CVPR 2021

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

Presented by Qing Lyu on June 16

DA vs DG

- Data domain shift impairs the performance of networks

- Domain adaptation
  - Bridges the gaps between source domains and a specific target domain with labelled or unlabeled target data
  - There are target domain data in training

- Domain generalization
  - Aims to train model with multiple source domains that can generalize to arbitrary unseen target domains
  - No target domain data in training

Fourier phase & amplitude information

- Phase: high-level semantics
- Amplitude: low-level statistics

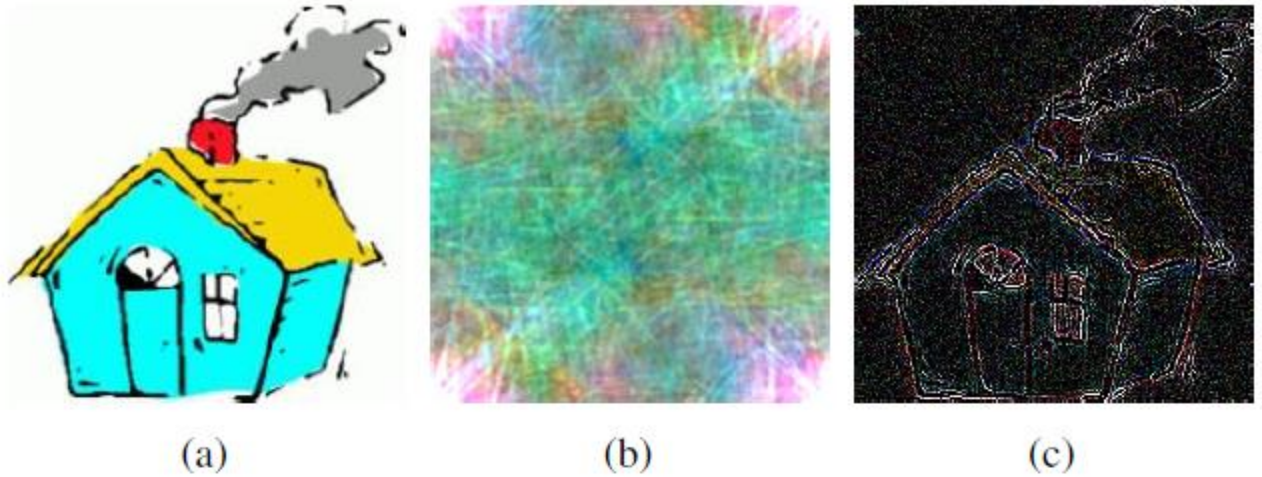


Figure 1. Examples of the amplitude-only and phase-only reconstruction: (a) original image; (b) reconstructed image with amplitude information only by setting the phase component to a constant; (c) reconstructed image with phase information only by setting the amplitude component to a constant.

Motivation

- Changing amplitude information for data augmentation
- Using augmented data for implementing data generalization

Contribution

- Introduce a novel Fourier-based perspective for domain generalization,
- Develop a novel Fourier-based data augmentation strategy called amplitude mix,
- Utilize a dual-formed consistency loss called co-teacher regularization during training process.

