A Illustration of Datasets

Datasets We evaluate the efficacy of our approach across three representative datasets for tasks involving OOD generalization: PACS [2] with 4 domains, $\{photos, art, cartoons, sketches\}$ and 7 classes; Office-Home [3] with 4 domains $\{art, clipart, product, real\}$ and 65 classes; Terra Incognita [1] with four of the camera locations $\{L100, L38, L43, L46\}$ and 10 classes. According to Ye et al. [4], these datasets are characterized by substantial diversity shifts, aligning with our objective of leveraging domain-specific information. Figure 1 shows the illustration of images in the PACS dataset, with 4 domains and 7 classes. We also evaluate the cosine-similarity of images between domains in the PACS dataset, and Figure 2 shows the results.

Figure 1. Illustration of images in the PACS dataset.

B Hyper-parameter Search Space

General hyperparameters We set the initial learning rate to lr=5e-5 for all datasets. We also set the dropout rate for the main object classifier to zero. Table 1 shows other basic hyperparameter values for ERM and RSC.

DFP hyperparameters The major noise-related hyperparameter for our proposed Domain Feature Perturbation (DFP) approach for domain generalization is ϵ, which regulates the standard deviation $\sigma = [|\nabla_z L_d|/\|\nabla_z L_d\|_p] \cdot \epsilon$. Furthermore, the loss weights α and $1-\alpha$ are flexible in order to manage the performance of the two classifiers. We experiment with numerous ϵ and $(\alpha, 1-\alpha)$ combinations for different datasets. Table 2 depicts the grid search space of these hyper-parameters.

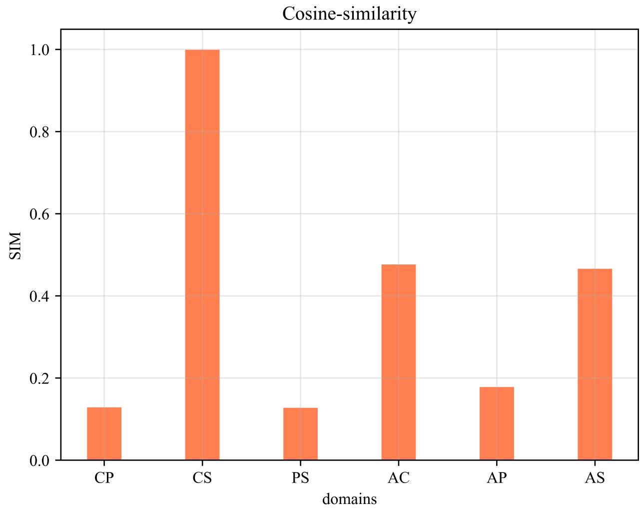

Figure 2. Cosine similarity of images between different domains in the PACS dataset.

Table 1. Basic hyper-parameters of experiments.

Parameter	Value
learning rate	5e-5
weight_decay	$10^{uniform(-6,-2)}$
resnet_dropout	0
data_augmentation	true
batch_size	default = 32, $2^{uniform(3,5.5)}$
rsc_f_drop_factor	default = $1/3$, $1^{uniform(0,0.5)}$
rsc_b_drop_factor	default = $1/3$, $1^{uniform(0,0.5)}$

Table 2. DFP hyper-parameters grid search space.

Parameter	Value
ϵ	0.2,0.1,0.05,0.01,0.005,0.001
$(\alpha, 1-\alpha)$	(0.99,0.01), (0.9,0.1)
p	∞,1,2

C More Results of DFP

C.1 Different Perturbation Levels

All of the results in this section are based on the Resnet-18 structure and are tested on the PACS dataset. Table 3 displays the complete results of several hyperparameter combinations in DFP.

C.2 Different Learning Rates

We train the model with the Resnet-18 structure and test it on the PACS dataset. We keep the main classifier's initial learning rate at 5e-5 and experiment with alternative learning rate settings for the domain classifier. The entire results of different learning rates of the domain classifier in DFP are shown in Table 4 and Table 5.

