Effectiveness of Fourier-Basis Noise on Improving Adversarial Robustness

YIFAN SUN, University of Twente, The Netherlands

ABSTRACT

Data augmentation is an important tool to improve the robustness of a model against adversarial attacks. This study is to evaluate the performance of the model trained with Fourier basis noise in terms of robustness against different adversarial attacks. The evaluation will mainly focus on the robustness of the model's accuracy. The results show that Fourier-basis augmentation has improved performance in robustness against FGSM, and PGD attacks compared to the baseline model. Furthermore, compare the performance of the Fourier-basis noise trained model with other defense mechanisms in terms of accuracy in robustness, demonstrating the positive effects of Fourier-basis augmentation to some extent.

Additional Key Words and Phrases: Data augmentation, Fourier-Basis noise, Robustness, Adversarial Attack

1 INTRODUCTION

Deep learning is a powerful and promising branch of machine learning that uses complex and flexible artificial neural networks to learn from data and perform various tasks, such as image recognition [29], natural language processing [3], speech synthesis [12], etc. However, it is found that machine learning models that perform well on real data are very vulnerable in the face of adversarial attacks [26]. An adversarial attack is the generation of adversarial examples by adding slight perturbations to an image that are imperceptible to the human visual system, fooling the deep neural network, and changing the classification of the image. Common adversarial attacks include FGSM [8], PGD [16], C&W [2] etc., which can be dangerous for high-stake applications such as autonomous vehicles [11], spam filters [21], face recognition [6], etc. For instance, if adversarial examples are applied to street signs and cones, the perception system of an autonomous vehicle may encounter difficulties and make errors in recognizing the obstacles. Consequently, the vehicle might mistakenly ignore the stop sign, leading to a collision with the cone. Therefore, there is a need to improve the robustness of machine learning models. Data augmentation techniques, like adversarial training [8] and various image transformations (e.g., flipping, cropping, adding random noise) including stylized image transformation [7], can be applied to the training data to enhance model robustness.

Traditional image augmentations mainly operate in the spatial domain, such as rotation, crop, flip, etc., which require manual designing and fixing the transformation for each dataset by experts. It can be inefficient when dealing with huge datasets. Therefore, some new augmentation methods, such as AutoAugment [4], and MixAug [10]. is proposed. Among that, Soklaski et al. [24] introduced Fourier-based augmentation by adding Fourier-basis noise to the AugMix image augment framework. This approach can improve model robustness by allowing for targeted distribution changes through customization of the AugMix framework. Zeng [31] demonstrated that using additive Fourier base noise can improve the robustness of convolutional networks against common computer vision corruptions. However, it is not known whether this method also performs well in improving the robustness of the model against adversarial attacks. This paper aims to investigate the problem by conducting experiments on the CIFAR-10 dataset to evaluate accuracy. Additionally, we will compare the performance of Fourier-basis augmentation with other defense methods.

Goal: Evaluate whether Fourier-basis data augmentation helps improve the robustness of models under adversarial attacks.

- **RQ1**: What is the impact of Fourier-basis noise on the robustness of a model against different adversarial attacks?
  - **RQ1.1**: How does the model trained with Fourier base noise perform in terms of robustness with respect to accuracy?
  - **RQ 1.2**: Which Fourier-basis augmentation policy improves the adversarial robustness most?
- **RQ2**: How do other defense mechanisms perform compared to Fourier-basis Data Augmentation in terms of adversarial robustness?

To answer **RQ1.1**, we compare the three Fourier-basis trained models with the baseline model and evaluate the accuracy in the face of adversarial attacks, which provides evidence to **RQ1.2**. Based on the answer of **RQ1.2**, we conduct the experiments by applying different defense methods to answer **RQ2**.

