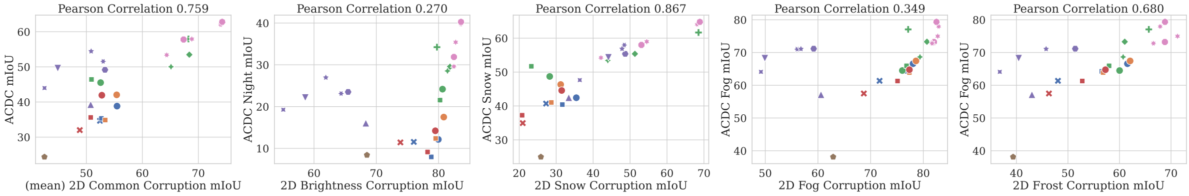

Figure 3. Correlation between model performance (legend as in Fig. 2) on ACDC (real-world corruptions) and 2D Common Corruptions (synthetic) for different corruption types. The left-most plot shows the correlation between mean mIoU across all 2D Common Corruptions and ACDC, with a strong Pearson correlation of 0.759, indicating that synthetic corruptions are generally a reasonable proxy for real-world robustness. The remaining plots analyze specific corruptions: brightness (synthetic) vs. night (real) with correlation 0.270, snow (synthetic) vs. snow (real) with correlation 0.867, fog (synthetic) vs. fog (real) with correlation 0.349, and frost (synthetic) vs. fog (real) with correlation 0.680. While some synthetic corruptions (e.g., snow) closely align with their real-world counterparts, others (e.g., brightness for night) exhibit weaker correlations, highlighting cases where synthetic corruptions may fail as accurate proxies.

can approximate real-world robustness trends, they are not universally reliable across all corruption types.

Interestingly, we observe a moderate positive correlation (Pearson correlation 0.680) in performance against ACDC Fog and 2D Common Corruption Frost. Since the Frost 2D Common Corruption involves superimposing a randomly chosen frost image on the input image with some transparency, one might hypothesize that the model finds the distribution shifts between the two to be moderately similar.

5. Conclusion

Our study provides the most comprehensive benchmarking to date on the reliability of synthetic corruptions as a proxy for real-world distribution shifts in semantic segmentation. Through extensive experiments, we observe a strong correlation in mean performance between synthetic and real-world corruptions, supporting their utility for robustness evaluation. However, a deeper analysis of individual corruption types reveals that while some synthetic corruptions (e.g., snow) closely align with real-world performance, others (e.g., brightness, fog) exhibit weak correlations, highlighting gaps in current benchmarking approaches.

These findings underscore the importance of refining synthetic corruption benchmarks to better capture real-world conditions. To promote OOD evaluations on synthetic datasets, we provide benchmarking of all 15 2D Common Corruptions on the most commonly used semantic segmentation datasets, namely, Cityscapes, ADE20k, and PASCAL VOC2012 datasets. We release our datasets and code to facilitate further research in this direction.

References

[1] Shashank Agnihotri, Steffen Jung, and Margret Keuper. CosPGD: an efficient white-box adversarial attack for pixel-wise prediction tasks. In *Proc. International Conference on Machine Learning (ICML)*, 2024. 3, 8

[2] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 888–897, 2018. 3

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 10

[4] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 2

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 10

[6] 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. 9, 10

[7] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1, 2, 7, 8, 9, 10, 12

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 9

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2, 8, 12

[10] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: a standardized adversarial robustness benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[11] Razvan-Gabriel Dumitru, Darius Peteleaza, and Catalin Craciun. Using duck-net for polyp image segmentation. *Scientific reports*, 13(1):9803, 2023. 2

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, 2010. 2

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012. 2, 9

[14] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *ECCV*, pages 308–325. Springer, 2022. 9

[15] Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17557–17568, 2023. 3

[16] Bharath Hariharan, Pablo Arbelaez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *International Conference on Computer Vision (ICCV)*, 2011. 9

[17] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, pages 447–456, 2015. 9

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 9, 10

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 3, 4, 9, 12

[20] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proc. of the International Conference on Learning Representations (ICLR)*, 2020. 2, 3

[21] J Hoffmann, S Agnihotri, Tonmoy Saikia, and Thomas Brox. Towards improving robustness of compressed cnns. In *ICML Workshop on Uncertainty and Robustness in Deep Learning (UDL)*, 2021. 2, 3

[22] 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. 9

[23] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8828–8838, 2020. 2, 3, 8, 9

[24] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18963–18974, 2022. 2, 3, 9

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[26] 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

Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 1, 10, 12

[27] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 2

[28] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 2, 3

[29] Patrick Müller, Alexander Braun, and Margret Keuper. Classification robustness to common optical aberrations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3632–3643, 2023. 9

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2, 9, 10

[31] 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 (ICCV)*, 2021. 1, 2, 12

[32] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14408–14419, 2023. 2, 9, 10

[33] 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. 10

[34] 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. 7, 8, 10

[35] Hengshuang Zhao. semseg. `https://github.com/hszhao/semseg`, 2019. 9

[36] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 3, 8, 9, 10

[37] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 2, 8

Are Synthetic Corruptions A Reliable Proxy For Real-World Corruptions?

