Supplementary Material:
Domain Generalization By Rejecting Bad Augmentations

Masih Aminbeidokhti, Fidel A. Guerrero Peña, Heitor Rapela Medeiros
Thomas Dubail, Eric Granger, Marco Pedersoli
LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
{masih.aminbeidokhti.1, fidel-alejandro.guerrero-pena}@ens.etsmtl.ca
{heitor.rapela-medeiros.1, thomas.dubail.1}@ens.etsmtl.ca
{eric.granger, marco.pedersoli}@etsmtl.ca

In this supplementary material, we give additional information to reproduce our work. Here, we provide implementation details, characterization of computational cost, the definition of affinity and diversity, visual changes of the selected by our proposed data augmentation schema, the effect of ViT-backbone, the effect of hyperparameter λ, and finally, show detailed results of Table 3 in the main manuscript.

A. Implementation details

The evaluation protocol by [15] is computationally too expensive, therefore we use the reduced search space from [7] for the common parameters. Table 5 summarizes the hyperparameter search space. We use the same search space for all datasets. To further reduce the hyperparameter search, we start by finding the optimal hyperparameter for TA and then use those to find the best λ for our proposed method.

Hyperparameter	Search Space
batch size	32
learning rate	{1e-5, 3e-5, 5e-5}
ResNet dropout	{0.0, 0.1, 0.5}
weight decay	{1e-4, 1e-6}
λ	{0.2, 0.5, 0.8}

Table 5. Hyperparameters used for all methods in and their respective distributions for grid search. λ refers to the balancing coefficient of the proposed reward function (Eq 2. of the manuscript).

A.1. Datasets

PACS: [25] is a 7-way object classification task with 4 domains: art, cartoon, photo, and sketch, with 9,991 samples.

VLCS: [11] is a 5-way classification task from 4 domains: Caltech101, LabelMe, SUN09, and VOC2007. There are 10,729 samples. This dataset mostly contains real photos. The distribution shifts are subtle and simulate real-life scenarios well.

OfficeHome: [45] is a 65-way classification task depicting everyday objects from 4 domains: art, clipart, product, and real, with a total of 15,588 samples.

TerraIncognita: [4] is a 10-way classification problem of animals in wildlife cameras, where the 4 domains are different locations, L100, L38, L43, L46. There are 24,788 samples. This represents a realistic use case where generalization is indeed critical.

DomainNet: [36] is a 345-way object classification task from 6 domains: clipart, infograph, painting, quickdraw, real, and sketch. With a total of 586,575 samples, it is larger than most of the other evaluated datasets in both samples and classes.

A.2. Code

Our work is built upon DomainBed[1] [15] and SWAD[2] [7] codebase, which is released under the MIT license.

B. Effect of the balancing coefficient λ.

Figures 6, 7, 8 and 9 show the effect of the balancing coefficient between diversity and consistency rewards terms, λ, on PACS, VLCS, OfficeHome and DomainNet datasets respectively. Figure 6 shows the obtained OOD accuracy for the PACS dataset. The best value of λ inside our searching space for all domains was 0.8. Thus the consistency value was 0.8 and 0.2 for the diversity value, which implies that for an improvement on the OOD accuracy, for the PACS, we should go towards a higher value of λ for the term

[1] https://github.com/facebookresearch/DomainBed
[2] https://github.com/khanrc/swad

Transform	Search Space Ranges		
	Default [9,33]	**Wide [33]**	**Wider (Ours)**
ShearX(Y)	[-0.3, 0.3]	[-1.0, 1.0]	[-1.0, 1.0]
TranslateX(Y)	[-32, 32]	[-32, 32]	[-224.0, 224.0]
Rotate	[-30.0, 30.0]	[-135.0, 135.0]	[-135.0, 135.0]
Posterize	[4, 8]	[2, 8]	[0, 8]
Solarize	[0, 255]	[0, 255]	[0, 255]
Contrast	[-1.0, 1.0]	[-1.0, 1.0]	[-10.0, 10.0]
Color	[-1.0, 1.0]	[-1.0, 1.0]	[-10.0, 10.0]
Sharpness	[-1.0, 1.0]	[-1.0, 1.0]	[-10.0, 10.0]
Brightness	[-1.0, 1.0]	[-1.0, 1.0]	[-1.0, 10.0]
AutoContrast	N/A	N/A	N/A
Equalize	N/A	N/A	N/A
Grey	N/A	N/A	N/A

Table 6. List of image transformations and their search space ranges. The table shows the Default range from RandAugment [9], the Wide range from TA [33] as well as our proposed Wider range.

