Supplemental Material to
Cross Contrasting Feature Perturbation for Domain Generalization

1. The Number of LDP Modules

We have studied the influence of the number of LDP modules in Table 1. It is observed that inserting the LDP modules in positions 1-5 results in the highest performance.

Numbers	1	2	3	4	5
PACS	85.1	84.7	85.4	85.6	**86.6**

Table 1. Effects of different numbers of LDP modules based on Training-domain model selection.

2. Evaluation Protocol Details

In this study, we follow the evaluation protocol given by DomainBed[8]. For the Colored MNIST and Rotated MNIST datasets, we train the model for 5000 steps and evaluate the model every 100 steps. For the PACS, VLCS, OfficeHome, and TerraIncognita datasets, we train the model for 5000 steps and evaluate the model every 300 steps. For the DomainNet dataset, we train the model for 5000 steps and evaluate the model every 1000 steps. The batch size and the learning rate are set to 32 for each domain and 5e-5 respectively. The dropout probability and the weight decay are set to zero.

The same procedure was applied for all methods: a random hyperparameter search of 20 trials over a joint distribution, described in Table 2. DomainBed conduct a hyperparameter search for each domain, which means we need to train 20 * n (number of domains) models to select the hyperparameters. However, in practice, it is non-trivial to apply a different model to a different domain. Hence, to reduce the search space for computational efficiency, we use the same hyperparameter for each domain in every dataset. Although the total runs are reduced, our method also outperforms the previous state-of-the-art.

For the comparison with SWAD-based methods, we build our code upon the open-source code of SWAD[4] and follow its training strategy.

Dataset	Parameter	Default value	distribution
PACS/VLCS/	learning rate	5e-5	$10^{Uniform(-5,-3.5)}$
OfficeHome/	batch size	32	$2^{Uniform(3,5.5)}$
TerraIncognita/	weight decay	0	$10^{Uniform(-6,-2)}$
DomainNet	dropout probability	0	RandomChoice$([0, 0.1, 0.5])$
CMNIST/	learning rate	1e-3	$10^{Uniform(-4.5,-3.5)}$
RMNIST	batch size	64	$2^{Uniform(3,9)}$
	weight decay	0	0
CCFP	domain specific strength λ_{dis}	1	Uniform(0.1,10)
	semantic regularization strength λ_{sem}	1	Uniform(0.1,10)

Table 2. Hyperparameter search space

3. Architecture Details

We specify the inserted position of the LDP modules in Table 3 for ResNet-50 and in Table 4 for MNIST ConvNet. The input size of LDP is the dimension of parameters $\gamma \in \mathbb{R}^{size}$ and $\beta \in \mathbb{R}^{size}$ in LDP modules.

4. Test-domain (Oracle) model selection

In the main paper, we focus on the 'Training-domain' model selection. For the oracle model selection, DomainBed claims that Oracle selection results can be either optimistic, because we access the test distribution, or pessimistic, because the query limit reduces the number of considered hyperparameter combinations. However, previous work[22, 17] argues that OOD performance cannot, by definition, be performed with a validation set from the same distribution as the training data. To this end, we still report the results based on the Test-domain model selection to prove the usefulness of our approach.

#	Layers
1	Conv2D (in=d, out=64)
2	**LDP** (size=64)
3	BatchNorm
4	Relu
5	Maxpooling
6	**LDP** (size=64)
7	Layer1(in=d, out=256)
8	**LDP** (size=256)
9	Layer2(in=d, out=512)
10	**LDP** (size=512)
11	Layer3(in=d, out=1024)
12	**LDP** (size=1024)
13	Layer4(in=d, out=2048)
14	Average pooling

Table 3. Details of ResNet-50 and the LDP modules.

#	Layers
1	Conv2D (in=d, out=64)
2	Relu
3	GroupNorm (groups=8)
4	**LDP** (size=64)
5	Conv2D (in=64, out=128, stride=2)
6	Relu
7	GroupNorm (groups=8)
8	**LDP** (size=64)
9	Conv2D (in=128, out=128)
10	Relu
11	GroupNorm (groups=8)
13	Conv2D (in=128, out=128)
14	Relu
15	GroupNorm (groups=8)
16	Average pooling

Table 4. Details of ConvNet and the LDP modules.

