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

Bingchen Zhao[1], Shaozuo Yu[2], Wufei Ma[3], Mingxin Yu[4], Shenxiao Mei[3], Angtian Wang[3], Ju He[3], Alan Yuille[3], and Adam Kortylewski[3,5,6]

[1]University of Edinburgh [2]The Chinese University of Hong Kong
[3]Johns Hopkins University [4]Peking University
[5]Max Planck Instutite for Informatics [6]University of Freiburg

Abstract. Enhancing the robustness of vision algorithms in real-world scenarios is challenging. One reason is that existing robustness benchmarks are limited, as they either rely on synthetic data or ignore the effects of individual nuisance factors. We introduce OOD-CV , a benchmark dataset that includes out-of-distribution examples of 10 object categories in terms of pose, shape, texture, context and the weather conditions, and enables benchmarking models for image classification, object detection, and 3D pose estimation. In addition to this novel dataset, we contribute extensive experiments using popular baseline methods, which reveal that: 1) Some nuisance factors have a much stronger negative effect on the performance compared to others, also depending on the vision task. 2) Current approaches to enhance robustness have only marginal effects, and can even reduce robustness. 3) We do not observe significant differences between convolutional and transformer architectures. We believe our dataset provides a rich testbed to study robustness and will help push forward research in this area.

1 Introduction

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 [4]. However, the problem largely remains unsolved.

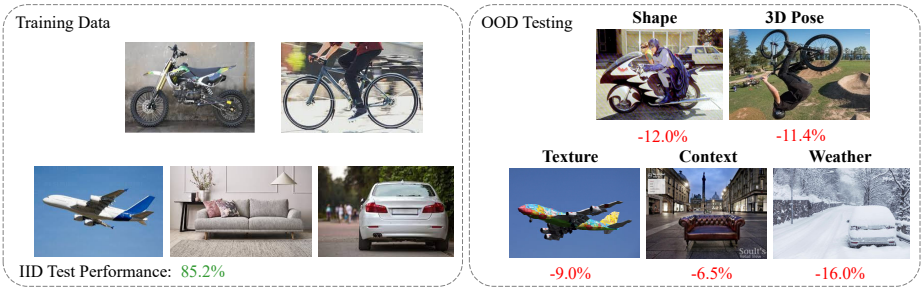


Fig. 1: Computer vision models are not robust to real-world distribution shifts at test time. For example, ResNet50 achieves 85.2% accuracy on our benchmark, when tested on images that are similarly distributed as the training data (IID). However, its performance deteriorates significantly when individual 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.

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 advancement of the computer vision field, e.g. in image classification [10], segmentation [31,13], pose estimation [50,44,52], and part detection [7]. However, benchmarks for OOD robustness have important limitations, which limit their usefulness for real-world scenarios. 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 [1,55,20,23]. 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 [35], synthetic noise [20] or partial occlusion [46]. 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 [6] or using synthetic data [26]. 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 [14] and ImageNet [10] datasets, and the respective labels for image classification, object detection, as well as the 3D pose annotation from the PASCAL3D+ dataset [50]. Our main contribution is the collection and annotation of a com-

prehensive 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 (Fig. 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 ROBIN 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, most importantly: 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 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.

2 Related works

Robustness benchmark on synthetic images. There has been a lot of recent work on utilizing synthetic images to test the robustness of neural networks [29,20,35]. For example, ImageNet-C [20] evaluates the performance of neural networks on images with synthetic noises such as JPEG compression, motion-blur and Gaussian noise by perturbing the standard ImageNet [10] test set with these noises. [35] extends this idea of perturbing images with synthetic noises to the task of object detection by adding these noises on COCO [31] and Pascal-VOC [13] test sets. Besides perturbation from image processing pipelines, there are also work [16] benchmarks the shape and texture bias of DNNs using images with artificially overwritten textures. Using style-transfer [15] as augmentation [16] or using a linear combination between strongly augmented images and the original images [22] have been shown as effective ways of improving the robustness against these synthetic image noises or texture changes. However, these benchmarks are limited in a way that synthetic image perturbations are not able to mimic real-world 3-dimensional nuisances such as novel shape or novel pose of objects. Our experiments in Sec. 4 also show that style-transfer [15] and strong augmentation [22] does not help with shape and pose changes. In addition, these benchmarks are limited to single tasks, for example, ImageNet-C [20] only evaluates the robustness on image classification, COCO-C [35] only evaluates on the tasks of object detection. DomainBed [17] also benchmarks algorithm on OOD

domain generalization on the task of classification. In our work, we evaluate the robustness on real world images, while also evaluate the robustness across different tasks including image classification, object detection, and pose estimation.