Fourier-based data augmentation

- Amplitude swap
  - Swap central part of amplitude images

- Amplitude mix
  - Linearly interpolate amplitude images

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

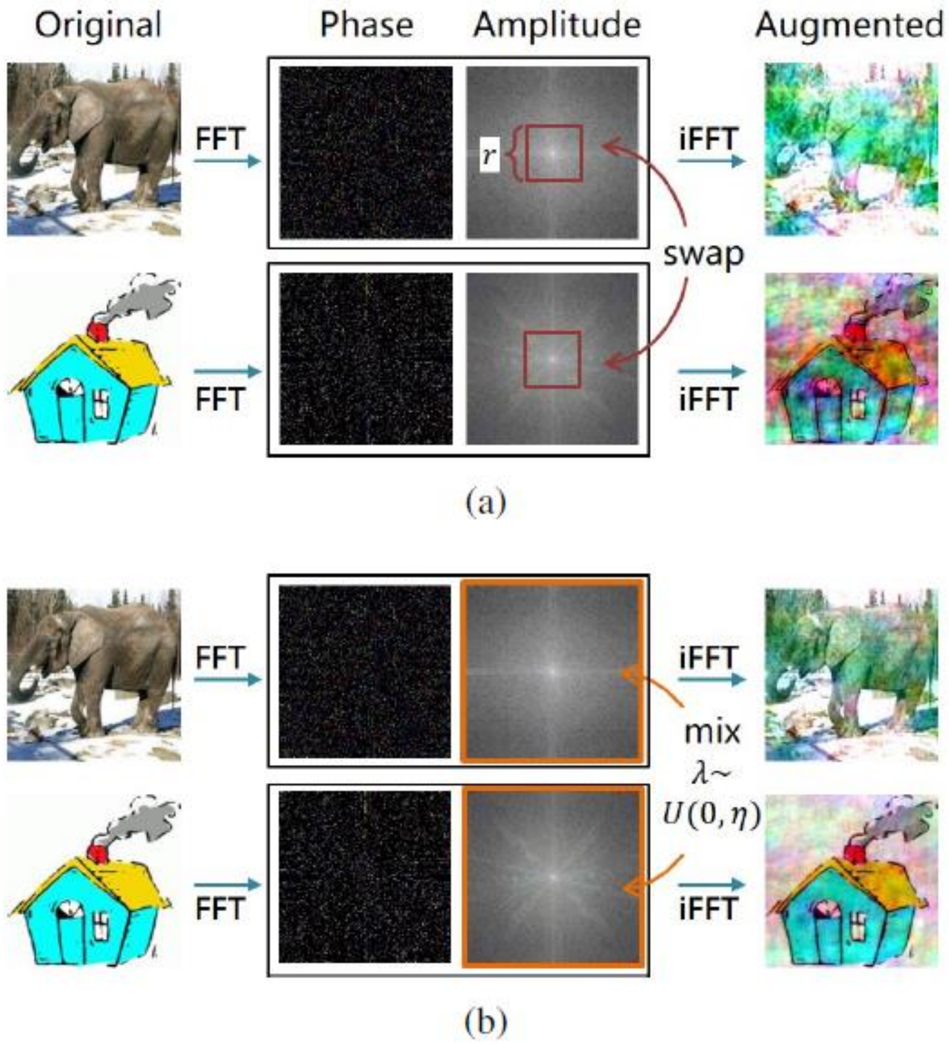
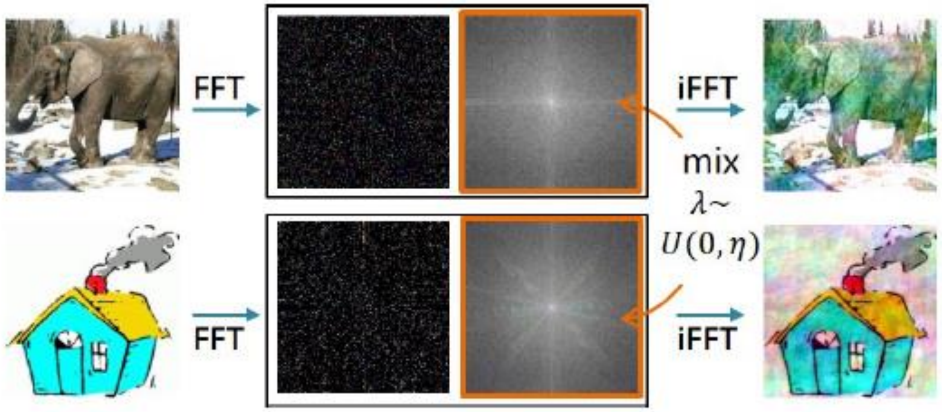


Figure 3. Illustration of (a) AS and (b) AM strategy.

