ALOFT: A Lightweight MLP-like Architecture with Dynamic Low-frequency Transform for Domain Generalization

Jintao Guo[1,2]   Na Wang[1,2]   Lei Qi[3*]   Yinghuan Shi[1,2*]

[1] State Key Laboratory for Novel Software Technology, Nanjing University
[2] National Institute of Healthcare Data Science, Nanjing University
[3] School of Computer Science and Engineering, Southeast University

{guojintao, wangna}@smail.nju.edu.cn, qilei@seu.edu.cn, syh@nju.edu.cn

Abstract

Domain generalization (DG) aims to learn a model that generalizes well to unseen target domains utilizing multiple source domains without re-training. Most existing DG works are based on convolutional neural networks (CNNs). However, the local operation of the convolution kernel makes the model focus too much on local representations (e.g., texture), which inherently causes the model more prone to overfit to the source domains and hampers its generalization ability. Recently, several MLP-based methods have achieved promising results in supervised learning tasks by learning global interactions among different patches of the image. Inspired by this, in this paper, we first analyze the difference between CNN and MLP methods in DG and find that MLP methods exhibit a better generalization ability because they can better capture the global representations (e.g., structure) than CNN methods. Then, based on a recent lightweight MLP method, we obtain a strong baseline that outperforms most state-of-the-art CNN-based methods. The baseline can learn global structure representations with a filter to suppress structure-irrelevant information in the frequency space. Moreover, we propose a dynAmic LOw-Frequency spectrum Transform (ALOFT) that can perturb local texture features while preserving global structure features, thus enabling the filter to remove structure-irrelevant information sufficiently. Extensive experiments on four benchmarks have demonstrated that our method can achieve great performance improvement with a small number of parameters compared to SOTA CNN-based DG methods. Our code is available at https://github.com/lingeringlight/ALOFT/.
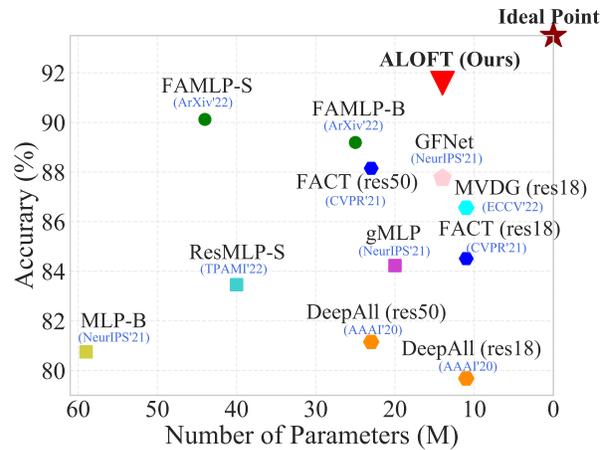

Figure 1. Comparison of the SOTA CNN-based methods, the latest MLP-like models, and our method on PACS. Among the SOTA CNN-based and MLP-based methods, our method can achieve the best performance with a relatively small-sized network.

1. Introduction

Most deep learning methods often degrade rapidly in performance if training and test data are from different distributions. Such performance degradation caused by distribution shift (*i.e.*, domain shift [3]) hinders the applications of deep learning methods in real world. To address this issue, unsupervised domain adaptation (UDA) assumes that the unlabeled target domain can be utilized during training to help narrow the potential distribution gap between source and target domains [12, 31, 57]. However, UDA methods cannot guarantee the performance of model on unknown target domains that could not be observed during training [39, 48]. Since the target domain could not always be available in reality, domain generalization (DG) is proposed as a more challenging yet practical setting, which aims to learn a model from observed source domains that performs well on arbitrary unseen target domains without re-training.

To enhance the robustness of model to domain shifts,

many DG methods intend to learn domain-invariant representations across source domains, mainly via adversarial learning [9, 63], meta-learning [5, 58], data augmentation [15, 24, 44], *etc*. Existing DG works are primarily built upon convolution neural networks (CNNs). However, due to the local processing in convolutions, CNN models inherently learn a texture bias from local representations [2, 29], which inevitably leads to their tendency to overfit source domains and perform unsatisfactorily on unseen target domains. To tackle this drawback, some pioneers propose to replace the backbone architecture of DG with transformer or MLP-like models, which can learn global representations with attention mechanisms [18, 61, 62]. Although these methods have achieved remarkable performance, few of them have analyzed how the differences between the MLP and CNN architectures affect the generalization ability of model in the DG task. These methods also suffer from excessive network parameters and high computational complexity, which hinders their applications in real-world scenarios.

In this paper, we first investigate the generalization ability of several MLP methods in the DG task and conduct the frequency analysis [2] to compare their differences with CNN methods. We observe that *MLP methods are better at capturing global structure information during inference, hence they can generalize better to unseen target domains than CNN methods.* Based on the observation, we propose an effective lightweight MLP-based framework for DG, which can suppress local texture features and emphasize global structure features during training. Specifically, based on the conventional MLP-like architecture [10, 35], we explore a strong baseline for DG that performs better than most state-of-the-art CNN-based DG methods. The strong baseline utilizes a set of learnable filters to adaptively remove structure-irrelevant information in the frequency space, which can efficiently help the model learn domain-invariant global structure features. Moreover, since the low-frequency components of images contain the most domain-specific local texture information, we propose a novel dynAmic LOw-Frequency spectrum Transform (ALOFT) to further promote the ability of filters to suppress domain-specific features. ALOFT can sufficiently simulate potential domain shifts during training, which is achieved by modeling the distribution of low-frequency spectrums in different samples and resampling new low-frequency spectrums from the estimated distribution. As shown in Fig. 1, our framework can achieve excellent generalization ability with a small number of parameters, proving its superiority in DG.

Our contributions are summarized as follows:

- We analyze how the MLP-like methods work in DG task from a frequency perspective. The results indicate that MLP-like methods can achieve better generalization ability because they can make better use of global structure information than CNN-based methods.

- We propose a lightweight MLP-like architecture with dynamic low-frequency transform as a competitive alternative to CNNs for DG, which can achieve a large improvement from the ResNet with similar or even smaller network size as shown in Fig. 1.

- For dynamic low-frequency transform, we design two variants to model the distribution of low-frequency spectrum from element-level and statistic-level, respectively. Both variants can enhance the capacity of the model in capturing global representations.

We demonstrate the effectiveness of our method on four standard domain generalization benchmarks. The results show that compared to state-of-the-art domain generalization methods, our framework can achieve a significant improvement with a small-sized network on all benchmarks.

2. Related Works

Domain generalization. Domain generalization (DG) aims to learn a robust model from multiple source domains that can generalize well to arbitrary unseen target domains. Many DG methods resort to aligning the distribution of different domains and learning domain-invariant features via domain-adversarial learning [52, 68] or feature disentanglement [6, 56]. Another popular way to address the DG problem is meta-learning, which splits the source domains into meta-train and meta-test domains to simulate domain shifts during training [21, 49, 58]. Data augmentation is also an effective technique to empower model generalization by generating diverse data invariants via domain-adversarial generation [39, 64], learnable augmentation networks [64, 65] or statistic-based perturbation [15, 20, 47]. Other DG methods also employ self-supervised learning [16, 17, 27], ensemble learning [1, 50, 67] and dropout regularization [13, 14, 38]. However, all of the above DG methods are based on CNNs and unavoidably learn a texture bias (*i.e.*, style) due to the limited receptive field of the convolutional layer. To tackle this problem, we explore an effective MLP-like architecture for DG to mitigate the texture bias of models and propose a novel dynamic low-frequency transform to enhance the ability of the model to capture global structure features.

