Supplementary Material for Localized Adversarial Domain Generalization

Wei Zhu[1,2] * Le Lu[3] Jing Xiao[4] Mei Han[1] Jiebo Luo[2] Adam P. Harrison[1]
[1] PAII Inc. [2] University of Rochester [3] Alibaba DAMO Academy [4] PingAn Insurance Group
{zwvews, tiger.lelu, jiebo.luo, adam.p.harrison}@gmail.com
xiaojing661@pingan.com.cn, hanmei613@paii-labs.com

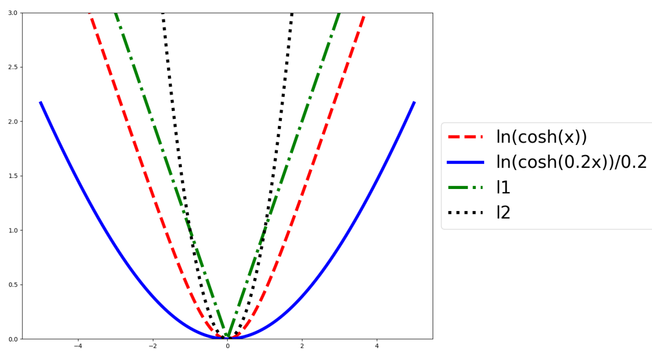


Figure 1. Visualization of log-cosh loss with a hyperparameter ρ.

1. Visualization of log-cosh loss

Since the loss that used to prevent space collapse is calculated with samples in the minibatch, we adopt a Log-Cosh loss considering the random permutation caused by minibatch training. Log-cosh loss imposes a small penalty for small permutation and a hyperparameter ρ is further adopted to smooth the loss curve around zero, which we set to 0.2. We visualize the log-cosh loss in Fig. 1.

2. More Experimental Demonstration on the ADG Incomplete Alignment

In addition to PACS, we provide visualization results on VLCS (Fig. 2) and Camelyon17 (Fig. 4) to show the incomplete alignment of existing adversarial domain generalizaiton methods as DANN [3] and CDANN [10].

3. More Experimental Demonstration on the ADG Space Collapse

In addition to PACS, we provide visualization results on VLCS (Fig. 3) and iWildCam (Fig. 5) to show the space collapse cased by DANN [3] and CDANN [10]. ADG leads to a smaller space collapse for datasets with a large number of domains, *e.g.* iWildCam with 323 domains. This should be attributed to the fact that the large number of domains makes the domain discriminator hard to train and thus cannot exert an significant influence on the representation learning.

*Work was done while Wei Zhu interned at PAII Inc.

4. Experimental Results on Domainbed

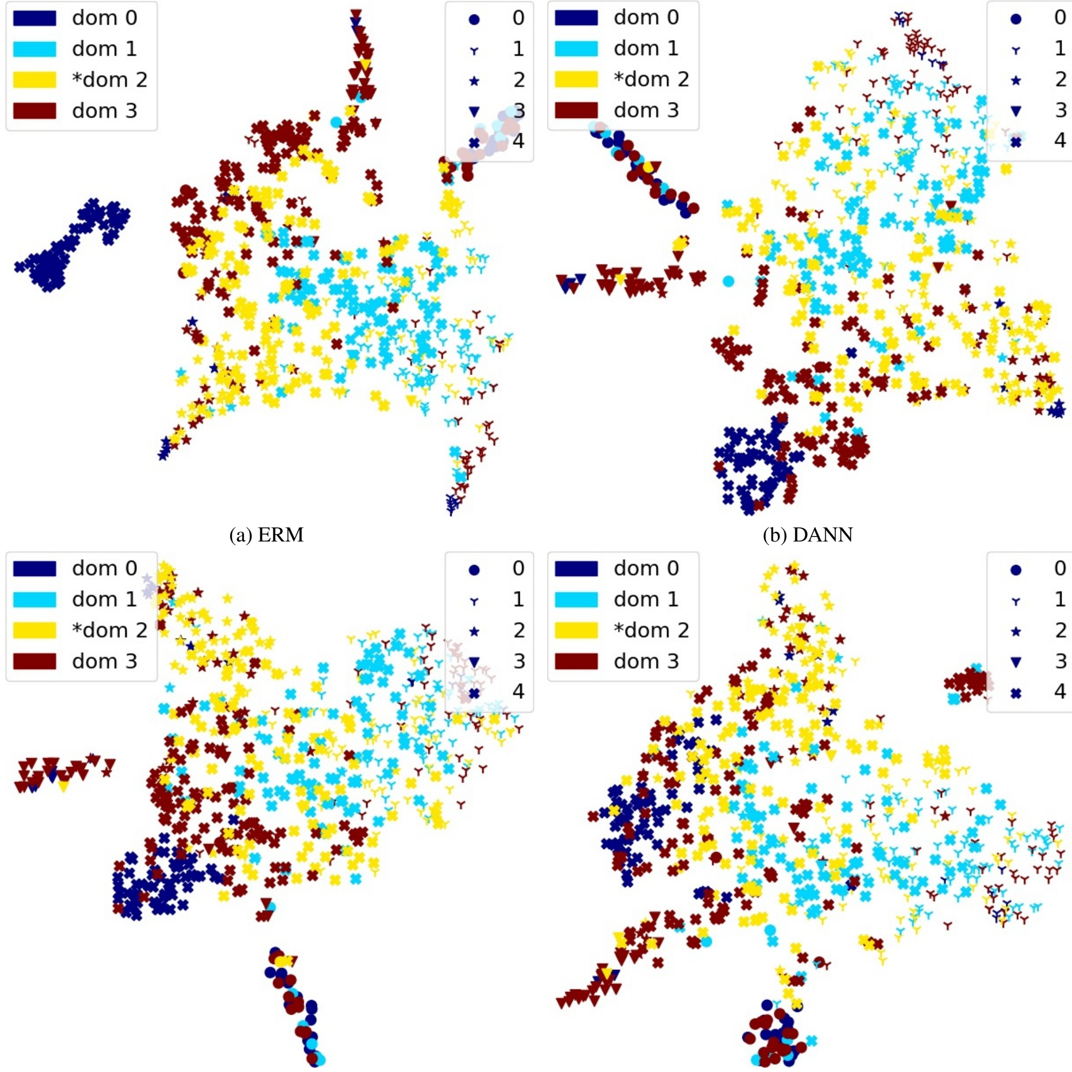
We conduct experiments on DomainBed to validate the effectiveness of our method.

