COCO-O: A Benchmark for Object Detectors under Natural Distribution Shifts

Xiaofeng Mao[†] Yuefeng Chen[†] Yao Zhu[‡] Da Chen[§] Hang Su[¶]
Rong Zhang [†] Hui Xue[†]
[†]Alibaba Group, [‡]Zhejiang University, [§]University of Bath, [¶]Tsinghua University
{mxf164419,yuefeng.chenyf}@alibaba-inc.com

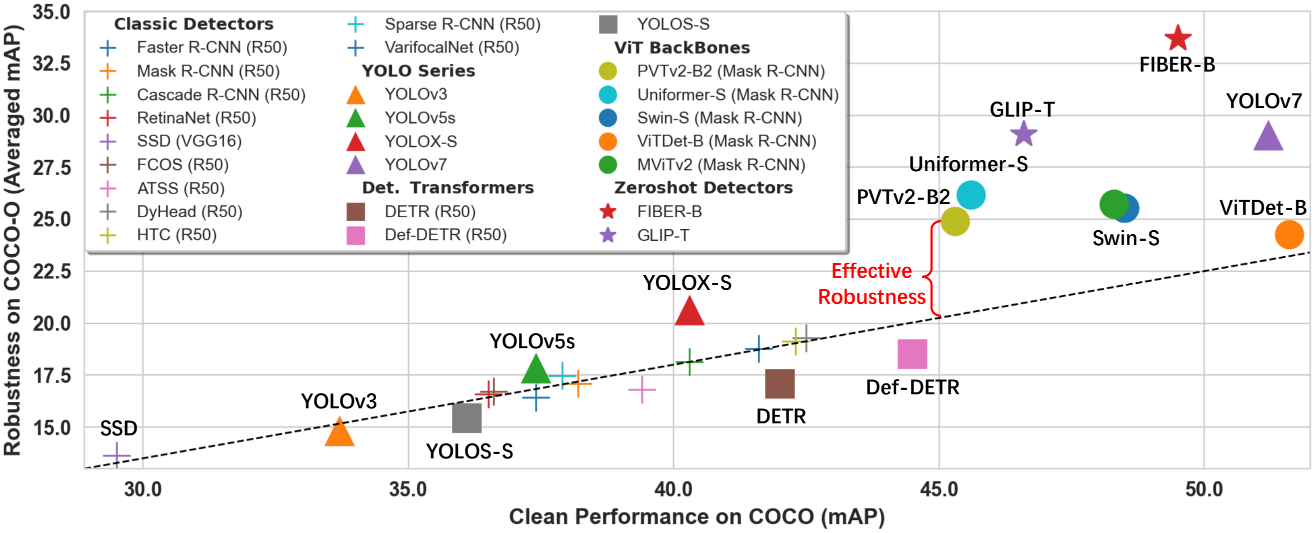

Figure 1: An overview of representative object detectors evaluated on COCO and our COCO-O dataset. The plotted dash line presents the linear function fitted on classic detectors, which demonstrates the trend of COCO-O mAP increasing along with COCO mAP. The *Effective Robustness* (red text in figure) measures how far the model lies above the linear fit.

Abstract

Practical object detection application can lose its effectiveness on image inputs with natural distribution shifts. This problem leads the research community to pay more attention on the robustness of detectors under Out-Of-Distribution (OOD) inputs. Existing works construct datasets to benchmark the detector's OOD robustness for a specific application scenario, e.g.*, Autonomous Driving. However, these datasets lack universality and are hard to benchmark general detectors built on common tasks such as COCO. To give a more comprehensive robustness assessment, we introduce* ***COCO-O****(ut-of-distribution), a test dataset based on COCO with 6 types of natural distribution shifts. COCO-O has a large distribution gap with training data and results in a significant* ***55.7%*** *relative performance drop on a Faster R-CNN detector. We leverage COCO-O to conduct experiments on* ***more than 100 modern object detectors*** *to investigate if their improvements are credible or just over-fitting to the COCO test set.* ***Unfortunately, most classic detectors in early years do not exhibit strong OOD generalization.*** *We further study the robustness effect on recent breakthroughs of detector's architecture design, augmentation and pre-training techniques. Some empirical findings are revealed: 1)* ***Compared with detection head or neck, backbone is the most important part for robustness;*** *2)* ***An end-to-end detection transformer design brings no enhancement, and may even reduce robustness;*** *3)* ***Large-scale foundation models have made a great leap on robust object detection.*** *We hope our COCO-O could provide a rich testbed for robustness study of object detection. The dataset will be available at* `https://github.com/alibaba/easyrobust/tree/main/benchmarks/coco_o`*.*

1. Introduction

Deep learning has achieved tremendous success in the field of computer vision. As a prerequisite, Deep Neural Networks (DNNs) rely on a rigorous assumption that train-

This research is supported in part by the National Key Research and Development Program of China under Grant No.2020AAA0140000.

ing and testing data are independent and identically distributed. This ideal hypothesis is hardly satisfied in real-world applications, where the model may encounter data with distribution drift due to environmental changes, resulting in a significant decrease in performance and posing potential security issues. To solve this problem, the robustness study [37, 57, 74, 10] of DNNs under distribution shifts has emerged in the research area of image classification.

However, most robustness researches merely focus on classification, and do not pay equal attention to other vision tasks, such as object detection. This phenomenon can be attributed to the lack of benchmark datasets. In contrast with holistic benchmarks [40, 63, 78, 38, 3] on ImageNet classification, the detection robustness benchmarks are limited. Previous work [58] benchmarks robustness using synthetic corruptions, however, it remains unclear if such simulated data can approximate real-world scenarios. Thus some other works collect images from internet to construct datasets. [16, 45, 36, 50, 98] use road scene datasets [89, 17, 44, 25] to benchmark domain generalization of detectors. Such scene-specific dataset lacks universality and domain diversity, leading to a biased assessment of robustness. For evaluation on common tasks, [95, 43] collect natural OOD images based on PASCAL VOC [21]. However, VOC is a small-scale detection task with limited number of categories, which has lagged behind the current standard evaluation protocol, *e.g.* COCO [53], LVIS [31] for detectors. We argue that more comprehensive and challenging benchmarks should be proposed to measure natural OOD robustness of modern detectors in 2020s.