Algorithm	CMNIST	RMNIST	VLCS	PACS	OfficeHome	TerraInc	DomainNet	Avg
ERM[23]	57.8 ± 0.2	97.8 ± 0.1	77.6 ± 0.3	86.7 ± 0.3	66.4 ± 0.5	53.0 ± 0.3	41.3 ± 0.1	68.7
IRM[1]	67.7 ± 1.2	97.5 ± 0.2	76.9 ± 0.6	84.5 ± 1.1	63.0 ± 2.7	50.5 ± 0.7	28.0 ± 5.1	66.9
GroupDRO[18]	61.1 ± 0.9	97.9 ± 0.1	77.4 ± 0.5	87.1 ± 0.1	66.2 ± 0.6	52.4 ± 0.1	33.4 ± 0.3	67.9
Mixup[25]	58.4 ± 0.2	98.0 ± 0.1	78.1 ± 0.3	86.8 ± 0.3	68.0 ± 0.2	**54.4** ± 0.3	39.6 ± 0.1	69.0
MLDG[11]	58.2 ± 0.4	97.8 ± 0.1	77.5 ± 0.1	86.8 ± 0.4	66.6 ± 0.3	52.0 ± 0.1	41.6 ± 0.1	68.7
CORAL[21]	58.6 ± 0.5	98.0 ± 0.0	77.7 ± 0.2	87.1 ± 0.5	68.4 ± 0.2	52.8 ± 0.2	**41.8** ± 0.1	69.2
MMD[12]	63.3 ± 1.3	98.0 ± 0.1	77.9 ± 0.1	87.2 ± 0.1	66.2 ± 0.3	52.0 ± 0.4	23.5 ± 9.4	66.9
DANN[7]	57.0 ± 1.0	97.9 ± 0.1	79.7 ± 0.5	85.2 ± 0.2	65.3 ± 0.8	50.6 ± 0.4	38.3 ± 0.1	67.7
CDANN[12]	59.5 ± 2.0	97.9 ± 0.0	79.9 ± 0.2	85.8 ± 0.8	65.3 ± 0.5	50.8 ± 0.6	38.5 ± 0.2	68.2
MTL[3]	57.6 ± 0.3	97.9 ± 0.1	77.7 ± 0.5	86.7 ± 0.2	66.5 ± 0.4	52.2 ± 0.4	40.8 ± 0.1	68.5
SagNet[13]	58.2 ± 0.3	97.9 ± 0.0	77.6 ± 0.1	86.4 ± 0.4	67.5 ± 0.2	52.5 ± 0.4	40.8 ± 0.2	68.7
ARM[27]	63.2 ± 0.7	**98.1** ± 0.1	77.8 ± 0.3	85.8 ± 0.2	64.8 ± 0.4	51.2 ± 0.5	36.0 ± 0.2	68.1
V-REx[10]	67.0 ± 1.3	97.9 ± 0.1	78.1 ± 0.2	87.2 ± 0.6	65.7 ± 0.3	51.4 ± 0.5	30.1 ± 3.7	68.2
RSC[9]	58.5 ± 0.5	97.6 ± 0.1	77.8 ± 0.6	86.2 ± 0.5	66.5 ± 0.6	52.1 ± 0.2	38.9 ± 0.6	68.2
AND-mask[15]	58.6 ± 0.4	97.5 ± 0.0	76.4 ± 0.4	86.4 ± 0.4	66.1 ± 0.2	49.8 ± 0.4	37.6 ± 0.6	67.5
SAND-mask[20]	62.3 ± 1.0	97.4 ± 0.1	76.2 ± 0.5	85.9 ± 0.4	65.9 ± 0.5	50.2 ± 0.1	32.2 ± 0.6	67.2
Fishr[17]	**68.8** ± 1.4	97.8 ± 0.1	78.2 ± 0.2	86.9 ± 0.2	68.2 ± 0.2	53.6 ± 0.4	**41.8** ± 0.2	70.8
CCFP (ours)	65.0 ± 0.1	97.9 ± 0.1	**80.0** ± 0.6	**88.4** ± 0.3	**69.7** ± 0.3	53.1 ± 0.4	**41.8** ± 0.1	**70.9**

Table 5. DomainBed with oracle model selection. We highlight the **best result**.