4.1. Experimental Settings

DomainBed contains an extensive set of domain generalization methods, including IRM [1], Group DRO [13], Mixup [16, 17], MLDG [8], CORAL [15], MMD [9], DANN [3], CDANN [10], MTL [2], SagNet [11], ARM [18], V-REx [7], RSC [5], Fish [14], and Fishr [12]. We use the training-domain validation set for model selection, and results for compared methods are retrieved from the DomainBed Benchmark [4] or [12]. We follow the configuration of DomainBed, and run all the experiments with three random seeds. For each seed, we make 20 hyperparameter queries and detail hyperparameter settings are summarized in Sec. 4.3. Note, DomainBed is a collection of synthetic datasets that do not necessarily reflect real-world domain shifts. Moreover, for some datasets the performance of ERM for in-domain vs out-of-domain data are very similar, suggesting that the domain shifts are not significant and that it may not be appropriate to apply approaches other than ERM. Thus, we focus our experimental attention more on the Wilds dataset, which represents real-world datasets with significant domain shifts. Given these issues, our goal with DomainBed is simply to demonstrate that LADG can perform competitively to other DG approaches.

4.2. Results for DomainBed

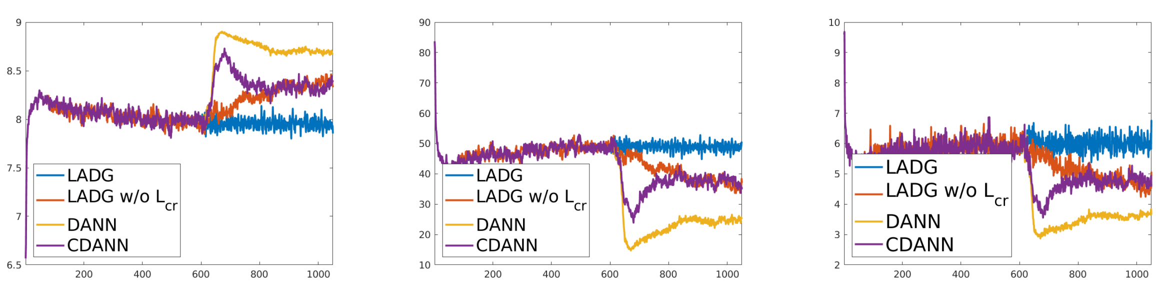
We summarize the results in Table 1 and detailed results on each dataset are shown in Tables 2-8. According to the results, the performance of our method are among the SOTA methods. Moreover, compared to other ADG methods, *i.e.*, DANN [3] and CDANN [10], LADG achieves significant improvements for almost all datasets. For exam-



