When Neural Networks Fail to Generalize? A Model Sensitivity Perspective

Jiajin Zhang[1], Hanqing Chao[1], Amit Dhurandhar[2], Pin-Yu Chen[2], Ali Tajer[3], Yangyang Xu[4], Pingkun Yan[1] *

[1]Department of Biomedical Engineering and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY, USA
[2]IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA
[3]Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA
[4]Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, USA

Abstract

Domain generalization (DG) aims to train a model to perform well in unseen domains under different distributions. This paper considers a more realistic yet more challenging scenario, namely Single Domain Generalization (Single-DG), where only a single source domain is available for training. To tackle this challenge, we first try to understand *when neural networks fail to generalize?* We empirically ascertain a property of a model that correlates strongly with its generalization that we coin as “model sensitivity”. Based on our analysis, we propose a novel strategy of Spectral Adversarial Data Augmentation (SADA) to generate augmented images targeted at the highly sensitive frequencies. Models trained with these hard-to-learn samples can effectively suppress the sensitivity in the frequency space, which leads to improved generalization performance. Extensive experiments on multiple public datasets demonstrate the superiority of our approach, which surpasses the state-of-the-art single-DG methods by up to 2.55%. The source code is available at https://github.com/DIAL-RPI/Spectral-Adversarial-Data-Augmentation.

1 Introduction

Deep learning models may perform poorly when tested on samples drawn from out-of-distribution (OoD) data. Applications encountering OoD problems commonly involve natural domain shift (Ben-David et al. 2010; Pan and Yang 2009) or image corruptions (Hendrycks and Gimpel 2016; Hendrycks and Dietterich 2019). To tackle the problem of performance degradation in unseen domains, extensive research has been carried out on domain generalization (DG), which attempts to extend a model to unseen target domains by regularizing the model and exposing it to more data.

Based on how the source domain knowledge gets transferred to an unseen target domain, the existing DG techniques can be divided into three categories (Wang et al. 2022), representation learning, constrained learning, and data manipulation. The former two categories explicitly regularize a model to improve its generalizability. These approaches aim to learn domain invariant predictors by enhancing the correlations between the domain invariant representations and the labels. We would like to point out that the data augmentation based methods are actually also regularizing models, but implicitly. One of the contributions of our work is to visualize and quantify the effect of implicit regularization of data augmentation strategies.

Most of the existing DG methods learn the representations from multiple source domains (Volpi et al. 2018; Dou et al. 2019; Muandet, Balduzzi, and Schölkopf 2013). However, in many applications, there is only one single source domain available for training (Volpi et al. 2018; Qiao, Zhao, and Peng 2020; Wang et al. 2021b). Despite the extensive literature on domain generalization, limited work deals with single source domain. In fact, many of the explicit regularization methods need multiple source domains to begin with and thus is inapplicable to this setting. Data augmentation, as an effective strategy in deep learning, has shown promising performance in single domain generalization (single-DG) problems (Volpi et al. 2018; Xu et al. 2020; Wang et al. 2021b). Such methods typically apply various operations to the source domain images to generate pseudo-novel domains (Wang et al. 2021b). Models will be trained using both the source domain images and the augmented images with designed constraints to learn invariant representations.

Despite the popularity of data augmentation in single-DG, the existing methods bear two major drawbacks. They are either model agnostic or provide very limited OoD augmentation. Recent study (Tan, Li, and Huang 2021) on measuring cross-domain transferability demonstrated that the characteristics of both model and training data are important factors when quantifying the model’s generalizability. However, the majority of data augmentation methods are model independent, which apply random generic image transformations to generate pseudo-domain images. Although they are helpful, those augmented images may not necessarily address the weaknesses of the models. In contrast, recent methods exploiting adversarial samples for domain generalization (Volpi et al. 2018; Qiao, Zhao, and Peng 2020; Zhang, Chao, and Yan 2020, 2023) learn to generate such augmentations by targeting the models’ weakness. However, the resulted minor perturbations to the samples in the image space only trivially enhance the appearance diversity. Therefore, these adversarial samples based augmentation methods lead to limited generalization performance improvement.

To tackle the above-mentioned challenges in data augmentation, we first ask a more fundamental question: *when*

[*] Corresponding author

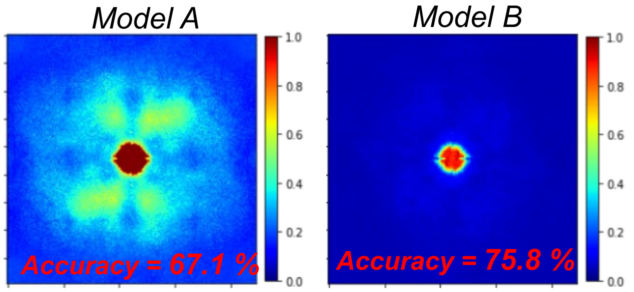

Figure 1: Introduction: we observed a clear correlation between the model generalization performance on an unseen target domain with the corresponding model sensitivity map.

do neural networks fail in domain generalization? In other words, we look into the model characteristics to quantify what aspects of the neural networks can reflect their generalizability. Inspired by the previous work on Fourier domain heatmap analysis (Yin et al. 2019), we first propose a model sensitivity analysis approach to compute sensitivity maps as surrogates to help quantify the spectral weaknesses of the model. Fig. 1 presents two example sensitivity maps of two models sharing the same architecture but were trained using different strategies. The corresponding prediction accuracy of the two models on an unseen target domain shows that *Model B* with less spectral sensitivity generalizes better than *Model A*, which has much higher sensitivity. More detailed analysis and results on the sensitivity maps and their association with model generalizability are included in the experiment part of this article.