Robustness benchmark on real world images. Distribution shift in real-world images are more than just synthetic noises, many recent works [39,23,20] focus on collecting real-world images to benchmark robustness of DNN performances. ImageNet-V2 [39] created a new test set for ImageNet [10] by downloading images from Flickr, and found this new test set causes the model performance to degrade, showing that the distribution shift in the real images has an important influence on DNN models. By leveraging an adversarial filtration technique that filtered out all images that a fixed ResNet-50 [18] model can correctly classifies, ImageNet-A [23] collected a new test set and shows that these adversarially filtered images can transfer across other architectures and cause the performance to drop by a large margin. Although ImageNet-A [23] shows the importance of evaluating the robustness on real-world images, but cannot isolate the nuisance factor. Most recently, ImageNet-R [19] collected four OOD testing benchmarks by collecting images with distribution shifts in texture, geo-location, camera parameters, and blur respectively, and shows that not one single technique can improve the model performance across all the nuisance factors. There are also benchmarks to test how well a model can learn invariant features from unbalanced datasets [43]. And benchmarks composed of many real world shifts [25]. We introduce a robustness benchmark that is complementary to prior datasets, by disentangling individual OOD nuisance factors that correspond to semantic aspects of an image, such as the object texture and shape, the context object, and the weather conditions. Due to rich annotation of our data, our benchmark also enables studying OOD robustness for various vision tasks.

Techniques for improving robustness. To close the gap between the performance of vision models on datasets and the performance in the real-world, many techniques has been proposed [37]. These techniques for improving robustness can be roughly categorized into two types: data augmentation and architectural changes. Adversarial training by adding the worst case perturbation to images at training-time [49], using stronger data augmentation [8,47], image mixtures [22,56,12], and image stylizations [16] during training, or augmenting in the feature space [19] are all possible methods for data augmentation. These data augmentation methods have been proven to be effective for synthetic perturbed images [22,16]. Architectural changes are another way to improve the robustness by adding additional inductive biases into the model. [54] proposed to perform de-noise to the feature representation for a better adversarial robustness. Analysis-by-synthesis appoaches [45,28] can handle scenarios like occlusion by leveraging a generative object model and through top-down feedback [53]. Transformers are a newly emerged architecture for computer vision [11,32,42], and there are works showing that transformers may have a better robustness than CNNs [5,34], although our experiments suggest that this is not the case. Object-centric representations [33,48] have also been show to improve robustness. Self-supervised learned representations also show improvement on OOD

examples [21,57,59,9] Our benchmark enables the comprehensive evaluation of such techniques to improve the robustness of vision models on realistic data, w.r.t. individual nuisances and vision tasks. We find that current approaches to enhance robustness have only marginal effects, and can even reduce robustness, thus highlighting the need for an enhanced effort in this research direction.

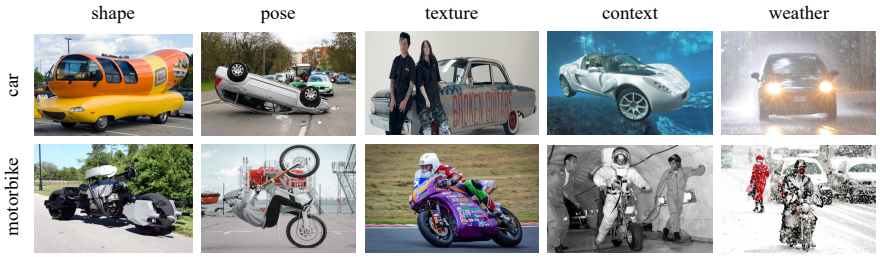


Fig. 2: Examples from our dataset with OOD variations of individual nuisance factors including the object shape, pose, texture, context and weather conditions.

3 Dataset collection

In this section, we introduce the design of the OOD-CV benchmark and discuss the data collection process to obtain the OOD images and annotations.

3.1 What are important nuisance factors?

The goal of the OOD-CV benchmark is to measure the robustness of vision models to realistic OOD shifts w.r.t. important individual nuisance factors. To achieve this, we define an ontology of nuisance factors that are relevant in real-world scenarios following related work on robust vision [36,41,38,2,27] and taking inspiration from the fact that images are 3D scenes with a hierarchical compositional structure, where each component can vary independently of the other components. In particular, we identify five important nuisance factors that vary strongly in real-world scenarios: the object shape, its 3D pose, and texture appearance, as well as the surrounding context and the weather conditions. These nuisance factors can be annotated by a human observer with reasonable effort, while capturing a large amount of the variability in real-world images. Notably, each nuisance can vary independently from the other nuisance factors, which will enable us to benchmark the OOD effect of each nuisance individually.