5. More Comparison with previous feature perturbation based methods

We conduct new extensive experiments on three additional datasets as shown in Table 6. It can be seen that our CCFP outperforms the previous feature perturbation based methods significantly.

Algorithm	VLCS	OfficeHome	TerraInc
Mixstyle	76.2 ± 0.4	64.2 ± 0.1	46.0 ± 1.0
Mixstyle (dual)	78.4 ± 0.2	66.7 ± 0.2	46.8 ± 0.8
DSU	76.8 ± 0.3	63.3 ± 0.2	42.9 ± 0.6
DSU (dual)	77.7 ± 0.3	66.3 ± 0.3	47.5 ± 0.3
pAdaIN	75.2 ± 0.6	64.6 ± 0.1	45.5 ± 0.8
pAdaIN (dual)	77.9 ± 0.2	68.1 ± 0.2	47.5 ± 0.3
CCFP (ours)	**78.9** ± 0.3	**68.9** ± 0.1	**48.6** ± 0.4

Table 6. The experiment setting is as same as Table 8 (page 7).

6. More Experimental results

We also validate our algorithm on DomainNet[16]. Our approach achieves a comparable performance against the previous state-of-the-art. As shown in Table 9, our approach shows state-of-the-art performances with significant margin of $+0.4\%$ averagely on accuracy.

Algorithm	A	C	P	S	Avg.
MASF[6]	82.9	80.5	95.0	72.3	82.7
DMG[5]	82.6	78.1	94.5	78.3	83.4
MetaReg[2]	87.2	79.2	97.6	70.3	83.6
ER[28]	87.5	79.3	**98.3**	76.3	85.3
pAdaIN[14]	85.8	81.1	97.2	77.4	85.4
EISNet[24]	86.6	81.5	97.1	78.1	85.8
DSON[19]	87.0	80.6	96.0	82.9	86.6
SWAD[4]	89.3	83.4	97.3	82.5	88.1
PCL[26]	90.2	83.9	98.1	82.6	88.7
CCFP (ours)	**90.3**	**84.0**	97.2	**83.7**	**88.8**

Table 7. Comparison with SWAD-based state-of-the-art methods on PACS benchmark. A: art, C: cartoon, P: photo, S: sketch, Avg.: average.

Algorithm	clip	info	paint	quick	real	sketch	Avg.
SWAD[4]	66.0	22.4	53.5	16.1	65.8	55.5	46.5
PCL[26]	67.9	24.3	55.3	15.7	66.6	56.4	**47.7**
CCFP (ours)	66.4	22.9	54.0	16.2	64.5	56.7	46.8

Table 8. Comparison with SWAD-based state-of-the-art methods on DomainBed benchmark.

Algorithm	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
SWAD[4]	88.1 ± 0.1	79.1 ± 0.1	70.6 ± 0.2	50.0 ± 0.3	46.5 ± 0.1	66.9
PCL[26]	88.7	78.0	71.6	52.1	**47.7**	67.6
CCFP (ours)	**88.8** ± 0.3	**79.4** ± 0.1	**72.1** ± 0.3	**53.0** ± 0.1	46.8 ± 0.2	**68.0**

Table 9. Comparison with SWAD-based state-of-the-art methods on DomainBed benchmark.

7. Full DomainBed Results

Tables below detail results for each dataset with 'Training-domain' model selection methods.