The correlation between the generalization performance and model sensitivity map inspired us to design a novel data augmentation strategy to suppress the model sensitivity for improved single-DG. We thus propose *spectral adversarial data augmentation* (SADA), which curbs model sensitivity with targeted perturbation to the source domain data samples in the frequency space. Fig. 2 shows an overview of our framework. More specifically, we first train a model using the original source domain data through empirical risk minimization (ERM) and then compute the model sensitivity map. Since randomly augmenting images like in (Sun et al. 2021; Xu et al. 2020; Hendrycks et al. 2019) may need generating a large number of images, the cost of data augmentation can be high and the following model training will be slow. To efficiently suppress the model sensitivity, instead of applying random operations, we target at each sensitive frequency point on the map and employ the adversarial techniques (Zhang et al. 2022b,a) to generate hard-to-learn samples. Such adversarial operation of the image amplitude spectrum allows us to largely augment samples with more appearance variation. The generated samples are then mixed with the original samples to finetune the original model. Compared with other methods, SADA trained models present less sensitivity to domain shift across the frequency space, thus guarantee the better generalization performance. Based on such observation, we further develop a quantitative measure, which helps predict model generalizability.

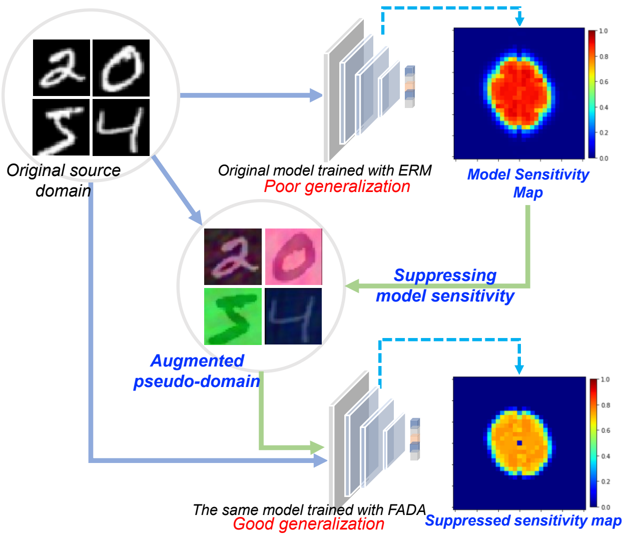

Figure 2: The overview of our proposed Spectral adversarial data augmentation (SADA). The proposed model sensitivity map presents as a spectral indicator to quantify the model generalizability. Augmented pseudo-domain images are generated by SADA to boost the model performance by suppressing the source model sensitivity map.

The major contributions of this work are as follows. *1)* We introduce spectral sensitivity map as an indicator to quantify the model generalizability, which also visualizes the effect of implicit regularization such as data augmentation. *2)* We propose SADA to improve the model generalization performance by suppressing the highly sensitive areas in the frequency space. SADA alleviates the drawbacks of the prior single-DG methods by targeting at model sensitivity and generating adversarial images with style variation. *3)* We present thorough empirical analysis to compare the proposed method with the baselines from multiple perspectives on public datasets.

2 Related Work

2.1 Explicit Regularization for DG

One line of works on DG aims to train domain invariant classifier with explicit regularization (Koyama and Yamaguchi 2020). A strategy that received significant attention in the last few years is invariant risk minimization (IRM) (Arjovsky et al. 2019). Given multiple environments, which correspond to different interventional distributions (viz. data from different sources) of a given data generating process, IRM promises to find invariant predictors that correspond to causal parents of a target variable. Efficient algorithms were designed (Ahuja et al. 2020) and further analysis in support of the principle (Ahuja et al. 2021b) have been done. However, it has been recently shown that the principle suffers from drawbacks in certain cases (Rosenfeld, Ravikumar, and Risteski 2021; Ahuja et al. 2021a), where it fails to uncover such predictors. Some studies adopt other strategies such as risk variance regularization (Krueger et al. 2021), domains gradients alignments (Koyama and Yamaguchi 2020), smoothing cross domain interpolation paths (Chuang and Mroueh 2021), and task-oriented techniques (Zhang et al. 2021). These approaches, however, generally require the target domain information, and cannot be directly adapted to the single-DG problem.

2.2 Implicit Regularization for DG

Data augmentation has been widely used to improve the generalization of deep learning models, which acts by implicitly regularization. Due to their effectiveness and simplicity, methods from the ***Mix*** family are the most commonly used approaches for data augmentation. They augment data by mixing images with different random combinations, *e.g.*, MixUp (Zhang et al. 2017), CutMix (Yun et al. 2019), AugMix (Hendrycks et al. 2019), PixMix (Hendrycks et al. 2022). In the single-DG settings, RandConv (Xu et al. 2020) augments images with a random convolutional layer. L2D (Wang et al. 2021b) diversifies the image styles via mutual information maximization. However, most of the methods are model independent and thus the augmented images may not necessarily address the weaknesses of the models.

Adversarial training generates hard-to-learn samples targeted at the model weakness. To improve the single-DG performance, DUG (Volpi et al. 2018) adversarially augments the representations of images to a fictitious domain. M-ADA (Qiao, Zhao, and Peng 2020) introduced a meta-learning framework to learn multi-adversarial domains with an autoencoder. AugMax (Wang et al. 2021a) generates adversarial samples by selecting the worst-case weights of AugMix. However, the resulting minor perturbations in the image space only trivially enhance the appearance diversity. Thus, adversarial-based augmentation methods usually lead to limited generalization improvement.

Frequency spectrum augmentation methods, including FDA (Yang and Soatto 2020), FDG (Xu et al. 2021) and FedDG (Liu et al. 2021), generate images by either mixing up or swapping the low-frequency components of the source and target domain amplitude spectrum. Because of requiring target domain data, these methods cannot be directly adapted to the single-DG problem. To enhance the adversarial robustness under domain shift, FourierMix (Sun et al. 2021) augments source images by adding noise to both amplitude and phase spectra. Our method instead, aiming to suppress the model spectral sensitivity, adversarially augments the image amplitude spectrum. We experimentally compare to typical baselines under a single-DG setting and demonstrate the superior performance.