Paper #0005 Supplementary Material

Table Of Content

The supplementary material covers the following information:

A. Correlation In Metrics

Here, we provide a comparison of mean accuracy across synthetic (2D Common Corruptions) and real-world (ACDC) corruptions. The top plot presents mAcc (mean class accuracy) with a stronger correlation of 0.782–0.858, while the bottom plot shows results for aAcc (all pixel accuracy) with a Pearson correlation of 0.688–0.767. These results indicate that synthetic corruptions serve as a reasonable proxy for real-world robustness. Thus, the analysis made using mIoU would also hold if made using other metrics.

B. Implementation Details Of The Benchmarking

Following, we provide details regarding the experiments done for creating the benchmark used in the analysis.

B.1. Compute Resources.

Most experiments were done on a single 40 GB NVIDIA Tesla V100 GPU each, however, SegFormer [34] and Mask2Former [7] with large backbones are more compute-intensive, and thus 80GB NVIDIA A100 GPUs or NVIDIA H100

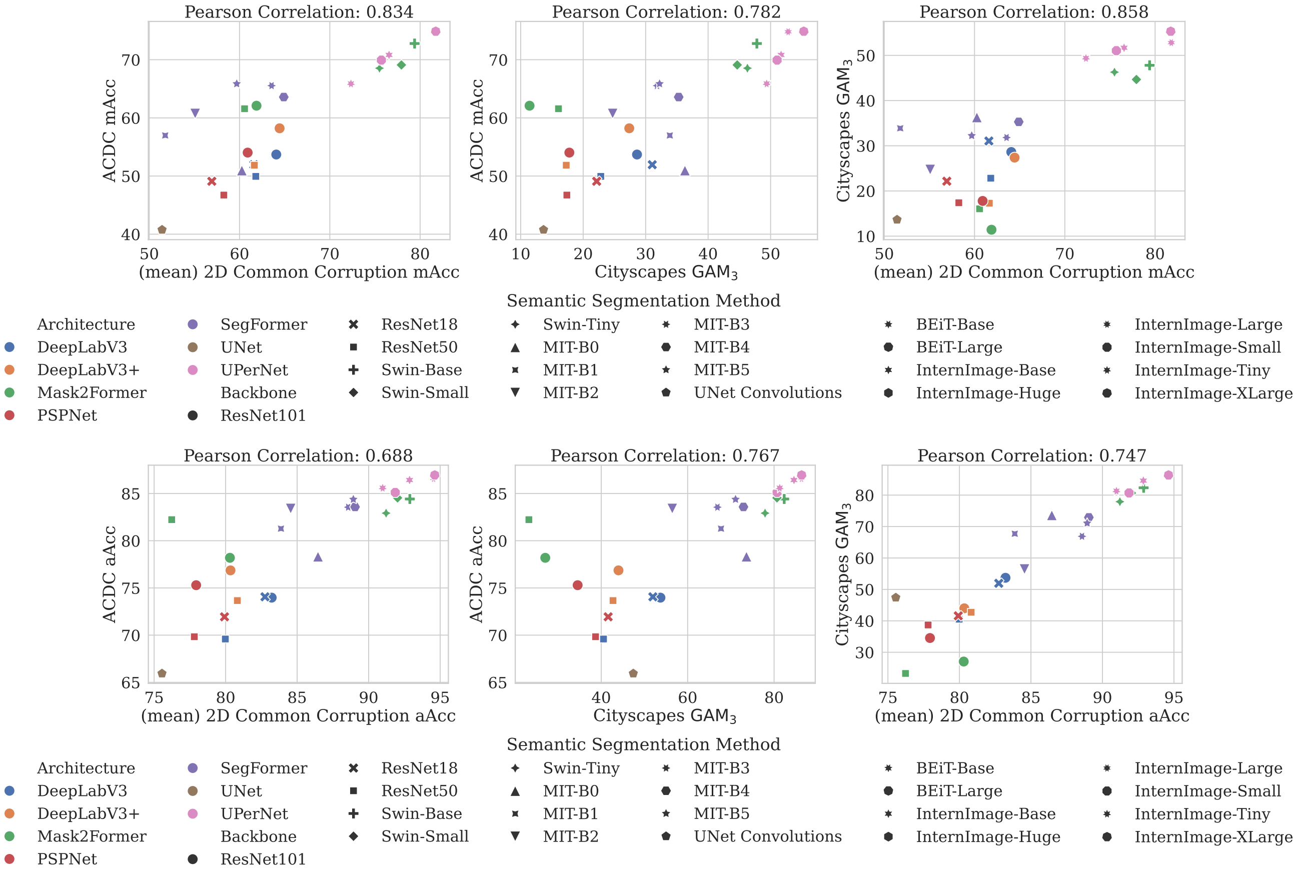

Figure 4. Comparison of mean accuracy across synthetic (2D Common Corruptions) and real-world (ACDC) corruptions. The top plot presents mAcc (mean class accuracy) with a stronger correlation of 0.782–0.858, while the bottom plot shows results for aAcc (all pixel accuracy) with a Pearson correlation of 0.688–0.767. These results indicate that synthetic corruptions serve as a reasonable proxy for real-world robustness, even when measured using metrics other than mIoU

were used for these models, a single GPU for each experiment. Training some of the architectures with large backbones required using two to four GPUs in parallel.

B.2. Dataset Details

Performing OOD robustness evaluations is very expensive and compute-intensive. Thus, for the benchmark, we only use ADE20k, Cityscapes, and PASCAL VOC2012 as these are the most commonly used datasets for evaluation [1, 7, 23, 34, 36].

B.2.1. ADE20K

ADE20K [37] dataset contains pixel-level annotations for 150 object classes, with a total of 20,210 images for training, 2000 images for validation, and 3000 images for testing. Following common practice [1, 34] we evaluate using the validation images.

B.2.2. Cityscapes

The Cityscapes dataset [9] comprises a total of 5000 images sourced from 50 different cities in Germany and neighboring countries. The images were captured at different times of the year and under typical meteorological conditions. Each image was subject to pixel-wise annotations by human experts. The dataset is split into three subsets: training (2975 images), validation (500 images), and testing (1525 images). This dataset has pixel-level annotations for 30 object classes.

B.2.3. PASCAL VOC2012

