A. Ablation Studies

In this paragraph, we first investigate the sensitivity of the model to batch size. Besides, we also conduct extensive ablation studies of our ALOFT-S on the PACS dataset, including the effects of different inserted positions in the network and the sensitivity of hyperparameters, *i.e.*, perturbation strength α and mask ratio r. The baseline is the GFNet [2] trained on the aggregation of source domains.

Model sensitivity to batch size. We here investigate the effect of different batch sizes on the performance of our ALOFT, which involves the modeling and resampling steps that are based on the samples of the current batch. As reported in Tab. 1, the results indicate that our methods perform relatively stably with different batch sizes, consistently exceeding the baseline model by approximately 2.7% (*e.g.*, achieving 91.67% accuracy compared to 87.93% with a batch size of 128). Moreover, we observe that as the batch size increases, the generalization ability of the model also improves due to the increased diversity of samples used to model the spectrum distribution. Interestingly, even with a small batch size of 4, our model still achieves promising results (*i.e.*, 90.16% accuracy of ALOFT-E). We speculate the reason to be that a small batch size could still provide some useful information for modeling the spectrum distribution. To maintain consistency with previous works [1, 5], we set the batch size as 64 for all our experiments.

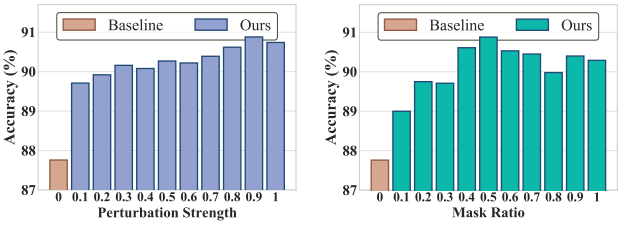
Table 1. Effect (%) of different batch sizes on the model performance. We conduct the experiments on the PACS dataset. The baseline is the GFNet model directly trained on source domains.

Batch size	4	8	16	32	64	128
Baseline	87.41	87.55	87.57	87.68	87.76	**87.93**
ALOFT-S	89.70	89.91	90.41	90.69	90.88	**90.92**
ALOFT-E	90.16	90.74	90.89	91.36	91.58	**91.67**

Different inserted positions of ALOFT-S. Here we explore the effectiveness of ALOFT-S in different positions of the network. The experimental results are reported in Tab. 2. The first line represents the results of the baseline model, which is trained using all source domains directly based on the strong baseline (*i.e.*, DeepAll [6] on GFNet). We observe that no matter which layer the ALOFT-S is inserted in, the model can consistently outperform the baseline by a significant margin, *e.g.*, 1.61% (89.37% vs. 87.76%) with ALOFT-S inserted in the first MLP block. The results indicate that our method is effective in enhancing the feature diversity at different layers. Moreover, applying ALOFT-S to all blocks of the network can achieve the best performance and exceed the baseline by 3.12% (90.88% vs 87.76%), verifying that ALOFT-S can generate diverse data variants to sufficiently simulate domain shifts during training. Therefore, ALOFT-S is inserted into all blocks in our experiments, which is the same as ALOFT-E.

Table 2. Effect (%) of different inserted positions on PACS. "Blo.1-4" represent four core MLP blocks of the network. The top shows the results of applying ALOFT-S to each block. The bottom is the results of the model with ALOFT-S in multiple blocks.

Position				**PACS**				
Blo.1	Blo.2	Blo.3	Blo.4	Art	Cartoon	Sketch	Photo	Avg.
-	-	-	-	89.37	84.74	79.01	97.94	87.76
✓	-	-	-	90.67	84.60	83.84	98.38	89.37
-	✓	-	-	90.09	84.77	82.67	98.68	89.05
-	-	✓	-	90.97	85.45	81.39	98.50	89.08
-	-	-	✓	91.31	84.64	82.69	98.44	89.27
✓	✓	-	-	90.58	85.84	84.30	98.74	89.86
✓	✓	✓	-	90.77	**86.09**	85.85	98.56	90.32
✓	✓	✓	✓	**91.70**	85.49	**87.58**	**98.76**	**90.88**



(a) Effects of perturbation strength. (b) Effects of mask ratio.