In this work, we present COCO-O, a novel test dataset for COCO detection task which benchmarks robustness of object detectors under natural distribution shifts. COCO-O consists of 6,782 online-collected images belonging to 6 test domains: sketch, weather, cartoon, painting, tattoo and handmake. We compare our COCO-O with previous robust detection benchmarks in Table 1. Compared to VOC-related datasets, our COCO-O is more comprehensive with richer types of OOD shifts and larger dataset scale. COCO-O is fully compatible with the modern COCO evaluation protocol. Moreover, compared with COCO-related benchmarks, COCO-O is more challenging and can lead to 55.7% relative performance drop on a Faster R-CNN detector. By calculating the Fréchet Inception Distance (FID) [41] to clean distribution, we show our COCO-O (with FID=132) has larger distribution shifts than COCO-C [58].

Taking advantage from the proposed COCO-O, we additionally contribute extensive experiments on more than 100 modern object detectors to investigate the credibility of their reported improvements and whether they are just overfitting to the COCO test set. An overview of some key results is shown in Figure 1. Through a more precise *Effective Robustness (ER)* metric [1] which eliminates extra impact brought by the variance of ID performance, we make a frustrating observation that most classic detectors have no great progress on robustness. However, recent breakthroughs in Visual Transformers (ViTs) [19] and large-scale vision foundation models have brought new hope for OOD robustness. Especially, zero-shot detectors [48, 20] pre-trained with massive image-language pairs exhibit great effectiveness on our COCO-O. Our results inspire future research to explore training data scaling or fusing external knowledge of human language to achieve more robust detection. Besides, we analyse how OOD robustness is influenced by detector architecture, augmentation, pre-training, *etc*. Some interesting findings are revealed, which can be summarized as: 1) Compared with the detection head or neck, backbone is the most important part for detector's robustness. Our empirical study shows scaling up backbone model or using advanced backbone design, *e.g.* ResNeXt [87], Swin [56] can bring greater robustness gains. 2) Detection transformers [7, 99] are more vulnerable than traditional non-end-to-end detectors under natural distribution shifts. Note that it is different from the previous experience [61, 2, 59] in classification tasks, where ViTs are regarded as a robust learner. We hope our COCO-O could provide a rich testbed for robustness study of object detection, and we appeal that detection algorithms proposed in future should also evaluate their OOD generalization ability.

Datasets	OOD Types	Class Num.	Natural Images	Performance Drop (%)	FID
VOC Scale Robustness Benchmarks					
OOD-CV [95]	**5**	10	2,632	↓ 26.6%	91
Clipart1k [43]	1	**20**	1,000	↓ 59.8%	**148**
Watercolor2k [43]	1	6	2,000	↓ 39.1%	113
Comic2k [43]	1	6	2,000	↓ **71.5%**	147
COCO Scale Robustness Benchmarks					
COCO-C [58]	**15**	80	0*	↓ 49.8%	41
COCO-O (Ours)	6	80	**6,782**	↓ **55.7%**	**132**

Table 1: Overview of existing general robust detection benchmark. *Note that COCO-C has only synthetic images.

Our contributions are summarized below:

- We propose COCO-O, the first COCO-scale test dataset for evaluating the robustness of detectors under natural distribution shifts.
- We benchmark the robustness of 100+ modern detectors and provide a thorough comparison in Section 4.
- Through analysing the impact factors of detector's robustness. We reveal some findings in Section 4.1 that can help to develop more robust detection algorithms.

2. Related Work

Object Detection Object detection task aims at classifying and localizing the objects in an image. Traditional de-

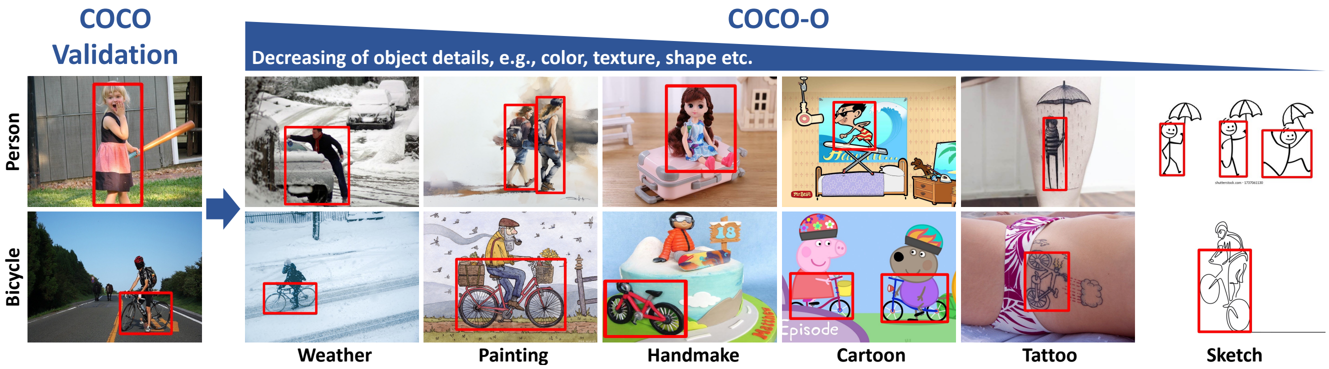

Figure 2: Visualization of our COCO-O. We adopt 6 domains, *i.e.* weather, painting, handmake, cartoon, tattoo, sketch. The domains are ordered by decreasing details of their contained objects. Each domain presents an abstract levels of the objects.

tection methods can be divided into two categories: single-stage detectors [55, 52, 96, 73, 75, 47] and two-stage detectors [29, 28, 51, 67, 34, 6, 72]. There is also a research branch extended from single-stage detection, which utilizes lightweight design for real-time detection [64, 65, 66, 24, 5]. Recently, the success of transformer models [76] in computer vision has led to the widespread use of transformer-based architectures [7, 99] in object detection, which replace the complex manual anchor design and non-maximum suppression procedure of previous methods. In contrast to previous closed-set detection methods, open-set object detection [48, 83] has also emerged as a mainstream research topic. By leveraging large-scale pre-training on image-language data, these methods can localize any object with only a given text description.

