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

Shift Happens Workshop @ ICML 2022
Accepted Paper @ ECCV 2022

Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei,
Angtian Wang, Ju He, Alan Yuille, Adam Kortylewski

We are also organizing a workshop at ECCV 2022, check it out at ood-cv.org

Motivation

Training Data





IID Test Performance: 85.2%

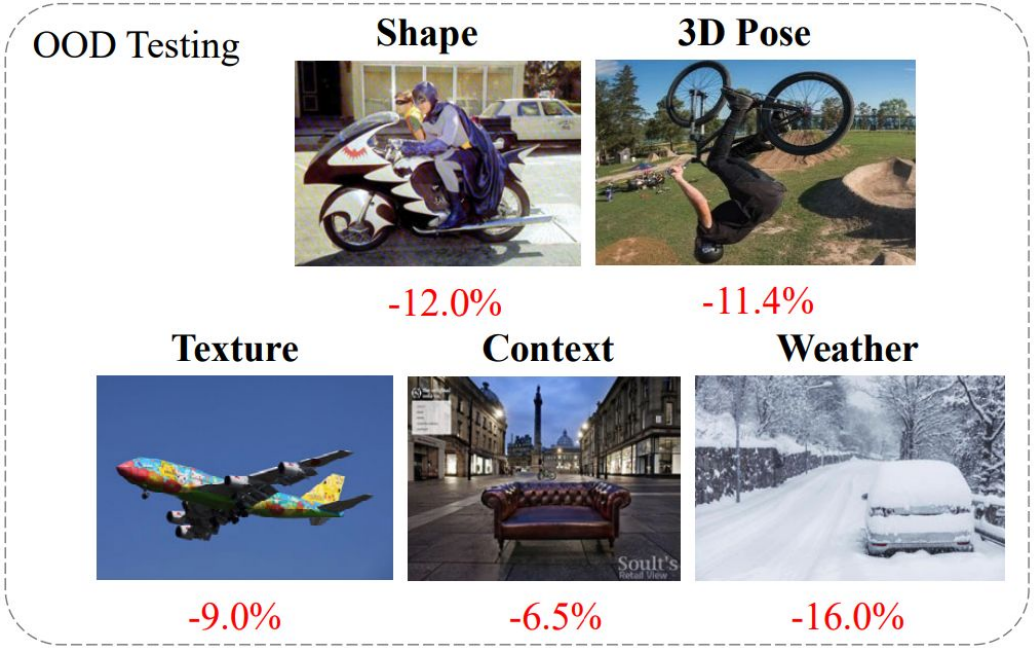
Deep learning is making progress.

But (mostly) only evaluated on IID test data.

Motivation

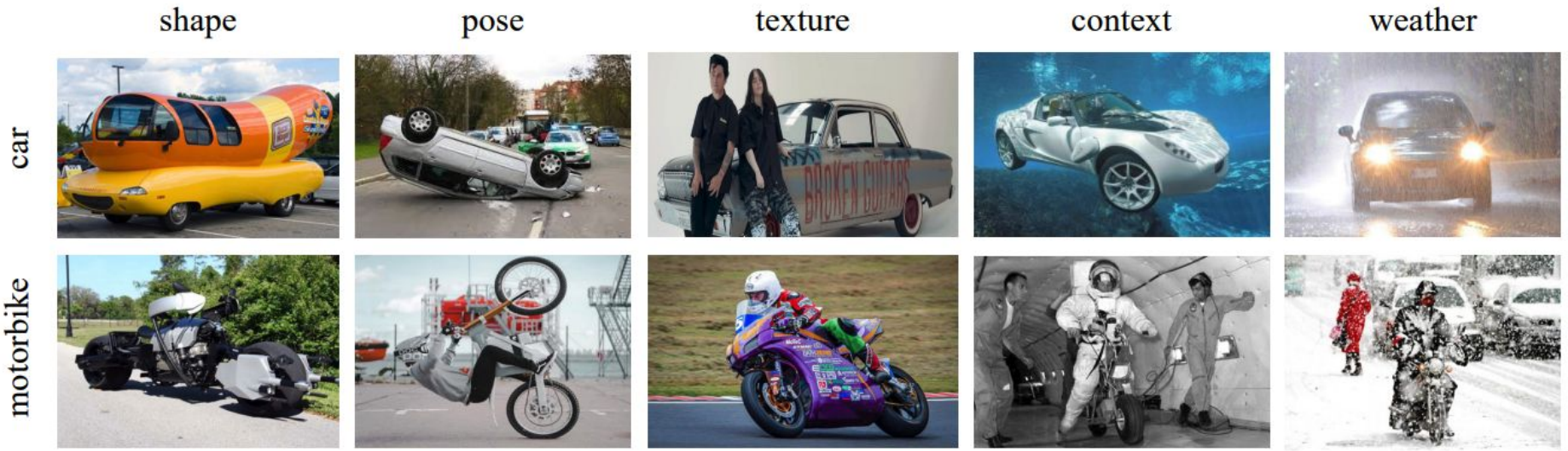
If the model is fed with these naturally occuring OOD examples,

what will the performance be?