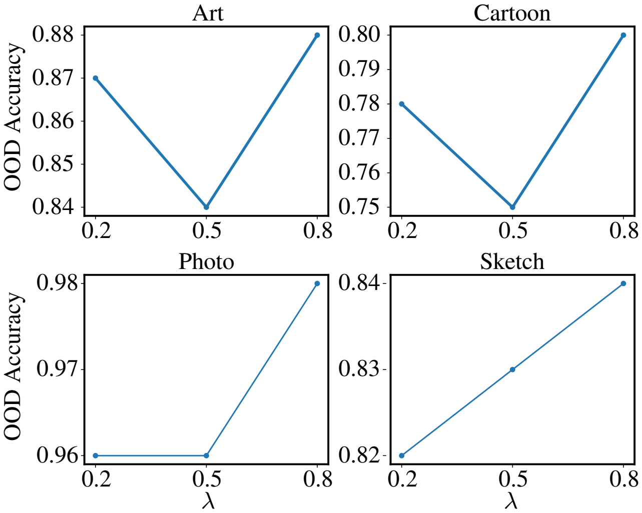

Figure 6. Effect of hyperparameter λ on PACS dataset. Each plot represents one domain: Art, Cartoon, Photo and Sketch. On x-axis is the λ and on y-axis the OOD accuracy.

of consistency than diversity. This has a positive impact on the performance with gains more than 0.01 for higher values of λ, when compared with smaller values for the Photo domain, i.e., for $\lambda = 0.2$ the OOD acc. is 0.96 and for $\lambda = 0.5$ the OOD acc. is 0.97. In which the best value of OOD acc. was 0.98 for the $\lambda = 0.8$ for this domain.

Regarding the Art and Cartoon domains, the extreme cases seem to be better than being too conservative with a λ of 0.5, so if it goes for the extremes $\lambda = 0.2$ or $\lambda = 0.8$, it is better, with better results for 0.8. In the Cartoon domain, the 0.8 λ had 0.02 gain in OOD acc. compared with performance for $\lambda = 0.2$, and 0.05 gain compared with the $\lambda = 0.5$. In the case of the Art domain, the previous behavior remained the same where $\lambda = 0.8$ had a gain of 0.01 when compared with the performance of $\lambda = 0.2$ and 0.04 gain when compared with $\lambda = 0.5$.

Considering Figure 7, the best λ inside our searching space for all domains was again 0.8, i.e., the consistency has the importance of 0.8 against the 0.2 for the diversity

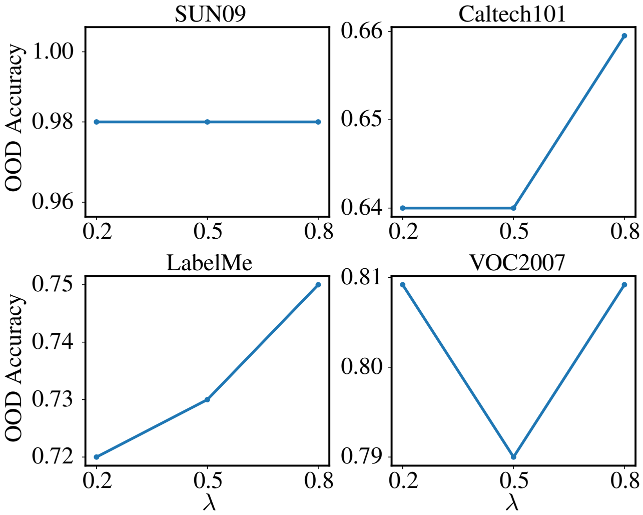

Figure 7. Effect of hyperparameter λ on VLCS dataset. Each plot represents one domain: SUN09, Caltech101, LabelMe, and VOC2007. On the x-axis is the λ and on the y-axis is the OOD accuracy.