The PASCAL VOC 2012 [13], contains 20 object classes and one background class, with 1464 training images, and 1449 validation images. We follow common practice [14, 17, 35, 36], and use work by Hariharan et al. [16], augmenting the training set to 10,582 images. We evaluate using the validation set.

Calculating the mIoU. mIoU is the mean Intersection over Union of the predicted segmentation mask with the ground truth segmentation mask.

B.3. Models Used

Table 1 presents a comprehensive reference table for all semantic segmentation models used in our benchmarking. These methods include some of the first efforts in DL-based semantic segmentation methods like UNet [30], and some of the most recent SotA methods like InterImage [32]. Each model is trained on the respective training subset of its dataset and evaluated on the corresponding validation set. The evaluations on 2D Common Corruptions are conducted using the validation sets.

C. 2D Common Corruptions

[19] propose introducing a distribution shift in the input samples by perturbing images with a total of 15 synthetic corruptions that could occur in the real world. These corruptions include weather phenomena such as fog, and frost, digital corruptions such as jpeg compression, pixelation, and different kinds of blurs like motion, and zoom blur, and noise corruptions such as Gaussian and shot noise amongst others corruption types. Each of these corruptions can perturb the image at 5 different severity levels between 1 and 5. The final performance of the model is the mean of the model's performance on all the corruptions, such that every corruption is used to perturb each image in the evaluation dataset. Since these corruptions are applied to a 2D image, they are collectively termed 2D Common Corruptions.

We show examples of perturbed images over some corruptions and the changed predictions in Figure 5.

In Figure 6, we extend the visualizations from Figure 1, additionally showing Night and Rain for ACDC, and Brightness and Frost for 2D Common Corruptions.

D. Benchmarking Results

Following, we include the results from the 2D Common Corruptions evaluations of all the semantic segmentation methods over all of the common corruptions, for PASCAL VOC2012 in Figure 7, for Cityscapes in Figure 8, and for ADE20K in Figure 9.

E. Extension To The Related Work

Kamann and Rother [23] provide an OOD robustness benchmark for semantic segmentation. While they use multiple backbone architectures, such as variants of ResNet [18], MobileNet [22], and Xception [8], their evaluations are limited to the DeepLabV3+ [6] architecture. Our evaluated benchmark extends to multiple architectures and backbones, including recently proposed SotA methods like Mask2Former [7] and InternImage [32].

