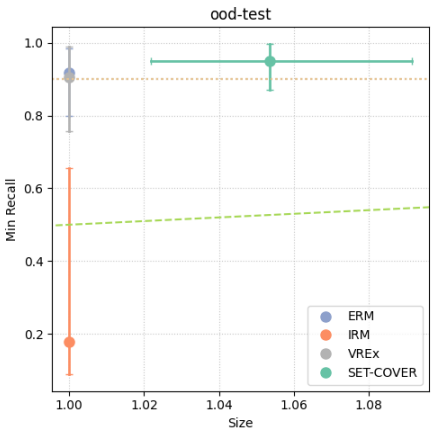
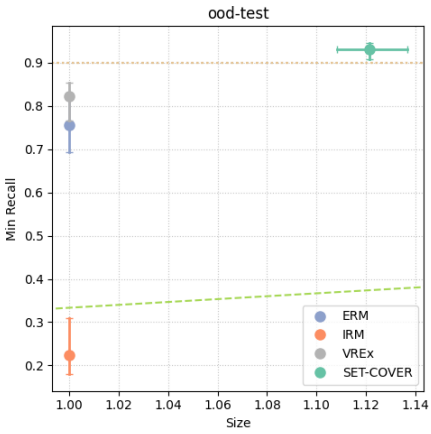
1 Experiments on Other OOD Benchmarks

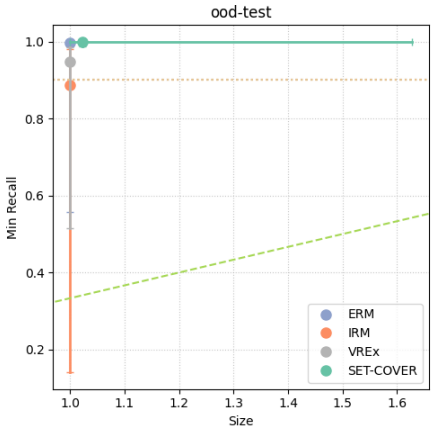
For the main experiments of this work we chose the ERM method as the single-prediction basekine, due to its vast popularity in real-world applications, and its superior, or at least compatible performance in various OOD baselines. In the next section we compare SET-COVER to additional common OOD baselines. These include IRM and VREx. We use the DomainBed package to train these models.



(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. **Blue** represents ERM predictor, **Orange** represents IRM, **Grey** represents VREx, and **Green** represents SET-COVER. The horizontal solid line represents the 90% recall target value, and dashed yellow diagonal line represents performance of a random predictor.

Table 1: Summary of OOD Results

Model	Camelyon			FMoW		
	Median Min Recall ↑	Median Avg Size ↓	Recall ≥ 90% Pctg ↑	Median Min Recall ↑	Median Avg Size ↓	Recall ≥ 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
SET-COVER	0.95	1.05	0.68	0.93	1.12	0.81

Model	iWildCam		
	Median Min Recall ↑	Median Avg Size ↓	Recall ≥ 90% Pctg ↑
ERM	0.99	1.0	0.71
IRM	0.88	1.00	0.50
VREx	0.94	1.00	0.60
SET-COVER	1.00	1.02	0.87