C.3 Different Datasets

The OfficeHome dataset has four training domain combinations {(C, P, R), (A, P, R), (A, C, R), (A, C, P)}, and four test domains {A,

Table 3. Test accuracies of DFP with resnet-18.

$(\alpha, 1-\alpha)$	ϵ	A	C	P	S	Avg
(0.9,0.1)	0.2	78.3 ± 0.6	72.8 ± 0.6	95.5 ± 0.4	76.5 ± 1.0	
(0.99,0.01)	0.2	77.6 ± 1.1	74.4 ± 0.5	95.5 ± 0.1	72.8 ± 1.2	
(0.9,0.1)	0.1	77.4 ± 0.1	73.6 ± 0.2	95.6 ± 0.1	76.4 ± 0.7	
(0.99,0.01)	0.1	78.3 ± 0.6	74.2 ± 2.3	**95.9 ± 0.4**	76.7 ± 1.2	
(0.9,0.1)	0.05	**80.6 ± 0.9**	72.2 ± 2.2	95.0 ± 0.3	74.6 ± 0.8	
(0.99,0.01)	0.05	75.9 ± 1.0	**74.7±0.6**	94.9 ± 0.2	**77.7 ± 0.8**	
(0.9,0.1)	0.01	79.3 ± 1.5	72.4 ± 1.0	95.2 ± 0.6	72.4 ± 1.3	
(0.99,0.01)	0.01	76.6 ± 1.8	72.9 ± 0.1	95.7 ± 0.5	70.9 ± 1.5	
(0.9,0.1)	0.005	77.0 ± 0.5	72.5± 0.2	95.6 ± 0.4	75.6± 1.2	
(0.99,0.01)	0.005	78.7 ± 1.8	74.2 ± 0.8	95.7 ± 0.2	76.2 ± 1.5	
(0.9,0.1)	0.001	79.0 ± 1.4	73.7 ± 0.2	95.7± 0.4	77.3 ± 1.4	
(0.99,0.01)	0.001	75.4 ± 0.9	73.3 ± 0.6	95.6 ±0.3	77.0 ± 1.0	
(0.9,0.1)	0.0005	78.0 ± 0.8	70.8 ± 1.0	95.4 ± 0.2	73.6± 0.9	
(0.99,0.01)	0.0005	77.4 ± 0.6	73.7 ± 0.8	95.7 ± 0.3	77.7 ± 1.1	
	best	**80.6 ± 0.9**	**74.7±0.6**	**95.9 ± 0.4**	**77.7 ± 0.8**	**82.2**

Table 4. Test accuracies of DFP with learning rates (1e−4).

$(\alpha, 1-\alpha)$	ϵ	A	C	P	S	Avg
(0.9,0.1)	0.1	78.1 ± 1.4	73.0 ± 2.3	95.4 ± 0.2	75.2 ± 0.9	
(0.99,0.01)	0.1	76.3 ± 1.1	73.4 ± 1.9	95.7 ± 0.2	75.3 ± 1.7	
(0.9,0.1)	0.05	**80.1 ± 0.5**	74.1 ± 0.7	95.3 ± 0.1	76.5 ± 0.1	
(0.99,0.01)	0.05	76.7 ± 1.0	72.1 ± 1.5	95.3 ± 0.3	75.4 ± 0.9	
(0.9,0.1)	0.01	80.0 ± 0.3	73.7 ± 0.2	95.5 ± 0.2	76.6 ± 0.8	
(0.99,0.01)	0.01	78.2 ± 0.6	73.2 ± 0.5	95.6 ± 0.5	75.6 ± 0.4	
(0.9,0.1)	0.005	76.4 ± 1.4	73.4 ± 0.5	95.3 ± 0.4	76.9 ± 1.3	
(0.99,0.01)	0.005	78.2 ± 0.7	74.3 ± 0.7	95.5 ± 0.1	**77.2 ± 0.5**	
(0.9,0.1)	0.001	78.8 ± 1.3	**74.5 ± 0.6**	95.6 ± 0.4	75.8 ± 0.4	
(0.99,0.01)	0.001	78.8 ± 0.8	74.0 ± 0.7	**96.2 ± 0.3**	74.8 ± 2.5	
	best	**80.1 ± 0.1**	**74.5 ± 0.6**	**96.2± 0.3**	**77.2 ± 0.5**	**82**

Table 5. Test accuracies of DFP with learning rates (1e−5).

