DFM-X: Augmentation by Leveraging Prior Knowledge of Shortcut Learning

Shunxin Wang Christoph Brune Raymond Veldhuis Nicola Strisciuglio
University of Twente, The Netherlands

Abstract

*Neural networks are prone to learn easy solutions from superficial statistics in the data, namely shortcut learning, which impairs generalization and robustness of models. We propose a data augmentation strategy, named DFM-X, that leverages knowledge about frequency shortcuts, encoded in **D**ominant **F**requencies **Maps** computed for image classification models. We randomly select **X**% training images of certain classes for augmentation, and process them by retaining the frequencies included in the DFMs of other classes. This strategy compels the models to leverage a broader range of frequencies for classification, rather than relying on specific frequency sets. Thus, the models learn more deep and task-related semantics compared to their counterpart trained with standard setups. Unlike other commonly used augmentation techniques which focus on increasing the visual variations of training data, our method targets exploiting the original data efficiently, by distilling prior knowledge about destructive learning behavior of models from data. Our experimental results demonstrate that DFM-X improves robustness against common corruptions and adversarial attacks. It can be seamlessly integrated with other augmentation techniques to further enhance the robustness of models. Codes are available at* https://github.com/nis-research/dfmX-augmentation.

1. Introduction

Neural networks are subject to shortcut learning, namely a tendency to relying on simple solutions to optimization problems, based on spurious correlations between data and ground truth. Shortcut solutions are thus one of the factors that negatively affect generalization and robustness performance of trained models [8, 22]. Mitigating shortcut learning was shown to be beneficial for enhancing the generalization performance and robustness of models [18, 2]. By enforcing models to learn from deeper task-related semantics instead of shallow correlations between data and ground truth that facilitate easy predictions during training, shortcut learning can be effectively addressed [15, 7, 14, 16, 17]. Existing methods to mitigate the learning of shortcut features include identifying and imitating shortcut features in the other class to reduce their specificity for classification [15], as well as measuring the amount of shortcut information present in the training data [7, 14, 16, 17]. However, these approaches are often limited to visually observable shortcut features (e.g. color patches and lines) or complex training strategies to learn image representations containing fewer shortcut features. Imitating or inducing shortcut features in the images of other classes [15] is a type of data augmentation. Commonly used data augmentation techniques, e.g. AugMix [10] and AutoAugment [5], do not usually take shortcut learning into account, but focus more on directly increasing data variety to bridge the distribution gap between training and testing data, improving the generalizability of models.

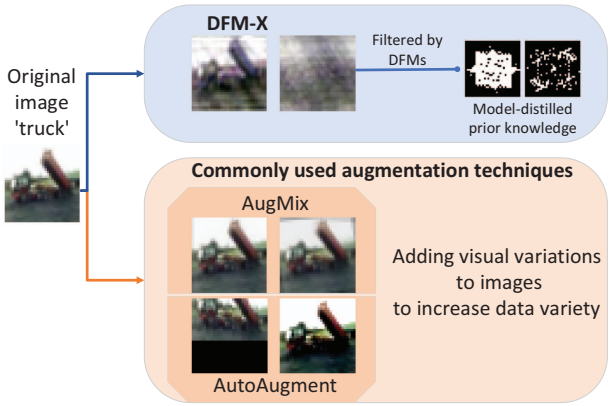


Figure 1: DFM-X exploits the original data efficiently, using model-distilled knowledge about shortcut learning behavior that impairs the generalization and robustness of models, rather than directly adding visual variations to the images like commonly used augmentation techniques.

In this work, we propose a data augmentation method called DFM-X. It is based on prior knowledge about frequency shortcuts [21, 23], which are identified as small sets of specific frequencies that contribute to achieving high-accuracy classification. We compute Dominant Frequency Maps (DFMs) for each class of a previously trained

model [21], and use them as prior knowledge of where destructive learning behavior happens in existing models to perform targeted data augmentation. Our work shares a similar idea with imitating shortcut features, like including color patches specific for a certain class [15], in the images of the other class. In this work, we imitate frequency shortcut features to reduce the reliance of models on specific frequency sets for classification, thus enforcing models to learn from a wider range of frequencies. We leverage the algorithm proposed in [21] to measure the dominant frequencies that play a crucial role in classifying each class, resulting in dominant frequency maps (DFMs). We utilize DFMs in our augmentation approach as prior knowledge (distilled by models from the data) to avoid unwanted learning behavior. This improves the generalizability and robustness of models to common corruptions and adversarial attacks in computer vision. We demonstrate the difference between DFM-X and other commonly used augmentation techniques in Figure 1. Compared with AugMix and AutoAugment, DFM-X makes an effort to exploit the original data in an efficient way, using the model-distilled knowledge about learning behavior that impairs the generalization and robustness of models, rather than directly adding variations to the images. Our contributions are:

- We propose a novel augmentation method called DFM-X to improve the generalization and robustness of models against common corruptions and adversarial attacks without sacrificing their performance on the original test images.
- DFM-X exploits model-distilled prior knowledge from data about frequency shortcuts, targeting the mitigation of destructive learning behavior which impairs generalization, unlike commonly used augmentation techniques that focus on increasing data variety directly but rarely consider implicit problems in the data.