(a) ERM (b) DANN

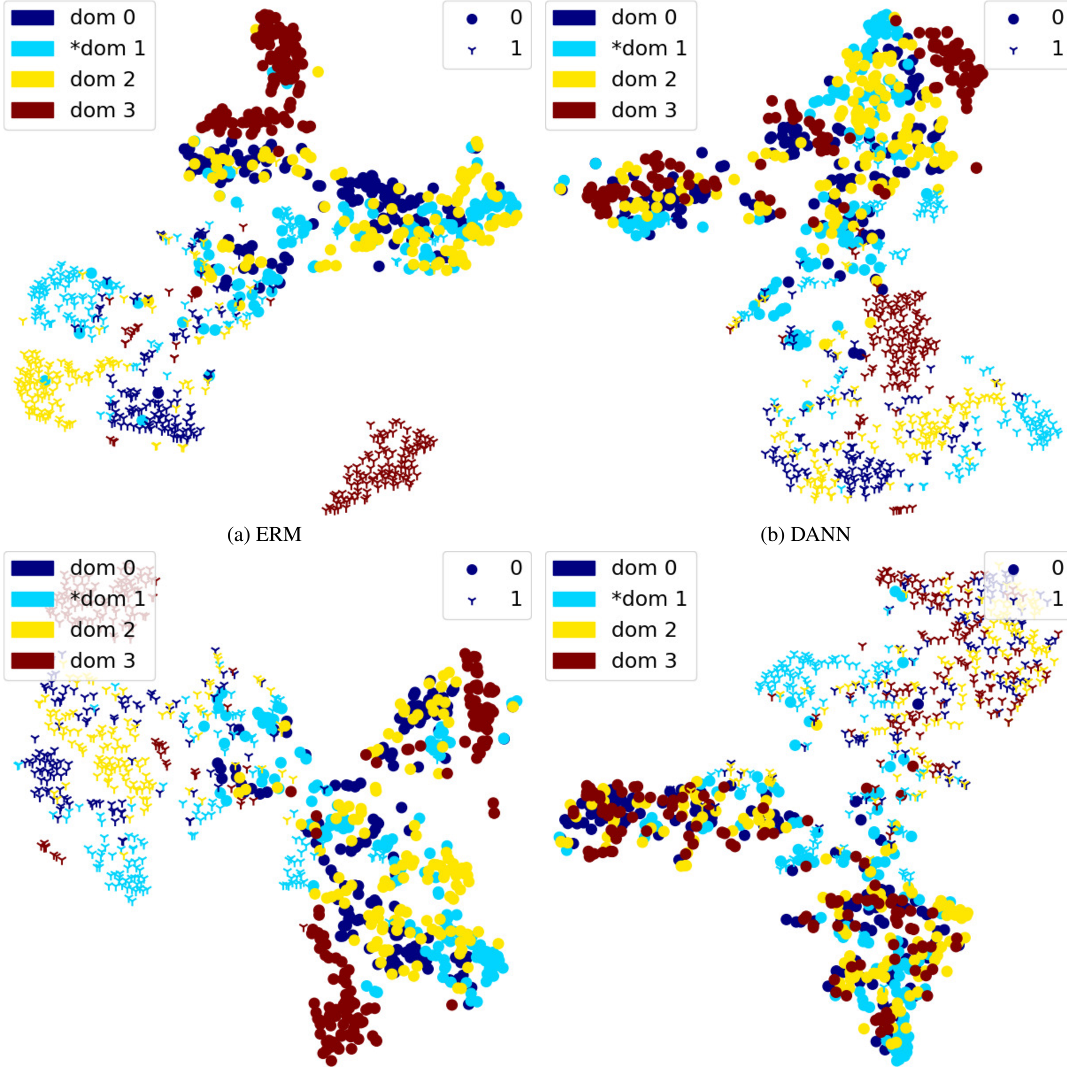
(c) CDANN (d) Ours

Figure 2. Visualization of learned representation on the VLCS dataset. * denotes testing domain. Different shapes (colors) represent different classes (domains).



(a) $V_k(H)$ (b) $R(H)$ (c) $R_C(H)$

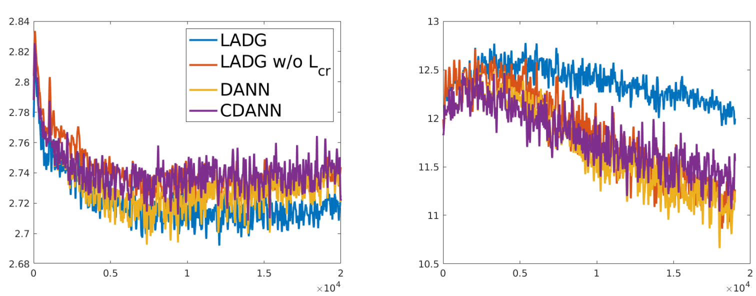
Figure 3. The feature space will collapse with adversarial domain alignment on VLCS. We pretrain the model for 600 steps with ERM and then ADG methods are applied. Our method trained with $\mathcal{L}_{cr}$ could avoid space collapse.



(a) ERM (b) DANN

(c) CDANN (d) Ours

Figure 4. Visualization of learned representation on the Camelyon17 dataset. * denotes OOD validation domain. Different shapes (colors) represent different classes (domains).



