DETECTBENCH: AN OBJECT DETECTION BENCHMARK FOR OOD GENERALIZATION ALGORITHMS

Anonymous authors
Paper under double-blind review

ABSTRACT

The consensus about machine learning tasks, such as object detection, is still the test data are drawn from the same distribution as the training data, which is known as IID (Independent and Identically Distributed). However, it can not avoid being confronted with OOD (Out-of-Distribution) scenarios in real practice. It is risky to apply an object detection algorithm without figuring out its OOD generalization performance. On the other hand, a plethora of OOD generalization algorithms has been proposed to amortize the gap between the in-house and open-world performances of machine learning systems. However, their effectiveness was only demonstrated in the image classification tasks. It is still an opening question of how these algorithms perform on more complex tasks. In this paper, we first specify the setting of OOD-OD (OOD generalization object detection). Then, we propose DetectBench consisting of four OOD-OD benchmark datasets to evaluate various object detection and OOD generalization algorithms. From extensive experiments on DetectBench, we find that existing OOD generalization algorithms fail dramatically when applied to the more complex object detection tasks. This raises questions over the current progress on a large number of these algorithms and whether they can be effective in practice beyond simple toy examples. For future work, we sincerely hope that DetectBench can serve as a foothold for OOD-OD research.

1 INTRODUCTION

Modern object detection methods (Liu et al., 2021; Huang et al., 2019; Pang et al., 2019; Wu et al., 2019; Zhang et al., 2020a; Sun et al., 2020; Zhu et al., 2021; Ge et al., 2021) have achieved many progresses on various applications, such as autonomous driving and industrial defect detection. Tremendous efforts have been devoted to improving an object detector's performance on standard datasets, such as MS-COCO (Lin et al., 2014). While these efforts have seen impacts on industry (Redmon et al., 2016; Redmon & Farhadi, 2017; 2018; Bochkovskiy et al., 2020; Ge et al., 2021), the improvements are becoming marginal recently and most achievements are accompanied by an inherent assumption, *i.e.* , the training data and the test data are IID (Independent and Identically Distributed). However, this assumption is unlikely to hold in real-world scenarios. For example, an autonomous system suffers from different environmental conditions (Dai & Gool, 2018; Volk et al., 2019); a medical system fails to work consistently among hospitals when data are collected from different equipment (de Castro et al., 2019; Albadawy et al., 2018; Perone et al., 2019). As a consequence, models trained on IID dataset are susceptible to a subtle disturbance in test data distribution (Out-of-Distribution) and fail to generalize to real scenarios (Torralba & Efros, 2011). Previous research devoted to encountering this train-test discrepancy can be summarized as either *"less complex"* or *"complex but not general"*. From the first perspective, a plethora of Domain Generalization (DG) algorithms (Arjovsky et al., 2019; Ahuja et al., 2021; Li et al., 2018b; Sun & Saenko, 2016; Xu et al., 2020c; Yan et al., 2020; Krueger et al., 2021; Pezeshki et al., 2020; Parascandolo et al., 2021; Koyama & Yamaguchi, 2021; Huang et al., 2020; Sagawa et al., 2019) concentrate on improving OOD generalization ability. But they are simply evaluated on the image classification. The effectiveness is unknown when applied to the complex task (object detection). On the other perspective, numerous Domain Adaption (DA) algorithms (Chen et al., 2018; He & Zhang, 2020; Rodriguez & Mikolajczyk, 2019; Xu et al., 2020a; Su et al., 2020; Xu et al., 2020b; Soviany et al., 2019; Deng et al., 2020; Chen et al., 2021) aim to build an optimal object detector that can be generalized into

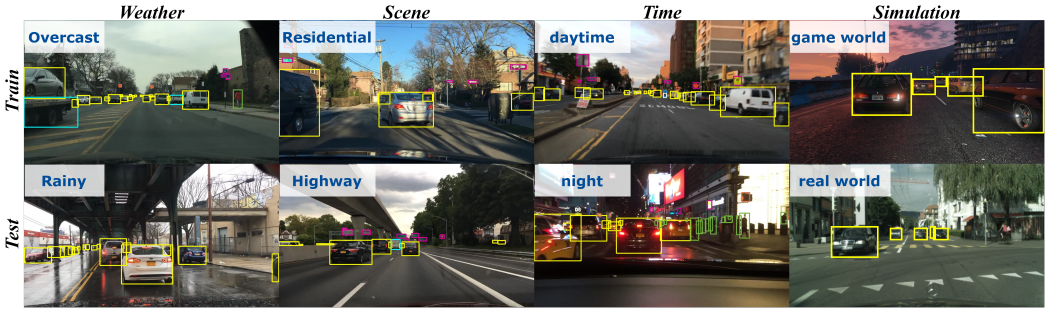


Figure 1: The setting illustration of out-of-distribution generalization object detection (OOD-OD).

a pre-specified target domain. However, it is hard to ensure performance consistency when dealing with unseen and infinite real-world domains. In this paper, we focus on **OOD generalization object detection (OOD-OD)** which aims at training detectors to generalize to the testing data drawn from an unseen distribution distinct from the training distribution. See Table 1 for more details and we provide the theoretical definition for OOD-OD in Appendix A.1.

In this work, we propose DetectBench, in which four OOD-OD benchmarks are constructed with existing datasets, including BDD100K (Yu et al., 2018), Cityscapes (Cordts et al., 2016) and Sim10K (Johnson-Roberson et al., 2016). As revealed by (Ye et al., 2021), data distribution shifts on classification datasets are dominated by correlation shift and diversity shift. We test whether a similar phenomenon also exists on detection datasets and we construct a synthetic dataset named CtrlShift to quantitatively analyze generalization ability over the two kinds of distribution shifts of OOD-OD, respectively. With the above benchmark datasets, numerous experiments are conducted with detectors ranging from one(two)-stage to transformer-based and diverse OOD generalization algorithms carefully implemented on popular detectors (Ren et al., 2015; Lin et al., 2017b).

Section 2 reviews the related work dispersed in different research areas. Section 3 provides clarification of different techniques and tasks with similar names. Section 4 introduces the implementation details of DetectBench, including the datasets, algorithms, and model selection methods. Finally, Section 5 discusses the experiment results on DetectBench and offers insightful recommendations for future work. Our main contributions can be summarized as followed:

1. We propose DetectBench, the first OOD generalization benchmark for object detection algorithms. Based on the extensive experiment results, we arrive at a surprising conclusion: *The enormous achievements in IID object detection are marginal on OOD generalization object detection, and the OOD generalization improvements on classification are hard to generalize to more complex (object detection).*
2. In DetectBench, we propose a Sim2real benchmark for OOD generalization object detection analysis which measures the possibility of training models with low-cost simulated data to generalize well on real scenarios.
3. To further analyze the generalization ability under the different types of shifts, we construct a synthetic dataset with designed shifts, namely CtrlShift. The synthetic dataset can systematically measure the OOD generalization algorithms' performances under different types of distribution shifts.
4. From Benchmark results and analysis, we recommend future research to clearly investigate the diversity shift and the correlation shift on OOD generalization object detection before designing algorithms and then evaluate them comprehensively on the two-dimensional shift using CtrlShift.

2 Related Work

2.1 Object Detection

The task of object detection aims at classifying and localizing the objects in an image based on the assumption that test data are drawn from the same distribution as training data. And the modern deep

Table 1: Comparison between OOD generalization object detection and other tasks. X^i, Y^i indicate the data drawn from distribution i and Y_1, Y_2 represent category labels and bounding box labels respectively.