MLP-like models. Recently, MLP-like models have achieved promising performance in various vision tasks [8, 10, 22, 23, 40, 42]. These works primarily focus on the high-complexity problem of the self-attention layer and attempt to replace it with pure MLP layers. Specifically, MLP-mixer proposes a simple MLP-like architecture with two MLP layers for performing token mixing and channel mixing alternatively [40], while ResMLP adopts a similar idea but replaces the Layer Normalization with a statistics-free Affine transformation [42]. The gMLP utilizes a spatial gating unit to re-weight tokens for enhancing spatial interactions [23]. And ViP explores the long-range dependen-

cies along the height and weight directions with linear projections [11]. These methods have been proven to mitigate the texture bias of model and shown excellent accuracy in traditional supervised learning [2, 32]. Inspired by this, we investigate how MLP-like methods work in DG task and compare the differences between the MLP-like and CNN methods. We observe that MLP-like models achieve better generalization ability because they can capture more global structure information than CNN methods. Thus, we develop a lightweight MLP-like architecture for DG with a non-parametric module to disturb the low-frequency component of samples, which can sufficiently extract domain-invariant features and generalize well to unseen target domains.

3. Proposed Method

3.1. Setting and Overview

The domain generalization task is defined as follows: given a training set of multiple observed source domains $\mathcal{D}_S = \{D_1, D_2, ..., D_K\}$ with N_k labeled samples $\{(x_i^k, y_i^k)\}_{i=1}^{N_k}$ in the k-th domain D_k, where K is the number of total source domains, x_i^k and y_i^k denote the samples and labels, respectively. The goal is to learn a model on multiple source domains $\mathcal{D}_S$ that generalizes well to arbitrary unseen target domains $\mathcal{D}_T$ with different distributions.

We first investigate the performance of MLP methods in the DG task from a frequency perspective, in which we observe that the excellent generalization ability of MLP methods is mainly owing to their stronger ability to capture global context than CNN methods. Inspired by this observation, we develop a lightweight MLP-like architecture for DG, which can effectively learn global structure information from images. As illustrated in Fig. 3, we build the architecture based on global filter network [35] and introduce a core module namely dynAmic LOw-frequency spectrum Transform (ALOFT) to simulate potential domain shifts during training. The key idea of ALOFT is to model the distribution of low-frequency components in different samples, from which we can resample new low-frequency spectrums that contain diverse local texture features. We consider different distribution modeling methods and propose two variants, *i.e.*, ALOFT-E which models by elements, and ALOFT-S which models by statistics. In the following parts, we first present the frequency analysis to compare the difference between MLP and CNN methods, and then introduce the main components of our method.

3.2. Qualitative Analysis for MLP Methods

In this paragraph, we analyze the differences in generalization ability between the MLP and CNN methods from a frequency perspective. We are motivated by the property of frequency spectrum [33, 45]: the high-frequency components preserve more global features (*e.g.*, shape) that

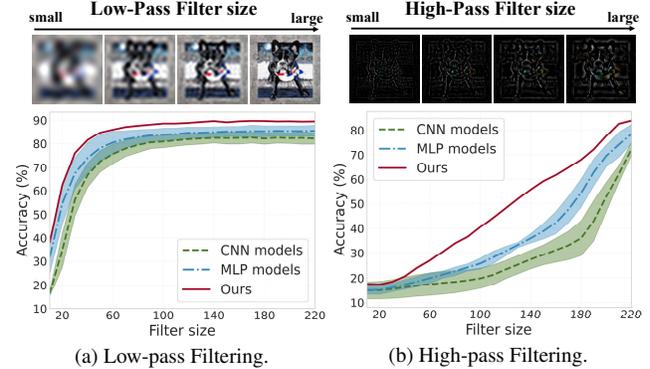(a) Low-pass Filtering.   (b) High-pass Filtering.

Figure 2. Comparison of CNN methods, MLP methods, and our methods on low- and high-pass filtered images in the target domain with different filter sizes. The experiment is conducted on PACS. A larger filter size for the low- and high-pass filtering means more low- and high-frequency components, respectively. We select three representative CNN-based DG methods, *i.e.*, DeepAll [64], FACT [51] and MVDG [58] with ResNet-18 as the backbone. For MLP methods, we employ three state-of-the-art methods, including GFNet [35], RepMLP [7] and ViP [11].

are domain-invariant, while the low-frequency components contain more local features (*e.g.*, texture) that are domain-specific. Therefore, we evaluate the performance of MLP and CNN methods on certain frequency components of test samples with discrete Fourier Transform (DFT). Below, we first describe how to obtain the high- and low-frequency components of images, and then conduct a detailed analysis of how MLP methods work in the DG task.

Extract high- and low-frequency components. Given an input image $x_i \in \mathbb{R}^{H \times W \times C}$, where H, W and C denote the height, width, and number of channels, respectively. we first obtain the Fourier transformations of input features x:

$$\mathcal{F}(x_i)(u, v, c) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x_i(h, w, c) e^{-j2\pi(\frac{h}{H}u + \frac{w}{W}v)},$$
(1)

where $j^2 = -1$. The low-frequency components are shifted to the center of the frequency spectrum by default in our experiments. Then, we introduce a binary mask $\mathcal{M} \in \mathcal{R}^{r \times r}$, whose value is zero except for the center region:

$$\mathcal{M}_{u,v} = \begin{cases} 1, \text{if} \max(|u - \frac{H}{2}|, |v - \frac{W}{2}|) \leq \frac{r \cdot \min(H, W)}{2} \\ 0, \text{otherwise} \end{cases},$$
(2)

where r is the ratio to control the size of $\mathcal{M}$ that distinguishes between high- and low-frequency components. Then we can obtain the low-pass filtered frequency $\mathcal{F}_l(x_i)$ and high-pass filtered frequency $\mathcal{F}_h(x_i)$ as follows:

$$\mathcal{F}_l(x_i) = \mathcal{M} \odot \mathcal{F}(x_i),$$
(3)

$$\mathcal{F}_h(x_i) = (I - \mathcal{M}) \odot \mathcal{F}(x_i),$$
(4)