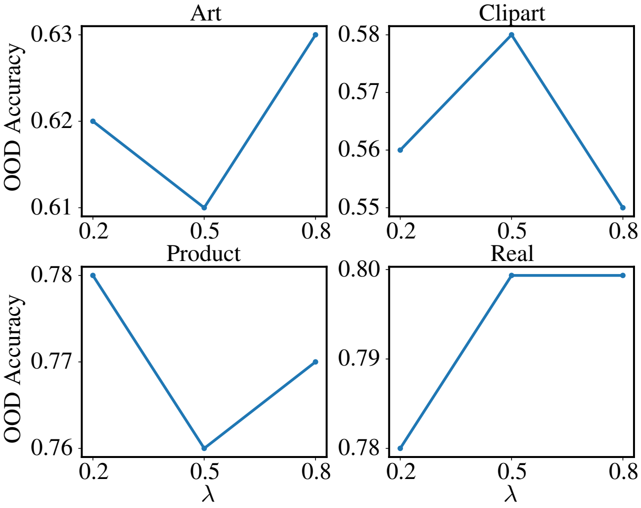

Figure 8. Effect of hyperparameter λ on OfficeHome dataset. Each plot represents one domain: Art, Clipart, Product, and Real. On the x-axis is the λ and on the y-axis is the OOD accuracy.

value. Robustness to the chosen λ was observed for the SUN09 domain. In Caltech101, $\lambda = 0.8$ had a gain of 0.01 compared with the other two. Similarly, for LabelMe domain the $\lambda = 0.8$ was the best, with improvement of 0.03 and 0.02 for $\lambda = 0.2$ and $\lambda = 0.5$, respectively. So for this domain, a high λ for consistency value is better than a lower one. For the VOC2007, both extreme cases are good, but $\lambda = 0.5$ was worse by 0.01 compared with $\lambda = 0.2$ and $\lambda = 0.8$. The OOD accuracy obtained in both cases was 0.80.

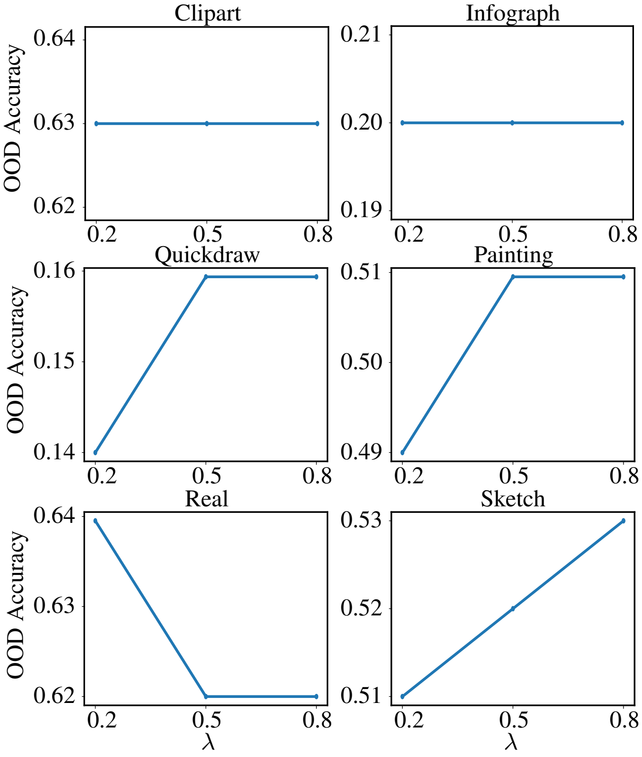

Figure 9. Effect of hyperparameter λ on DomainNet dataset. Each plot represents one domain: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. On the x-axis is the λ and on the y-axis is the OOD accuracy.

