A Fourier-based Framework for Domain Generalization

Qinwei Xu[1] Ruipeng Zhang[1] Ya Zhang[1,2] ✉ Yanfeng Wang[1,2] Qi Tian[3]

[1] Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[2] Shanghai AI Laboratory [3] Huawei Cloud & AI

{qinweixu, zhangrp, ya_zhang, wangyanfeng}@sjtu.edu.cn, tian.qi1@huawei.com

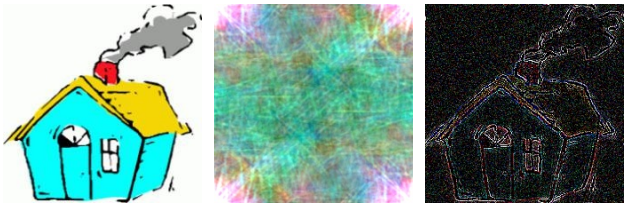
Abstract

Modern deep neural networks suffer from performance degradation when evaluated on testing data under different distributions from training data. Domain generalization aims at tackling this problem by learning transferable knowledge from multiple source domains in order to generalize to unseen target domains. This paper introduces a novel Fourier-based perspective for domain generalization. The main assumption is that the Fourier phase information contains high-level semantics and is not easily affected by domain shifts. To force the model to capture phase information, we develop a novel Fourier-based data augmentation strategy called amplitude mix which linearly interpolates between the amplitude spectrums of two images. A dual-formed consistency loss called co-teacher regularization is further introduced between the predictions induced from original and augmented images. Extensive experiments on three benchmarks have demonstrated that the proposed method is able to achieve state-of-the-arts performance for domain generalization.

1. Introduction

Over the past few years, deep learning have made tremendous progress on various tasks. Under the assumption that training and testing data share the same distribution, deep neural networks (DNNs) have shown great promise for a wide spectrum of applications [18, 10, 12]. However, DNNs have demonstrated quite poor generalizability for out-of-distribution data. Such performance degeneration caused by distributional shift (a.k.a. domain shift) impairs the applications of DNNs, as in reality training and testing data often come from different distributions.

In order to address the problem of domain shift, domain adaptation (DA) bridges the gap between source domain(s) and a specific target domain with the help of some labelled or unlabeled target data. However, DA methods fail to generalize to unknown target domains that have not been seen during training. Collecting data from every possible target domain and training DA models with every source-target pair are expensive and impractical. As a result, a more challenging yet practical problem setting, *i.e.*, domain generalization (DG) [29, 20] is proposed. Unlike DA, DG aims to train model with multiple source domains that can generalize to arbitrary unseen target domains. To tackle the DG problem, many existing methods utilize adversarial training [24, 23, 38], meta learning [21, 1, 25, 4], self-supervised learning [2] or domain augmentation [44, 37, 50, 51] techniques and have shown promising results.



(a) (b) (c)

Figure 1. Examples of the amplitude-only and phase-only reconstruction: (a) original image; (b) reconstructed image with amplitude information only by setting the phase component to a constant; (c) reconstructed image with phase information only by setting the amplitude component to a constant.

In this paper, we introduce a novel Fourier-based perspective for DG. Our motivation comes from a well-known property of the Fourier transformation: the phase component of Fourier spectrum preserves high-level semantics of the original signal, while the amplitude component contains low-level statistics [32, 33, 35, 11]. For better understanding, we present an example of the images reconstructed from only amplitude information and only phase information, as well as the corresponding original image in Fig. 1. As shown in Fig. 1(c), the phase-only reconstruction reveals the important visual structures, from which one can easily recognize the "house" conveyed in the original image. On the other hand, it is hard to tell the exact object from the amplitude-only reconstruction in Fig. 1(b). Based on

these observations, Yang *et al*. [48] have recently developed a Fourier-based method for DA. They propose a simple image translation strategy by replacing the amplitude spectrum of a source image with that of a random target image. By simply training on the amplitude-transferred source images, their method achieves a remarkable performance.

Inspired by the above work, we further explore Fourier-based methods for domain generalization and introduce a *Fourier Augmented Co-Teacher* (FACT) framework, which consists of an implicit constraint induced by Fourier-based data augmentation and an explicit constraint in terms of co-teacher regularization, as shown in Fig. 2. 1) *Fourier-based data augmentation*. Since the phase information is known for carrying the essential features to define an object, it is reasonable to assume that by learning more from the phase information, the model can better extract the semantic concepts of different objects that are robust to domain shifts. However, when dealing with DG, we have no access to the target domain, thus the amplitude transfer strategy as [48] is not applicable. To overcome this, we propose to augment the training data by distorting the amplitude information while keeping the phase information unchanged. Specifically, a linear interpolation strategy similar to MixUp [49] is adopted to generate augmented images. But instead of the whole images, only the amplitude of the images are mixed. Through this Fourier-based data augmentation, our model can avoid overfitting to low-level statistics carried in the amplitude information, thus pay more attention on the phase information when making decisions. 2) *Co-teacher regularization*. In addition to the above implicit constraint induced by Fourier-based data augmentation, we further introduce an explicit constraint to force the model to maintain the predicted class relationships between the original image and its amplitude-perturbed counterpart. This explicit constraint is designed in a form of dual consistency regularization equipped with a momentum-updated teacher [41]. Through the co-teacher regularization, the model is further constrained to focus on the invariant phase information in order to minimize the prediction discrepancy between original and augmented images.

