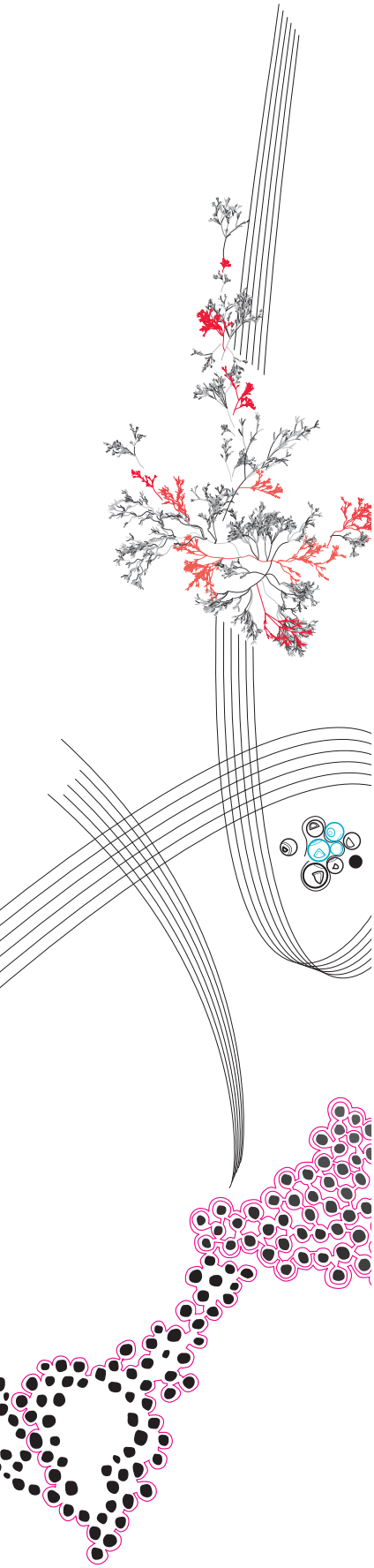


MSc Applied Mathematics
Final Project

Fourier Insights in Machine Learning: Bridging the Augmentation Gap through Frequency-basis Functions

Puru Vaish

Graduation Committee:
Dr. Nicola Strisciuglio (UT)
Dr. Sophie Langer (UT)
Dr. Marcello Carioni (UT)
Shunxin Wang, MSc (UT)
Chair:
Prof. Dr. Christoph Brune (UT)

February, 2024

Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

Fourier Insights in Machine Learning: Bridging the Augmentation Gap through Frequency-basis Functions

Puru Vaish
University of Twente
Enschede, Netherlands
p.vaish@student.utwente.nl, puruvaish24@gmail.com

Abstract

For neural networks, challenges arise when deploying models in real-world scenarios, as unforeseen changes in inputs can lead to diminished performance. While data augmentation is a common remedy to bridge the gap between training and test data, its efficacy in enhancing the robustness of computer vision models is not guaranteed. This paper introduces Auxiliary Fourier-basis Augmentation (AFA), a novel approach that extends beyond visual augmentations to address this limitation by focusing on neural networks.

AFA leverages Fourier-basis additive noise as a complementary technique in the frequency domain, filling the robustness gap left by conventional visual augmentations. Our method demonstrates its effectiveness in an adversarial setting, showcasing its utility in enhancing model robustness. Notably, AFA contributes to reducing the impact of common corruptions, facilitates out-of-distribution (OOD) generalisation, and ensures consistent model performance against increasing perturbations. Importantly, it introduces a unique capability to minimise frequency shortcuts, further fortifying the overall resilience of neural network models.

The results affirm that AFA seamlessly integrates with existing augmentation techniques, providing a comprehensive enhancement to model performance. This work presents a valuable contribution to the broader pursuit of robust neural networks, extending beyond the conventional focus on computer vision models.

1. Introduction

In real-world deployments, computer vision models commonly experience diminished performance owing to unanticipated variations in images. Enhancing the resilience of computer vision models to out-of-distribution (OOD) data becomes imperative for ensuring their dependable functionality in practical applications. In the realm of enhancing the robustness and generalization of computer vision models, various methods have been explored [2, 7, 8, 10, 46, 49, 52]. Data augmentation emerges as a widely adopted strategy due to its user-friendly application and effectiveness in minimizing the distribution gap between training and test data [45]. Noteworthy augmentation techniques, including AugMix [16], AugMax [42], AutoAugment [3], TrivialAugment [35], and PRIME [34], have demonstrated substantial advancements in benchmarks for corruption and perturbation robustness, as well as out-of-distribution (OOD) datasets for generalization, such as ImageNet-C, ImageNet-$\bar{\text{C}}$, ImageNet-3DCC, ImageNet-P, ImageNet-R, and ImageNet-v2 [14, 15, 19, 33, 37]. These approaches predominantly concentrate on introducing visual variations to images through either random or policy-based combinations [3, 16, 17, 27, 28, 31, 32, 35] of visual transformations. This aims to augment the diversity of training images by expanding their domain, as illustrated in Fig. 1. Additionally, adversarial-based augmentations address the difficulty of training samples, albeit at a higher computational cost, as depicted in Tab. 1 for AugMax.

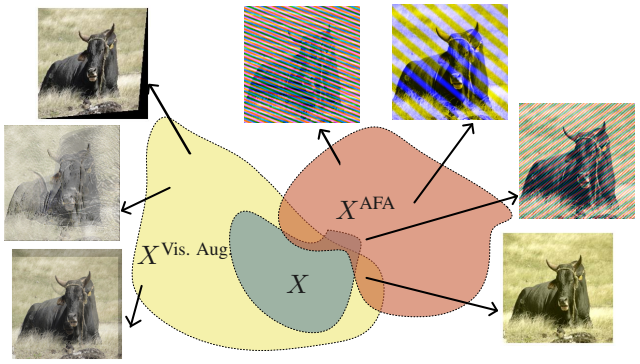


Figure 1. Frequency augmentation with Fourier-basis functions is complementary to common visual augmentations. They appear *unnatural* and can be used as adversarial examples.

Despite being trained with visual augmentations, models remain susceptible to image variations not accounted for during training [26] and frequency perturbations [48]. This vulnerability arises from the predefined frequency characteristics of visual transformations, which fail to guarantee comprehensive model robustness against noise exhibiting different frequency characteristics than those encountered in the training data. This gap in frequency robustness can be exploited by attackers, leading to potential performance degradation in operational settings [24].

This raises a question: *Is there a complementary augmentation technique that can bridge the gap left by visual augmentations?*

Traditional visual augmentations concurrently impact various frequency components in images, making explicit control challenging and potentially missing certain frequency variations present in unforeseen corruptions or real-world scenarios [38]. In response, we propose a reevaluation of image augmentation by delving into the frequency domain. Our approach complements visual augmentation strategies by incorporating Fourier basis functions in an adversarial setting. The exploration of frequency-based augmentations aims to unlock capabilities beyond the reach of traditional visual augmentations.

Researchers have investigated diverse frequency-based augmentation techniques to broaden the scope of augmentation capabilities. For instance, studies such as [1, 41, 47] involve swapping or mixing partial amplitude spectra between images to enhance phase-reliance for classification. In another approach, [43] augments images with shortcut features to reduce their specificity for classification. AugSVF [39] introduces frequency noise within the AugMix framework, while [25, 29] adversarially perturb the frequency components of images. It is important to note that these augmentations, while offering enhanced capabilities, often come with computational complexity. This complexity arises from intricate augmentation frameworks [39], the computation of multiple Fourier transforms for training images and their augmented versions [1, 41, 47], the identification of learned frequency shortcuts [43], or the adoption of adversarial training strategies [25, 29].

This research introduces the concept of Auxiliary Fourier-basis Augmentation (AFA), employing additive noise based on Fourier-basis functions to efficiently augment the frequency spectrum. AFA's approach stands out for its effectiveness and computational efficiency compared to other methods utilizing frequency manipulations [1, 39, 43].

The impact of additive Fourier-basis functions on image appearance is distinct and orthoginal to conventional augmentations, as illustrated in Fig. 1. These images serve as representative samples of an adversarial distribution, deviating from those augmented through typical visual transformations. This work expands on the conventional notion of adversarial augmentation, transcending the generation of imperceptible noise via gradient back-propagation.

Our proposed training architecture and strategy incorporate an auxiliary component to address the adversarial distribution, alongside a main component for the original distribution, following a similar paradigm to AugMax [42]. Notably, the adversarial distribution created by additive Fourier-basis is significantly less computationally expensive compared to AugMax and adds minimal additional burden to other visual augmentation methods when used as complimentary (refer to Tab. 1). This approach yields comparable or superior generalisation results while enabling the training of larger models on more extensive datasets, such as ImageNet.

1.1. Contributions

In this work, we present two key contributions aimed at enhancing the robustness of computer vision models in real-world scenarios. Firstly, we introduce a novel augmentation technique named Auxiliary Fourier-basis Augmentation (AFA). This technique, designed for straightforward implementation and computational efficiency, proves to be highly effective in improving model robustness against common image corruptions. Through a series of experiments, we demonstrate that AFA not only enhances resistance to visual perturbations but also significantly contributes to out-of-distribution (OOD) generalisation. Moreover, AFA showcases a notable capability in maintaining prediction consistency in the face of various perturbations, thus addressing a crucial aspect of model reliability in dynamic environments.

Secondly, we extend the existing augmentation space by introducing amplitude- and phase-adjustable frequency noise, a distinctive feature of AFA. By not limiting this frequency-based augmentation to be visually palatable, extending it into the realm of adversarial examples, we successfully reduce the augmentation gap associated with common visual augmentations. This expansion of the augmentation space provides a more comprehensive and complementary approach to traditional visual augmentations, further fortifying the model against unforeseen variations in input data. The proposed methodology not only improves the overall robustness of computer vision models but also sets the stage for a broader exploration of frequency-driven perspectives and extreme image augmentations in the realm of data augmentation.

In summary, our contributions encompass the introduction of AFA as a practical and efficient augmentation technique, showcasing its effectiveness in bolstering model robustness against image corruptions, OOD scenarios, and perturbations. Additionally, we expand the augmentation space through frequency-based adjustments, revealing a

	APR-SP	AFA (ours) w/o aux.	AFA (ours)	AugMix†	AFA w/ AugMix	PRIME	AFA w/ PRIME	AugMax
FLOPs	×1	×1	×2	×3	×2	×1	×2	×8
Memory	×1.02	×1.02	×1.62	×2.66	×1.83	×2.50	×3.06	×2.35

Table 1. Computational resources of different combinations compared to standard training. Methods with † are reported with JSD.

promising avenue for advancing the field of data augmentation and reinforcing the resilience of computer vision models in challenging real-world conditions.

1.2. Thesis Outline

Beginning with an exploration of related works in the field, detailed in the "Related Works" section. Here, we delve into existing literature, frameworks, and methodologies that form the foundation for our research, including known mathematical results. Following this, the "Method" section delineates the proposed Auxiliary Fourier-basis Augmentation (AFA) technique and its integration into computer vision models. This section provides a comprehensive understanding of the augmentation process, emphasizing its straightforward implementation and computational efficiency. Subsequently, the "Experiments" section constitutes the empirical validation of our approach, where we present and analyse results obtained from various scenarios. This section offers insights into the robustness, generalization, and consistency improvements achieved through AFA, supported by comprehensive experimental evidence. We also include results on how this method can be used to reduce bias in privacy-sensitive context. Finally, the thesis culminates in the "Conclusion" section, summarizing key findings, discussing the implications of our contributions, and suggesting potential avenues for future research in the realm of frequency-driven perspectives in data augmentation for computer vision.

2. Background

In this section provides a comprehensive overview of the foundational research and key concepts that underpin our study. We initiate our exploration by delving into the realm of data augmentation, distinguishing between traditional visual augmentation techniques and the emerging frontier of Fourier-based augmentation. This demarcation sets the stage for understanding the significance of augmenting in the frequency domain, laying the groundwork for our proposed technique, Auxiliary Fourier-basis Augmentation (AFA). Moving beyond augmentation, we delve into the landscape of Convolutional Neural Networks (CNNs) and Vision Transformers. This novel machine learning model offers insights into the potential efficacy of frequency-driven perspectives in enhancing the capabilities of vision models. Additionally, we scrutinize the fundamentals of 2D Fourier Basis Functions, recognizing their crucial role in shaping the frequency spectrum employed in AFA. As we progress, we explore empirical evidence supporting the claim that machine learning models exhibit a preference for learning low frequencies first, a phenomenon foundational to the motivation behind our proposed augmentation technique. Lastly, we introduce the concept of Dominant Frequency Maps, a technique pivotal in substantiating our argument that vision models indeed learn specific frequencies, elucidating the importance of frequency-centric analysis over traditional structural considerations.