Robust Detection Benchmarks Object detectors can fail under various conditions such as blur, occlusion, weather changes, deformation, *etc*. To study the impact of these conditions, previous studies have constructed benchmark datasets via synthetic or online collected images. For instance, COCO-C [58] adds synthetic corruptions such as JPEG compression, gaussian noise to COCO [53] test set. In this work, we do not consider image synthesis technique for benchmark construction since it has two inherent drawbacks: 1) it is hard to synthesize objects with pose or shape changes; 2) noise or artifacts will be introduced in synthesis process, leading to the deviation from natural image distribution. Another line of work proposes benchmarks for specific problems, *e.g.*, environmental changes in autonomous driving [44], object variation in aerial imagery [85], *etc*. However we believe a general robustness benchmark should be built on some common detection tasks such as COCO or VOC [21]. [95, 43] collected OOD images based on VOC, but their task scale and domain diversity are still limited. To the best of our knowledge, COCO-O is the first natural OOD benchmark for COCO task. It has larger test set with more object categories and OOD types.

Robust Detectors Training robust detectors generalizing to unknown domain has been extensively studied in the literature. Most domain adaptation based methods [16, 43, 36, 45, 98, 46] require target domain data for adapting detectors. However, for online-deployed detectors, the test domain is open and indeterminate. [94] first studies the domain generalization problem in object detection. They eliminate the dependence within RoI features to improve the generalization of detection models under distribution shifts. To make detectors robust to image corruptions, [58] proposes to transfer styles of training images for data augmentation. Further, Det-AdvProp [15] follows AdvProp [86] to train detectors on clean and adversarial examples using two-way batchnorm. Such adversarially learned feature makes detector less sensitive to unknown distortions. Another branch of works [18, 91, 12] aim at improving the adversarial robustness of detectors. Due to the well-known adversarial robustness and accuracy trade-off [92], these methods suffer from a drop of clean mAP. Meanwhile, our experiment in Section 4.2 suggests their OOD generalization ability has also decreased.

3. COCO-O

3.1. Choice of Test Domains

As depicted in Figure 2, in COCO-O, the selection of test domains is carried out by first dividing the objects into six abstract levels based on decreasing levels of details such as color, texture, and shape. For each abstract level, an appropriate domain is chosen. Most domain designs are motivated by ImageNet-R [37]. We introduce them as follows: 1) Weather contains objects in challenging weather conditions, *e.g.* rain, snow and fog. It is the easiest domain which has only appearance-based shifts and reserves most of the object details; 2) Painting includes most watercolor paintings which provide a realistic description of objects in a different image style; 3) Handmake consists of real-world human handicrafts, *e.g.*, origami, toy, sculpture, *etc*. The material of the object is changed in this domain; 4) Cartoon has images of 2D or 3D digital animation. It only preserves the rough structure and color information of the object. 5) Tattoo involves art drawing on human bodies. It can even include less image details than cartoon images, and some tattoos are black-only; 6) Sketch is considered the hardest case in COCO-O. It contains a set of line-drawing images missing texture and color. As a high-level abstraction of objects, detecting sketch objects requires more external knowledge or human priors. It should be noted that since traditional factors of small size objects, occlusion, illumination, image quality has been studied before [70, 42, 88, 13, 32], we do not adopt them as an individual test domain in COCO-O, but implicitly include them (Figure 3). For instance, cars on a rainy night have poor illumination conditions or bicycles covered by snow are seriously occluded.

3.2. Data Collection

We collect COCO-O images by searching the internet using a combinations of OOD scenario keywords and object categories from COCO. For instance, “cartoon + dog” aims to gather a collection of animation dog images. Generally, most images searched by “cartoon + dog” are iconic [4], where single high quality object is centered in the image and can be localized easily. To obtain more non-iconic images, we follow the way used in COCO [53] and add more object categories into keywords combinations, such as “cartoon + dog + car”. We manually control the number of images retrieved by each keyword combination to ensure a balance among categories. For combinations that return only a few images, such as “fog + bowl + tv”, we try to use multiple search engines for collecting more images. A list of the search queries is provided in Supplementary G.

3.3. Dataset Statistics

The annotated COCO-O has a total of 6,782 images and 26,624 labelled bounding boxes. It includes six test domains: Sketch (992 images, 3,707 objects), Weather (961 images, 4,509 objects), Cartoon (1,996 images, 8,774 objects), Painting (954 images, 4,879 objects), Tattoo (918 images, 1,489 objects) and Handmake (961 images, 3,266 objects). Original 80 COCO categories are adopted in our dataset. We additionally visualize the number of instances per image and class distribution in Supplementary A. Compared to COCO, COCO-O has roughly 5% more images with only one single object, which may introduce potential gaps. However, the analysis in Supplementary A has demonstrated that the performance change brought by more iconic images can be negligible.