We validate the effectiveness of FACT on three domain generalization benchmarks, namely Digits-DG [50], PACS [20], and OfficeHome [42]. Extensive experimental results have shown that FACT outperforms several state-of-the-arts DG methods with regards to its capability to generalize to unseen domains, indicating that learning more from the phase information does help model generalize better across domains. We further carry out detailed ablation studies to show the superiority of our framework design. We also conduct an in-depth analysis about the rationales behind our hypothesis and method, which demonstrates that the visual structures in phase information contain rich semantics and our model can learn efficiently from them.

2. Related Work

Domain generalization: Domain generalization (DG) aims to extract knowledge from multiple source domains so as to generalize well to arbitrary unseen target domains. Early DG studies mainly follow the distribution alignment idea in domain adaptation by learning domain-invariant features via either kernel methods [29, 9] or domain-adversarial learning [24, 23, 38]. Later on, Li *et al*. [21] propose a meta learning approach that simulates the training strategy of MAML [5] by back propagating the second-order gradients calculated on a random meta-test domain split from the source domains at each iteration. Subsequent meta learning-based DG methods utilize a similar strategy to meta-learn a regularizer [1], a feature-critic network [25], or how to maintain semantic relationships [4]. Another popular way to address DG problem is domain augmentation, which creates samples from fictitious domains via gradient-based adversarial perturbations [44, 37] or adversarially trained image generators [50, 51]. Recently, inspired by the robustness of a shape-biased model to out-of-distributions [8], Shi *et al*. [39] bias their model to shape features by filtering out texture features according to local self-information. Similarly, Carlucci *et al*. [2] introduce a self-supervised jigsaw task to help the model learn global shape features and Wang *et al*. [45] further extend this work by incorporating a momentum metric learning scheme. Other DG methods also employ low-rank decomposition [20, 36] and gradient-guided dropout [16]. Different from all the methods above, our work takes a new Fourier-based perspective for DG. By emphasizing the Fourier phase information, our method achieves a remarkable performance compared with current DG methods.

The importance of phase information: Many early studies [32, 33, 35, 11] have shown that in the Fourier spectrum of signals, the phase component retains most of the high-level semantics in the original signals, while the amplitude component mainly contains low-level statistics. Most recently, Yang *et al*. [48] introduce the Fourier perspective into domain adaptation. By simply replacing a small area in the centralized amplitude spectrum of a source image with that of a target image, they can generate target-like images for training. Another concurrent work of Yang *et al*. [47] proposes to keep a phase consistency in source-target image translations, which is shown to be a better choice than the commonly used cycle consistency [15] for segmentation tasks under DA scenario. Inspired from the above work, we develop a novel Fourier-based framework for DG to encourage our model to focus on the phase information.

Consistency regularization: Consistency regularization (CR) is widely-used in supervised and semi-supervised learning. Laine and Aila [17] first introduce a consistency loss between outputs from two differently perturbed models. Tarvainen and Valpola [41] further extend this work

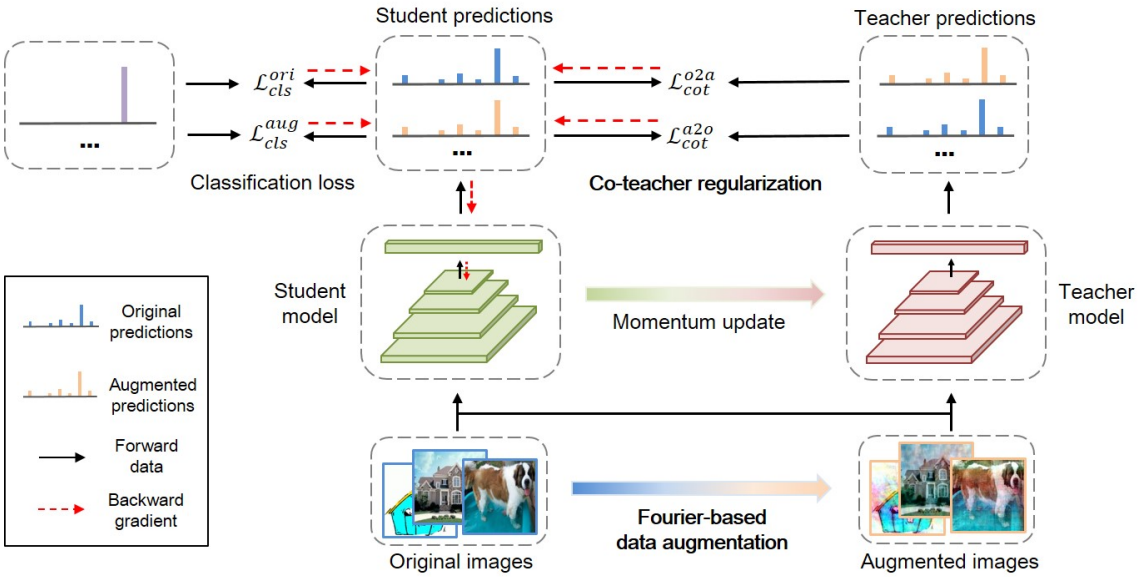


Figure 2. The framework of the proposed FACT. Our framework contains two key components, namely Fourier-based data augmentation and co-teacher regularization, which are highlighted in bold. Both the original and augmented data are sent to the student model and a momentum-updated teacher model. The co-teacher regularization then imposes a dual consistency between the predictions from original data and augmented data. Note that the parameters of teacher model are not updated during back propagation.

by using a momentum-updated teacher to provide better targets for consistency alignment. Miyato *et al*. [27] and Park *et al*. [34] develop different techniques by replacing the stochastic perturbations with adversarial ones. Verma *et al*. [43] also demonstrate that a interpolation consistency with MixUp [49] samples could be helpful. Some recent work [6, 40, 46] also use consistency regularization in UDA to improve the performance on target domain. In our work, we design a dual-formed consistency loss to further bias our model on the phase information.