$(\alpha, 1-\alpha)$	ϵ	A	C	P	S	Avg
(0.9,0.1)	0.1	76.4 ± 0.4	74.4 ± 0.6	95.5 ± 0.4	75.7 ± 1.2	
(0.99,0.01)	0.1	77.8 ± 1.2	75.5 ± 0.2	95.3 ± 0.3	77.8 ± 0.9	
(0.9,0.1)	0.05	77.5 ± 1.3	73.5 ± 1.1	95.1 ± 0.3	77.2 ± 0.3	
(0.99,0.01)	0.05	77.2 ± 0.4	74.7 ± 1.0	95.2 ± 0.1	75.3 ± 0.7	
(0.9,0.1)	0.01	77.5 ± 1.4	74.1 ± 0.8	94.2 ± 0.6	74.1 ± 1.2	
(0.99,0.01)	0.01	76.9 ± 1.3	**75.8 ± 1.4**	95.0 ± 0.4	76.4 ± 1.5	
(0.9,0.1)	0.005	**78.9 ± 1.3**	73.5 ± 0.9	95.2 ± 0.1	76.2 ± 1.6	
(0.99,0.01)	0.005	78.5 ± 0.9	74.5 ± 1.1	95.7 ± 0.3	**78.3 ± 1.2**	
(0.9,0.1)	0.001	76.3 ± 0.9	72.1 ± 0.9	**96.1 ± 0.3**	75.7 ± 0.3	
(0.99,0.01)	0.001	75.7 ± 1.9	74.9 ± 0.5	95.6 ± 0.5	76.8 ± 0.2	
	best	**78.9 ± 1.3**	**75.8 ± 1.4**	**96.1± 0.3**	**78.3 ± 1.2**	**82.275**

C, P, R}. The Terra Incognita dataset also includes four training domain combinations {(L38, L43, L46), (L100, L43, L46), (L100, L38, L46), (L100, L38, L43)}, as well as four test domain types {L100, L38, L43, L46}. Table 6 displays the outcomes of the OfficeHome dataset with various hyperparameter combinations. Table 7 illustrates the Terra Incognita dataset results with various hyperparameter combinations.

Table 6. Test accuracies of DFP on OfficeHome.

$(\alpha, 1-\alpha)$	ϵ	A	C	P	S	Avg
(0.9,0.1)	0.1	57.4 ± 0.9	50.1 ± 0.4	73.4 ± 0.1	74.3 ± 0.4	
(0.9,0.1)	0.01	56.3 ± 0.2	50.2 ± 0.4	73.0 ± 0.3	74.3 ± 0.4	
(0.9,0.1)	0.001	56.0 ± 0.1	51.0 ± 0.4	72.9 ± 0.3	74.1 ± 0.2	
	best	**57.4 ± 0.9**	**51.0 ± 0.4**	**73.4 ± 0.1**	**74.3 ± 0.4**	**64.0**

Table 7. Test accuracies of DFP on Terra Incognita.

$(\alpha, 1-\alpha)$	ϵ	L100	L38	L43	L46	Avg
(0.9,0.1)	0.1	45.3 ± 4.1	36.4 ± 1.5	52.6 ± 0.0	33.6 ± 1.1	
(0.9,0.1)	0.01	44.4 ± 4.2	36.7 ± 2.6	50.5 ± 1.1	36.8 ± 0.1	
(0.99,0.01)	0.01	47.4 ± 1.9	39.7 ± 2.7	51.2 ± 0.4	37.2 ± 0.9	
	best	**47.4 ± 1.9**	**39.7 ± 2.7**	**52.6 ± 0.0**	**37.2 ± 0.9**	**44.2**

C.4 Different Model Architectures

Table 8 displays results based on the Resnet-50 structure with various hyperparameter combinations. We test the model on the PACS dataset.