2.1. Data Augmentation

Data augmentation includes a set of techniques to increase data variety, thus reducing the distribution gap between training and test data. Generalization and robustness performance of models normally benefits from the use of data augmentation for training [45] or at test-time [20].

Visual Image Augmentation Common image augmentation techniques include transformations, e.g. cropping, flipping, rotation, among others [45]. Applying the transformations with fixed configuration lacks flexibility when the models encounter more variations in the inputs at testing time. Thus, algorithms were designed to combine transformations randomly, e.g. AugMix [16], RandAug [4], TrivialAugment [35], MixUp [51], and CutMix [50]. However, random combinations might not be optimal. In [3], AutoAugment was proposed, based on using reinforcement learning to find the best policy on how to combine basic transformations for augmentation. AugMax [42] instead combines transformations adversarially, aiming at complementing augmentations based on diversity with others that favour hardness of training data. PRIME [34] samples transformations with maximum-entropy distributions. [40] augments images based on knowledge distilled by a teacher model. However, these approaches address variations limited by visually-plausible transformations only.

Frequency-based augmentations. In [48], it was discovered that models trained with visual transformations might be vulnerable to noise impacting certain parts of the frequency spectrum (e.g. high-frequency components), demonstrating that visual augmentations do not completely guarantee robustness. Complementary augmentation techniques are thus required to fill the augmentation gap left by visual augmentations. The straightforward approach is augmentation in the frequency domain. For example, [1] mixes the amplitude spectrum of images to reduce reliance on the amplitude part of the spectrum and induce phase-reliance for classification. [41, 47] swap or mix the amplitude spectrum of images. [43] augments images with shortcut features to reduce their specificity for classification, mit-

igating frequency shortcut learning. [39] introduces frequency noise in the AugMix framework. [25, 30] adversarially perturb images in the frequency domain. While these techniques address what visual augmentations may overlook, they also have limitations. Most frequency augmentation methods are based on manipulation of the frequency components of images. They usually have high computational requirements to identify frequency shortcuts [43], implement adversarial training setup [25] or calculate multiple Fourier transforms of original and augmented images [1, 41, 43, 47] and do not directly address the sensitivity of models to single-frequency noise. For instance, the methods based on amplitude mixing/swapping might result in overfitting to the changed amplitude spectrum if the datasets are small. The methods targeting the frequency characteristics of images usually have large computational requirement, e.g. DFM-X [43] and AdvWavAug [25], to identify frequency shortcuts or implement adversarial training setup.

We instead propose to use Fourier-basis functions as additive noise in the frequency domain. Our augmentation technique requires only one extra step during training rather than multiple pre-processing and expensive computations during training time as in other methods [1, 41, 43, 47], and works to complement image-based augmentations. Furthermore, we simplify the adversarial training framework of AugMax [42], not requiring an optimization process to maximize the hardness of adversarial augmentation, and achieving comparable or higher robustness. This allows the use of adversarial augmentations at larger-scale. We account for the induced distribution shifts in the frequency domain via an auxiliary component. The benefit of AFA is complementary to visual augmentations, and we can incorporate them seamlessly to further boost model robustness.

2.2. Model Architectures

Convolutional Neural Networks Convolutional Neural Networks (CNNs) represent a pivotal advancement in the field of deep learning, particularly tailored for image classification tasks. CNNs leverage convolutional layers to automatically learn hierarchical features from input images, capturing spatial hierarchies and patterns. One notable architecture within the realm of CNNs is the Residual Neural Network (ResNet). Introduced by [13], ResNets revolutionized deep learning by introducing residual blocks that allow the network to learn residual functions, making it easier to train very deep networks. The key innovation lies in the use of skip connections, or shortcuts, which enable the network to bypass certain layers . This not only facilitates the training of deeper networks but also mitigates the vanishing gradient problem, leading to improved convergence and performance. ResNets have demonstrated remarkable success in various computer vision tasks, including image recognition and object detection, earning them a prominent position in the deep learning landscape.

Batch Normalization [18] is a crucial component in the training of CNNs, including ResNets, and plays a pivotal role in stabilizing and accelerating the convergence of deep networks. Batch Normalization operates by normalizing the input to a layer across mini-batches, reducing internal covariate shift. This normalization process helps address issues related to vanishing or exploding gradients during training, enabling the network to be more robust and converge faster. Furthermore, Batch Normalization acts as a regularizer, reducing the need for techniques like dropout and contributing to improved generalization. In the context of ResNets, Batch Normalization facilitates the training of very deep networks by providing stable and normalized inputs to each layer, which is essential for the effective learning of hierarchical features.

Vision Transformers Vision Transformers (ViTs) [6] represent a paradigm shift in computer vision by adopting a transformer architecture, originally designed for natural language processing, to directly process image data. ViTs discard the conventional convolutional layers found in traditional Convolutional Neural Networks (CNNs) and instead rely on self-attention mechanisms to capture long-range dependencies within the image . The transformer architecture in ViTs divides the image into fixed-size patches, linearly flattens them, and feeds them through self-attention mechanisms, allowing for holistic context understanding across the entire image.

Within the domain of Vision Transformers, the Compact Convolution Transformer (CCT) is a noteworthy development. CCT combines the strengths of convolutional layers and transformer architectures to create a more efficient and scalable model. CCT employs a compact convolutional backbone to process image patches, which are then fed into a transformer for capturing global context . This hybrid design retains the advantages of convolutional operations for local feature extraction while leveraging the transformer's ability to capture long-range dependencies.

2.3. 2D Fourier-basis Functions

We utilize Fourier-basis functions in our augmentation strategy as an additive perturbation to the images. They are sinusoidal wave functions used as basic components of the Fourier transform to represent signals and images. A real Fourier basis function has two parameters, namely a frequency f and direction ω, and is denoted as:

$$A_{f,\omega}(u,v) = R\sin(2\pi f(u\cos(\omega) + v\sin(\omega) - \phi)), \quad (1)$$

where $A_{f,\omega}(u,v)$ represents the amplitude of the wave at position (u,v). The function involves the sine of a 2D spatial frequency $2\pi f$ to produce a planar wave with a spe-

cific frequency f, and angle ω that indicates the direction of propagation. R is chosen such that the planar wave has unit l_2-norm. A particular Fourier basis function, characterized by specific frequency (f) and direction (ω), can be associated with a Dirac delta function in the spectral domain. Therefore, when employed in an additive manner, as in our augmentation strategy, this Fourier-basis function facilitates the targeted modification of particular frequency components of images. ϕ is the phase offset of the directional wave set to $\pi/4$. Examples of Fourier-basis waves superimposed on images are shown in Fig. 2.

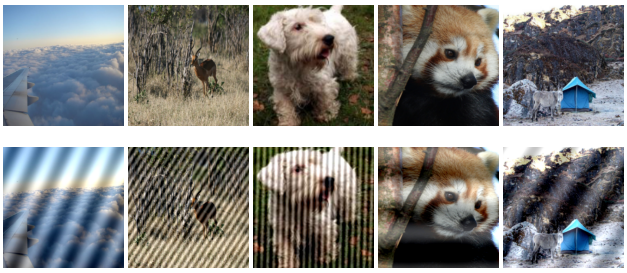


Figure 2. Example of Fourier-basis functions added to natural images. They appear as *gratings* that obscure spatial information.

2.4. Spectral Bias of Neural Networks

[36] investigated the properties of ReLU neural networks, particularly focusing on their learning bias revealed through Fourier analysis. While neural networks are recognized for their high expressivity and capability to fit random input-output mappings with perfect accuracy, the study uncovers a notable learning bias towards low-frequency functions. These low-frequency functions exhibit global variations without local fluctuations, indicating that deep networks prioritize learning simple patterns that generalize across different data samples. While the paper focussed on sequential data trained on with MLPs, the same analysis holds for image dataset and convolutional neural networks as they can be reduced to the prior case.

Additionally, the paper delves into the role of the data manifold's shape in the learning process. Contrary to intuition, the study provides both empirical and theoretical evidence indicating that learning higher frequencies becomes easier as the manifold complexity increases.

Considering the neural network's affinity for learning global variations without local fluctuations, the introduction of Fourier noise can act as a means to guide the learning process. Fourier noise introduces controlled variations in frequency components, influencing how the network prioritizes and adapts to different patterns in the dataset. This augmentation strategy aligns with the notion that over-parameterized networks often prioritize simpler, globally varying features. Therefore, the introduction of Fourier noise may provide a valuable means to explore and exploit the frequency-dependent learning bias for improved performance in terms of robustness, generalisability and consistency of neural networks.

2.5. Frequency Shortcut Learning

In our pursuit of generalisation and robustness, while learning a general function for complicated data manifolds might sound advantageous, we quickly run into issues of shortcut learning [9] where essentially neural networks rely on spurious information rather than deeper semantic information or task-related cues [44]. The same is true in the Fourier domain, where [44] show this frequency shortcut learning phenomenon.

To address the challenges associated with mitigating implicit shortcuts, we draw inspiration from the advancements in Fourier-based perspectives and the spectral bias paper, introducing Auxiliary Fourier-basis Augmentation (AFA) as a powerful tool for reshaping the learning dynamics of neural networks. AFA strategically integrates Fourier-based perturbations into the training process, hypothetically allowing the network to develop a more nuanced understanding of the frequency domain and mitigating its reliance on superficial statistics or biases by acting as a regularisation shown in Fig. 10.

Through the application of AFA, we aim to disrupt the learned frequency shortcuts by introducing controlled variations in the frequency components of the training data. This process not only acts as a form of regularization but also guides the network towards prioritizing more semantically relevant features, diminishing its inclination to exploit simple, non-semantic cues.

3. Auxiliary Fourier-basis Augmentation

The Auxiliary Fourier-basis Augmentation (AFA) that we propose is based on two lines of augmentations, one considered in-distribution (using visual augmentations) and another considered out-of-distribution or adversarial (using frequency-based noise) as shown in Fig. 3. We generate the adversarial augmented images by sampling a Fourier-basis and a strength parameter per colour channel, and adding them to the original images. Visually augmented and adversarially augmented training images are then processed using a main component and an auxiliary component, respectively. Joint optimisation of two cross-entropy functions encourages robust and consistent classification, as it promotes correctness under adversarially augmented images. Details of the different parts of the method are reported below.

Generation of adversarial augmented images. Randomly sampling augmentations and applying them to images with random strengths was shown to be sufficient to outperform more complex strategies [35].

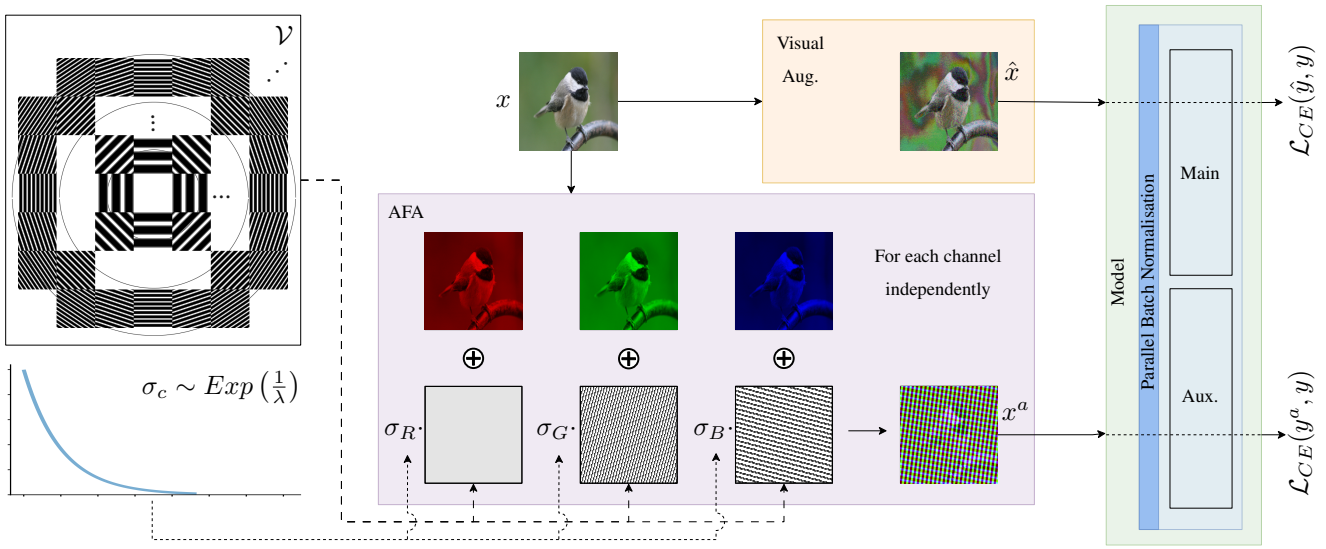


Figure 3. Schema of the AFA augmentation pipeline. The image x is augmented using AFA, which adds a planar wave per channel c of the image at a strength value σ_c sampled from an exponential distribution (eq.2). The AFA augmented image x^a is used for training, processed through the auxiliary component of the parallel batch normalisation layer (for models that use batch normalization to track batch statistics, e.g. ResNet). Other visual augmentations are applied in parallel, and used for training via the main component of the normalization layer. Finally, we train via optimizing two cross-entropy losses, one for the main and the other for the auxiliary component.