Setup	Task	Training inputs	Test inputs	Outputs
Unsupervised learning	classify / detect	X^1	X^1	Y^1
Supervised learning	classify / detect	X^1, Y^1	X^1	Y^1
Semi-supervised learning	classify / detect	$X^1, (Y^1)'$	X^1	Y^1
Transfer learning	classify / detect	$X^{1,\ldots,d}, X^{d+1}, Y^{d+1}$	X^{d+1}	Y^{d+1}
Domain generalization	classify	$X^{1,\ldots,d}, Y^{1,\ldots,d}$	X^{d+1}	Y_1^{d+1}
Domain adaption	detect	$X^{1,\ldots,d}, Y^{1,\ldots,d}, X^{d+1}$	X^{d+1}	$Y_{1,2}^{d+1}$
OOD-OD	detect	$X^{1,\ldots,d}, Y^{1,\ldots,d}$	$X^{d+1,\ldots}$	$Y_{1,2}^{d+1,\ldots}$

learning-based object detection models can be divided into three categories: two-stage detectors (Girshick et al., 2014; Grauman & Darrell, 2005; Girshick, 2015; Ren et al., 2015; Lin et al., 2017a; Dai et al., 2016; He et al., 2017; Qiao et al., 2021; Cai & Vasconcelos, 2019; Huang et al., 2019; Pang et al., 2019; Wu et al., 2019; Sun et al., 2020), one-stage detectors (Redmon et al., 2016; Redmon & Farhadi, 2017; 2018; Bochkovskiy et al., 2020; Ge et al., 2021; Liu et al., 2016; Lin et al., 2017b; Zhou et al., 2019; Tan et al., 2020; Law & Deng, 2018; Tian et al., 2019; Zhang et al., 2020a; Zhu et al., 2021; Liu et al., 2021) and lightweight detectors with small components (Howard et al., 2017; Sandler et al., 2018; Howard et al., 2019; Zhang et al., 2018; Ma et al., 2018; Wang et al., 2018; Iandola et al., 2016). Compared to one-stage detectors, two-stage detectors are equipped with a separate differentiable module to generate region proposals which are possible to contain objects. Lightweight detectors are usually proposed to improve real-time performance with a small and efficient network. Recently, with the enormous success of applying transformer (Vaswani et al., 2017) on computer vision, a branch of transformer-based detector (Zhu et al., 2021; Liu et al., 2021) has shaped up.

2.2 OOD Generalization

The task of OOD generalization is training on multiple datasets sampled from distinct domains and then generalizing to an unseen test domain. Models with OOD generalization ability typically have access to multiple training datasets for the same task obtained from various environments. The purpose of OOD generalization algorithms is to learn from these diverse but relevant training settings before being applied to unknown testing environments. Driven by this motivation, many algorithms have been proposed throughout these years. These algorithms can be divided into: empirical risk learning (Vapnik, 1998; Sagawa et al., 2019), invariant risk optimization (Arjovsky et al., 2019), domain adversarial learning (Ajakan et al., 2014; Li et al., 2018c; Ruan et al., 2021), meta-learning (Zhang et al., 2020b; Li et al., 2018a), kernel function (Li et al., 2018b; Sun & Saenko, 2016), gradient-based approach (Shi et al., 2021; Pezeshki et al., 2020; Bai et al., 2020; Parascandolo et al., 2021; Shahtalebi et al., 2021; Koyama & Yamaguchi, 2021; Rame et al., 2021), risk extrapolation (Krueger et al., 2021), data processing (Xu et al., 2020c; Yan et al., 2020), transfer learning (Blanchard et al., 2017; Xu & Jaakkola, 2021), information bottleneck (Ahuja et al., 2021) and self-supervised learning (Wang et al., 2020; Zhou et al., 2020).

OOD generalization for object detection is currently underexplored. Region Aware Proposal reweighTing (RAPT) (Zhang et al., 2022) aims to eliminate dependence within RoI features for domain generalization. Cyclic-Disentangled Self-Distillation (Wu & Deng, 2022) aims at disentangling domain-invariant representations. 3D-VField (Lehner et al., 2022) improves generalization on 3D object detection.

2.3 OOD Benchmark

Different domains data (Zhou et al., 2021; Wang et al., 2022) can be viewed as data drawn from different distributions and the distinct train-test domains are Out-of-Distribution. DomainBed (Gulrajani & Lopez-Paz, 2020) is a large-scale benchmark suite for reproducing domain generalization research and facilitating the implementation of new algorithms. With the experiment results of fourteen algorithms on seven datasets, the authors found that empirical risk minimization (Vapnik, 1998)

Table 2: Details of datasets used in benchmarks. **Quantity** indicates the number of images in each domain. **Total** counts the total number of training and testing domains respectively.

Benchmark	Dataset	Domain	Train	Test	Quantity	Total
Weather	BDD100K	clear	√		42690	52699
		overcast	√		10009	
		foggy		√	143	17888
		cloudy		√	5619	
		rainy		√	5808	
		snowy		√	6318	
Scene	BDD100K	city street	√		49628	69506
		highway	√		19878	
		gas station		√	34	9943
		parking lot		√	426	
		residential		√	9327	
		tunnel		√	156	
Time	BDD100K	daytime	√		41986	47791
		dawn dusk	√		5805	
		night		√	31900	31900
Sim2real	Cityscapes	simulation	√		8500	8500
	Sim10K	reality		√	3975	3975

achieves state-of-the-art performance across all datasets. OoD-Bench (Ye et al., 2021) identifies and measures two distinct kinds of distribution shifts in various OOD datasets. With tremendous empirical learning results, the authors recommend that algorithms should be comprehensively evaluated on two types of datasets dominated by correlation shift and diversity shift respectively.

3 Clarification of Tasks

Domain Randomization techniques (Tobin et al., 2017; Tremblay et al., 2018; Zakharov et al., 2019; Yue et al., 2019; Huang et al., 2021) aim at providing enough simulated domains at training data so that models are possible to generalize to real-world scenarios based on a hypothesis that with enough variability in the data simulator, the real world may appear to be a specific variation of the simulation data which exists in the training set.
OOD Detection for Object Detection (Joseph et al., 2021; Du et al., 2022a; Harakeh & Waslander, 2021; Riedlinger et al., 2021; Dhamija et al., 2020; Miller et al., 2019; Hall et al., 2020; Deepshikha et al., 2021) can be formulated as a binary classification problem which distinguishes whether the distribution of the incoming data is out of the distribution of the training data.
Open-World Object Detection (Joseph et al., 2021; Zhao et al., 2022) initially learns a model which can detect all the previously encountered categories, and incrementally updates the model when unseen classes come.
Open-Vocabulary Object Detection (Gu et al., 2021; Zareian et al., 2021; Du et al., 2022b; Bravo et al., 2022) aims to train an detector which can detect various objects in any novel categories described by arbitrary texts.

4 DetectBench: Implementation Details

4.1 Benchmarking Datasets

In DetectBench, we choose datasets to cover as many types of variations between training and test datasets as possible. Figure 2 lists some samples of the four benchmark datasets.

BDD100K (Yu et al., 2018) contains 80,000 labeled images (70,000 for training and 10,000 for validation) with ten annotated object categories, including bike, bus, car, motor, person, rider, traffic light, traffic sign, train and truck. Each image has three attribute labels which indicate the condition, including the weather, scene and time for data collection and we remove the images with an undefined attribute label. Specifically, we construct three OOD environments using the attribute labels, including **Weather**, **Scene**, and **Time**.

Figure 2: Some samples of datasets included in DetectBench. Note that BDD100K has ten categories while Sim10K and Cityscapes only use the annotated cars.

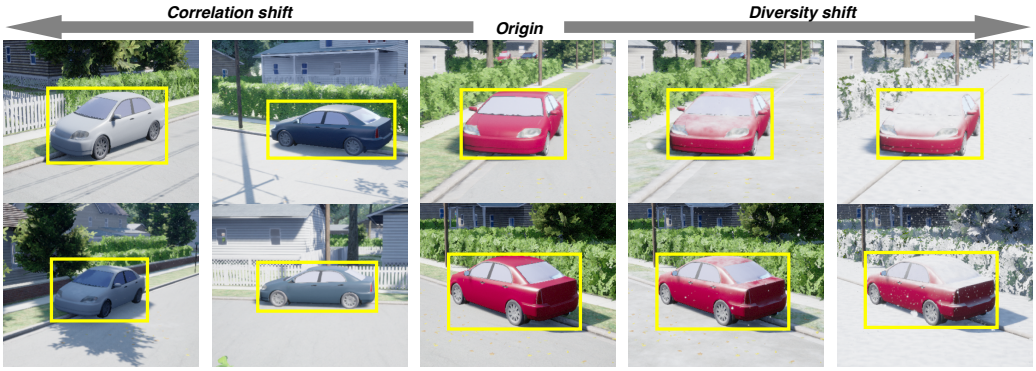


Figure 3: Some samples of CtrlShift and the illustration of the two-dimensional shift.

