Supplementary Material for Domain Generalization via Rationale Invariance

Liang Chen[1] Yong Zhang[2*] Yibing Song[3] Anton van den Hengel[1] Lingqiao Liu[1*]
[1] The University of Adelaide [2] Tencent AI Lab [3] AI[3] Institute, Fudan University
{liangchen527, zhangyong201303, yibingsong.cv}@gmail.com
{anton.vandenhenge, lingqiao.liu}@adelaide.edu.au

In this supplementary material, we provide,
1. Visualizations of Values from Eq. (3) in the manuscript in Sec. 1;
2. Sensitive analysis regarding the hyper-parameters in Sec. 2;
3. Comparison regarding the setting of combing logit and features in Sec. 3;
4. Evaluations on the DomainBed benchmark using the ResNet50 backbone in Sec. 4;
5. Detailed results in the DomainBed benchmark in Sec. 5.

1. Visualizations of Values from Eq. (3)

In this section, we plot the changes in the sample-to-center-difference (SCD) values for rationales, features, and logits in Fig. 1 (a)-(c) in settings of with and without $\mathcal{L}_{inv}$. Our observations are as follows: (1) Using $\mathcal{L}_{inv}$ tends to decrease the three SCD values, which is significant compared to disabling $\mathcal{L}_{inv}$. The results indicate that ERM fails to summarize shared clues to make a robust decision for samples from the same class, explaining why it is less effective in generalizing than ours. (2) When compared to the case of rationales, features, and logits, the SCD values exhibit larger variances throughout iterations, indicating that our $\mathcal{L}_{inv}$ allows for some flexibility, enabling features and logits to deviate from their centers. This observation aligns with our suggestion: the contribution of each feature dimension should be jointly modulated by both the feature itself and its corresponding classifier weight.

We perform vectorization on rationale matrices for different samples and use t-SNE for dimension reduction. In Fig. 1 (d), we show that, rationales from different domains will be mixed together when with $\mathcal{L}_{inv}$, indicating the adopted $\mathcal{L}_{inv}$ can ensure using the same rational for samples from the same class despite the varying domains.

2. Sensitive Analysis Regarding the Hyper-Parameter Settings

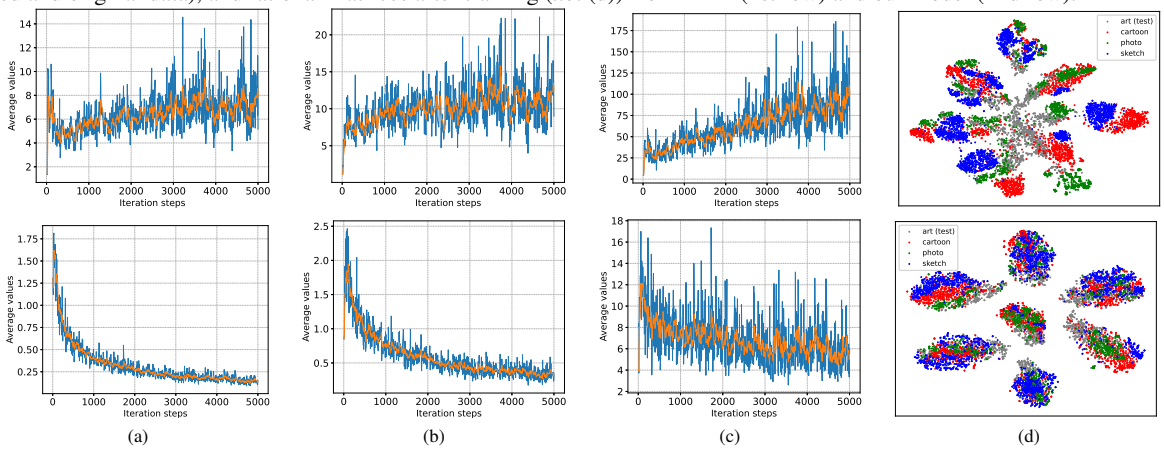
Our implementation involves two hyper-parameters: the momentum value m in Eq. (4) and the positive weight α in Eq. (5) in the manuscript. This section evaluates our method with different settings of these two hyper-parameters by conducting experiments on the widely-used PACS dataset [13] with a ResNet18 backbone [9] using the same setting illustrated in Sec. 4.1 in the manuscript, similar to that in [5]. Note we fix the value for one hyper-parameter when analyzing another. Results are listed in Table 1. We observe that our method performs consistently well when the hyper-parameter m in the range of $[0.0001, 0.1]$ and α in the range of $[0.001, 0.1]$.