We follow this design principle in our method to generate adversarial augmented images with Fourier basis functions, which allows us to avoid optimization steps to determine the worst-case combination of augmentations as in AugMax [42]. We produce adversarial augmented images by adding a different Fourier basis function $A_{f,\omega}$ per channel of the original RGB image. We generate the Fourier basis functions by sampling f and ω from uniform distributions as $f \sim \mathcal{U}_{[1,\mathrm{M}]}$ and $\omega \sim \mathcal{U}_{[0,\pi]}$, where M is the image size. The sampling space of all Fourier-basis is denoted as $\mathcal{V}$. We add the generated Fourier basis functions per channel c with a weight factor sampled from an exponential distribution $\sigma_c \sim \mathrm{Exp}(1/\lambda)$, with $c \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$. The selection of the exponential distribution for sampling augmentation magnitude is motivated by the concept of event rate, where perturbations with larger magnitudes become progressively less likely, albeit still possible. This is controlled by adjusting λ, ensuring a balance between maintaining diversity in sampled values while minimizing the occurrence of extremely large augmentation perturbations. In Sec. 4.3, we show how the parameter λ affects the augmentation results.

The proposed augmentation process results in a 3-channel image $x^a = [x^a_R, x^a_G, x^a_B]$, where:

$$x^a_c = \mathrm{Clamp}_{[0,1]}(x_c + \sigma_c A_{f_c,\omega_c}), \quad c \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}. \quad (2)$$

An example of image x^a augmented with additive Fourier-basis functions is shown in our method schema in Fig. 3.

Auxiliary component for distribution shifts. As shown in Figs. 2 and 3, the Fourier-basis augmentations result in images with an unnatural appearance due to substantial frequency perturbations. The presence of planar waves across the augmented images determines the *unnaturalness* of image appearance, which can be seen as adversarial attacks on the images. These augmentations disrupt the learned mean and variance in batch normalization layers, which are inconsistent with the distribution shifts induced by our augmentation and lead to inconsistent activations. This results in a negative impact on model convergence and generalization abilities.

We address these issues by deploying architectural components in the training, capable of handling distribution shifts explicitly by tracking statistics and adjusting the loss function accordingly. Namely, we incorporate auxiliary components into the model, such as Parallel Batch Normalization layers and an additional cross-entropy term in the loss function to specifically account for these adversarial augmented images. These modifications to the model architecture and training enhance performance, particularly in the presence of distribution shifts, contributing to better generalization, robustness to common corruptions and consistency to time-dependent increasing perturbations. The introduction of parallel batch normalization layers is motivated by the need to account for distribution shifts induced by adversarial (Fourier-basis) augmentations, as observed in [42]. With the parallel batch normalisation, the affine parameters and statistics of main and auxiliary distributions are recorded separately. This allows independent learning of distribution of the visually and adversarially augmented

images. Without these additional normalization layers, the model training assumes a single-modal sample distribution, limiting its ability to differentiate between the main and the adversarial distribution, thus negatively affecting overall performance. In Sec. 4.3, we show the result of not employing the auxiliary components.

It is worth noting that for models that do not employ batch normalization layers (e.g. CCT that uses layer normalization and does not track statistics), the parallel normalization layers are not needed. However, the extra term in the loss function (see next paragraph) to generate consistent predictions across distribution shifts serves as a regularization mechanism.

Loss function. We work in the supervised learning setting with a training dataset $\mathcal{D}$ consisting of clean images x with labels y. We train the model in the main architecture stream (see Fig. 3) using a cross-entropy loss $\mathcal{L}_{\text{CE}}(\hat{y}, y)$, where y is the ground-truth label and $\hat{y}$ is the predicted label for images augmented with a given visual augmentation strategy (e.g. standard, PRIME, etc.). Under the non-auxiliary setting, models thus optimise the standard cross entropy loss.

In the auxiliary setting, we add an extra cross-entropy loss term $\mathcal{L}_{\text{CE}}(y^a, y)$, which optimise the model to predict the correct label on adversarial augmented images whose predicted label is denoted by y^a, contributing to robustness of the model w.r.t. aggressive distribution shifts. We refer to the combined loss function $\mathcal{L}_{\text{ACE}}$, taking the average of the two cross-entropy terms, as the Auxiliary Cross Entropy (ACE) Loss:

$$\mathcal{L}_{\text{ACE}}(\hat{y}, y^a, y) = \frac{1}{2}\left[\mathcal{L}_{\text{CE}}(\hat{y}, y) + \mathcal{L}_{\text{CE}}(y^a, y)\right]. \quad (3)$$

It contributes to achieve comparable performance, with lower training time and complexity, than using the Jensen-Shannon Divergence (JSD) loss [16, 42]. Our motivation to not employ the JSD loss is the reduced training time due to less computational complexity. In our experiments, for comparison purposes, we also use the JSD loss in the auxiliary setting, where training batches are augmented using AFA and go through auxiliary components. We report results in Sec. 4.3 (Fig. 6).

3.1. Suitability of Fourier-basis Functions

According to the Fourier Transform theory, any signal can be represented as the sum of sinusoidal functions (i.e. planar waves in 2D). Adding such a function to the image space (with parameters f and ω) corresponds to augmenting the amplitude of a specific frequency component (f,ω) in the 2D Fourier transform of the image. Therefore, adding sinusoids is the same as augmenting the corresponding frequency and amplitude in the 2D Fourier transform of the image. As mentioned before neural networks exhibit spectral biases and therefore this makes AFA able to bridge the gap left by visual augmentations, which are usually carried out in the spatial domain and might not address well the spectral bias of models [44]. We use the real part of the planar waves in the visual domain to 1) avoid explicit computations of Fourier transforms for efficiency, and 2) reduce amplitude-reliance and encourage phase-reliance of models for classification, useful to improve generalisation capabilities [1]. While it is possible to explore other shape/pattern functions, they would not possess the same characteristics of sinusoidal waves according to the Fourier Transform theory, thus undermining the validity and specificity of frequency-based augmentations.

3.2. Proof of Augmenting Fourier Domain

Lemma 1 (Linearity)**.** *Let f, g be functions of a real variable and let $\mathscr{F}(f)$ and $\mathscr{F}(g)$ be their Fourier transforms. Then for complex numbers a and b*

$$\mathscr{F}(af + bg) = a\mathscr{F}(f) + b\mathscr{F}(g), \quad (4)$$

therefore, Fourier transform $\mathscr{F}$ is a linear transformation.

Lemma 2 (Fourier Transform of Plane Wave)**.** *The Fourier transform of the planar wave given by the frequency f and the direction ω, $A_{f,\omega}$ has a fourier transform*

$$\mathscr{F}(A_{f,\omega}) = \mathscr{F}\left(R\cos(2\pi f(u\cos(\omega) + v\sin(\omega)))\right) \quad (5)$$

$$= \frac{R}{2}\left(\delta(\hat{x}, \hat{y}) + \delta(\bar{x}, \bar{y})\right), \quad (6)$$

where, $\hat{x} = x - f\cos(\omega)$, $\hat{y} = y - f\sin(\omega)$ and $\bar{x} = x + f\cos(\omega)$, $\bar{y} = y + f\sin(\omega)$.

Theorem 1 (AFA Augments the Fourier Domain)**.** *Given an image sample s, an augmentation using AFA produces as augmentation in the Fourier domain of the image for one specific frequency and orientation of the wave (f, ω).*

Proof. Given image s and the randomly sampled planar wave using AFA, $\sigma A_{f,\omega}$, dropping the subscript for the channels for clarity, we have:

$$\mathscr{F}(\text{AFA}(s)) = \mathscr{F}(s + \sigma A_{f,\omega})$$

$$= \mathscr{F}(s) + \sigma\mathscr{F}(A_{f,\omega}) \quad (7)$$

(using Lemma 1)

$$= \mathscr{F}(s) + \frac{\sigma R}{2}\left(\delta(\hat{x}, \hat{y}) + \delta(\bar{x}, \bar{y})\right). \quad (8)$$

(using Lemma 2)

Therefore, we prove augmenting an image s with AFA corresponds to augmenting the amplitude of a specific frequency component (f,ω) in the 2D Fourier transform of the image. □

4. Experiments and results

We compare AFA with other popular augmentation techniques, evaluating robustness to common corruptions, generalization abilities and consistency to time-dependent increasing perturbations, on benchmark datasets.

4.1. Experimental Setup

Datasets We trained models on the CIFAR-10 (C10) [21], CIFAR-100 (C100) [22], TinyImageNet (TIN) [23] and ImageNet (IN) [5] datasets and evaluate them on the corresponding robustness benchmark datasets, namely C10-C, C100-C, TIN-C, IN-C [15], IN-$\bar{\text{C}}$ [33], and IN-3DCC [19]. For ImageNet-trained models, we further evaluate their generalisation performance on the IN-v2 [37] and IN-R datasets [14], and consistency of performance on time-dependent increasing perturbations on the IN-P dataset [15].

Architectures and training details. We train ResNet [13] and Compact Convolution Transformers (CCTs) [12]. We train ResNet-18 and CCT-7/3x1 (32 resolution) on C-10, C-100, and only ResNet-18 on TIN. In the case of ImageNet, we train ResNet-18, ResNet-50 and CCT-14/7x2 (224 resolution). Under auxiliary setting, we use the DuBIN variant of ResNet [42]. We always use standard transforms [13] before other augmentations. Implementation details and hyperparameter configurations are in the Appendix A. We release code and models[1].

Evaluation metrics. We evaluate the classification accuracy on the original test set, which we refer to as standard accuracy (SA), and the average classification accuracy over all corruptions in the robustness benchmarks as robustness accuracy (RA). This provides direct comparison between model performance on original and corruption benchmark datasets. We also compute the mean corruption error (mCE) [15] for TIN and IN (for CIFAR there are no baselines advised) to evaluate the normalized robustness of models against image corruptions, the mean flip rate (mFR) and the mean top-5 distance (mT5D) to evaluate the consistency performance of models against increasing perturbations. For the evaluation of generalization performance, we compute the accuracy on the ImageNet-R and ImageNet-v2 test sets (note that ImageNet-v2 has 3 test sets, and we report the average accuracy on them). More details about the metrics are in the Appendix B.

4.2. Results

Comparison with AugMax. We first report a direct comparison with AugMax [42] in Tab. 5, as AFA addresses the computational shortcomings of generating adversarial augmentations via PGD iterations, and of using a JSD loss for alignment of the distribution of original and (adversarially) augmented images. We use AugMix as main augmentation, as in AugMax, and ablate on the use of JSD and ACE loss.

We show that AFA achieves comparable (or better) performance than AugMax, despite it being much less computational intensive. We indeed demonstrate that we can generate adversarial augmentations by only adding (weighted) Fourier-basis waves per color channel, not requiring PGD steps, and can train the models using an extra cross-entropy instead of the expensive JSD loss. The improvements granted by our approach are particularly evident in the case of ImageNet (using ACE), where we gain 1.6% of standard accuracy and 4.1% of robust accuracy (5.6% mCE) performance w.r.t. AugMax. Considering the increased computational efficiency and the simplicity of adversarial augmentation method, AFA is a more versatile and effective tool than AugMax. Hence, in the rest of the paper, we do not report further results of the AugMax framework, due to its high computational requirements, which complicate the training of larger models (e.g. ResNet-50 and CCT).