Sim10K (Johnson-Roberson et al., 2016) is a synthetic dataset containing 10,000 images (8,000 for training, 1,000 for validating and 1,000 for testing) with bounding box annotations for cars, which is rendered with the Grand Theft Auto V (GTA5) game engine.

Cityscapes (Cordts et al., 2016) is a large-scale database which focuses on urban street scenes. The dataset consists of around 5000 fine annotated images (2975 for training, 500 for validating and the rest for testing) with eight annotated instance categories. On DetectBench, we consider the car recognition task to construct the Sim2real benchmark for simplicity and without loss of generality. We construct the **Sim2Real** benchmark which covers the "sim2real" scenario. The simulated images of Sim10K are used for training and the real images of Cityscapes are used for testing.

CtrlShift is a synthetic dataset to analyze the two-dimension shift in OOD generalization object detection. The Airsim simulator (Shah et al., 2017) based on the high-fidelity rendering software Unreal Engine 4 is used to generate samples in CtrlShift. We totally sample over 2000 simulated images from both the rural and urban environments which contain common objects, including buildings, traffic lights, and vehicles. Moreover, every image comprises two attribute labels to construct a controllable distribution shift. One is car color which indicates the color of the car in the image,

Table 3: Experimental results of detectors performance measured by AP(%) on the four benchmarks on DetectBench. All models are implemented by mmdetection (Chen et al., 2019) and loaded pretrained weights provided by open-mmlab (Contributors, 2018). Note that @Faster in Libra R-CNN row represents applying Faster R-CNN as architecture. X-101-64x4d represents a modified ResNeXt-101 network architecture from (Xie et al., 2017), R-50 and R-101 represents ResNet backbones with 50 or 101 layers (He et al., 2016).

Detector	Backbone	Sim2real	Weather	Scene	Time	Avg.
Faster R-CNN	X-101-64x4d	52.1	35.3	36.0	26.4	37.5
RetinaNet	X-101-64x4d	49.3	34.7	35.8	24.6	36.1
Mask R-CNN	X-101-64x4d	52.9	34.8	35.8	26.8	37.6
CornetNet	Hourglass-104	27.8	29.7	30.8	22.5	27.7
YOLOv3	DarkNet-53	38.2	27.0	28.2	19.2	28.2
FCOS	X-101-64x4d	50.8	35.6	36.5	24.8	36.9
Cascade R-CNN	X-101-64x4d	53.1	36.3	36.7	24.1	37.6
MS R-CNN	X101-64x4d	52.5	35.5	35.7	25.5	37.3
Libra R-CNN@Faster	X-101-64x4d	51.4	34.9	35.5	25.5	36.8
Double-Head R-CNN	R-50	52.2	35.3	35.2	25.5	37.1
VarifocalNet	X-101-64x4d	56.1	38.7	39.0	27.4	40.3
Sparse R-CNN	R-101	54.1	36.3	36.4	28.2	38.8
DETR	R-50	37.8	23.5	24.4	15.8	25.4
Deformable DETR	R-50	53.3	35.1	35.3	26.1	37.5
Swin Transformer	Swin-B	58.4	36.6	32.0	28.9	39.0
YOLOX	YOLOX-x	51.1	33.1	34.8	29.4	37.1

the other is snow intensity which indicates the intensity of the weather snow added using Airsim's plugin. Cars are annotated in the dataset since the car is a common object in existing datasets. To construct CtrlShift dominated by correlation shift, we restrict the white cars only existing in the training set and control the quantity ratio $\rho_{correlation}$ of white car data in the training set. In a way that spuriously correlates strongly with the class label since the color white will be more relevant to the car label in the training set when the $\rho_{correlation}$ increases, which increases the correlation shifts in the dataset. And for CtrlShift dominated by diversity shift, we render snow weather effects on each training datum with a certain intensity $\rho_{diversity}$ and the increase of $\rho_{diversity}$ corresponds to a larger diversity shift. See Figure 3 for some examples. Moreover, we provide an API to generate the training set and the testing set with specific choices of $\rho_{diversity}$ and $\rho_{correlation}$.

4.2 Detection Methods for Comparisons

We choose widely-used detectors trained with the empirical risk minimization (ERM) or OOD generalization algorithms on DetectBench. The comparisons of different detectors trained with ERM on DetectBench can help answer the question that whether the progress made by recently proposed detectors is generalizable to OOD data. The benchmarks of detectors trained with proposed OOD generalization algorithms can indicate whether the OOD generalization algorithms proposed recently are still effective for object detection beyond toy image classification tasks.

Detectors. Object detection models (detectors) generally can be categorized into two genres: one-stage methods and two-stage methods. One-stage detectors predict the bounding boxes as well as the categories of the objects. Two-stage detectors predict the bounding boxes first to indicate the possible locations of objects. Then, two-stage detectors conduct classifications on the images within the bounding boxes to predict the categories of these images. Recently, with the tremendous success of transformer (Vaswani et al., 2017), transformer-based detectors become popular. We have selected one/two-stage and transformer-based algorithms ranging from 2015 to 2021 for our Object Detection OOD generalization benchmark. One-stage detectors: **RetinaNet** (Lin et al., 2017b), **CornerNet** (Law & Deng, 2018), **YOLOv3** (Redmon & Farhadi, 2018), **FCOS** (Tian et al., 2019), **VarifocalNet** (Zhang et al., 2020a), **YOLOX** (Ge et al., 2021). Two-stage detectors: **Faster R-CNN** (Ren et al., 2015), **Mask R-CNN** (He et al., 2017), **Cascade R-CNN** (Cai & Vasconcelos, 2019), Mask Scoring R-CNN (**MS R-CNN**) (Huang et al., 2019), **Libra R-CNN** (Pang et al., 2019), **Double-Head R-CNN** (Wu et al., 2019) and **Sparse R-CNN** (Sun et al., 2020). Transformer-based detectors: **DETR**

Table 4: Experimental results of domain generalization algorithms on OOD generalization object detection. All algorithms are implemented by ourselves based on Faster R-CNN (Girshick, 2015) with ResNet-50 (He et al., 2016) backbone. Hyper-parameter of each algorithm is chosen among 0.1, 1 and 10 according to the average AP over the four benchmarks.

Algorithm	Hyper-parameters	Sim2real	Weather	Scene	Time	Average
ERM	-	44.1	24.4	24.1	17.7	27.6
IB-ERM	$\lambda_{ib} = 0.1$	42.5	24.7	23.4	18.2	27.2
IRM	$\lambda_{irm} = 0.1$	42.3	24.9	23.2	17.5	27.0
MMD	$\gamma_{mmd} = 1$	42.9	24.1	23.9	18.4	27.3
CORAL	$\gamma_{mmd} = 10$	42.6	24.4	22.8	17.7	26.9
VREx	$\lambda_{vrex} = 1$	43.1	24.7	24.3	18.0	27.5
GS	$\lambda_{reg} = 10$	40.3	19.9	18.4	15.2	23.5
IGA	$\lambda_{penalty} = 1$	42.9	24.6	24.3	18.6	27.6
GroupDRO	$\eta_{groupdro} = 1$	42.8	24.5	23.6	18.1	27.3
RSC	$\lambda_{rsc} = 10$	41.0	8.7	8.8	7.0	16.4
CAD	$\lambda_{cad} = 1$	42.8	24.2	23.1	18.4	27.1
CausIRL	$\lambda_{cirl} = 1$	42.4	23.8	23.4	17.1	26.7

(Carion et al., 2020), **Deformable DETR** (Zhu et al., 2021) and **Swin Transformer** (Liu et al., 2021).