Figure 1. Effects of hyper-parameters including the perturbation α and the low-frequency mask ratio r in ALOFT-S. The experiments are conducted on PACS with GFNet as the backbone architecture.

Different inserted positions of ALOFT-S. Here we explore the effectiveness of ALOFT-S in different positions of the network. The experimental results are reported in Tab. 2. The first line represents the results of the baseline model, which is trained using all source domains directly based on the strong baseline (*i.e.*, DeepAll [6] on GFNet). We observe that no matter which layer the ALOFT-S is inserted in, the model can consistently outperform the baseline by a significant margin, *e.g.*, 1.61% (89.37% vs. 87.76%) with ALOFT-S inserted in the first MLP block. The results indicate that our method is effective in enhancing the feature diversity at different layers. Moreover, applying ALOFT-S to all blocks of the network can achieve the best performance and exceed the baseline by 3.12% (90.88% vs 87.76%), verifying that ALOFT-S can generate diverse data variants to sufficiently simulate domain shifts during training. Therefore, ALOFT-S is inserted into all blocks in our experiments, which is the same as ALOFT-E.

Effects of the perturbation strength in ALOFT-S. We also investigate the effects of perturbation strength α in ALOFT-S. Recall that α is used to control the magnitude of changing the low-frequency components of images. The larger α, the greater the low-frequency spectrums change. We evaluate α on PACS and report the results in Fig. 1a, where $\alpha = 0$ means the baseline model trained merely with original frequency spectrums. As shown in Fig. 1a, when α goes up from 0.1 to 1.0, the accuracy rises from

89.71% to 90.74%, indicating that relatively strong perturbations can synthesize diverse data variants to sufficiently simulate domain shifts during training. Thus, we recommend setting α to a relatively large value, *i.e.*, selecting α from $\{0.8, 0.9, 1.0\}$ as the default value.

Effects of the mask ratio in ALOFT-S. The mask ratio r denotes the size of the binary mask $\mathcal{M} \in \mathbb{R}^{r \times r}$, which represents the scale of low-frequency components to be disturbed. As presented in Fig. 1b, with r increasing from 0.1 to 0.5, the performance slides from 89.00% to 90.88%, indicating that a relatively small could lead to insufficient perturbations of the low-frequency components. However, further increasing r causes performance degradation because the high-frequency components are disturbed, which hinders the model learning of domain-invariant features. Thus, we suggest practitioners to choose r from $\{0.4, 0.5, 0.6\}$, with $r = 0.5$ being the default setting in our experiments.

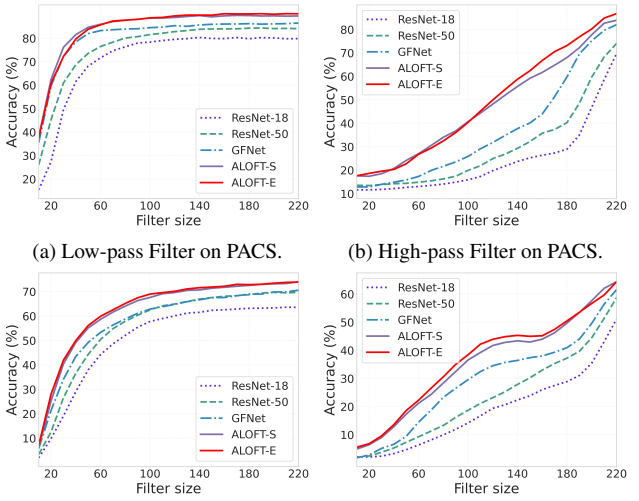
B. Further Analysis

We here conduct experiments to analyze the effectiveness of our methods, including: 1) We analyze the impact of low- and high-frequency components of frequency features; 2) We compare our methods with other low-frequency transforms; 3) We provide detailed qualitative analysis for our methods from the frequency perspective.