2 RELATED WORK

2.1 Adversarial Attacks

Fast Gradient Sign Method (FGSM) [8] is designed to quickly find the direction of the anti-perturbation in a given input sample, increasing the likelihood of model misclassification. Iterative-FGSM (IFGSM) [13] expands on FGSM by introducing two iterative methods, namely the Basic-iterative method, and the Iterative least-likely class method. The authors conducted the experiment by printing multiple pairs of clean and adversarial examples to verify whether the cell phone camera can successfully detect the QR code in the corner. Based on IFGSM, Dong et al. [5] proposed a method called introduced momentum iterative gradient-based methods (MI-FGSM) that can improve the success rate of the generated adversarial samples MI-FGSM. The method accumulates the gradient of the loss function at each iteration to stabilize the optimization and avoid undesirable local maxima. They evaluated L_∞ norm bound for non-target attacks, the effect of the number of iterations on the success rate. Unlike FGSM, which only performs one iteration with a large step, Projected Gradient Descent (PGD) [16] performs multiple iterations, each time the step is small. The study of C&W [2] demonstrated that the adversarial perturbations can be generated by solving a norm-restricted constrained optimization problem:

$$min\|\rho\|_p + c \cdot f(x+\rho), s.t. x+\rho \in [0,1]^m$$

TScIT 39, July 7, 2023, Enschede, The Netherlands

Carlini and Wagner also show that C&W is sufficient to pass the defensive distillation [20] methods.

2.2 Adversarial Training

Goodfellow et al. [8] proposed adversarial training as a defense method, which is adding the generated adversarial examples to the train set as data augmentation. They proposed FGSM attack to speed the generation of adversarial examples for training. However, Seeyed-Mohensen [18] found that FGSM adversarial training does not always enhance the adversarial robustness of the model. Madry et al. [16] proposed PGD attack to conduct adversarial training. The method is to find a model θ such that it can correctly classify adversarial examples with perturbation δ within a certain range S.

$$min_{\theta}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\|\delta\leq\epsilon\|} L(f_{\theta}(X+\delta),y)\right],$$

where (x, y) denotes the original data and the corresponding labels, $\mathcal{D}$ denotes the distribution of the data, L denotes the loss function. FreeAT [23] modified on PGD, which uses the gradient obtained each time to update both the perturbation and the parameters.

2.3 Data Augmentation

Unlike traditional data automation techniques that require manual design, AutoAugment(AA) [4], can automatically search for optimized data augmentation policies, which performed well in image classification. AA consists of algorithm and search space, where Reinforcement Learning is employed as a search algorithm. In the search space, each policy consists of 5 sub-policies, for each of them two image operations are applied sequentially. The operation depends on the probability of applying the operation and the magnitude of the operation. The operations are ShearX/Y, TranslateX/Y, Rotate, AutoContrast, Invert, Equalize, Solarize, Posterize, Contrast, Color, Brightness, Sharpness, Cutout, and Sample Pairing. The search algorithm is composed of the recurrent neural network as a controller and the Proximal Policy Optimization algorithm [22] as a training algorithm. The RNN controller is used to sample an augmented policy which will be used to train the model. And then the validation accuracy can be used to update the controller as a reward. However, AA requires a huge amount of time, which is not efficient. To deal with this problem, Lim et al [14]. proposed Fast AutoAugment(FAA), which applies Bayesian Optimization [1] to explore the policy. Instead of repeated training child models, FAA uses a single model to find the improved augmentation strategies between the one distribution of augmented split and another distribution of unaugmented split.

2.4 Fourier-based Augmentation

Yin et al. [30] proposed a method that can generate a perturbed image with Fourier-basis noise.

$$\widetilde{X_{i,j}} = X + rvF(U_{i,j})$$

Where X is the original image, $F(U_{i,j})$ has at most two non-zero elements located at (i, j) with symmetric coordinates with respect to the image center, r is a random number from {-1,1}, $v > 0$ denotes the norm of perturbations.

2.5 Relationship Between Frequency and Adversarial Robustness

Some studies analyze adversarial examples from the signal-processing perspective of frequencies. Tsuzuku and Sato [27] found that convolution networks are sensitive to the direction of Fourie basis functions. Based on their work, Yin et al. [30] examined the model sensitivity to additive noise, and pointed out that the adversarial perturbations of a naturally trained model are more high frequency, whereas for the adversarial training the adversarial perturbations towards the lower frequencies. Guo et al. [9] demonstrated that attacks on images in a black-box environment can be achieved by especially perturbing the low-frequency part of the input signal. Lorenz et al. [15] showed that frequency features can be used to detect adversarial attacks. Wang et al. [28] conjectured that high-frequency components may be associated with adversarial attacks. Maiya et al.[17] claimed that adversarial examples cannot be simply classified as either high-frequency or low-frequency phenomena.