3. Comparisons with the Setting Combining Logit and Feature

As stated in the manuscript, analyzing the decision-making process from either the perspective of feature or logit has intrinsic limitations. Specifically, since the classifier is not taken into account, the model may emphasize heavily on feature elements that with large values but correspond to small weights in the classifier if only consider the feature. Although logit can ease the issue to a certain extent, it only provides a coarse representation for the decision-making process, thus difficult to ensure robust outputs. One may wonder if the combination of feature and logit could avoid the limitation of each other and lead to certain improvements. To answer this question, we conduct further analysis by substituting the rationale invariance constraint with the regularization term that enforces invariance for both the feature and logit (*i.e.* **W/ fea. & log.**), which

*Corresponding authors. This work is done when L. Chen is an intern in Tencent AI Lab.

Figure 1. Plots of different terms by iterations (*i.e.* Values of Eq. (3), feature, and logit differences in (a)-(c), yellow and blue lines are smoothed and original data), and rational matrices after training (*i.e.* (d)) from ERM (1st row) and our model (2nd row).



(a) (b) (c) (d)

Table 1. Evaluations regarding different hyper-parameter (*i.e.* m in Eq. (4) and α in Eq. (5) from the manuscript) settings. We fix one parameter and tune another when conducting the experiments which are examined in PACS [13] with the leave-one-out training-test strategy. The reported accuracies (%) and standard deviations are computed from 60 trials in each target domain.

hyper-parameters		art	cartoon	photo	sketch	avg
$\alpha \in [0.001, 0.1]$	$m = 0$	82.3 ± 0.1	73.4 ± 1.2	95.0 ± 0.5	75.8 ± 0.8	81.6 ± 0.4
	$m = 0.0001$	82.3 ± 0.5	76.0 ± 0.5	94.6 ± 0.4	75.9 ± 1.1	82.2 ± 0.4
	$m = 0.001$	82.1 ± 1.4	75.5 ± 1.2	94.9 ± 0.5	76.5 ± 0.3	82.2 ± 0.5
	$m = 0.01$	82.9 ± 0.8	76.2 ± 1.2	94.6 ± 0.6	75.9 ± 1.5	82.4 ± 5.9
	$m = 0.1$	82.2 ± 0.7	75.9 ± 0.9	95.3 ± 0.2	78.1 ± 0.9	82.9 ± 5.9
	$m = 1$	81.6 ± 1.9	76.2 ± 0.6	94.9 ± 0.4	75.3 ± 1.6	82.0 ± 0.3
$m \in [0.0001, 0.1]$	$\alpha = 0.0001$	78.3 ± 0.6	74.2 ± 1.8	94.0 ± 0.4	76.4 ± 2.4	80.7 ± 0.8
	$\alpha = 0.001$	82.3 ± 0.5	74.7 ± 1.4	93.7 ± 0.8	75.5 ± 0.3	81.6 ± 0.4
	$\alpha = 0.01$	81.9 ± 1.0	75.0 ± 1.1	94.9 ± 0.3	75.7 ± 0.4	81.9 ± 0.1
	$\alpha = 0.1$	82.3 ± 0.8	76.2 ± 0.5	94.9 ± 0.4	76.4 ± 1.0	82.4 ± 0.4
	$\alpha = 1$	81.6 ± 0.7	74.0 ± 0.5	94.5 ± 0.3	73.5 ± 1.4	80.9 ± 0.2

reformulates Eq. (3) into $\mathcal{L}_{inv} = \frac{1}{N_b} \sum_k \sum_{\{n|y_n=k\}} (\|\mathbf{z}_n - \overline{\mathbf{z}}_k\|^2 + \|\mathbf{o}_n - \overline{\mathbf{o}}_k\|^2)$, where $\mathbf{z}$, $\mathbf{o}$, $\overline{\mathbf{z}}$ and $\overline{\mathbf{o}}$ are the feature, logit, and their corresponding momentum updated mean values, respectively. We use the same setting as the original design and test the model in the widely-used PACS dataset [13] to evaluate its effectiveness.

Experimental results are listed in Table 2. We note that combining the feature and logit can lead to improvements for both the two invariance constraints (*i.e.* W/ fea. and W/ log.) in almost all target domains. This finding is not surprising since the combined setting considers both the classifier and the feature, thereby mitigating some of the limitations of the two individual settings. However, our rationale invariance regularization still outperforms the combined approach. This is because our rationale concept provides a direct characterization of the decision-making process, encompassing the fine-grained representations of both the features and the weights in the classifier, while the latter can only be coarsely represented in the combined setting.