Algorithm	+90%	+80%	+10%	Avg.
ERM[23]	71.7 ± 0.4	72.9 ± 0.2	10.0 ± 0.1	51.5
IRM[1]	72.5 ± 0.1	73.3 ± 0.5	10.2 ± 0.3	52.0
GroupDRO[18]	73.1 ± 0.3	73.2 ± 0.2	10.0 ± 0.2	52.1
Mixup[25]	72.7 ± 0.4	73.4 ± 0.1	10.1 ± 0.1	52.1
MLDG[11]	71.5 ± 0.2	73.1 ± 0.2	9.8 ± 0.1	51.5
CORAL[21]	71.6 ± 0.3	73.1 ± 0.1	9.9 ± 0.1	51.5
MMD[12]	71.4 ± 0.3	73.1 ± 0.2	9.9 ± 0.3	51.5
DANN[7]	71.4 ± 0.9	73.1 ± 0.1	10.0 ± 0.0	51.5
CDANN[12]	72.0 ± 0.2	73.0 ± 0.2	10.2 ± 0.1	51.7
MTL[3]	70.9 ± 0.2	72.8 ± 0.3	10.5 ± 0.1	51.4
SagNet[13]	71.8 ± 0.2	73.0 ± 0.2	10.3 ± 0.0	51.7
ARM[27]	**82.0** ± 0.5	**76.5** ± 0.3	10.2 ± 0.0	**56.2**
V-REx[10]	72.4 ± 0.3	72.9 ± 0.4	10.2 ± 0.0	51.8
RSC[9]	71.9 ± 0.3	73.1 ± 0.2	10.0 ± 0.2	51.7
AND-mask[15]	70.7 ± 0.5	73.3 ± 0.2	10.0 ± 0.1	51.3
SAND-mask[20]	72.0 ± 0.5	73.2 ± 0.4	**10.3** ± 0.2	51.8
Fishr[17]	72.3 ± 0.9	73.5 ± 0.2	10.1 ± 0.2	52.0
CCFP (ours)	72.4 ± 0.0	73.1 ± 0.2	10.2 ± 0.2	51.9

Table 10. Comparison with state-of-the-art methods on Colored MNIST benchmark.

Algorithm	0	15	30	45	60	75	Avg.
ERM[23]	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[1]	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[18]	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[25]	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[11]	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[21]	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[12]	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
DANN[7]	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[12]	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
MTL[3]	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[13]	95.9 ± 0.3	98.9 ± 0.1	99.0 ± 0.1	**99.1** ± 0.1	99.0 ± 0.1	96.3 ± 0.1	98.0
ARM[27]	**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**
V-REx[10]	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[9]	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
AND-mask[15]	94.8 ± 0.2	98.8 ± 0.1	98.9 ± 0.0	98.7 ± 0.0	98.7 ± 0.1	95.5 ± 0.4	97.6
SAND-mask[20]	94.5 ± 0.4	98.6 ± 0.1	98.8 ± 0.1	98.7 ± 0.1	98.6 ± 0.0	95.5 ± 0.2	97.4
Fishr[17]	95.0 ± 0.3	98.5 ± 0.0	**99.2** ± 0.1	98.9 ± 0.0	98.9 ± 0.1	**96.5** ± 0.0	97.8
CCFP (ours)	95.6 ± 0.1	98.7 ± 0.0	98.8 ± 0.0	98.8 ±0.0	98.8 ± 0.1	95.7 ± 0.1	97.8

Table 11. Comparison with state-of-the-art methods on Rotated MNIST benchmark.