where $\odot$ is element-wise multiplication. Finally, we use the inverse DFT to convert the frequency back to the spatial domain and obtain the low- and high-pass filtered images:
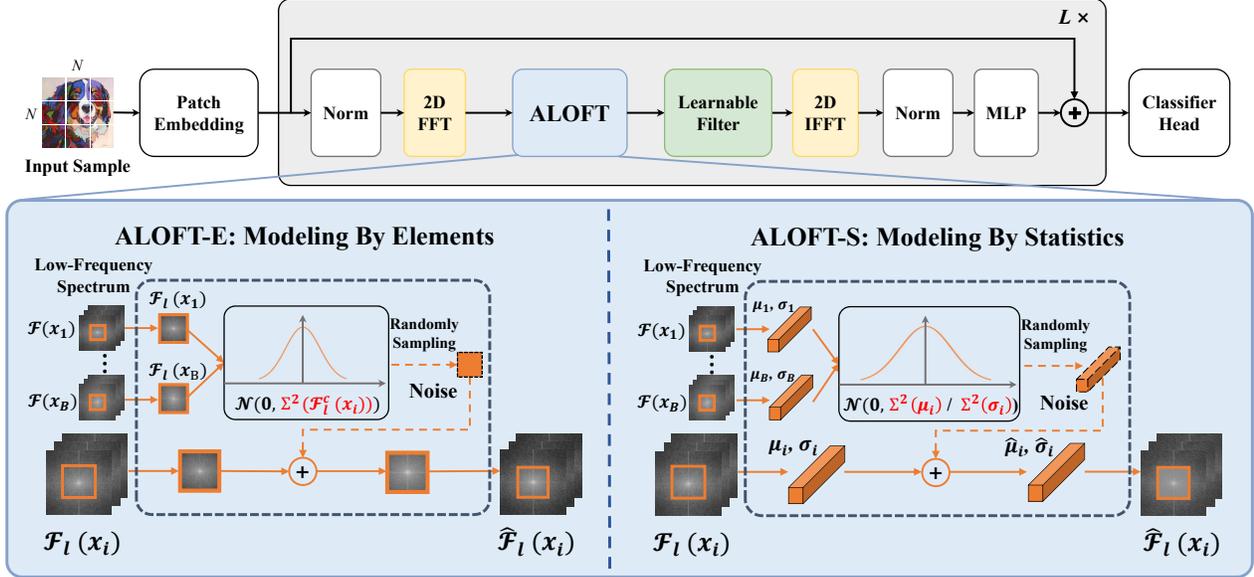
Figure 3. The overall architecture of the proposed ALOFT Framework. The core MLP-like module of our architecture contains 1) a 2D Fast Fourier Transform (FFT) to map the input spatial features to the frequency space; 2) a dynAmic LOw-Frequency Transform (ALOFT) that perturbs local texture features while preserving global structure features; 3) a learnable filter to further remove structure-irrelevant features; 4) a 2D Inverse Fast Fourier Transform (IFFT) to convert the features back to the spatial domain. We design two variants of ALOFT, *i.e.*, ALOFT-E that models distribution at element level, and ALOFT-S that models distribution at statistic level, respectively.

$$x_i^l = \mathcal{F}^{-1}(\mathcal{F}_l(x_i)), \quad x_i^h = \mathcal{F}^{-1}(\mathcal{F}_h(x_i)). \quad (5)$$

Comparision of MLP and CNN models. We here compare the performance of CNN methods with ResNet-18 and MLP methods on PACS. The results are presented in Fig. 2. The accuracy of DeepAll, FACT and MVDG on the original PACS test set is 79.68%, 84.51%, and 86.56%, respectively. And the accuracy of RepMLP, GFNet, and ViP is 84.12%, 87.76%, and 88.27%, respectively. From Fig. 2, we observe that *MLP methods perform significantly better than CNN methods on high-frequency components, while the performance is relatively close to that of CNN on low-frequency components.* Since the high-frequency components primarily retain global structure features that are consistent across different domains, MLP methods can be more robust to domain shifts and achieve better generalization ability than CNN methods. It also makes sense that MLP methods can learn long-term spatial dependencies among different patches by attention mechanisms, which can reduce the texture bias and promote the shape bias of models.

3.3. Dynamic Low-frequency Spectrum Transform

To facilitate model learning of global structure information, we propose a novel frequency transformation method namely dynAmic LOw-Frequency spectrum Transform (ALOFT). Since the high-frequency components preserve more global structure features, the ALOFT can help the model emphasize the high-frequency components by disturbing the low-frequency components. Different from

previous methods that treat low-frequency spectrum as deterministic values [51, 53], our ALOFT models the distribution of low-frequency spectrum from different samples as a Gaussian distribution, from which we resample new low-frequency spectrums to replace the original ones and simulate diverse domain shifts. Specifically, we design two practical variants for modeling the distribution of low-frequency spectrum, including ALOFT-E which estimates the element distribution in different samples, and ALOFT-S which estimates the statistic distribution in different samples.

Low-frequency spectrum transform by element. As low-frequency spectrums contain most energy distributions, they can explicitly reflect the style information of images that changes with domains [45, 51]. Therefore, it is reasonable to directly modify the element value of low-frequency spectrum to generate new data with diverse styles. To this end, we propose the ALOFT-E that models the distribution of low-frequency spectrum by element in different samples, and then randomly samples new element values to obtain new low-frequency spectrums to replace the original ones.

Concretely, given a mini-batch of input features $\{x_i\}_{i=1}^B$, where B denotes the batch size, we first obtain their Fourier transformations by Eq. (1) and extract the low-frequency components as Eq. (3). For simplicity, we denote the low- and high-Frequency components as $\{\mathcal{F}_i^l\}_{i=1}^B$ and $\{\mathcal{F}_i^h\}_{i=1}^B$, respectively. Then, we model the per-element distribution of low-frequency spectrum as a multivariate Gaussian distribution. The Gaussian distribution is centered on the original value of the corresponding element, and its variance can

be computed by the element values in different samples:

$$\Sigma^2(\mathcal{F}_i^l(u,v,c)) = \frac{1}{B}\sum_{i=1}^{B}[\mathcal{F}_i^l(u,v,c) - \mathbb{E}[\mathcal{F}_i^l(u,v,c)]]^2. \tag{6}$$

The magnitude of variance Σ^2 represents the variant intensity of elements considering underlying domain shifts. Then we resample the probabilistic value of each element in low-frequency spectrums from the estimated distribution:

$$\hat{\mathcal{F}}_i^l = \mathcal{F}_i^l + \epsilon \cdot \Sigma(\mathcal{F}_i^l), \epsilon \sim \mathcal{N}(0,\alpha), \tag{7}$$

where $\alpha \in [0,1]$ is the perturbation strength. The technique protects original low-frequency spectrum while introducing diversity noise, thus promoting semantic representation. Finally, we compose the perturbed low-frequency component $\hat{\mathcal{F}}_i^l$ and the original high-frequency component $\hat{\mathcal{F}}_i^h$ to a new frequency $\hat{\mathcal{F}}(x_i)$, and pass it to the learnable filter.