3 Methodology

The objective of single-DG is to train a model in one source domain, that can generalize well in many unseen target domains. We denote the source domain by $\boldsymbol{X}_S = \{(\boldsymbol{x}, \boldsymbol{y})\}$. $\boldsymbol{x} \in \mathbb{R}^{w \times h}$ is the source image, where w and h is the width and height. $\boldsymbol{y}$ is the corresponding label. As introduced in Fig 2, to tackle this challenge, we propose the framework of *spectral adversarial data augmentation* (SADA) to boost the model's generalizability by suppressing its spectral sensitivity. SADA first computes a model sensitivity map as a surrogate of the model vulnerability in the frequency space. Then it uses the model sensitivity map as guidance to synthesize spectral adversarial images, which encodes the model sensitivity into hard-to-learn augmentation images. In this section, the model sensitivity measurement and spectral adversarial augmentation processes are discussed in detail.

3.1 Amplitude-modulated Sensitivity Map

To quantify the model's vulnerability/weakness to the different frequency corruptions, (Yin et al. 2019) previously proposed the Fourier sensitivity analysis. Briefly, a *Fourier basis* $\boldsymbol{A}_{i,j} \in \mathbb{R}^{w \times h}$ is defined as a Hermitian matrix with only two non-zero elements at (i, j) and $(-i, -j)$, where the origin is at the image center. A *Fourier basis image* $\boldsymbol{U}_{i,j}$ is a real-valued matrix in the pixel space. It is defined as the ℓ_2-normalized Inverse Fast Fourier Transform (IFFT) of $\boldsymbol{A}_{i,j}$, *i.e.*, $\boldsymbol{U}_{i,j} = \frac{\mathcal{IFFT}(\boldsymbol{A}_{i,j})}{||\mathcal{IFFT}(\boldsymbol{A}_{i,j})||_2}$. Perturbed images are generated by adding the *Fourier basis noise*

$$\boldsymbol{N}_{i,j} = r \cdot \epsilon \cdot \boldsymbol{U}_{i,j} \tag{1}$$

to the original image $\boldsymbol{x}$ as $\boldsymbol{x} + \boldsymbol{N}_{i,j}$, where ϵ is a frequency-independent constant value to control the ℓ_2-norm of the perturbation and r is randomly sampled to be either -1 or 1. The *Fourier basis noise* $\boldsymbol{N}_{i,j}$ only introduces perturbations at the frequency components (i, j) and $(-i, -j)$ to the original images. The constant ϵ guarantees that images are uniformly perturbed across all frequency components. For RGB images, we add $\boldsymbol{N}_{i,j}$ to each channel independently following (Yin et al. 2019). The sensitivity at frequency (i, j) of a given model F trained on source domain is defined as the prediction error rate over the whole dataset $\boldsymbol{X}_S$:

$$\boldsymbol{M}_{org}(i,j) = 1 - \underset{(\boldsymbol{x},\boldsymbol{y}) \in \boldsymbol{X}_S}{\mathrm{Acc}} (F(\boldsymbol{x} + \boldsymbol{N}_{i,j}, \boldsymbol{y})), \tag{2}$$

where Acc is the model prediction accuracy. By aggregating all the model sensitivity entries $\boldsymbol{M}_{org}(i, j)$ across the frequency space, a 2D model sensitivity map can be obtained as shown by the examples in Fig. 1. The lowest frequency is at the center of the map and higher frequencies are closer to the edges.

Since ϵ is a frequency-independent constant, the original model sensitivity map defined by Eq. 2 describes model's local vulnerability by uniformly perturbing all frequency components of the source images. Instead of a uniform distribution, the amplitude spectrum of natural images generally follows a power-law distribution (Tolhurst, Tadmor, and Chao 1992). Low-frequency amplitudes have much higher values than the high-frequency ones, and can vary more significantly across domains (Yang and Soatto 2020). Models generally would generalize poorly if such low-frequency variability is not presented in the training set (Yang and Soatto 2020; Liu et al. 2021). These observations indicate that the low-frequency components of images with large amplitude should be perturbed more significantly to truly reflect the model vulnerability *w.r.t.* the domain shift problem.

Thus, we propose to enhance the model sensitivity map by using the source domain amplitude spectrum as domain

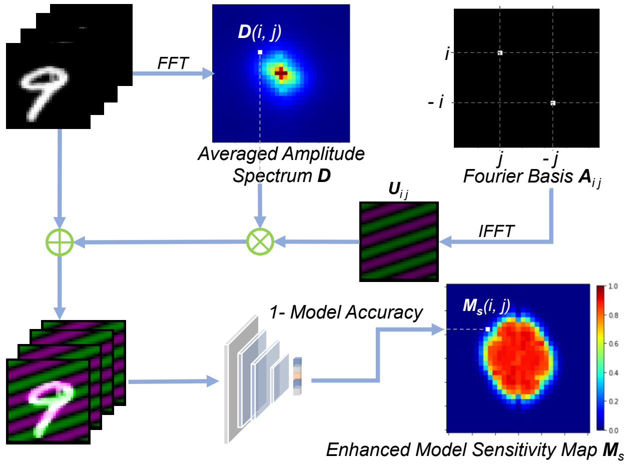

Figure 3: Enhancing model sensitivity map with source domain amplitude distribution. The averaged source amplitude spectrum D is encoded into the perturbed images.

prior. As shown in Fig. 3, a mean amplitude spectrum $\boldsymbol{D}$ is first computed by averaging the amplitude spectrum of all images in the source domain. Then, we reformulate the original *Fourier basis noise* $\boldsymbol{N}_{i,j}$ by

$$\hat{\boldsymbol{N}}_{i,j} = r \cdot \boldsymbol{D}(i,j) \cdot \boldsymbol{U}_{i,j}, \tag{3}$$

where the frequency-independent ϵ in Eq. 1 is replaced with the $(i,j)_{th}$ entry of $\boldsymbol{D}$ to control the noise level.