(a) $V_k(H)$ (b) $R(H)$

Figure 5. The feature space on iWildCam. We skip the pretraining stage with ERM. iWildCam contains much larger number of domains and the space collapse is not significant.

ple, our method gains 2.3% and 3.3% improvements over DANN and CDANN, respectively, on PACS. This should be attributed to the fine-grained domain alignment by the localized domain discriminator and also the loss to prevent the space collapse.

4.3. Hyperparameter Settings

We pretrain the featurizer ϕ, primary task predictor w, and domain discriminator for our method. To ensure a fair comparison, we subtract the total number of training steps and epochs by that of pretrainining to train our method. The discriminator is composed of a GatedGCN layer and two fully connected layers.

We provide a hyperparameter selection range for DomainBed. We search the τ from $\{1, 2\}$, γ from $\{0.1, 1\}$, and λ from $\{0.1, 0.5\}$. We follow other configurations as the ERM of DomainBed but fix the batchsize as default for all datasets [4]. That is, for non small-scale datasets, we fix the batchsize as 32, and randomly select the learning rate from $10^{\text{Uniform}(-5,-3.5)}$, weight decay from $10^{\text{Uniform}(-6,-2)}$, and dropout rate from $\{0, 0.1, 0.5\}$ following the default settings of DomainBed [4]. For small-scale datasets including variants of MNIST, we fix the batchsize as 64 and randomly select the learning rate from $10^{\text{Uniform}(-4.5,-3.5)}$.

For Wilds [6], we search the τ from $\{1, 2\}$, γ from $\{0.1, 1\}$, and λ from $\{0.1, 0.5\}$. We basically follow the default settings of Wilds for other hyperparameters which are summarized in Table 9.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1

[2] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017. 1

[3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 1

[4] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 1, 4

[5] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 124–140. Springer, 2020. 1

[6] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 4

[7] 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. 1

[8] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1

[9] 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. 1

[10] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017. 1

[11] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 1

[12] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. *arXiv preprint arXiv:2109.02934*, 2021. 1

[13] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 1

[14] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021. 1

[15] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. In *Domain Adaptation in Computer Vision Applications*, pages 153–171. Springer, 2017. 1

[16] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 1

[17] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1

[18] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group distribution shift. *arXiv preprint arXiv:2007.02931*, 2020. 1

Algorithm	Colored MNIST	Rotated MNIST	VLCS	PACS	OfficeHome	TerraIncognita	DomainNet	Avg
ERM	51.5 ± 0.1	98.0 ± 0.0	77.5 ± 0.4	85.5 ± 0.2	66.5 ± 0.3	46.1 ± 1.8	40.9 ± 0.1	66.6
IRM	52.0 ± 0.1	97.7 ± 0.1	78.5 ± 0.5	83.5 ± 0.8	64.3 ± 2.2	47.6 ± 0.8	33.9 ± 2.8	65.4
GroupDRO	52.1 ± 0.0	98.0 ± 0.0	76.7 ± 0.6	84.4 ± 0.8	66.0 ± 0.7	43.2 ± 1.1	33.3 ± 0.2	64.8
Mixup	52.1 ± 0.2	98.0 ± 0.1	77.4 ± 0.6	84.6 ± 0.6	68.1 ± 0.3	47.9 ± 0.8	39.2 ± 0.1	66.7
MLDG	51.5 ± 0.1	97.9 ± 0.0	77.2 ± 0.4	84.9 ± 1.0	66.8 ± 0.6	47.7 ± 0.9	41.2 ± 0.1	66.7
CORAL	51.5 ± 0.1	98.0 ± 0.1	**78.8** ± 0.6	86.2 ± 0.3	**68.7** ± 0.3	47.6 ± 1.0	**41.5** ± 0.1	67.5
MMD	51.5 ± 0.2	97.9 ± 0.0	77.5 ± 0.9	84.6 ± 0.5	66.3 ± 0.1	42.2 ± 1.6	23.4 ± 9.5	63.3
MTL	51.4 ± 0.1	97.9 ± 0.0	77.2 ± 0.4	84.6 ± 0.5	66.4 ± 0.5	45.6 ± 1.2	40.6 ± 0.1	66.2
SagNet	51.7 ± 0.0	98.0 ± 0.0	77.8 ± 0.5	**86.3** ± 0.2	68.1 ± 0.1	**48.6** ± 1.0	40.3 ± 0.1	67.2
ARM	**56.2** ± 0.2	**98.2** ± 0.1	77.6 ± 0.3	85.1 ± 0.4	64.8 ± 0.3	45.5 ± 0.3	35.5 ± 0.2	66.1
VREx	51.8 ± 0.1	97.9 ± 0.1	78.3 ± 0.2	84.9 ± 0.6	66.4 ± 0.6	46.4 ± 0.6	33.6 ± 2.9	65.6
RSC	51.7 ± 0.2	97.6 ± 0.1	77.1 ± 0.5	85.2 ± 0.9	65.5 ± 0.9	46.6 ± 1.0	38.9 ± 0.5	66.1
Fish	51.6 ± 0.1	98.0 ± 0.0	77.8 ± 0.3	85.5 ± 0.3	68.6 ± 0.4	45.1 ± 1.3	42.7 ± 0.2	67.1
Fishr	52.0 ± 0.2	97.8 ± 0.0	77.8 ± 0.1	85.5 ± 0.4	67.8 ± 0.1	47.4 ± 1.6	41.7 ± 0.0	67.1
DANN	51.5 ± 0.3	97.8 ± 0.1	78.6 ± 0.4	83.6 ± 0.4	65.9 ± 0.6	46.7 ± 0.5	38.3 ± 0.1	66.1
CDANN	51.7 ± 0.1	97.9 ± 0.1	77.5 ± 0.1	82.6 ± 0.9	65.8 ± 1.3	45.8 ± 1.6	38.3 ± 0.3	65.6
Ours	52.0 ± 0.2	97.8 ± 0.1	77.7 ± 0.4	85.9 ± 0.8	66.7 ± 1.0	47.7 ± 1.2	40.2 ± 0.6	66.9