3.4. Potential Difficulties in COCO-O

COCO-O is a challenging benchmark as it not only has large distribution shifts, but also contains some potential hard cases frequently encountered in detection tasks. We provide visualizations of these potential challenges in Figure 3, most of which are specific to a particular test domain. For instance, weather changes bring additional difficulties for the detector, such as object occlusion caused by snow cover, poor illumination conditions in fog. In tattoo domain, geometric deformation caused by human bodies can pose an extra challenge. These potential challenges further enhance the difficulty of COCO-O. A robust detector should not only maintain consistent performance under distribution shifts, but also be able to tackle these potential challenges.

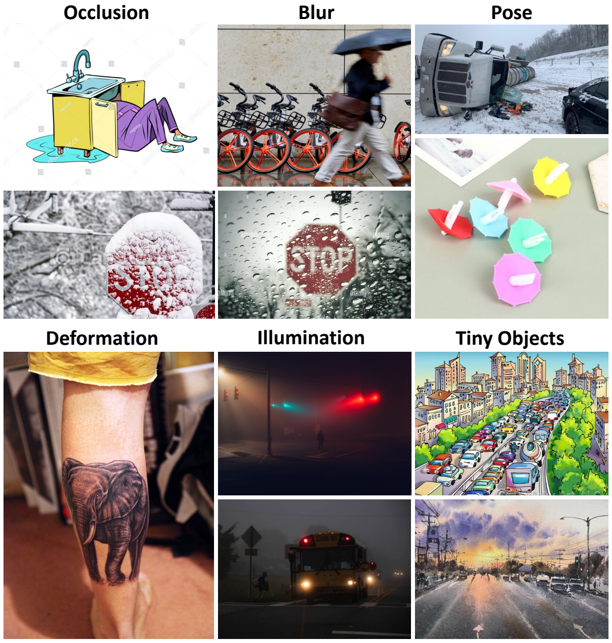

Figure 3: Some potential difficulties in COCO-O dataset.

3.5. Other Applicable Tasks

In addition to evaluating the OOD robustness of detectors, our dataset has potential applications for other detection-related tasks. Given the domain labels provided in COCO-O, one direct application is domain adaptation [60] or generalization [94] research. Few-shot learning [9] and incremental learning [81] are alternative solutions for OOD problem. Specifically, COCO-O can also be leveraged for cross-domain few-shot detection [23, 30], where only a limited number of samples are available for training detectors. Overall, our dataset offers abundant resources and challenges, and facilitates the advancement of various tasks in the object detection field.

4. Experiments

In Section 4.1, we study how some basic components, such as detection architecture, augmentation and pretraining effect on the OOD robustness of traditional detection algorithms. We utilize COCO-O to examine previously proposed robust and SOTA detectors in Section 4.2 and 4.3.

			COCO	COCO-O (mAP)							Effective
			mAP	Sketch	Weather	Cartoon	Painting	Tattoo	Handmake	Avg.	Robustness
Faster R-CNN	BackBone	RN-50 [35]	37.4	9.8	25.1	13.9	23.3	10.1	16.3	16.4	-0.41
		RN-101 [35]	39.4	10.7	30.3	16.1	27.6	11.3	18.6	19.1	+1.37
		RX-101-32x4d [87]	41.2	12.1	31.6	16.3	28.9	11.3	19.7	20.0	+1.44
		Swin-T [56]	42.5	12.8	34.2	17.3	31.1	10.6	20.4	21.1	+1.94
		PVTv2-B2 [80]	45.6	16.3	37.8	22.1	35.9	13.0	24.0	24.9	+4.33
	Neck	FPN [51]	37.4	9.8	25.1	13.9	23.3	10.1	16.3	16.4	-0.41
		PAFPN [54]	37.5	9.4	26.9	13.9	23.8	10.7	15.8	16.8	-0.13
		NAS-FPN [27]	38.0	8.7	23.8	12.7	22.1	8.7	14.6	15.1	-2.00
	Det. Head	Standard	37.4	9.8	25.1	13.9	23.3	10.1	16.3	16.4	-0.41
		Cascade [6]	40.3	10.8	27.8	14.9	25.5	12.5	17.4	18.2	+0.02
		SABL [79]	39.9	10.6	27.3	15.0	25.3	11.8	18.2	18.0	+0.08
		2_Heads [84]	40.0	10.6	30.4	14.6	25.7	12.0	18.4	18.6	+0.62
		Groie [69]	38.3	9.8	28.1	14.0	24.4	10.8	16.2	17.2	-0.02
RetinaNet	BackBone	RN-50 [35]	36.5	9.8	25.9	13.8	23.7	10.4	16.0	16.6	+0.18
		RN-101 [35]	38.5	10.9	30.2	15.3	27.3	11.6	18.9	19.0	+1.71
		RX-101-32x4d [87]	39.9	12.2	32.1	16.1	28.0	11.2	19.7	19.9	+1.93
		Swin-T [56]	41.4	11.1	33.7	16.4	31.0	11.1	20.0	20.6	+1.92
		PVTv2-B2 [80]	44.6	17.6	38.9	22.0	35.1	14.2	23.2	25.2	+5.10
	Neck	FPN [51]	36.5	9.8	25.9	13.8	23.7	10.4	16.0	16.6	+0.18
		PAFPN [54]	36.7	9.5	27.0	13.5	24.5	10.9	16.2	16.9	+0.42
		NAS-FPN [27]	36.1	9.0	27.3	11.5	21.7	8.7	13.9	15.4	-0.90
	Det. Head	Standard	36.5	9.8	25.9	13.8	23.7	10.4	16.0	16.6	+0.18
		SABL [79]	37.7	9.0	26.3	13.3	24.1	11.6	16.4	16.8	-0.18
		FSAF [97]	37.4	9.4	25.2	13.5	23.2	11.9	15.9	16.5	-0.31
		FreeAnchor [93]	38.7	10.2	26.3	14.3	24.6	12.1	16.5	17.3	-0.08
FCOS [75]			36.6	9.8	25.9	13.3	24.2	10.4	16.7	16.7	+0.25
DETR [7]			42.0	9.0	30.0	12.3	23.9	11.6	15.7	17.1	-1.82
Deformable DETR [99]			44.5	10.5	30.2	15.1	26.2	10.6	18.6	18.5	-1.49

Table 2: Comparison of object detectors with different backbone, neck and head. The effective robustness lower than the linear trend in Figure 1 is highlighted by red. Number in green means the model is above the linear trend.