OOD generalization algorithms. We have adapted eleven algorithms from different OOD research areas to the classification branch in object detection, including Empirical Risk Minimization (**ERM**) (Vapnik, 1998) which aims to minimize the loss function overall the training domains, (**IB-ERM**) (Ahuja et al., 2021) which applies an information bottleneck constraint to address OOD generalization, Invariant Risk Minimization (**IRM**) (Arjovsky et al., 2019) which aims at estimating invariant correlations across different domains, adversarial feature learning (**MMD**) (Li et al., 2018b) which imposes Maximum Mean Discrepancy (Gretton et al., 2012) to align the distributions among different domains, correlation alignment (**CORAL**) (Sun & Saenko, 2016) which aims at matching the mean and covariance of feature distributions, Variance Risk Extrapolation (**VREx**) (Krueger et al., 2021) which performs a form of robust optimization over extrapolated domains, Gradient Starvation (**GS**) (Pezeshki et al., 2020) which derives a regularization to overcome the gradient descent phenomenon across different domains, (**IGA**) (Koyama & Yamaguchi, 2021) which uses a parametrization trick to conduct feature searching and predictor training, Group Distributionally Robust Optimization (**GroupDRO**) (Sagawa et al., 2019) which increases the importance of each domain with penalty loss, and Representation Self-Challenging (**RSC**) (Huang et al., 2020) iteratively challenges the dominant features to force the model to activate the remaining features. Optimal Representations (**CAD**) (Ruan et al., 2021) designs self-supervised objectives to obtain representations on which risk is minimal to any distribution. Invariant Causal Mechanisms (**CausIRL**) (Chevalley et al., 2022) learns the invariant features by viewing the learning process as a causal process and introduces a unifying framework.

4.3 Model Selection Methods

Model selection methods can influence the final rankings of methods to a large extent, especially in OOD generalization tasks (Gulrajani & Lopez-Paz, 2020). However, there is no consensus on what parameters selection strategy should be used in OOD generalization research for object detection. In DetectBench, we choose the models trained at the last epoch as our model selection method. This is because the testing data is inaccessible and selecting models based on the trainset's performances may lead to excessive over-fitting for current methods since there is a huge distribution gap between the training set and testing set. For future research, we strongly recommend that researchers should detail and include the model selection methods in OOD generalization object detection research.

5 Experiment Results

5.1 Benchmark Results

Table 5: Ablation study of OOD algorithms with different detectors on Sim2real. Faster R-CNN (Ren et al., 2015), RetinaNet (Lin et al., 2017b) and DETR (Carion et al., 2020) are all with ResNet-50 (He et al., 2016) backbone.

Algorithm	Hyper-parameters	Faster R-CNN	RetinaNet	DETR	Average
ERM	-	44.1	38.5	36.0	39.5
IB-ERM	$\lambda_{ib} = 0.1$	42.5	38.8	35.1	38.8
IRM	$\lambda_{irm} = 0.1$	42.3	38.4	34.7	38.5
MMD	$\gamma_{mmd} = 1$	42.9	14.0	36.2	31.0
CORAL	$\gamma_{mmd} = 10$	42.6	35.8	36.9	38.4
VREx	$\lambda_{vrex} = 1$	43.1	37.6	36.9	39.2
GS	$\lambda_{reg} = 10$	40.3	33.0	33.0	35.4
IGA	$\lambda_{penalty} = 1$	42.9	38.2	29.5	36.9
GroupDRO	$\eta_{groupdro} = 1$	42.8	37.3	36.2	38.8
RSC	$\lambda_{rsc} = 10$	41.0	38.4	36.9	38.8
CAD	$\lambda_{cad} = 1$	42.8	37.9	35.5	38.7
CausIRL	$\lambda_{cirl} = 1$	42.4	38.3	36.3	39.0

In this section, we conduct numerical experiments on our benchmark to reveal the OOD generalization ability for existing algorithms. All experiments are conducted on a Pytorch platform with eight Tesla V100 GPUs. We evaluate each algorithm using the Average Precision (AP) from MS COCO (Lin et al., 2014). For object detection algorithms, our codes are based on mmdetection (Chen et al., 2019) and for domain generalization algorithms, our codes are stemmed from DomainBed (Gulrajani & Lopez-Paz, 2020). We draw several conclusions from the results.

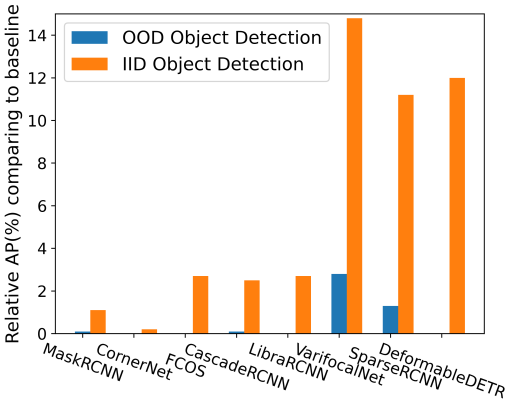
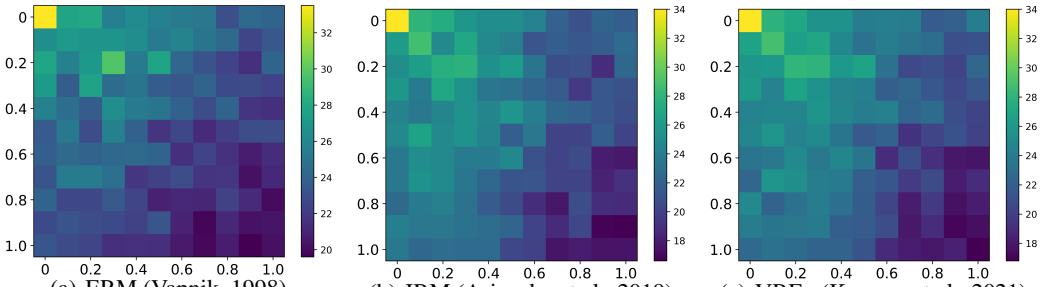


Figure 4: The improvements of recent object detection methods over the baseline on IID and OOD respectively. While the improvements on IID datasets (MS COCO) are prominent, it is not generalizable to OOD scenarios. The compared baseline method is Faster R-CNN.

The enormous achievements of object detection on IID datasets are marginal on the OOD condition. Table 3 summarizes the OOD results of various object detection algorithms. If we use the classic Faster R-CNN (Ren et al., 2015) as our baseline, all successful algorithms on object detection only marginally improve OOD generalization ability by 2.8 AP (VarifocalNet (Zhang et al., 2020a)) overall the four OOD benchmarks. While on the IID object detection benchmark (MS COCO test-dev (Lin et al., 2014)), this performance improvement can be up to 14.8 AP comparing to VarifocalNet (Zhang et al., 2020a). Figure 4 intuitively displaces the significant discrepancy between IID and OOD. What is responsible for these results? We suspect two factors: One is that current researches simply stem from the ideal assumption of IID regardless of whether it can be satisfied in real scenarios. The other is that the improvement on IID datasets may be a phenomenon of over-fitting since few works provide sufficient evidence that the causal features have been learned during the training process without evaluating on OOD benchmarks.

The tremendous success of domain generalization algorithms confronting OOD is inconsistent between classification and object detection. We draw this conclusion from Table 4 and the experimental results reported on OOD-bench (Ye et al., 2021). The OOD results on the four benchmarks in Table 4 suggest that the domain generalization algorithms degenerate or slightly outperform the ERM (Vapnik, 1998) which can be attributed to the hyper-parameters bias. Moreover, as for VREx (Krueger et al., 2021) which is the best models on Correlation-Bench (Ye et al., 2021), AP drops by 0.1 comparing to ERM (Vapnik, 1998) while VREx (Krueger et al., 2021) outperforms ERM (Vapnik, 1998) by 8.6 AP on Correlation-Bench (Ye et al., 2021). RSC (Huang et al., 2020) which is the



(a) ERM (Vapnik, 1998) (b) IRM (Arjovsky et al., 2019) (c) VREx (Krueger et al., 2021)

Figure 5: Controlled distribution shifts experiments results of ERM, IRM and VREx. X-axis is Diversity shift and Y-axis is Correlation shift. Each block indicates the AP(%).

best models on Diversity-Bench (Ye et al., 2021) degenerates 11.2 AP while improves 0.6 accuracy comparing to ERM (Vapnik, 1998).