Robustness, generalization and consistency. In Tab. 3, we report results achieved by AFA combined with different visual augmentation methods, AugMix, PRIME, TrivialAugment (TA), to train different architectures (ResNet, CCT). We evaluate robustness to common corruptions on IN-C, IN-$\bar{\text{C}}$ and IN-3DCC, OOD generalisation on IN-v2 and IN-R, and consistency w.r.t. increasing perturbations on IN-P.

AFA generally contributes to a boost of performance (green colored results in Tab. 3) when combined with different visual augmentation techniques, reducing the robustness and generalization gap for different model architectures. Even compared to another Fourier basis augmen-

-	Main	Auxiliary	SA↑	RA↑	mCE↓
C10	AugMix†	✗	95.47	86.48	-
	AugMix†	AugMax	95.76	**90.36**	-
	AugMix†	AFA	95.24	89.96	-
	AugMix	AFA	95.44	89.81	-
C100	AugMix†	✗	78.72	61.61	-
	AugMix†	AugMax	78.69	65.75	-
	AugMix†	AFA	78.99	65.96	-
	AugMix	AFA	77.80	**66.69**	-
TIN	AugMix†	✗	64.65	36.30	83.90
	AugMix†	AugMax	62.21	**38.67**	**80.72**
	AugMix†	AFA	64.34	38.53	**80.79**
	AugMix	AFA	62.51	**38.67**	80.83
IN	AugMix†	✗	65.2	31.5	87.1
	AugMix†	AugMax	66.5	36.5	80.6
	AugMix†	AFA	65.0	36.8	80.4
	AugMix	AFA	68.1	**41.1**	**75.0**

Table 2. Comparison of AFA and AugMax (with AugMix for visual augmentation [42]), with a ResNet18 backbone. The mark † indicates the use of the JSD loss, otherwise the ACE loss is used.

[1]Code and models available at https://ANONYMOUS

				Robustness						Generalisation		Consistency	
				IN-C		IN-$\bar{C}$		IN-3DCC		IN-R	IN-v2	IN-P	
	Main	Aux	SA (↑)	RA (↑)	mCE (↓)	RA (↑)	mCE (↓)	RA (↑)	mCE (↓)	Acc. (↑)	Avg. Acc. (↑)	mFP (↓)	mT5D (↓)
ResNet18	-	✗	**68.9**	32.9	84.7	34.8	87.0	34.9	84.4	33.1	**64.3**	72.8	87.0
	-	AFA	68.2	35.9	81.0	41.7	78.3	37.1	81.7	32.8	63.7	64.2	76.8
	AugMix†	✗	65.2	31.5	87.1	34.6	87.3	32.1	88.3	28.2	59.5	80.2	86.2
	AugMix†	AFA	65.0	36.8	80.4	40.9	79.3	36.0	83.2	30.6	60.9	60.1	68.5
	AugMix	AFA	68.1	41.1	75.0	45.2	73.3	38.9	79.4	35.2	63.2	68.5	81.7
	PRIME	✗	66.0	43.6	72.0	42.0	78.1	42.4	75.2	36.9	61.4	54.7	65.3
	PRIME	AFA	67.2	**47.2**	**67.8**	**47.3**	**71.1**	**43.8**	**73.5**	**37.8**	63.0	**52.3**	**63.7**
	TA$^+$	✗	**68.9**	36.9	80.1	35.9	85.6	38.6	79.7	32.6	63.7	68.1	81.4
	TA$^+$	AFA	67.8	41.4	74.7	42.9	76.7	41.1	76.5	35.4	62.7	59.9	72.3
ResNet50	-	✗	75.6	39.2	76.7	39.9	79.4	41.2	76.1	36.2	70.8	58.0	78.4
	APR-SP	✗	71.9	42.9	72.7	45.9	72.5	39.8	78.4	34.9	67.2	60.2	75.4
	-	AFA	76.5	46.2	68.0	47.6	69.4	46.2	69.8	38.1	72.0	48.0	67.2
	AugMix†	✗	74.7	43.4	72.0	44.6	73.3	41.9	75.5	33.0	70.0	60.9	72.5
	AugMix†	AFA	75.6	50.6	62.9	51.8	64.0	47.6	68.3	36.3	71.2	44.5	56.1
	AugMix	AFA	76.6	49.1	64.7	52.5	62.9	46.3	69.6	41.0	71.8	52.2	72.2
	PRIME	✗	72.1	49.2	64.9	46.4	71.5	47.2	68.8	38.5	67.8	45.4	58.1
	PRIME	AFA	74.5	**53.9**	**59.2**	**54.2**	**61.3**	**50.2**	**65.0**	**40.9**	69.8	**40.4**	**54.8**
	TA$^+$	✗	75.9	43.4	71.7	41.8	77.1	44.7	71.6	37.1	70.3	51.9	70.4
	TA$^+$	AFA	76.6	50.3	63.1	49.7	66.7	49.6	65.4	40.0	**72.2**	45.1	64.5
CCT	-	✗	76.4	43.9	70.7	50.3	65.6	43.4	73.2	35.6	71.2	48.3	72.9
	-	AFA	76.9	51.9	61.0	58.5	55.4	50.7	64.4	39.0	71.9	38.4	61.8
	AugMix	✗	76.1	47.3	66.8	52.2	63.1	45.3	71.0	37.9	70.7	49.3	72.8
	AugMix	AFA	**77.4**	56.5	55.6	60.8	52.2	51.8	62.8	41.0	**72.5**	37.9	59.9
	PRIME	✗	73.6	54.1	58.6	54.5	60.8	50.7	64.4	39.2	68.7	36.1	53.0
	PRIME	AFA	76.6	**58.7**	**52.8**	**61.2**	**52.0**	**54.5**	**59.4**	**43.2**	71.9	**31.9**	**51.2**
	TA$^+$	✗	77.1	50.2	63.2	54.1	60.7	49.3	65.8	38.2	72.1	41.8	66.3
	TA$^+$	AFA	76.9	56.0	56.0	59.1	54.6	53.1	61.1	41.1	72.1	36.4	58.5

Table 3. Robustness, generalization and consistency results on ImageNet-based benchmarks. Models with † use the JSD loss. TrivialAugment (TA) has overlapping augmentations with IN-C ($^+$), and no other overlaps with other datasets. The green colour indicates an improvement when the main augmentation is combined with AFA, while red indicates no improvement. Results marked with **bold/bold** are the best for a particular architecture.

tation technique APR-SP [1] AFA out performs it on all benchmarks when trained without any other augmentation techniques. Also in the case of ResNet50 trained with AugMix and AFA, we record better overall performance even after dropping the JSD term in the loss. For the transformer architecture CCT, training with AFA contributes to an even stronger improvement in all tests. These results stay consistent for smaller resolution datasets (CIFAR and TIN), as we report at the end of this section.

Robustness to high-severity corruptions. AFA contributes to a consistent improvement of robustness of models at increasing corruption severity (example images with different corruptions are in the supplementary material). We compute the relative corruption error, namely the difference between the corruption error of models trained with a visual augmentation technique only and those trained with both visual augmentations and AFA, and report it in Fig. 4 for different corruption severity. A positive value indicates that models trained with the addition of AFA have better robustness. For higher corruption severity, AFA contributes to stronger robustness, measured by an increase in the relative corruption error in Fig. 4. The improvements obtained by AFA on IN-3DCC are slightly less pronounced than those on IN-C and IN-$\bar{C}$. This is attributable to the specific corruptions in IN-3DCC that concern 3D geometric information, and are somewhat more complicated image transformations. However, AFA contributes to a substantial improvement w.r.t. to models trained without it. We thus highlight that AFA is very beneficial for increasing robustness to aggressive corruptions of the test images. Details of the results at different severity are in the supplementary material.

Fourier heatmap: robustness in the frequency spectrum. We further evaluate the robustness of models to perturbations at specific frequencies, using test images perturbed with frequency noises according to [48]. We present the results in the form of Fourier heatmaps, see Fig. 5 for heatmaps of ResNet18 models (trained on ImageNet), and the supplementary material for the heatmaps of CCT models. The intensity of a pixel at location (u, v) in the heatmap indicates the classification error of a model tested on images perturbed by Fourier noise at frequency (u, v) in the

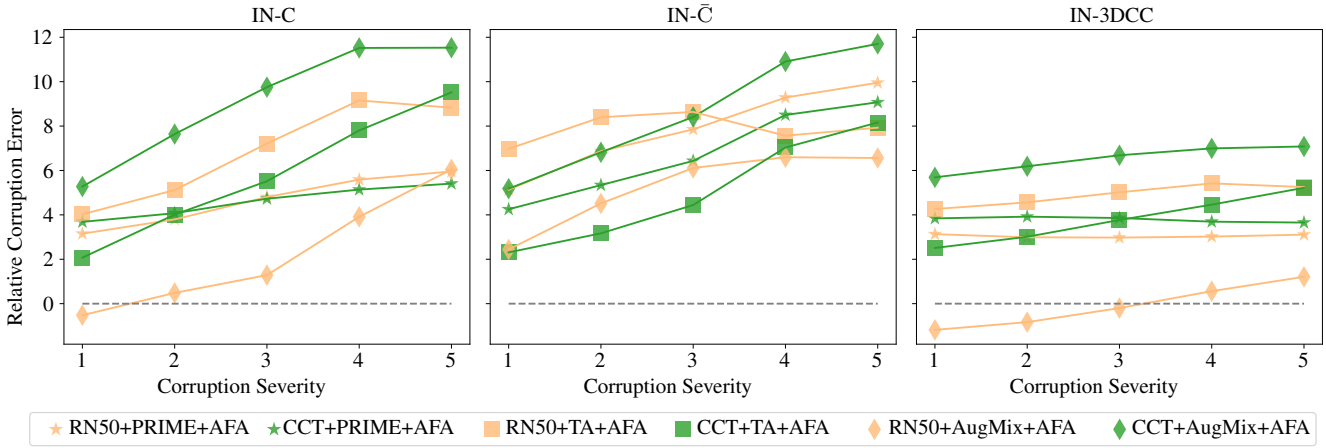


Figure 4. Relative error per corruption severity, computed by subtracting the classification error of models trained with PRIME, TrivialAugment, and AugMix with that of corresponding models trained with PRIME+AFA, TrivialAugment+AFA, and AugMix+AFA.

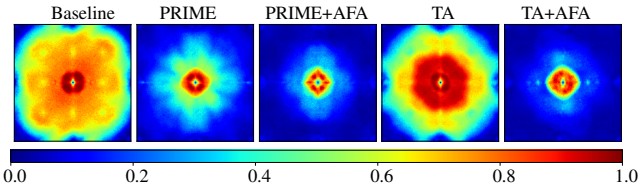


Figure 5. Fourier heatmaps of ResNet18 trained with standard setup, and PRIME and TrivialAugment, with and without AFA.

frequency spectrum (implementation details are in the supplementary material). ResNet18 trained with standard augmentations setting (baseline) is very sensitive to perturbations at low and middle-high frequency (see Fig. 5), while those trained with visual augmentations like PRIME and TrivialAugment (TA) still show vulnerability at low and middle-high frequency noise. When training models with AFA, i.e. PRIME+AFA and TA+AFA, the models become more robust to frequency pertubations, especially at middle-high frequency. AFA can provide extensive robustness to frequency perturbations and bridge the robustness gap that visual augmentation might not cover.

Results on CIFAR and TIN. In Tab. 4, we present the robustness results on smaller resolution datasets, C10 and C100. The results on TIN are in the Appendix C.1. There results are inline with those reported on IN in Tab. 3.