3 METHODS

3.1 Fourier-basis Noise Augmentation Policies

Based on the method mentioned in Section 2.3, Zeng [31] designed a method to search for augmentation policies. The noise is divided into 22 groups based on the radius of central noise, which leads to 22 different transformation options. The search space is composed of options, the possibility of adding each noise frequency and added noise magnitudes. The search space is increased by dividing noise by frequency and phase, where the phase indicates the difference between the phase representation signal and the standard phase reference. By dividing the phases of the same frequency into quadrants every 90° or 45°, the number of candidates can be raised to 85 or 165 accordingly. To get the optimal augmentation policy, FAA is employed.

To evaluate the effectiveness of Fourier-noise augmentation, we employed the methods described above to search the Fourier-basis augmentation policy. The FB1, FB2, and FB3 are selected by the search strategy, and the applied search spaces contained 22, 85, and 165 transformation candidates, respectively. Table 1 presents the results of searched augmentation Policy FB1.

	Operation1	**Operation2**
1	FB_5(0.4,4.24)	FB_15(0.71,1.64)
2	FB_9(0.02,2.2)	FB_22(0.86,3)
3	FB_4(0.77,2.48)	FB_8(0.55,3.45)
4	FB_15(0.2,1.31)	FB_21(0.64,2.94)
5	FB_13(0.39,4.56)	FB_3(0.96,3.59)
6	FB_2(0.53,4.39)	FB_15(0.7,3.58)
7	FB_22(0.47,4.63)	FB_14(0.18,2.21)
8	FB_2(0.66,3.68)	FB_9(0.1,4.26)

Table 1. Selected augmentation Policy FB1. FB_5(0.4,4.24) denotes the addition of a 4.24-magnitude Fourier-basis noise in group 5 with a probability of 0.4. [31]

4 EXPERIMENT

This section introduces two experiments, which are aimed at investigating the effectiveness of Fourier-Basis noise on improving the adversarial robustness of a model. The first experiment focused on the effect of Fourier-basis noise augmentation on adversarial attacks. The second experiment compares the performance of Fourier-basis noise augmentation with other adversarial training methods in the face of adversarial attacks.

4.1 Experiment 1: Evaluating Fourier-basis Noise Augmentation

The experiment compares the performance of four models against FGSM, PGD L_∞, and PGD L_2 on the CIFAR-10 Dataset. Because this study is based on Zeng's [31] previous work, we employed the same architecture for training. The baseline model is Wide Resnet-28-10 without augmentation. The other three models applied searched augmentation policies introduced in Section 3.1, named **FB1**, **FB2**, and **FB3**.

4.1.1 Augmentation Setup. The transformation for all models comprises padding, random horizontal flipping, and random cropping before adding Fourier-basis noise chosen according to the policies.

4.1.2 Training. The training procedure is implemented in PyTorch. The Wide Resnet-28-10 architecture is used. The optimizer uses Adam, has a learning rate of 0.0001, a weight decay of 1e-4, and employs cross-entropy loss as the loss function. Every experiment contains 100 training epochs, in case the validation loss does not improve after 30 epochs, the training process will stop early. The training and validation sets of the CIFAR-10 data are split 90:10 between the training and validation data. Nividia A16 is used during the training process.

4.1.3 Generate Adversarial Perturbations. The adversarial robust toolbox library [19] is used to generate attacks. Due to the time limitation, the adversarial attacks FGSM, PGD L_∞, and PGD L_2 are involved in this experiment.

For FGSM, 10 sets of perturbations are generated based on epsilon values from 1/255 to 10/255 sequentially. For PGD L_∞, the step size is set to 2/255, the maximum iteration is 10, and the random initial number is 0, the epsilon also increases from 1/255 to 10/255. For PGD L_2, we fix the value of epsilon to 128/255, set the step size to 0.05, and the maximum iteration to 100.

4.1.4 Testing. During the test, we add generated perturbations to the test sets and then evaluate the accuracy of the prediction.