Adopted from Eq. 2, the enhanced model sensitivity at frequency (i,j) is computed by evaluating the prediction error rate on the perturbed source images as by

$$\boldsymbol{M}_S(i,j) = 1 - \underset{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{X}_S}{\mathrm{Acc}} (F(\boldsymbol{x} + r \cdot \boldsymbol{D}(i,j) \cdot \boldsymbol{U}_{i,j}), \boldsymbol{y}), \tag{4}$$

where F is a model trained with empirical risk minimization (ERM) by minimizing the cross entropy loss $\mathcal{L}_{ERM} = \underset{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{X}_S}{\boldsymbol{E}} \ell_{CE}(F(\boldsymbol{x}), \boldsymbol{y})$. In the experiment section, we quantitatively compared the enhanced model sensitivity map to the original one from different perspectives.

3.2 Spectral Adversarial Data Augmentation

The model sensitivity $\boldsymbol{M}_S$ describes the model spectral weakness *w.r.t.* the domain shift, which strongly correlates with the model cross-domain generalizability. The model sensitivity of certain frequencies can be suppressed if the diversity of the training data increases at those frequency elements. Random spectral perturbation to the source images may help increase the overall diversity, however, generally lacks efficiency to sufficiently cover all potential pseudo domains. Following this direction, we propose a *spectral adversarial data augmentation* (SADA) method, which curbs model sensitivity with targeted perturbation to the source domain data samples in the spectral space. Instead of random transformation, SADA employs an adversarial technique to directly search for hard-to-learn samples by adding specially designed perturbations to the source images.

The entire pipeline of SADA is summarized in Alg. 1. More specifically, given a source domain image $\boldsymbol{x}$, its spectral amplitude $\boldsymbol{A}_{org}$ and phase $\boldsymbol{P}_{org}$ are computed by the Fast Fourier Transform (FFT) as

$$\boldsymbol{A}_{org}, \boldsymbol{P}_{org} = \mathcal{FFT}[\boldsymbol{x}]. \tag{5}$$

Then the original amplitude spectrum $\boldsymbol{A}_{org}$ is initialized with random perturbation as

$$\boldsymbol{A}_0 = \boldsymbol{A}_{org} \odot (1 + \mathrm{Unif}(-\epsilon, \epsilon)), \tag{6}$$

where $\mathrm{Unif}(-\epsilon,\epsilon) \in \mathbb{R}^{\mathrm{w}\times\mathrm{h}}$ represents 2D matrix with each entry sampled uniformly from $[-\epsilon, \epsilon]$, and $\odot$ denotes the Hadamard product.

To target at each sensitive frequency component, as in Eq. 7, the amplitude spectrum $\boldsymbol{A}_{t+1}$ is optimized iteratively by adding the $\boldsymbol{M}_S$-weighted sign gradient of the cross-entropy loss to the amplitude spectrum $\boldsymbol{A}_t$ with δ as the perturbation step size.

$$\begin{aligned} \boldsymbol{A}_{t+1} = \boldsymbol{A}_t \cdot \{1+ \\ \delta \cdot sign[\frac{\partial \ell_{CE}(F(\mathcal{IFFT}[\boldsymbol{A}_t, \boldsymbol{P}_{org}]), \boldsymbol{y})}{\partial \boldsymbol{A}_t}] \odot \boldsymbol{M}_S\} \end{aligned} \tag{7}$$

Previous studies (Piotrowski and Campbell 1982; Hansen and Hess 2007; Oppenheim and Lim 1981; Oppenheim et al. 1979) have demonstrated that the phase spectrum retains most of the semantic structure information of the original signals, while the amplitude mainly contains the style/domain-related statistics. Since the data augmentation objective is to diversify the image styles without affecting the original semantic meaning, we adversarially perturb the amplitude spectrum while keeping the original phase spectrum. That is, in each iteration, the augmented image is reconstructed from the updated amplitude $\boldsymbol{A}_{t+1}$ and the original phase spectrum $\boldsymbol{P}_{org}$. The reconstructed image is then clamped into the definition region $[0,1]$ by $\boldsymbol{x}_{t+1} = \mathrm{Clamp}(\mathcal{FFT}[\boldsymbol{A}_{\mathrm{t+1}}, \boldsymbol{P}_{\mathrm{org}}], 0, 1)$.

Algorithm 1: Spectral adversarial data augmentation.

```
Input: model $F$; source data $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^{N}$; initial perturbation
    level $\epsilon$; max steps $T$ and step size $\delta$; sensitivity map $\boldsymbol{M}_S$.
Output: augmented images $\{\tilde{\boldsymbol{x}}_k\}_{k=1}^{N}$
for $k \leftarrow 1$ to $N$ do
    Compute the original spectrum $\boldsymbol{A}_{org}, \boldsymbol{P}_{org}$ by Eq. 5
    Randomly initialize amplitude $\boldsymbol{A}_0$ by Eq. 6
    for $t \leftarrow 0$ to $T$ do
        $\boldsymbol{x}_{k,t} \leftarrow \mathrm{Clamp}(\mathcal{IFFT}[\boldsymbol{A}_{\mathrm{t}}, \boldsymbol{P}_{\mathrm{org}}], 0, 1)$
        if model prediction is changed by $\boldsymbol{x}_{k,t}$ then
            break                          ▷ Early stop for acceleration
        end if
        Update amplitude spectrum $\boldsymbol{A}_t$ by Eq. 7
        $\boldsymbol{A}_t = \mathrm{Max}(\boldsymbol{A}_{\mathrm{t}}, 0)$          ▷ constrain $\boldsymbol{A}_t > 0$
    end for
    $\tilde{\boldsymbol{x}}_k \leftarrow \mathrm{Clamp}(\boldsymbol{x}_{\mathrm{k,t}}, 0, 1)$
end for
```