F. Future Work

Distribution shifts in the real world can be caused by multiple factors, one such factor is lens aberrations. [29] presents many such lens aberrations. Additionally, Kar et al. [24] recently proposed 3D Common Corruptions that take scene depth into account to make corruptions more realistic-looking. We intend to extend our analysis to include these, enabling a more comprehensive robustness study.

F.1. Limitations

Benchmarking the robustness of semantic segmentation methods is a computationally and labor-intensive endeavor. Thus, best utilizing available resources, we benchmark a limited number of settings. While more evaluations like correlation with different severity levels would be interesting, this is the most comprehensive robustness benchmark to date and instills interest to further improve our synthetic corruptions.

Table 1. An Overview of all the semantic segmentation methods used in the benchmark in this work made using SemSegBench. Each of the mentioned backbones has been evaluated using each of the architectures and datasets mentioned in the row in this table.

Backbone	**Architecture**	**Datasets**	**Time Proposed (yyyy-mm-dd)**
ResNet101 [18]	DeepLabV3 [5], DeepLabV3+ [6], Mask2Former [7], PSPNet [36]	ADE20K, Cityscapes, PASCAL VOC 2012	2017-12-05
ResNet18 [18]	DeepLabV3 [5], DeepLabV3+ [6], PSPNet [36]	Cityscapes	2017-12-05
ResNet50 [18]	DeepLabV3 [5], DeepLabV3+ [6], Mask2Former [7], PSPNet [36]	ADE20K, Cityscapes, PASCAL VOC 2012	2017-12-05
Swin-Base [26]	Mask2Former [7]	ADE20K, Cityscapes, PASCAL VOC 2012	2022-06-15
Swin-Small [26]	Mask2Former [7]	ADE20K, Cityscapes, PASCAL VOC 2012	2022-06-15
Swin-Tiny [26]	Mask2Former [7]	ADE20K, Cityscapes, PASCAL VOC 2012	2022-06-15
MIT-B0 [34]	SegFormer [34]	ADE20K, Cityscapes, PASCAL VOC 2012	2021-10-28
MIT-B1 [34]	SegFormer [34]	ADE20K, Cityscapes, PASCAL VOC 2012	2021-10-28
MIT-B2 [34]	SegFormer [34]	ADE20K, Cityscapes, PASCAL VOC 2012	2021-10-28
MIT-B3 [34]	SegFormer [34]	ADE20K, Cityscapes, PASCAL VOC 2012	2021-10-28
MIT-B4 [34]	SegFormer [34]	ADE20K, Cityscapes, PASCAL VOC 2012	2021-10-28
MIT-B5 [34]	SegFormer [34]	ADE20K, Cityscapes, PASCAL VOC 2012	2021-10-28
UNet Convolutions	UNet [30]	Cityscapes	2015-05-18
BEiT-Base [3]	UPerNet [33]	ADE20K	2022-09-03
BEiT-Large [3]	UPerNet [33]	ADE20K	2022-09-03
InternImage-Base [32]	UPerNet [33]	ADE20K, Cityscapes, PASCAL VOC 2012	2023-04-17
InternImage-Huge [32]	UPerNet [33]	ADE20K	2023-04-17
InternImage-Large [32]	UPerNet [33]	ADE20K, Cityscapes	2023-04-17
InternImage-Small [32]	UPerNet [33]	ADE20K, Cityscapes, PASCAL VOC 2012	2023-04-17
InternImage-Tiny [32]	UPerNet [33]	ADE20K, Cityscapes, PASCAL VOC 2012	2023-04-17
InternImage-XLarge [32]	UPerNet [33]	ADE20K, Cityscapes	2023-04-17