The generalization inconsistency between classification and object detection of domain generalization algorithms happens among different detectors. As shown in Table 4.2, we choose the popular two-stage detector, Faster R-CNN (Ren et al., 2015), one-stage detector, RetinaNet (Lin et al., 2017b), and transformer-based detector, DETR (Carion et al., 2020), as base models for implementing domain generalization algorithms. Obviously, ERM achieves the best average generalization ability on the three detectors and we can conclude that the degeneration of domain generalization algorithms has little relevance to the detectors.

5.2 Controlled Distribution Shifts Experiments

Previous experiments provide performance analysis on the real scenarios for OOD generalization object detection. But it is hard to see which kind of distribution shift leads to performance degeneration. To systematically analyze the generalization performance under the influence of the two-dimensional distribution shift, we test the performance of Faster-RCNN trained by ERM (Vapnik, 1998), and the top performers on previous datasets (IRM (Arjovsky et al., 2019) and VREx (Krueger et al., 2021)) on CtrlShift dataset with different settings of the correlation shift and diversity shift. The results are shown in Figure 5, all methods consistently (Vapnik, 1998; Arjovsky et al., 2019; Krueger et al., 2021) achieve the best AP when both correlation shift and diversity shift are low. For ERM (Vapnik, 1998), the performance evenly degenerates on two dimensions. In Figure 5(b), with the increase of the two-dimension shift, the performance of IRM (Arjovsky et al., 2019) in the horizontal direction tends to degenerate faster than in the vertical direction. This indicates that IRM (Arjovsky et al., 2019) confronts correlation shift better than diversity shift. From Figure 5(c), we can observe the similar phenomenon happens for VREx (Krueger et al., 2021) on the two-dimensional shift. This phenomenon demonstrates that existing OOD generalization algorithms may help mitigate performance degradation when confronted with correlation shifts. Whereas for diversity shift, key components are missing to improve the generalization abilities, let alone the complex mixture of both shifts in real datasets. For future research, we recommend that both shifts should be included in new benchmark datasets and algorithms should be evaluated on both types of distribution shifts simultaneously.

6 Conclusion and Discussion

In this paper, we propose the first benchmark for OOD-OD tasks, named DetectBench. The benchmark suite includes four benchmark datasets along with a synthetic dataset to generate controlled distribution shifts. The experimental results conducted on DetectBench suggest that the enormous achievements in classical IID object detection are marginal on OOD generalization object detection. And the OOD generalization methods mainly tested on classification cannot generalize to object detection tasks. This raises questions about existing progress on object detection and OOD generalization algorithms. We appeal for more attention from the community for this problem to propose an OOD-OD method that is undoubtedly effective.

REFERENCES

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *NeurIPS*, 34, 2021.

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv:1412.4446*, 2014.

Ehab Albadawy, Ashirbani Saha, and Maciej A. Mazurowski. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical Physics*, 2018.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.

Haoyue Bai, Rui Sun, Lanqing Hong, Fengwei Zhou, Nanyang Ye, Han-Jia Ye, S-H Gary Chan, and Zhenguo Li. Decaug: Out-of-distribution generalization via decomposed feature representation and semantic augmentation. *arXiv:2012.09382*, 2020.

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv:1711.07910*, 2017.

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

Maria A Bravo, Sudhanshu Mittal, and Thomas Brox. Localized vision-language matching for open-vocabulary object detection. *arXiv preprint arXiv:2205.06160*, 2022.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *TPAMI*, pp. 1–1, 2019. ISSN 1939-3539. doi: 10.1109/tpami.2019.2956516. URL `http://dx.doi.org/10.1109/tpami.2019.2956516`.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv:1906.07155*, 2019.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018.

Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive faster r-cnn. *IJCV*, 2021.

Mathieu Chevalley, Charlotte Bunne, Andreas Krause, and Stefan Bauer. Invariant causal mechanisms through distribution matching. *arXiv preprint arXiv:2206.11646*, 2022.

MMCV Contributors. MMCV: OpenMMLab computer vision foundation. `https://github.com/open-mmlab/mmcv`, 2018.

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 *CVPR*, pp. 3213–3223, 2016.

Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. *international conference on intelligent transportation systems*, 2018.

Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *NeurIPS*, 29, 2016.

Daniel Coelho de Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 2019.

Kumari Deepshikha, Sai Harsha Yelleni, PK Srijith, and C Krishna Mohan. Monte carlo dropblock for modelling uncertainty in object detection. *arXiv preprint arXiv:2108.03614*, 2021.

Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *CVPR*, 2020.

Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1021–1030, 2020.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022a.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14084–14093, 2022b.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding yolo series in 2021. *arXiv:2107.08430*, 2021.

Ross Girshick. Fast r-cnn. In *ICCV*, pp. 1440–1448, 2015.

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*, pp. 580–587, 2014.

Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, volume 2, pp. 1458–1465. IEEE, 2005.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *JMLR*, 2012.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv:2007.01434*, 2020.

David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1031–1040, 2020.

Ali Harakeh and Steven L Waslander. Estimating and evaluating regression predictive uncertainty in deep object detectors. *arXiv preprint arXiv:2101.05036*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *ICCV*, Oct 2017.

Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *ECCV*, 2020.

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 *ICCV*, pp. 1314–1324, 2019.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 2017.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6891–6902, 2021.

Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, pp. 124–140. Springer, 2020.

Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *CVPR*, 2019.

Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv:1602.07360*, 2016.

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 *ICRA*, 2016.

KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5830–5840, 2021.

Masanori Koyama and Shoichiro Yamaguchi. Out-of-distribution generalization with maximal invariant predictor. 2021.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, pp. 5815–5826. PMLR, 2021.

Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, pp. 765–781. Springer Verlag, 2018.

Alexander Lehner, Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Mohammad-Ali Nikouei Mahani, Nassir Navab, Benjamin Busam, and Federico Tombari. 3d-vfield: Adversarial augmentation of point clouds for domain generalization in 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17295–17304, 2022.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018a.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, pp. 5400–5409, 2018b.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, pp. 624–639, 2018c.

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 *ECCV*, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pp. 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017b.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pp. 21–37. Springer, 2016.

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. *arXiv:2103.14030*, 2021.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, pp. 116–131, 2018.

Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 2348–2354. IEEE, 2019.

Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *CVPR*, 2019.

Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. In *ICLR*, 2021.

Christian S. Perone, Pedro Ballester, Rodrigo C. Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 2019.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *arXiv: Learning*, 2020.

Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *CVPR*, pp. 10213–10224, 2021.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. *arXiv:2109.02934*, 2021.

Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pp. 7263–7271, 2017.

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

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015.

Tobias Riedlinger, Matthias Rottmann, Marius Schubert, and Hanno Gottschalk. Gradient-based quantification of epistemic uncertainty for deep object detectors. *arXiv preprint arXiv:2107.04517*, 2021.

Adrian Lopez Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency. *BMVC*, 2019.

Yangjun Ruan, Yann Dubois, and Chris J Maddison. Optimal representations for covariate shift. *arXiv:2201.00057*, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv: Learning*, 2019.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pp. 4510–4520, 2018.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor, 2017. URL `http://arxiv.org/abs/1705.05065`.

Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv:2106.02266*, 2021.

Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv:2104.09937*, 2021.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum self-paced learning for cross-domain object detection. *CVIU*, 2019.

Peng Su, Kun Wang, Zeng Xingyu, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. In *ECCV*, 2020.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pp. 443–450. Springer, 2016.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. SparseR-CNN: End-to-end object detection with learnable proposals. *arXiv:2011.12450*, 2020.

Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *CVPR*, pp. 10781–10790, 2020.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *arXiv:1904.01355*, 2019.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 969–977, 2018.

Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017.

Georg Volk, Stefan Müller, Alexander von Bernuth, Dennis Hospach, and Oliver Bringmann. Towards robust cnn-based object detection through augmentation with synthetic rain variations. *international conference on intelligent transportation systems*, 2019.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Robert J Wang, Xiang Li, and Charles X Ling. Pelee: A real-time object detection system on mobile devices. *NeurIPS*, 31, 2018.

Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *ECCV*, 2020.

Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 847–856, 2022.

Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. *arXiv:1904.06493*, 2019.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pp. 1492–1500, 2017.

Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, 2020a.

Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *CVPR*, 2020b.

Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, volume 34, pp. 6502–6509, 2020c.

Yilun Xu and Tommi Jaakkola. Learning representations that support robust transfer of predictors. *arXiv:2110.09940*, 2021.

Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv:2001.00677*, 2020.

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:2106.03721*, 2021.

Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv:1805.04687*, 2(5):6, 2018.

Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2100–2110, 2019.

Sergey Zakharov, Wadim Kehl, and Slobodan Ilic. Deceptionnet: Network-driven domain randomization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 532–541, 2019.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14393–14402, 2021.

Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. Varifocalnet: An iou-aware dense object detector. *arXiv:2008.13367*, 2020a.

Marvin Mengxin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. 2020b.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pp. 6848–6856, 2018.

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.

Xiaowei Zhao, Xianglong Liu, Yifan Shen, Yuqing Ma, Yixuan Qiao, and Duorui Wang. Revisiting open world object detection. *arXiv preprint arXiv:2201.00471*, 2022.

Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. *national conference on artificial intelligence*, 2020.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

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

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. URL `https://openreview.net/forum?id=gZ9hCDWe6ke`.

A APPENDIX

A.1 THEORETICAL ANALYSIS OF OOD GENERALIZATION OBJECT DETECTION

The OOD generalization object detection has been fragmentally researched in previous research, however, no rigorous definition of OOD generalization object detection has been given. We give a definition and taxonomy as follows:

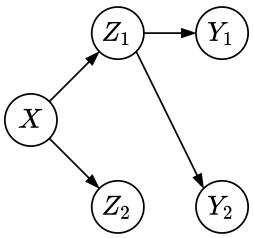


Figure 6: The causal influence among the concerned variables.

OOD generalization object detection (OOD-OD): In object detection tasks, algorithms learn a mapping function f to predict the category ($\mathbf{y}$) and location of interested objects in an image $\mathbf{x}$. In OOD generalization object detection tasks, training and test data pairs $(\mathbf{X}, \mathbf{Y})$ are not necessarily drawn from the same distribution. This poses great challenges for existing machine learning methods as most methods are reliant on exploiting the correlation between $\mathbf{X}$ and $\mathbf{Y}$. Due to the distribution change, the correlation might not be generalizable. More specifically, we depict the causal data generating process in Figure 6. When the data $\mathbf{X}$ are given, the causal features $\mathbf{Z}_1$ and the non-causal features $\mathbf{Z}_2$ are given. The causal features reliably determine the location and categories of interested objects in the input images. The non-causal features are irrelevant features for predictions. An intuitive example is that if we want to recognize a dog in an image, the causal features are dogs' shapes. The non-causal features are the environment features, such as weather or captured time of the image. We improve the definitions in (Ye et al., 2021) and propose the following mathematical definitions for $\mathbf{Z}_1$ and $\mathbf{Z}_2$ given overall semantic features $\mathbf{z}$[1]:

$$\forall\, z \in \mathbf{Z}_1 : p(z) \cdot q(z) \neq 0 \wedge \forall\, y_1 \in \mathbf{Y}_1 : p(y_1|z) = q(y_1|z) \wedge \forall\, y_2 \in \mathbf{Y}_2 : p(y_2|z) = q(y_2|z) \quad (1)$$

$$\exists\, z \in \mathbf{Z}_2 : p(z) \cdot q(z) = 0 \vee \exists\, y_1 \in \mathbf{Y}_1 : p(y_1|z) \neq q(y_1|z) \vee \exists\, y_2 \in \mathbf{Y}_2 : p(y_2|z) \neq q(y_2|z) \quad (2)$$

where p is the training distribution and q is the test distribution. Since $\mathbf{Z}_1$ is the stable and reliable predictor for the category and location of objects, there are two kinds of shifts intuitively because of the discrepancy of $\mathbf{Z}_2$ in training and test distribution. Diversity shift stems from the first kind of features in $\mathbf{Z}_2$ since the diversity of data is embodied by novel features not shared by the environments; whereas correlation shift is caused by the second kind of features in $\mathbf{Z}_2$ which is spuriously correlated with some $\mathbf{Y}_1$ or $\mathbf{Y}_2$. Based on this, we partition $\mathbf{Z}_2$ into two subsets:

$$\mathbf{S} := \{\mathbf{z} \in \mathbf{Z}_2 | p(\mathbf{z}) \cdot q(\mathbf{z}) = 0\} \quad \mathbf{T} := \{\mathbf{z} \in \mathbf{Z}_2 | p(\mathbf{z}) \cdot q(\mathbf{z}) \neq 0\} \quad (3)$$

Definition A.1. ***(Definition of Diversity shift and Correlation shift for OOD-OD.)*** *Given* $\mathbf{S}$ *and* $\mathbf{T}$ *defined in Equation 3, the definitions of diversity shift and correlation shift are given as follows:*

$$D_{diversity} := \frac{1}{2} \int_{\mathbf{S}} \mid p(z) - q(z) \mid dz \quad (4)$$

$$D_{correlation} := \frac{1}{2} \int_{\mathbf{T}} \sqrt{p(z) \cdot q(z)} \int_{y_2} \sum_{y_1} \mid p(y_1, y_2|z) - q(y_1, y_2|z) \mid dz \quad (5)$$

It can be seen that both $D_{\text{diversity}}$ and $D_{\text{correlation}}$ are within the range of $[0, 1]$. $D_{\text{diversity}}$ measures the support difference non-causal features for object detection. While $D_{\text{correlation}}$ gauges the variations of conditional probabilities of the object category $\mathbf{Y}_1$ given non-causal features and object locations. This serves as an indicator for spurious correlations existing in datasets. The proposed definition first provides a quantitative way for measuring the distributional shifts for OOD-OD to the best of our knowledge. We leave it for future work to compute the numerical values of shifts given an object detection dataset.

A.2 PROOF

Proposition A.2. *For any probability functions* p *and* q *of training distribution and testing distribution,* $D_{diversity}$ *and* $D_{correlation}$ *are inclusively bounded between 0 and 1.*

[1]We assume no category label ($\mathbf{Y}_1$) shifts for simplicity and without loss of generality

Table 6: Illustration of the two settings on Sim2real. $\text{sim}_{\text{train}}$ and sim_{val} indicate the training set and validating set from original Sim10K (Johnson-Roberson et al., 2016) while $\text{city}_{\text{train}}$ and city_{val} are from original Cityscapes (Cordts et al., 2016). **Quantity** indicates the number of images. **Total** counts the total number of training and testing domains respectively.

Setting	Split	Train	Test	Quantity	Total
part-sim-part-real	$\text{sim}_{\text{train}}$	√		8000	8500
	sim_{val}		√	1000	
	$\text{city}_{\text{train}}$		√	2975	3975
	city_{val}	√		500	
all-sim-all-real	$\text{sim}_{\text{train}}$	√		8000	9000
	sim_{val}	√		1000	
	$\text{city}_{\text{train}}$		√	2975	3475
	city_{val}		√	500	

Proof. Obviously, both $D_{\text{diversity}}$ and $D_{\text{correlation}}$ are positive. Then, we prove the upper bound by the triangle inequality as followed:

$$D_{\text{diversity}} = \frac{1}{2}\int_{\mathbf{S}} \mid p(z) - q(z) \mid dz \leq \frac{1}{2}\int_{\mathbf{S}} [p(z) + q(z)] dz \leq 1 \tag{6}$$

Similarly, we have the following inequality:

$$\begin{aligned} D_{\text{correlation}} &= \frac{1}{2}\int_{\mathbf{T}} \sqrt{p(z)\cdot q(z)} \sum_{y_1, y_2} \mid p(y_1, y_2|z) - q(y_1, y_2|z) \mid dz \\ &\leq \frac{1}{2}\int_{\mathbf{T}} \sqrt{p(z)\cdot q(z)} \sum_{y_1, y_2} [p(y_1, y_2|z) + q(y_1, y_2|z)] dz \\ &= \frac{1}{2}\int_{\mathbf{T}} 2\sqrt{p(z)\cdot q(z)} dz = \int_{\mathbf{T}} \sqrt{p(z)\cdot q(z)} dz \leq 1 \end{aligned} \tag{7}$$

The second inequality is due to triangle inequality. □

A.3 Implementation Details

To evaluate the object detection algorithms, we use the models and the pre-trained weights provided by mmdetection (Chen et al., 2019). For domain generalization algorithms on OOD generalization object detection, we derive the implementations using Faster R-CNN (Ren et al., 2015) with ResNet-50 FPN backbone (He et al., 2016) from torchvision. The whole network is optimized by Stochastic Gradient Descent with learning rate 0.02, momentum 0.9 and weight decay 0.0005.

