A Proof

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

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

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

Similarly, we have the following inequality:

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

The second inequality is due to triangle inequality. □

B Datasets

BDD100K The original BDD100K [40] 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. We remove the images with an undefined attribute label and separate the rest into three OOD environments.

Sim10K Sim10k [19] 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. On DetectBench, we use the training and validating data to construct Sim2real benchmark.

Cityscapes Cityscapes [6] 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 Sim2real benchmark for simplicity and without loss of generality.

CtrlShift CtrlShift is a synthetic dataset to analyze the two-dimension shift on OOD object detection. It provides an API to generate the training set and the testing set with specific choices of $\rho_{diversity}$ and $\rho_{attribute}$.

C Clarification of Tasks

Domain Randomization techniques [36, 37, 42, 41, 17] aim at providing enough simulated domains at training data so that models are possible to generalize to real-world scenarios.
OOD Detection for Object Detection [20, 9, 14, 31, 8, 26, 13, 7] can be formulated as a binary classification problem which distinguishes the Out-of-Distribution data.
Open-World Object Detection [20, 45] 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 [12, 43, 10, 3] aims to train an detector which can deal with text inputs to detect objects in any novel categories.

Table 1: Illustration of the two settings on Sim2real. $\text{sim}_{\text{train}}$ and sim_{val} indicate the training set and validating set from original Sim10K [19] while $\text{city}_{\text{train}}$ and city_{val} are from original Cityscapes [6]. **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	

D Implementation Details

To evaluate the object detection algorithms, we use the models and the pre-trained weights provided by mmdetection [5].

For domain generalization algorithms on OOD object detection, we derive the implementations using Faster R-CNN [30] with ResNet-50 FPN backbone [16] from torchvision. The whole network is optimized by Stochastic Gradient Descent with learning rate 0.02, momentum 0.9 and weight decay 0.0005.

E Further Results

Task complexity. To analyse the IID condition on CtrlShift, which indicates both $D_{\text{attribute}}$ 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 1. 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 [19] and the validating data from Cityscapes [6], while the testing set comprises the training data from Cityscapes [6] and the validating data from Sim10K [19] (noted by part-sim-part-real, more details can be found in Table E). We reported the experimental results on all-sim-all-real in Table 2 and Table 3.

Table 2: The experimental results of object detection algorithms on the all-sim-all-real of Sim2real. Mem (GB)† and Inf time (fps)† are from mmdetection [5].

Detector	Backbone	Mem†	fps†	AP
Faster R-CNN [30]	X-101	10.3	9.4	35.6
RetinaNet [25]	X-101	10.0	8.7	38.0
Mask R-CNN [15]	X-101	10.7	8.0	36.7
CornetNet [23]	Hourglass104	13.9	4.2	21.6
YOLOv3 [29]	DarkNet-53	7.4	48.1	28.2
FCOS [35]	X-101	10.0	9.7	37.9
Cascade R-CNN [4]	X-101	10.7	-	40.5
MS R-CNN [18]	R-X101	11.0	8.0	35.7
Libra R-CNN [27]	X-101	10.8	8.5	35.3
DH R-CNN [39]	R-50	6.8	9.5	33.8
VarifocalNet [44]	X-101	-	-	42.3
Sparse R-CNN [34]	R-101	-	-	40.3
Deformable [46]	R-50	-	-	37.4
YOLOX [11]	YOLOX-x	28.1	-	36.4

Table 3: The experimental results of domain generalization algorithms on the all-sim-all-real of Sim2real.

Algorithm	hyper-parameters	AP
ERM [38]	-	32.8
IB-ERM [1]	$\lambda_{ib} = 100$	18.3
IRM [2]	$\lambda_{irm} = 1$	32.7
MMD [24]	$\gamma_{mmd} = 1$	33.2
CORAL [33]	$\gamma_{mmd} = 1$	32.5
VREx [22]	$\lambda_{vrex} = 1$	32.4
GS [28]	$\lambda_{reg} = 0.1$	31.4
IGA [21]	$\lambda_{penalty} = 1000$	33.4
GroupDRO [32]	$\eta_{groupdro} = 0.01$	31.9

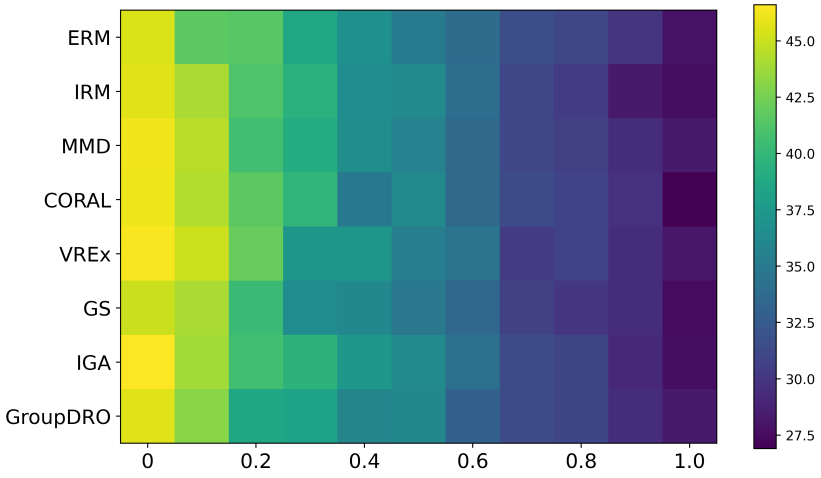

Figure 1: X-axis is task complexity α. Each block indicates the AP(%).

References

[1] 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.

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

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

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *TPAMI*, page 1–1, 2019.

[5] 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.

[6] 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*, pages 3213–3223, 2016.

[7] 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.

[8] 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*, pages 1021–1030, 2020.

[9] 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*, 2022.

[10] 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*, pages 14084–14093, 2022.

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

[12] 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.

[13] 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*, pages 1031–1040, 2020.

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

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

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

[17] 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*, pages 6891–6902, 2021.

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

[19] 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.

[20] 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*, pages 5830–5840, 2021.

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

[22] 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*, pages 5815–5826. PMLR, 2021.

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

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

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

[26] 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)*, pages 2348–2354. IEEE, 2019.

[27] 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.

[28] 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.

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

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

[31] 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.

[32] 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.

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

[34] 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.

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

[36] 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)*, pages 23–30. IEEE, 2017.

[37] 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*, pages 969–977, 2018.

[38] Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[39] 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.

[40] 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.

[41] 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*, pages 2100–2110, 2019.

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

[43] 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*, pages 14393–14402, 2021.

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

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

[46] 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.