Low-frequency spectrum transform by statistic. Inspired by [20, 47] using spatial feature statistics to represent style information, we also propose ALOFT-S to model the distribution of the channel-level statistics in low-frequency spectrums. Specifically, similar as ALOFT-E, we first utilize Eq. (3) to obtain the low-frequency components $\{\mathcal{F}_i^l\}_{i=1}^B$. Then we compute channel-wise statistics (*i.e.*, mean and standard deviation) as:

$$\mu(\mathcal{F}_i^l) = \frac{1}{HW}\sum_{u=1}^{H}\sum_{v=1}^{H}\mathcal{F}_i^l(u,v,c), \tag{8}$$

$$\sigma(\mathcal{F}_i^l) = \frac{1}{HW}\sum_{u=1}^{H}\sum_{v=1}^{H}[\mathcal{F}_i^l(u,v,c) - \mu(\mathcal{F}_i^l)]^2. \tag{9}$$

We assume that the distribution of each statistic follows a Gaussian distribution, and compute the standard deviations of the statistics are computed as follows:

$$\Sigma_\mu^2(\mathcal{F}_i^l) = \frac{1}{B}\sum_{i=1}^{B}[\mu(\mathcal{F}_i^l) - \mathbb{E}[\mu(\mathcal{F}_i^l)]]^2, \tag{10}$$

$$\Sigma_\sigma^2(\mathcal{F}_i^l) = \frac{1}{B}\sum_{i=1}^{B}[\sigma(\mathcal{F}_i^l) - \mathbb{E}[\sigma(\mathcal{F}_i^l)]]^2. \tag{11}$$

In this way, we establish the Gaussian distribution for probabilistic statistics of low-frequency spectrum, from which we randomly sample new mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$:

$$\hat{\mu}(\mathcal{F}_i^l) = \mu(\mathcal{F}_i^l) + \epsilon_\mu\Sigma_\mu(\mathcal{F}_i^l), \epsilon_\mu \sim \mathcal{N}(0,\alpha), \tag{12}$$

$$\hat{\sigma}(\mathcal{F}_i^l) = \sigma(\mathcal{F}_i^l) + \epsilon_\sigma\Sigma_\mu(\mathcal{F}_i^l), \epsilon_\sigma \sim \mathcal{N}(0,\alpha), \tag{13}$$

where $\alpha \in (0,1]$ represents the strength of the perturbation. Finally, we reconstruct the low-frequency spectrum:

$$\hat{\mathcal{F}}_i^l = \hat{\mu}(\mathcal{F}_i^l)(\frac{\mathcal{F}_i^l - \mu(\mathcal{F}_i^l)}{\sigma(\mathcal{F}_i^l)}) + \hat{\sigma}(\mathcal{F}_i^l). \tag{14}$$

The above-resampled low-frequency component $\hat{\mathcal{F}}_i^l$ and the original high-frequency component $\hat{\mathcal{F}}_i^h$ are combined to form the augmented frequency $\hat{\mathcal{F}}(x_i)$ of input features.

Learnable filter. To further promote the extraction of global structure features, we utilize a learnable frequency filter $W \in \mathcal{C}^{H \times W \times C}$ to remove the structure-irrelated feature [35]. We conduct element-wise multiplication between the perturbed frequency $\mathcal{F}(x_i)$ and the learnable filter W:

$$\hat{\mathcal{F}}_{filtered}(x_i) = \hat{\mathcal{F}}(x_i) \odot W. \tag{15}$$

The filtered frequency features are finally mapped back to the spatial domain and passed to the subsequent layers.

In summary, we propose a dynamic low-frequency spectrum transform with two variants, *i.e.*, ALOFT-E and ALOFT-S that model the distribution by element and statistic, respectively. Note that both our ALOFT-E and ALOFT-S use Gaussian distribution to model the distributions of low-frequency spectrums in different samples, but we can also use other distributions, *e.g.*, Uniform distribution, to estimate the distributions. We experimentally analyze the effect of different distributions in Sec. 4.5 and find that the Gaussian distribution can produce diverse data variants to improve model performance. Besides, the ALOFT is essentially different from previous augmentation-based DG methods [20, 47, 66] that directly modify feature statistics in the spatial domain, which still could disturb the semantic features and negatively influence classification tasks. By contrast, we convert the representations into the frequency space and only perturb the low-frequency components, which can generate features with diverse styles while preserving the semantic features. In this way, our method can simulate various domain shifts and promote the ability of the model to extract domain-invariant features.

4. Experiment

4.1. Datasets

- **PACS** [19] consists of images from 4 domains: Photo, Art Painting, Cartoon, and Sketch, including 7 object categories and $9,991$ images total. We adopt the official split provided by [19] for training and validation.

- **VLCS** [41] comprises of 5 categories selected from 4 domains, VOC 2007 (Pascal), LabelMe, Caltech and Sun. We use the same setup as [3] and divide the dataset into training and validation sets based on $7:3$.

- **Office-Home** [43] contains around $15,500$ images of 65 categories from 4 domains: Artistic, Clipart, Product and Real-World. As in [3], we randomly split each domain into 90% for training and 10% for validation.

- **Digits-DG** [64] is a digit recognition benchmark consisting of four datasets MNIST, MNIST-M, SVHN, and SYN. Following [64], we randomly select 600 images per class from each domain and split the data into 80% for training and 20% for validation.

4.2. Implementation Details

Basic details. We closely follow the implementation of [35] and use the hierarchical model of GFNet, *i.e.*, the GFNet-H-Ti that has similar computational costs with the ResNet model, as the backbone. We denote the GFNet-H-Ti as GFNet for simplicity. The backbone is pre-trained on the ImageNet [36] in all of our experiments. We use 4×4 patch embedding to form the input token and utilize a non-overlapping convolution layer to downsample tokens following [35, 46]. The network depth of the GFNet backbone is 4 the same as the ResNet model. The 1-st, 2-nd, 4-th stage all contain 3 core MLP blocks. For the 3-rd stage, the number of blocks is set to 10. The embeddings dimensions of blocks in 1-st, 2-nd, 3-rd and 4-th stages are fixed as 64, 128, 256 and 512, respectively. We train the model for 50 epochs with a batch size of 64 using AdamW [25]. As in [35], We set the initial learning rate as $6.25e^{-5}$ and decay the learning rate to $1e^{-5}$ using the cosine schedule. We also use the standard augmentation protocol as in [51, 61], which consists of random resized cropping, horizontal flipping, and color jittering in our experiments.

Method-specific details. We obtain a strong baseline by directly training the GFNet on the data aggregation of source domains without other DG methods. For all experiments, we set the perturbation strength α for generating diverse low-frequency spectrums to 1.0 in ALOFT-E and 0.9 in ALOFT-S. We set the ratio r of binary mask $\mathcal{M}$, which controls the scale of low-frequency components to be disturbed, is set to 0.5 for PACS, VLCS, and Digits-DG, and 0.25 for OfficeHome. We apply the leave-one-domain-out protocol for all benchmarks. We train our model on source domains and test the model on the remaining domain. We select the best model on the validation splits of all source domains and report the top-1 classification accuracy. All the reported results are the averaged value over five runs.

4.3. Comparison with SOTA Methods

Results on PACS are presented in Tab. 1. We first compare our model with the state-of-the-art CNN-based DG methods on ResNet-18 and ResNet-50, respectively. We notice that the strong baseline (GFNet [35]) can get a promising performance, which exceeds the ResNet-18 by 8.08% (87.76% vs. 79.68%) and the ResNet-50 by 6.61 (87.76% vs 81.15%), indicating the superiority of this network structure. Furthermore, we apply our ALOFT-E and ALOFT-E to GFNet and build the advanced models, which both achieve significant improvements without introducing any extra parameters. With ALOFT-E as a representative, our method outperforms the best CNN-based method MVDG, the state-of-the-art DG method utilizing a multi-view regularized meta-learning algorithm to solve the DG problem, by 5.02% (91.58% vs. 86.56%) on ResNet-18 with the similar network sizes (14M vs. 11M) and 2.25% (91.58% vs.

Table 1. Leave-one-domain-out results on PACS. The best and second-best are **bolded** and <u>underlined</u> respectively.

