Supplemental Material

Construction Details of Synthetic Dataset

In this section, we will give more details on constructing the **Cross Line** and **Vertical Line** experiments.

Cross Lines Experiment

Based on CIFAR10 dataset, we create ten-valued spurious feature and add a vertical line passing through the middle of each channel, and a horizontal line passing through the first channel. For each line added to the channel, we implement by adding the value taken of $0.5 \pm 0.5\mathcal{B}$ where $\mathcal{B} \in [-1, 1]$. Four lines, each with two choices of $+\mathcal{B}$ or $-\mathcal{B}$. Then we have in total $2^4 = 16$ choices. We select 10 of the 16 configurations to map each configuration into one specific class. In detail, we select the images with specific class (e.g. bird), and add the the line with specific configuration (e.g. 0.5 + 0.5 $\mathcal{B}$, $\mathcal{B} = 0.8$). Similar with (Nagarajan, Andreassen, and Neyshabur 2021), we add the line of i-th configuration to corresponding class images with a probability of $p_{ii} = 0.5$; for other j-th class, we set $p_{ij} = (1 - p_{ii})/10 = 0.05$. We call the i-th class images added with i-th configuration line with p_{ii} the majority group. We call the i-th class images added with other j-th configuration line with p_{ij} the minority group. The specific configurations are in the Table 1.

Vertical Line Experiment

Based on CIFAR10 dataset, we add a vertical line to the last channel of all images. In detail, we add the line with value from $\mathcal{B} \in [-4, 4]$. To avoid negative values, all pixels in last channel are added by 4, and then added by $\mathcal{B}$, and then divided by 9 to ensure the pixels lie in the range of [0, 1]. Such operations will add non-orthogonal componenets to images, where each data-point is represented on the plane of $(x_{\text{inv}}, x_{\text{env}} + x_{\text{inv}})$ because of the added line with constant value in last channel. In (Nagarajan, Andreassen, and Neyshabur 2021), they show the non-orthogonal versus orthogonal experiment, the results suggest that the non-orthogonal images $(x_{\text{inv}}, x_{\text{env}} + x_{\text{inv}})$ are more hard-to-disentangle than orthogonal ones $(x_{\text{inv}}, x_{\text{env}})$, and would cause geometric skews of a max-margin classifier. In our experiment, during training, we set the $\mathcal{B}$ = 4 and 0, and test on domains with different $\mathcal{B} \in \{-4, -2, 0, 2, 4\}$.

More Results on DomainBed

We report the rest experiment results in this section. In training-domain validation set, the validation set is subset of training set, we choose the model that performs best on the overall validation set for each domain. This strategy characterizes the in-distribution generalization capability of the model.

The results are recorded in Table 2. From the Table, we can see that IIB achieves 67.5% accuracy across 7 datasets on average, which is comparable to the best algorithm CORAL on DomainBed. Also IIB shows better performance on larger datasets (e.g. OfficeHome, DomainNet). The results demonstrate IIB's in-domains generalization ability.

Table 1: 10 configurations in **Cross Lines** experiments.

Configuration #	**Channel**	**Line's Position**	**Sign**
0	0	Vertical	+
	1	Vertical	+
	2	Vertical	+
	0	Horizontal	+
1	0	Vertical	-
	1	Vertical	+
	2	Vertical	+
	0	Horizontal	+
2	0	Vertical	+
	1	Vertical	-
	2	Vertical	+
	0	Horizontal	+
3	0	Vertical	+
	1	Vertical	+
	2	Vertical	+
	0	Horizontal	+
4	0	Vertical	+
	1	Vertical	+
	2	Vertical	-
	0	Horizontal	+
5	0	Vertical	+
	1	Vertical	+
	2	Vertical	+
	0	Horizontal	-
6	0	Vertical	-
	1	Vertical	-
	2	Vertical	+
	0	Horizontal	+
7	0	Vertical	+
	1	Vertical	-
	2	Vertical	-
	0	Horizontal	+
8	0	Vertical	+
	1	Vertical	+
	2	Vertical	-
	0	Horizontal	-
9	0	Vertical	-
	1	Vertical	+
	2	Vertical	+
	0	Horizontal	-
10	0	Vertical	-
	1	Vertical	+
	2	Vertical	-
	0	Horizontal	+

Table 2: Performance comparison (Acc. %) between the proposed IIB method and the state-of-the-art domain generalization methods with *training-domain validation set* model selection strategy. The best accuracy in each dataset is presented in boldface. The average accuracy over all the datasets is also reported.

Algorithm	**ColoredMNIST**	**RotatedMNIST**	**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
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
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
IIB(Ours)	52.0 ± 0.3	98.1 ± 0.2	77.6 ± 0.2	85.7 ± 0.6	**69.0** ± 0.1	48.5 ± 0.4	**41.6** ± 0.8	**67.5**

References

Nagarajan, V.; Andreassen, A.; and Neyshabur, B. 2021. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*.