3.3 Model training

To learn invariant representations, we regularize the prediction consistency among the original image and all augmented images through a Jensen-Shannon (JS) divergence (Hendrycks et al. 2019). The total training loss is

$$\mathcal{L} = \mathcal{L}_{ERM} + \lambda \cdot \mathrm{JS}(\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_\mathrm{n}), \tag{8}$$

where λ is the trade-off parameter and $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $...\boldsymbol{x}_n$ are the n augmented images from the same original image $\boldsymbol{x}_0$. The JS divergence is defined as $\mathrm{JS}(\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_\mathrm{n}) = \frac{1}{\mathrm{n+1}} \sum_{\mathrm{i=0}}^{\mathrm{n}} \mathrm{KL}(\mathrm{F}(\boldsymbol{x}_\mathrm{i}) || \bar{\boldsymbol{p}})$, where KL is the Kullback-Leibler divergence, $F(\boldsymbol{x}_i)$ is the model prediction probability of $\boldsymbol{x}_i$ and $\bar{\boldsymbol{p}} = \frac{1}{n+1} \sum_{i=0}^{n} F(\boldsymbol{x}_i)$.

4 Experiments

In this section, we conduct comprehensive experiments to evaluate SADA from different perspectives. Specifically, we aim to answer the following questions: **Q1**: Compared with prior methods, can SADA effectively improve the single-DG performance? (Sec. 4.1) **Q2**: Can the proposed model sensitivity map suggest when neural networks may generalize well? (Sec. 4.2) **Q3**: How is the data efficiency of SADA compared with other data augmentation methods? (Sec. 4.3)

Datasets To answer those questions, we evaluated our method and other benchmarks on three benchmark datasets.

1) **DIGITS** consists of 5 domains, including **MNIST** (LeCun et al. 1998), **SVHN** (Netzer et al. 2011), **MNIST-M** (Ganin and Lempitsky 2015), **SYNTH** (Ganin and Lempitsky 2015) and **USPS** (LeCun et al. 1989). We converted all the gray scale images to RGB images.

2) **PACS** (Li et al. 2017) is a more challenging domain generalization dataset including four domains, **P**hoto, **A**rt painting, **C**artoon, and **S**ketch. We follow the official dataset split for training validation and testing.

3) **CIFAR-10-C** (Hendrycks and Dietterich 2019) is the corrupted version of CIFAR-10 (Krizhevsky and Hinton 2009) by four categories of corruption, *i.e.*, weather, blur, noise, and digital. Each corruption has 5 level severity.

Implementation details In all the experiments, we set the weighting factor $\lambda = 0.25$, perturbation steps $T = 5$, step size $\delta = 0.08$ and random initialization range $\epsilon = 0.2$. The number of augmented images per training sample in Eq. 8 is set to 3, based on our empirical evaluation results. We also include three SADA variants, 1SADA+2Mix, 2SADA+1Mix and 3SADA+0Mix, for comparison. The '#' indicates the number of SADA and AugMix images included in the 3 augmented images per training sample.

For a fair comparison with other methods, we directly adopted the same network architectures of the previous works (Volpi et al. 2018; Qiao, Zhao, and Peng 2020; Wang et al. 2021b). For **DIGITS** dataset, we trained a ConvNet (LeCun et al. 1998) with SGD optimizer (default settings) for 50 epochs. The initial learning rate is 0.001, which decays by 0.1 for every 20 epochs. The batch size is 128. For the **PACS** dataset, ResNet-18 (He et al. 2016) is pretrained on Imagenet and finetuned in the source domain by SGD for 80 epochs. The initial learning of 0.01 is scheduled to decay by 0.1 for every 20 epochs. The batch size is 256. For **CIFAR-10-C**, a Wide Residual Network (Zagoruyko and Komodakis 2016) with 16 layers and width of 4 (WRN-16-4) was optimized with SGD for 200 epochs with batch size 256. The initial learning rate of 0.1 linearly decays by 0.1 for every 40 epochs.

4.1 Method Effectiveness

We compared SADA with ERM, CCSA (Motiian et al. 2017), JiGen (Carlucci et al. 2019), d-SNE (Xu et al. 2019), AugMix (Hendrycks et al. 2019), GUD (Volpi et al. 2018), M-ADA (Qiao, Zhao, and Peng 2020), RandConv (Xu et al. 2020) and L2D (Wang et al. 2021b). The same model architecture was used for all the approaches.

DIGITS Table 1 shows the 3-run average accuracy of all the methods trained on **MNIST** and evaluated in each target domain. All the variants of SADA achieved better accuracy than the baselines. Specifically, significant improvements of 5.50% and 9.08% are observed on the two very challenging target domains, **SVHN** and **SYNTH**, respectively. This performance gain mainly contributes to the spectrally augmented samples with large appearance/style variation. As we pointed out earlier, adversarial-based methods, such as GUD and M-ADA, generates only minor perturbations in the image space to enhance the appearance diversity, and thus couldn't outperform the random data augmentation methods, such as RandConv.