Method	Params.	A	C	S	P	Avg.
CNN: ResNet-18						
DeepAll [64] (AAAI'20)	11M	78.63	75.27	68.72	96.08	79.68
FACT [51] (CVPR'21)	11M	85.37	78.38	79.15	95.15	84.51
EFDMix [60] (CVPR'22)	11M	83.90	79.40	75.00	96.80	83.90
StyleNeophile [15] (CVPR'22)	11M	84.41	79.25	83.27	94.93	85.47
I^2-ADR [30] (ECCV'22)	11M	82.90	80.80	83.50	95.00	85.60
COMEN [5] (CVPR'22)	11M	82.60	81.00	84.50	94.60	85.70
CIRL [26] (CVPR'22)	11M	86.08	80.59	82.67	95.93	86.32
XDED [17] (ECCV'22)	11M	85.60	84.20	79.10	96.50	86.40
MVDG [58] (ECCV'22)	11M	85.62	79.98	85.08	95.54	86.56
CNN: ResNet-50						
DeepAll [64] (AAAI'20)	23M	81.31	78.54	69.76	94.97	81.15
EFDMix [60] (CVPR'22)	23M	90.60	82.50	76.40	98.10	86.90
FACT [51] (CVPR'22)	23M	89.63	81.77	84.46	96.75	88.15
I^2-ADR [30] (ECCV'22)	23M	88.50	83.20	85.80	95.20	88.20
StyleNeophile [15] (CVPR'22)	23M	90.35	84.20	85.18	96.73	89.11
MVDG [58] (ECCV'22)	23M	89.31	84.22	86.36	97.43	89.33
CIRL [26] (CVPR'22)	23M	90.67	84.30	87.68	97.84	90.12
MLP-like models						
MLP-B [40] (NeurIPS'21)	59M	85.00	77.86	65.72	94.43	80.75
ResMLP-S [42] (TPAMI'22)	40M	85.50	78.63	72.64	97.07	83.46
RepMLP [7] (ArXiv'22)	38M	82.28	78.80	79.49	95.93	84.12
gMLP-S [23] (NeurIPS'21)	20M	86.72	80.80	72.13	97.54	84.23
ViP-S [11] (TPAMI'22)	25M	88.09	84.22	82.41	98.38	88.27
FAMLP-B [61] (ArXiv'22)	25M	92.06	82.49	84.09	98.10	89.19
FAMLP-S [61] (ArXiv'22)	44M	**92.63**	<u>87.03</u>	82.69	98.14	90.12
Strong Baseline	14M	89.37	84.74	79.01	97.94	87.76
ALOFT-S (Ours)	14M	91.70	85.49	**87.58**	<u>98.76</u>	<u>90.88</u>
ALOFT-E (Ours)	14M	<u>92.24</u>	**87.84**	<u>87.38</u>	**98.86**	**91.58**

89.33%) on ResNet-50 with the nearly half network sizes (14M vs. 23M). Among the SOTA MLP-like models, our model still achieves the best performance with the least parameters, *e.g.*, achieving 1.46% (91.58% vs. 90.12%) improvement while decreasing 30M (14M vs. 44M) parameters compared with the second-best method FAMLP-S [61]. The experimental results demonstrate the effectiveness and superiority of our method for domain generalization.

Results on OfficeHome are portrayed in Tab. 2. The OfficeHome is a more challenging benchmark than PACS for domain generalization because of its larger number of categories and samples. Even so, our methods can still achieve significant improvements compared with CNN-based methods, *e.g.*, ALOFT-E outperforms the state-of-the-art DG method I^2-ADR [30] by 7.55% (75.05% vs. 67.50%) on ResNet-18. Our ALOFT-E also precedes the best method ATSRL [52] on ResNet-50, which proposes a teacher-student adversarial learning scheme for DG, with a large improvement of 1.75% (75.05% vs. 73.30%). Besides, we also observe that the MLP-like models show comparable or even better results than most mainstream CNN-based models, indicating their great potential in the DG task. Our model achieves competitive performance with the SOTA MLP-like model FAMLP-S (75.05% vs. 74.82%) with a much smaller network size (14M vs. 44M). The above results further justify the efficacy of our method.

Results on Digits-DG are presented in Tab. 3. Among all the competitors, our ALOFT-E achieves the best performance, exceeding the best CNN-based method STEAM [6]

Table 2. Leave-one-domain-out results on OfficeHome. The best and second-best are **bolded** and underlined respectively.

Method	Params.	A	C	P	R	Avg.
CNN: ResNet-18						
DeepAll [64] (AAAI'20)	11M	52.06	46.12	70.45	72.45	60.27
StyleNeophile [15] (CVPR'22)	11M	59.55	55.01	73.57	75.52	65.89
COMEN [5] (CVPR'22)	11M	57.60	55.80	75.50	76.90	66.50
FACT [51] (CVPR'21)	11M	60.34	54.85	74.48	76.55	66.56
MVDG [58] (ECCV'22)	11M	60.25	54.32	75.11	77.52	66.80
CIRL [26] (CVPR'22)	11M	61.48	55.28	75.06	76.64	67.12
XDED [17] (ECCV'22)	11M	60.80	57.10	75.30	76.50	67.40
I^2-ADR [30] (ECCV'22)	11M	66.40	53.30	74.90	75.30	67.50
CNN: ResNet-50						
Fishr [34] (ICML'22)	23M	63.40	54.20	76.40	78.50	68.20
SWAD [4] (NeurIPS'21)	23M	66.10	57.70	78.40	80.20	70.60
ATSRL [52] (NeurIPS'21)	23M	69.30	60.10	81.50	82.10	73.30
MLP-like models						
ResMLP-S [42] (TPAMI'22)	40M	62.42	51.94	75.40	77.21	66.74
MLP-B [40] (NeurIPS'21)	59M	63.45	56.31	77.81	79.76	69.33
gMLP-S [23] (NeurIPS'21)	20M	64.81	58.33	75.78	79.3	69.56
ViP-S [11] (TPAMI'22)	25M	69.55	61.51	79.34	83.11	73.38
FAMLP-B [61] (ArXiv'22)	25M	69.34	62.61	79.82	82.00	73.44
FAMLP-S [61] (ArXiv'22)	44M	70.53	**64.63**	81.32	82.79	74.82
Strong Baseline	14M	66.83	55.58	78.86	80.29	70.39
ALOFT-S (Ours)	14M	71.49	60.94	82.03	83.15	74.40
ALOFT-E (Ours)	14M	**73.30**	61.12	**82.32**	**83.45**	**75.05**

Table 3. Leave-one-domain-out results on Digits-DG. The best and second-best are **bolded** and underlined respectively.