2. Related works

We review existing research related to shortcut learning mitigation and data augmentation in the frequency domain.

Shortcut learning mitigation. Avoiding learning shortcuts in the data is a promising approach to improve generalization and corruption robustness by encouraging models to learn more meaningful task-related semantics. Existing work has explored different strategies to address shortcut learning and its impact on model performance.

One approach is to explicitly identify and manipulate shortcuts present in the data. The authors in [15] identified shortcuts (i.e. color patches) in a class and induced similar patches in the other class. This forces the models to ignore spurious correlations between the color patches and the class, thus effectively mitigating the influence of shortcuts. Another line of research focuses on addressing shortcut learning without explicitly identifying shortcuts. The work of [16] proposed a regularization term that decouples feature learning dynamics, allowing the networks to learn from as many features as possible rather than a subset of features that easily minimizes cross-entropy loss. Similarly, [6] used an auxiliary network with low capacity to measure the degree of shortcut information in images, because image classes containing shortcuts are easier to learn in early training stages and a low-capacity network is more prone to shortcut learning than a high-capacity one. Leveraging this, the target network with high capacity can selectively learn less from images with high shortcut degrees. Other methods use gradient-based scores [1] to measure the shortcut degree of training samples or adversarial training [14, 17] to learn image representations containing less shortcut information.

Existing methods mainly focus on mitigating learning shortcut features that are visually observable. Our work, instead, aims at the mitigation of shortcut implicit in the data from a frequency perspective. We exploit the learned frequency shortcuts as prior knowledge of unwanted learning behaviors, and learn to avoid them by using the proposed DFM-X augmentation strategy in the training.

Frequency-based data augmentation. Data augmentations applied to images are usually spatial transformations, such as flipping, rotation, and cropping. These are commonly used in augmentation techniques such as AugMix [10], AutoAugment [5], AugMax [20], among others. Inspired by the research analyzing the learning behavior in the frequency domain of neural networks (NNs), there is a trend in developing frequency-based augmentation techniques. Chen *et al.* [4] discovered that enforcing NNs to learn more from the phase spectrum than the amplitude spectrum can improve model robustness toward common corruptions. Xu *el al.* [25] proposed amplitude-mixed augmentation, where NNs are trained with phase-invariant images with fused amplitude spectrum because the phase information is considered to be robust to domain change.

Rather than mixing frequency information, the work of [11] drops frequency components of images if their discrete cosine transformation coefficients are below a randomly selected threshold. Inspired by the work [27] which demonstrates how noise consisting of different frequency affect classification performance, Soklaski *et al.* [19] added Fourier-basis noise to the operation candidate pool in the AugMix framework [10]. The work in [24, 26] proposed Fourier domain adaptation for segmentation tasks and deep metric learning respectively, which replaces the low frequency of target images with that of source images. As low frequency contains shape information, the annotations of the source images are used as ground truth for training.

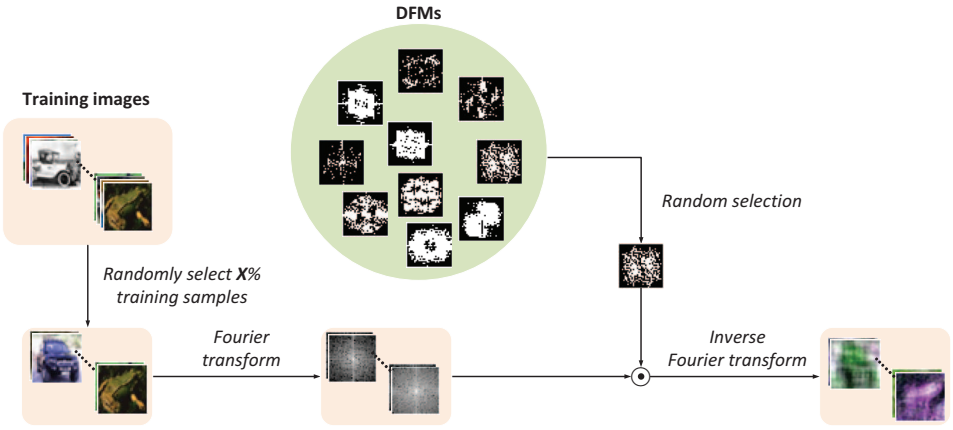


Figure 2: The scheme of DFM-X augmentation. For each training epoch, X% training images are randomly selected for augmentation. A randomly chosen DFM serves as a mask to filter the Fourier spectrum of the images. If the DFM and the image belong to the same class, the image is not processed. Thus, the images of class i are filtered with the DFM of class k ($i \neq k$). This reduces the specificity of the frequency sets to the classification of the corresponding classes, thus mitigating frequency shortcut learning.

Most augmentation approaches focus on increasing data variety to bridge the distribution gap between training and testing data, or enforcing specific characteristics that benefit the performance. However, they do not consider shortcuts in the data during training. We develop an augmentation strategy based on prior knowledge about frequency shortcuts that we gain by analyzing models trained for image classification. We devise a form of augmentation, in which the models are induced to exploit a larger amount of frequency components and avoid learning shortcut solutions, thus improving model robustness against common corruptions and adversarial attacks.