4. Results in the DomainBed Benchmark with the ResNet50 Backbone

To comprehensively examine the effectiveness of the proposed method, we also evaluate our method and the baseline ERM and the top-3 arts in Table 1 in the manuscript using the larger ResNet50 backbone [9]. Results are listed in Table 3, which are directly cited from [11]. We note that our method surpasses the baseline ERM model in all datasets and leads by 1.4 in average, while the top 3 methods of the compared models (*i.e.* CORAL [24], SagNet [17], and SelfReg [11]) lead their ERM model by 1.3, 0.9 and 0.9 in average. These results indicate that our method can consistently improve the baseline model and perform favorably against existing arts when implemented with a larger ResNet50 backbone.

Table 2. Comparison between different invariance constraint and mean value updating schemes in the unseen domain from the PACS benchmark [13]. Here the "Z", "O", and "R" denotes the feature-invariance, logits-invariance, and the proposed rationale invariance constraints. The reported accuracies (%) and standard deviations are computed from 60 trials in each target domain.

Models	Invariance			Target domain				Avg.
	Z	O	R	Art	Cartoon	Photo	Sketch	
ERM	–	–	–	78.0 ± 1.3	73.4 ± 0.8	94.1 ± 0.4	73.6 ± 2.2	79.8 ± 0.4
W/ fea.	✓	–	–	82.3 ± 0.4	74.3 ± 1.0	94.3 ± 0.4	73.6 ± 1.3	81.1 ± 0.5
W/ log.	–	✓	–	81.9 ± 0.5	75.5 ± 0.5	94.8 ± 0.2	73.9 ± 1.3	81.5± 0.3
W/ fea & log.	✓	✓	–	82.2 ± 1.0	75.8 ± 0.6	95.6 ± 0.4	74.7 ± 0.8	82.1 ± 0.4
Ours	–	–	✓	82.4 ± 1.0	76.7 ± 0.6	95.3 ± 0.1	76.7 ± 0.3	82.8 ± 0.3

Table 3. Average accuracies on the DomainBed [8] benchmark using the ResNet50 [9] backbones. Results without † are directly cited from a previous work [11]. Improve. denotes the average improvements with respect to the corresponding ERM model.

	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.	Improve.
ERM [25]	85.5 ± 0.2	77.5 ± 0.4	66.5 ± 0.3	46.1 ± 1.8	40.9 ± 0.1	63.3	-
IRM [1]	83.5 ± 0.8	78.5 ± 0.5	64.3 ± 2.2	47.6 ± 0.8	33.9 ± 2.8	61.2	-2.1
GroupGRO [22]	84.4 ± 0.8	76.7 ± 0.6	66.0 ± 0.7	43.2 ± 1.1	33.3 ± 0.2	60.7	-2.6
Mixup [27]	84.6 ± 0.6	77.4 ± 0.6	68.1 ± 0.3	47.9 ± 0.8	39.2 ± 0.1	63.4	+0.1
MLDG [14]	84.9 ± 1.0	77.2 ± 0.4	66.8 ± 0.6	47.7 ± 0.9	41.2 ± 0.1	63.6	+0.3
CORAL [24]	86.2 ± 0.3	78.8 ± 0.6	68.7 ± 0.3	47.6 ± 1.0	41.5 ± 0.1	64.6	+1.3
MMD [15]	84.6 ± 0.5	77.5 ± 0.9	66.3 ± 0.1	42.2 ± 1.6	23.4 ± 9.5	58.8	-4.5
DANN [7]	83.6 ± 0.4	78.6 ± 0.4	65.9 ± 0.6	46.7 ± 0.5	38.3 ± 0.1	62.6	-0.7
CDANN [16]	82.6 ± 0.9	77.5 ± 0.1	65.8 ± 1.3	45.8 ± 1.6	38.3 ± 0.3	62.0	-1.3
MTL [3]	84.6 ± 0.5	77.2 ± 0.4	66.4 ± 0.5	45.6 ± 1.2	40.6 ± 0.1	62.9	-0.4
SagNet [17]	86.3 ± 0.2	77.8 ± 0.5	68.1 ± 0.1	48.6 ± 1.0	40.3 ± 0.1	64.2	+0.9
ARM [28]	85.1 ± 0.4	77.6 ± 0.3	64.8 ± 0.3	45.5 ± 0.3	35.5 ± 0.2	61.7	-1.6
VREx [12]	84.9 ± 0.6	78.3 ± 0.2	66.4 ± 0.6	46.4 ± 0.6	33.6 ± 2.9	61.9	-2.4
RSC [10]	85.2 ± 0.9	77.1 ± 0.5	65.5 ± 0.9	46.6 ± 1.0	38.9 ± 0.5	62.7	-0.6
SelfReg [11]	85.6 ± 0.4	77.8 ± 0.9	67.9 ± 0.7	47.0 ± 0.3	42.8 ± 0.0	64.2	+0.9
ERM† [25]	83.1 ± 0.9	77.7 ± 0.8	65.8 ± 0.3	46.5 ± 0.9	40.8 ± 0.2	62.8	-
Fish† [23]	84.0±0.3	78.6±0.1	67.9±0.5	46.6±0.4	40.6±0.2	63.5	+0.7
CORAL† [24]	85.0±0.4	77.9±0.2	68.8±0.3	46.1±1.2	41.4±0.0	63.9	+1.1
SD† [19]	84.4±0.2	77.6±0.4	68.9±0.2	46.4±2.0	42.0±0.2	63.9	+1.1
Ours†	84.7 ± 0.2	77.8 ± 0.4	68.6 ± 0.2	47.8 ± 1.1	41.9 ± 0.3	64.2	+1.4