3. Method

Given a training set of multiple source domains $\mathcal{D}_s = \{\mathcal{D}_1, \ldots, \mathcal{D}_S\}$ with N_k labelled samples $\{(x_i^k, y_i^k)\}_{i=1}^{N_k}$ in the k-th domain $\mathcal{D}_k$, where x_i^k and $y_i^k \in \{1, \ldots, C\}$ denote the inputs and labels respectively, the goal of domain generalization is to learn a domain-agnostic model $f(\cdot;\theta)$ on source domains that is expected to perform well on unseen target domains $\mathcal{D}_t$.

Inspired by the semantic-preserving property of Fourier phase component [32, 33, 35, 11], we assume that models highlight the phase information have better generalization ability across domains. To this end, we design a novel Fourier-based data augmentation strategy by mixing the amplitude information of images. We further add a dual-formed consistency loss, named as co-teacher regularization, to reach consensuses between predictions derived from augmented and original inputs. The consistency loss is implemented with a momentum-updated teacher model to provide better targets for consistency alignment as in [41]. The overall *Fourier-based Augmented Co-Teacher* (FACT) framework is illustrated in Fig. 2. Below we introduce the main components of FACT, *i.e*., Fourier-based data augmentation and co-teacher regularization.

3.1. Fourier-based data augmentation

For a single channel image x, its Fourier transformation $\mathcal{F}(x)$ is formulated as:

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

and $\mathcal{F}^{-1}(x)$ defines the inverse Fourier transformation accordingly. Both the Fourier transformation and its inverse can be calculated with the FFT algorithm [31] efficiently. The amplitude and phase components are then respectively expressed as:

$$\begin{aligned} \mathcal{A}(x)(u,v) &= \left[R^2(x)(u,v) + I^2(x)(u,v)\right]^{1/2} \\ \mathcal{P}(x)(u,v) &= \arctan\left[\frac{I(x)(u,v)}{R(x)(u,v)}\right], \end{aligned} \quad (2)$$

where $R(x)$ and $I(x)$ represent the real and imaginary part of $\mathcal{F}(x)$, respectively. For RGB images, the Fourier transformation for each channel is computed independently to get the corresponding amplitude and phase information.

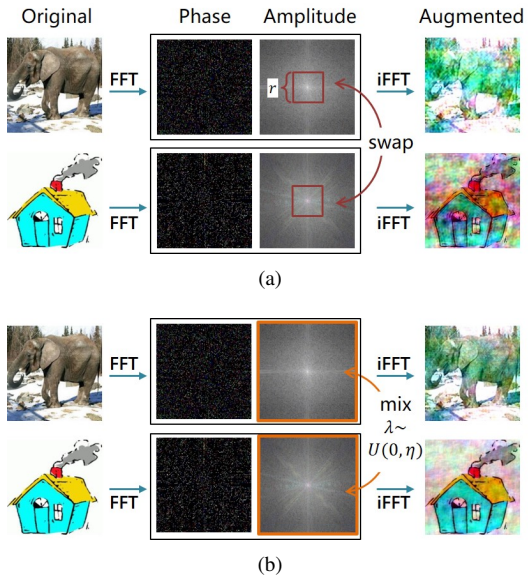


Figure 3. Illustration of (a) AS and (b) AM strategy.

With the semantic-preserving property of Fourier phase component, we here attempt to build models that specifically highlight the phase information, which are expected to have better generalization ability across domains. To achieve this goal, a natural choice is perturbing the amplitude information in the original images via a certain form of data augmentation. Inspired from MixUp [49], we design a novel data augmentation strategy by linearly interpolating between the amplitude spectrums of two images from arbitrary source domains:

$$\hat{\mathcal{A}}(x_i^k) = (1-\lambda)\mathcal{A}(x_i^k) + \lambda\mathcal{A}(x_{i'}^{k'}), \tag{3}$$

where $\lambda \sim U(0,\eta)$, and the hyperparameter η controls the strength of the augmentation. The mixed amplitude spectrum is then combined with the original phase spectrum to form a new Fourier representation:

$$\mathcal{F}(\hat{x}_i^k)(u,v) = \hat{\mathcal{A}}(x_i^k)(u,v) * e^{-j*\mathcal{P}(x_i^k)(u,v)}, \tag{4}$$

which is then fed to inverse Fourier transformation to generate the augmented image. $\hat{x}_i^k = \mathcal{F}^{-1}[\mathcal{F}(\hat{x}_i^k)(u,v)]$. This Fourier-based augmentation strategy, named *amplitude mix* (AM) thereafter, is illustrated in Fig. 3 (b).