Table 8. Test accuracies of ERM and DFP with Resnet-50.

$(\alpha, 1-\alpha)$	ϵ	A	C	P	S	Avg
	ERM	82.6 ± 1.1	79.7 ± 0.4	97.3 ± 0.2	74.7 ± 1.3	83.575
(0.9, 0.1)	0.001	84.3 ± 0.9	76.9 ± 0.3	97.0 ± 0.3	75.5 ± 1.0	
	0.005	82.0 ± 1.3	78.2 ± 0.8	96.8 ± 0.2	74.1 ± 2.4	
	0.01	81.5 ± 0.8	77.9 ± 0.9	96.5 ± 0.1	**80.3 ± 0.8**	
	0.05	84.1 ± 1.2	79.6 ± 0.4	96.6 ± 0.4	77.8 ± 0.3	
	0.1	**84.7 ± 1.6**	79.6 ± 0.3	96.7 ± 0.2	74.1 ± 3.1	
(0.99, 0.01)	0.001	82.8 ± 0.8	79.5 ± 0.4	**97.0 ± 0.1**	79.9 ± 1.1	
	0.005	80.2 ± 1.2	75.5 ± 1.0	96.7 ± 0.2	77.6 ± 0.5	
	0.01	82.7 ± 0.5	77.4 ± 1.2	96.6 ± 0.2	78.0 ± 0.3	
	0.05	83.4 ± 0.8	**79.7 ± 1.2**	96.8 ± 0.1	76.8 ± 1.7	
	0.1	84.0 ± 0.5	79.1 ± 0.4	96.8 ± 0.2	74.8 ± 2.3	
	best	**84.7 ± 1.6**	**79.7 ± 1.2**	**97.0 ± 0.1**	**80.3 ± 0.8**	**85.425**

D More Results of Ablation Studies

D.1 Noise injection point

Table 9 shows the outcomes with different positions for adding the perturbations. The results are based on ResNet-18 model and the PACS dataset.

Table 9. Test accuracies of ERM with random perturbation.

$(\alpha, 1-\alpha)$	ϵ	A	C	P	S	Avg
(0.9, 0.1)	0.2	78.5 ± 0.7	72.6 ± 0.9	95.6 ± 0.0	72.6 ± 1.8	
	0.1	76.5 ± 1.0	70.9 ± 1.3	94.9 ± 0.4	75.4 ± 0.5	
	0.05	78.1 ± 0.2	73.5 ± 0.9	95.2 ± 0.7	77.2 ± 0.9	
	0.01	78.5 ± 1.1	73.3 ± 1.4	95.3 ± 0.2	76.8 ± 1.4	
	0.005	77.4 ± 0.9	73.5 ± 0.8	95.1 ± 0.4	75.4 ± 0.7	
	0.001	76.5 ± 0.1	71.4 ± 1.2	94.9 ± 0.5	75.5 ± 0.6	
(0.99, 0.01)	0.2	74.5 ± 0.9	72.6 ± 0.3	95.8 ± 0.3	74.8 ± 0.6	
	0.1	78.5 ± 0.9	76.0 ± 1.0	95.3 ± 0.2	73.0 ± 3.3	
	0.05	78.1 ± 1.2	73.7 ± 1.1	95.3 ± 0.5	75.4 ± 0.2	
	0.01	76.0 ± 0.5	75.6 ± 1.0	95.8 ± 0.2	76.3 ± 1.0	
	0.005	79.8 ± 1.0	72.8 ± 0.7	95.4 ± 0.1	73.0 ± 2.7	
	0.001	79.9 ± 0.4	73.4 ± 2.1	96.0 ± 0.3	74.1 ± 2.2	
	best	**79.9 ± 0.4**	**76.0 ± 1.0**	**96.0 ± 0.3**	**77.2 ± 0.9**	**82.3**

D.2 Sensitivity Analysis

To scrutinize the sensitivity of our proposed approach to different loss weights, we conduct experiments on the PACS dataset, maintaining consistent parameters such as the initial learning rate lr=5e-5 and a training duration of 7000 steps. In addition, we run one random search of basic hyperparameters for each of the five independent training series. We set the random noise $n \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.05$. And the loss weights $(\alpha, 1-\alpha) \in \{(0.5, 0.5), (0.6, 0.4), (0.7, 0.3), (0.8, 0.2), (0.9, 0.1), (0.99, 0.01)\}$. The results are shown in Table 10 and Figure 3.According to the results, we primarily utilize loss weights $(\alpha, 1-\alpha) \in \{(0.9, 0.1), (0.99, 0.01)\}$ for other investigations.