5. Detailed Results in the DomainBed Benchmark [8]

this section presents the average accuracy in each domain from different datasets. As shown in Table 4, 5, 6, 7, and 8, these results are detailed illustrations of the results in Table 1 in our manuscript. For all the experiments, we use the "training-domain validate set" as the model selection method. A total of 23 methods are examined for 60 trials in each unseen domain, and all methods are trained with the leave-one-out strategy using the ResNet18 [9] backbones.

We note the MIRO [4] method performs inferior to other arts when evaluated in the PACS dataset, this is mainly because their approach specifically enforces similarity between intermediate features from the model and that from the pretrained backbone, which can be detrimental to the performance when there is a significant distribution shift between the target data and samples used for pertaining. In this case, the distribution shift is particularly noticeable between data from the 'cartoon' and 'sketch' domains and real photos in imagenet that are adopted for pretraining.

Table 4. Average accuracies on the PACS [13] datasets using the default hyper-parameter settings in DomainBed [8].

	art	cartoon	photo	sketch	Average
ERM [25]	78.0 ± 1.3	73.4 ± 0.8	94.1 ± 0.4	73.6 ± 2.2	79.8 ± 0.4
IRM [1]	76.9 ± 2.6	75.1 ± 0.7	94.3 ± 0.4	77.4 ± 0.4	80.9 ± 0.5
GroupGRO [22]	77.7 ± 2.6	76.4 ± 0.3	94.0 ± 0.3	74.8 ± 1.3	80.7 ± 0.4
Mixup [27]	79.3 ± 1.1	74.2 ± 0.3	94.9 ± 0.3	68.3 ± 2.7	79.2 ± 0.9
MLDG [14]	78.4 ± 0.7	75.1 ± 0.5	94.8 ± 0.4	76.7 ± 0.8	81.3 ± 0.2
CORAL [24]	81.5 ± 0.5	75.4 ± 0.7	95.2 ± 0.5	74.8 ± 0.4	81.7 ± 0.0
MMD [15]	81.3 ± 0.6	75.5 ± 1.0	94.0 ± 0.5	74.3 ± 1.5	81.3 ± 0.8
DANN [7]	79.0 ± 0.6	72.5 ± 0.7	94.4 ± 0.5	70.8 ± 3.0	79.2 ± 0.3
CDANN [16]	80.4 ± 0.8	73.7 ± 0.3	93.1 ± 0.6	74.2 ± 1.7	80.3 ± 0.5
MTL [3]	78.7 ± 0.6	73.4 ± 1.0	94.1 ± 0.6	74.4 ± 3.0	80.1 ± 0.8
SagNet [17]	82.9 ± 0.4	73.2 ± 1.1	94.6 ± 0.5	76.1 ± 1.8	81.7 ± 0.6
ARM [28]	79.4 ± 0.6	75.0 ± 0.7	94.3 ± 0.6	73.8 ± 0.6	80.6 ± 0.5
VREx [12]	74.4 ± 0.7	75.0 ± 0.4	93.3 ± 0.3	78.1 ± 0.9	80.2 ± 0.5
RSC [10]	78.5 ± 1.1	73.3 ± 0.9	93.6 ± 0.6	76.5 ± 1.4	80.5 ± 0.2
SelfReg [11]	82.5 ± 0.8	74.4 ± 1.5	95.4 ± 0.5	74.9 ± 1.3	81.8 ± 0.3
MixStyle [29]	82.6 ± 1.2	76.3 ± 0.4	94.2 ± 0.3	77.5 ± 1.3	82.6 ± 0.4
Fish [23]	80.9 ± 1.0	75.9 ± 0.4	95.0 ± 0.4	76.2 ± 1.0	82.0 ± 0.3
SD [19]	83.2 ± 0.6	74.6 ± 0.3	94.6 ± 0.1	75.1 ± 1.6	81.9 ± 0.3
CAD [21]	83.9 ± 0.8	74.2 ± 0.4	94.6 ± 0.4	75.0 ± 1.2	81.9 ± 0.3
CondCAD [21]	79.7 ± 1.0	74.2 ± 0.9	94.6 ± 0.4	74.8 ± 1.4	80.8 ± 0.5
Fishr [20]	81.2 ± 0.4	75.8 ± 0.8	94.3 ± 0.3	73.8 ± 0.6	81.3 ± 0.3
MIRO [4]	79.3 ± 0.6	68.1 ± 2.5	95.5 ± 0.3	60.6 ± 3.1	75.9 ± 1.4
Ours	82.4 ± 1.0	76.7 ± 0.6	95.3 ± 0.1	76.7 ± 0.3	82.8 ± 0.3