PACS We train a model in a single source domain and test on the other three target domains. The averaged accuracy on the three target domains are reported in Table 2. The proposed SADA variants achieve the best performance in 3 out of the 4 source domains, i.e., **P**hoto, **A**art, and **S**ketch. Both 2SADA+1Mix and 3SADA+0Mix achieved over 3.9% improvement with **S**ketch as the source domain, which contains the largest domain shift from the other three colored domains. In addition, 2SADA+1Mix and 3SADA+0Mix consistently outperform 1SADA+2Mix, which indicates the importance of SADA augmented images. These observa-

	Target Domain				
Method	**U**	**M**	**V**	**S**	Average
ERM	76.90	52.74	27.85	39.65	49.29
CCSA	83.72	49.29	25.89	37.31	49.05
JiGen	77.16	57.80	33.81	43.79	53.14
d-SNE	**93.16**	50.98	26.22	37.83	52.05
AugMix	80.24	75.86	63.85	69.84	72.45
GUD	77.26	60.41	35.51	45.32	55.67
M-ADA	78.53	67.94	42.55	48.95	59.49
RandConv	84.37	**87.77**	57.56	62.85	72.88
L2D	83.95	87.32	62.85	63.72	74.45
1SADA+2Mix	81.92	80.88	67.66	70.65	75.28
2SADA+1Mix	89.34	75.74	68.34	72.10	76.38
3SADA+0Mix	89.29	75.61	**68.45**	**72.90**	**76.56**

Table 1: 3-run average accuracy of MNIST-trained models evaluated on USPS(U), MNIST-M(M), SVHN(V), and SYNTH(Y). Best performance is in bold.

Method	Source Domain				Average
	Photo	**A**rt	**C**atoon	**S**ketch	
ERM	33.52	57.86	67.84	25.12	46.09
CCSA	42.77	61.89	67.46	26.43	51.08
JiGen	43.49	63.66	70.08	32.47	52.43
d-SNE	46.28	63.20	26.22	37.83	52.05
AugMix	48.27	72.92	73.81	54.88	62.47
GUD	45.62	69.47	73.46	41.67	57.56
M-ADA	48.22	70.46	75.67	43.26	59.40
RandConv	50.86	75.82	75.46	48.90	62.76
L2D	51.17	76.90	**77.80**	53.68	64.74
1SADA+2Mix	51.26	76.98	76.26	55.91	65.20
2SADA+1Mix	**51.22**	**77.82**	76.94	**57.76**	**66.18**
3SADA+0Mix	51.18	77.68	76.35	57.61	65.71

Table 2: 3-run average accuracy of models trained in each single domain (Photo, Art, Catoon, Sketch). Best performance is in bold.

tions verified that model generalization performance can improve, if SADA is included for suppressing the model's sensitivity.

CIFAR-10-C Besides the natural domain shift in **DIGITS** and **PACS**, we further evaluated the method on the image corruption dataset. Table 3 shows the average accuracy of all the methods trained on **CIFAR-10** and evaluated on four types of corruption under the severest level 5. The averages accuracy of 2SADA+1Mix surpasses the best baseline AugMix by 2.55%. More detailed performance of all five-level corruptions are included in appendix, where the proposed SADA consistently outperforms other baseline methods at different severity levels. The results validate that SADA not only handles the natural domain shift but is resilient to artificial corruptions.

4.2 Model Sensitivity Perspective

To verify if the enhanced model sensitivity map can indicate the model's generalizability, we further computed the sensitivity maps of ConvNet trained with different strategies on the **MNIST** dataset as shown by the examples in Fig. 4. *First*, different from the original sensitivity maps (Eq. 2) in Fig. 4 **(a)**, the enhanced sensitivity maps (Eq. 4) in Fig. 4 **(b)** show that source models are more vulnerable to the perturbations in the low frequency region. This result matches with the observations in the previous studies (Yang and Soatto 2020; Liu et al. 2021), that the models cannot generalize well due to the low-frequency amplitude difference between the source and target domains is large. *Second*, in Fig. 4 **(b)**, comparing the model sensitivity map of ERM with the sensitivity maps of other single-DG approaches, our enhanced sensitivity computation clearly shows how the single-DG approaches can help improve model performance by suppressing the model sensitivity, especially in the low-frequency space. To better visualize the observations, we present the scattering plot of the accuracy versus the averaged ℓ_1-norm of model sensitivity map. As shown in Fig. 4**(c)**, the original sensitivity computation method fails to correlate the model performance and sensitivity. In contrast, Fig. 4**(d)** shows that the enhanced model sensitivity computation provides strong correlation between the model performance and sensitivity. The model prediction accuracy degrades significantly when the ℓ_1-norm of model sensitivity maps increases. These results demonstrate that the enhanced model sensitivity map in Eq. 4 could be used for *visualizing and quantifying the effect of implicit regularization* on model generalizability.

Method	Corruption Category				Average
	Weather	**Blur**	**Noise**	**Digits**	
ERM	67.21	56.73	30.26	62.30	54.08
CCSA	67.66	57.81	28.73	61.96	54.04
JiGen	67.20	58.06	30.37	62.05	54.43
d-SNE	67.90	56.59	33.97	61.83	55.07
AugMix	78.53	82.04	64.45	76.17	75.28
GUD	69.94	60.57	48.66	60.37	59.91
M-ADA	75.54	63.76	54.21	65.10	64.65
RandConv	76.87	55.36	**75.19**	77.51	71.23
L2D	75.98	70.21	73.29	72.02	72.88
1SADA+2Mix	78.69	82.10	67.95	77.32	75.52
2SADA+1Mix	79.14	**82.38**	71.42	78.38	**77.83**
3SADA+0Mix	**79.44**	80.68	70.77	**78.42**	77.33

Table 3: 3-run average accuracy of models trained on CIFAR-10 and evaluated on CIFAR-10-C dataset. Best performance is in bold.

