OOD-CV: A Benchmark for Robustness to Individual Nuisances in Real-World Out-of-Distribution Shifts

Bingchen Zhao[1] **Shaozuo Yu**[1] **Wufei Ma**[2] **Mingxin Yu**[3] **Shenxiao Mei**[4] **Angtian Wang**[4] **Ju He**[4]
Alan Yuille[4] **Adam Kortylewski**[4 5]

Deep learning sparked a tremendous increase in the performance of computer vision systems over the past decade, under the implicit assumption that the training and test data are drawn independently and identically distributed (IID) from the same distribution. However, Deep Neural Networks (DNNs) are still far from reaching human-level performance at visual recognition tasks in real-world environments. The most important limitation of DNNs is that they fail to give reliable predictions in unseen or adverse viewing conditions, which would not fool a human observer, such as when objects have an unusual pose, texture, shape, or when objects occur in an unusual context or in challenging weather conditions (Figure 1). The lack of robustness of DNNs in such out-of-distribution (OOD) scenarios is generally acknowledged as one of the core open problems of deep learning, for example by the Turing award winners Yoshua Bengio, Geoffrey Hinton, and Yann LeCun (Bengio et al., 2021). However, the problem largely remains unsolved.

One reason for the limited progress in OOD generalization of DNNs is the lack of benchmark datasets that are specifically designed to measure OOD robustness. Historically, datasets have been pivotal for the advancement of the computer vision field, for example in large-scale image classification (Deng et al., 2009) and segmentation (Lin et al., 2014; Everingham et al., 2015).However, current benchmarks for OOD robustness have important limitations, which limit their usefulness for real-world scenarios. The limitations of OOD benchmarks can be categorized into three types: Some works measure robustness by training models on one dataset and testing them on another dataset without fine-tuning (a; Hendrycks & Dietterich, 2019). However, cross-dataset performance is only a very coarse measure of robustness, which ignores the effects of OOD changes to individual nuisance factors such as the object texture, shape or context. Other approaches artificially generate corruptions of individual nuisance factors, such as weather (Michaelis et al., 2019), synthetic noise (Hendrycks & Dietterich, 2019) or partial occlusion (Wang et al., 2020). However, some nuisance factors are difficult to simulate, such as changes in the object shape or 3D pose. Moreover, artificial corruptions only have limited generalization ability to real-world scenarios. The third type of approach obtains detailed annotation of nuisance variables by recording objects in fully controlled environments, such as in a laboratory (Borji et al., 2016) or using synthetic data (Kortylewski et al., 2018). But such controlled recording can only be done for limited amount of objects and it remains unclear if the conclusions made transfer to real-world scenarios.

In this work, we introduce OOD-CV, a dataset for benchmarking OOD robustness on real images with annotations of individual nuisance variables and labels for several vision tasks. Specifically, the training and IID testing set in OOD-CV consists of 10 rigid object categories from the PASCAL VOC 2012 (Everingham et al.) and ImageNet (Deng et al., 2009) datasets, and the respective labels for image classification, object detection, as well as the 3D pose annotation from the PASCAL3D+ dataset (Xiang et al., 2014). Our main contribution is the collection and annotation of a comprehensive out-of-distribution test set consisting of images that vary w.r.t. the training data in PASCAL3D+ in terms individual nuisance variables, i.e. images of objects with an unseen shape, texture, 3D pose, context or weather (Figure 1). Importantly, we carefully select the data such that each of our OOD data samples only varies w.r.t. one nuisance variable, while the other variables are similar as observed in the training data. We annotate data with class labels, object bounding boxes and 3D object poses, resulting in a total dataset collection and annotation effort more than 650 hours. Our OOD-CV dataset, for the first time, enables studying the influence of individual nuisances on the OOD performance of vision models. In addition to the dataset, we contribute an extensive experimental evaluation of popular baseline methods for each vision task and make several interesting observations which cannot be presented in this extended abstract but can be found in the full paper (Zhao et al., 2021):

[1]Tongji University [2]Purdue University [3]Peking University [4]Johns Hopkins University [5]Max Planck Institute for Informatics. Correspondence to: Bingchen Zhao <zhaobc.gm@gmail.com>, Adam Kortylewski <akortyle@mpi-inf.mpg.de>.

