Supplementary Material

1. Visualization

In this section, we provide interpretable visualizations of deep neural networks trained by selected algorithms to get a better understanding of the learned representations. In Figure 1, we visualize the class attention maps [4] of samples that W2D correctly predicts. Since W2D discards the most predictive representations and forces the model to predict with remaining information, it tends to capture more structural feature information in order to make correct predictions; thus it exhibits broader attention during inference.

On the other hand, in Figure 2, we visualize the class attention maps [4] of samples that W2D incorrectly predicts. We observe that in the first three columns, W2D fails to make correct predictions while ERM or W2D's components (sample dimension and feature dimension) can predict correctly in these columns. Although it appears W2D's performances are degraded over these samples, we believe these samples are fairly difficult to predict correctly (even by human) in the first place.

In addition, we visualize the worst-case samples from different domains during training in PACS. We observe that worse-case samples often have rare shapes or textures. Also, the objects in these samples are often partially occluded or viewed from an unusual angle.

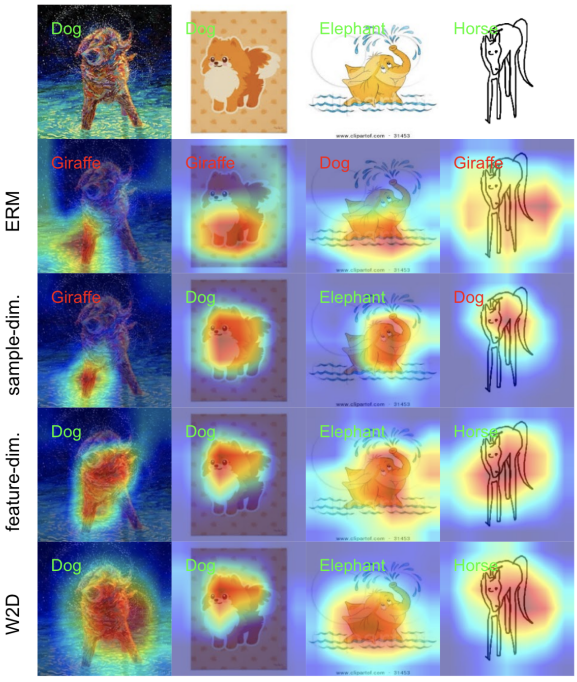

Figure 1. Attention heatmaps of selected algorithms trained and evaluated on PACS [2] visualized by class activation map [4]. Green label means correct prediction and red label means wrong prediction.

2. Additional Implementation Details

For network architecture, models trained on CMNIST adopt the two-layer convolution network, while other datasets use ResNet-18 as the backbone following Oodbench. For hyperparameter search protocol, we use the same as in Ood-bench except for batchsize search space. As we motioned in the discussion section, the batchsize range goes as small as 8 in both Ood-bench and Domainbed, limiting the potential of the DRO-family methods to take advantage of the hard samples. To avoid this issue, we increase the minimum batchsize to 16 in the implementation.

3. Additional Empirical Results

Recall that we evaluate the results in CMNIST using the -90 as testing environment in Table 2 following Ood-Bench [3]. In this section, we report the results averaged over three environments (+90, +80 and -90) in CMNIST, which is the protocol used in DomainBed [1]. The choice of the settings does not affects our ranking score. W2D is still among the top three the datasets dominated by correlation shift.

References

[1] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 1, 2

[2] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 1, 2

[3] Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Ood-bench: Benchmarking and understanding out-of-distribution generalization