3.2 Collecting images

OOD data can only be defined w.r.t. some reference distribution of training data. For our dataset, the reference training data is based on the PASCAL3D+ [51] dataset which is composed of images from Pascal-VOC [13] and ImageNet [10] datasets, and contains annotations of the object class, bounding box and 3D pose. Our goal is to collect images where only one nuisance factor is OOD w.r.t. training data, while other factors are similar as in training data.

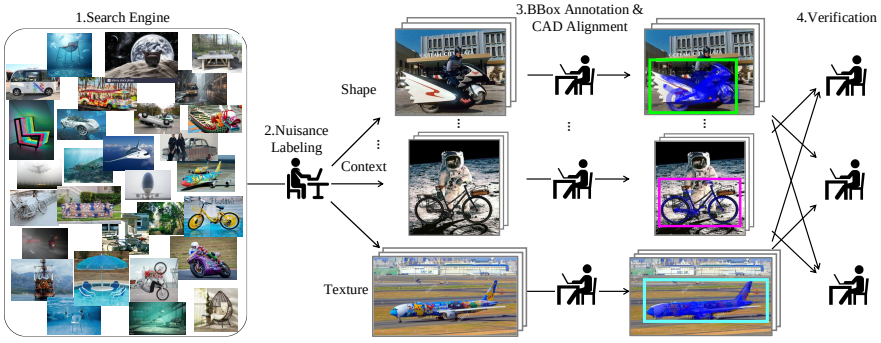


Fig. 3: The data is collected from internet using a predefined set of search keywords, we then manually filter out all images that do not have OOD nuisances or have multiple nuisances. After collecting and splitting the data into different collections with different nuisance, we label images with object bounding boxes and align a CAD model to estimate the 3D pose of the object. The CAD models are overlaid on the images in blue. After each image annotation has been verified by at least two other annotators, we include it in our final dataset.

To collect data with OOD nuisance factors, we search the internet using a curated set of search keywords that are combinations of the object class from the PASCAL3D+ dataset and attribute words that may retrieve images with OOD attributes, e.g. "car+hotdog" or "motorbike+batman", a comprehensive list of our search keywords used can be found in the supplementary material. Note that we only use 10 object categories from PASCAL3D+ , as we could not find sufficient OOD test samples for all nuisances for the categories "bottle" and "television". We manually filtered images with multiple nuisances and put an effort in retaining images that significantly vary in terms of one nuisance only. Following this approach, we collect 2632 images with OOD nuisances in terms of shape, texture, context and weather. Examples images are shown in Fig. 2.

To create OOD dataset splits regarding 3D pose and shape we leverage the shape and pose annotations from PASCAL3D+. These allow us to split the dataset such that 3D pose and shape of training and testing set do not overlap. We augment these OOD splits in pose and shape with additional data that we collect from internet. Statistics of our dataset are shown in the supplementary.On average we have 52 images per nuisance and object class which is comparable to other datasets, e.g. ImageNet-C with an average of 50 images.

Overall, the OOD-CV benchmark is an image collection with a total of 13297 images composed from PASCAL3D+ and internet where 10665 images are from PASCAL3D+ and 2632 images are collected and annotated by us. To ensure that test data is really OOD, three annotators went through all training data from PASCAL3D+ and filtered out images from training set that were too similar to OOD test data. To enable us to benchmark OOD robustness, the nuisance factors and vision tasks were annotated as discussed in the next section.

3.3 Data annotation

A schematic illustration of the annotation process is shown in Fig. 3. After collecting the images from internet, we first classify the images according to OOD nuisance factor following the ontology discussed in Sec. 3.1. Subsequently, we annotate the images to enable benchmarking of a variety of vision tasks. In particular, we annotate the object class, 2D bounding box and 3D object pose. Note that we include the 3D pose, despite the large additional annotation effort compared to class labels and 2D bounding boxes, because we believe that extracting 3D information from images is an important computer vision task.

The annotation of the bounding boxes follows the coco format [31]. We used a web-based annotation tool [1] that enables the data annotation with multiple annotators in parallel. The 3D pose annotation mainly follows the pipeline of PASCAL3D+ [51] and we use a slightly modified annotation tool from the one used in the PASCAL3D+ toolkit [2]. Specifically, to annotate the 3D pose each annotator selects a CAD model from the ones provided in PASCAL3D+, which best resembles the object in the input image. Subsequently, the annotator labels several keypoints to align the 6D pose of the CAD model to the object in input image. After we have obtained annotations for the images, we count the distribution of number of images in each category and for categories with less images than average, we continue to collect additional images from internet for the minority categories. Following this annotation process, we collected labels for all 2632 images covering all nuisance factors. Finally, the annotations produced by every annotator are verified by at least two other annotators to ensure the annotation is correct. We have a total of 5 annotators, and it took about 15 minutes per image, resulting in more than 650 hours of annotation effort.