Table 5. Average accuracies on the VLCS [6] datasets using the default hyper-parameter settings in DomainBed [8].

	Caltech	LabelMe	Sun	VOC	Average
ERM [25]	97.7 ± 0.3	62.1 ± 0.9	70.3 ± 0.9	73.2 ± 0.7	75.8 ± 0.2
IRM [1]	96.1 ± 0.8	62.5 ± 0.3	69.9 ± 0.7	72.0 ± 1.4	75.1 ± 0.1
GroupGRO [22]	96.7 ± 0.6	61.7 ± 1.5	70.2 ± 1.8	72.9 ± 0.6	75.4 ± 1.0
Mixup [27]	95.6 ± 1.5	62.7 ± 0.4	71.3 ± 0.3	75.4 ± 0.2	76.2 ± 0.3
MLDG [14]	95.8 ± 0.5	63.3 ± 0.8	68.5 ± 0.5	73.1 ± 0.8	75.2 ± 0.3
CORAL [24]	96.5 ± 0.3	62.8 ± 0.1	69.1 ± 0.6	73.8 ± 1.0	75.5 ± 0.4
MMD [15]	96.0 ± 0.8	64.3 ± 0.6	68.5 ± 0.6	70.8 ± 0.1	74.9 ± 0.5
DANN [7]	97.2 ± 0.1	63.3 ± 0.6	70.2 ± 0.9	74.4 ± 0.2	76.3 ± 0.2
CDANN [16]	95.4 ± 1.2	62.6 ± 0.6	69.9 ± 1.3	76.2 ± 0.5	76.0 ± 0.5
MTL [3]	94.4 ± 2.3	65.0 ± 0.6	69.6 ± 0.6	71.7 ± 1.3	75.2 ± 0.3
SagNet [17]	94.9 ± 0.7	61.9 ± 0.7	69.6 ± 1.3	75.2 ± 0.6	75.4 ± 0.8
ARM [28]	96.9 ± 0.5	61.9 ± 0.4	71.6 ± 0.1	73.3 ± 0.4	75.9 ± 0.3
VREx [12]	96.2 ± 0.0	62.5 ± 1.3	69.3 ± 0.9	73.1 ± 1.2	75.3 ± 0.6
RSC [10]	96.2 ± 0.0	63.6 ± 1.3	69.8 ± 1.0	72.0 ± 0.4	75.4 ± 0.3
SelfReg [11]	95.8 ± 0.6	63.4 ± 1.1	71.1 ± 0.6	75.3 ± 0.6	76.4 ± 0.7
MixStyle [29]	97.3 ± 0.3	61.6 ± 0.1	70.4 ± 0.7	71.3 ± 1.9	75.2 ± 0.7
Fish [23]	97.4 ± 0.2	63.4 ± 0.1	71.5 ± 0.4	75.2 ± 0.7	76.9 ± 0.2
SD [19]	96.5 ± 0.4	62.2 ± 0.0	69.7 ± 0.9	73.6 ± 0.4	75.5 ± 0.4
CAD [21]	94.5 ± 0.9	63.5 ± 0.6	70.4 ± 1.2	72.4 ± 1.3	75.2 ± 0.6
CondCAD [21]	96.5 ± 0.8	62.6 ± 0.4	69.1 ± 0.2	76.0 ± 0.2	76.1 ± 0.3
Fishr [20]	97.2 ± 0.6	63.3 ± 0.7	70.4 ± 0.6	74.0 ± 0.8	76.2 ± 0.3
MIRO [4]	97.5 ± 0.2	62.0 ± 0.5	71.3 ± 1.0	74.8 ± 0.6	76.4 ± 0.4
Ours	96.7 ± 0.5	63.2 ± 1.0	70.3 ± 0.8	73.4 ± 0.3	75.9 ± 0.3

Table 6. Average accuracies on the OfficeHome [26] datasets using the default hyper-parameter settings in DomainBed [8].