4.3. Ablation

Auxiliary components. We investigate the contribution and importance of the auxiliary components in improving model robustness. We trained models with AFA-augmented images, passing through only the main components or the auxiliary components. The results in Tab. 5, i.e. lower RA and higher mCE of models trained with AFA applied

-	Main	Auxiliary	C10-C SA↑	C10-C RA↑	C100-C SA↑	C100-C RA↑
ResNet18	-	✗	94.15	73.67	78.27	48.30
	-	AFA	94.69	88.22	77.91	62.53
	AugMix†	✗	**95.47**	86.48	78.72	61.61
	AugMix†	AFA	95.24	89.96	**78.99**	65.96
	PRIME	✗	94.38	89.81	75.49	66.16
	PRIME	AFA	94.54	**90.64**	76.16	**68.48**
CCT	-	✗	95.67	80.45	**78.37**	54.20
	-	AFA	**95.94**	88.13	77.47	61.40
	AugMix	✗	95.10	85.42	75.79	60.83
	AugMix	AFA	**95.93**	90.57	77.22	66.18
	PRIME	✗	95.30	90.56	76.65	**67.92**
	PRIME	AFA	95.49	**91.40**	76.50	**67.89**
CVT	-	✗	94.31	77.02	75.53	48.25
	-	AFA	**94.53**	**87.03**	**76.96**	**60.12**
VIT	-	✗	94.46	75.97	74.26	50.88
	-	AFA	**94.58**	**86.71**	**75.13**	**58.25**

Table 4. Results for C10-C and C100-C with ResNet18, CCT, CVT and ViT-Light. Models with † use loss with JSD.

only in the main components, highlight the importance of AFA auxiliary components. The auxiliary components play a crucial role in mitigating the impact of aggressive adversarial distribution shifts induced by AFA. By doing so, they contribute to model ability to learn from the original distribution, while AFA facilitates learning robustness to distribution shifts. This is also highlighted in the substantial decrease in SA for models not employing auxiliary components. While model robustness improves under both settings, the performance gain for the auxiliary setting is three to five percentage points higher across all datasets.

-	Main	Auxiliary	SA↑	RA↑	mCE↓
C10	-	✗	94.15	73.67	-
	AFA	✗	92.36	83.25	-
	-	AFA	**94.69**	**88.22**	-
C100	-	✗	78.27	48.30	-
	AFA	✗	72.34	58.70	-
	-	AFA	**77.91**	**62.53**	-
TIN	-	✗	63.56	25.86	97.34
	AFA	✗	59.04	28.87	93.45
	-	AFA	**62.52**	**33.35**	**87.58**
IN	-	✗	68.9	32.9	84.7
	AFA	✗	66.7	33.3	84.4
	-	AFA	**68.2**	**35.9**	**81.0**

Table 5. Ablation results ResNet18 trained with and without Auxiliary Components on C10, C100, TinyImageNet and ImageNet.

ACE vs JSD. As part of our method, we replaced the use of JSD with ACE which is less computationally burdening. We thus performed an ablation analysis of the tradeoff of using JSD. We report results for robustness using mCE and Robust Accuracy (RA) in Fig. 6, and observe that JSD does not significantly improve the robustness of our model to image corruptions, despite it being more computationally heavy than using ACE. Using JSD also results in slightly worse robustness on C100. Given the minimal differences, we opt for the simpler ACE loss for training with the AFA augmentation pipeline and only using JSD if other techniques (e.g. AugMix) employ them.

Effect of hyperparameter $1/\lambda$. We studied also the contribution of the mean $1/\lambda$ of the exponential distribution that we use to sample the weight factor for the channel-wise application of the Fourier-basis augmentations. We provide the results in Fig. 7, and observe that our method has low sensitivity to the choice of the rate parameter. This is attributable to the choice of the exponential distribution that allows larger values to be sampled even if they are less likely. We indeed observe that larger values of $1/\lambda$, which result in larger perturbations (in the range of 10 to 15), result in stronger gains in robustness. At the same time, there is no clear trend in the standard accuracy on the clean dataset, with only minimal variations for the larger values, indicating that the choice of the $1/\lambda$ value does not have a specific influence on the correct functioning of AFA.

5. Discussion

This section discusses on three key aspects of our proposed method, Auxiliary Fourier-basis Augmentation (AFA). Firstly, we provide compelling evidence of the adversarial nature inherent in AFA, highlighting its impact on model robustness. Secondly, we delve into the strong regularization mechanisms of AFA, elucidating its role in enhancing

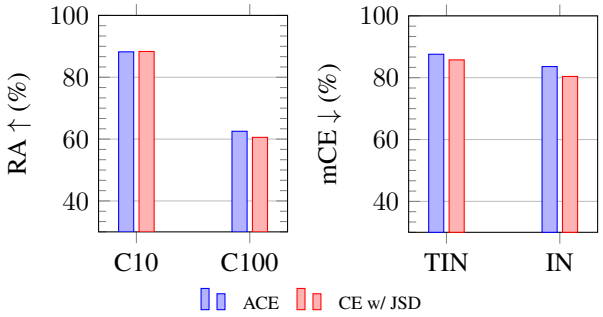


Figure 6. Comparison of using objective with and without the JSD term. All models are ResNet-18 trained with only AFA in the auxiliary component and no other augmentations. When used with JSD two batches passed through Auxiliary components and there was no main augmentation (in total 3 batches, 1 clean and 2 AFA).

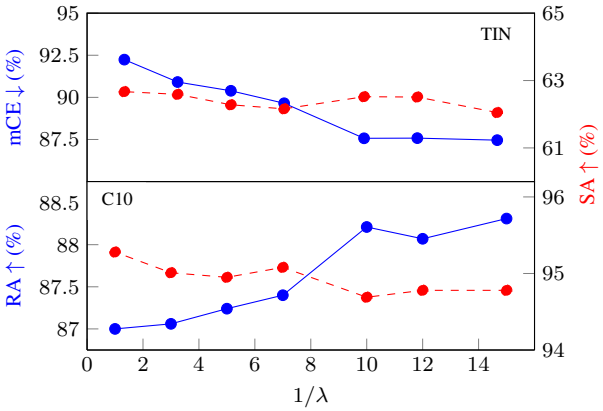


Figure 7. Trend of the mCE and SA with respect to the rate parameter. The models were trained using AFA in the auxiliary setting and no other augmentations for the main.

model stability and generalization. In the final subsection, we explore how AFA addresses and mitigates frequency shortcuts to fortify the resilience of neural network models.

5.1. Evidence of Adversarial Augmentation

Main and auxiliary batch normalisation For the ResNet architecture, which includes Batch Normalisation layers, we had replaced the Batch Normalisation layers with DuBIN layers [42] while operating the Auxiliary setting. Assuming that there is no difference in the distribution of images augmented using AFA and a typical visual augmentation technique, there should be no difference in the affine parameters learnt for each individual batch normalisation parameter (the main and the auxiliary).

We show in Fig. 8 the Mean Absolute Difference of the same parameter between the main and the auxiliary component of the DuBIN layer at different depths of the model. We show the results for models trained with ACE loss for ResNet-50 where AFA is paired with just standard transforms, AugMix, PRIME and Trivial Augment (TA).

We can see that at earlier depths the parameter differ largely, which is explained by the difference in distribution of a visually augmented and AFA augmented image. This difference converges to a lower value, which is again explained by the model attempting to extract similar features from the differently augmented images.

Embedding Space Visualization We compare how diverse are the augmentations of AFA are with respect to other methods. We follow the procedure in [34]. To reiterate the procedure, we randomly select 3 images from ImageNet, each one belonging to a different class. For each image, we generate 100 transformed instances using Standard Transform, Trivial Augment, PRIME, PGD attack with the following parameters: 5 steps, epsilon of $8/255$ and alpha of $2/255$, and with AFA. Then, we pass the transformed instances of each method through a ResNet-50 pre-trained on ImageNet using standard transform and training setup, and extract the features of its embedding space from the penultimate layer before the dense layer. On the features extracted for each method, we perform PCA after whitening and then visualize the projection of the features onto the first two principal components. We visualize the projected augmented space in Fig. 9, which demonstrates that AFA generates which are more akin to an adversarial attack rather than a standard augmentation. This is clear from a visual similarity of AFA's result in Fig. 9e to PGD's result in Fig. 9d and dissimilarity to the other Visual Augmentation techniques.

Finally, we also add in Fig. 9f the embedding space visualisation for the Auxiliary Trained model with AFA augmentation and standard transform for main, following the same procedure as above. We see that the model learns more separable embeddings for images augmented with AFA using the auxiliary setting, therefore is less sensitive to Frequency perturbation. The embeddings also retain a large variance and hardness, therefore showcasing the diversity of the augmentations of AFA.

5.2. Strong Regularisation Effect

In Fig. 10 we show the norm of the weights of the convolutional kernels for the ResNet50 models trained with and without AFA at each depth. We see that AFA provides a strong regularisation effect that is akin to the regularisation effect of PRIME. Meanwhile, we see that AugMix does not regularise the weights at all compared to the baseline model with only the standard transforms. The weights are however regularised to when AFA is paired with AugMix. Combined with PRIME, there does not seem to be further regularisation of the weights.

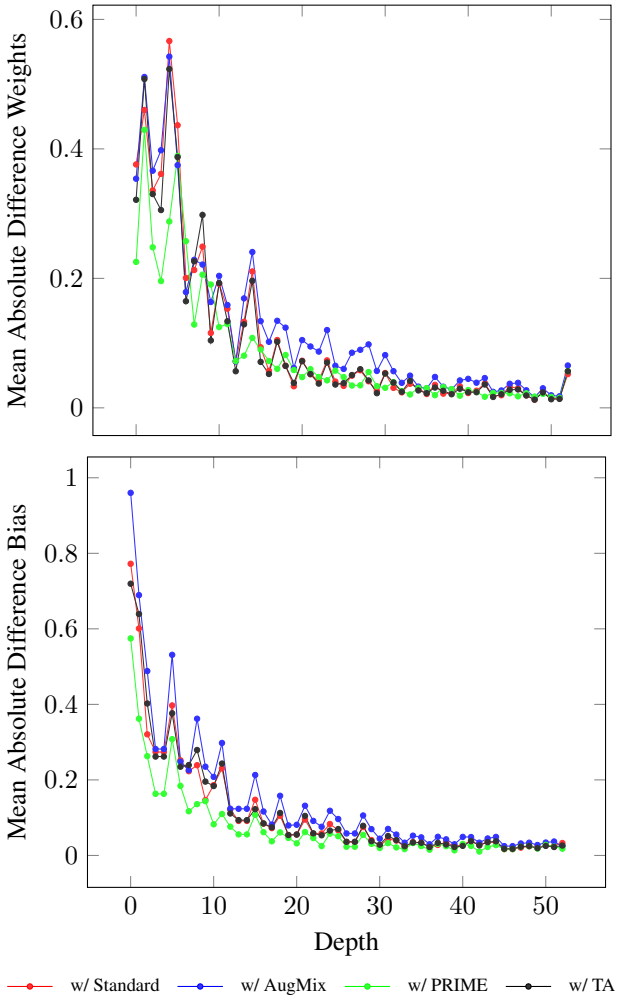


Figure 8. Comparison of the mean absolute difference of the learnt affine parameters for the two batch normalisations in the Dual Batch Norm Layers of ResNet50-DuBIN architecture at different depths.

5.3. Frequency Shortcuts

We assessed the effectiveness of our novel method, AFA, specifically tailored for addressing Frequency Shortcuts. The evaluation was conducted on a dataset characterized by binary targets, and we subsequently computed Dominant Frequency Maps (DFMs) to elucidate the impact of AFA on model performance.