Dataset splits. To benchmark the IID performance, we split the 10665 images that we retained from the PASCAL3D+ dataset into 8532 training images and 2133 test images. The OOD dataset splits for the nuisances "texture", "context", and "weather" can be directly used from our collected data. As the Pascal3D+ data is highly variable in terms of 3D pose and shape, we create OOD splits w.r.t. the nuisances "pose" and "shape" by biasing the training data using the pose and shape annotations, such that the training and test set have no overlap in terms of shape and pose variations. These initial OOD splits are further enhanced using the data we collected from the internet. The dataset and a detailed documentation of the dataset splits is available online[3].

4 Experiments

We test the robustness of vision models w.r.t. out-of-distribution shifts of individual nuisance factors in Sec. 4.1 and evaluate popular methods for enhancing the model robustness of vision models using data augmentation techniques (Sec. 4.2) and changes to the model architecture (Sec. 4.3). Finally, we study

[1] https://github.com/jsbroks/coco-annotator
[2] https://cvgl.stanford.edu/projects/pascal3d.html
[3] http://ood-cv.org/, Also see the supplementary material.

the effect when multiple nuisance factors are subject to OOD shifts in Sec. 4.4 and give a comprehensive discussion of our results in Sec. 5.

Experimental Setup. Our OOD-CV dataset enables benchmark vision models for three popular vision tasks: image classification, object detection, and 3D pose estimation. We study robustness of popular methods for each task w.r.t. OOD shifts in five nuisance factors: object shape, 3D pose, object texture, background context and weather conditions. We use the standard evaluation process of mAP@50 and Acc@$\frac{\pi}{6}$ for object detection and 3D pose estimation respectively. For image classification, we crop the objects in the images based on their bounding boxes to create object-centric images, and use the commonly used Top-1 Accuracy to evaluate the performance of classifiers. In all our experiments, we control variables such as the number of model parameters, model architecture, and training schedules to be comparable and only modify those variables we wish to study. The models for image classification are pre-trained on ImageNet [10] and fine-tuned on our benchmark. As datasets for a large-scale pre-training are not available for 3D pose estimation, we randomly initialize the pose estimation models and directly train them on the OOD-CV training split. Detailed training settings for vision models and data splits can be found in our supplementary.

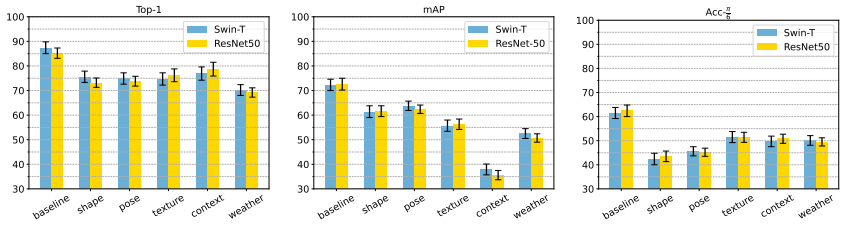
Table 1: Robustness to individual nuisances of popular vision models for different vision tasks. We report the performance on i.i.d. test data and OOD shifts in the object shape, 3D pose, texture, context and weather. Note that image classification models are most affected by OOD shifts in the weather, 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.

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%

4.1 Robustness to individual nuisances

The OOD-CV benchmarks enables, for the first time, to study the influence of OOD shifts in individual nuisance factors on tasks of classification, detection and pose estimation. We first study the robustness of one representative methods for each task. In Tab. 1, we report the test performance on a test set with i.i.d. data, as well as the performance under OOD shifts to all five nuisance factors that are annotated in the OOD-CV benchmark. We observe that for image classification, the performance of the classic ResNet50 architecture [18] drops significantly for every OOD shift in the data. The largest drop is observed under OOD shifts in the weather conditions (-16.0%), while the performance

drop for OOD context is only −6.5%. The results suggests that the model does not rely very much on contextual cues but rather focuses more on the overall gist of the image, which is largely affected by changing weather conditions. Moreover, the classification model is more affected by OOD shifts in geometric cues such as the shape and the 3D pose, compared to the object texture. On the contrary, for object detection the performance of a Faster-RCNN [40] model drops the most under OOD context (−37% mAP@50), showing that detection models rely strongly on contextual cues. While the performance of the detection model also decreases significantly across all OOD shifts, the appearance-based shifts like texture, context and weather have a stronger influence compared to OOD shifts in the shape and pose of the object. For the task of 3D pose estimation, we study a ResNet50-Specific [58] model, which is a common pose estimation baseline that treats pose estimation as a classification problem (discretizing the pose space and then classifying an image into one of the pose bins). We observe that the performance for 3D pose estimation drops significantly, across all nuisance variables and most prominently for OOD shifts in the shape and pose.