A.4 Further Results

Task complexity. To analyse the IID condition on CtrlShift, which indicates both $D_{\text{correlation}}$ and $D_{\text{diversity}}$ equal zero, we propose a hyper-parameter task complexity α to measure the difficulty of the task. The difficulty is adjusted by using $1 - \alpha$ percent novel data in the testing set in addition to the original training data. The experimental results are shown in Figure 7. The generalization ability of each algorithm drops with the increase of task complexity.

Sim2real benchmark. The training set of the Sim2real results reported in the main manuscript comprises the training data from Sim10K (Johnson-Roberson et al., 2016) and the validating data from Cityscapes (Cordts et al., 2016), while the testing set comprises the training data from Cityscapes (Cordts et al., 2016) and the validating data from Sim10K (Johnson-Roberson et al., 2016) (noted by part-sim-part-real, more details can be found in Table A.4). We reported the experimental results on all-sim-all-real in Table 7 and Table 8.

Full results. Table 9, 10, 11, 12 and 13 are the full experimental results of Table 3 and 4, including AP, AP_{50}, AP_{75}, AP_{s}, AP_{m} and AP_{l} evaluation metrics.

Table 7: The experimental results of object detection algorithms on the all-sim-all-real of Sim2real. Mem (GB)† and Inf time (fps)† are from mmdetection (Chen et al., 2019).

Detector	Backbone	Mem†	fps†	AP
Faster R-CNN (Ren et al., 2015)	X-101	10.3	9.4	35.6
RetinaNet (Lin et al., 2017b)	X-101	10.0	8.7	38.0
Mask R-CNN (He et al., 2017)	X-101	10.7	8.0	36.7
CornetNet (Law & Deng, 2018)	Hourglass104	13.9	4.2	21.6
YOLOv3 (Redmon & Farhadi, 2018)	DarkNet-53	7.4	48.1	28.2
FCOS (Tian et al., 2019)	X-101	10.0	9.7	37.9
Cascade R-CNN (Cai & Vasconcelos, 2019)	X-101	10.7	-	40.5
MS R-CNN (Huang et al., 2019)	R-X101	11.0	8.0	35.7
Libra R-CNN (Pang et al., 2019)	X-101	10.8	8.5	35.3
DH R-CNN (Wu et al., 2019)	R-50	6.8	9.5	33.8
VarifocalNet (Zhang et al., 2020a)	X-101	-	-	42.3
Sparse R-CNN (Sun et al., 2020)	R-101	-	-	40.3
Deformable (Zhu et al., 2021)	R-50	-	-	37.4
YOLOX (Ge et al., 2021)	YOLOX-x	28.1	-	36.4

Table 8: The experimental results of domain generalization algorithms on the all-sim-all-real of Sim2real.

Algorithm	hyper-parameters	AP
ERM (Vapnik, 1998)	-	32.8
IB-ERM (Ahuja et al., 2021)	$\lambda_{ib} = 100$	18.3
IRM (Arjovsky et al., 2019)	$\lambda_{irm} = 1$	32.7
MMD (Li et al., 2018b)	$\gamma_{mmd} = 1$	33.2
CORAL (Sun & Saenko, 2016)	$\gamma_{mmd} = 1$	32.5
VREx (Krueger et al., 2021)	$\lambda_{vrex} = 1$	32.4
GS (Pezeshki et al., 2020)	$\lambda_{reg} = 0.1$	31.4
IGA (Koyama & Yamaguchi, 2021)	$\lambda_{penalty} = 1000$	33.4
GroupDRO (Sagawa et al., 2019)	$\eta_{groupdro} = 0.01$	31.9

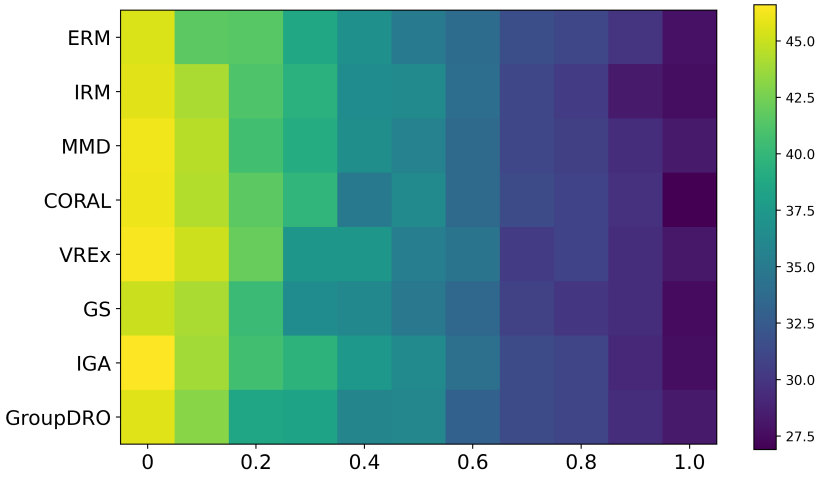


Figure 7: X-axis is task complexity α. Each block indicates the AP(%).

Table 9: Experimental results of detectors on Sim2real.

Detector	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Faster R-CNN	X-101	52.1	73.6	55.2	22.3	58.6	82.3
RetinaNet	X-101	49.3	72.2	51.0	16.8	56.2	81.6
Mask R-CNN	X-101	52.9	74.2	56.4	23.2	60.0	82.7
CornetNet	Hourglass104	27.8	40.2	28.7	6.6	34.4	46.7
YOLOv3	DarkNet-53	38.2	62.7	40.0	10.2	42.2	70.0
FCOS	X-101	50.8	72.1	53.1	21.4	56.6	82.1
Cascade R-CNN	X-101	53.1	74.1	56.2	23.2	59.2	83.6
MS R-CNN	R-X101	52.5	73.7	56.0	23.0	59.5	82.3
Libra R-CNN	X-101	51.4	72.1	55.0	22.0	59.7	81.5
DH R-CNN	R-50	52.2	73.4	56.1	23.6	59.2	81.2
VarifocalNet	X-101	56.1	75.3	59.4	25.9	63.9	85.9
Sparse R-CNN	R-101	54.1	76.8	57.7	26.7	59.3	81.9
DETR	R-50	37.8	63.9	38.3	10.2	38.0	72.1
Deformable	R-50	53.3	78.0	57.0	25.6	59.7	80.2
Swin Transformer	Swin-B	58.4	78.5	63.5	31.4	65.5	84.2
YOLOX	YOLOX-x	51.1	70.3	53.7	19.3	58.4	84.4

Table 10: Experimental results of detectors on Weather.

Detector	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Faster R-CNN	X-101	35.3	56.1	36.3	14.5	37.5	49.5
RetinaNet	X-101	34.7	54.3	35.8	12.8	36.9	49.4
Mask R-CNN	X-101	34.8	55.2	35.8	14.9	36.9	47.9
CornetNet	Hourglass104	29.7	46.1	30.6	12.2	34.3	38.4
YOLOv3	DarkNet-53	27.0	47.9	26.1	8.7	29.4	40.5
FCOS	X-101	35.6	55.7	36.4	14.8	37.9	48.8
Cascade R-CNN	X-101	36.3	56.4	37.8	15.1	38.2	51.7
MS R-CNN	R-X101	35.5	56.1	36.6	14.6	37.4	50.2
Libra R-CNN	X-101	34.9	54.8	36.3	14.0	37.4	49.5
DH R-CNN	R-50	35.3	55.7	36.5	14.5	37.2	50.5
VarifocalNet	X-101	38.7	59.3	39.8	16.0	41.0	54.1
Sparse R-CNN	R-101	36.3	57.5	37.1	15.6	38.5	50.8
DETR	R-50	23.5	41.7	22.5	4.9	23.2	40.1
Deformable	R-50	35.1	56.7	35.3	14.7	37.2	49.8
Swin Transformer	Swin-B	36.6	56.2	38.2	15.8	39.7	51.1
YOLOX	YOLOX-x	33.1	51.7	34.1	11.5	35.4	47.6

Table 11: Experimental results of detectors on Scene.