Why not remove the low-frequency components? We train the model only with the low-frequency components of features by filtering out the high-frequency components (namely Only LowF), and so is the model trained only with the high-frequency components (namely Only HighF), with a mask ratio r of 0.5. We also use ALOFT-S and ALOFT-E to transform the high-frequency spectrum (HighF-S and HighF-E) and both the low- and high-frequency spectrums (Both-S and Both-E), respectively. As shown in Tab. 3, compared to the baseline trained on original data, the model trained with only low-frequency components of features suffers from large performance degradation, indicating that low-frequency components contain limited global semantics. In contrast, the model trained with only high-frequency components performs better than the baseline, suggesting that high-frequency spectrums contain meaningful semantics for generalizing to unseen domains. We notice that the model trained with only high-frequency components suffers performance degradation when generalizing to cartoon and photo domains. We conjecture it is because the low-frequency components contain some semantic information, with which the model can achieve better performance. Moreover, we observe that perturbing the high-frequency spectrum can bring a slight improvement from the baseline, as it encourages the model to explore semantic information from the low-frequency components. However, directly perturbing the entire spectrum may result in a loss of important semantic information and thus hurt the model performance. Therefore, we do not remove low-frequency components but explore the ALOFT-S and ALOFT-E methods to dynamically transform the low-frequency spectrums while preserving the high-frequency spectrums.

Table 3. Effects (%) of different components of images. The experiments are conducted on the PACS dataset. The baseline is the GFNet directly trained on the aggregation of source domains.

Method	A	C	S	P	Avg.
Baseline	89.37	84.74	79.01	97.94	87.76
Only LowF	62.30	65.15	42.25	85.39	63.77
Only HighF	91.21	83.84	82.32	97.23	88.65
Swap LowF	90.31	85.73	85.09	98.17	89.82
Mix LowF	91.99	85.67	86.10	97.96	90.43
HighF-S	88.33	85.75	81.90	98.56	88.64
Both-S	91.50	85.78	85.44	98.44	90.29
HighF-E	90.72	85.79	81.85	98.32	89.17
Both-E	92.19	85.88	84.91	98.80	90.44
ALOFT-S (Ours)	91.70	85.49	87.18	98.56	90.73
ALOFT-E (Ours)	**92.24**	**87.84**	**87.38**	**98.86**	**91.58**



(a) Low-pass Filter on PACS. (b) High-pass Filter on PACS.

(c) Low-pass Filter on OfficeHome. (d) High-pass Filter on OfficeHome.

Figure 2. Comparison of ResNet-18, ResNet-50, GFNet, and our ALOFT-S and ALOFT-E on the PACS and OfficeHome datasets. A larger filter size for the low- and high-pass filtering means more low- and high-frequency components, respectively.

Comparison with other low-frequency transforms. We consider the schemes that directly exchange or mix low-frequency components between any two samples, namely Swap LowF and Mix LowF, respectively. The results in Tab. 3 show that both Swap LowF and Mix LowF can achieve significant improvements from the Only-HighF, verifying that the presence of low-frequency components can help the model generalize well to the cartoon and photo domains. Among these results, our methods still achieve the best performance, *e.g.*, ALOFT-E exceeds Mix LowF by 1.15% (91.58% vs 90.43%), demonstrating that our meth-

ods can simulate domain shifts more sufficiently than other methods. Besides, since ALOFT-E directly models and resamples each element in the low-frequency spectrums, it can synthesize more diverse data variants, thus helping the model generalize better to target domains than ALOFT-S.

Qualitative analysis for ALOFT-S and ALOFT-E. To study the effectiveness of our ALOFT-S and ALOFT-E, we here conduct detailed qualitative analysis from the frequency perspective, *i.e.*, evaluate the model performance on certain frequency components of test samples. We compare our methods with ResNet-18, ResNet-50, and GFNet which are trained directly on the aggregation of source domains. Fig. 2 present the results on PACS and OfficeHome. As shown in Fig. 2a and Fig. 2b, both ALOFT-S and ALOFT-E can remarkably improve the model performance on the high-frequency components of images, verifying their effectiveness in promoting the ability of the model to capture global structure information. We notice that our methods also perform well on the low-frequency components of images, which suggests that our methods help the model sufficiently mine the semantic features in the low-frequency components. Specifically, ALOFT-E performs better on the high-frequency components, thus it can achieve better generalization ability than ALOFT-S. The results in Fig. 2c and Fig. 2d justify the effectiveness of our methods again.