Through a comparative analysis, we contrasted the DFMs derived from models trained with AFA against those trained without it. The results unveiled a compelling trend: a discernible decrease in the reliance on High Frequency components in models utilizing AFA as shown in Fig. 11. We can see that for the standardly trained model, the identified dominant frequencies are quite broad, ranging from high frequencies to lower frequencies. However, when the model is trained with AFA, we see these dominant frequen-

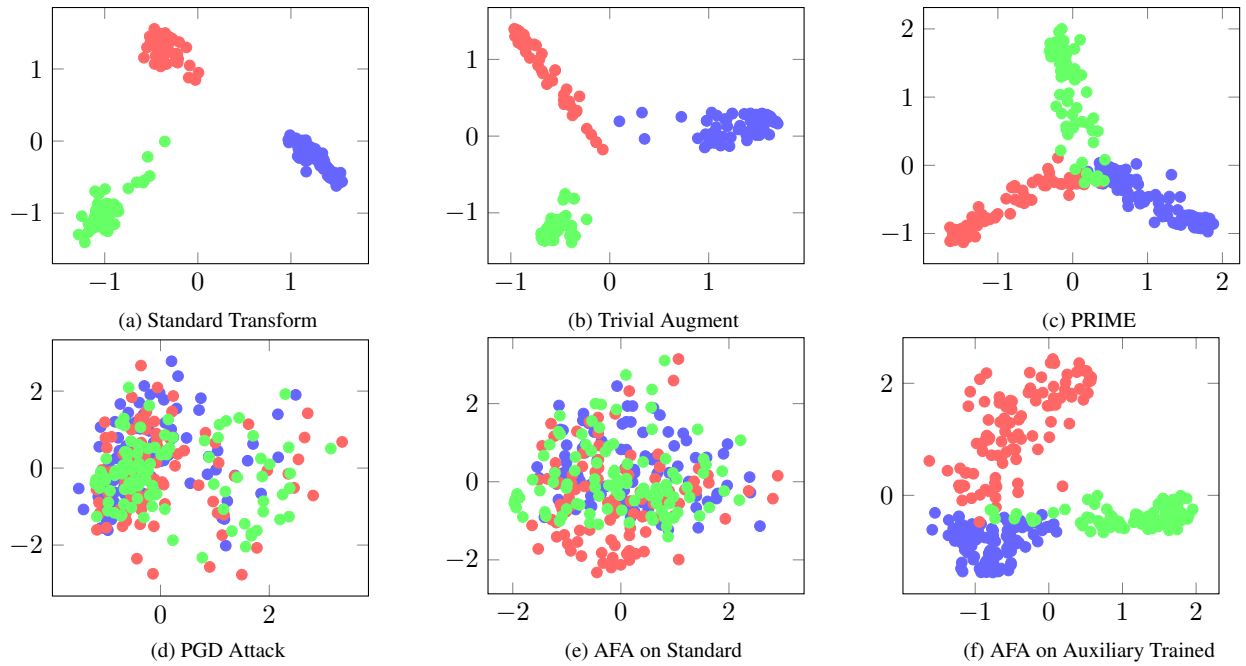


Figure 9. Differences in the Embedding Space for Different Methods and PGD Attack. From (a)-(e) the standardly trained model is used, and for (f) the model trained in the auxiliary setting is used.

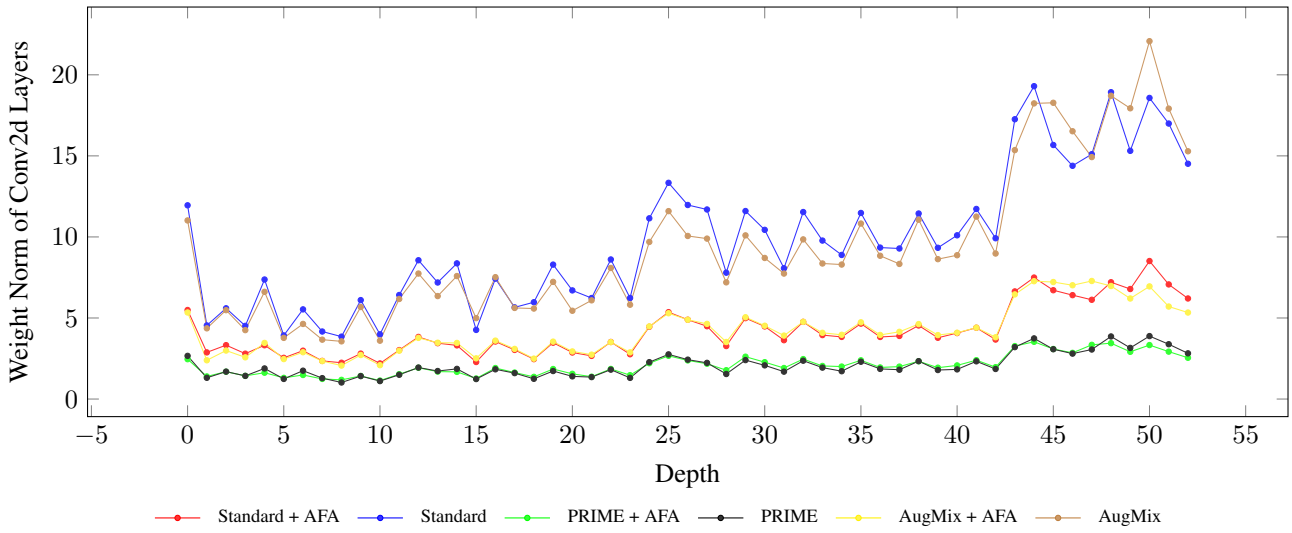
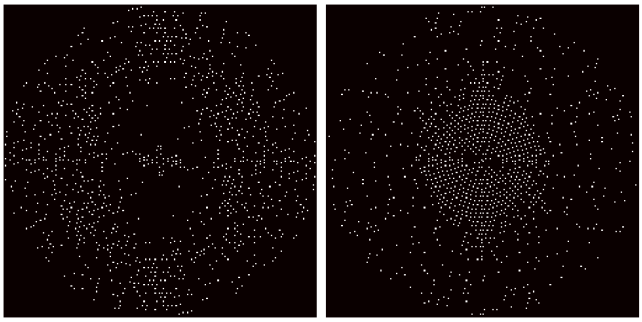


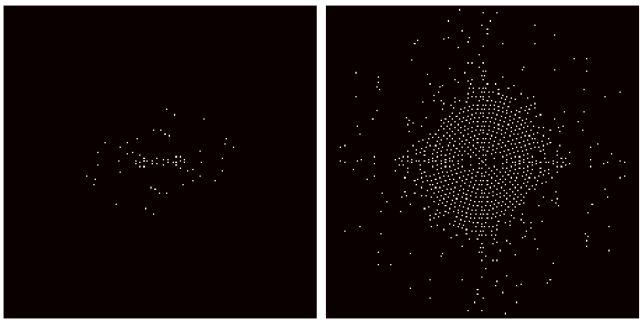
Figure 10. The norm of the Conv2d Layers for ResNet 50 trained with different augmentation techniques with and without AFA. The plot highlights the regularisation effect the methods have on the model weights.

cies become more sparse for the examples with the positive class (left) and the negative class (right) with a profound decrease in the positive class. The difference between classes also imply that models can learn more frequency shortcuts for a particular class. As explained in [44], these are tied with textures pertaining specifically to those frequencies. Therefore, we can conclude a classification relying on more semantic information has been made. The same effect was seen in Robustness to Fourier Attacks in Fig. 5.

This reduction implies a heightened robustness to changes or deletions in these frequencies, indicative of the method's efficacy in mitigating the adverse effects of frequency shortcuts. Furthermore, intriguingly, our findings suggest that models incorporating AFA demonstrate a remarkable ability to maintain performance even when certain frequencies are omitted or altered, thereby emphasizing the



(a) Standardly Trained



(b) Trained with AFA

Figure 11. Comparison of DFMs for a standardly trained model and model trained with AFA for a binary classification task

significance of our proposed augmentation technique in enhancing model resilience and generalization capabilities.

6. Future Work

The exploration of future avenues in our research encompasses several intriguing possibilities. Firstly, for semantic segmentation, further experiments are warranted to validate our assertion that models now leverage enhanced semantic information. This investigation can offer deeper insights into the mechanisms through which models process and incorporate semantic cues, thereby refining our understanding of their evolving capabilities in image analysis.

Moreover, the application of our proposed method extends beyond image-based tasks. The realm of NLP, encompassing sequence-based tasks (like generative AI), shares common challenges with image processing, particularly in confronting frequency shortcuts. While tokenization in text presents a distinct challenge for such augmentation, future work should delve into adapting our framework to these domains, exploring how similar strategies can be employed to enhance robustness and generalizability in the face of data corruption.

Furthermore, our current study predominantly provides an empirical analysis of regularization without delving into rigorous mathematical proofs. Future investigations could delve into formalizing the underlying principles, substantiating our findings with mathematical rigour to strengthen the theoretical foundation of the proposed techniques.

While our proposed method has demonstrated proficiency in overcoming frequency shortcuts and pinpointing frequency biases, a critical dimension yet to be explored involves the editing of these biases. Specifically, we have not delved into the feasibility of interventions aimed at mitigating or eliminating correlations introduced by identified biases. Future research should investigate strategies for editing frequency biases, addressing the intriguing challenge of potentially modifying or removing correlations to enhance the overall interpretability and fairness of neural networks.

Lastly, an intriguing avenue for future research lies in the exploration of unsupervised learning scenarios, where targets are inherently absent. Employing (information-theoretic) losses within the framework and training setup we propose could shed light on novel approaches to unsupervised learning, offering valuable insights into the inherent structure and representations learned by models in the absence of explicit target guidance. These suggested directions collectively contribute to the ongoing advancement of our understanding of model behaviour and performance in diverse domains.

7. Conclusions

We proposed an efficient data augmentation technique called AFA, which complements existing visual augmentation techniques by filling the augmentation gap, that they do not cover in the Fourier domain. AFA perturbs the frequency components of images and generates adversarial samples. By leveraging Fourier-basis functions and the auxiliary augmentation setting we demonstrate that AFA allows the models to learn from aggressive/adversarial input changes. We performed extensive experiments on benchmark datasets, and demonstrated that AFA benefits the robustness of models against common image corruptions, the consistency of predictions when facing increasing perturbations, and the OOD generalization performance. The promising results underscore AFA's potential in fortifying models against frequency shortcuts as well, offering a valuable enhancement to their adaptability and performance stability. Being complementary to other augmentation techniques, AFA can further boost the robustness of models, especially against strong corruptions and perturbation, and it also results in better robustness in the frequency spectrum. We foresee that investigating the use of Fourier-basis functions on the training process of neural networks would provide promising improvement to model performance, thus encouraging their reliability in real scenarios.

References

[1] 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, 2021. 3, 4, 5, 8, 10

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 2

[3] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data, 2019. 2, 4

[4] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc., 2020. 4

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 9

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*, Oct. 2020. 5

[7] Fartash Faghri, Hadi Pouransari, Sachin Mehta, Mehrdad Farajtabar, Ali Farhadi, Mohammad Rastegari, and Oncel Tuzel. Reinforce data, multiply impact: Improved model accuracy and robustness with dataset reinforcement, 2023. 2

[8] Zhiqiang Gao, Kaizhu Huang, Rui Zhang, Dawei Liu, and Jieming Ma. Towards better robustness against common corruptions for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18882–18893, October 2023. 2

[9] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut Learning in Deep Neural Networks. *arXiv*, Apr. 2020. 6

[10] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. Mixgen: A new multi-modal data augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 379–389, January 2023. 2

[11] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the Big Data Paradigm with Compact Transformers. *arXiv*, Apr. 2021. 18

[12] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers, 2022. 9

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 5, 9

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2021. 2, 9

[15] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. 2, 9, 18

[16] 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. *arXiv*, Dec. 2019. 2, 4, 8, 17

[17] Ignacio Hounie, Luiz F. O. Chamon, and Alejandro Ribeiro. Automatic data augmentation via invariance-constrained learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 13410–13433. PMLR, 23–29 Jul 2023. 2

[18] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv*, Feb. 2015. 5

[19] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation, 2022. 2, 9

[20] Ildoo Kim, Younghoon Kim, and Sungwoong Kim. Learning loss for test-time augmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4163–4174. Curran Associates, Inc., 2020. 4

[21] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 9

[22] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). 9

[23] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015. 9

[24] Xiu-Chuan Li, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. F-mixup: Attack cnns from fourier perspective. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 541–548, 2021. 3

[25] Chang Liu, Wenzhao Xiang, Yuan He, Hui Xue, Shibao Zheng, and Hang Su. Improving model generalization by on-manifold adversarial augmentation in the frequency domain, 2023. 3, 5

[26] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey, 2023. 3

[27] Siao Liu, Zhaoyu Chen, Yang Liu, Yuzheng Wang, Dingkang Yang, Zhile Zhao, Ziqing Zhou, Xie Yi, Wei Li, Wenqiang Zhang, and Zhongxue Gan. Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23436–23446, October 2023. 2

[28] Yang Liu, Shen Yan, Laura Leal-Taixé, James Hays, and Deva Ramanan. Soft augmentation for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16241–16250, June 2023. 2

[29] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In Shai

Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 549–566, Cham, 2022. Springer Nature Switzerland. 3

[30] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 549–566, Cham, 2022. Springer Nature Switzerland. 5