3. Methods

As discovered in [21], convolutional neural networks can use small, specific sets of frequencies, i.e. frequency shortcuts, to classify images of certain classes. Because shortcuts harm the generalization of models, we aim to develop an augmentation technique, to improve model robustness and generalization performance by mitigating the learning of frequency shortcuts (prior knowledge distilled from data). We achieve this by reducing the reliance of models on specific frequency sets for the classification of shortcut-affected classes. The models thus rely on a wider range of frequencies to classify images and are induced to learn more semantics. We further evaluate the benefits on corruption robustness and adversarial robustness of models.

3.1. DFM-X augmentation

CNNs can be biased toward specific sets of frequencies to achieve classification [21]. Our goal is to reduce such bias and enforce the models to learn more semantics, by inducing the use of larger sets of frequencies. Hereby, we design a DFM-based augmentation technique.

Obtaining DFMs: model-distilled prior knowledge. DFMs record the importance of each frequency to the classification of a certain class. They can carry knowledge of shortcuts in the data which are learned by a model. We use them as priors in our augmentation approach to avoid unwanted shortcut learning behavior, exploiting data more efficiently and resulting in robust models.

The algorithm in [21] computes DFMs by evaluating the importance of frequencies from images based on the degradation of classification performance. To compute the DFM of a certain class, they iteratively remove an individual frequency from the Fourier spectrum of images of the class in the test dataset, and measure the loss in classification. If the degradation is above a certain threshold, the frequency is considered important to the classification of the class and it is kept in the Fourier spectrum of images for the following iterations. Otherwise, less important frequencies are removed. We limit the performance degradation to be within 30% when the models classify the images of the class retaining only the dominant frequencies, compared to the standard performance. One can obtain DFMs after the importance of each frequency of the images of the corresponding classes is measured, which are in the form of binary masks demonstrating the specific sets of frequencies possibly used as shortcuts for classification (see examples of DFMs in Figure 2). Through leveraging the information contained in the DFMs, we guide the learning behavior of models, aiming to reduce their reliance on specific sets of

frequencies associated with shortcut learning.

Augmentation strategy. We show the schematic of our augmentation strategy in Figure 2. Given a dataset containing images $\{x_m^c\}$ where c is the class of the m^{th} image in the dataset, we compute the DFM of each class for a model f. We use the DFMs as priors to guide the training of new models, as they can carry information about unwanted learning behavior. We randomly select X% training images to be augmented. This helps to control the impact of augmentations by adjusting the number of images being augmented. The selected images are augmented by retaining the dominant frequencies of other classes. That is, we use the DFMs as masks to filter the Fourier spectrum of the images:

$$\hat{x}_m^i = \mathcal{F}^{-1}[\mathcal{F}[x_m^i] \odot DFM^k] \;\; (i \neq k),$$

where x_m^i is the m^{th} image in the dataset of class i, DFM^k is the dominant frequency map of a randomly selected class k, $\mathcal{F}$ and $\mathcal{F}^{-1}$ indicate the Fourier transform and the inverse transform, and $\hat{x}_m^i$ is the augmented version being filtered by the DFM. We apply element-wise multiplication of the Fourier spectrum of x_m^i and DFM^k (the process of filtering), and obtain the augmented image $\hat{x}_m^i$ through computing the inverse Fourier transform of the filtered Fourier spectrum of x_m^i. Note that, i is not equal to k. This enforces that the models learn from the dominant frequencies of class k to classify the images of class i. To sum up, DFM-X augmentation mitigates frequency shortcut learning by highlighting features or visual cues across the whole dataset that are originally specific for certain classes.

3.2. Evaluation of corruption robustness

Common image corruptions are visual transformations applied to images and might affect the ability of models to extract semantic features (e.g. Gaussian noise and defocus blur [9]), thus negatively influencing model robustness. We utilize the mean corruption error (mCE) and the relative corruption error (rCE) to evaluate the corruption robustness of models on datasets containing sub-datasets, e.g. CIFAR-C that are corrupted by one corruption [9], computed as:

$$\mathrm{mCE} = \frac{1}{|C|} \sum_{c \in C} \frac{\sum_{s=1}^{5} \mathrm{CE}_{s,c}^{f}}{\sum_{s=1}^{5} \mathrm{CE}_{s,c}^{baseline}},$$

$$\mathrm{rCE} = \frac{1}{|C|} \sum_{c \in C} \frac{\sum_{s=1}^{5} (E_{s,c}^{f} - E_{clean}^{f})}{\sum_{s=1}^{5} (E_{s,c}^{baseline} - E_{clean}^{baseline})},$$

where $\mathrm{CE}_{s,c}^{f}$ is the classification error of model f on a test set corrupted by c (e.g. defocus blur and shot noise) with severity $s \in \{1, 2, 3, 4, 5\}$. The higher the severity, the more influence the corruption effect has on the images. C is the set of corruptions in the entire test set and $baseline$ is the baseline model for comparison. The mCE measures the relative classification performance of a model normalized by that of the baseline. The rCE additionally measures the performance degradation of model f on corrupted images w.r.t. their clean version. When mCE and rCE are less than one, this indicates that model f is more robust than the baseline, as it has less classification error in general. Additionally, we use standard accuracy (SA) to evaluate the performance of models on the original test dataset. The robust accuracy (RA), instead, is the average accuracy of the models tested on the corrupted versions of the test set.