C. Additional Experiments

Domain discrepancy of extracted features To investigate the influence of our methods, we calculate the inter-domain distance (across all source domains) of the feature maps extracted by different models, including ResNet-18, GFNet [2], ALOFT-S, and ALOFT-E. We conduct the experiments on both the PACS and OfficeHome datasets. We calculate the inter-domain distance as below:

$$d = \frac{2}{K(K-1)} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} ||\overline{f}_{k_1} - \overline{f}_{k_2}||_2, \qquad (1)$$

where K is the number of source domains, $\overline{f}_{k_1}$ and $\overline{f}_{k_2}$ denote the averaged feature maps of all samples from the k_1 and k_2 domain, respectively. The results are reported in Tab. 4, from which we observe that compared to the CNN-based method (*i.e.*, ResNet-18), the strong baseline (*i.e.*, GFNet) can inherently narrow the domain gap because of its better ability to capture global structure features. Moreover, our ALOFT-S and ALOFT-E can achieve smaller domain gaps than other methods, *e.g.*, ALOFT-E reduces the domain gap of GFNet by 2.62 (11.28 vs. 13.90) on the PACS dataset. Even on the OfficeHome, a more challenging dataset with a larger number of classes than the PACS dataset, our methods can still effectively narrow the inter-domain gap among source domains. The reduced intra-domain discrepancy among source domains indicates that our methods can guide the model to extract more domain-invariant information, thus helping the model generalize better to unseen target domains than other methods.

Table 4. The inter-domain distribution gap ($\times 100$) of the extracted features by different methods. For the PACS dataset, we take Art Painting as the target domain and the others as all source domains. For OfficeHome, the target domain is Real-World and the others are source domains. The smaller the inter-domain distance, the better the generalization performance of the model.

Method	ResNet-18	GFNet	ALOFT-S	ALOFT-E
PACS	15.97	13.90	11.76	**11.28**
OfficeHome	11.56	9.95	8.88	**8.08**

Table 5. The FLOPs (G) of ALOFT compared with other models.

Method	ResNet-18	ResNet-50	RepMLP-S	GFNet	ViP-S	ALOFT-S	ALOFT-E
FLOPs (G)	1.82	4.13	2.85	2.05	6.92	2.05	2.05

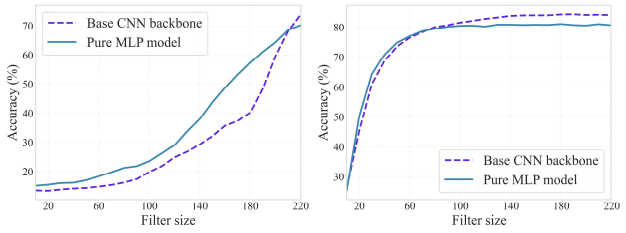
Table 6. Effects (%) of ALOFT on the ResNet architectures. The experiments are conducted on the PACS dataset.

Method	Baseline	ALOFT-S	ALOFT-E
ResNet-18	79.68	84.80	**85.13**
ResNet-50	81.15	87.52	**88.59**

Comparison of FLOPs with other models. We here compare the FLOPs of our ALOFT-S and ALOFT-E with other CNN-based or MLP-like models and report the results in Tab. 5. We observe that most existing MLP-like models suffer relatively large FLOPs, *e.g.*, the FLOPs of RepMLP-S is 2.85 and the FLOPs of ViP-S is 6.92. In contrast, the FLOPs of our ALOFT methods are comparable to the small-sized network ResNet-18, while our methods can achieve the SOTA performance and exceed the ResNet-18 by a significant magnitude, *e.g.*, 11.90% (91.58% vs. 79.68%) on the PACS dataset, proving the superiority of our ALOFT.

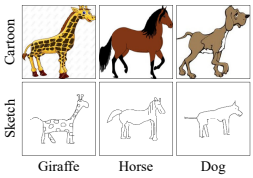
Effects of ALOFT on the ResNet architectures. To validate the generalization of our ALOFT-S and ALOFT-E modules, we insert the two modules into the ResNet-18 and ResNet-50, respectively. The experiments are conducted on the PACS dataset, and the results are reported in Tab. 6. Our ALOFT modules can improve the generalization ability of the model significantly on both the ResNet-18 and ResNet-50 networks, *e.g.*, for the ALOFT-E module, boosting 5.45% (85.13% vs. 79.68%) on ResNet-18 and 7.44% (88.59% vs. 81.15%) on ResNet-50, respectively. The above results suggest that the ALOFT modules are effective and can be generalized to various networks.