[31] Guozheng Ma, Linrui Zhang, Haoyu Wang, Lu Li, Zilin Wang, Zhen Wang, Li Shen, Xueqian Wang, and Dacheng Tao. Learning better with less: Effective augmentation for sample-efficient visual reinforcement learning, 2023. 2

[32] Juliette Marrie, Michael Arbel, Diane Larlus, and Julien Mairal. Slack: Stable learning of augmentations with cold-start and kl regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24306–24314, June 2023. 2

[33] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness, 2021. 2, 9

[34] Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. PRIME: A few primitives can boost robustness to common corruptions. *arXiv*, Dec. 2021. 2, 4, 13

[35] Samuel G. Müller and Frank Hutter. TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation. *arXiv*, Mar. 2021. 2, 4, 6

[36] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the Spectral Bias of Neural Networks. *arXiv*, June 2018. 6

[37] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. 2, 9

[38] Tonmoy Saikia, Cordelia Schmid, and Thomas Brox. Improving robustness against common corruptions with frequency biased models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10211–10220, October 2021. 3

[39] Ryan Soklaski, Michael Yee, and Theodoros Tsiligkaridis. Fourier-Based Augmentations for Improved Robustness and Uncertainty Calibration. *arXiv*, Feb. 2022. 3, 5

[40] Teppei Suzuki. Teachaugment: Data augmentation optimization using teacher knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10904–10914, June 2022. 4

[41] An Wang, Mobarakol Islam, Mengya Xu, and Hongliang Ren. Curriculum-based augmented fourier domain adaptation for robust medical image segmentation. *IEEE Transactions on Automation Science and Engineering*, pages 1–13, 2023. 3, 4, 5

[42] Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. AugMax: Adversarial Composition of Random Augmentations for Robust Training. *arXiv*, Oct. 2021. 2, 3, 4, 5, 7, 8, 9, 12, 18

[43] Shunxin Wang, Christoph Brune, Raymond Veldhuis, and Nicola Strisciuglio. DFM-x: Augmentation by leveraging prior knowledge of shortcut learning. In *4th Visual Inductive Priors for Data-Efficient Deep Learning Workshop*, 2023. 3, 4, 5

[44] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. Frequency Shortcut Learning in Neural Networks. *OpenReview*, Oct. 2022. 6, 8, 14

[45] 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. *arXiv*, May 2023. 2, 4

[46] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[47] Qinwei Xu, Ruipeng Zhang, Ziqing Fan, Yanfeng Wang, Yi-Yan Wu, and Ya Zhang. Fourier-based augmentation with applications to domain generalization. *Pattern Recognition*, 139:109474, 2023. 3, 4, 5

[48] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A Fourier Perspective on Model Robustness in Computer Vision. *arXiv*, June 2019. 3, 4, 10, 18

[49] Mehmet Kerim Yucel, Ramazan Gokberk Cinbis, and Pinar Duygulu. Hybridaugment++: Unified frequency spectra perturbations for model robustness, 2023. 2

[50] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019. 4

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

[52] Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the Robustness of Deep Neural Networks via Stability Training. *arXiv*, Apr. 2016. 2

A. Implementation Details

Below, we report the training setup in detail. For all methods, and a particular dataset and architecture, the same training setup was used unless stated otherwise.

Convolution Neural Networks For CIFAR-100 and Tiny ImageNet we use the SGD optimiser with an initial learning rate of 0.2, Nesterov momentum of 0.9 with a batch size of 128 training for 100 epochs. We use a weight decay of 0.0005 and we do not decay the affine parameters of normalisation. For CIFAR-10, we follow the same setup as above, except we train for 200 epochs with a batch size of 256 and an initial learning rate of 0.1. The learning rate is decayed with a cosine annealing schedule to 0 which is stepped step-wise. For all models, we always employ the standard transformation of random crop with a padding of 4 and random horizontal flip. For ImageNet, we follow [16]

in that we use SGD optimiser with an initial learning rate of 0.1 and Nesterov momentum of 0.9 and train for 90 epochs. We use a weight decay of 0.0001 and we do not decay affine parameters of normalisation. The learning rate decays with a by a factor of 0.1 every 30 epochs. For all models, we employ the standard transformation of random resized crop to image size of 224×224 with bilinear interpolation and random horizontal flip, before other augmentations. We choose to train all models from scratch (no fine-tuning using AFA) so that we can study the effects of AFA without other underlying factors. Therefore, for fair comparison, we retrain PRIME from scratch as well using our setup. For models trained with JSD, we follow [42] for the regularising coefficient, mainly: $\lambda = 10$ for CIFAR-10 and Tiny ImageNet, $\lambda = 1$ for CIFAR-100 and $\lambda = 12$ for ImageNet.

Compact Convolution Transformer For CIFAR-10/100 and ImageNet we also train a transformer architecture. For all datasets we use CutMix (alpha=1.0) and MixUp (alpha=0.2 for ImageNet and alpha=1.0 for CIFAR-10/100) with an equal chance of applying one of the two. For CIFAR-10/100, we follow [11]. We train using the AdamW optimiser with max learning rate of 0.0006 and weight decay of 0.06, and we do not decay the affine parameters of the normalisation modules. We train with an effective batch size of 256, and apply learning rate decay following a cosine decay with a warm-up period of 10 epochs and the learning rate scheduler is stepped step-wise. For ImageNet, we use a max learning rate of 0.0005, effective batch size of 1024 and a weight decay of 0.05. The learning rate decay follows a cosine annealing schedule with a warm-up of 25 epochs. The same standard transformations as for convolutional neural networks were applied.

B. Evaluation metrics

Mean corruption error (mCE) measures the robustness of models against image corruptions [15], computed as:

$$\mathrm{mCE} = \frac{1}{|C|} \sum_{c \in C} \frac{\sum_{s=1}^{5} E_{s,c}^{f}}{\sum_{s=1}^{5} E_{s,c}^{baseline}}, \tag{9}$$

where the sum of classification error E of five severity $s \in \{1, 2, 3, 4, 5\}$ per corruption c of model f is normalized by that of a baseline model. The normalized classification errors of all corruptions C in the dataset are averaged to obtain mCE. We use AlexNet as baseline in ImageNet experiments and ResNet-18 for Tiny ImageNet. For CIFAR-10/100 there are no baselines advised so we do not report the mCE for these datasets.

Mean flip rate (mFR) evaluates the consistency of model predictions with increasing perturbations [15], computed as follows:

$$\mathrm{mFR} = \frac{1}{|C|} \sum_{c \in C} \mathrm{FR}_c^f = \frac{1}{|C|} \sum_{c \in C} \frac{\mathrm{FP}_c^f}{\mathrm{FP}_c^{baseline}}, \tag{10}$$

with

$$\mathrm{FP}_c^f = \frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{j=2}^{n} \mathbb{1}(f(x_j^{(i)}) \neq f(x_{j-1}^{(i)})). \tag{11}$$

$\mathbb{1}(f(x_j^{(i)}) \neq f(x_{j-1}^{(i)}))$ measures whether the prediction of the model f on a frame x_j is the same as its previous perturbed frame in the i^{th} sequence. If the predictions are the same, $\mathbb{1}(f(x_j^{(i)}) \neq f(x_{j-1}^{(i)}))$ equals to zero, and thus the performance of the model is not affected by the considered perturbations. FP_c^f measures the consistency of predictions over m perturbed sequences, each with n of frames. For a sequence corrupted by noise, the predictions are compared with those of the first frame, as noise is not temporally related. The mFR is obtained by averaging the normalized FP_c^f by that of a baseline model across all the perturbations C. The value of mFR is expected to be close to zero for a robust model.

Mean top-5 distance (mT5D) also measures the consistency of model predictions in terms of increasing perturbations [15]. For a robust model, the top-5 predictions of frames over a sequence should be relevant to those of the previous frames in the sequence. The top-5 distance thus measures the inconsistency of top-5 predictions under consecutive perturbations, computed as follows:

$$\mathrm{T5D}_c^f = \frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{j=2}^{n} d(\tau(x_j), \tau(x_{j-1})), \tag{12}$$

with

$$d(\tau(x_j), \tau(x_{j-1})) = \sum_{i=1}^{5} \sum_{j=min\{i,\rho(i)\}+1}^{max\{i,\rho(i)\}} \mathbb{1}(1 \leq j-1 \leq 5), \tag{13}$$

where $\rho(\tau(x_j)(k)) = \tau(x_{j-1})(k)$, $\tau(x_j)$ is the ranking of predictions for a perturbed frame x_j and $\tau(x_j)(k)$ indicates the rank of the prediction being k. If $\tau(x_j)$ and $\tau(x_{j-1})$ are the same, then $d(\tau(x_j), \tau(x_{j-1})) = 0$. Averaging the normalized T5D by that of the baseline over all corruptions obtain $\mathrm{mT5D} = \frac{1}{|C|} \sum_{c \in C} \frac{\mathrm{T5D}_c^f}{\mathrm{T5D}_c^{baseline}}$.

Fourier heatmap evaluates model robustness from a Fourier perspective [48] exploiting Fourier basis functions to perturb test images and measuring the classification error of models. They are constructed as follows. Let $U_{i,j} \in \mathbb{R}^{d_1 \times d_2}$ be a real-valued matrix such that its norm equals

-	Main	Auxiliary	TIN-C SA↑	TIN-C RA↑	TIN-C mCE↓
ResNet18	-	✗	63.56	25.86	97.34
	AFA	✗	59.04	28.87	93.45
	-	AFA	62.52	33.35	87.58
	AugMix	✗	62.95	36.26	84.05
	AugMix	AFA	62.51	38.67	80.83
	AugMix†	✗	**64.65**	36.30	83.90
	AugMix†	AFA	64.34	38.52	80.79
	PRIME	✗	63.07	39.67	79.42
	PRIME	AFA	62.48	41.09	77.55
	PRIME†	✗	63.24	41.22	77.44
	PRIME†	AFA	62.65	**43.00**	**73.11**

Table 6. Results for TIN-C with ResNet18. Models with † use loss with JSD.

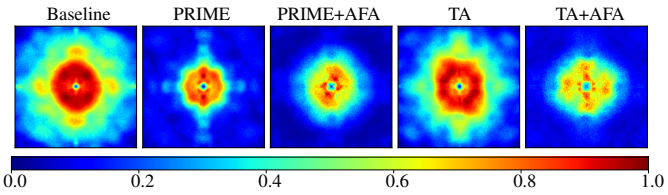


Figure 12. Fourier heatmaps of CCT trained with standard setting, PRIME, PRIME+AFA, TA and TA+AFA.

to 1. The Fourier transform of $U_{i,j}$ has only two non-zero elements located at (i, j) and the corresponding symmetric coordinate with respect to the image center. Given an image X, a perturbed image with Fourier basis noise can be generated by $\tilde{X}_{i,j} = X + rvU_{i,j}$, where r is chosen randomly from a uniform distribution ranging from -1 to 1, and v controls the strength of the added noise. Each channel of the images is perturbed independently with different r and v. The model robustness against Fourier basis noise $U_{i,j}$ is evaluated by the classification error, and the final outcome is in a form of heatmap which records the error of the evaluated model under different Fourier basis noise. Examples are in Fig. 12.

C. Supplementary results

C.1. Results on Tiny ImageNet

In Tab. 6 we provide the robustness results on Tiny ImageNet (TIN), which are consistent with those presented on other datasets. Models trained with AFA show robustness improvements consistently by significant margin with only negligible reduction of the clean accuracy. We again see that JSD improves robustness slightly, and in AugMix it improves clean accuracy greatly.

C.2. Robustness in the frequency spectrum.

The Fourier heatmaps of CCT trained with standard setting, PRIME, PRIME+AFA, TA and TA+AFA are provided in Fig. 12. Our observations are consistent with those in the main paper. Also CCT models trained with the contribution of AFA have better robustness to low and middle-high frequency corruptions.

C.3. Robustness per corruption severity.

We report the classification error of models tested under corruptions with different severity levels Fig. 13. The models trained with AFA have consistently lower error than their counterpart trained without AFA, showing that AFA can further boost the robustness of models against common image corruptions, especially in difficult testing conditions with high severity. Figures provided by Shunxin Wang, MSc.

C.4. Robustness to each image corruption.