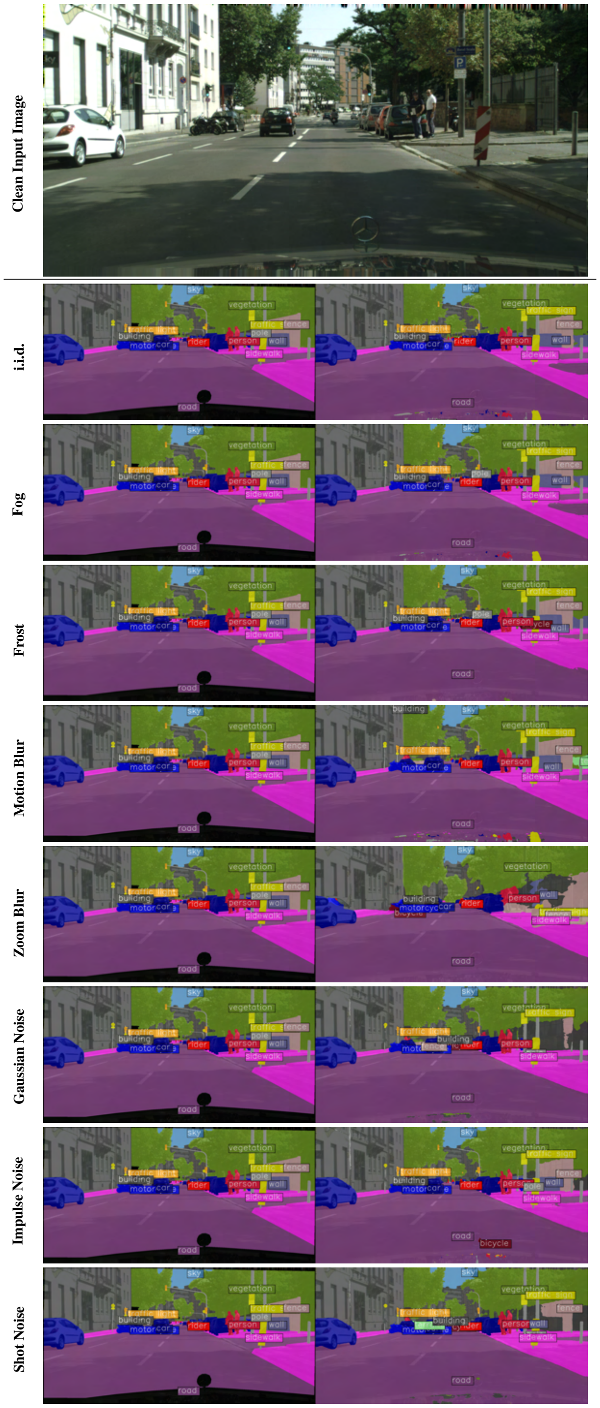

Figure 5. Illustrating changes in prediction due to different 2D Common Corruptions on a randomly chosen input image from the **Cityscapes dataset**, when attaching the semantic segmentation method **InterImage-Base**. In the subfigures with semantic segmentation mask predictions, **Left: Ground Truth Mask**, and **Right: Predicted Mask**.