Algorithm	C	L	V	S	Avg.
ERM[23]	97.7 ± 0.4	64.3 ± 0.9	73.4 ± 0.5	74.6 ± 1.3	77.5
IRM[1]	98.6 ± 0.1	64.9 ± 0.9	73.4 ± 0.6	77.3 ± 0.9	78.5
GroupDRO[18]	97.3 ± 0.3	63.4 ± 0.9	69.5 ± 0.8	76.7 ± 0.7	76.7
Mixup[25]	98.3 ± 0.6	64.8 ± 1.0	72.1 ± 0.5	74.3 ± 0.8	77.4
MLDG[11]	97.4 ± 0.2	65.2 ± 0.7	71.0 ± 1.4	75.3 ± 1.0	77.2
CORAL[21]	98.3 ± 0.1	**66.1** ± 1.2	73.4 ± 0.3	77.5 ± 1.2	78.8
MMD[12]	97.7 ± 0.1	64.0 ± 1.1	72.8 ± 0.2	75.3 ± 3.3	77.5
DANN[7]	**99.0** ± 0.3	65.1 ± 1.4	73.1 ± 0.3	77.2 ± 0.6	78.6
CDANN[12]	97.1 ± 0.3	65.1 ± 1.2	70.7 ± 0.8	77.1 ± 1.5	77.5
MTL[3]	97.8 ± 0.4	64.3 ± 0.3	71.5 ± 0.7	75.3 ± 1.7	77.2
SagNet[13]	97.9 ± 0.4	64.5 ± 0.5	71.4 ± 1.3	77.5 ± 0.5	77.8
ARM[27]	98.7 ± 0.2	63.6 ± 0.7	71.3 ± 1.2	76.7 ± 0.6	77.6
V-REx[10]	98.4 ± 0.3	64.4 ± 1.4	74.1 ± 0.4	76.2 ± 1.3	78.3
RSC[9]	97.9 ± 0.1	62.5 ± 0.7	72.3 ± 1.2	75.6 ± 0.8	77.1
AND-mask[15]	97.8 ± 0.4	64.3 ± 1.2	73.5 ± 0.7	76.8 ± 2.6	78.1
SAND-mask[20]	98.5 ± 0.3	63.6 ± 0.9	70.4 ± 0.8	77.1 ± 0.8	77.4
Fishr[17]	98.9 ± 0.3	64.0 ± 0.5	71.5 ± 0.2	76.8 ± 0.7	77.8
CCFP (ours)	98.1 ± 0.2	64.9 ± 0.1	**74.5** ± 1.5	**78.3** ± 0.2	**78.9**

Table 12. Comparison with state-of-the-art methods on VLCS benchmark.

Algorithm	A	C	P	S	Avg.
ERM[23]	84.7 ± 0.4	80.8 ± 0.6	97.2 ± 0.3	79.3 ± 1.0	85.5
IRM[1]	84.8 ± 1.3	76.4 ± 1.1	96.7 ± 0.6	76.1 ± 1.0	83.5
GroupDRO[18]	83.5 ± 0.9	79.1 ± 0.6	96.7 ± 0.3	78.3 ± 2.0	84.4
Mixup[25]	86.1 ± 0.5	78.9 ± 0.8	**97.6** ± 0.1	75.8 ± 1.8	84.6
MLDG[11]	85.5 ± 1.4	80.1 ± 1.7	97.4 ± 0.3	76.6 ± 1.1	84.9
CORAL[21]	88.3 ± 0.2	80.0 ± 0.5	97.5 ± 0.3	78.8 ± 1.3	86.2
MMD[12]	86.1 ± 1.4	79.4 ± 0.9	96.6 ± 0.2	76.5 ± 0.5	84.6
DANN[7]	86.4 ± 0.8	77.4 ± 0.8	97.3 ± 0.4	73.5 ± 2.3	83.6
CDANN[12]	84.6 ± 1.8	75.5 ± 0.9	96.8 ± 0.3	73.5 ± 0.6	82.6
MTL[3]	87.5 ± 0.8	77.1 ± 0.5	96.4 ± 0.8	77.3 ± 1.8	84.6
SagNet[13]	87.4 ± 1.0	80.7 ± 0.6	97.1 ± 0.1	80.0 ± 0.4	86.3
ARM[27]	86.8 ± 0.6	76.8 ± 0.5	97.4 ± 0.3	79.3 ± 1.2	85.1
V-REx[10]	86.0 ± 1.6	79.1 ± 0.6	96.9 ± 0.5	77.7 ± 1.7	84.9
RSC[9]	85.4 ± 0.8	79.7 ± 1.8	**97.6** ± 0.3	78.2 ± 1.2	85.2
AND-mask[15]	85.3 ± 1.4	79.2 ± 2.0	96.9 ± 0.4	76.2 ± 1.4	84.4
SAND-mask[20]	85.8 ± 1.7	79.2 ± 0.8	96.3 ± 0.2	76.9 ± 2.0	84.6
Fishr[17]	**88.4** ± 0.2	78.7 ± 0.7	97.0 ± 0.1	77.8 ± 2.0	85.5
CCFP (ours)	87.5 ± 0.1	**81.3** ± 0.3	96.4 ± 0.3	**81.4** ± 0.8	**86.6**