Experimental Setup. To study the robustness effect of detection architecture and pre-training, we adopt typical two-stage Faster R-CNN [67] and one-stage RetinaNet [52] for baselines. For analyzing the impact of data augmentation on robustness, YOLOX [24] is used as baseline. All the above experiments are implemented by mmdetection [11] and use the consistent training configuration for fair comparison. Besides, for benchmark experiments on robust and SOTA detectors, we by default adopt the optimal training settings reported in their papers. Pre-trained models are directly used for robust and SOTA detection methods who have released their official weights.

Metrics. A frequently-used metric of object detection is the mean Average Precision (mAP) which averages over Intersection over Unions (IoUs) between 50% and 95%. Similarly, we also adopt mAP to measure the robustness on each OOD case in COCO-O. The averaged mAP on 6 test domains is used as the overall performance. To exclude the impact of the linear trend of performance improvement on in- versus out-of-distribution data, we also adopt the Effective Robustness (ER) metric proposed by [1]. Given a set of classic detectors $\mathcal{F}$, we approximate the linear trend by $\beta(\mathbf{mAP}_{id}(\cdot))$, where $\mathbf{mAP}_{id}(\cdot)$ is the COCO mAP metric, β is a learnable linear function fitted using observations on $\mathcal{F}$. Thus for any detector f, the performance on COCO-O can be predicted by $\beta(\mathbf{mAP}_{id}(f))$. We use Scipy [77] for linear regression on 11 classic detectors in Figure 1. The slope is calculated as 0.45. Finally the effective robustness of a detector f can be defined as:

$$\mathbf{ER}(f) = \mathbf{mAP}_{ood}(f) - 0.45 \times \mathbf{mAP}_{id}(f), \qquad (1)$$

where $\mathbf{mAP}_{ood}(\cdot)$ is our COCO-O mAP metric.

4.1. Analysis of Traditional Detectors

Robustness vs. Detection Architecture. Classic object detectors consist of three components: backbone, neck and head, each of which has a distinct role in feature extraction, feature map fusion, object localization and classification. In order to investigate the impact of each component on the overall robustness, we adopt default setting with ResNet-50 [35] backbone, FPN [51] neck, standard head for baseline and modify each part using some advanced designs. For backbone, we compare three different architectures: ResNet series [35], ViT-based Swin-T [56] and PVTv2-B2 [80]. For neck, FPN [51] and its two variants: PAFPN [54], NAS-FPN [27] are compared. For detection head, we analyse 3 and 4 advanced head designs for single- and two-stage detectors respectively. The results are reported in Table 2. Surprisingly, it suggests that advanced detection architectures with higher clean COCO mAP do not imply better robustness. Some methods, such as NAS-FPN [27], FSAF [97] even make the model more fragile under natural distribution shifts. The best architectures of neck and head on Faster R-CNN are PAFPN [54] and 2_Heads [84], which yield -0.13 and +0.62 effective robustness. Such improvement is marginal, indicating that advanced techniques on neck and head have a limited effect on robustness. In contrast, the backbone plays a more important role. Simply replacing ResNet50 [35] with PVTv2-B2 [80] can achieve +8.5 and +8.6 COCO-O mAP, +4.33 and +5.10 effective robustness on Faster R-CNN and RetinaNet respectively. Motivated by this phenomenon, for the first time we emphasize the vital role of feature extractor in detectors to enhance OOD robustness.

In addition to classic detectors, some works have innovated the entire detection framework by modeling localization with point prediction [47] or self-attention [7]. It is still lacking sufficient robustness analysis on these methods. We evaluate the OOD robustness of three novel frameworks, namely FCOS [75], DETR [7], Deformable DETR [99] on COCO-O. However, the results of Table 2 show that most frameworks do not bring a great promotion of effective robustness. An interesting phenomenon is that detection transformers have the worst OOD generalization ability, contrary to the general conclusion in the field of image classification that transformers can enhance OOD robustness [61, 2, 59].

Robustness vs. Augmentations. Data augmentation is a widely used technique for enhancing generalization. Especially in object detection, where heavy augmentations have been a crucial factor for the success of YOLO series models. To study its effect on OOD robustness, we adopt YOLOX-S [24], one of the strongest detectors which adopts diverse augmentations including MixUp [90], ColorJitter, Mosaic [5], Random Affine, *etc*. Then we re-train the detector by iteratively deleting one augmentation to observe the variance of robustness. In Figure 4, it can be suggested that all used augmentations contribute to the enhanced robustness. Among them, MixUp plays a principal role for OOD robustness. Removing it causes a significant drop in mAP on COCO-O from 19.8 to 17.7, as well as a decrease in effective robustness. This implies that, in addition to classification tasks, MixUp can also help for domain generalization on object detection. By comparison, ColorJitter is the least effective augmentation, it merely promotes 0.06 ER and 0.3 mAP on COCO-O.