	art	clipart	product	real	Average
ERM [25]	52.2 ± 0.2	48.7 ± 0.5	69.9 ± 0.5	71.7 ± 0.5	60.6 ± 0.2
IRM [1]	49.7 ± 0.2	46.8 ± 0.5	67.5 ± 0.4	68.1 ± 0.6	58.0 ± 0.1
GroupGRO [22]	52.6 ± 1.1	48.2 ± 0.9	69.9 ± 0.4	71.5 ± 0.8	60.6 ± 0.3
Mixup [27]	54.0 ± 0.7	49.3 ± 0.7	70.7 ± 0.7	72.6 ± 0.3	61.7 ± 0.5
MLDG [14]	53.1 ± 0.3	48.4 ± 0.3	70.5 ± 0.7	71.7 ± 0.4	60.9 ± 0.2
CORAL [24]	55.1 ± 0.7	49.7 ± 0.9	71.8 ± 0.2	73.1 ± 0.5	62.4 ± 0.4
MMD [15]	50.9 ± 1.0	48.7 ± 0.3	69.3 ± 0.7	70.7 ± 1.3	59.9 ± 0.4
DANN [7]	51.8 ± 0.5	47.1 ± 0.1	69.1 ± 0.7	70.2 ± 0.7	59.5 ± 0.5
CDANN [16]	51.4 ± 0.5	46.9 ± 0.6	68.4 ± 0.5	70.4 ± 0.4	59.3 ± 0.4
MTL [3]	51.6 ± 1.5	47.7 ± 0.5	69.1 ± 0.3	71.0 ± 0.6	59.9 ± 0.5
SagNet [17]	55.3 ± 0.4	49.6 ± 0.2	72.1 ± 0.4	73.2 ± 0.4	62.5 ± 0.3
ARM [28]	51.3 ± 0.9	48.5 ± 0.4	68.0 ± 0.3	70.6 ± 0.1	59.6 ± 0.3
VREx [12]	51.1 ± 0.3	47.4 ± 0.6	69.0 ± 0.4	70.5 ± 0.4	59.5 ± 0.1
RSC [10]	49.0 ± 0.1	46.2 ± 1.5	67.8 ± 0.7	70.6 ± 0.3	58.4 ± 0.6
SelfReg [11]	55.1 ± 0.8	49.2 ± 0.6	72.2 ± 0.3	73.0 ± 0.3	62.4 ± 0.1
MixStyle [29]	50.8 ± 0.6	51.4 ± 1.1	67.6 ± 1.3	68.8 ± 0.5	59.6 ± 0.8
Fish [23]	54.6 ± 1.0	49.6 ± 1.0	71.3 ± 0.6	72.4 ± 0.2	62.0 ± 0.6
SD [19]	55.0 ± 0.4	51.3 ± 0.5	72.5 ± 0.2	72.7 ± 0.3	62.9 ± 0.2
CAD [21]	52.1 ± 0.6	48.3 ± 0.5	69.7 ± 0.3	71.9 ± 0.4	60.5 ± 0.3
CondCAD [21]	53.3 ± 0.6	48.4 ± 0.2	69.8 ± 0.9	72.6 ± 0.1	61.0 ± 0.4
Fishr [20]	52.6 ± 0.9	48.6 ± 0.3	69.9 ± 0.6	72.4 ± 0.4	60.9 ± 0.3
MIRO [4]	57.4 ± 0.9	49.5 ± 0.3	74.0 ± 0.1	75.6 ± 0.2	64.1 ± 0.4
Ours	56.6 ± 0.7	50.3 ± 0.6	72.5 ± 0.0	73.8 ± 0.3	63.3 ± 0.1

Table 7. Average accuracies on the TerraInc [2] datasets using the default hyper-parameter settings in DomainBed [8].