As shown in Figure 8, the domain Art had 0.01 gain with $\lambda = 0.8$ when compared with $\lambda = 0.2$ and 0.02 gain compared with $\lambda = 0.5$, so far for this domain art, the consistency improved more than the diversity. Considering domain Clipart, the $\lambda = 0.5$ was the best with 0.58 acc. OOD, so consistency and diversity are important for this domain, which has 0.02 gain compared with $\lambda = 0.2$ and 0.03 gain compared with $\lambda = 0.8$. For domain Product, the $\lambda = 0.2$ had 0.78 OOD acc., and it was better than $\lambda = 0.5$ and $\lambda = 0.8$ by 0.01 and 0.02 respectively. For the Real domain, it required a balance of consistency and diversity, or higher λ, i.e., greater than 0.5, to gain on the OOD acc.. The best performance obtained was of 0.80.

As illustrated in Figure 9, the Clipart and Infograph domains were not impacted by the value of λ in terms of OOD accuracy. On the other hand, in the Painting domain, a λ greater than 0.5 is preferable, with an increase on the OOD acc. of 0.01 for $\lambda = 0.5$ when compared with $\lambda = 0.2$. The same trend occurred with Quickdraw domain with 0.15 OOD acc. Regarding the Real domain, more diversity can increase the OOD acc., so the $\lambda = 0.2$ was better than values of 0.5 and 0.8. For such a value of λ the performance was 0.63. Finally, in the Sketch domain, the performance was linearly correlated with the value of λ. So $\lambda = 0.8$ had the best OOD acc. with 0.53 value, which represents an increase of 0.02 compared with $\lambda = 0.2$ and 0.01 increase compared with $\lambda = 0.5$.

C. Computational cost

Method	**Minibatch Time (s)**
ERM	0.13
DCAug^{domain}	0.25
DCAug^{label}	0.21
TeachDCAug^{label}	0.21

Table 7. Training iteration time for a minibatch of 32 samples on PACS dataset for ERM and our methods.

Compared to ERM, DCAug has a small additional computational cost. In particular, DCAug^{domain}, other than updating the parameters of both domain and label classifier, for each sample computes the loss of the domain classifier twice without the need to calculate the gradients. As we can the in Table 7, on an NVIDIA-A100 GPU, this roughly amounts to twice the slower step time than regular ERM.

D. Quantifying mechanisms of data Augmentation using affinity and diversity

We use the following definition of Affinity and Diversity (as defined in [14]):

Affinity (Consistency): *Let a be an augmentation and D_{train} and D_{val} be training and validation datasets drawn IID from the same clean data distribution, and let D'_{val} be derived from D_{val} by applying a stochastic augmentation strategy, a, once to each image in D_{val}, $D'_{val} = \{(a(x), y) : \forall(x, y) \in D_{val}\}$. Further let m be a model trained on D_{train} and $A(m, D)$ denote the model's accuracy when evaluated on dataset D. The Affinity, $T[a; m; D_{val}]$, is given by*

$$T[a; m; D_{val}] = A(m, D_{val}) - A(m, D'_{val}). \qquad (7)$$

Diversity: *Let a be an augmentation and D'_{train} be the augmented training data resulting from applying the augmentation, a, stochastically. Further, let L_{train} be the training loss for a model, m, trained on D'_{train}. We define the Diversity, $D[a; m; D_{train}]$, as*

$$D[a; m; D_{train}] = \mathbb{E}_{D'_{train}}[L_{train}]. \qquad (8)$$

E. DCAug with ViT backbone

In this section, we investigate the robustness of the proposed method to the choice of pretrained models, particularly the ViT backbone [10]. To be able to compare with Resnet50, we use the ViT-B-16 variant which is the base model with a patch size of 16. We use the same experimental setup as before and use the PACS dataset to evaluate the models. As we can see from Table 8, while TA seems to lose accuracy when using ViT, our approach shows consistent improvements over the ERM baseline.