Table 1. Results on Domainbed. We use training-domain validation set for model selection.

Algorithm	+90%	+80%	-90%	Avg
ERM	71.7 ± 0.1	72.9 ± 0.2	10.0 ± 0.1	51.5
IRM	72.5 ± 0.1	73.3 ± 0.5	10.2 ± 0.3	52.0
GroupDRO	73.1 ± 0.3	73.2 ± 0.2	10.0 ± 0.2	52.1
Mixup	72.7 ± 0.4	73.4 ± 0.1	10.1 ± 0.1	52.1
MLDG	71.5 ± 0.2	73.1 ± 0.2	9.8 ± 0.1	51.5
CORAL	71.6 ± 0.3	73.1 ± 0.1	9.9 ± 0.1	51.5
MMD	71.4 ± 0.3	73.1 ± 0.2	9.9 ± 0.3	51.5
MTL	70.9 ± 0.2	72.8 ± 0.3	10.5 ± 0.1	51.4
SagNet	71.8 ± 0.2	73.0 ± 0.2	10.3 ± 0.0	51.7
ARM	82.0 ± 0.5	76.5 ± 0.3	10.2 ± 0.0	56.2
VREx	72.4 ± 0.3	72.9 ± 0.4	10.2 ± 0.0	51.8
RSC	71.9 ± 0.3	73.1 ± 0.2	10.0 ± 0.2	51.7
Fish	-	-	-	51.6
Fishr	72.3 ± 0.9	73.5 ± 0.2	10.1 ± 0.2	52.0
DANN	71.4 ± 0.9	73.1 ± 0.1	10.0 ± 0.0	51.5
CDANN	72.0 ± 0.2	73.0 ± 0.2	10.2 ± 0.1	51.7
Ours	72.9 ± 0.3	73.0 ± 0.2	10.0 ± 0.1	52.0

Table 2. Results on Colored MNIST. We use training-domain validation set for model selection.