3.3. Evaluation of adversarial robustness

We evaluate the accuracy of models under FGSM and PGD attacks. These attacks usually have a bias toward high-frequency. As shown in [3], inducing low-frequency bias to models during training can improve adversarial robustness. Differently, our augmentation approach enforces models to learn from a wider range of frequencies with the model-distilled prior knowledge from data, which might be beneficial for adversarial robustness. We use L_∞-norm bounded perturbation ϵ ranging from $1/255$ to $10/255$. For the PGD attack, we use 10 steps and set the step size $2.5\epsilon/10$ to ensure that the boundary of the ϵ-ball is reached.

4. Experiments and results

4.1. Datasets

We use CIFAR-10 [12], which contains 10 classes of 50000 training images and 10000 testing images. For the evaluation of corruption robustness, we use its corrupted variant CIFAR-C [9], which includes 19 corrupted subsets. The 19 corruptions are categorized into four groups, including noise (Gaussian, impulse, shot, speckle), blur (defocus, glass, Gaussian, motion, zoom), weather (brightness, fog, frost, snow, spatter), and digital transformation (contrast, elastic, JPEG compression, pixelate, saturate). For each corruption, there are five levels of severity. High severity indicates a high impact of corruption on images.

4.2. Training setup

We train ResNets for 200 epochs on the CIFAR-10 dataset. The initial learning rate is 0.01, reduced by a factor of 10 if the validation loss does not decrease for 10 epochs. We use batch size 64 and an SGD optimizer with momentum 0.9 and weight decay 10^{-4}. Note that, low-capacity models are more prone to shortcut learning than high-capacity models [6] and shortcuts in the data are supposed to be architecture-agnostic. Thus, we compute the DFMs of ResNet18, a relatively low-capacity model, in DFM-X augmentation. To compare DFM-X with other commonly used augmentation techniques, we train mod-

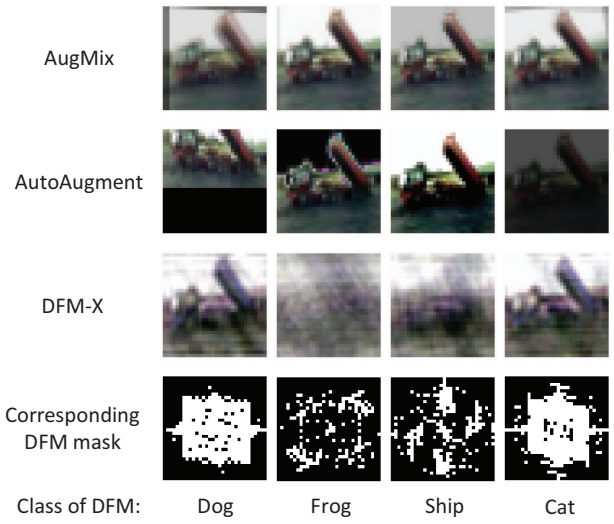


Figure 3: An image of truck is augmented by AugMix (the first row), AutoAugment (the second row), and DFM-X (the third row). The fourth row demonstrates the corresponding DFM used to obtain the images in the third row.

Model	SA	RA	mCE (%)	rCE (%)
ResNet18	92.15	77.49	100	100
+ DFM-30	92.11	80.43	87.94	**75.92**
+ DFM-50	92.10	**81.4**	86.61	78.73
+ DFM-70	**92.27**	81.2	**85.36**	76.74
+ AugMix	**93.39**	83.47	**76.42**	72.71
+ AugMix + DFM-30	91.44	83.65	80.68	58.18
+ AugMix + DFM-50	91.99	**84.56**	77.33	62.58
+ AugMix + DFM-70	91.08	84.2	80.3	**57.19**
+ AutoAugment	**93.47**	82.78	75.81	65.86
+ AutoAugment + DFM-30	92.43	83.29	78.28	64.72
+ AutoAugment + DFM-50	92.72	**84.37**	**73.23**	59.48
+ AutoAugment + DFM-70	91.39	83.65	79.97	**59.45**
ResNet34	93.02	79.84	90.16	93.34
+ DFM-30	**93.75**	**82.32**	**78.65**	80.76
+ DFM-50	93.51	81.57	80.59	79.07
+ DFM-70	92.7	81.93	81.7	**72.26**
+ AugMix	92.18	83.71	79.69	66.63
+ AugMix + DFM-30	**93.49**	**86.42**	**65.81**	**54.1**
+ AugMix + DFM-50	92.74	85.77	70.06	**54.1**
+ AugMix + DFM-70	93.28	85.45	69.27	58.14
+ AutoAugment	**94.08**	83.66	72.97	70.32
+ AutoAugment + DFM-30	93.71	85.72	**66.38**	57.58
+ AutoAugment + DFM-50	93.46	85.55	67.42	**56.36**
+ AutoAugment + DFM-70	93.35	**85.96**	67.53	57.06

Table 1: Performance of ResNets on CIFAR-10 and CIFAR-C. The best values of each group of models are in bold and the best values for ResNet18 and ResNet34 are underlined.

els with AugMix or AutoAugment. Example images augmented by AugMix, AutoAugment and the proposed DFM-X are shown in Figure 3.