Method	**OOD Accuracy**
ERM	85.0±1.1
TA	83.8±1.0
DCAug^{domain}	85.4±0.6
DCAug^{label}	84.3±0.9
TeachDCAug^{label}	88.4±0.5

Table 8. Out-of-domain performance of models based on ViT-B-16 backbone on PACS dataset. Our experiments are repeated three times.

F. Full Results

In this section, we show detailed results of Table 3 of the main manuscript. Tables 9, 10, 11, 12 13 show full results on PACS, VLCS, OfficeHome, TerraIncognita and DomainNet datasets, respectively. The provided tables summarize the obtained out-of-distribution accuracy for every domain within the four datasets. Standard deviations are reported from three trials, when available. The results for the methods were gathered from [15] and [7]. To guarantee the comparability of the results, we followed the same experimental setting as in DomainBed [15].

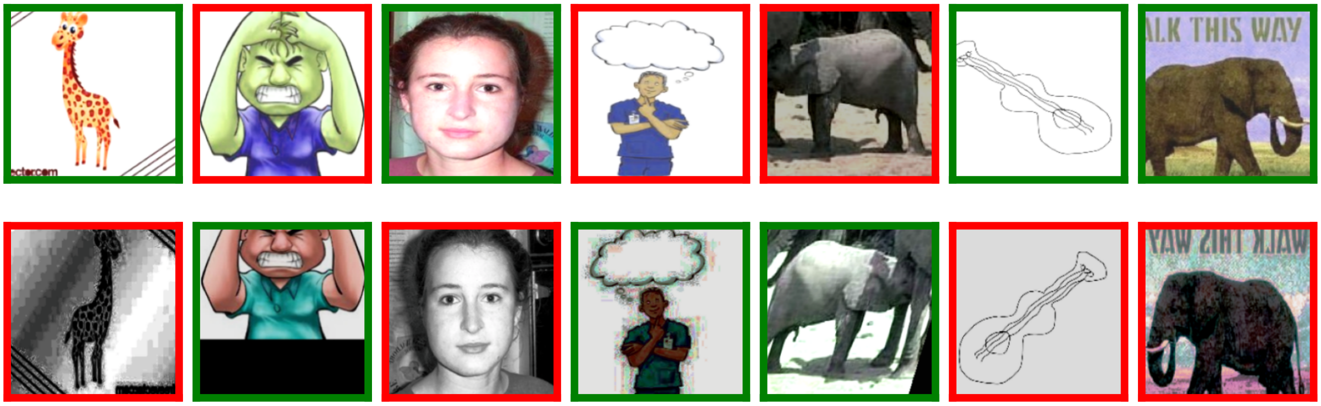

Figure 10. Visual changes of the selected images by our proposed data augmentation schema. For each sample within a minibatch, our method produces two augmentations of $\mathcal{T}_{weak}$ (top row) and $\mathcal{T}_{wider}$ (bottom row). After calculating R_{div} and R_{con} for each, our method selects the transformation with the highest reward (green box) and rejects the other one (red box).

Method	**Category**	**Domain**				
		Art	Cartoon	Photo	Sketch	**Avg.**
ERM [44]	*Baseline*	85.7±0.6	77.1±0.8	97.4±0.4	76.6±0.7	84.2
MMD [26]	*Domain-Invariant*	86.1±1.4	79.4±0.9	96.6±0.2	76.5±0.5	84.7
IRM [2]		84.8±1.3	76.4±1.1	96.7±0.6	76.1±1.0	83.5
GroupDRO [39]		83.5±0.9	79.1±0.6	96.7±0.3	78.3±2.0	84.4
DANN [12]		86.4±0.8	77.4±0.8	97.3±0.4	73.5±2.3	83.7
CORAL [41]		88.3±0.2	80.0±0.5	97.5±0.3	78.8±1.3	86.2
mDSDI [6]		87.7±0.4	80.4±0.7	98.1±0.3	78.4±1.2	86.2
DDAIG [55]	*Data Augmentation*	84.2	78.1	95.3	74.7	83.1
MixStyle [56]		86.8±0.5	79.0±1.4	96.6±0.1	78.5±2.3	85.2
RSC [19]		85.4±0.8	79.7±1.8	97.6±0.3	78.2±1.2	85.2
Mixup [48]		86.1±0.5	78.9±0.8	97.6±0.1	75.8±1.8	84.6
SagNets [34]		87.4±1.0	80.7±0.6	97.1±0.1	80.0±0.4	86.3
DCAug^{domain} (Ours)		87.5±0.7	79.0±1.5	96.3±0.1	81.5±0.9	86.1
DCAug^{label} (Ours)		88.5±0.8	78.8±1.5	96.3±0.1	80.8±0.5	86.1
TeachDCAug^{label} (Ours)		89.6±0.0	81.8±0.5	97.7±0.0	84.5±0.2	88.4