Co-teacher regularization

- The categorical relations predicted from original and augmented images with the same phase information may be different



- Co-teacher regularization is used for alleviate this disagreement

Co-teacher regularization

- Dual consistency loss

- Momentum-updated teacher model
  - Teacher model receives parameters from the student model via exponential moving average

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

Objective function

- Classification loss
  - Cross-entropy

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

- Co-teacher regularization
  - Softened softmax at temperature T

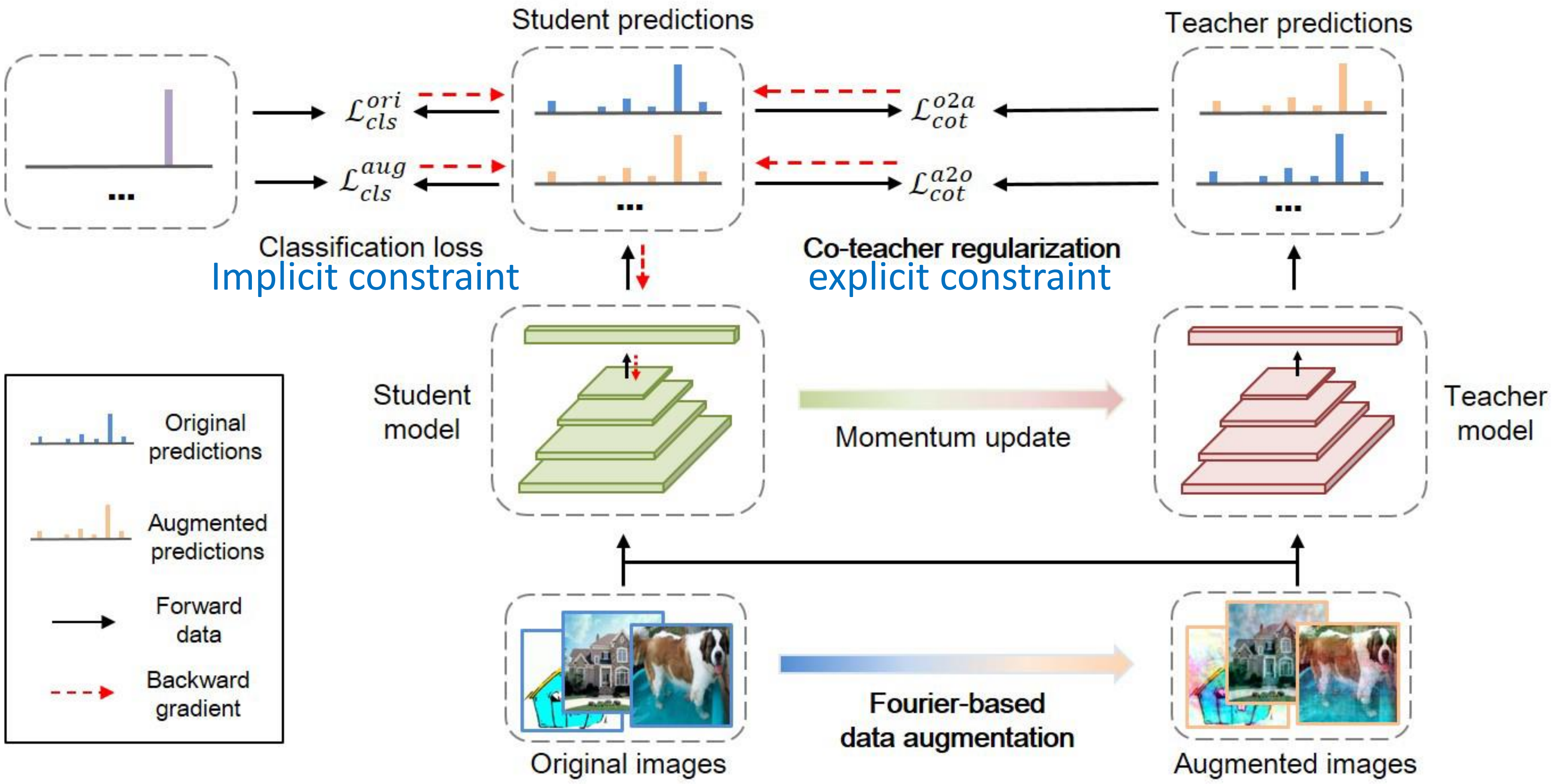
$$\mathcal{L}_{cot}^{a2o} = \mathrm{KL}(\sigma(f_{stu}(\hat{x}_i^k)/T)||\sigma(f_{tea}(x_i^k)/T))$$