Algorithm	CMNIST	NICO	CelebA	Average	Prev score	Ranking score
GroupDRO	61.2 ± 0.6	71.8 ± 0.8	87.5 ± 1.1	73.5	-1	+1
W2D	59.1 ± 0.3	71.6 ± 0.9	87.7 ± 0.4	72.8	+3	+1
ERM	58.5 ± 0.3	71.4 ± 1.3	87.2 ± 0.2	72.3	0	0
ERDG	59.2 ± 0.7	70.6 ± 1.3	84.5 ± 0.2	71.4	-2	0
ARM	63.2 ± 0.1	63.9 ± 1.8	86.6 ± 0.7	71.2	-3	0
IRM	70.2 ± 0.2	67.6 ± 1.4	85.4 ± 1.2	74.4	-1	-1
MMD	63.4 ± 0.7	68.3 ± 1.0	86.0 ± 0.5	72.5	+2	-1
ANDMask	58.3 ± 0.4	72.2 ± 1.2	86.2 ± 0.2	72.2	-2	-1
IGA	58.7 ± 0.5	70.5 ± 1.2	86.2 ± 0.7	71.8	0	-1
MTL	57.6 ± 0.3	70.2 ± 0.6	87.0 ± 0.7	71.6	-2	-1
VREx	56.3 ± 1.9	71.0 ± 1.3	87.3 ± 0.2	71.5	-1	-1
Mixup	58.4 ± 0.2	66.6 ± 0.9	87.5 ± 0.5	70.8	-2	-1
RSC	58.5 ± 0.2	69.7 ± 0.9	85.9 ± 0.2	71.4	+2	-2
SagNet	58.2 ± 0.3	69.3 ± 1.0	85.8 ± 1.4	71.1	+1	-2
DANN	58.3 ± 0.1	68.6 ± 1.1	86.0 ± 0.4	71.0	-2	-2
MLDG	58.4 ± 0.2	51.6 ± 6.1	85.4 ± 1.3	65.1	-4	-2
CORAL	57.6 ± 0.5	68.3 ± 1.4	86.3 ± 0.5	70.7	-1	-3

Table 1. Performance of domain generalization algorithms on datasets dominated by correlation shift. Note, the CMNIST results here are adopted from DomainBed [1].

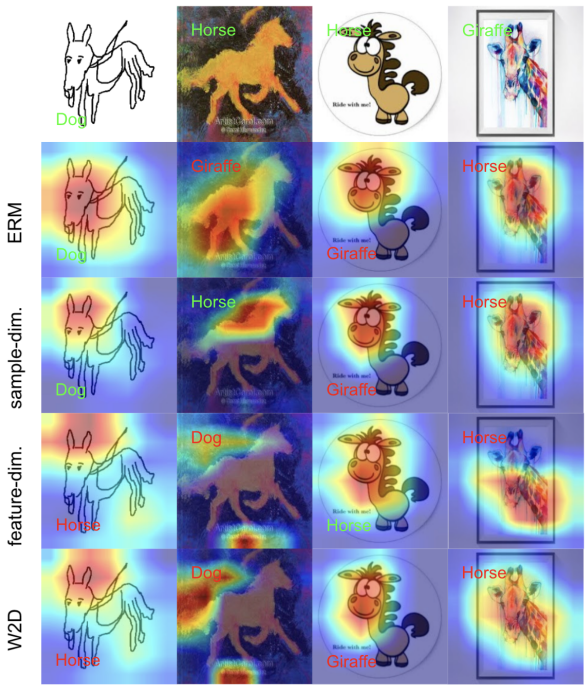

Figure 2. Attention heatmaps of selected algorithms trained and evaluated on PACS [2] visualized by class activation map [4]. Green label means correct prediction and red label means wrong prediction.

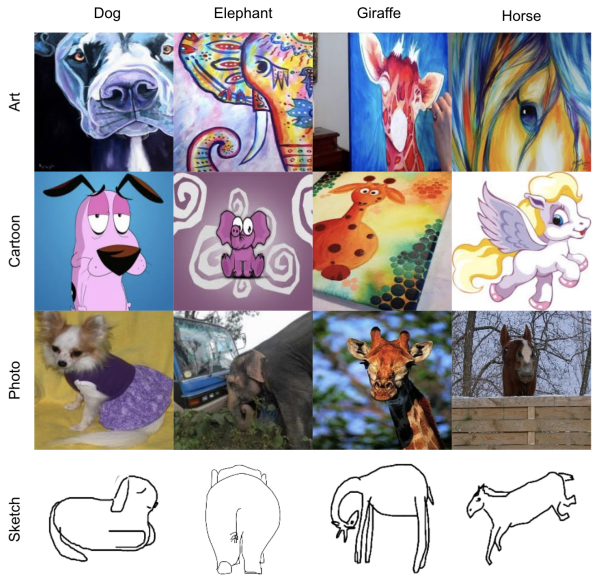

Figure 3. Worst-case samples during training in PACS [2].

datasets and algorithms. *arXiv preprint arXiv:2106.03721*, 2021. 1

[4] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1, 2