Rotated MNIST	0	15	30	45	60	75	Avg
ERM	95.9 ± 0.1	98.9 ± 0.0	98.8 ± 0.0	98.9 ± 0.0	98.9 ± 0.0	96.4 ± 0.0	98.0
IRM	95.5 ± 0.1	98.8 ± 0.2	98.7 ± 0.1	98.6 ± 0.1	98.7 ± 0.0	95.9 ± 0.2	97.7
GroupDRO	95.6 ± 0.1	98.9 ± 0.1	98.9 ± 0.1	99.0 ± 0.0	98.9 ± 0.0	96.5 ± 0.2	98.0
Mixup	95.8 ± 0.3	98.9 ± 0.0	98.9 ± 0.0	98.9 ± 0.0	98.8 ± 0.1	96.5 ± 0.3	98.0
MLDG	95.8 ± 0.1	98.9 ± 0.1	99.0 ± 0.0	98.9 ± 0.1	99.0 ± 0.0	95.8 ± 0.3	97.9
CORAL	95.8 ± 0.3	98.8 ± 0.0	98.9 ± 0.0	99.0 ± 0.0	98.9 ± 0.1	96.4 ± 0.2	98.0
MMD	95.6 ± 0.1	98.9 ± 0.1	99.0 ± 0.0	99.0 ± 0.0	98.9 ± 0.0	96.0 ± 0.2	97.9
MTL	95.6 ± 0.1	99.0 ± 0.1	99.0 ± 0.0	98.9 ± 0.1	99.0 ± 0.1	95.8 ± 0.2	97.9
SagNet	95.9 ± 0.3	98.9 ± 0.1	99.0 ± 0.1	99.1 ± 0.0	99.0 ± 0.1	96.3 ± 0.1	98.0
ARM	96.7 ± 0.2	99.1 ± 0.0	99.0 ± 0.0	99.0 ± 0.1	99.1 ± 0.1	96.5 ± 0.4	98.2
VREx	95.9 ± 0.2	99.0 ± 0.1	98.9 ± 0.1	98.9 ± 0.1	98.7 ± 0.1	96.2 ± 0.2	97.9
RSC	94.8 ± 0.5	98.7 ± 0.1	98.8 ± 0.1	98.8 ± 0.0	98.9 ± 0.1	95.9 ± 0.2	97.6
Fish	-	-	-	-	-	-	98.0
Fishr	95.0 ± 0.3	98.5 ± 0.0	99.2 ± 0.1	98.9 ± 0.1	98.9 ± 0.1	96.5 ± 0.0	97.8
DANN	95.0 ± 0.5	98.9 ± 0.1	99.0 ± 0.0	99.0 ± 0.1	98.9 ± 0.0	96.3 ± 0.2	97.8
CDANN	95.7 ± 0.2	98.8 ± 0.0	98.9 ± 0.1	98.9 ± 0.1	98.9 ± 0.1	96.1 ± 0.3	97.9
Ours	94.9 ± 0.2	98.9 ± 0.1	99.1 ± 0.0	99.0 ± 0.1	98.6 ± 0.1	96.4 ± 0.2	97.8

Table 3. Results on Rotated MNIST. We use training-domain validation set for model selection.

Algorithm	A	C	P	S	Avg
ERM	84.7 ± 0.4	80.8 ± 0.6	97.2 ± 0.3	79.3 ± 1.0	85.5
IRM	84.8 ± 1.3	76.4 ± 1.1	96.7 ± 0.6	76.1 ± 1.0	83.5
GroupDRO	83.5 ± 0.9	79.1 ± 0.6	96.7 ± 0.3	78.3 ± 2.0	84.4
Mixup	86.1 ± 0.5	78.9 ± 0.8	97.6 ± 0.1	75.8 ± 1.8	84.6
MLDG	85.5 ± 1.4	80.1 ± 1.7	97.4 ± 0.3	76.6 ± 1.1	84.9
CORAL	88.3 ± 0.2	80.0 ± 0.5	97.5 ± 0.3	78.8 ± 1.3	86.2
MMD	86.1 ± 1.4	79.4 ± 0.9	96.6 ± 0.2	76.5 ± 0.5	84.6
MTL	87.5 ± 0.8	77.1 ± 0.5	96.4 ± 0.8	77.3 ± 1.8	84.6
SagNet	87.4 ± 1.0	80.7 ± 0.6	97.1 ± 0.1	80.0 ± 0.4	86.3
ARM	86.8 ± 0.6	76.8 ± 0.5	97.4 ± 0.3	79.3 ± 1.2	85.1
VREx	86.0 ± 1.6	79.1 ± 0.6	96.9 ± 0.5	77.7 ± 1.7	84.9
RSC	85.4 ± 0.8	79.7 ± 1.8	97.6 ± 0.3	78.2 ± 1.2	85.2
Fish	-	-	-	-	85.5
Fishr	88.4 ± 0.2	78.7 ± 0.7	97.0 ± 0.1	77.8 ± 2.0	85.5
DANN	86.4 ± 0.8	77.4 ± 0.8	97.3 ± 0.4	73.5 ± 2.3	83.6
CDANN	84.6 ± 1.8	75.5 ± 0.9	96.8 ± 0.3	73.5 ± 0.6	82.6
Ours	85.5 ± 0.5	81.3 ± 0.8	97.0 ± 0.9	79.7 ± 1.7	85.9

Table 4. Results on PACS. We use training-domain validation set for model selection.