	L100	L38	L43	L46	Average
ERM [25]	42.1 ± 2.5	30.1 ± 1.2	48.9 ± 0.6	34.0 ± 1.1	38.8 ± 1.0
IRM [1]	41.8 ± 1.8	29.0 ± 3.6	49.6 ± 2.1	33.1 ± 1.5	38.4 ± 0.9
GroupGRO [22]	45.3 ± 4.6	36.1 ± 4.4	51.0 ± 0.8	33.7 ± 0.9	41.5 ± 2.0
Mixup [27]	49.4 ± 2.0	35.9 ± 1.8	53.0 ± 0.7	30.0 ± 0.9	42.1 ± 0.7
MLDG [14]	39.6 ± 2.3	33.2 ± 2.7	52.4 ± 0.5	35.1 ± 1.5	40.1 ± 0.9
CORAL [24]	46.7 ± 3.2	36.9 ± 4.3	49.5 ± 1.9	32.5 ± 0.7	41.4 ± 1.8
MMD [15]	49.1 ± 1.2	36.4 ± 4.8	50.4 ± 2.1	32.3 ± 1.5	42.0 ± 1.0
DANN [7]	44.3 ± 3.6	28.0 ± 1.5	47.9 ± 1.0	31.3 ± 0.6	37.9 ± 0.9
CDANN [16]	36.9 ± 6.4	32.7 ± 6.2	51.1 ± 1.3	33.5 ± 0.5	38.6 ± 2.3
MTL [3]	45.2 ± 2.6	31.0 ± 1.6	50.6 ± 1.1	34.9 ± 0.4	40.4 ± 1.0
SagNet [17]	36.3 ± 4.7	40.3 ± 2.0	52.5 ± 0.6	33.3 ± 1.3	40.6 ± 1.5
ARM [28]	41.5 ± 4.5	27.7 ± 2.4	50.9 ± 1.0	29.6 ± 1.5	37.4 ± 1.9
VREx [12]	48.0 ± 1.7	41.1 ± 1.5	51.8 ± 1.5	32.0 ± 1.2	43.2 ± 0.3
RSC [10]	42.8 ± 2.4	32.2 ± 3.8	49.6 ± 0.9	32.9 ± 1.2	39.4 ± 1.3
SelfReg [11]	46.1 ± 1.5	34.5 ± 1.6	49.8 ± 0.3	34.7 ± 1.5	41.3 ± 0.3
MixStyle [29]	50.6 ± 1.9	28.0 ± 4.5	52.1 ± 0.7	33.0 ± 0.2	40.9 ± 1.1
Fish [23]	46.3 ± 3.0	29.0 ± 1.1	52.7 ± 1.2	32.8 ± 1.0	40.2 ± 0.6
SD [19]	45.5 ± 1.9	33.2 ± 3.1	52.9 ± 0.7	36.4 ± 0.8	42.0 ± 1.0
CAD [21]	43.1 ± 2.6	31.1 ± 1.9	53.1 ± 1.6	34.7 ± 1.3	40.5 ± 0.4
CondCAD [21]	44.4 ± 2.9	32.9 ± 2.5	50.5 ± 1.3	30.8 ± 0.5	39.7 ± 0.4
Fishr [20]	49.9 ± 3.3	36.6 ± 0.9	49.8 ± 0.2	34.2 ± 1.3	42.6 ± 1.0
MIRO [4]	46.0 ± 0.7	34.4 ± 0.4	51.2 ± 1.0	33.6 ± 0.9	41.3 ± 0.2
Ours	46.2 ± 4.0	39.7 ± 2.4	53.0 ± 0.6	36.0 ± 0.3	43.7 ± 0.5

Table 8. Average accuracies on the DomainNet [18] datasets using the default hyper-parameter settings in DomainBed [8].