4.3. Robustness against common corruption

We report the results of the models trained with different augmentation strategies in Table 1. The best values of models with the same architecture trained with namely DFM-X, AugMix + DFM-X and AutoAugment + DFM-X, are highlighted in bold, and the best values for ResNet18 and ResNet34 are underlined.

DFM-X benefits corruption robustness. ResNets trained with DFM-X augmentation are more robust against common corruptions than ResNets trained without DFM-X. They have higher or comparable standard accuracy than ResNets trained without DFM-X, as well as higher robust accuracy. This indicates that DFM-X benefits the robustness of models to common corruptions without impairing their performance on the clean dataset. We conjecture that DFM-X enforces models to learn from a wider range of frequencies with the prior knowledge provided, and thus more meaningful and task-related semantics is used by the models, benefiting their corruption robustness.

Comparison with existing augmentations. We compare DFM-X to existing and largely-used augmentation techniques like AugMix [10] and AutoAugment [5]. The models trained with DFM-X, AugMix, and AutoAugment have similar SA and RA, while the model trained with DFM-X has higher mCE than the models trained with AugMix or AutoAugment. The rCE of models trained with DFM-X is also higher than that of models trained with AutoAugment or AugMix. We attribute this to the fact that other augmentation techniques focus on increasing data variety to reduce the distribution gap between training and testing data, and may use augmentations that are visually similar to the corruptions in CIFAR-C. Our approach, instead, focuses on exploiting as much information as possible from the clean training data without additionally overlaying corruption-like variations, learning more meaningful and task-related semantics. We thus investigate the effectiveness of combining DFM-X with another augmentation technique, as they augment images differently (DFM-X exploits data efficiently while the others increase data variety by adding corruption-like variations).

Boosted robustness with a complementary technique. We apply DFM-X together with either AugMix or AutoAugment during training, and observe that this contributes to further improving corruption robustness (Table 1). For example, ResNet18 trained with AugMix/AutoAugment and DFM-50 augmentations generally achieve higher robust accuracy than the models trained with only AugMix or AutoAugment. Moreover, they have lower or comparable values of mCE and rCE, compared to

those trained with AugMix or AutoAugment only. When ResNet34 is trained with DFM-30 and AugMix or AutoAugment augmentations, the models demonstrate more robustness than those trained with only one of the augmentation techniques. This indicates that combining augmentation techniques that are complementary can further benefit corruption robustness. For the future design of augmentation techniques, we should focus on inspecting data itself and exploiting it more efficiently, rather than directly increasing the variety of data by adding visual variations.

Intriguingly, when ResNet18 and ResNet34 are trained with AugMix, both models show similar corruption robustness, though ResNet34 has a larger model capacity than ResNet18. ResNet34 + AugMix even has a slightly worse mCE than ResNet18 + AugMix. However, when models are trained with AugMix + DFM-X, ResNet18s have worse mCEs than that trained solely with AugMix. ResNet34 trained with AugMix + DFM-30, instead, demonstrates significantly improved corruption robustness (mCE equal to 65.81). As low-capacity models are more prone to shortcut learning than high-capacity models, DFM-X together with AugMix might not be enough to mitigate shortcut learning in ResNet18 but benefits the robustness of ResNet34. We conjecture that for low-capacity models, there needs extra regularization to overcome shortcuts. Moreover, AugMix results in displacement effects that are visually similar to those in some DFM-augmented images (see Figure 3). Thus, there is an partial overlap in the augmentation effects. When combining DFM-X with other augmentation techniques, it is important to consider what kinds of operation are appropriate and complementary to DFM-X.

The choice of X. Our results show that the percentage of training images subject to augmentation via DFM-X does not influence significantly on the corruption robustness when the models are trained with DFM-X solely (they have close RA, mCE and rCE). However, when DFM-X is incorporated with AugMix or AutoAugment, models with a different capacity might perform better as a different value of X is chosen. For instance, ResNet18 prefers DFM-50 while ResNet34 prefers DFM-30. We attribute this to model capacity. As a relatively low-capacity model, ResNet18 suffers more from shortcut learning than ResNet34, and thus it needs more regularization during training. The more images are augmented, the more regularization is imposed on the training process.

Robustness against different corruption types. In Table 2, we present a detailed overview of the robustness of ResNet18 trained with different augmentation techniques against the four corruption categories in CIFAR-C, namely noise, blur, weather conditions, and digital transformation. The best results of ResNet18 trained with DFM-X, AugMix + DFM-X and AutoAugment + DFM-X are highlighted in bold respectively. Observed from values in bold, models trained with DFM-X demonstrate better robustness against all types of corruption than the models trained without DFM-X. Among the four corruption types, the improvement in the robustness toward blur corruption is relatively lower than that of the other three types. As demonstrated in [27], blur corruptions, such as defocus blur and Gaussian blur, have energy highly concentrating on middle-high frequencies. Our augmentation technique enforces models to look into a wider range of frequencies for classification, and thus, the models are relatively less robust to corruptions having a specific energy concentration in the Fourier spectrum than those having a rather even energy distribution over the spectrum, e.g. Gaussian and shot noise.