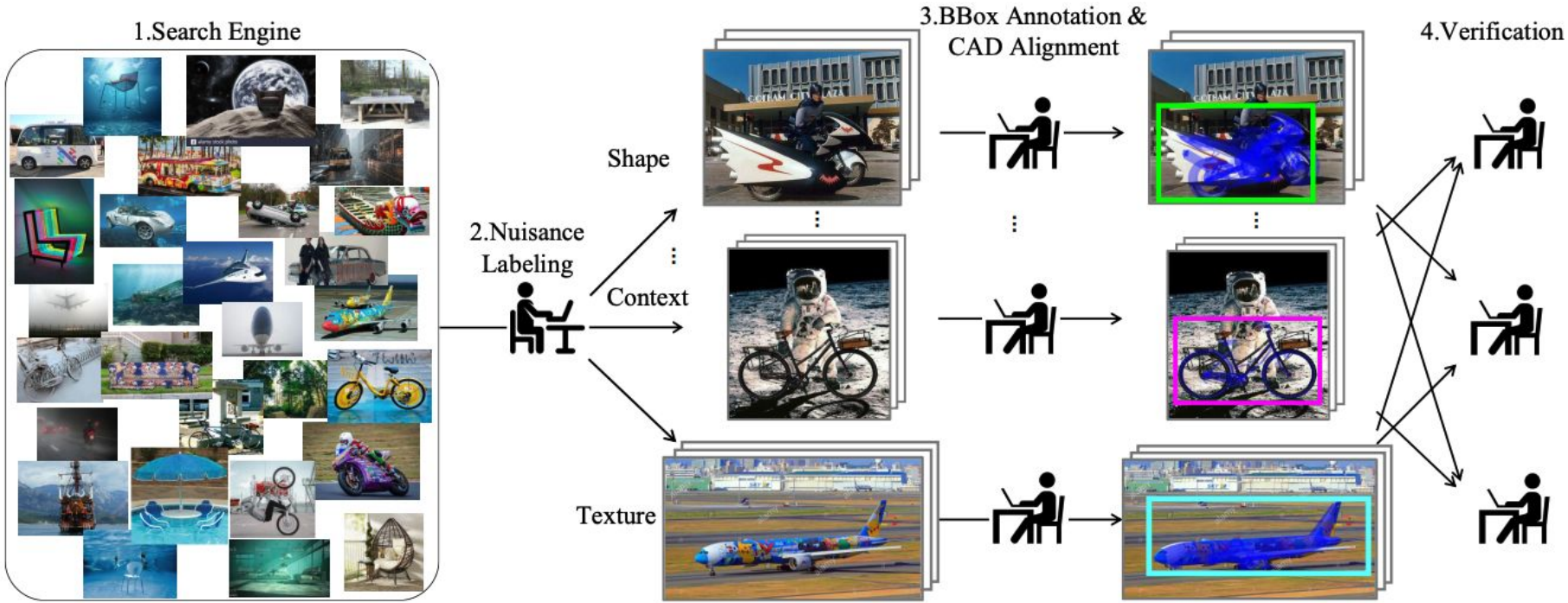


What are important OOD shifts for CV?

We select five attributes that can vary independently as the nuisance for our study.



How do we collect the images and annotations?



We collected BBox and 3D pose annotations for ~500 images per nuisances, total images number are ~2600

Findings

- The influence of nuisance depends on the task

Task		i.i.d	shape	pose	texture	context	weather
Image Classification	ResNet50	85.2%±2.1%	73.2%±1.9%	73.8%±2.0%	76.2%±2.6%	78.7%±2.8%	69.2%±1.9%
	MbNetv3-L	81.5%±1.7%	68.2%±2.0%	71.4%±1.6%	72.1%±2.4%	75.9%±2.9%	66.5%±2.5%
Object Detection	Faster-RCNN	72.6%±1.7%	61.6%±2.4%	62.4%±1.7%	56.3%±1.1%	35.6%±1.8%	50.7%±1.6%
	RetinaNet	74.7%±1.6%	64.1%±2.0%	65.8%±1.9%	61.5%±2.0%	40.3%±2.2%	54.2%±2.0%
3D Pose Estimation	Res50-Specific	62.4%±2.4%	43.5%±2.5%	45.2%±2.8%	51.4%±1.8%	50.8%±1.9%	49.5%±2.1%
	NeMo	66.7%±2.3%	51.7%±2.3%	56.9%±2.7%	52.6%±2.0%	51.3%±1.5%	49.8%±2.0%

Table 2. Robustness to individual nuisances of popular vision models for different vision tasks. We report the performance on i.i.d. test data and OOD shifts in the object shape, 3D pose, texture, context and weather. Note that image classification models are most affected by OOD shifts in the weather, while detection and pose estimation models mostly affected by OOD shifts in context and shape, suggesting that vision models for different tasks rely on different visual cues.