Algorithm	C	L	S	V	Avg
ERM	97.7 ± 0.4	64.3 ± 0.9	73.4 ± 0.5	74.6 ± 1.3	77.5
IRM	98.6 ± 0.1	64.9 ± 0.9	73.4 ± 0.6	77.3 ± 0.9	78.5
GroupDRO	97.3 ± 0.3	63.4 ± 0.9	69.5 ± 0.8	76.7 ± 0.7	76.7
Mixup	98.3 ± 0.6	64.8 ± 1.0	72.1 ± 0.5	74.3 ± 0.8	77.4
MLDG	97.4 ± 0.2	65.2 ± 0.7	71.0 ± 1.4	75.3 ± 1.0	77.2
CORAL	98.3 ± 0.1	66.1 ± 1.2	73.4 ± 0.3	77.5 ± 1.2	78.8
MMD	97.7 ± 0.1	64.0 ± 1.1	72.8 ± 0.2	75.3 ± 3.3	77.5
MTL	97.8 ± 0.4	64.3 ± 0.3	71.5 ± 0.7	75.3 ± 1.7	77.2
SagNet	97.9 ± 0.4	64.5 ± 0.5	71.4 ± 1.3	77.5 ± 0.5	77.8
ARM	98.7 ± 0.2	63.6 ± 0.7	71.3 ± 1.2	76.7 ± 0.6	77.6
VREx	98.4 ± 0.3	64.4 ± 1.4	74.1 ± 0.4	76.2 ± 1.3	78.3
RSC	97.9 ± 0.1	62.5 ± 0.7	72.3 ± 1.2	75.6 ± 0.8	77.1
Fish	-	-	-	-	77.8
Fishr	98.9 ± 0.3	64.0 ± 0.5	71.5 ± 0.2	76.8 ± 0.7	77.8
DANN	99.0 ± 0.3	65.1 ± 1.4	73.1 ± 0.3	77.2 ± 0.6	78.6
CDANN	97.1 ± 0.3	65.1 ± 1.2	70.7 ± 0.8	77.1 ± 1.5	77.5
Ours	97.6 ± 0.8	66.0 ± 0.2	70.4 ± 2.4	76.8 ± 0.4	77.7

Table 5. Results on VLCS. We use training-domain validation set for model selection.

Algorithm	L100	L38	L43	L46	Avg
ERM	49.8 ± 4.4	42.1 ± 1.4	56.9 ± 1.8	35.7 ± 3.9	46.1
IRM	54.6 ± 1.3	39.8 ± 1.9	56.2 ± 1.8	39.6 ± 0.8	47.6
GroupDRO	41.2 ± 0.7	38.6 ± 2.1	56.7 ± 0.9	36.4 ± 2.1	43.2
Mixup	59.6 ± 2.0	42.2 ± 1.4	55.9 ± 0.8	33.9 ± 1.4	47.9
MLDG	54.2 ± 3.0	44.3 ± 1.1	55.6 ± 0.3	36.9 ± 2.2	47.7
CORAL	51.6 ± 2.4	42.2 ± 1.0	57.0 ± 1.0	39.8 ± 2.9	47.6
MMD	41.9 ± 3.0	34.8 ± 1.0	57.0 ± 1.9	35.2 ± 1.8	42.2
MTL	49.3 ± 1.2	39.6 ± 6.3	55.6 ± 1.1	37.8 ± 0.8	45.6
SagNet	53.0 ± 2.9	43.0 ± 2.5	57.9 ± 0.6	40.4 ± 1.3	48.6
ARM	49.3 ± 0.7	38.3 ± 2.4	55.8 ± 0.8	38.7 ± 1.3	45.5
VREx	48.2 ± 4.3	41.7 ± 1.3	56.8 ± 0.8	38.7 ± 3.1	46.4
RSC	50.2 ± 2.2	39.2 ± 1.4	56.3 ± 1.4	40.8 ± 0.6	46.6
Fish	-	-	-	-	45.1
Fishr	50.2 ± 3.9	43.9 ± 0.8	55.7 ± 2.2	39.8 ± 1.0	47.4
DANN	51.1 ± 3.5	40.6 ± 0.6	57.4 ± 0.5	37.7 ± 1.8	46.7
CDANN	47.0 ± 1.9	41.3 ± 4.8	54.9 ± 1.7	39.8 ± 2.3	45.8
Ours	50.6 ± 3.0	45.5 ± 1.7	55.0 ± 1.4	39.7 ± 3.1	47.7

Table 6. Results on TerraIncognita. We use training-domain validation set for model selection.