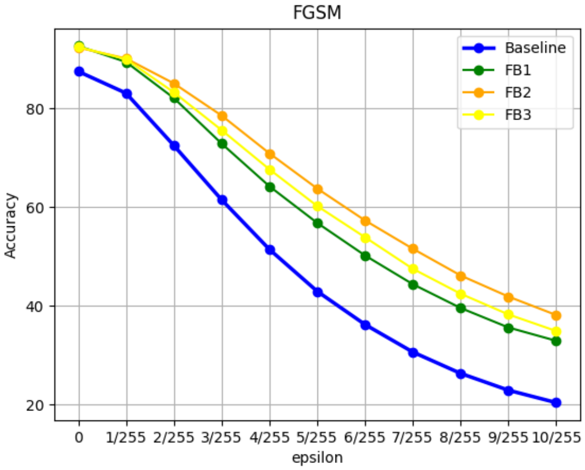
4.2 Experiment 2: Comparing with other works

This experiment compares the Fourier-basis noise augmentation with other adversarial training methods, such as FGSM adversarial training, and PGD adversarial training.

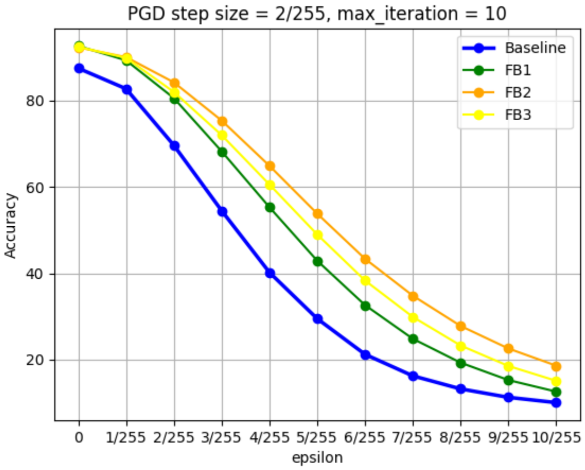
4.2.1 Training. The training processes are slightly different from Section 4.1.2. After getting input and target from the batch and setting them to the device, we add FGSM perturbation with epsilon 8/255 for FGSM adversarial training. And we set epsilon to 8/255, iteration to 7, and α to 0.00784 for the PGD method.

4.2.2 Testing. The processes of generating adversarial attacks and testing are identical to Experiment 1.

5 RESULT



(a)



(b)

Fig. 1. The accuracy of four models against (a) FGSM and (b) PGD L_∞ (b) attack with different epsilon

Table 2 shows the four models' tested accuracy for clean images, FGSM, and PGD L_∞ attacks with various epsilon values. **FB1**, **FB2**, and **FB3** all improve the accuracy against FGSM and PGD L_∞ attacks compared to the baseline model. Among them, the highest clean accuracy occurs in **FB1**, while **FB2** performs best in the face of all adversarial attacks. To visualize the results, we represent the results as a line plot in Fig.1.

(a) Test accuracy on clean images and FGSM adversarial examples of 4 models.

model	acc	FGSM									
	(clean)	1/255	2/255	3/255	4/255	5/255	6/255	7/255	8/255	9/255	10/255
Baseline	87.52	83.11	72.51	61.5	51.44	42.94	36.21	30.63	26.3	22.89	20.38
FB1	**92.68**	89.43	82.12	72.89	64.23	56.81	50.21	44.38	39.56	35.61	32.91
FB2	92.36	**90.14**	**85.09**	**78.57**	**70.91**	**63.77**	**57.31**	**51.57**	**46.15**	**41.87**	**38.13**
FB3	92.48	90.0	83.28	75.62	67.67	60.23	53.9	47.5	42.45	38.25	34.89

(b) Test accuracy on clean images and PGD L_∞ and PGD L_2 adversarial examples of 4 models.

model	PGD L_∞										PGD L_2
	1/255	2/255	3/255	4/255	5/255	6/255	7/255	8/255	9/255	10/255	128/255
Baseline	82.76	69.66	54.51	40.2	29.47	21.21	16.22	13.2	11.24	10.02	44.36
FB1	89.35	80.6	68.26	55.31	42.9	32.62	24.81	19.27	15.24	12.57	59.77
FB2	**90.09**	**84.25**	**75.4**	**64.91**	**53.81**	**43.35**	**34.8**	**27.78**	**22.56**	**18.58**	**68.69**
FB3	89.93	82.02	72.02	60.57	48.98	38.32	29.88	23.25	18.51	15.07	64.44

Table 2. Accuracy of each model tested on the CIFAR-10 dataset for clean images and the added adversarial perturbations, where FGSM with epsilon = [1/255,...10/255], PGD L_∞ with epsilon = [1/255,...10/255], step size = 2/255, iteration = 10, and PGD L_2 with epsilon = 128/255, step size = 0.05, iteration = 100. The highest accuracy is highlighted, which will be used in experiment 2.