Detector	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Faster R-CNN	X-101	36.0	56.3	38.0	15.9	38.8	54.7
RetinaNet	X-101	35.8	56.1	37.4	13.5	39.5	55.9
Mask R-CNN	X-101	35.8	56.2	37.5	15.7	38.2	54.3
CornetNet	Hourglass104	30.8	46.1	32.6	13.8	32.7	45.9
YOLOv3	DarkNet-53	28.2	49.7	27.1	10.0	32.3	44.2
FCOS	X-101	36.5	56.7	38.7	15.9	39.1	54.8
Cascade R-CNN	X-101	36.7	56.7	38.1	15.0	39.3	58.9
MS R-CNN	R-X101	35.7	56.1	37.2	15.3	38.4	54.4
Libra R-CNN	X-101	35.5	55.6	36.8	15.9	38.3	54.9
DH R-CNN	R-50	35.2	55.3	36.9	15.9	38.2	54.7
VarifocalNet	X-101	39.0	58.8	40.8	18.1	41.4	59.4
Sparse R-CNN	R-101	36.4	57.4	37.4	18.4	39.1	56.0
DETR	R-50	24.4	42.8	23.8	6.3	25.3	47.4
Deformable	R-50	35.3	55.8	36.2	15.3	37.9	54.6
Swin Transformer	Swin-B	32.0	50.1	33.7	14.7	36.7	48.3
YOLOX	YOLOX-x	34.8	54.1	36.2	13.4	37.7	53.4

Table 12: Experimental results of detectors on Time.

Detector	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Faster R-CNN	X-101	26.4	45.2	26.3	9.7	25.2	39.7
RetinaNet	X-101	24.6	43.1	24.3	7.8	23.5	37.8
Mask R-CNN	X-101	26.8	45.8	26.5	9.5	25.8	39.8
CornetNet	Hourglass104	22.5	37.8	22.3	8.6	23.4	29.6
YOLOv3	DarkNet-53	19.2	34.5	18.7	5.1	18.3	31.3
FCOS	X-101	24.8	42.5	24.4	9.2	23.6	36.2
Cascade R-CNN	X-101	24.1	41.0	23.6	8.7	21.9	37.6
MS R-CNN	R-X101	25.5	43.4	25.3	9.2	23.8	39.0
Libra R-CNN	X-101	25.5	43.7	25.2	8.9	24.6	39.2
DH R-CNN	R-50	25.5	43.9	25.4	9.4	24.9	37.6
VarifocalNet	X-101	27.4	45.6	27.0	10.4	26.0	39.3
Sparse R-CNN	R-101	28.2	48.3	27.7	10.6	27.2	41.4
DETR	R-50	15.8	30.9	13.7	3.2	14.7	28.2
Deformable	R-50	26.1	46.2	25.1	10.1	25.4	39.0
Swin Transformer	Swin-B	28.9	47.6	29.4	10.5	28.6	42.3
YOLOX	YOLOX-x	29.4	48.0	29.2	9.4	28.2	44.0

Table 13: Generalization performance of detectors with OOD algorithms.

Algorithm	hyper-parameters	Dataset	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
ERM	-	Sim2real	44.1	64.0	47.1	13.8	41.2	68.4
IB-ERM	$\lambda_{ib} = 0.1$		42.5	64.4	44.9	12.6	39.2	67.0
IRM	$\lambda_{irm} = 0.1$		42.3	64.9	45.0	12.7	39.4	66.1
MMD	$\gamma_{mmd} = 1$		42.9	64.8	45.4	13.2	39.8	67.2
CORAL	$\gamma_{mmd} = 10$		42.6	64.6	45.3	12.8	39.7	66.8
VREx	$\lambda_{vrex} = 1$		43.1	65.4	45.4	13.0	40.0	67.4
GS	$\lambda_{reg} = 10$		40.3	64.2	42.5	11.5	37.6	65.3
IGA	$\lambda_{penalty} = 1$		42.9	64.7	45.6	13.7	40.0	66.6
GroupDRO	$\eta_{groupdro} = 1$		42.8	65.0	45.4	12.8	39.9	67.1
RSC	$\lambda_{rsc} = 10$		41.0	63.6	43.2	10.8	38.6	64.9
ERM	-	Weather	24.4	47.8	21.8	9.5	29.3	41.9
IB-ERM	$\lambda_{ib} = 0.1$		24.7	47.7	21.9	9.7	29.6	41.8
IRM	$\lambda_{irm} = 0.1$		24.9	48.3	22.1	9.7	29.9	41.9
MMD	$\gamma_{mmd} = 1$		24.1	47.2	21.5	9.4	29.0	41.6
CORAL	$\gamma_{mmd} = 10$		24.4	47.3	21.9	9.7	29.2	41.6
VREx	$\lambda_{vrex} = 1$		24.7	48.4	21.8	9.6	29.6	41.4
GS	$\lambda_{reg} = 10$		19.9	39.2	17.5	8.0	24.6	35.4
IGA	$\lambda_{penalty} = 1$		24.6	47.9	21.7	9.9	29.6	41.2
GroupDRO	$\eta_{groupdro} = 1$		24.5	47.8	21.6	9.6	29.2	41.3
RSC	$\lambda_{rsc} = 10$		8.7	17.6	7.5	3.3	11.1	17.9
ERM	-	Scene	24.1	46.7	21.2	9.7	29.4	46.5
IB-ERM	$\lambda_{ib} = 0.1$		23.4	45.6	20.4	9.5	28.6	45.7
IRM	$\lambda_{irm} = 0.1$		23.2	45.8	20.3	9.4	28.4	45.0
MMD	$\gamma_{mmd} = 1$		23.9	46.4	21.2	9.4	29.2	47.5
CORAL	$\gamma_{mmd} = 10$		22.8	44.5	20.1	9.4	27.8	45.4
VREx	$\lambda_{vrex} = 1$		24.3	47.0	21.6	9.7	29.7	46.6
GS	$\lambda_{reg} = 10$		18.4	35.9	16.4	8.0	23.7	38.0
IGA	$\lambda_{penalty} = 1$		24.3	47.6	21.0	10.2	29.3	45.9
GroupDRO	$\eta_{groupdro} = 1$		23.6	46.1	21.1	10.1	29.0	45.7
RSC	$\lambda_{rsc} = 10$		8.8	17.9	7.4	3.5	11.6	21.5
ERM	-	Time	17.7	37.5	14.6	6.6	19.5	32.0
IB-ERM	$\lambda_{ib} = 0.1$		18.2	38.3	15.2	6.8	20.1	32.1
IRM	$\lambda_{irm} = 0.1$		17.5	37.5	14.2	6.6	19.3	30.6
MMD	$\gamma_{mmd} = 1$		18.4	38.4	15.4	7.0	20.2	31.4
CORAL	$\gamma_{mmd} = 10$		17.7	37.6	14.6	6.4	19.5	30.7
VREx	$\lambda_{vrex} = 1$		18.0	38.3	14.9	6.8	19.9	31.2
GS	$\lambda_{reg} = 10$		15.2	32.4	12.5	5.6	17.4	27.7
IGA	$\lambda_{penalty} = 1$		18.6	39.1	15.4	7.2	20.4	32.0
GroupDRO	$\eta_{groupdro} = 1$		18.1	38.1	15.1	6.9	20.1	31.4
RSC	$\lambda_{rsc} = 10$		7.0	15.0	5.7	2.5	8.2	14.7