We then feed the augmented images and the original labels to the model for classification. The loss function is formulated as standard cross entropy[1]:

$$\mathcal{L}_{cls}^{aug} = -y_i^k \log\left(\sigma(f(\hat{x}_i^k;\theta))\right) \tag{5}$$

where σ is the softmax activation. We also use the original images for training, the classification loss $\mathcal{L}_{cls}^{ori}$ can then be defined similarly as Eq. 5.

[1]The expectation is omitted for brevity for all the loss functions.

Note that the AM strategy is essentially different from the spectral transfer operation proposed in [48]. Specifically, the spectral transfer operation aims to adapt low-level statistics from the source domain to the target domain by replacing the low-frequency amplitude information of source images with that of target images. However, in domain generalization, since we have no access to target data, such an adaptive operation is impossible. Nevertheless, we can still swap the amplitude spectrums between two source images directly to create augmented images. This *amplitude swap* (AS) strategy is illustrate in Fig. 3 (a). Like [48], the proportion of the swapped area is controlled by a hyperparameter r. However, only swapping a small area of the centralized amplitude spectrum (*i.e.* low-frequency amplitude information) could still cause the model to overfit on the remaining middle-frequency and high-frequency amplitude information, while swapping the whole amplitude spectrum (*i.e.* setting $r = 1$) may be too aggressive for the model to learn. On the other hand, the AM strategy perturbs each frequency component in the amplitude spectrum equally and bridge the model to the most aggressively augmented images via linear interpolation. Therefore, the model can efficiently learn from the phase information by comparing between the original and augmented images through AM augmentation.

3.2. Co-teacher Regularization

The above Fourier-based data augmentation imposes an implicit constraint which requires the model to predict the same object before and after augmentation. However, the categorical relations predicted from original and augmented images with the same phase information may be different. For example, a model may learn from the original images that horses are more similar to giraffes than houses. However, this learned knowledge may conflict to that learned from augmented images due to a different augmentation. To alleviate this kind of disagreement, we add an explicit constraint in the form of a dual consistency loss. As suggested in [41], we use a momentum-updated teacher model to provide better targets for consistency alignment. During training, the teacher model receives parameters from the student model via exponential moving average:

$$\theta_{tea} = m\theta_{tea} + (1-m)\theta_{stu} \tag{6}$$

where m is the momentum parameter. Note that no gradient flows through the teacher model during back propagation. To make full use of the knowledge learned from data, we compute the outputs with a softened softmax at a temperature T [14] for both the teacher and student models. We then force the model to be consistent between the outputs derived from original images and augmented images:

$$\mathcal{L}_{cot}^{a2o} = \mathrm{KL}(\sigma(f_{stu}(\hat{x}_i^k)/T)||\sigma(f_{tea}(x_i^k)/T)), \tag{7}$$

$$\mathcal{L}_{cot}^{o2a} = \mathrm{KL}(\sigma(f_{stu}(x_i^k)/T)||\sigma(f_{tea}(\hat{x}_i^k)/T)). \tag{8}$$

Here we align the student outputs of augmented images to the teacher outputs of original images, as well as the student outputs of original images to the teacher outputs of augmented images. Since the consistency loss takes a dual form and incorporates a momentum teacher, we rename the loss as *co-teacher regularization* for brevity. Through the co-teacher regularization, we want our model to learn equally from the original and augmented images. More specifically, the original image and its augmented counterpart can be seen as two views of a same object. When learning from the "original view", the student model is not only guided by the ground-truth, but also taught by the teacher model that learns from the "augmented view". So is the case when the student model learns from the "augmented view". Such a simultaneous co-teaching process enables a comprehensive knowledge sharing between the original and augmented view, and further direct the model to focus on the invariant phase information in order to reach a consistency between the two views.

Combining all the loss functions together, we can get our full objective as:

$$\mathcal{L}_{FACT} = \mathcal{L}_{cls}^{ori} + \mathcal{L}_{cls}^{aug} + \beta(\mathcal{L}_{cot}^{a2o} + \mathcal{L}_{cot}^{o2a}) \qquad (9)$$

where β controls the trade-off between the classification loss and the co-teacher regularization loss.

4. Experiment

In this section, we demonstrate the superiority of our method on several DG benchmarks. We also carry out detailed ablation studies about the impact of different components and augmentation types.

4.1. Datasets and settings

We evaluate our method on three conventional DG benchmarks, which cover various recognition scenes. Details of these benchmarks are as follows: (1) **Digits-DG** [50]: a digit recognition benchmark consisted of four classical datasets *MNIST* [19], *MNIST-M* [7], *SVHN* [30], SYN [7]. The four datasets mainly differ in font style, background and image quality. We use the original train-validation split in [50] with 600 images per class per dataset. (2) **PACS** [20]: an object recognition benchmark including four domains (*photo*, *art-painting*, *cartoon*, *sketch*) with large discrepancy in image styles. It contains seven classes and 9,991 images totally. We use the original train-validation split provided by Li *et al*. [20]. (3) **Office-Home** [42]: an object recognition benchmark including 15,500 images of 65 classes from four domains (*Art*, *Clipart*, *Product*, *Real-World*). The domain shift mainly comes from image styles and viewpoints, but is much smaller than PACS. Follow [2], we randomly split each domain into 90% for training and 10% for validation.

Table 1. Leave-one-domain-out results on Digits-DG. The best and second-best results are bolded and underlined respectively.