Table 13. Comparison with state-of-the-art methods on PACS benchmark.

Algorithm	**A**	**C**	**P**	**R**	**Avg.**
ERM[23]	61.3 ± 0.7	52.4 ± 0.3	75.8 ± 0.1	76.6 ± 0.3	66.5
IRM[1]	58.9 ± 2.3	52.2 ± 1.6	72.1 ± 2.9	74.0 ± 2.5	64.3
GroupDRO[18]	60.4 ± 0.7	52.7 ± 1.0	75.0 ± 0.7	76.0 ± 0.7	66.0
Mixup[25]	62.4 ± 0.8	54.8 ± 0.6	76.9 ± 0.3	78.3 ± 0.2	68.1
MLDG[11]	61.5 ± 0.9	53.2 ± 0.6	75.0 ± 1.2	77.5 ± 0.4	66.8
CORAL[21]	**65.3** ± 0.4	54.4 ± 0.5	76.5 ± 0.1	78.4 ± 0.5	68.7
MMD[12]	60.4 ± 0.2	53.3 ± 0.3	74.3 ± 0.1	77.4 ± 0.6	66.3
DANN[7]	59.9 ± 1.3	53.0 ± 0.3	73.6 ± 0.7	76.9 ± 0.5	65.9
CDANN[12]	61.5 ± 1.4	50.4 ± 2.4	74.4 ± 0.9	76.6 ± 0.8	65.8
MTL[3]	61.5 ± 0.7	52.4 ± 0.6	74.9 ± 0.4	76.8 ± 0.4	66.4
SagNet[13]	63.4 ± 0.2	54.8 ± 0.4	75.8 ± 0.4	78.3 ± 0.3	68.1
ARM[27]	58.9 ± 0.8	51.0 ± 0.5	74.1 ± 0.1	75.2 ± 0.3	64.8
V-REx[10]	60.7 ± 0.9	53.0 ± 0.9	75.3 ± 0.1	76.6 ± 0.5	66.4
RSC[9]	60.7 ± 1.4	51.4 ± 0.3	74.8 ± 1.1	75.1 ± 1.3	65.5
AND-mask[15]	59.5 ± 1.2	51.7 ± 0.2	73.9 ± 0.4	77.1 ± 0.2	65.6
SAND-mask[20]	60.3 ± 0.5	53.3 ± 0.7	73.5 ± 0.7	76.2 ± 0.3	65.8
Fishr[17]	62.4 ± 0.5	54.4 ± 0.4	76.2 ± 0.5	78.3 ± 0.1	67.8
CCFP (ours)	63.7 ± 0.3	**55.5** ± 0.3	**77.2** ± 0.4	**79.2** ± 0.3	**68.9**

Table 14. Comparison with state-of-the-art methods on OfficeHome benchmark.