Furthermore, we show the classification error averaged over five corruption severity levels per corruption type in Fig. 15. The error points of model trained with visual augmentations only, and with further use of AFA are connected by a line. A downward trend means models trained with AFA have better robustness performance on specific corruption types. We observe that, in general, models with AFA have better corruption robustness than models trained only with visual augmentations. Significant improvements are especially evident on noise corruptions (Gaussian noise, impulse noise, iso noise, plasma noise, shot noise, single frequency grayscale noise and cocentric sine waves). One exception is ResNet50 trained with AugMix and AFA, for which the model trained without AFA performs better except on few cases. This can be attributed to the less training time (90 epoch vs 180 epochs) than that of ResNet50+AugMix.

Figures provided by Shunxin Wang, MSc.

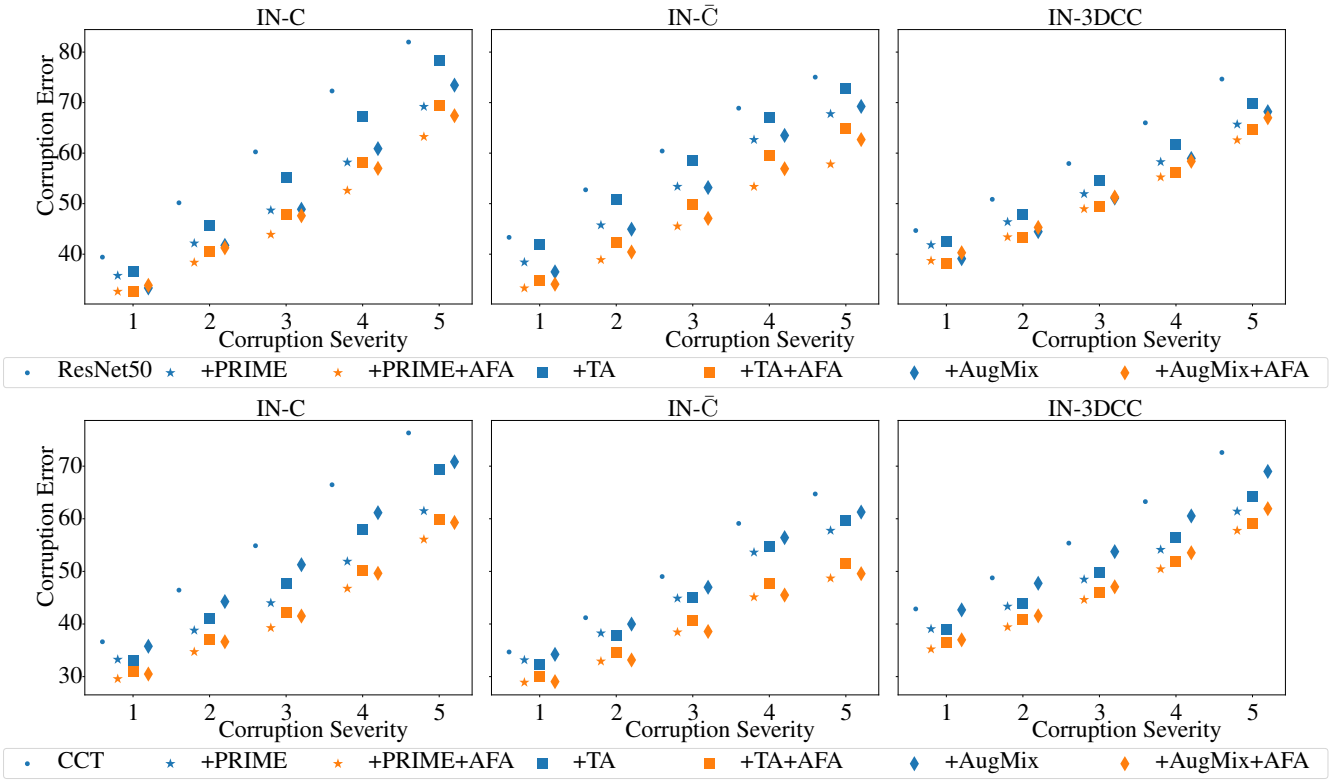


Figure 13. Corruption error of ResNet50 and CCT trained with PRIME, PRIME+AFA, TA, TA+AFA, AugMix and AugMix+AFA. Models trained with AFA (orange points) have lower error at each severity than their counterpart trained with only visual augmentation (blue points), demonstrating the benefit of AFA to corruption robustness.

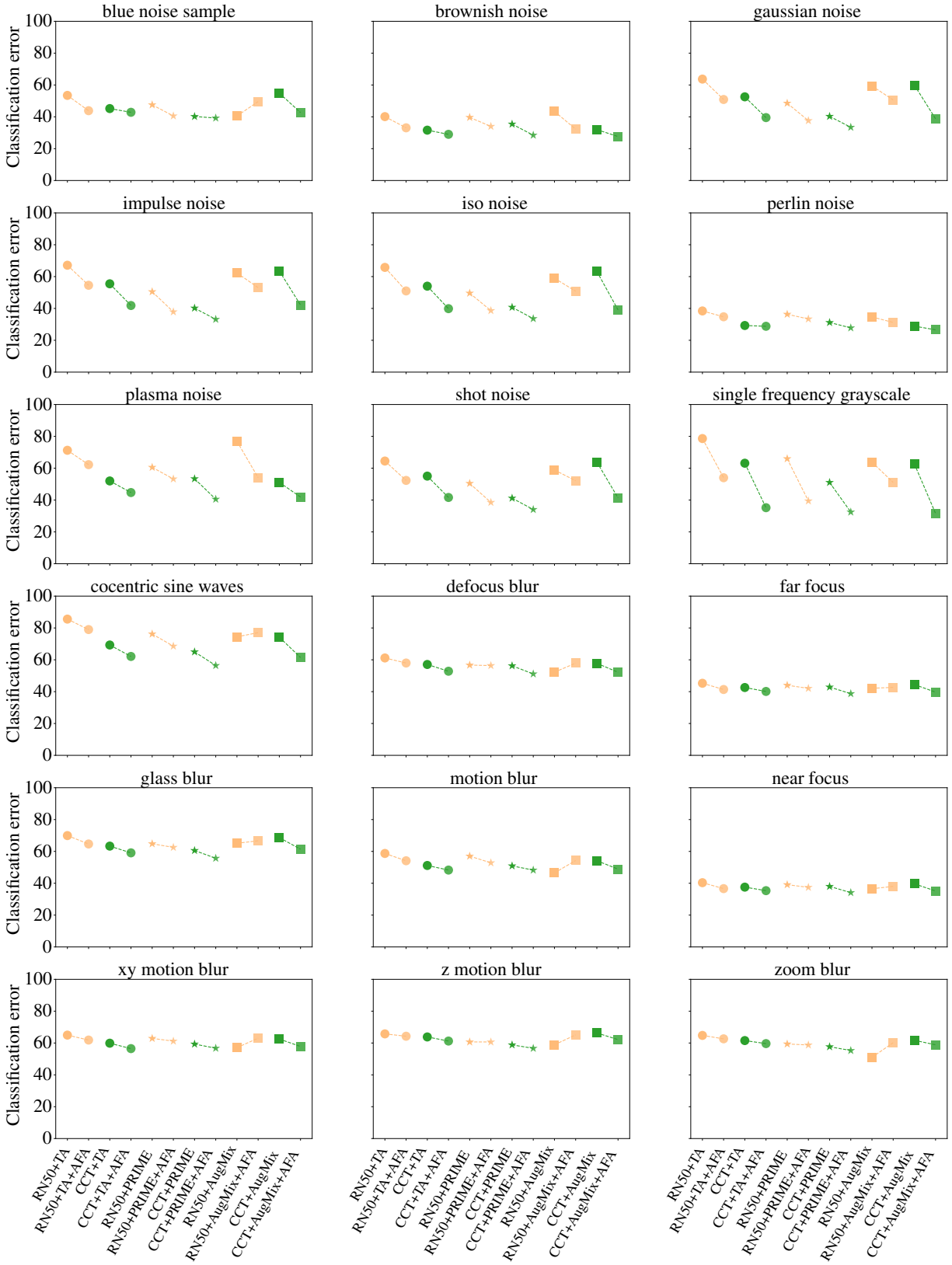


Figure 14. Averaged classification error per corruption of ResNet50s (orange) and CCTs (green). The error points of model trained with visual augmentations and additionally with AFA are connected. A decreasing line indicates better performance when trained additionally with AFA (a).

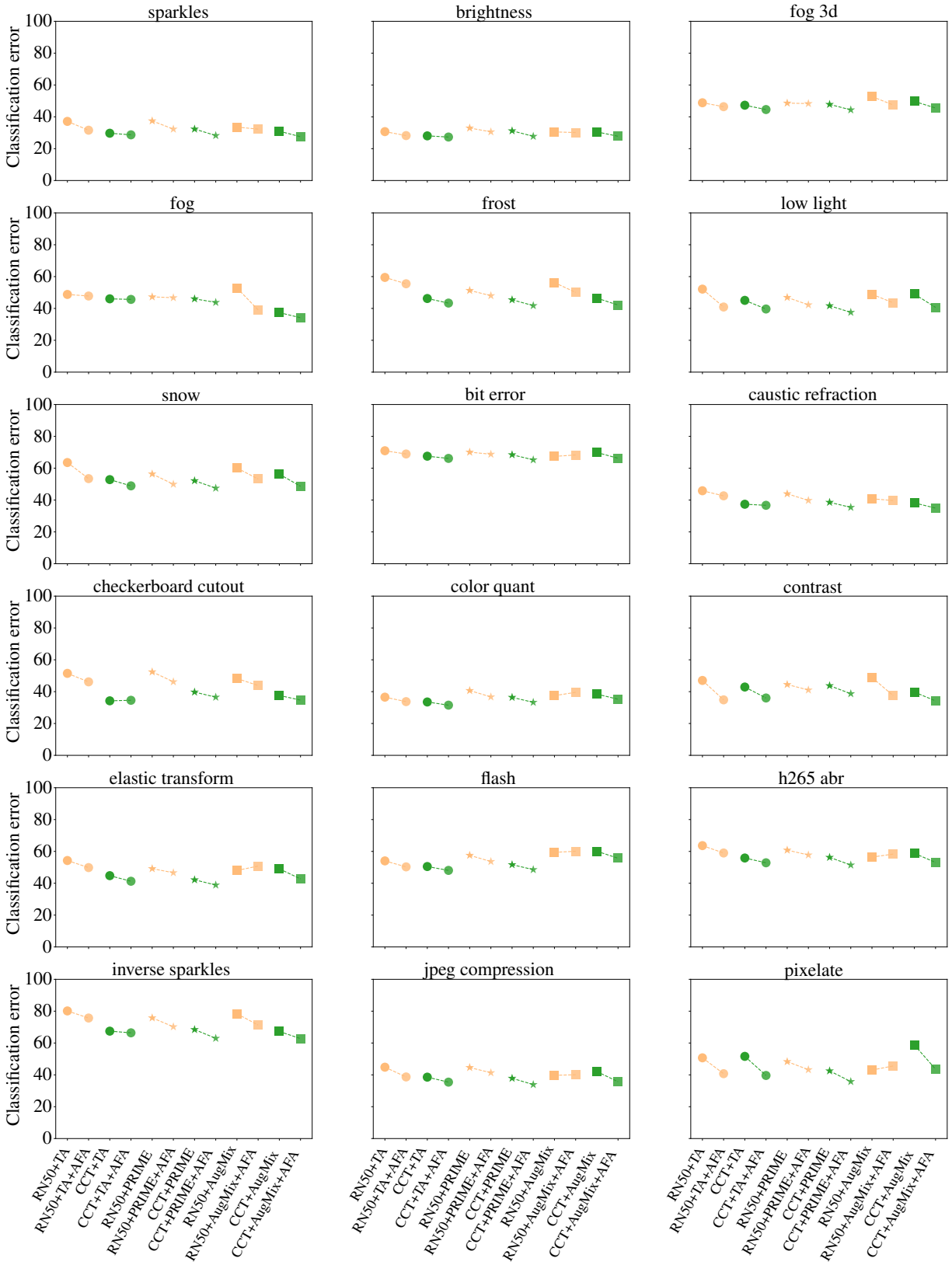


Figure 15. Averaged classification error per corruption of ResNet50s (orange) and CCTs (green). The error points of model trained with visual augmentations and additionally with AFA are connected. A decreasing line indicates better performance when models are trained additionally with AFA (b).