Methods	MNIST	MNIST-M	SVHN	SYN	Avg.
DeepAll [50]	95.8	58.8	61.7	78.6	73.7
CCSA [28]	95.2	58.2	65.5	79.1	74.5
MMD-AAE [23]	96.5	58.4	65.0	78.4	74.6
CrossGrad [37]	96.7	61.1	65.3	80.2	75.8
DDAIG [50]	96.6	<u>64.1</u>	68.6	81.0	77.6
Jigen [2]	96.5	61.4	63.7	74.0	73.9
L2A-OT [51]	<u>96.7</u>	63.9	<u>68.6</u>	<u>83.2</u>	<u>78.1</u>
FACT (*ours*)	**97.9**	**65.6**	**72.4**	**90.3**	**81.5**

For all benchmarks, we conduct the leave-one-domain-out evaluation. We train our model on the training splits and select the best model on the validation splits of all source domains. For testing, we evaluate the selected model on all images of the held-out target domain. All the results are reported in terms of classification accuracy and averaged over three runs. We employ a vanilla ConvNet trained from a simple aggregation of all source data as our baseline, which is named as DeepAll in the remaining sections.

4.2. Implementation details

We closely follow the implementations of [2, 50]. Here we briefly introduce the implementation details for training our model, and more details can be found in the supplementary. The source code is released at https://github.com/MediaBrain-SJTU/FACT.

Basic details: For Digits-DG, we use the same backbone network as [50, 51]. We train the network from scratch using SGD, batch size of 128 and weight decay of 5e-4 for 50 epochs. The initial learning rate is set to 0.05 and decayed by 0.1 every 20 epochs. For PACS and OfficeHome, we use the ImageNet pretrained ResNet [12] as our backbone. We train the network with SGD, batch size of 16 and weight decay of 5e-4 for 50 epochs. The initial learning rate is 0.001 and decayed by 0.1 at 80% of the total epochs. We also use the standard augmentation protocol as in [2], , which consists of random resized cropping, horizontal flipping and color jittering.

Method-specific details: For all experiments, we set the momentum m for the teacher model to 0.9995 and the temperature T to 10. The weight β of the consistency loss is set to 2 for Digits-DG and PACS, and 200 for OfficeHome. We also use a sigmoid ramp-up [41] for β with a length of 5 epochs. The augmentation strength η of AM is chosen as 1.0 for Digits-DG and PACS, and 0.2 for OfficeHome.

4.3. Evaluation on Digits-DG

We present the results in Table 1. Among all the competitors, our method achieves the best performance, exceeding the second best method L2A-OT [51] by more than

Table 2. Leave-one-domain-out results on PACS. The best and second-best results are bolded and underlined respectively. †: results are reported based on the best models on test splits.

Methods	Art	Cartoon	Photo	Sketch	Avg.
ResNet18					
DeepAll	77.63	76.77	95.85	69.50	79.94
MetaReg [1]	83.70	77.20	95.50	70.30	81.70
JiGen [2]	79.42	75.25	96.03	71.35	80.51
Epi-FCR [22]	82.10	77.00	93.90	73.00	81.50
MMLD [26]	81.28	77.16	96.09	72.29	81.83
DDAIG [50]	84.20	78.10	95.30	74.70	83.10
CSD [36]	78.90	75.80	94.10	76.70	81.40
InfoDrop [39]	80.27	76.54	96.11	76.38	82.33
MASF [4]†	80.29	77.17	94.99	71.69	81.04
L2A-OT [51]	83.30	78.20	**96.20**	73.60	82.80
EISNet [45]	81.89	76.44	95.93	74.33	82.15
RSC [16]	83.43	80.31	95.99	80.85	85.15
RSC (*our imple.*)	80.55	**78.60**	94.43	76.02	82.40
FACT (*ours*)	**85.37**	78.38	95.15	**79.15**	**84.51**
ResNet50					
DeepAll	84.94	76.98	**97.64**	76.75	84.08
MetaReg [1]	87.20	79.20	97.60	70.30	83.60
MASF [4]†	82.89	80.49	95.01	72.29	82.67
EISNet [45]	86.64	81.53	97.11	78.07	85.84
RSC [16]	87.89	82.16	97.92	83.35	87.83
RSC (*our imple.*)	83.92	79.52	95.15	82.20	85.20
FACT (*ours*)	**89.63**	**81.77**	96.75	**84.46**	**88.15**

3% on average. Specifically, on the hardest target domains SVHN and SYN, where involve cluttered digits and low image qualities, our method outperforms L2A-OT with a large margin of 4% and 7% respectively. The success of our method indicates that training the model to focus more on the spectral phase information can significantly promote its performance on out-of-domain images.

4.4. Evaluations on PACS

The results are shown in Table 2. It is clearly that our method is among the top performing ones. We notice that the naive DeepAll baseline can get a remarkable accuracy on the photo domain, due to its similarity to the pretrained dataset ImageNet. However, DeepAll performs poorly on the art-painting, cartoon and sketch domains, which bear a larger domain discrepancy. Nevertheless, our FACT can lift the performance of DeepAll by a huge margin of 7.52% on art-painting, 3.52% on cartoon and 11.41% on sketch. Meanwhile, the performance of our model on photo domain drops a little. This is reasonable since domains like photo contain complicated and redundant details, and the model may ignore some possibly useful low-level cues by only highlighting the phase information. Nevertheless, our model still achieves a better overall performance.