Robustness vs. Pre-training. "Pre-training and fine-tuning" is still the de facto paradigm for object detection task. Although previous study [39] have demonstrated the efficacy of pre-training for constructing reliable models, they have not considered object detection tasks. To build a detector, there will be multiple ways that 1) training from scratch without any pre-training data; 2) using pre-trained checkpoint to initialize the backbone; 3) leveraging datasets like Object365 [71] to pre-train the overall detector and then fine-tuning. We list and compare the most pre-training settings in Table 3. For detectors trained from scratch, we employ a longer 6× schedules adviced by [33]. Besides, all other compared methods are using ResNet-50 backbone and same training hyper-parameters for ensuring fairness. The results suggest that the detector can becomes more robust when ImageNet-1K is used for pre-training. However the obtained robustness depends on an appropriate pre-training method. For example, we compare two self-supervised approaches for backbone pre-training: SwAV [8] and MoCov2 [14], and find that SwAV pre-trained detector performs poorly on OOD robustness. Instead, an elaborate supervised pre-training procedure [82] can even beat SwAV and MoCov2. Moreover, using more data for pre-training can further improve the robustness. It is foreseeable, but we need to remark that big data policy works in both backbone and detector training stage. On Fatser R-CNN, using backbone pre-trained on ImageNet-21K [68] and detector pre-trained on Object365 [71] can get +1.20 and +0.32 effective robustness respectively.

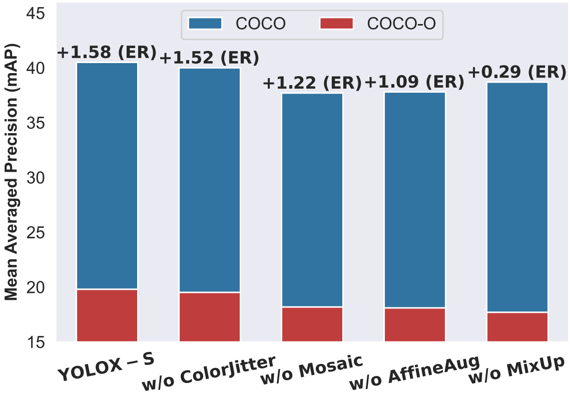

Figure 4: The robustness effect of different used augmentations in YOLOX-S detector.

Others. In addition to the techniques mentioned above, there are many others tricks for improving detectors during training. In this study, we investigate the impact of two common practices, multi-scale training and longer-epoch training, on the robustness of detectors. The results are reported in Figure 5. Interestingly, we do not find any trend of robustness increasing with longer training times. It suggests the improvement of clean mAP brought by longer epochs of training may indicate over-fitting on the COCO test set. On the other hand, training with auxiliary multi-scale inputs can slightly improve the effective robustness of Faster R-CNN and RetinaNet by +0.26 and +0.43 respectively.

	BackBone Pre-training		Detector	COCO	COCO-O (mAP)							Effective
	Method	Data	Pre-training	mAP	Sketch	Weather	Cartoon	Painting	Tattoo	Handmake	Avg.	Robustness
Faster R-CNN	-	-	-	37.2	10.6	23.3	14.6	23.1	9.2	16.0	16.1	-0.65
	Sup	IN-1K	-	37.4	9.8	25.1	13.9	23.3	10.1	16.3	16.4	-0.41
	Sup_RSB [82]	IN-1K	-	40.8	11.0	29.8	14.5	27.3	12.2	18.9	19.0	+0.59
	SwAV [8]	IN-1K	-	38.6	8.3	26.5	11.4	22.3	8.8	15.7	15.5	-1.9
	MoCov2 [14]	IN-1K	-	37.5	9.8	26.8	13.8	22.8	10.1	16.6	16.7	-0.23
	Sup	IN-21K [68]	-	38.9	9.9	30.1	14.8	26.0	11.8	19.6	18.7	+1.20
	Sup	IN-1K	Obj365 [71]	42.1	11.5	30.1	16.4	27.0	11.3	19.3	19.3	+0.32
RetinaNet	Sup	IN-1K	-	36.5	9.8	25.9	13.8	23.7	10.4	16.0	16.6	+0.18
	Sup_RSB [82]	IN-1K	-	39.0	11.0	28.4	14.5	25.3	11.4	17.7	18.1	+0.50
	SwAV [8]	IN-1K	-	38.7	8.5	28.4	11.3	22.2	9.4	16.0	16.0	-1.45
	MoCov2 [14]	IN-1K	-	36.2	11.2	26.6	13.1	23.6	10.0	15.1	16.6	+0.31
	Sup	IN-21K [68]	-	38.2	9.8	28.9	14.3	24.8	11.6	18.9	18.1	+0.86
	Sup	IN-1K	Obj365 [71]	41.0	11.9	28.7	16.3	25.5	10.5	19.1	18.7	+0.22

Table 3: The reported COCO-O performance of detectors with different pre-training methods.

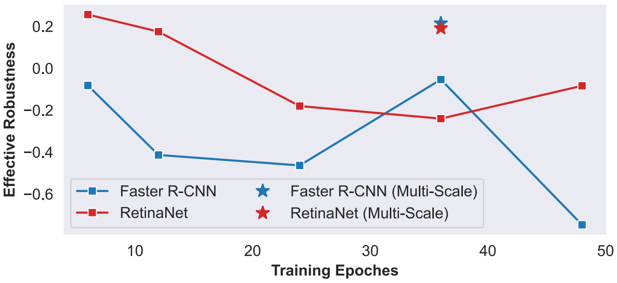

Figure 5: Multi-scale and longer training vs. robustness

4.2. Results on Robust Detectors

There have been several works proposed to train robust object detectors. For resisting adversarial attacks, RobustDet [18] trains an adversarially robust object detector on clean and adversarial images via adversarially-aware convolution. Det-AdvProp [15] also uses adversarial images for training but it aims at better clean performance through a separate batch norm design similar to AdvProp. [58] has shown that style transfer augmentation can significantly improve corruption robustness. However, most of these methods do not evaluate their models under natural distribution shifts due to the lack of OOD benchmark datasets on COCO. Take advantage from our COCO-O, in this work we present an assessment of their natural OOD robustness in Table 4. Both Stylized-Aug and Det-AdvProp can improve performance on COCO-O effectively, but it should be note that the former method sacrifices 0.7 clean mAP while the latter even achieves +0.6. As an adversarial defense method, RobustDet has a greater impact on clean performance, meanwhile, such adversarially robust detectors have low generalization ability on our COCO-O.