Model	Noise	Blur	Weather	Digital
ResNet18	100	100	100	100
+ DFM-30	88.75	91.6	85.2	86.6
+ DFM-50	**78**	94.6	**84.8**	87.2
+ DFM-70	86.5	**91**	84	**79.7**
+ AugMix	62	**86.2**	**75.2**	**79.8**
+ AugMix + DFM-30	53.5	92.2	85.6	86.2
+ AugMix + DFM-50	50.75	91	79.6	82.6
+ AugMix + DFM-70	**50.5**	95.4	83.8	85.8
+ AutoAugment	72.75	**85.4**	**63.8**	80.8
+ AutoAugment + DFM-30	62	95.8	73.4	78.2
+ AutoAugment + DFM-50	61.75	89	68.6	**71**
+ AutoAugment + DFM-70	**61.5**	97.8	77.4	79.8

Table 2: The corruption error (CE) (%) of ResNet18s trained with different augmentation techniques on each corruption type (Baseline: ResNet18). The best values in each group are highlighted in bold.

4.4. Robustness against adversarial attacks

We evaluate the adversarial robustness of the models trained with different augmentations under FGSM and PGD attacks. We select ResNet18 + AugMix/AutoAugment + DFM-50 and ResNet34 + AugMix/AutoAugment + DFM-30, as models with a different capacity prefer a different percentage of images to be augmented. We report the classification accuracy of the models under the FGSM and PGD attacks in Tables 3 and 4, respectively.

Learning from more frequencies improves robustness. We observe that ResNets trained with DFM-X show improved adversarial robustness to the FGSM and PGD attacks. Unlike PGD and FGSM adversarial training which sacrifices performance on natural images [13], DFM-X maintains high model performance on the original test set while improving the robustness of models to both FGSM and PGD attacks. We attribute this to the wider learning range of frequencies, when compared to models trained with standard setups. DFM-X augments images based on

Model	L_∞-norm bounded perturbation of size ϵ									
	1/255	2/255	3/255	4/255	5/255	6/255	7/255	8/255	9/255	10/255
ResNet18	86.47	75.66	64.03	54.3	46.46	40.05	35.04	31.83	28.92	26.47
+ DFM-50	87.71	78.97	69.43	60.7	53.05	46.55	41.04	37.18	34.09	31.61
+ AugMix	88.02	79.27	69.28	60.38	52.52	46.03	41.17	36.93	33.57	30.69
+ AugMix + DFM-50	**88.29**	**79.97**	**71.07**	**62.41**	**55.23**	**48.75**	43.38	38.92	35.03	31.67
+ AutoAugment	87.72	76.13	65.26	55.7	48.99	43.85	39.91	36.64	33.97	32.15
+ AutoAug + DFM-50	87.7	78.51	69.03	60.68	54.03	48.33	**43.94**	**40.66**	**37.8**	**35.43**
ResNet34	87.7	77.53	67.61	58.12	50.71	45.21	40.32	36.69	33.52	30.91
+ DFM-30	88.45	79.17	70.48	63.14	57.32	52.96	48.88	46.03	43.72	41.81
+ AugMix	88.71	79.56	70.14	60.99	52.62	45.78	40.12	35.58	31.85	28.72
+ AugMix + DFM-30	88.8	81.16	72.23	63.34	55.37	48.41	42.17	37.45	33.55	30.35
+ AutoAugment	87.39	76.76	65.98	57.06	50.48	45.3	41.04	38.05	35.55	33.68
+AutoAugment + DFM-30	**89.23**	**81.96**	**74.92**	**68.34**	**63.01**	**58.81**	**54.93**	**51.64**	**49.13**	**46.94**

Table 3: Accuracy (%) of models under FGSM attack. The best values of ResNet18 and ResNet34 are highlighted in bold.