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

Methods	Art	Clipart	Product	Real	Avg.
DeepAll	57.88	52.72	73.50	74.80	64.72
CCSA [28]	59.90	49.90	74.10	75.70	64.90
MMD-AAE [23]	56.50	47.30	72.10	74.80	62.70
CrossGrad [37]	58.40	49.40	73.90	75.80	64.40
DDAIG [50]	59.20	52.30	74.60	76.00	65.50
L2A-OT [51]	**60.60**	50.10	**74.80**	**77.00**	65.60
Jigen [2]	53.04	47.51	71.47	72.79	61.20
RSC [16]	58.42	47.90	71.63	74.54	63.12
Jigen (*our imple.*)	57.95	49.21	72.61	74.90	63.67
RSC (*our imple.*)	57.67	48.48	72.62	74.16	63.23
FACT (*ours*)	60.34	**54.85**	74.48	76.55	**66.56**

Compared with the SOTA, our FACT clearly beats the methods based on adversarial data augmentation or meta learning, including the latest MASF [4], DDAIG [50] and L2A-OT [51], yet FACT enjoys an efficient training process without any additional adversarial or episodic training steps. The performance of our method also exceeds that of RSC [16], a recent proposed method utilizing a simple yet powerful gradient-guided dropout, by 2.11% on average[2]. All the above comparisons reveal the effectiveness of our method and further demonstrate that the emphasis on phase information improves generalizability across domains.

4.5. Evaluations on OfficeHome

We report the results in Table 3. Due to a relatively smaller domain discrepancy and the similarity to the pretrained dataset ImageNet, DeepAll acts as a strong baseline on OfficeHome. Many previous DG methods, such as CCSA [28], MMD-AAE [23], CrossGrad [37] and Jigen [2], can not improve much over the simple DeepAll baseline. Nevertheless, our FACT achieves a consistent improvement over DeepAll on all the held-out domains. Moreover, FACT also the surpasses the latest DDAIG [50] and L2A-OT [50] in terms of average performance. This again justifies the superiority of our method.

4.6. Ablation studies

Impact of different components: We conduct an extensive ablation study to investigate the role of each component in our FACT model in Table 4. Starting from baseline, model A is trained with the AM augmentation only and already works better than the strongest competitor DDAIG [50] in Table 2. Based on model A, we add

[2]We rerun the source codes of RSC under the same hyperparameters, but we are unable to reproduce the reported results in original paper [16]. We conjecture this may attribute to the difference in hardware environment. For fairness, we compare our method and RSC under our environment.

Table 4. Ablation studies on different components of our method on the PACS dataset with ResNet18.

Method	AM	$\mathcal{L}_{cot}^{a2o}$	$\mathcal{L}_{cot}^{o2a}$	Teacher	Art	Cartoon	Photo	Sketch	Avg.
Baseline	-	-	-	-	77.63±0.84	76.77±0.33	**95.85±0.20**	69.50±1.26	79.94
Model A	✓	-	-	-	83.90±0.50	76.95±0.45	95.55±0.12	77.36±0.71	83.44
Model B	✓	✓	✓	-	83.71±0.30	77.84±0.49	94.73±0.12	78.55±0.46	83.71
Model C	-	✓	✓	✓	82.68±0.44	78.06±0.39	95.35±0.44	74.76±0.67	82.71
Model D	✓	✓	-	✓	83.97±0.77	77.04±0.86	94.59±0.03	79.08±0.56	83.67
Model E	✓	-	✓	✓	84.07±0.43	77.70±0.65	95.28±0.34	78.29±0.61	83.84
FACT	✓	✓	✓	✓	**85.37±0.29**	**78.38±0.29**	95.15±0.26	**79.15±0.69**	**84.51**

a vanilla dual-formed consistency loss to obtain model B, which improves over model A slightly. Further incorporating the momentum teacher results in our FACT, which performs best in all variants. This indicates the importance of the momentum teacher that provides better targets for consistency loss. We also create a model C by excluding the AM augmentation from FACT and the performance drops a lot, showing that the Fourier-based data augmentation plays a crucial role. We further validate the necessity of the dual form in co-teacher regularization by using only $\mathcal{L}_{cot}^{a2o}$ or $\mathcal{L}_{cot}^{o2a}$, and resulting in model D and E respectively. As in Table 4, neither model D or E outperforms the full FACT, suggesting the effectiveness of incorporating both original-to-augmented and augmented-to-original consistency alignment through co-teacher regularization.

Other choices of Fourier-based data augmentation: Next we show the advantages of our AM augmentation over its alternatives in Table 5. Specifically, we compare the AM strategy with the AS strategy which is mentioned in Sec. 3.1. Follow [48], we first choose to swap only a small area in the centralized amplitude spectrum by setting ratio $r = 0.09$, and the resulting strategy is called AS-partial. As in Table 5, the performance of AS-partial is inferior to that of AM. This is reasonable, as choosing a small value of r in AS only perturbs the low-frequency components in the amplitude spectrum, while the model still has risks to overfit on the remaining amplitude information. Nevertheless, we can still perturb all frequency components with AS by setting $r = 1.0$, and the resulting strategy is called AS-full. This choice brings some improvement but is still worse than AM. We attribute this to the negative effect caused by directly swapping the entire amplitude spectrums of two images, which may be too aggressive for the model to learn.