Algorithm	**L100**	**L38**	**L43**	**L46**	**Avg.**
ERM[23]	49.8 ± 4.4	42.1 ± 1.4	56.9 ± 1.8	35.7 ± 3.9	46.1
IRM[1]	54.6 ± 1.3	39.8 ± 1.9	56.2 ± 1.8	39.6 ± 0.8	47.6
GroupDRO[18]	41.2 ± 0.7	38.6 ± 2.1	56.7 ± 0.9	36.4 ± 2.1	43.2
Mixup[25]	**59.6** ± 2.0	42.2 ± 1.4	55.9 ± 0.8	33.9 ± 1.4	47.9
MLDG[11]	54.2 ± 3.0	**44.3** ± 1.1	55.6 ± 0.3	36.9 ± 2.2	47.7
CORAL[21]	51.6 ± 2.4	42.2 ± 1.0	57.0 ± 1.0	39.8 ± 2.9	47.6
MMD[12]	41.9 ± 3.0	34.8 ± 1.0	57.0 ± 1.9	35.2 ± 1.8	42.2
DANN[7]	51.5 ± 3.5	40.6 ± 0.6	57.4 ± 0.5	37.7 ± 1.8	46.7
CDANN[12]	47.0 ± 1.9	41.3 ± 4.8	54.9 ± 1.7	39.8 ± 2.3	45.8
MTL[3]	49.3 ± 1.2	39.6 ± 6.3	55.6 ± 1.1	37.8 ± 0.8	45.6
SagNet[13]	53.0 ± 2.9	43.0 ± 2.5	57.9 ± 0.6	40.4 ± 1.3	**48.6**
ARM[27]	49.3 ± 0.7	38.3 ± 2.4	55.8 ± 0.8	38.7 ± 1.3	45.5
V-REx[10]	48.2 ± 4.3	41.7 ± 1.3	56.8 ± 0.8	38.7 ± 3.1	46.4
RSC[9]	50.2 ± 2.2	39.2 ± 1.4	56.3 ± 1.4	**40.8** ± 0.6	46.6
AND-mask[15]	50.0 ± 2.9	40.2 ± 0.8	53.3 ± 0.7	34.8 ± 1.9	44.6
SAND-mask[20]	45.7 ± 2.9	31.6 ± 4.7	55.1 ± 1.0	39.0 ± 1.8	42.9
Fishr[17]	50.2 ± 3.9	43.9 ± 0.8	55.7 ± 2.2	39.8 ± 1.0	47.4
CCFP (ours)	56.4 ± 1.8	42.3 ± 0.1	**58.0** ± 0.7	37.5 ± 0.4	**48.6**

Table 15. Comparison with state-of-the-art methods on TerraIncognita benchmark.