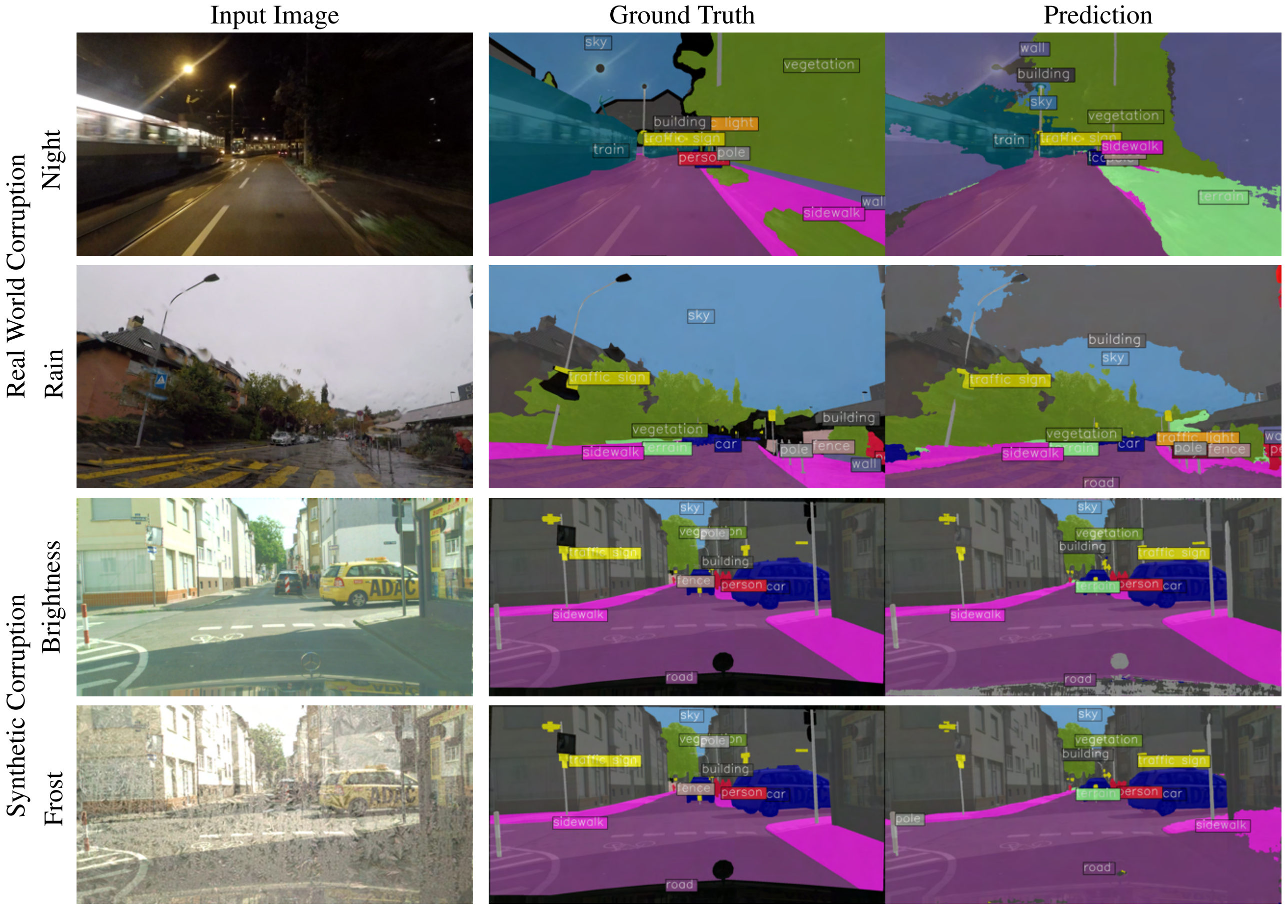

Figure 6. An extension to Figure 1, comparing images with weather corruptions captured in the wild (ACDC [31] and images corrupted using synthetic corruptions [19] and the predictions using a Mask2Former [7] with a Swin-Base [26] backbone trained on the Cityscapes [9] dataset.