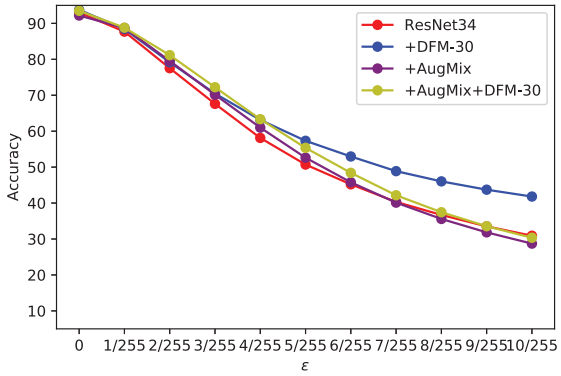
Model	L_∞-norm bounded perturbation of size ϵ									
	1/255	2/255	3/255	4/255	5/255	6/255	7/255	8/255	9/255	10/255
ResNet18	85.39	70.79	53.5	38.43	26.09	18.13	12.98	9.95	7.98	6.96
+ DFM-50	86.93	75.4	60.63	45.88	33.61	24.8	18.51	14.15	11.16	9.09
+ AugMix	87.36	74.91	58.75	43.35	30.54	21.49	15.52	11.57	9.07	7.41
+ AugMix +DFM-50	**87.71**	**76.86**	**63.81**	**51.01**	**38.22**	**29.99**	**22.95**	**17.3**	**13.63**	**11.14**
+ AutoAugment	86.08	67.37	46.96	32.17	21.48	15.03	11.29	8.63	7.01	6.1
+ AutoAugment + DFM-50	86.45	71.45	54.86	40.13	28.75	20.79	15.79	12.27	9.89	8.15
ResNet34	86.85	73.35	57.26	42.31	30.72	21.72	15.42	11.55	8.74	7.38
+ DFM-30	87.94	75.03	58.78	44.29	33.2	25.45	20.16	16.35	13.87	11.72
+ AugMix	88.18	76.39	62.5	48.74	36.88	27.59	20.52	15.6	12.17	9.8
+ AugMix + DFM-30	88.25	**78.38**	**65.2**	**51.62**	**39.6**	**29.55**	**22.58**	**17.35**	13.69	10.93
+ AutoAugment	86.05	68.98	49.95	34.35	24.14	17.62	13.11	10.42	8.69	7.39
+ AutoAugment + DFM-30	**89.41**	76.97	61.64	46.6	35.1	26.33	20.89	16.94	**14.32**	**12.35**

Table 4: Accuracy (%) of models under PGD attack. The best values of ResNet18 and ResNet34 are highlighted in bold.

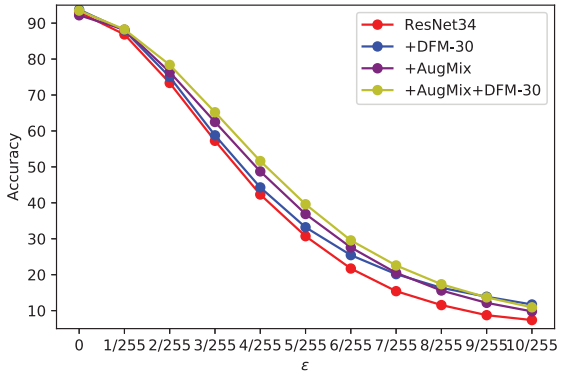
the prior knowledge of the reliance of models on specific frequencies, with the aim of reducing it during training. Thus, the frequency bias of models associated with their vulnerability to adversarial noise is reduced, benefiting adversarial robustness.

DFM-X vs. AugMix. In Figure 4 we compare the results of ResNet34 trained without augmentation or with DFM-30, AugMix, and AugMix + DFM-30. We observe that the model trained with AugMix (see the purple line) demonstrates less robustness to FGSM attack, compared to the model trained with DFM-30 (see the blue line). Regarding the PGD attack, both models show comparable robustness. Combining AugMix with DFM-30 does not obtain more robustness to the FGSM attack when ϵ becomes large, but the model is robust to the PGD attack. From Table 3, ResNet18 trained with AugMix shows similar robustness to the FGSM attack to the one trained with DFM-50, but it is less robust to the PGD attack. We conjecture that AugMix, resulting in images with similar displacement effects to those augmented by DFM-X (see Figure 3), might indirectly augment the frequency information of images like DFM-X. But it is less effective than DFM-X, which employs model-distilled prior knowledge from the data in augmentation, rather than randomly adding augmentations to images. Combining DFM-X with AugMix avoids unwanted learning behavior and increases data variety, thus obtaining more adversarial robustness.

Reduced negative impact of AutoAugment. We demonstrate in Figure 5 that ResNet34 trained with DFM-30 (see the blue line) has better adversarial robustness than ResNet34 and ResNet34+AutoAugment. Interestingly, training solely with AutoAugment impairs adversarial robustness slightly to FGSM attack and significantly to PGD attack (see green lines in Figure 5). When combining DFM-30 and AutoAugment (see the orange line), the model gains more robustness to the FGSM attack, compared with the one trained only with DFM-30. We observe from Tables 3 and 4 that AutoAugment also impairs the adversarial robustness of ResNet18 to the FGSM and PGD attacks. Training the model with AutoAugment and DFM-X augmentations benefits the adversarial robustness of the models. Although AutoAugment alone harms the robustness, using

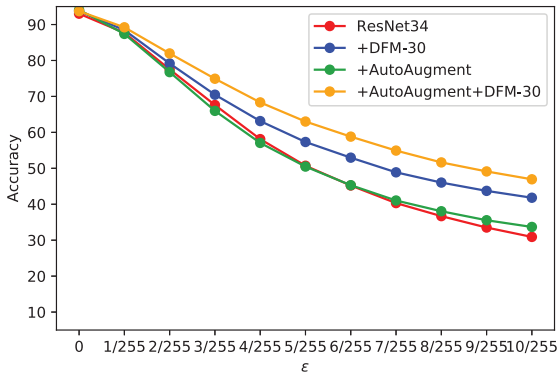


(a)

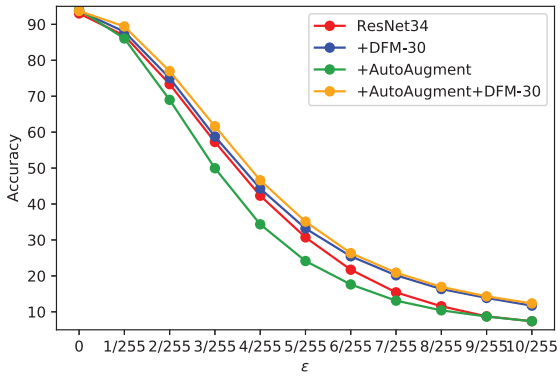


(b)

Figure 4: ResNet34 trained without augmentation or with DFM-30, AugMix, or AugMix + DFM-30 under (a) FGSM attack and (b) PGD attack.