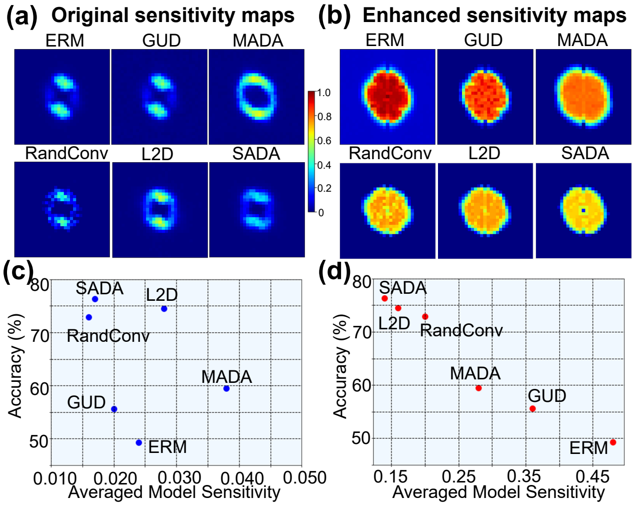

Figure 4: (a) and (b): original and our proposed model sensitivity maps of different single-DG methods. (c) and (d): model performance versus the model sensitivity maps.

4.3 Data Efficiency

Due to the limited availability of data, data efficiency is an important indicator of data augmentation performance. We evaluated the model accuracy by gradually decreasing the amount of augmented images used in the training process. Fig. 5 shows the analysis results of using **MNIST** and **P**hoto as source domains on **DIGITS** and **PACS**, respectively. Our method consistently outperforms other baselines when only a proportion of the augmented data are used for training. It is

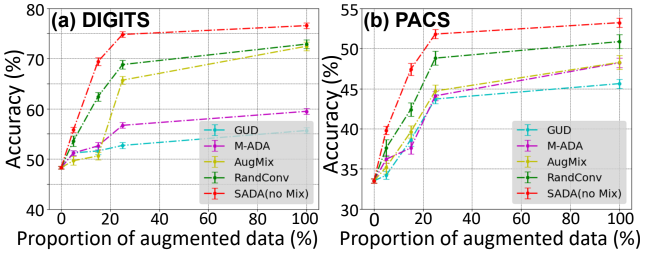

Figure 5: Evaluation of data efficiency of SADA.

also very impressive to see that, with only 25% augmented data for training, our method can generalize better than the baselines trained with the fully augmented dataset.

5 Further Analysis and Discussion

5.1 Ablation Studies

We conducted ablation studies on both **DIGITS** and **PACS** datasets to verify the effectiveness of each component in SADA. Table 4 reports the performance with 2SADA+1Mix as example. We first removed the two SADA augmented images (w/o SADA) and the performance degraded more than 20% in all the target domains, which clearly shows the significance of SADA in the whole framework. Second, without the AugMix (w/o Mix), *i.e.,* 2SADA+0Mix, the model performance decreased by 7% in each target domain. That is because AugMix includes several random image style transfer operations, such as 'solarization' and 'autocontrast', which diversify the augmented images to complement our targeted spectral augmentation. Third, we also observed the performance drop if the models are trained without the JS divergence, which helps learn invariant representations for improved generalizability.

5.2 Effectiveness of Targeted Augmentation

This section examines the effectiveness of the proposed model sensitivity map and the targeted adversarial perturbation. We evaluate the model performance with 3SADA+0Mix by 1) using the original model sensitivity map; 2) replacing the adversarial spectral augmentation with the random spectral perturbation($\epsilon = 0.2$) following (Sun et al. 2021). The results in Fig. 6 show that 3SADA+0Mix generalizes worse to unseen domains if our proposed components are replaced by the two alternative approaches on different source domains. In addition, we also evaluate the time consumption regarding the sensitivity map generation and the spectral adversarial augmentation in the appendix.

Component	**USPS**	**MNIST-M**	**SVHN**	**SYNTH**	Avg
2SADA+1Mix	89.34	75.74	68.34	72.10	76.38
w/o SADA	69.62	53.57	47.16	49.02	60.27
w/o Mix	82.79	69.43	62.44	63.18	69.46
w/o JS	81.68	70.35	64.32	65.11	70.37

Table 4: Ablation of SADA(1Mix) on DIGITS dataset.

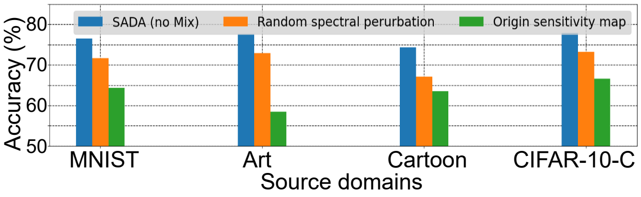

Figure 6: 3SADA+0Mix performance comparison on model sensitivity map and spectral perturbation.

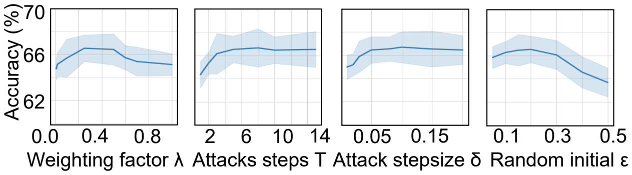

Figure 7: Sensitivity analysis of different hyperparameters.

5.3 Hyperparameter Sensitivity Analysis

To validate the significance of weighting factor λ, perturbation steps T and step size δ, and random initialization ϵ, we conduct sensitivity analysis of 2SADA+1Mix on **PACS** dataset as presented in Fig. 7. In the experiments, we initially set $\lambda = 0.25$, $T = 5$, $\delta = 0.08$ and $\epsilon = 0.20$. When analyzing the sensitivity to one parameter, the other parameters are fixed. When λ is within $[0.1, 0.6]$, our method consistently outperforms other baselines (Fig. 7**a)**). That is due to the balance between the JS loss and the ERM loss. When the perturbation gets stronger, Fig. 7**b)** and **c)** show that the performance increases initially, and then stays stable. It is because the early-stop acceleration is adopted to control the perturbation strength. Fig. 7**d)** shows that the model performance is stable if the perturbation strength $\epsilon < 0.30$, which decreases if the randomization gets too strong.