Method	Params.	MN	MN-M	SVHN	SYN	Avg.
CNN: ResNet-18						
DeepAll [64] (AAAI'20)	11M	95.80	58.80	61.70	78.60	73.70
FACT [51] (CVPR'21)	11M	97.90	65.60	72.40	90.30	81.50
COMEN [5] (CVPR'22)	11M	97.10	67.60	75.10	91.30	82.30
CIRL [26] (CVPR'22)	11M	96.08	69.87	76.17	87.68	82.50
STEAM [6] (ECCV'21)	11M	96.80	67.50	76.00	92.20	83.10
MLP-like models						
FAMLP-B [61] (ArXiv'22)	25M	98.00	83.30	84.10	96.90	90.60
Strong Baseline	14M	97.95	74.05	80.83	96.71	87.39
ALOFT-S (Ours)	14M	98.18	83.21	84.38	97.20	90.74
ALOFT-E (Ours)	14M	**98.45**	**83.35**	**84.55**	**97.37**	**90.93**

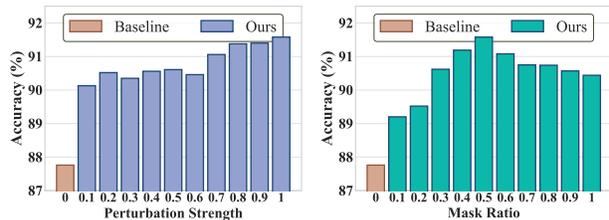Table 4. Leave-one-domain-out results on VLCS. The best and second-best are **bolded** and underlined respectively.

Method	Params.	C	L	P	S	Avg.
DeepAll [64] (AAAI'20)	11M	91.86	61.81	67.48	68.77	72.48
RSC [14] (ECCV'20)	11M	95.83	63.74	71.86	72.12	75.89
MMLD [28] (AAAI'20)	11M	97.01	62.20	73.01	72.49	76.18
StableNet [59] (CVPR'21)	11M	96.67	65.36	73.59	74.97	77.65
MVDG [58] (ECCV'22)	11M	98.40	63.79	75.26	71.05	77.13
Strong Baseline	14M	98.85	62.65	78.11	74.81	78.60
ALOFT-S (Ours)	14M	98.92	65.36	82.20	75.32	80.45
ALOFT-E (Ours)	14M	**99.36**	**65.96**	**82.91**	**77.03**	**81.31**

Table 5. Effect (%) of different inserted positions on PACS. Blo.1-4 represents four core MLP blocks of GFNet. The top reports the results of applying ALOFT-E to each block. The bottom shows the results of the model with ALOFT-E in multiple blocks.

Position				PACS				
Blo.1	Blo.2	Blo.3	Blo.4	A	C	S	P	Avg.
-	-	-	-	89.37	84.74	79.01	97.94	87.76
✓	-	-	-	91.06	83.79	83.94	98.56	89.34
-	✓	-	-	90.33	84.81	83.66	98.26	89.27
-	-	✓	-	90.67	86.35	80.96	98.62	89.15
-	-	-	✓	90.97	87.03	80.33	98.50	89.21
✓	✓	-	-	90.43	86.65	84.98	98.50	90.12
✓	✓	✓	-	91.43	86.67	86.24	98.68	90.75
✓	✓	✓	✓	**92.24**	**87.84**	**87.38**	**98.86**	**91.58**

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

Figure 4. Effects of hyper-parameters including the perturbation α and the low-frequency mask ratio r. The experiments are conducted on PACS with GFNet as the backbone architecture.

StableNet, the sophisticated method that discards the task-irrelevant features for stable learning, by 3.66% (81.31% vs. 77.65%) on average. The above results indicate that our method can effectively capture domain-invariant features, thus generalizing well to arbitrary unseen target domains.

4.4. Ablation Studies

We here conduct extensive ablation studies of ALOFT-E on the PACS dataset. We analyze the effects of different inserted positions and hyper-parameters of ALOFT-E. The ablation studies of ALOFT-S and more experiments can be found in supplementary material. The baseline is the GFNet directly trained on the aggregation of source domains.

Effect of different inserted positions. We conduct experiments on PACS using the GFNet architecture. Given that a standard GFNet model has four core MLP blocks denoted by $\mathrm{block}1-4$, we train different models with ALOFT-E inserted at different blocks. As shown in Tab. 9, no matter where the modules are inserted, the model consistently achieves higher performance than the baseline. The results show that inserting the modules at every block in GFNet has the best performance, indicating that increasing the frequency diversity in all training stages will achieve the best generalization ability. Based on the analysis, we plug the ALOFT-E module into $\mathrm{block}1, 2, 3, 4$ in all experiments.

Effects of the perturbation strength. The hyper-parameter of the perturbation strength α in Eq. (7), Eq. (12) and Eq. (13) is to control the strength of low-frequency spectrum augmentation. The larger α, the greater the mag-

by 7.83% (90.93% vs. 83.10%) on average. Our method also outperforms the SOTA MLP-based method FAMLP-B by 0.33% (90.93% vs. 90.60%) with nearly half the amount of network parameters. All the above comparisons indicate the effectiveness of our method and further demonstrate that emphasizing the high-frequency components of images can improve model generalizability across domains.

Results on VLCS are summarised in Tab. 4. We compare our models with the state-of-the-art DG methods and the results show that our models outperform existing approaches by a significant margin, e.g., ALOFT-E exceeds

nitude of low-frequency spectrum changes. We evaluate α on PACS and present the results in Fig. 5a. The results show that with α increasing from 0.1 to 1.0, the accuracy slides from 90.13% to 91.58% and consistently exceeds the baseline by a large margin, which verifies the stability of our method. The performance achieves the best value when setting α as 1.0, indicating that perturbing the low-frequency components relatively strongly can effectively enhance the generalization ability of the model. Therefore, we adopt $\alpha = 1.0$ for ALOTF-E in all experiments.

Effects of the mask ratio. The hyper-parameter of the mask ratio r denotes the size of the binary mask $\mathcal{M}$ in Eq. (2), which controls the scale of low-frequency components of images to be perturbed. The larger the mask ratio r, the more low-frequency representations are augmented. As shown in Fig. 5b, ALOFT achieves the best performance when the mask ratio is 0.5, which is also adopted as the default setting in all experiments if not specified. The results indicate that distorting the low-frequency part of features can effectively enhance the robustness of model to domain shift. We also observe that a relatively large mask ratio causes a decrease in model performance, suggesting that distorting the high-frequency components of images could hinder the model from learning domain-invariant features.

4.5. Further Analysis

In this paragraph, we compare our ALOFT with other augmentation methods on the GFNet backbone to verify the superiority of our method. We also investigate the effect of distributions other than Gaussian distribution to model low-frequency spectrums in different samples. More analytical experiments could be found in supplementary material.

Comparisons with other augmentation methods. In our experiments, we adopt the GFNet as the backbone and design a dynamic low-frequency spectrum transform to improve generalization ability of model. We also conduct other augmentation methods for comparison, including two popular image-level augmentation methods, *i.e.*, Mixup [55] and CutMix [54], and two SOTA feature-level augmentation methods, *i.e.*, MixStyle [66] and DSU [20]. As shown in Tab. 6, both the image- and feature-level augmentation methods bring performance improvements, indicating that enhancing data diversity is beneficial for the generalization ability of MLP-like models. Besides, compared to image-level augmentation methods, ALOFT generalizes better to unseen target domains, suggesting that our approach can generate more diverse data during training. Our method also outperforms the feature-level augmentation methods, *i.e.*, MixStyle and DSU, that manipulate the feature statistics in the spatial domain. The results verify that ALOFT can perturb domain-specific features while protecting domain-invariant features in the frequency space, thus helping the model generalize well to target domains.