Algorithm	A	C	P	R	Avg
ERM	61.3 ± 0.7	52.4 ± 0.3	75.8 ± 0.1	76.6 ± 0.3	66.5
IRM	58.9 ± 2.3	52.2 ± 1.6	72.1 ± 2.9	74.0 ± 2.5	64.3
GroupDRO	60.4 ± 0.7	52.7 ± 1.0	75.0 ± 0.7	76.0 ± 0.7	66.0
Mixup	62.4 ± 0.8	54.8 ± 0.6	76.9 ± 0.3	78.3 ± 0.2	68.1
MLDG	61.5 ± 0.9	53.2 ± 0.6	75.0 ± 1.2	77.5 ± 0.4	66.8
CORAL	65.3 ± 0.4	54.4 ± 0.5	76.5 ± 0.1	78.4 ± 0.5	68.7
MMD	60.4 ± 0.2	53.3 ± 0.3	74.3 ± 0.1	77.4 ± 0.6	66.3
MTL	61.5 ± 0.7	52.4 ± 0.6	74.9 ± 0.4	76.8 ± 0.4	66.4
SagNet	63.4 ± 0.2	54.8 ± 0.4	75.8 ± 0.4	78.3 ± 0.3	68.1
ARM	58.9 ± 0.8	51.0 ± 0.5	74.1 ± 0.1	75.2 ± 0.3	64.8
VREx	60.7 ± 0.9	53.0 ± 0.9	75.3 ± 0.1	76.6 ± 0.5	66.4
RSC	60.7 ± 1.4	51.4 ± 0.3	74.8 ± 1.1	75.1 ± 1.3	65.5
Fish	-	-	-	-	68.6
Fishr	62.4 ± 0.5	54.4 ± 0.4	76.2 ± 0.5	78.3 ± 0.1	67.8
DANN	59.9 ± 1.3	53.0 ± 0.3	73.6 ± 0.7	76.9 ± 0.5	65.9
CDANN	61.5 ± 1.4	50.4 ± 2.4	74.4 ± 0.9	76.6 ± 0.8	65.8
Ours	63.9 ± 1.1	52.5 ± 0.45	73.2 ± 0.6	77.4 ± 0.7	66.7

Table 7. Results on OfficeHome. We use training-domain validation set for model selection.

DomainNet	clip	info	paint	quick	real	sketch	Avg
ERM	58.1 ± 0.3	18.8 ± 0.3	46.7 ± 0.3	12.2 ± 0.4	59.6 ± 0.1	49.8 ± 0.4	40.9
IRM	48.5 ± 2.8	15.0 ± 1.5	38.3 ± 4.3	10.9 ± 0.5	48.2 ± 5.2	42.3 ± 3.1	33.9
GroupDRO	47.2 ± 0.5	17.5 ± 0.4	33.8 ± 0.5	9.3 ± 0.3	51.6 ± 0.4	40.1 ± 0.6	33.3
Mixup	55.7 ± 0.3	18.5 ± 0.5	44.3 ± 0.5	12.5 ± 0.4	55.8 ± 0.3	48.2 ± 0.5	39.2
MLDG	59.1 ± 0.2	19.1 ± 0.3	45.8 ± 0.7	13.4 ± 0.3	59.6 ± 0.2	50.2 ± 0.4	41.2
CORAL	59.2 ± 0.1	19.7 ± 0.2	46.6 ± 0.3	13.4 ± 0.4	59.8 ± 0.2	50.1 ± 0.6	41.5
MMD	32.1 ± 13.3	11.0 ± 4.6	26.8 ± 11.3	8.7 ± 2.1	32.7 ± 13.8	28.9 ± 11.9	23.4
MTL	57.9 ± 0.5	18.5 ± 0.4	46.0 ± 0.1	12.5 ± 0.1	59.5 ± 0.3	49.2 ± 0.1	40.6
SagNet	57.7 ± 0.3	19.0 ± 0.2	45.3 ± 0.3	12.7 ± 0.5	58.1 ± 0.5	48.8 ± 0.2	40.3
ARM	49.7 ± 0.3	16.3 ± 0.5	40.9 ± 1.1	9.4 ± 0.1	53.4 ± 0.4	43.5 ± 0.4	35.5
VREx	47.3 ± 3.5	16.0 ± 1.5	35.8 ± 4.6	10.9 ± 0.3	49.6 ± 4.9	42.0 ± 3.0	33.6
RSC	55.0 ± 1.2	18.3 ± 0.5	44.4 ± 0.6	12.2 ± 0.2	55.7 ± 0.7	47.8 ± 0.9	38.9
Fish	-	-	-	-	-	-	42.7
Fishr	58.2 ± 0.5	20.2 ± 0.2	47.7 ± 0.3	12.7 ± 0.2	60.3 ± 0.2	50.8 ± 0.1	41.7
DANN	53.1 ± 0.2	18.3 ± 0.1	44.2 ± 0.7	11.8 ± 0.1	55.5 ± 0.4	46.8 ± 0.6	38.3
CDANN	54.6 ± 0.4	17.3 ± 0.1	43.7 ± 0.9	12.1 ± 0.7	56.2 ± 0.4	45.9 ± 0.5	38.3
Ours	55.8 ± 0.7	18.5 ± 0.8	46.5 ± 1.2	11.9 ± 0.2	59.4 ± 0.9	49.1 ± 0.7	40.2

Table 8. Results on DomainNet. We use training-domain validation set for model selection.

	iWildCam	Camelyon17	PovertyMap	FMoW	CivilComments	Amazon
batchsize	32	120	64	64	32	16
# domains per batch	4	3	8	4	4	4
Featurizer	resnet50	densenet121	resnet18	densenet121	distilbert	distilbert
lr	3e-5	0.001	0.001	0.0001	1e-5	1e-5

Table 9. Settings for Wilds. We basically follow the default settings of Wilds.