6 Conclusion and Discussion

In this paper, an enhanced model sensitivity map is proposed to empirically ascertain *when the neural networks may fail in domain generalization* from a new perspective of spectral sensitivity. Our analysis shows that models with high sensitivity may not generalize well. Based on our analysis, we develop a novel framework of *Spectral Adversarial Data Augmentation* (SADA) to tackle single-DG by generating adversarially augmented images targeted at the highly sensitive frequencies. By successfully suppressing the model sensitivity in the frequency space, the experimental results on three public benchmarking datasets demonstrate that SADA can efficiently train a high-performance model resilient to various unseen domain shifts.

Acknowledgements

This research was partially supported by the National Science Foundation (NSF) under the CAREER award OAC 2046708 and the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons).

References

Ahuja, K.; Caballero, E.; Zhang, D.; Bengio, Y.; Mitliagkas, I.; and Rish, I. 2021a. Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization. In *Advances in Neural Inf. Proc. Systems*.

Ahuja, K.; Shanmugam, K.; Varshney, K.; and Dhurandhar, A. 2020. Invariant risk minimization game. In *International Conference on Machine Learning*.

Ahuja, K.; Wang, J.; Dhurandhar, A.; Shanmugam, K.; and Varshney, K. R. 2021b. Empirical or Invariant Risk Minimization? A Sample Complexity Perspective. In *International Conference on Learning Representations*.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning*, 79(1): 151–175.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2229–2238.

Chuang, C.-Y.; and Mroueh, Y. 2021. Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*.

Dou, Q.; Coelho de Castro, D.; Kamnitsas, K.; and Glocker, B. 2019. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Hansen, B. C.; and Hess, R. F. 2007. Structural sparseness and spatial phase alignment in natural scenes. *JOSA A*, 24(7): 1873–1885.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. CVPR. 2016. *arXiv preprint arXiv:1512.03385*.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. AugMix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*.

Hendrycks, D.; Zou, A.; Mazeika, M.; Tang, L.; Li, B.; Song, D.; and Steinhardt, J. 2022. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16783–16792.

Koyama, M.; and Yamaguchi, S. 2020. When is invariance useful in an Out-of-Distribution Generalization problem? *arXiv preprint arXiv:2008.01883*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario, Canada.

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Le Priol, R.; and Courville, A. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, 5815–5826. PMLR.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4): 541–551.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

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

Liu, Q.; Chen, C.; Qin, J.; Dou, Q.; and Heng, P.-A. 2021. FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1013–1023.

Motiian, S.; Piccirilli, M.; Adjeroh, D. A.; and Doretto, G. 2017. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, 5715–5725.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, 10–18. PMLR.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Oppenheim, A.; Lim, J.; Kopec, G.; and Pohlig, S. 1979. Phase in speech and pictures. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, 632–637. IEEE.

Oppenheim, A. V.; and Lim, J. S. 1981. The importance of phase in signals. *Proceedings of the IEEE*, 69(5): 529–541.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359.

Piotrowski, L. N.; and Campbell, F. W. 1982. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3): 337–346.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12556–12565.

Rosenfeld, E.; Ravikumar, P.; and Risteski, A. 2021. The Risks of Invariant Risk Minimization. In *International Conference on Learning Representations*.

Sun, J.; Mehra, A.; Kailkhura, B.; Chen, P.-Y.; Hendrycks, D.; Hamm, J.; and Mao, Z. M. 2021. Certified adversarial defenses meet out-of-distribution corruptions: Benchmarking robustness and simple baselines. *arXiv preprint arXiv:2112.00659*.

Tan, Y.; Li, Y.; and Huang, S.-L. 2021. OTCE: A transferability metric for cross-domain cross-task representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15779–15788.

Tolhurst, D. J.; Tadmor, Y.; and Chao, T. 1992. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2): 229–232.

Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J. C.; Murino, V.; and Savarese, S. 2018. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31.

Wang, H.; Xiao, C.; Kossaifi, J.; Yu, Z.; Anandkumar, A.; and Wang, Z. 2021a. AugMax: Adversarial composition of random augmentations for robust training. *Advances in neural information processing systems*, 34: 237–250.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Yu, P. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, Z.; Luo, Y.; Qiu, R.; Huang, Z.; and Baktashmotlagh, M. 2021b. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 834–843.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14383–14392.

Xu, X.; Zhou, X.; Venkatesan, R.; Swaminathan, G.; and Majumder, O. 2019. d-SNE: Domain adaptation using stochastic neighborhood embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2497–2506.

Xu, Z.; Liu, D.; Yang, J.; Raffel, C.; and Niethammer, M. 2020. Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003*.

Yang, Y.; and Soatto, S. 2020. FDA: Fourier Domain Adaptation for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4085–4095.

Yin, D.; Gontijo Lopes, R.; Shlens, J.; Cubuk, E. D.; and Gilmer, J. 2019. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, J.; Chao, H.; Dasegowda, G.; Wang, G.; Kalra, M.; and Yan, P. 2022a. Quantifying Trustworthiness of Explainability in Medical AI.

Zhang, J.; Chao, H.; Dasegowda, G.; Wang, G.; Kalra, M. K.; and Yan, P. 2022b. Overlooked Trustworthiness of Saliency Maps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 451–461. Springer.

Zhang, J.; Chao, H.; Xu, X.; Niu, C.; Wang, G.; and Yan, P. 2021. Task-oriented low-dose CT image denoising. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 441–450. Springer.

Zhang, J.; Chao, H.; and Yan, P. 2020. Robustified Domain Adaptation. *arXiv preprint arXiv:2011.09563*.

Zhang, J.; Chao, H.; and Yan, P. 2023. Towards Adversarial Robustness in Unlabeled Target Domains. *IEEE Transactions on Image Processing*, in press.