Different distributions for modeling. With the Gaus-

Table 6. Comparisons with existing augmentation methods on PACS with GFNet as the backbone. 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
Mixup [55]	91.00	84.78	78.10	98.74	88.16
CutMix [54]	90.87	83.15	81.57	98.56	88.54
MixStyle [66]	88.72	85.32	84.88	97.49	89.10
DSU [20]	90.48	85.62	84.12	98.38	89.64
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**

Table 7. Comparisons of different distributions to model low-frequency spectrum in different samples on the PACS dataset.

Method	A	C	S	P	Avg.
Baseline	89.37	84.74	79.01	97.94	87.76
ALOFT-S					
Random	17.68	21.16	19.29	29.04	21.79
Uniform	91.80	86.48	83.63	98.74	90.16
Gaussian	91.70	85.49	87.18	98.56	90.73
ALOFT-E					
Random	52.69	38.23	33.60	27.49	38.00
Uniform	91.16	86.18	84.35	98.68	90.09
Gaussian	**92.24**	**87.84**	**87.38**	**98.86**	**91.58**

sian distribution as the default setting, we also explore other distributions for comparisons, including the Random Gaussian Distribution (denoted as Random) and the Uniform Distribution (denoted as Uniform). The random Gaussian distribution means that we directly sample random noises from $\mathcal{N}(0, 1)$ and add them to the low-frequency components. The uniform distribution means that we sample noise from $U(-\Sigma, \Sigma)$, where Σ is the variance in Eq. (6), Eq. (10) and Eq. (11). As shown in Tab. 7, the model suffers a severe performance degradation with noise drawn from random Gaussian distribution, indicating that unconstrained noise is detrimental to model learning. We notice that utilizing the noise drawn from the Uniform Distribution can also improve the model performance, suggesting that the constrained noise is beneficial to model generalization. Among all results, the Gaussian distribution achieves the best performance on both ALOFT-S and ALOFT-E, demonstrating its effectiveness in generating diverse data variants.

5. Conclusion

In this paper, we study the performance difference between MLP and CNN methods in DG and find that MLP methods can capture more global structure information than CNN methods. We further propose a lightweight MLP-like architecture with dynamic low-frequency transform for DG, which can outperform the SOTA CNN-based methods by a significant margin with a small-sized network. Our architecture can be a competitive alternative to ResNet in DG, which we hope can bring some light to the community.

References

[1] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. In *NeurIPS*, 2022. 2

[2] Jiawang Bai, Li Yuan, Shu-Tao Xia, Shuicheng Yan, Zhifeng Li, and Wei Liu. Improving vision transformers by revisiting high-frequency components. In *ECCV*, 2022. 2, 3

[3] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019. 1, 5

[4] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *NeurIPS*, 2021. 7

[5] Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. Compound domain generalization via meta-knowledge encoding. In *CVPR*, 2022. 2, 6, 7

[6] Yang Chen, Yu Wang, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. A style and semantic memory mechanism for domain generalization. In *ICCV*, 2021. 2, 6, 7

[7] Xiaohan Ding, Chunlong Xia, Xiangyu Zhang, Xiaojie Chu, Jungong Han, and Guiguang Ding. Repmlp: Reparameterizing convolutions into fully-connected layers for image recognition. *arXiv preprint arXiv:2105.01883*, 2021. 3, 6

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[9] Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. Adversarially adaptive normalization for single domain generalization. In *CVPR*, 2021. 2

[10] John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021. 2

[11] Qibin Hou, Zihang Jiang, Li Yuan, Ming-Ming Cheng, Shuicheng Yan, and Jiashi Feng. Vision permutator: A permutable mlp-like architecture for visual recognition. *IEEE TPAMI*, 2022. 3, 6, 7

[12] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. Category contrast for unsupervised domain adaptation in visual tasks. In *CVPR*, 2022. 1

[13] Zeyi Huang, Haohan Wang, Dong Huang, Yong Jae Lee, and Eric P Xing. The two dimensions of worst-case training and their integrated effect for out-of-domain generalization. In *CVPR*, 2022. 2

[14] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, 2020. 2, 7

[15] Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *CVPR*, 2022. 2, 6, 7

[16] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *ICCV*, 2021. 2

[17] Kyungmoon Lee, Sungyeon Kim, and Suha Kwak. Cross domain ensemble distillation for domain generalization. In *ECCV*, 2022. 2, 6, 7, 11

[18] Bo Li, Jingkang Yang, Jiawei Ren, Yezhen Wang, and Ziwei Liu. Sparse fusion mixture-of-experts are domain generalizable learners. *arXiv preprint arXiv:2206.04046*, 2022. 2

[19] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 5

[20] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and LINGYU DUAN. Uncertainty modeling for out-of-distribution generalization. In *ICLR*, 2021. 2, 5, 8

[21] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*. PMLR, 2019. 2

[22] Dongze Lian, Zehao Yu, Xing Sun, and Shenghua Gao. Asmlp: An axial shifted mlp architecture for vision. In *ICLR*, 2021. 2

[23] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. In *NeurIPS*, 2021. 2, 6, 7

[24] Xiao-Chang Liu, Yong-Liang Yang, and Peter Hall. Geometric and textural augmentation for domain gap reduction. In *CVPR*, 2022. 2

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[26] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *CVPR*, 2022. 6, 7

[27] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *ICML*, 2021. 2

[28] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, 2020. 7

[29] Sachin Mehta and Mohammad Rastegari. Mobilevit: Lightweight, general-purpose, and mobile-friendly vision transformer. In *ICLR*, 2022. 2

[30] Rang Meng, Xianfeng Li, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Mingli Song, Di Xie, and Shiliang Pu. Attention diversification for domain generalization. In *ECCV*, 2022. 6, 7

[31] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *CVPR*, 2022. 1

[32] Namuk Park and Songkuk Kim. How do vision transformers work? In *ICLR*, 2022. 3

[33] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 1982. 3

[34] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *ICML*, 2022. 7

[35] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *NeurIPS*, 2021. 2, 3, 5, 6, 11, 13

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 6

[37] 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. 14

[38] Baifeng Shi, Dinghuai Zhang, Qi Dai, Zhanxing Zhu, Yadong Mu, and Jingdong Wang. Informative dropout for robust representation learning: A shape-bias perspective. In *ICML*, 2020. 2

[39] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *CVPR*, 2021. 1, 2

[40] 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. 2, 6, 7, 13

[41] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 5

[42] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE TPAMI*, 2022. 2, 6, 7

[43] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017. 5

[44] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, 2018. 2

[45] Jingye Wang, Ruoyi Du, Dongliang Chang, Kongming Liang, and Zhanyu Ma. Domain generalization via frequency-domain-based feature disentanglement and interaction. In *ACM MM*, 2022. 3, 4

[46] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 6

[47] Yue Wang, Lei Qi, Yinghuan Shi, and Yang Gao. Feature-based style randomization for domain generalization. *IEEE TCSVT*, 2022. 2, 5

[48] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *ICCV*, 2021. 1

[49] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Metaalign: Coordinating domain alignment and classification for unsupervised domain adaptation. In *CVPR*, 2021. 2

[50] Guile Wu and Shaogang Gong. Collaborative optimization and aggregation for decentralized domain generalization and adaptation. In *ICCV*, 2021. 2