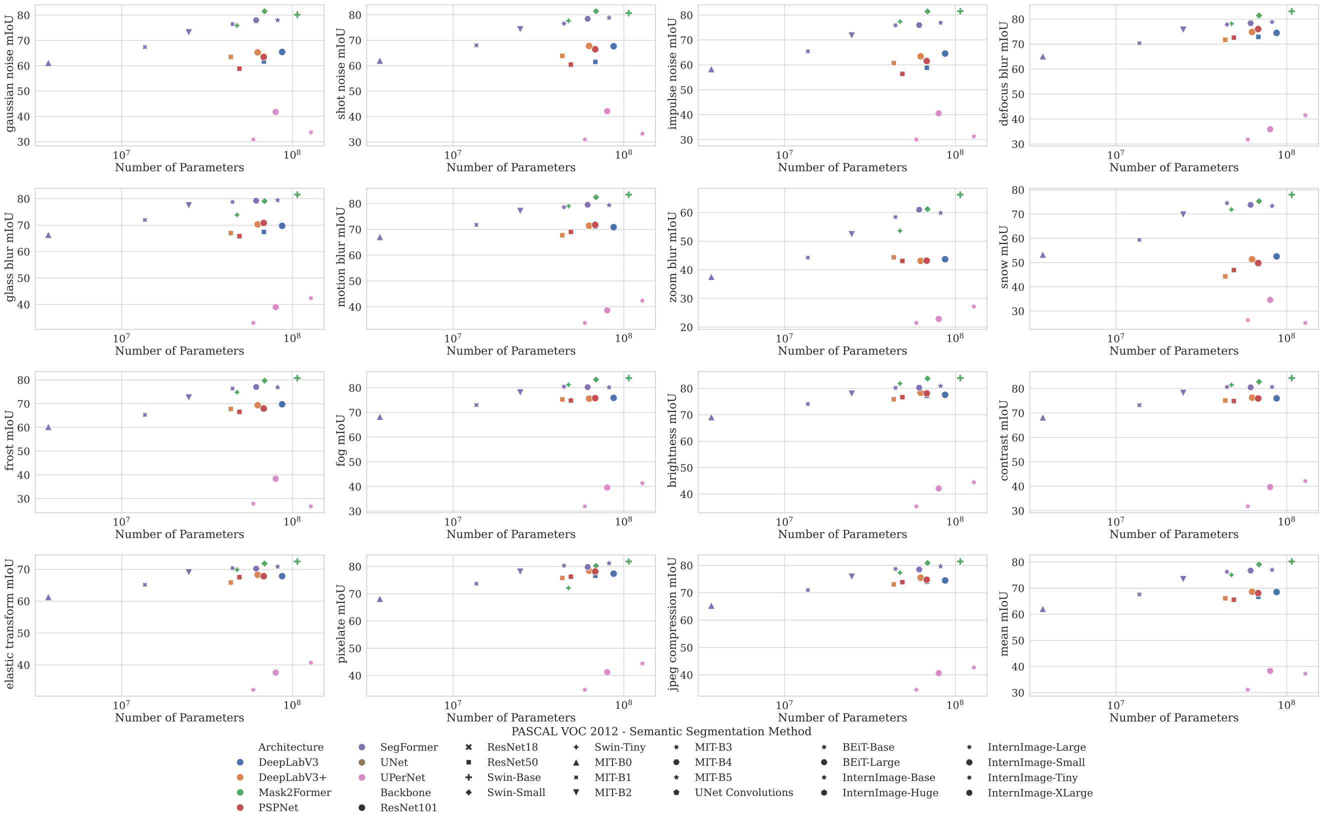

Figure 7. **Dataset used: PASCAL VOC2012**. The correlation in the performance of semantic segmentation methods against different 2D Common Corruptions. The respective axis shows the name of the common corruption used. Colors are used to show different architectures and marker styles are used to show different backbones used by the semantic segmentation methods. For the limited PASCAL VOC2012 evaluations we observe some correlation between the number of learnable parameters and the performance against common corruptions, however, more evaluations (more publicly available checkpoints) are required for a meaningful analysis.

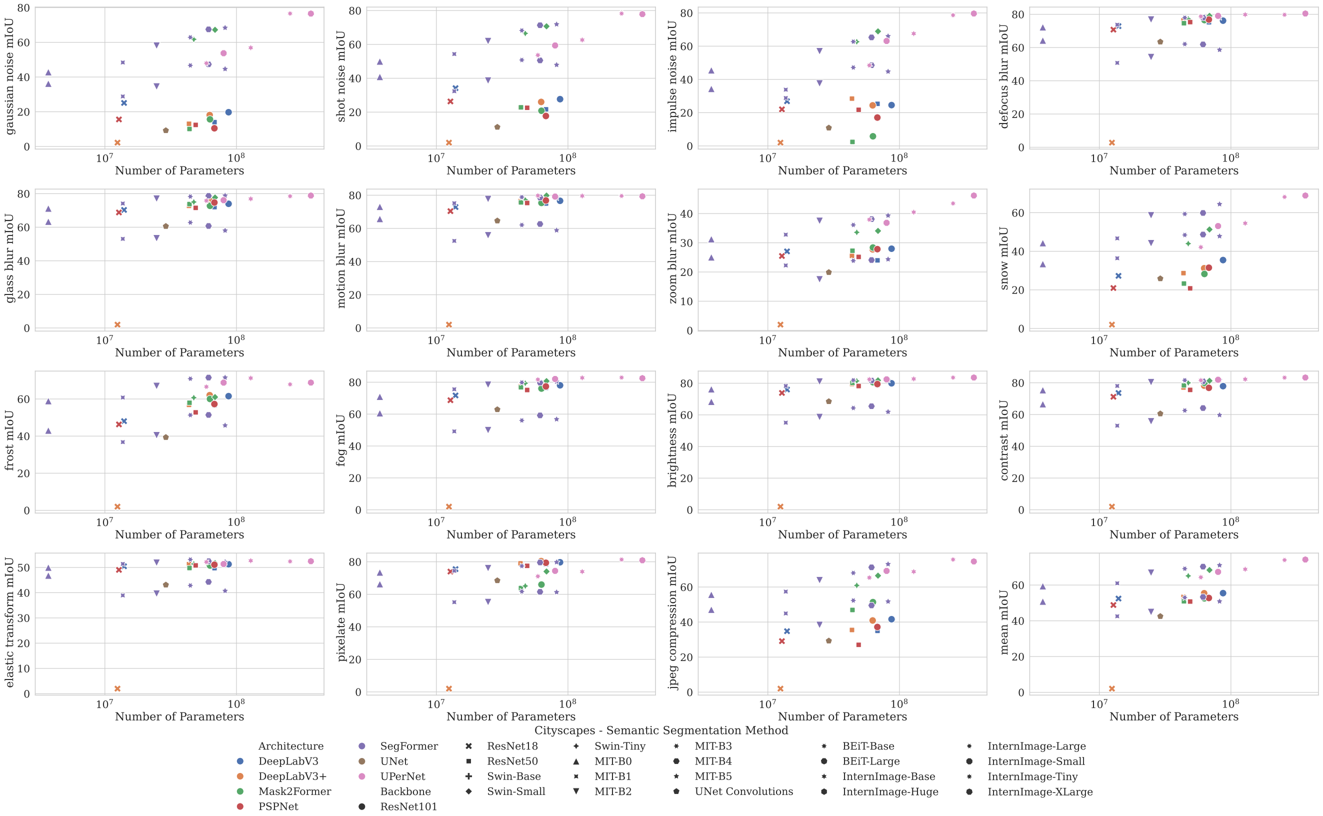

Figure 8. **Dataset used: Cityscapes**. The correlation in the performance of semantic segmentation methods against different 2D Common Corruptions. The respective axis shows the name of the common corruption used. Colors are used to show different architectures and marker styles are used to show different backbones used by the semantic segmentation methods. Except for DeepLabV3+ with a ResNet18 backbone, most other methods show a weak positive correlation between the number of learnable parameters used by a method and its performance against most of the common corruption. Multiple occurrences of an Architecture and Backbone pair are due to their evaluations being performed at two different crop sizes i.e. 512×512, and 512×1024.