5. Discussion

Phase information contains meaningful semantics and helps generalization. In previous sections, we have seen that by learning more from the phase information, the model can generalize well on unseen domains. Here we further verify the importance of phase information through single domain evaluations on PACS. Specifically, we first generate phase-only reconstructed images by setting the amplitude component as a constant, and so is the amplitude-only reconstructed images. Then we train three models on original images, phase-only images and amplitude-only images respectively, and compare their performance in Table 6. For fair comparison, all the models here are trained from scratch without ImageNet pretraining. As shown in Table 6, the model trained with phase-only images performs better, or at least comparable with the baseline trained with original images in 11 out of 16 cases. This indicates that the phase information does contain useful semantics to help the

Table 5. Ablation studies of different choices of the Fourier data augmentation on the PACS dataset with ResNet18.

Methods	Art	Cartoon	Photo	Sketch	Avg.
DeepAll with					
AS-partial	82.00	76.19	93.89	77.27	82.34
AS-full	83.50	76.07	94.49	77.13	82.80
AM	**83.90**	**76.95**	**95.55**	**77.36**	**83.44**
FACT with					
AS-partial	81.61	76.95	93.83	78.30	82.67
AS-full	83.46	77.37	94.10	78.63	83.39
AM	**85.37**	**78.38**	**95.15**	**79.15**	**84.51**

Table 6. The performance changes of training with phase-only reconstructed images and amplitude-only reconstructed images when compared with original images. The values greater than zero (meaning an improvement) are in bold.

Data	Train \ Test	Photo	Art	Cartoon	Sketch
Phase only	Photo	-4.68	**3.16**	**4.07**	**2.38**
	Art	-5.35	**1.28**	**5.97**	**15.87**
	Cartoon	-11.53	**0.29**	-4.08	**18.55**
	Sketch	**10.66**	**14.56**	**21.26**	-1.09
Amplitude only	Photo	-14.03	-4.15	-4.41	-0.08
	Art	-18.40	-21.96	-5.59	-10.72
	Cartoon	-13.95	-7.48	-15.89	**1.36**
	Sketch	-4.79	-0.73	-1.99	-13.99

model generalize to unseen domains. On the other hand, the model trained with amplitude-only images suffers from large performance degradation in almost all the cases, suggesting that the amplitude information hardly contains any meaningful semantics. Another interesting finding is that the model trained with phase-only images has some performance drops when generalizing to photo domain from cartoon and art domain. We conjecture that a desirable performance on the photo domain may also require the presence of amplitude information. Therefore, in our FACT framework, we do not completely eliminate the amplitude information, but instead shift model's attention to the phase information in an amplitude-perturbation fashion. It worths nothing that the performance of the model trained with phase-only images also declines in some in-domain generalization cases. This is reasonable considering the loss of amplitude information.

Amplitude perturbation constrains the model to focus more on phase information. Our Fourier-based data augmentation are implemented via perturbing the amplitude information. Here we present a brief theoretical analysis to demonstrate that amplitude perturbation does make the model to focus more on phase information. For simplicity, we consider the case of a linear softmax classifier together with a feature extractor $\mathbf{h}$. Suppose the distribution of Fourier-based data augmentation is $g \sim \mathcal{G}$, and the risk of training on the augmented data is:

$$\hat{R}_{\text{aug}} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{g\sim\mathcal{G}}\left[\ell\left(\mathbf{W}^{\top}\mathbf{h}\left(g\left(x\right)\right), y\right)\right] \tag{10}$$

Similar as in [3, 13], we expand $\hat{R}_{\text{aug}}$ with Taylor expansion:

$$\begin{aligned}\mathbb{E}_{g\sim\mathcal{G}}\left[\ell\left(\mathbf{W}^{\top}\mathbf{h}(g(x)), y\right)\right] = \\ \ell\left(\mathbf{W}^{\top}\overline{\mathbf{h}}, y\right) + \tfrac{1}{2}\mathbb{E}_{g\sim\mathcal{G}}\left[\Delta^{\top}\mathbf{H}(\tau, y)\Delta\right]\end{aligned} \tag{11}$$

where $\overline{\mathbf{h}} = \mathbb{E}_{g\sim\mathcal{G}}[\mathbf{h}(g(x))]$, $\Delta = \mathbf{W}^{\top}(\overline{\mathbf{h}} - \mathbf{h}(g(x)))$ and $\mathbf{H}$ is the Hessian matrix. For cross-entropy loss with softmax, $\mathbf{H}$ is semi-definite and independent of y. Then, minimizing the second-order term in Eq. 11 requires that for some feature h_d, if its variance $h_d(g(x))$ is large, the weight $w_{i,d}$ will approach 0. Suppose that the features induced from phase information and amplitude information is h_p and h_a respectively, since we only perturb the amplitude information and keep the phase information unchanged, it is reasonable that:

$$\begin{cases} |h_p(g(x)) - h_p(x)| < \zeta \\ |h_a(g(x)) - h_a(x)| > \epsilon \end{cases} \tag{12}$$

where $\zeta > 0$ is a small value, and $\epsilon \geq \zeta$. Therefore, minimizing $\hat{R}_{\text{aug}}$ restricts $w_{i,a} \to 0$ for those features h_a derived from the amplitude information. As a result, the classifier would pay more attention to the features h_p that derived from the phase information when making decisions.

Table 7. The cosine similarities between the phase-only reconstructed images (P) and the edge images detected by Sobel operator (S) and Laplacian operator (L).