(a)



(b)

Figure 5: ResNet34 trained without augmentation or with DFM-30, AutoAugment, or AutoAugment + DFM-30 under (a) FGSM attack and (b) PGD attack.

DFM-X avoids much performance degradation under the attacks. From the augmented images in Figure 3, DFM-X results in different variations from those augmented by AutoAugment. We attribute the models having better adversarial robustness than those trained with only one of them to the complementarity between DFM-X and AutoAugment in terms of augmentation effects.

5. Conclusions

We propose DFM-X, an augmentation approach that leverages prior knowledge about frequency shortcuts. Motivated by shortcut mitigation, our method aims at avoiding unwanted shortcut solutions by enforcing models to learn from a wider range of frequencies and thus more semantics. DFM-X exploits data efficiently, as it targets implicit problems in the data that might impair the generalization and robustness of models, unlike other commonly used augmentation techniques focusing on increasing data variety by adding visual variations. Our experimental results show that DFM-X enhances model robustness against common corruptions and adversarial attacks without sacrificing the standard performance on the original test set. Combining DFM-X with other commonly used augmentation techniques, e.g. AugMix and AutoAugment, gains more robustness than using only one of them. DFM-X compensates for the weakness of AutoAugment in impairing adversarial robustness. We observe that the complementarity between augmentation techniques is important to model performance. Distilling prior knowledge about destructive learning behavior from data helps exploit data more efficiently. We suggest future research on designing augmentation strategies that consider data characteristics instead of directly increasing the visual variations of images to bridge the distribution gap between training and testing data.

Acknowledgements. This work was supported by the SEARCH project, UT Theme Call 2020, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente.

References

[1] Sumyeong Ahn, Seongyoon Kim, and Se-Young Yun. Mitigating dataset bias by using per-sample gradient. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. 2

[2] Kaoutar Ben Ahmed, Lawrence O. Hall, Dmitry B. Goldgof, and Ryan Fogarty. Achieving multisite generalization for cnn-based disease diagnosis models by mitigating shortcut learning. *IEEE Access*, 10:78726–78738, 2022. 1

[3] Alvin Chan, Yew-Soon Ong, and Clement Tan. How does frequency bias affect the robustness of neural image classifiers against common corruption and adversarial perturbations?, 2022. 4

[4] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 458–467, October 2021. 2

[5] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data, 2019. 1, 2, 5

[6] Nikolay Dagaev, Brett D. Roads, Xiaoliang Luo, Daniel N. Barry, Kaustubh R. Patil, and Bradley C. Love. A too-good-to-be-true prior to reduce shortcut reliance, 2021. 2, 4

[7] Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. Towards interpreting and mitigating shortcut learning behavior of nlu models, 2021. 1

[8] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, nov 2020. 1

[9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 4

[10] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty, 2020. 1, 2, 5

[11] Md Tahmid Hossain, Shyh Wei Teng, Dengsheng Zhang, Suryani Lim, and Guojun Lu. Distortion robust image classification using deep convolutional neural network with discrete cosine transform. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 659–663, 2019. 2

[12] Alex Krizhevsky. Learning multiple layers of features from tiny images, Apr 2009. 4

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. 6

[14] Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic shortcut removal for self-supervised representation learning, 2020. 1, 2

[15] Meike Nauta, Ricky Walsh, Adam Dubowski, and Christin Seifert. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, 12(1), 2022. 1, 2

[16] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 1, 2

[17] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions?, 2021. 1, 2

[18] Piyapat Saranrittichai, Chaithanya Kumar Mummadi, Claudia Blaiotta, Mauricio Munoz, and Volker Fischer. Overcoming shortcut learning in a target domain by generalizing basic visual factors from a source domain. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 294–309, Cham, 2022. Springer Nature Switzerland. 1

[19] Ryan Soklaski, Michael Yee, and Theodoros Tsiligkaridis. Fourier-based augmentations for improved robustness and uncertainty calibration. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. 2

[20] Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training, 2022. 2

[21] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. Frequency shortcut learning in neural networks. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. 1, 2, 3

[22] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. Larger is not better: A survey on the robustness of computer vision models against common corruptions. 2023. 1

[23] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. What do neural networks learn in image classification? a frequency shortcut perspective. In *International Conference on Computer Vision (ICCV)*, 2023. 1

[24] Zheng Wang, Zhenwei Gao, Guoqing Wang, Yang Yang, and Heng Tao Shen. Visual embedding augmentation in fourier domain for deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 2

[25] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14383–14392, June 2021. 2

[26] 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 (CVPR)*, June 2020. 2

[27] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, ed-

itors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2, 6