[51] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, 2021. 3, 4, 6, 7

[52] Fu-En Yang, Yuan-Chia Cheng, Zu-Yun Shiau, and Yu-Chiang Frank Wang. Adversarial teacher-student representation learning for domain generalization. In *NeurIPS*, 2021. 2, 6, 7

[53] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020. 4

[54] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 8

[55] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 8

[56] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *CVPR*, 2022. 2

[57] Jingyi Zhang, Jiaxing Huang, Zichen Tian, and Shijian Lu. Spectral unsupervised domain adaptation for visual recognition. In *CVPR*, 2022. 1

[58] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Mvdg: A unified multi-view framework for domain generalization. In *ECCV*, 2022. 2, 3, 6, 7

[59] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *CVPR*, 2021. 7, 11

[60] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *CVPR*, 2022. 6

[61] Kecheng Zheng, Yang Cao, Kai Zhu, Ruijing Zhao, and Zheng-Jun Zha. Famlp: A frequency-aware mlp-like architecture for domain generalization. *arXiv preprint arXiv:2203.12893*, 2022. 2, 6, 7

[62] Zangwei Zheng, Xiangyu Yue, Kai Wang, and Yang You. Prompt vision transformer for domain generalization. *arXiv preprint arXiv:2208.08914*, 2022. 2

[63] Dawei Zhou, Nannan Wang, Chunlei Peng, Yi Yu, Xi Yang, and Xinbo Gao. Towards multi-domain face synthesis via domain-invariant representations and multi-level feature parts. *IEEE TMM*, 2021. 2

[64] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, 2020. 2, 3, 5, 6, 7, 11, 13

[65] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, 2020. 2

[66] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2020. 5, 8

[67] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE TIP*, 2021. 2

[68] Wei Zhu, Le Lu, Jing Xiao, Mei Han, Jiebo Luo, and Adam P Harrison. Localized adversarial domain generalization. In *CVPR*, 2022. 2

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 [35] 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. 8, 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 [17, 59], we set the batch size as 64 for all our experiments.

Table 8. 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. 9. 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 [64] 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 9. 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**

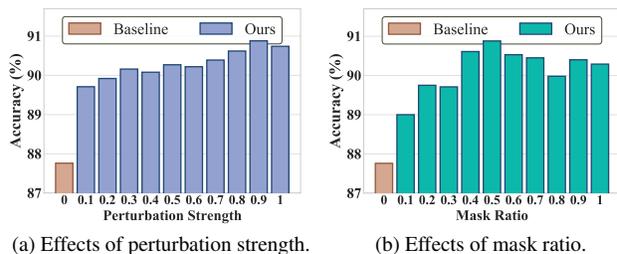(a) Effects of perturbation strength.   (b) Effects of mask ratio.

Figure 5. 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.

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. 5a, where $\alpha = 0$ means the baseline model trained merely with original frequency spectrums. As shown in Fig. 5a, 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. 5b, 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. 10, 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.

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. 10 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 methods 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

Table 10. 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.
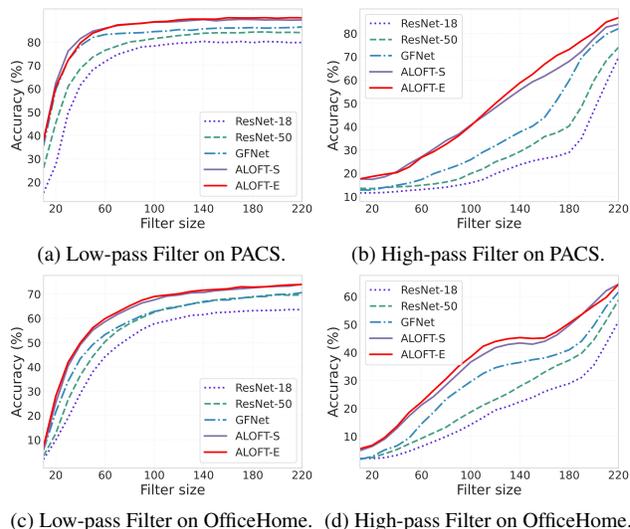
Figure 6. Comparison of ResNet-18, ResNet-50, GFNet, and our ALOFT-S and ALOFT-E on the PACS and OfficeHome datasets. A larger filter size for the low- and high-pass filtering means more low- and high-frequency components, respectively.

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. 6 present the results on PACS and OfficeHome. As shown in Fig. 6a and Fig. 6b, 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

Table 11. 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**

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. 6c and Fig. 6d 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 [35], ALOFT-S, and ALOFT-E. We conduct the experiments on both the PACS and OfficeHome datasets. We calculate the inter-domain distance as below:

$$d = \frac{2}{K(K-1)} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} ||\overline{f}_{k_1} - \overline{f}_{k_2}||_2, \quad (16)$$

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. 11, 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.

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. 12. We observe that most existing MLP-like models

Table 12. 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 13. 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**

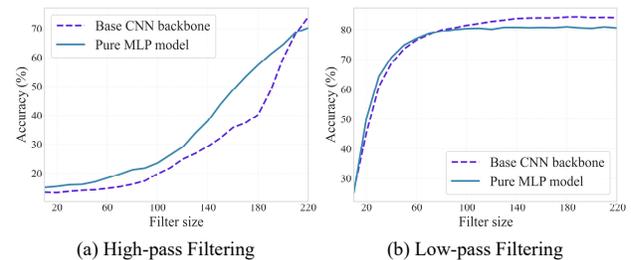(a) High-pass Filtering   (b) Low-pass Filtering

Figure 7. Comparison of the base CNN backbone (*i.e.*, ResNet-18) and the pure MLP backbone (*i.e.*, MLP-mixer [40]) on the PACS dataset. A larger filter size for the low- and high-pass filtering means more low- and high-frequency components, respectively.

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. 13. 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 [64] and the pure MLP model [40]. As shown in Fig. 7, we can observe that the pure MLP model achieves a better performance than the base CNN backbone, which indicates the effectiveness of the MLP model to capture global structure information.

For objects with similar shapes but different tex-

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 8. 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.
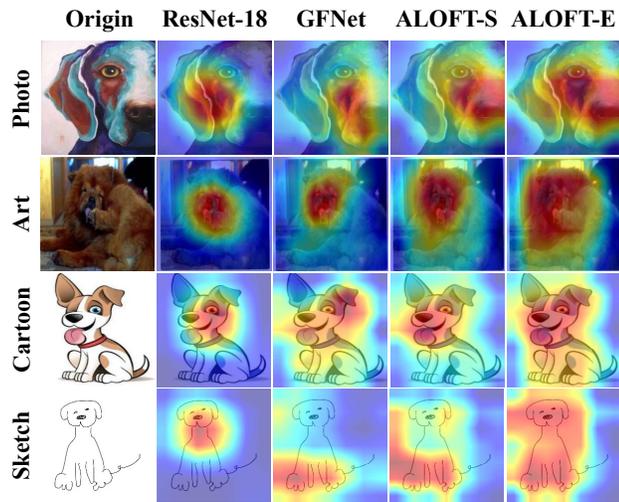

Figure 9. Visualization of attention maps of the last convolutional layer using GradCAM [37] on PACS with Sketch as the target domain. Note that the redder the area indicates the higher attention.

tures. 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. 8, 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 [37]. The results are presented in Fig. 9. 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.