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

- Overall loss

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

Fourier Augmented Co-Teacher (FACT)



Dataset

- Digits-DG
  - Four datasets
  - MNIST, MNIST-M, SVHN, SYN

- PACS
  - Four domains
  - Photo, art-painting, cartoon, sketch

- Office-Home
  - Four domains
  - Art, clipart, product, real-world

Training strategy

- Leave-one-domain-out
  - Train on data in all domain but one
  - Test on the held-out domain

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 [52]	95.8	58.8	61.7	78.6	73.7
CCSA [29]	95.2	58.2	65.5	79.1	74.5
MMD-AAE [24]	96.5	58.4	65.0	78.4	74.6
CrossGrad [38]	96.7	61.1	65.3	80.2	75.8
DDAIG [52]	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 [53]	<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**

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 [23]	82.10	77.00	93.90	73.00	81.50
MMLD [27]	81.28	77.16	96.09	72.29	81.83
DDAIG [52]	<u>84.20</u>	78.10	95.30	74.70	<u>83.10</u>
CSD [37]	78.90	75.80	94.10	<u>76.70</u>	81.40
InfoDrop [40]	80.27	76.54	<u>96.11</u>	76.38	82.33
MASF [4]†	80.29	77.17	94.99	71.69	81.04
L2A-OT [53]	83.30	78.20	**96.20**	73.60	82.80
EISNet [46]	81.89	76.44	95.93	74.33	82.15
RSC [17]	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**	<u>78.38</u>	95.15	**79.15**	**84.51**
		ResNet50			
DeepAll	84.94	76.98	**97.64**	76.75	84.08
MetaReg [1]	<u>87.20</u>	79.20	<u>97.60</u>	70.30	83.60
MASF [4]†	82.89	80.49	95.01	72.29	82.67
EISNet [46]	86.64	<u>81.53</u>	97.11	78.07	<u>85.84</u>
RSC [17]	87.89	82.16	97.92	83.35	87.83
RSC (*our imple.*)	83.92	79.52	95.15	<u>82.20</u>	85.20
FACT (*ours*)	**89.63**	**81.77**	96.75	**84.46**	**88.15**

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	<u>52.72</u>	73.50	74.80	64.72
CCSA [29]	59.90	49.90	74.10	75.70	64.90
MMD-AAE [24]	56.50	47.30	72.10	74.80	62.70
CrossGrad [38]	58.40	49.40	73.90	75.80	64.40
DDAIG [52]	59.20	52.30	<u>74.60</u>	76.00	65.50
L2A-OT [53]	**60.60**	50.10	**74.80**	**77.00**	<u>65.60</u>
Jigen [2]	53.04	47.51	71.47	72.79	61.20
RSC [17]	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*)	<u>60.34</u>	**54.85**	74.48	<u>76.55</u>	**66.56**

Ablation study: model component

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**

Ablation study: data augmentation

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**

Discussion

- Phase information contains meaningful semantics and helps generalization

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

Discussion

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, 14], 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 $w_{i,a} \rightarrow 0$ 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.