Comparisons of CNN and MLP backbones. To avoid the impact of the method itself, we here compare the difference between the base CNN backbone [6] and the pure MLP model [4]. As shown in Fig. 3, we can observe that the pure MLP model achieves a better performance than the



(a) High-pass Filtering (b) Low-pass Filtering

Figure 3. Comparison of the base CNN backbone (*i.e.*, ResNet-18) and the pure MLP backbone (*i.e.*, MLP-mixer [4]) on the PACS dataset. A larger filter size for the low- and high-pass filtering means more low- and high-frequency components, respectively.



Category	Giraffe	Horse	Dog
Baseline	80.97	95.66	57.41
ALOFT-S	84.32	97.11	65.74
ALOFT-E	**88.43**	**98.84**	**72.84**

Figure 4. Effects (%) of ALOFT for the objects with similar shapes but different textures. The figures on the left show some categories in PACS, including dogs, horses, and giraffes that have similar shapes but different textures. The right table presents the accuracy of ALOFT-S and ALOFT-E in these categories.

base CNN backbone, which indicates the effectiveness of the MLP model to capture global structure information.

For objects with similar shapes but different textures. In real-world scenes, there are instances of object categories that have similar shapes but different textures, making it difficult to distinguish between them. The key distinguishable information for these categories is often contained in the low-frequency spectrums. To resist this challenge, it is crucial to preserve semantic information by focusing on the low-frequency spectrums. Therefore, our ALOFT adopts a *perturb-while-preserve* strategy during training, where generated perturbations are applied to the original low-frequency spectrums to enhance semantic information. This strategy preserves the original low-frequency spectrums while introducing diverse noise, resulting in a more effective enhancement of semantic information. We also conduct an experiment to validate the effectiveness of the *perturb-while-preserve* strategy. Specifically, we select three representative classes from PACS with similar shapes but different textures, *i.e.*, Giraffes, Horses, and Dogs. As shown in Fig. 4, our ALOFT methods outperform the baseline model in these challenging classes.

Visual explanation. To visually verify the claim that our ALOFT can encourage the model to learn global structure information, we provide the attention maps of the last convolutional layer for ResNet-18, GFNet, ALOFT-S, and ALOFT-E utilizing the visualization technique in [3]. The

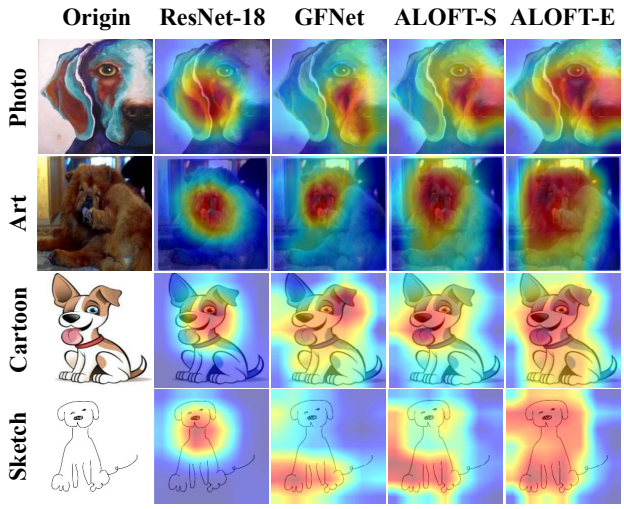


Figure 5. Visualization of attention maps of the last convolutional layer using GradCAM [3] on PACS with Sketch as the target domain. Note that the redder the area indicates the higher attention.

results are presented in Fig. 5. We can observe that the representations learned by ALOFT contain more global structure information than those learned by ResNet-18 and GFNet, which suggests that our ALOFT methods can help the model learn comprehensive domain-invariant features, enabling it to generalize well to target domains.

References

[1] Kyungmoon Lee, Sungyeon Kim, and Suha Kwak. Cross domain ensemble distillation for domain generalization. In *ECCV*, 2022. 1

[2] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *NeurIPS*, 2021. 1, 3

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 4

[4] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. In *NeurIPS*, 2021. 3, 4

[5] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *CVPR*, 2021. 1

[6] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, 2020. 1, 3