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

Fig. 4: Performance of CNN and Transformer on our benchmark. Transformers have a higher in-domain performance, but CNNs and transformers degrades mostly the same on OOD testing examples.

In summary, our experimental results show that **OOD nuisances have different effect on vision models for different visual tasks**. This suggests OOD robustness should not be simply treated as a domain transfer problem between datasets, but instead it is important to study the effects of individual nuisance factors. Moreover, OOD robustness might require different approaches for each vision tasks, as we observe clear differences in the effect of OOD shifts in individual nuisance factors between vision tasks.

4.2 Data Augmentation for Enhancing Robustness

Data augmentation techniques have been widely adopted as an effective means of improving the robustness of vision models. Among such data augmentation methods, stylizing images with artistic textures [16], mixing up the original image with a strongly augmented image (AugMix [22]), and adversarial training [49] are the most effective methods. We test these data augmentation methods on

Table 2: Effect of data augmentation techniques on OOD robustness for three vision tasks. We report the performance of one baseline model for each task, as well as the same model trained with different augmentation techniques: Stylizing , AugMix[22] and Adversarial Training [49]. We evaluate all models on i.i.d. test data and OOD shifts in the object shape, 3D pose, texture, context and weather. Strong data augmentation only improves robustness to appearance-based nuisances but even decreases the performance to geometry-based nuisances like shape and 3D pose.

top-1	i.i.d	shape	pose	texture	context	weather
ResNet-50	85.2%	73.2%	73.8%	76.2%	78.7%	69.2%
Style Transfer	86.4%	72.8%	72.6%	78.9%	78.8%	73.6%
AugMix	87.6%	73.0%	73.3%	82.1%	82.6%	75.2%
Adv. Training	83.7%	72.1%	71.6%	77.9%	79.9%	72.6%

(a) Top-1 accuracy results on image classification

mAP-50	i.i.d	shape	pose	texture	context	weather
Faster-RCNN	72.6%	61.6%	62.4%	56.3%	35.6%	50.7%
Style Transfer	73.1%	59.8%	61.3%	58.4%	39.4%	53.5%
Adv. Training	71.3%	60.4%	60.1%	57.4%	36.9%	52.8%

(b) mAP@50 results on object detection

Acc-$\frac{\pi}{6}$	i.i.d	shape	pose	texture	context	weather
Res50-Spec.	62.4%	43.5%	45.2%	51.4%	50.8%	49.5%
Style Transfer	63.1%	41.8%	44.7%	55.8%	54.3%	53.8%
AugMix	64.8%	44.1%	44.8%	56.7%	54.7%	55.6%
Adv. Training	61.1%	41.7%	43.5%	52.4%	51.7%	50.9%

(c) Acc-$\frac{\pi}{6}$ results on pose estimation

OOD-CV to find out if and how they affect the OOD robustness. The experimental results are summarized in Tab. 2. Overall, AugMix [22] improves the OOD robustness the most for image classification and pose estimation. While AugMix is not directly applicable to object detection, we observe that strong data augmentation style transfer [15] leads to a better improvement compared to adversarial training. Importantly, these data augmentation methods improve the OOD robustness mostly w.r.t. appearance-based nuisances like texture, context, and weather.However, in all our experiments *data augmentation slightly reduces the performance* under OOD shape and 3D pose. We suspect that this happens because data augmentation techniques mostly change appearance-based properties of the image and do not change the geometric properties of the object (i.e. shape and 3D pose). Similar trends are observed across all three of the tasks

Table 3: OOD robustness of models with different capacities. While the performance degradation of MobileNetv3-Large (MbNetv3-L) are about the same as those of ResNet-50, training with data augmentation technique has smaller effect on MbNetv3-L due to the limited capacity.

	i.i.d	shape	pose	texture	context	weather
ResNet50	85.2%	73.2%	73.8%	76.2%	78.7%	69.2%
+AugMix [22]	87.6%	73.0%	73.3%	82.1%	82.6%	75.2%
MbNetv3-L [24]	81.5%	68.2%	71.4%	72.1%	75.9%	66.5%
+AugMix [22]	83.1%	67.8%	71.6%	74.3%	76.8%	69.7%

we tested, image classification, object detection, and pose estimation. These results suggest that two categories of nuisances exists, namely *appearance-based* nuisances like novel texture, context, and weather, and *geometric-based* nuisances like novel shape and pose. We observe that **data augmentation only improves robustness of appearance-based nuisances but can even decrease the performance w.r.t. geometry-based nuisances**.

4.3 Effect of Model Architecture on Robustness

In this section, we investigate four popular architectural changes that have proven to be useful in real world applications. Paricularly, we evaluate *CNNs vs Transformers*, the *model capacity*, *one stage vs two stage* detectors, and models with *integrated 3D priors*. Note that when we change the model architecture we keep other parameters such as number of parameters and capacity the same.