ICML 2022 Shift Happens Workshop: Crowdsourcing metrics and test datasets beyond ImageNet., Baltimore, Maryland, USA, 2022.

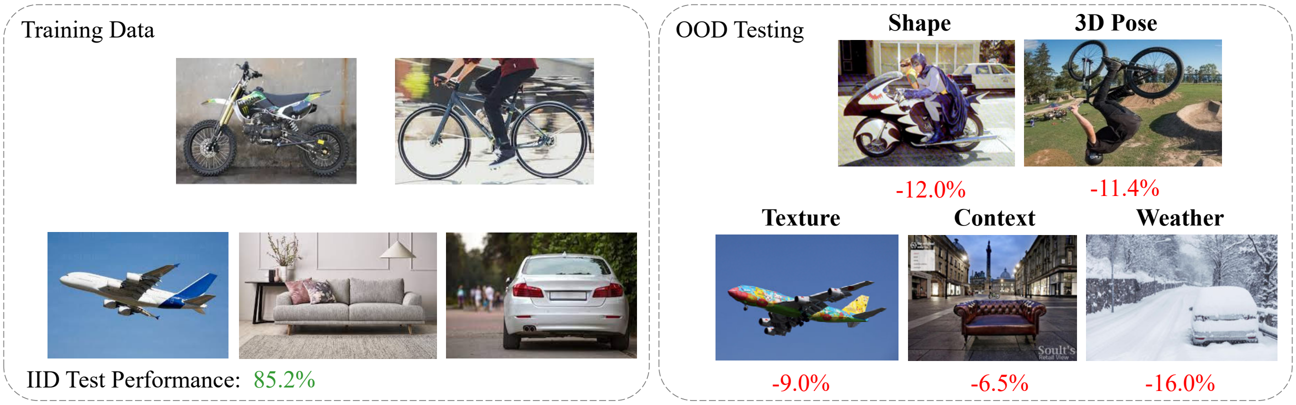

Figure 1. Computer vision models are not robust to real-world distribution shifts at test time. For example, a ResNet50 achieves a test accuracy of 85.2% on our OOD-CV benchmark, when tested on images that are similarly distributed as the training data (IID). However, its performance deteriorates significantly when individual nuisance factors in the test images break the IID assumption. Our benchmark makes it possible, for the first time, to study the robustness of image classification, object detection and 3D pose estimation to OOD shifts in individual nuisances.

1) Some nuisance factors have a much stronger negative effect on the model performance compared to others. Moreover, the negative effect of a nuisance depends on the downstream vision task, because different tasks rely on different visual cues.
2) Current approaches to enhance robustness using strong data augmentation have only marginal effects in real-world OOD scenarios, and sometimes even reduce the OOD performance. Instead, some results suggest that architectures with 3D object representations have an enhanced robustness to OOD shifts in the object shape and 3D pose.
3) We do not observe any significant differences between convolutional and transformer architectures in terms of OOD robustness
We believe our dataset provides a rich testbed for researchers to benchmark and discuss novel approaches to OOD robustness in real-world scenarios and we expect the benchmark to play a pivotal role in driving the future of research on robust computer vision.

References

a. Robust Vision Challenge 2020. http://www.robustvision.net/.

Bengio, Y., Lecun, Y., and Hinton, G. Deep learning for ai. *Communications of the ACM*, 2021.

Borji, A., Izadi, S., and Itti, L. ilab-20m: A large-scale controlled object dataset to investigate deep learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009.

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

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 2015.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *Int. Conf. Learn. Represent.*, 2019.

Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., and Vetter, T. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2014.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M., and Brendel, W. Benchmarking robustness in object detection: Autonomous driving when winter is coming. In *Adv. Neural Inform. Process. Syst.*, 2019.

Wang, A., Sun, Y., Kortylewski, A., and Yuille, A. L. Robust object detection under occlusion with context-aware compositionalnets. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

Xiang, Y., Mottaghi, R., and Savarese, S. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conf. on Applications of Comput. Vis.*, 2014.

Zhao, B., Yu, S., Ma, W., Yu, M., Mei, S., Wang, A., He, J., Yuille, A., and Kortylewski, A. Robin: A benchmark for robustness to individual nuisances in real-world out-of-distribution shifts. *arXiv preprint arXiv:2111.14341*, 2021.