FGSM: From Fig.1(a), the accuracy of the four models decreases noticeably as the perturbation increases. And the falling trend is similar. Initially, **FB1**, **FB2**, and **FB3** performed close to each other, but as the epsilon increased, **FB2** the accuracy of **FB2** was about 7% higher than **FB1** after epsilon = 5/255. And the accuracy of **FB2** improved by about 20% compared to the baseline model.

PGD L_∞: For the PGD L_∞ attacks, the result demonstrates a similar trend to FGSM attacks in Fig.1(b). Especially, when epsilon equals 5/255, the difference between **FB2** and the baseline model reaches a maximum of +24.71% when epsilon is 4/255. After that, as the perturbation increases, the accuracy of models tends to converge.

PGD L_2: From Table 1, **FB2** has a highest accuracy 68.69%, which improves 24.33% compares to baseline model. And **FB3** performs slightly inferior.

Overview: The trend of baseline and **FB** models are consistent in Fig.1. The models trained with Fourier basis noise improve the robustness against FGSM, PGD L_∞, and PGD L_2 attacks, where **FB2** has the best performance.

5.1 Experiment 2

From experiment 1, we can see that **FB2** performed better than **FB1** and **FB3** in the face of FGSM, PGD L_∞, and PGD L_2 attacks. Therefore, in this section, we only compared **FB2** with other works.

FGSM: Fig.2 shows the accuracy of the four models in the face of different levels of FGSM attacks. When the perturbation is very slight, the **FB2** model performed better than the FGSM adversarial trained model (M_{FGSM}) and PGD adversarial trained model (M_{PGD}). However, when the epsilon grows to greater than 3/255, **FB2** does not perform as well as M_{FGSM}, but it outperforms M_{PGD}. After epsilon becomes larger than 7/255, the accuracy of **FB2** is lower than M_{PGD}.

PGD L_∞: Regarding the PGD L_∞ attacks, Fig.3 presents the result. The overall trend is similar to the FGSM attack. **FB2** reached a high accuracy when the epsilon is small. As the perturbation increases **FB2** is worse than M_{FGSM} and M_{PGD} at epsilon greater than 3/255 and 5/255, respectively.

PGD L_2 Compared with M_{FGSM} and M_{PGD}, **FB2** performed better than M_{PGD} but not well as M_{FGSM}. However, the difference between the accuracy of **FB2** and M_{FGSM} is not remarkable.

Overview: Unlike M_{FGSM} and M_{PGD}, the accuracy decreases slowly in the face of increasing perturbations, while the accuracy of **FB2** decreases dramatically. Therefore, **FB2** outperforms M_{FGSM} only when the perturbations are slight enough and gradually inferior to M_{PGD} as the perturbation increases.

model	acc (clean)	PGD L_2
Baseline	87.52	44.36
FB2	**92.36**	68.69
M_{FGSM}	79.7	**73.88**
M_{PGD}	63.13	57.65

Table 3. Accuracy of four models in the face of clean images and PGD L_2 (step size = 0.05, epsilon = 128/255, iteration = 100) attacks.