Algorithm	clip	info	paint	quick	real	sketch	Avg.
ERM[23]	58.1 $\pm$ 0.3	18.8 $\pm$ 0.3	46.7 $\pm$ 0.3	12.2 $\pm$ 0.4	59.6 $\pm$ 0.1	49.8 $\pm$ 0.4	40.9
IRM[1]	48.5 $\pm$ 2.8	15.0 $\pm$ 1.5	38.3 $\pm$ 4.3	10.9 $\pm$ 0.5	48.2 $\pm$ 5.2	42.3 $\pm$ 3.1	33.9
GroupDRO[18]	47.2 $\pm$ 0.5	17.5 $\pm$ 0.4	33.8 $\pm$ 0.5	9.3 $\pm$ 0.3	51.6 $\pm$ 0.4	40.1 $\pm$ 0.6	33.3
Mixup[25]	55.7 $\pm$ 0.3	18.5 $\pm$ 0.5	44.3 $\pm$ 0.5	12.5 $\pm$ 0.4	55.8 $\pm$ 0.3	48.2 $\pm$ 0.5	39.2
MLDG[11]	59.1 $\pm$ 0.2	19.1 $\pm$ 0.3	45.8 $\pm$ 0.7	**13.4** $\pm$ 0.3	59.6 $\pm$ 0.2	50.2 $\pm$ 0.4	41.2
CORAL[21]	**59.2** $\pm$ 0.1	19.7 $\pm$ 0.2	46.6 $\pm$ 0.3	**13.4** $\pm$ 0.3	59.8 $\pm$ 0.2	50.1 $\pm$ 0.6	41.5
MMD[12]	32.1 $\pm$ 13.3	11.0 $\pm$ 4.6	26.8 $\pm$ 11.3	8.7 $\pm$ 2.1	32.7 $\pm$ 13.8	28.9 $\pm$ 11.9	23.4
DANN[7]	53.1 $\pm$ 0.2	18.3 $\pm$ 0.1	44.2 $\pm$ 0.7	12.1 $\pm$ 0.7	55.5 $\pm$ 0.4	46.8 $\pm$ 0.5	38.3
CDANN[12]	54.6 $\pm$ 0.4	17.3 $\pm$ 0.1	43.7 $\pm$ 0.9	12.1 $\pm$ 0.7	56.2 $\pm$ 0.4	45.9 $\pm$ 0.5	38.3
MTL[3]	57.9 $\pm$ 0.5	18.5 $\pm$ 0.4	46.0 $\pm$ 0.1	12.5 $\pm$ 0.1	59.5 $\pm$ 0.3	49.2 $\pm$ 0.1	40.6
SagNet[13]	57.7 $\pm$ 0.3	19.0 $\pm$ 0.2	45.3 $\pm$ 0.3	12.7 $\pm$ 0.5	58.1 $\pm$ 0.5	48.8 $\pm$ 0.2	40.3
ARM[27]	49.7 $\pm$ 0.3	16.3 $\pm$ 0.5	40.9 $\pm$ 1.1	9.4 $\pm$ 0.1	53.4 $\pm$ 0.4	43.5 $\pm$ 0.4	35.5
V-REx[10]	47.3 $\pm$ 3.5	16.0 $\pm$ 1.5	35.8 $\pm$ 4.6	10.9 $\pm$ 0.3	49.6 $\pm$ 4.9	42.0 $\pm$ 3.0	33.6
RSC[9]	55.0 $\pm$ 1.2	18.3 $\pm$ 0.5	44.4 $\pm$ 0.6	12.2 $\pm$ 0.2	55.7 $\pm$ 0.7	47.8 $\pm$ 0.9	38.9
AND-mask[15]	52.3 $\pm$ 0.8	16.6 $\pm$ 0.3	41.6 $\pm$ 1.1	11.3 $\pm$ 0.1	55.8 $\pm$ 0.4	45.4 $\pm$ 0.9	37.2
SAND-mask[20]	43.8 $\pm$ 1.3	14.8 $\pm$ 0.3	38.2 $\pm$ 0.6	9.0 $\pm$ 0.3	47.0 $\pm$ 1.1	39.9 $\pm$ 0.6	32.1
Fishr[17]	58.2 $\pm$ 0.5	**20.2** $\pm$ 0.2	**47.7** $\pm$ 0.3	12.7 $\pm$ 0.2	**60.3** $\pm$ 0.2	**50.8** $\pm$ 0.1	**41.7**
CCFP (ours)	58.7 $\pm$ 0.2	19.4 $\pm$ 0.3	47.1 $\pm$ 0.3	**13.4** $\pm$ 0.4	58.1 $\pm$ 0.4	50.5 $\pm$ 0.1	41.2

Table 16. Comparison with state-of-the-art methods on DomainNet benchmark.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *stat*, 1050:27, 2020.

[2] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018.

[3] Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1):46–100, 2021.

[4] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

[5] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, pages 301–318. Springer, 2020.

[6] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] 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.

[8] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

[9] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020.

[10] 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.

[11] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[12] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[13] 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.

[14] Oren Nuriel, Sagie Benaim, and Lior Wolf. Permuted adain: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9482–9491, 2021.

[15] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

[16] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[17] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022.

[18] 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.

[19] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *European Conference on Computer Vision*, pages 68–83. Springer, 2020.

[20] Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv preprint arXiv:2106.02266*, 2021.

[21] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[22] Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton Van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16761–16772, 2022.

[23] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

[24] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *European Conference on Computer Vision*, pages 159–176. Springer, 2020.

[25] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *stat*, 1050:3, 2020.

[26] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7097–7107, 2022.

[27] 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.

[28] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems*, 33:16096–16107, 2020.