Findings

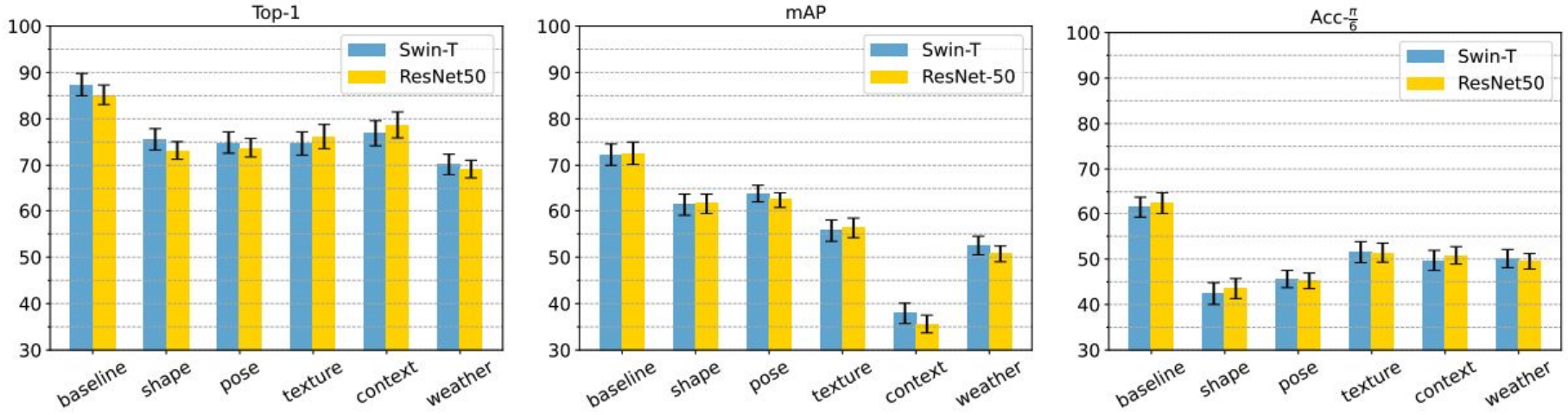
- Strong data augmentations does not help much with these OOD shifts

Classification	i.i.d improvement	Avg. OOD improvements
ResNet-50	0.0%	0.0%
Style Transfer	1.2%	1.1%
AugMix	2.4%	3.0%
Adv. Training	-1.5%	0.6%

Detection	i.i.d improvement	Avg. OOD improvements
ResNet-50	0.0%	0.0%
Style Transfer	0.5%	1.0%
Adv. Training	-1.3%	0.2%

Findings

- CNNs and Transformers perform roughly the same on these nuisances



(a) Image Classification (b) Object Detection (c) 3D Pose Estimation

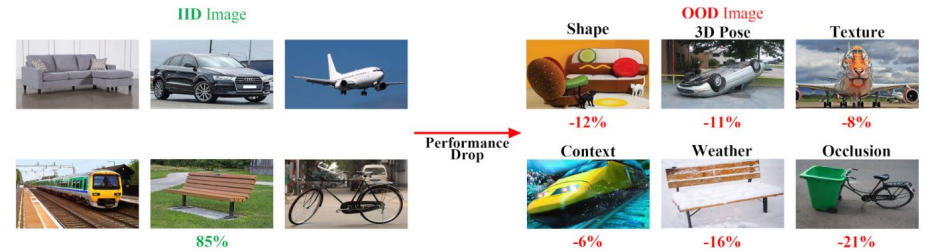
Figure 4. Performance of CNN and Transformer on our robustness benchmark. Transformers have a higher in-domain testing performance, but when it comes to testing on OOD examples, the performance degradation of both CNNs and transformers are mostly the same.

Summary

1. Our dataset enables testing on individual nuisance like:
    a. Object shape
    b. Object 3D pose
    c. Texture appearance
    d. Surrending context
    e. Weather condition
2. The influence of nuisance depends on the task
3. Strong data augmentations does not help much with these OOD shifts
4. CNNs and Transformers perform roughly the same on these nuisances

OOD-CV workshop at ECCV 2022

Check out [ood-cv.org](ood-cv.org).
We have a paper track and a challenge track, the challenge winners have the opportunity to win $10k and be invited to submit to an IJCV special issue!



Deep learning models are usually developed and tested under the implicit assumption that the training and test data are drawn independently and identically distributed (**IID**) from the same distribution. Overlooking out-of-distribution (**OOD**) images can result in poor performance in unseen or adverse viewing conditions, which is common in real-world scenarios.