Table 9. Out-of-domain accuracies (%) on PACS.

Method	**Category**	**Domain**				
		Caltech101	LabelMe	SUN09	VOC2007	**Avg.**
ERM [44]	*Baseline*	98.0±0.3	64.7±1.2	71.4±1.2	75.2±1.6	77.3
MMD [26]	*Domain-Invariant*	97.7±0.1	64.0±1.1	72.8±0.2	75.3±3.3	77.5
IRM [2]		98.6±0.1	64.9±0.9	73.4±0.6	77.3±0.9	78.6
GroupDRO [39]		97.3±0.3	63.4±0.9	69.5±0.8	76.7±0.7	76.7
DANN [12]		99.0±0.3	65.1±1.4	73.1±0.3	77.2±0.6	78.6
CORAL [41]		98.3±0.1	66.1±1.2	73.4±0.3	77.5±1.2	78.8
mDSDI [6]		97.6±0.1	66.4±0.4	74.0±0.6	77.8±0.7	79.0
MixStyle [56]	*Data Augmentation*	98.6±0.3	64.5±1.1	72.6±0.5	75.7±1.7	77.9
RSC [19]		97.9±0.1	62.5±0.7	72.3±1.2	75.6±0.8	77.1
Mixup [48]		98.3±0.6	64.8±1.0	72.1±0.5	74.3±0.8	77.4
SagNets [34]		97.9±0.4	64.5±0.5	71.4±1.3	77.5±0.5	77.8
DCAug^{domain} (Ours)		98.3±0.3	64.7±0.2	74.2±0.6	78.3±0.8	78.9
DCAug^{label} (Ours)		98.3±0.1	64.2±0.4	74.4±0.6	77.5±0.3	78.6
TeachDCAug^{label} (Ours)		98.5±0.1	63.7±0.3	75.6±0.5	77.0±0.7	78.7

Table 10. Out-of-domain accuracies (%) on VLCS.

Method	**Category**	**Domain**				
		Art	Clipart	Product	Real	**Avg.**
ERM [44]	*Baseline*	63.1±0.3	51.9±0.4	77.2±0.5	78.1±0.2	67.6
MMD [26]	*Domain-Invariant*	60.4±0.2	53.3±0.3	74.3±0.1	77.4±0.6	66.4
IRM [2]		58.9±2.3	52.2±1.6	72.1±2.9	74.0±2.5	64.3
GroupDRO [39]		60.4±0.7	52.7±1.0	75.0±0.7	76.0±0.7	66.0
DANN [12]		59.9±1.3	53.0±0.3	73.6±0.7	76.9±0.5	65.9
CORAL [41]		65.3±0.4	54.4±0.5	76.5±0.1	78.4±0.5	68.7
mDSDI [6]		68.1±0.3	52.1±0.4	76.0±0.2	80.4±0.2	69.2
DDAIG [55]	*Data Augmentation*	59.2	52.3	74.6	76.0	65.5
MixStyle [56]		51.1±0.3	53.2±0.4	68.2±0.7	69.2±0.6	60.4
RSC [19]		60.7±1.4	51.4±0.3	74.8±1.1	75.1±1.3	65.5
Mixup [48]		62.4±0.8	54.8±0.6	76.9±0.3	78.3±0.2	68.1
SagNets [34]		63.4±0.2	54.8±0.4	75.8±0.4	78.3±0.3	68.1
DCAug^{domain} (Ours)		62.4±0.4	56.7±0.5	77.0±0.4	79.0±0.1	68.8
DCAug^{label} (Ours)		61.8±0.6	55.4±0.6	77.1±0.3	78.9±0.3	68.3
TeachDCAug^{label} (Ours)		66.2±0.2	57.0±0.3	78.3±0.1	80.1±0.0	70.4

Table 11. Out-of-domain accuracies (%) on OfficeHome.