	clip	info	paint	quick	real	sketch	Average
ERM [25]	50.4 ± 0.2	14.0 ± 0.2	40.3 ± 0.5	11.7 ± 0.2	52.0 ± 0.2	43.2 ± 0.3	35.3 ± 0.1
IRM [1]	43.2 ± 0.9	12.6 ± 0.3	35.0 ± 1.4	9.9 ± 0.4	43.4 ± 3.0	38.4 ± 0.4	30.4 ± 1.0
GroupGRO [22]	38.2 ± 0.5	13.0 ± 0.3	28.7 ± 0.3	8.2 ± 0.1	43.4 ± 0.5	33.7 ± 0.0	27.5 ± 0.1
Mixup [27]	48.9 ± 0.3	13.6 ± 0.3	39.5 ± 0.5	10.9 ± 0.4	49.9 ± 0.2	41.2 ± 0.2	34.0 ± 0.0
MLDG [14]	51.1 ± 0.3	14.1 ± 0.3	40.7 ± 0.3	11.7 ± 0.1	52.3 ± 0.3	42.7 ± 0.2	35.4 ± 0.0
CORAL [24]	51.2 ± 0.2	15.4 ± 0.2	42.0 ± 0.2	12.7 ± 0.1	52.0 ± 0.3	43.4 ± 0.0	36.1 ± 0.2
MMD [15]	16.6 ± 13.3	0.3 ± 0.0	12.8 ± 10.4	0.3 ± 0.0	17.1 ± 13.7	0.4 ± 0.0	7.9 ± 6.2
DANN [7]	45.0 ± 0.2	12.8 ± 0.2	36.0 ± 0.2	10.4 ± 0.3	46.7 ± 0.3	38.0 ± 0.3	31.5 ± 0.1
CDANN [16]	45.3 ± 0.2	12.6 ± 0.2	36.6 ± 0.2	10.3 ± 0.4	47.5 ± 0.1	38.9 ± 0.4	31.8 ± 0.2
MTL [3]	50.6 ± 0.2	14.0 ± 0.4	39.6 ± 0.3	12.0 ± 0.3	52.1 ± 0.1	41.5 ± 0.0	35.0 ± 0.0
SagNet [17]	51.0 ± 0.1	14.6 ± 0.1	40.2 ± 0.2	12.1 ± 0.2	51.5 ± 0.3	42.4 ± 0.1	35.3 ± 0.1
ARM [28]	43.0 ± 0.2	11.7 ± 0.2	34.6 ± 0.1	9.8 ± 0.4	43.2 ± 0.3	37.0 ± 0.3	29.9 ± 0.1
VREx [12]	39.2 ± 1.6	11.9 ± 0.4	31.2 ± 1.3	10.2 ± 0.4	41.5 ± 1.8	34.8 ± 0.8	28.1 ± 1.0
RSC [10]	39.5 ± 3.7	11.4 ± 0.8	30.5 ± 3.1	10.2 ± 0.8	41.0 ± 1.4	34.7 ± 2.6	27.9 ± 2.0
SelfReg [11]	47.9 ± 0.3	15.1 ± 0.3	41.2 ± 0.2	11.7 ± 0.3	48.8 ± 0.0	43.8 ± 0.3	34.7 ± 0.2
MixStyle [29]	49.1 ± 0.4	13.4 ± 0.0	39.3 ± 0.0	11.4 ± 0.4	47.7 ± 0.3	42.7 ± 0.1	33.9 ± 0.1
Fish [23]	51.5 ± 0.3	14.5 ± 0.2	40.4 ± 0.3	11.7 ± 0.5	52.6 ± 0.2	42.1 ± 0.1	35.5 ± 0.0
SD [19]	51.3 ± 0.3	15.5 ± 0.1	41.5 ± 0.3	12.6 ± 0.2	52.9 ± 0.2	44.0 ± 0.4	36.3 ± 0.2
CAD [21]	45.4 ± 1.0	12.1 ± 0.5	34.9 ± 1.1	10.2 ± 0.6	45.1 ± 1.6	38.5 ± 0.6	31.0 ± 0.8
CondCAD [21]	46.1 ± 1.0	13.3 ± 0.4	36.1 ± 1.4	10.7 ± 0.2	46.8 ± 1.3	38.7 ± 0.7	31.9 ± 0.7
Fishr [20]	47.8 ± 0.7	14.6 ± 0.2	40.0 ± 0.3	11.9 ± 0.2	49.2 ± 0.7	41.7 ± 0.1	34.2 ± 0.3
MIRO [4]	50.9 ± 0.2	15.6 ± 0.1	41.9 ± 0.4	10.4 ± 0.1	55.1 ± 0.1	42.5 ± 0.3	36.1 ± 0.1
Ours	50.2 ± 0.3	15.9 ± 0.1	42.0 ± 0.5	12.6 ± 0.2	51.3 ± 0.1	43.8 ± 0.3	36.0 ± 0.2

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 3, 4, 5, 6

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, 2018. 5

[3] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *arXiv preprint arXiv:1711.07910*, 2017. 3, 4, 5, 6

[4] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *ECCV*, 2022. 3, 4, 5, 6

[5] Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *CVPR*, 2023. 1

[6] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, 2013. 4

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016. 3, 4, 5, 6

[8] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021. 3, 4, 5, 6

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2, 3

[10] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, 2020. 3, 4, 5, 6

[11] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *ICCV*, 2021. 2, 3, 4, 5, 6

[12] 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*, 2021. 3, 4, 5, 6

[13] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 1, 2, 3, 4

[14] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 3, 4, 5, 6

[15] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 3, 4, 5, 6

[16] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018. 3, 4, 5, 6

[17] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *CVPR*, 2021. 2, 3, 4, 5, 6

[18] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 6

[19] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *NeurIPS*, 2021. 3, 4, 5, 6

[20] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *ICML*, 2022. 4, 5, 6

[21] Yangjun Ruan, Yann Dubois, and Chris J Maddison. Optimal representations for covariate shift. In *ICLR*, 2022. 4, 5, 6

[22] 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. In *ICLR*, 2020. 3, 4, 5, 6

[23] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *ICLR*, 2021. 3, 4, 5, 6

[24] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016. 2, 3, 4, 5, 6

[25] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 3, 4, 5, 6

[26] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017. 5

[27] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 3, 4, 5, 6

[28] 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. 3, 4, 5, 6

[29] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021. 4, 5, 6