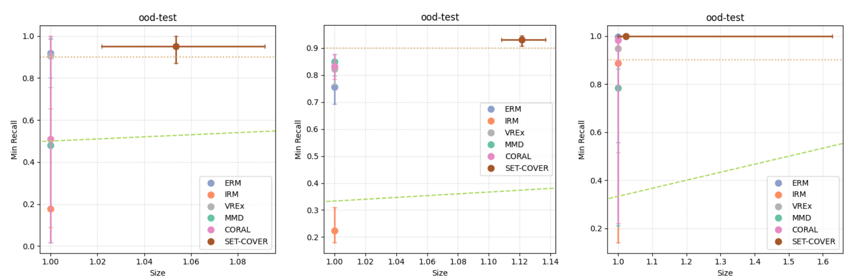
1 Experiments on Other OOD Baselines

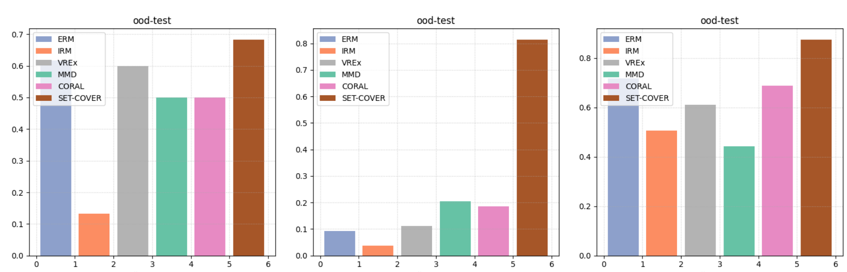
For the main experiments of this work we chose the ERM method as the single-prediction basekine, due to its vast popularity in real-world applications, and its superior, or at least compatible performance in various OOD baselines [Koh et al., 2021, Gulrajani and Lopez-Paz, 2020]. In the next section we compare SET-COVER to additional common OOD baselines. These include IRM [Arjovsky et al., 2019], VREx [Krueger et al., 2021], MMD [Li et al., 2018], and CORAL [Sun and Saenko, 2016]. We use the DomainBed [Gulrajani and Lopez-Paz, 2020] package to train these models.

The results show an advantage for SET-COVER over common single-prediction baselines in maintaining the 90% min-recall target across unseen domains. These results suggest that set-valued predictors may be a step in the right direction for robust OOD generalization.



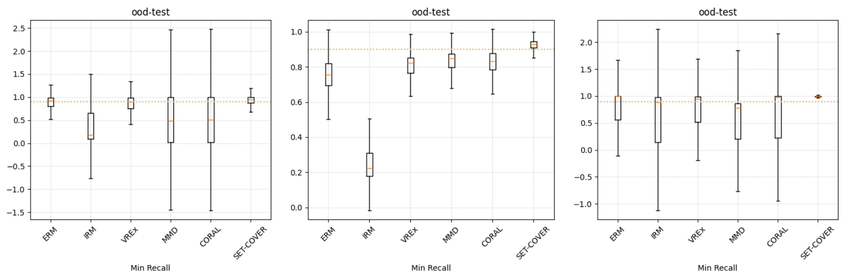
(a) Camelyon (b) FMoW (c) iWildCam

Figure 1: Each figure represents Min-Recall over Avg Set Size cross. y-axis represents min-recall, and x-axis represents average set size. Each cross shows the median and the 25th and 75th percentiles for both metrics across domain. The horizontal solid line represents the 90% recall target value, and dashed yellow diagonal line represents performance of a random predictor.



(a) Camelyon (b) FMoW (c) iWildCam

Figure 2: Percentage of OOD domains where the min-recall is higher than 90%. Each bar represents a different model.



(a) Camelyon (b) FMoW (c) iWildCam

Figure 3: Boxplots represent the distribution of min-recall across OOD domains.

Table 1: Summary of OOD Results for different OOD baselines.

Model	**Camelyon**			**FMoW**		
	Median Min Recall ↑	Median Avg Size ↓	Recall $\geq$ 90% Pctg ↑	Median Min Recall ↑	Median Avg Size ↓	Recall $\geq$ 90% Pctg ↑
ERM	0.91	1.0	0.61	0.75	1.0	0.09
IRM	0.17	1.00	0.13	0.22	1.00	0.03
VREx	0.90	1.00	0.60	0.82	1.00	0.11
MMD	0.48	1.00	0.50	0.84	1.00	0.20
CORAL	0.50	1.00	0.50	0.83	1.00	0.18
SET-COVER	0.95	1.05	0.68	0.93	1.12	0.81

Model	**iWildCam**		
	Median Min Recall ↑	Median Avg Size ↓	Recall $\geq$ 90% Pctg ↑
ERM	0.99	1.0	0.71
IRM	0.88	1.00	0.50
VREx	0.94	1.00	0.60
MMD	0.78	1.00	0.44
CORAL	0.98	1.00	0.68
SET-COVER	1.00	1.02	0.87

References

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

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

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts, 2021. URL `https://arxiv.org/abs/2012.07421`.

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 *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016.