Method	**Category**	**Domain**				
		L100	L38	L43	L46	**Avg.**
ERM [44]	*Baseline*	54.3±0.4	42.5±0.7	55.6±0.3	38.8±2.5	47.8
MMD [26]	*Domain-Invariant*	41.9±3.0	34.8±1.0	57.0±1.9	35.2±1.8	42.2
IRM [2]		54.6±1.3	39.8±1.9	56.2±1.8	39.6±0.8	47.6
GroupDRO [39]		41.2±0.7	38.6±2.1	56.7±0.9	36.4±2.1	43.2
DANN [12]		51.1±3.5	40.6±0.6	57.4±0.5	37.7±1.8	46.7
CORAL [41]		51.6±2.4	42.2±1.0	57.0±1.0	39.8±2.9	47.7
mDSDI [6]		53.2±3.0	43.3±1.0	56.7±0.5	39.2±1.3	48.1
MixStyle [56]		54.3±1.1	34.1±1.1	55.9±1.1	31.7±2.1	44.0
RSC [19]		50.2±2.2	39.2±1.4	56.3±1.4	40.8±0.6	46.6
Mixup [48]		59.6±2.0	42.2±1.4	55.9±0.8	33.9±1.4	47.9
SagNets [34]		53.0±2.9	43.0±2.5	57.9±0.6	40.4±1.3	48.6
DCAug^{domain} (Ours)		59.0±0.5	42.7±1.1	54.2±1.5	38.9±0.2	48.7
DCAug^{label} (Ours)		56.1±1.3	44.5±1.7	57.1±1.3	39.4±1.7	49.3
TeachDCAug^{label} (Ours)		60.6±0.6	43.0±2.0	58.5±0.3	42.3±1.4	51.1

Table 12. Out-of-domain accuracies (%) on TerraIncognita.

Method	**Category**	**Domain**						
		Clipart	Infograph	Painting	Quickdraw	Real	Sketch	**Avg.**
ERM [44]	*Baseline*	63.0±0.2	21.2±0.2	50.1±0.4	13.9±0.5	63.7±0.2	52.0±0.5	44.0
MMD [26]	*Domain-Invariant*	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
IRM [2]		48.5±2.8	15.0±1.5	38.3±4.3	10.9±0.5	48.2±5.2	42.3±1.1	33.9
GroupDRO [39]		42.7±0.5	17.5±0.4	33.8±0.5	9.3±0.3	51.6±0.4	40.1±0.6	33.3
DANN [12]		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
CORAL [41]		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
mDSDI [6]		62.1±0.3	19.1±0.4	49.4±0.4	12.8±0.7	62.9±0.3	50.4±0.4	42.8
MixStyle [56]	*Data Augmentation*	51.9±0.4	13.3±0.2	37.0±0.5	12.3±0.1	46.1±0.3	43.4±0.4	34.0
RSC [19]		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
Mixup [48]		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
SagNets [34]		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
DCAug^{domain} (Ours)		62.8±0.2	19.9±0.2	50.6±0.3	13.5±0.3	63.0±0.1	52.3±0.4	43.7
DCAug^{label} (Ours)		62.5±0.2	20.0±0.2	50.4±0.1	13.9±0.3	62.9±0.2	53.2±0.4	43.8
TeachDCAug^{label} (Ours)		65.5±0.0	22.2±0.0	53.7±0.0	15.6±0.1	65.8±0.1	55.9±0.1	46.4

Table 13. Out-of-domain accuracies (%) on DomainNet.