CNNs vs Transformers. Transformers have emerged as a promising alternative to convolutional neural networks (CNNs) as an architecture for computer vision tasks recently [11,32]. While CNNs have been extensively studied for robustness, the robustness of vision transformers are still under-explored. Some works [5,34] have shown that transformer architecture maybe more robust to adversarial examples, but it remains if this result holds for OOD robustness. In the following, we compare the performance of CNNs and transformers on the tasks of image classification, object detection and 3D pose estimation on the OOD-CV benchmark. Specifically, we replace the backbone the vision models for each task from ResNet-50 to Swin-T [32].Our experimental results are presented in Fig. 4. Each experiment is performed five times and we report mean performance and standard deviation. It can be observed that CNNs and vision transformers have a comparable performance across all tasks as the difference between their performances are within the margin of error. Particularly, we do not observe any enhanced robustness as OOD shifts in individual nuisance factors lead to a similar decrease in performance in both the transformer and the CNN architecture. While we observe a slight performance gain on i.i.d. data in image classification (as reported in many other works), our results suggest that **Transformers do not have any enhanced OOD robustness compared to CNNs**. Note our findings here contrast with previous work on this topic [3],

Table 4: Comparison between one-stage method and two-stage object detection methods. One-stage methods are more robust compared to two-stage methods.

	i.i.d	shape	pose	texture	context	weather
RetinaNet [30]	74.7%	64.1%	65.8%	61.5%	40.3%	54.2%
+Style Transfer [16]	75.8%	62.7%	64.2%	63.7%	44.7%	55.8%
Faster-RCNN [40]	72.6%	61.6%	62.4%	56.3%	35.6%	50.7%
+Style Transfer [16]	73.1%	59.8%	61.3%	58.4%	39.4%	53.5%

Table 5: Robustness of 3D pose estimation methods. We compare "Res50-Specific", which treats pose estimation as classification problem, and "NeMo", which represents the 3D object geometry explicitly. We observe OOD shifts in shape and pose leads to more performance degradation. NeMo has a significantly enhanced performance to OOD shifts in object shape and pose.

	i.i.d	shape	pose	texture	context	weather
Res50-Specific	62.4%	43.5%	45.2%	51.4%	50.8%	49.5%
+AugMix [22]	64.8%	44.1%	44.8%	56.7%	54.7%	55.6%
NeMo [45]	66.7%	51.7%	56.9%	52.6%	51.3%	49.8%
+AugMix [22]	67.9%	53.1%	58.6%	57.8%	55.1%	56.7%

we argue that this is because our benchmark enables the study for individual nuisance factors on real world images, and the control over different individual nuisances give us opportunity to observe more errors in current vision models.

Model capacity. For deployment in real applications, smaller models are preferred because they can yield better efficiency than regular models. In the following, we compare image classification performance of MobileNetV3 [24] in Tab. 3. Compared to ResNet-50, MobileNetv3 suffers a similar performance degradation under OOD shifts in the data. However, data augmentations does not improve the robustness of MobileNetV3 [24] as much as for ResNet-50, *e.g.*, performance on context nuisances improved by 3.9% for ResNet-50, but the improvement is only 0.9% for MobileNetV3. This suggests that **OOD robustness is more difficult to achieve for efficient models with a limaited capacity**.

One stage vs two stage for detection. It is a common belief in object detection community that two-stage detectors are more accurate, while one-stage detectors are more efficient. For object detection task, two popular types of architecture exist, namely one-stage and two stage models. We tested two representative models from these architecture types, RetinaNet [30], a one-stage detector, and Faster-RCNN [40], which is a two-stage detector. From our results in Tab. 4, we observe that RetinaNet achieves a higher performance compared to Faster-RCNN on the OOD-CV benchmark. However, when accounting for improved i.i.d performance, the OOD performance degradation are similar between two models.

These initial result suggests that **two-stage methods achieve a higher score than one-stage methods, but are not necessarily more robustness**.

Models with explicit 3D object geomtery. Recently, Wang et al. [45] introduced NeMo, a neural network architecture for 3D pose estimation that explicitly models 3D geometery, and they demonstrated promising results on enhancing robustness to partial occlusion and unseen 3D poses. In Tab. 5, we compare NeMo [45] model and a general Res50-Specific model on task of pose estimation on OOD-CV benchmark. NeMo [45] shows a stronger robustness against geometric-based nuisances (shape and pose), while robustness on appearance-based nuisances is comparable. This result suggests that, **neural networks with an explicit 3D object representation have a largely enhanced robustness to OOD shifts in geometry-based nuisances**. These results seem complementary to our experiments in the previous section, which demonstrate that strong data augmentation can help to improve the robustness of vision models to appearance-based nuisances, but not to geometry-based nuisances.