4.3. Results on SOTA Detectors

To investigate whether the latest developments in the field of object detection have made progress in closing the OOD distortion robustness gap, we collected 53 powerful detectors based on the COCO Leaderboard and evaluated their performance on our COCO-O. The core results are shown in Figure 6. Currently, EVA [22] stands out from thousands of open-sourced detectors and holds the first place on COCO-O. EVA is a billion-scale vision foundation model, which shows model&data scaling is still the most direct and effective way towards OOD generalization in object detection. However, without the help of additional data, all detectors are facing a giant decline on COCO-O (shown in grey bars in Figure 6). A frustrate situation is that the most effectively robust detector ViTDet-H [49], which is trained on standard data (COCO, ImageNet-1K), achieves merely 7.885 and even cannot enter the top@10 of the ranking list. This finding suggests that most recent progression may be due to the use of more training data.

Baseline Models	Robust Methods	COCO mAP	COCO-O mAP	Effective Robustness
Faster R-CNN	-	37.4	16.4	-0.41
	Stylized-Aug	36.1	20.4	+4.12
EfficientDet-D1	-	40.2	22.0	+3.86
	Det-AdvProp	40.8	22.9	+4.56
SSD	-	42.0	18.7	-0.18
	RobustDet	31.0	13.7	-0.28

Table 4: The performance of robust detectors on COCO-O.

5. Discussion

Large-scale foundation models have made the greatest progress in robust object detection. Since CLIP [62] has been firstly verified its success on comprehensive vision benchmarks, large-scale foundation models gradually become the mainstream of visual research. [26] have shown classification models trained on billion-scale data achieve significant robustness under distribution shifts, and non-trivial progress on closing the gap between human and machine vision. Our paper discovers a similar phenomenon on object detection. Figure 6 shows large-scale pre-trained detectors have made considerable progress in OOD robustness. However, we must remain cautious as this success

https://paperswithcode.com/sota/object-detection-on-coco

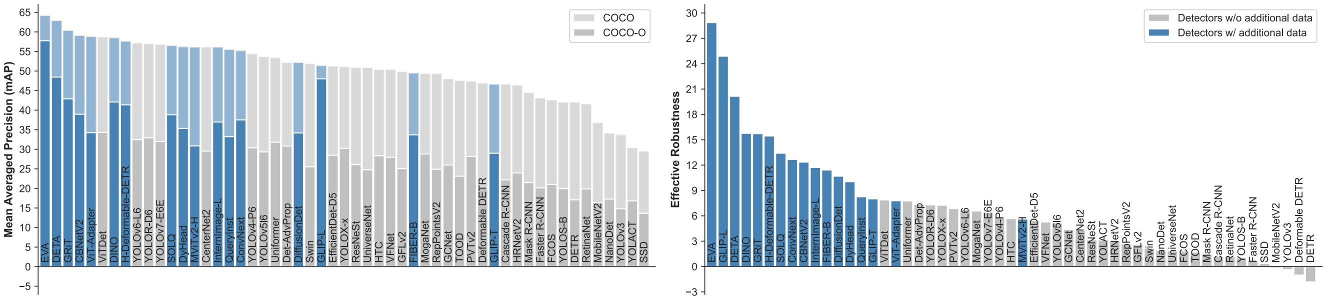

Figure 6: The reported COCO-O mAP (left) and effective robustness (right) on 53 SOTA detectors.

may be attributed to the fact that the detector has seen OOD data during training. For the research community, it is still more meaningful to focus on innovating robust detection algorithms rather than solely relying on larger training datasets.

Why DETRs underperform traditional detectors on robustness? In our results, the DETRs [7] are shown to be vulnerable. To achieve end-to-end detection, DETR introduces two major modifications: 1) replacing the conventional one-to-many label assignment with one-to-one Hungarian matching, and 2) learning a set of object queries for localization without any prior anchor information. We suspect that the label assignment rules or localization queries learned by DETR may heavily rely on the training data, which is unfavorable for generalization. Instead, human-designed priors could have stronger generalization and interpretability. Till now, the exact reason for DETR's poor robustness remains unsolved. Further research is needed to study this problem.

Analysis on different test domains. By using COCO-O, we can study the influence of each type of natural distribution shift. We adopt detectors in Figure 1, and evaluate them on each test domain of COCO-O. The results are shown in Figure 7. Same with the order in Figure 2, sketch and tattoo objects are the hardest to detect, as they lost important feature *e.g.* colors for detection. In contrast, appearance-based shifts such as weather are relatively easier to handle. Detectors have the lowest performance variance on tattoo objects, most of them have the mAP below 13. It suggests the necessity of designing specialized detectors for tattoo images.

Our difference with COCO-C. People may concern about the necessity of our COCO-O since previous COCO-C [58] has been taken as a generally accepted robustness metric for object detection. Here we must reaffirm our difference and superiority with COCO-C: 1) COCO-C is a synthetic dataset and has limitations as discussed in Section 2. In contrast, COCO-O includes realistic images, which are more representative of the real-world scenarios; 2) The purposes are different. COCO-C measures robustness under image corruptions, while corruptions may not cover all real-world OOD shifts such as artificial creation in different styles and forms. COCO-O contains more diverse realistic images that cover such OOD shifts; 3) As demonstrated in Supplementary C, our COCO-O has a lower correlation with COCO mAP. It implies that our proposed COCO-O, which covers robustness evaluation scenarios that are not considered in COCO validation set, can be a meaningful metric complement with COCO mAP to reflect the overall performance.