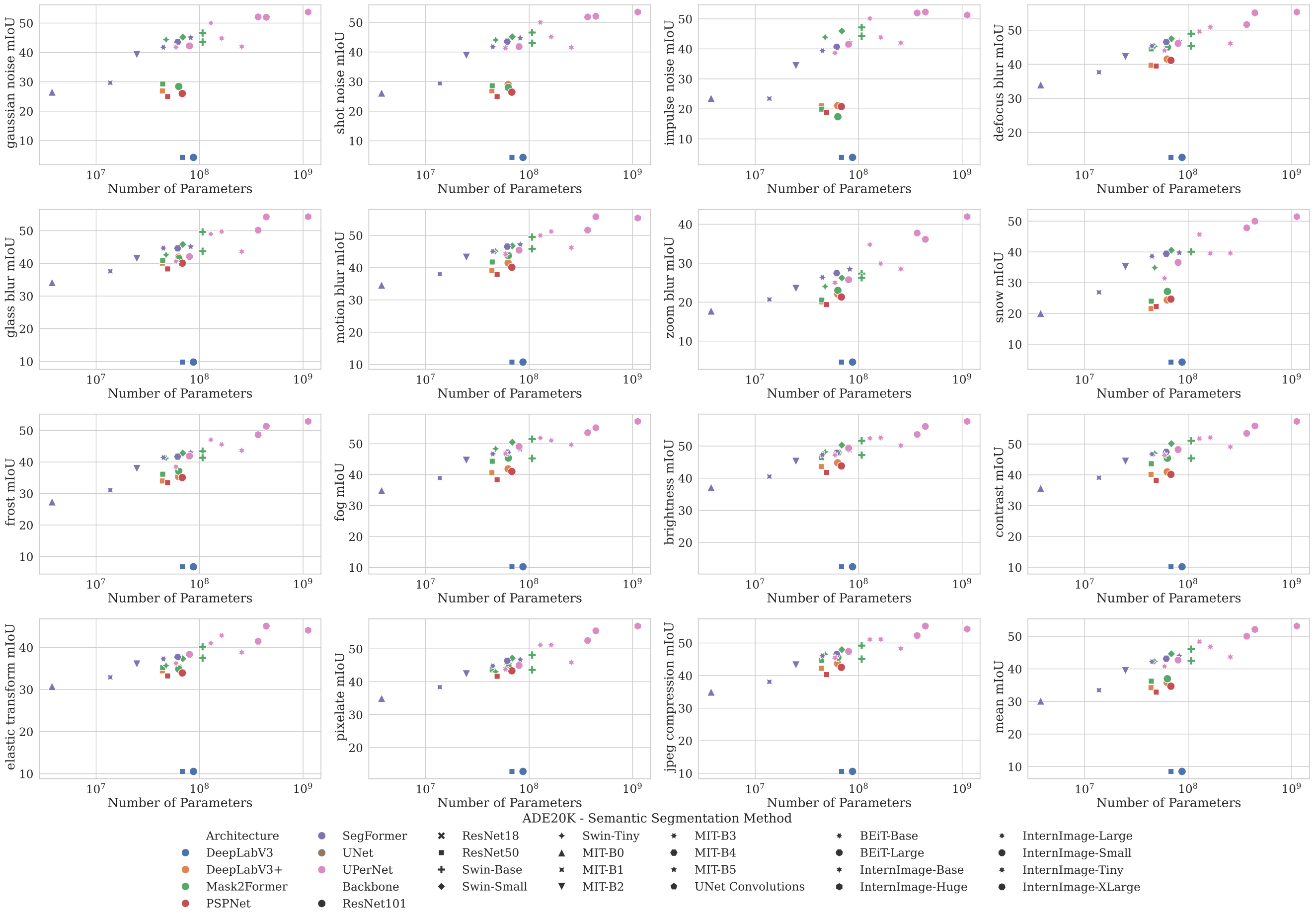

Figure 9. **Dataset used: ADE20K**. The correlation in the performance of semantic segmentation methods against different 2D Common Corruptions. The respective axis shows the name of the common corruption used. Colors are used to show different architectures and marker styles are used to show different backbones used by the semantic segmentation methods. Except for DeepLabV3, all other methods show some positive correlation between the number of learnable parameters used by a method and its performance against any common corruption.