Table 10. Test accuracies of DFP with different loss weights.

ϵ	$(\alpha, 1-\alpha)$	A	C	P	S	Avg
0.05	(0.99, 0.01)	79.1 ± 0.8	73.9 ± 0.5	95.5 ± 0.4	75.2 ± 1.4	80.9
	(0.9, 0.1)	77.4 ± 0.6	72.1 ± 0.8	95.4 ± 0.1	77.1 ± 0.9	80.5
	(0.8, 0.2)	77.7 ± 0.6	73.4 ± 0.8	95.5 ± 0.2	74.1 ± 0.9	80.2
	(0.7, 0.3)	77.6 ± 0.9	71.4 ± 1.2	95.5 ± 0.2	73.1 ± 1.3	79.4
	(0.6, 0.4)	78.3 ± 0.7	70.7 ± 0.6	95.3 ± 0.3	74.3 ± 0.9	79.5
	(0.5, 0.5)	76.6 ± 0.7	70.0 ± 1.1	95.9 ± 0.3	75.3 ± 1.0	79.7

D.3 Random Perturbations

We set the random noise $n \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.2\}$. And the results are shown in Table 11.

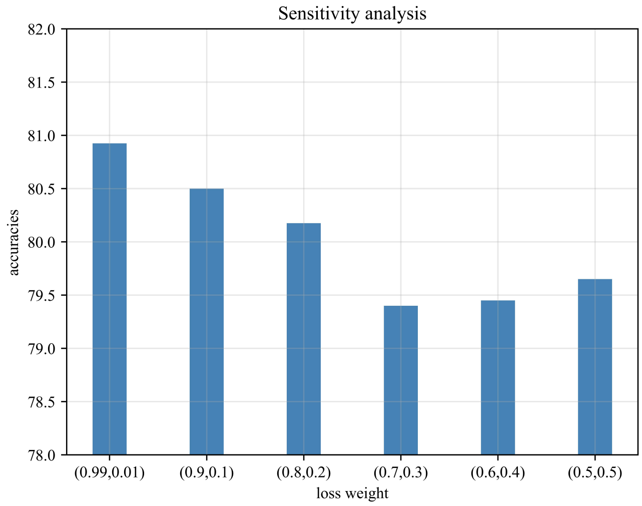

Figure 3. Accuracy results of different loss weights.

Table 11. Test accuracies of ERM with random perturbations.

ϵ	**A**	**C**	**P**	**S**	**Avg**
0.001	74.4 ± 0.3	74.3 ± 0.3	95.4 ± 0.2	74.7 ± 1.2	
0.005	75.6 ± 1.3	**75.3 ± 0.9**	**95.4 ± 0.5**	72.6 ± 1.5	
0.01	76.5 ± 1.3	73.0 ± 1.2	95.1 ± 0.1	**74.8 ± 0.7**	
0.05	77.4 ± 1.2	72.5 ± 1.2	95.3 ± 0.1	74.2 ± 1.8	
0.1	**79.6 ± 0.5**	67.8 ± 0.5	91.6 ± 0.6	62.7 ± 2.0	
0.2	70.6 ± 0.2	53.8 ± 0.8	83.6 ± 0.9	46.4 ± 5.3	
best	**79.6 ± 0.5**	**75.3 ± 0.9**	**95.4 ± 0.2**	**74.8 ± 0.7**	**81.3**

References

[1] S. Beery, G. Van Horn, and P. Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[2] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[3] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[4] N. Ye, K. Li, H. Bai, R. Yu, L. Hong, F. Zhou, Z. Li, and J. Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7947–7958, 2022.