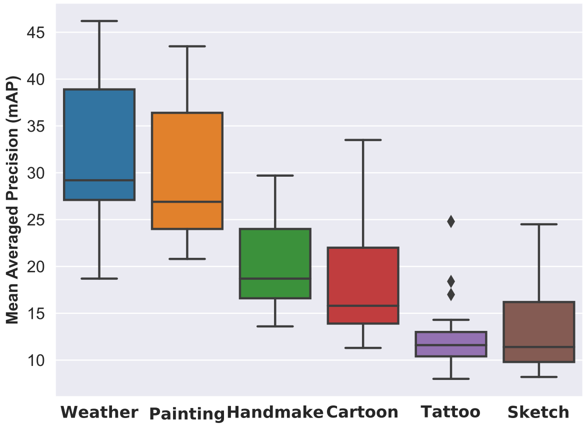

Figure 7: Robustness level on six test domains of COCO-O. The statistical result is counted from detectors in Figure 1.

6. Conclusion

In this paper, we propose a novel dataset called COCO-O to benchmark object detection under natural distribution shifts. With a thorough diagnosis of more than 100 modern object detectors, we demonstrate that detecting objects with OOD shifts remains a challenge and requires further attention from the research community. Additionally, we empirically investigate how OOD robustness is influenced by various factors, including detector architecture, augmentation, pre-training, *etc*. With our COCO-O dataset, innovative techniques can be developed to enhance the OOD robustness of existing detection algorithms, which will be the focus of our future work.

References

[1] Anders Johan Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *Transactions on Machine Learning Research.*

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

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

[4] Tamara L Berg and Alexander C Berg. Finding iconic images. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2009.

[5] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[6] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[9] Da Chen, Yuefeng Chen, Yuhong Li, Feng Mao, Yuan He, and Hui Xue. Self-supervised learning for few-shot image classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1745–1749. IEEE, 2021.

[10] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 458–467, 2021.

[11] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[12] Pin-Chun Chen, Bo-Han Kung, and Jun-Cheng Chen. Class-aware robust adversarial training for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10420–10429, 2021.

[13] Winston Chen and Tejas Shah. Exploring low-light object detection techniques. *arXiv preprint arXiv:2107.14382*, 2021.

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

[15] Xiangning Chen, Cihang Xie, Mingxing Tan, Li Zhang, Cho-Jui Hsieh, and Boqing Gong. Robust and accurate object detection via adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16622–16631, 2021.

[16] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.

[17] 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.

[18] Ziyi Dong, Pengxu Wei, and Liang Lin. Adversarially-aware robust object detector. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 297–313. Springer, 2022.

[19] 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 *International Conference on Learning Representations*.

[20] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *Advances in Neural Information Processing Systems*.

[21] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015.

[22] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

[23] Yipeng Gao, Lingxiao Yang, Yunmu Huang, Song Xie, Shiyong Li, and Wei-Shi Zheng. Acrofod: An adaptive method for cross-domain few-shot object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 673–690. Springer, 2022.

[24] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[25] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[26] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A

Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

[27] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7036–7045, 2019.

[28] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[29] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[30] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 124–141. Springer, 2020.

[31] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[32] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part V 28*, pages 387–395. Springer, 2021.

[33] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019.

[34] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

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

[36] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6668–6677, 2019.

[37] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[38] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.

[39] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019.

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

[41] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[42] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 951–959, 2017.

[43] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018.

[44] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 746–753. IEEE, 2017.

[45] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 480–490, 2019.

[46] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019.

[47] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[48] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[49] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 280–296. Springer, 2022.

[50] Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8771–8780, 2021.

[51] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[52] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[53] 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.

[54] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

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

[56] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[57] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12042–12051, 2022.

[58] 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.

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

[60] Poojan Oza, Vishwanath A Sindagi, Vibashan VS, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. *arXiv preprint arXiv:2105.13502*, 2021.

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

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

[63] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[64] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[65] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[66] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[67] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[68] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

[69] Leonardo Rossi, Akbar Karimi, and Andrea Prati. A novel region of interest extraction layer for instance segmentation. In *2020 25th international conference on pattern recognition (ICPR)*, pages 2203–2209. IEEE, 2021.

[70] Kaziwa Saleh, Sándor Szénási, and Zoltán Vámossy. Occlusion handling in generic object detection: A review. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000477–000484. IEEE, 2021.

[71] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.

[72] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021.

[73] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[74] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[75] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[77] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[78] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

[79] Jiaqi Wang, Wenwei Zhang, Yuhang Cao, Kai Chen, Jiangmiao Pang, Tao Gong, Jianping Shi, Chen Change Loy, and Dahua Lin. Side-aware boundary localization for more precise object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 403–419. Springer, 2020.

[80] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[81] Kun Wei, Da Chen, Yuhong Li, Xu Yang, Cheng Deng, and Dacheng Tao. Incremental embedding learning with disentangled representation translation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[82] Ross Wightman, Hugo Touvron, and Herve Jegou. Resnet strikes back: An improved training procedure in timm. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*.

[83] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.

[84] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10186–10195, 2020.

[85] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.

[86] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020.

[87] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[88] Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Rfla: Gaussian receptive field based label assignment for tiny object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 526–543. Springer, 2022.

[89] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[90] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

[91] Haichao Zhang and Jianyu Wang. Towards adversarially robust object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 421–430, 2019.

[92] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[93] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. *Advances in neural information processing systems*, 32, 2019.

[94] Xingxuan Zhang, Zekai Xu, Renzhe Xu, Jiashuo Liu, Peng Cui, Weitao Wan, Chong Sun, and Chen Li. Towards domain generalization in object detection. *arXiv preprint arXiv:2203.14387*, 2022.

[95] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 163–180. Springer, 2022.

[96] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[97] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 840–849, 2019.

[98] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019.

[99] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*.