We further investigate, if robustness against all nuisance types can be improved by combining data augmentation with architectures that explicitly represent the 3D object geometry. Specifically, we train NeMo [45] with strong augmentations like AugMix [22] and our results in Tab. 5 show that this indeed largely enhances the robustness to OOD shifts in appearance-based nuisances, while retaining (and slightly improving) the robustness to geometry-based nuisances. Result suggests that enhancements of robustness to geometry-based nuisances can be developed independently to those for appearance-based nuisances.

4.4 OOD shifts in Multiple Nuisances

In our experiments, we observed that geometry-based nuisances have different effects compared to appearance-based nuisances. In the following, we test the effect when OOD shifts happen in both of these nuisance types. Specifically, we introduce new dataset splits, which combine appearance-based nuisances, including texture, context, or weather, with the geometry-based nuisances shape and pose. From Tab. 6, we observe **OOD shifts in multiple nuisances amplify each other**. For example, for image classification, an OOD shift in only the 3D pose reduces the performance by 11.4% from 85.2% to 73.8%, and an OOD shift in the context reduces the performance by 6.6%. However, when pose and context are combined the performance reduces by 24.5%. We observe a similar amplification behaviour across all three tasks, suggesting that it is a general effect that is likely more difficult to address compared to single OOD shifts.

5 Conclusion

We have shown that proposed OOD-CV benchmark enables a thorough diagnosis of robustness of vision models to realistic OOD shifts in individual nuisance factors. Overall, we observe that OOD shifts poses a great challenge to current state-of-the-art vision models and requires significant attention from the research community to be resolved. Notably, we found that nuisance factors have a different effect on different vision tasks, suggesting that we might need different solutions for enhancing the OOD robustness for different vision tasks.

Table 6: Robustness to OOD shifts in multiple nuisances. When combined, OOD shifts in appearance-based nuisances and geometric-based nuisances amplifies each other, leads to further decrease compared to effects in individual nuisances.

	i.i.d	texture	context	weather
Classification	85.2%	76.2%	78.7%	69.2%
+ shape	73.2%	62.8%	63.6%	51.2%
+ pose	73.8%	61.9%	60.7%	49.8%
Detection	72.6%	56.3%	35.6%	50.7%
+ shape	61.6%	41.2%	24.3%	30.7%
+ pose	62.4%	45.6%	26.1%	29.8%
Pose estimation	62.4%	51.4%	50.8%	49.5%
+ shape	43.5%	33.1%	31.0%	29.8%
+ pose	45.2%	30.2%	29.7%	28.1%

In our experiments, it can also be clearly observed that the nuisances can be roughly separated into two categories, *appearance-based nuisances* like texture, context, or weather, and another one is *geometry-based nuisances* such as shape or pose. We showed that strong data augmentation enhances the robustness against appearance-based nuisances, but has very little effect on geometric-based nuisances. On the other hand, neural network architectures with an explicit 3D object representation achieve an enhanced robustness against geometric-based nuisances. While we observe that OOD robustness is largely an unsolved and severe problem for computer vision models, our results also suggest a way forward to address OOD robustness in the future. Particularly, that approaches to enhance the robustness may need to be specifically designed for each vision tasks, as different vision tasks focus on different visual cues. Moreover, we observed a promising way forward to a largely enhanced OOD robustness is to develop neural network architectures that represent the 3D object geometry explicitly and are trained with strong data augmentation to address OOD shifts in both geometry-based and appearance-based nuisances combined.

Acknowledgements. AK acknowledges support via his Emmy Noether Research Group funded by the German Science Foundation (DFG) under Grant No. 468670075. BZ acknowledges compute support from LunarAI. AY acknowledges grants ONR N00014-20-1-2206 and ONR N00014-21-1-2812.

References

1. Robust Vision Challenge 2020. http://www.robustvision.net/. 2
2. Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2019. 5
3. Yutong Bai, Jieru Mei, Alan Yuille, and Cihang Xie. Are transformers more robust than cnns? In *Adv. Neural Inform. Process. Syst.*, 2021. 11

4. Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for ai. *Communications of the ACM*, 2021. 1
5. Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Int. Conf. Comput. Vis.*, 2021. 4, 11
6. Ali Borji, Saeed Izadi, and Laurent Itti. ilab-20m: A large-scale controlled object dataset to investigate deep learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2
7. Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 2
8. Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 4
9. Quan Cui, Bingchen Zhao, Zhao-Min Chen, Borui Zhao, Renjie Song, Jiajun Liang, Boyan Zhou, and Osamu Yoshie. Discriminability-transferability trade-off: An information-theoretic perspective. In *Eur. Conf. Comput. Vis.*, 2022. 5
10. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. 2, 3, 4, 5, 8
11. 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. In *Int. Conf. Learn. Represent.*, 2020. 4, 11
12. N Benjamin Erichson, Soon Hoe Lim, Francisco Utrera, Winnie Xu, Ziang Cao, and Michael W Mahoney. Noisymix: Boosting robustness by combining data augmentations, stability training, and noise injections. *arXiv preprint arXiv:2202.01263*, 2022. 4
13. Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 2015. 2, 3, 5
14. 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. 2
15. Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3, 10
16. Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Int. Conf. Learn. Represent.*, 2019. 3, 4, 9, 12
17. Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *Int. Conf. Learn. Represent.*, 2021. 3
18. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 4, 8
19. Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Int. Conf. Comput. Vis.*, 2021. 4

20. Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Int. Conf. Learn. Represent.*, 2019. 2, 3, 4
21. Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Adv. Neural Inform. Process. Syst.*, 2019. 5
22. 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. In *Int. Conf. Learn. Represent.*, 2020. 3, 4, 9, 10, 11, 12, 13
23. Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2, 4
24. Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Int. Conf. Comput. Vis.*, 2019. 11, 12
25. Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021. 4
26. Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. 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. 2
27. Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 5
28. Adam Kortylewski, Qing Liu, Angtian Wang, Yihong Sun, and Alan Yuille. Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. *International Journal of Computer Vision*, 2021. 4
29. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *Int. Conf. Learn. Represent.*, 2017. 3
30. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Int. Conf. Comput. Vis.*, 2017. 12
31. 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 *Eur. Conf. Comput. Vis.*, 2014. 2, 3, 7
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 *Int. Conf. Comput. Vis.*, 2021. 4, 11
33. Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Adv. Neural Inform. Process. Syst.*, 2020. 4
34. Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Int. Conf. Comput. Vis.*, 2021. 4, 11
35. Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Bench-

marking robustness in object detection: Autonomous driving when winter is coming. In *Adv. Neural Inform. Process. Syst.*, 2019. 2, 3
36. 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. 5
37. Sina Mohseni, Haotao Wang, Zhiding Yu, Chaowei Xiao, Zhangyang Wang, and Jay Yadawa. Practical machine learning safety: A survey and primer. *ArXiv*, 2021. 4
38. Weichao Qiu and Alan Yuille. Unrealcv: Connecting computer vision to unreal engine. In *European Conference on Computer Vision*, pages 909–916. Springer, 2016. 5
39. Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Int. Conf. Machine Learning*, 2019. 4
40. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*, 2015. 9, 12
41. Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018. 5
42. Jie Shao, Xin Wen, Bingchen Zhao, and Xiangyang Xue. Temporal context aggregation for video retrieval with contrastive learning. In *IEEE Winter Conf. on Applications of Comput. Vis.*, 2021. 4
43. Kaihua Tang, Mingyuan Tao, Jiaxin Qi, Zhenguang Liu, and Hanwang Zhang. Invariant feature learning for generalized long-tailed classification. In *Eur. Conf. Comput. Vis.*, 2022. 4
44. Jonathan Tremblay, Thang To, and Stan Birchfield. Falling Things: A synthetic dataset for 3D object detection and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 2
45. Angtian Wang, Adam Kortylewski, and Alan Yuille. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. In *Int. Conf. Learn. Represent.*, 2021. 4, 12, 13
46. Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2
47. Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training. In *NeurIPS*, 2021. 4
48. Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. *arxiv: 2205.15288*, 2022. 4
49. Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *Int. Conf. Learn. Represent.*, 2020. 4, 9, 10
50. Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conf. on Applications of Comput. Vis.*, 2014. 2
51. Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conf. on Applications of Comput. Vis.*, 2014. 5, 7
52. Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018. 2

53. Mingqing Xiao, Adam Kortylewski, Ruihai Wu, Siyuan Qiao, Wei Shen, and Alan Yuille. Tdmpnet: Prototype network with recurrent top-down modulation for robust object classification under partial occlusion. In *European Conference on Computer Vision*, pages 447–463. Springer, 2020. 4
54. Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 4
55. Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*, 2021. 2
56. 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 *Int. Conf. Comput. Vis.*, 2019. 4
57. Bingchen Zhao and Xin Wen. Distilling visual priors from self-supervised learning. In *Eur. Conf. Comput. Vis.*, 2020. 5
58. Xingyi Zhou, Arjun Karpur, Linjie Luo, and Qixing Huang. Starmap for category-agnostic keypoint and viewpoint estimation. In *Eur. Conf. Comput. Vis.*, 2018. 9
59. Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *Int. Conf. Comput. Vis.*, 2021. 5