	S&P	L&P	S&L
Similiarity	0.790	0.914	0.873

Why does the phase information provides model with meaningful semantics? The answer may be that the phase information records the "location" of events [33] (or small local structures) and reveals the spatial relationships between them within a given image [11]. The model can then aggregate these cues to gain a correct knowledge about the objects conveyed in the image. A similar mechanism can also be found in human vision systems [11]. We also notice that the phase-only reconstructed image in Fig. 1(c) mainly preserves the contours and edges of the original image. For further investigation, we compute the cosine similarities between phase-only images and edge-detected images produced by Sobel or Laplacian operators in Table 7. As we can seen, the similarity scores between phase-only images and edge-detected images are very high, implying that the visual structures such as edges and contours are carried in the phase information. Since the visual structures are the keys to describe different objects, regardless of the underlying domain distributions, learning from such information can facilitate model to extract high-level semantics.

6. Conclusions

In this paper, we introduce a novel perspective based on Fourier transformation into domain generalization. The main idea is that learning more from the spectral phase information can help the model capture domain-invariant semantic concepts. We then propose a framework composed of an implicit constraint induced by Fourier-based data augmentation and an explicit constraint induced by co-teacher regularization. Extensive experiments on three benchmarks demonstrate that our method is able to achieve state-of-the-art performance for domain generalization. Furthermore, we conduct an in-depth analysis about the mechanisms and rationales behind our method, which gives us a better knowledge about why focusing on the phase information can help domain generalization. Considering the mainstream of related work is still domain-adversarial learning or meta learning, we hope our work can shed some lights into the community.

Acknowledgement: This work is supported by the National Key Research and Development Program of China (No. 2019YFB1804304), SHEITC (No. 2018-RGZN-02046), 111 plan (No. BP0719010), and STCSM (No. 18DZ2270700), and State Key Laboratory of UHD Video and Audio Production and Presentation.

References

[1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, pages 998–1008, 2018. 1, 2, 6

[2] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 1, 2, 5, 6

[3] Tri Dao, A. Gu, Alexander J. Ratner, V. Smith, C. D. Sa, and C. Ré. A kernel theory of modern data augmentation. *Proceedings of machine learning research*, 97:1528–1537, 2019. 8

[4] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, pages 6450–6461, 2019. 1, 2, 6

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 2

[6] Geoff French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018. 3

[7] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 5

[8] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 2

[9] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016. 2

[10] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 1

[11] Bruce C Hansen and Robert F Hess. Structural sparseness and spatial phase alignment in natural scenes. *JOSA A*, 24(7):1873–1885, 2007. 1, 2, 3, 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5

[13] Zhuoxun He, Lingxi Xie, X. Chen, Y. Zhang, Yanfeng Wang, and Q. Tian. Data augmentation revisited: Rethinking the distribution gap between clean and augmented data. *ArXiv*, abs/1909.09148, 2019. 8

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018. 2

[16] Zeyi Huang, Haohan Wang, E. Xing, and D. Huang. Self-challenging improves cross-domain generalization. In *ECCV*, 2020. 2, 6

[17] S. Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017. 2

[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[20] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 1, 2, 5

[21] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 1, 2

[22] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1446–1455, 2019. 6

[23] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 1, 2, 5, 6

[24] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 1, 2

[25] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. volume 97 of *Proceedings of Machine Learning Research*, pages 3915–3924, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 1, 2

[26] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, pages 11749–11756, 2020. 6

[27] Takeru Miyato, S. Maeda, Masanori Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1979–1993, 2019. 3

[28] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5715–5725, 2017. 5, 6

[29] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013. 1, 2

[30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[31] Henri J Nussbaumer. The fast fourier transform. In *Fast Fourier Transform and Convolution Algorithms*, pages 80–111. Springer, 1981. 3

[32] A Oppenheim, Jae Lim, Gary Kopec, and SC Pohlig. Phase in speech and pictures. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 632–637. IEEE, 1979. 1, 2, 3

[33] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. 1, 2, 3, 8

[34] Sungrae Park, JunKeon Park, Su-Jin Shin, and Il-Chul Moon. Adversarial dropout for supervised and semi-supervised learning. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 3917–3924. AAAI press, 2018. 3

[35] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346, 1982. 1, 2, 3

[36] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International Conference on Machine Learning*, 2020. 2, 6

[37] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018. 1, 2, 5, 6

[38] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10023–10031, 2019. 1, 2

[39] Baifeng Shi, Dinghuai Zhang, Qi Dai, Zhanxing Zhu, Y. Mu, and J. Wang. Informative dropout for robust representation learning: A shape-bias perspective. In *International Conference on Machine Learning*, 2020. 2, 6

[40] Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations*, 2018. 3

[41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017. 2, 3, 4, 5

[42] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 2, 5

[43] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3635–3641. International Joint Conferences on Artificial Intelligence Organization, 7 2019. 3

[44] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in neural information processing systems*, pages 5334–5344, 2018. 1, 2

[45] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsicsupervisions for domain generalization. In *ECCV*, 2020. 2, 6

[46] Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Dual mixup regularized learning for adversarial domain adaptation. In *ECCV*, 2020. 3

[47] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9011–9020, 2020. 2

[48] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 2, 4, 7

[49] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 2, 3, 4

[50] Kaiyang Zhou, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, pages 13025–13032, 2020. 1, 2, 5, 6

[51] K. Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, 2020. 1, 2, 5, 6