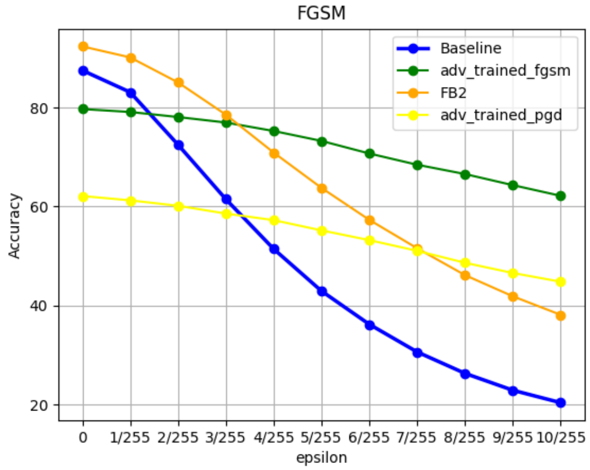


Fig. 2. The accuracy of four models against FGSM attack with different epsilon

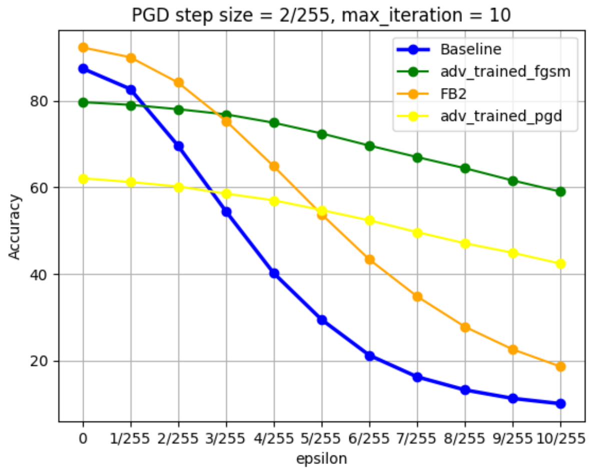


Fig. 3. The accuracy of four models against PGD L_∞ attack with different epsilon

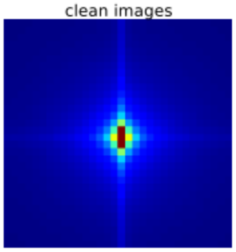


Fig. 4. Fourier spectrum of natural images in CIFAR-10 [30]

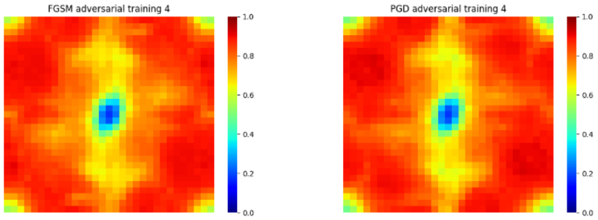


Fig. 5. Fourier heat map from M_{FGSM} and M_{PGD}. The magnitude of Fourier-basis noise is 4. M_{FGSM} and M_{PGD} perform similarly, while M_{FGSM} is slightly better than M_{PGD} in mid-frequency. Both models are robust to low-frequency noise.

6 DISCUSSION

6.1 Explaination of results

The results of Experiment 1 demonstrate that Fourier-based noise augmentation effectively improves model robustness against adversarial attacks. As shown in Fig.5 and Table 1, **FB1** mainly incorporates noise at low and high frequencies. In contrast, **FB2** introduces mid-to-high frequency noise, while **FB3** includes a substantial quantity of high-frequency noise. The following analysis is conducted based on the frequency distribution of the three Fourier-basis noise augmentation policies. The CIFAR-10 clean images have a high concentration at low frequencies (Fig.4), which explains why **FB1** has the highest clean accuracy. According to the previous works mentioned in Section 2.5, We initially hypothesized that adversarial perturbations are more often associated with high frequency. Therefore, training the model with more high-frequency noise such as **FB2** and **FB3** can reach a better performance than **FB1**. However, **FB3**, which has more high-frequency noise does not perform as well as **FB2**. It indicates that the robustness does not increase as expected with the addition of high-frequency noise, as Yin et al. [30] claimed that adversarial examples are not limited to high frequencies.

Compared with other adversarial training methods, the results of Experiment 2 show that the Fourier-basis noise trained model outperforms other methods only in a limited range of adversarial perturbations. It is interesting to observe that the curves of M_{FGSM} and M_{PGD} show similarities, and M_{FGSM} always performs better than M_{PGD}. This result can be partly explained in Fig.5, the heat maps of M_{FGSM} and M_{PGD} show a remarkable resemblance, with M_{FGSM} performing slightly better than M_{PGD}. Both models are only robust to Fourier-basis noise toward low frequency, which means they tend to have a low-frequency noise. Therefore, the performance of the **FB2** with mid-high-frequency noise is weaker. However, based on the results of Experiment 1, it is observed that **FB1**, which adds a certain amount of low-frequency noise, is less effective. The insight provided by Maiya et al. [17], adversarial examples cannot be simply characterized by low-frequency and high-frequency phenomena helps explain the situation.

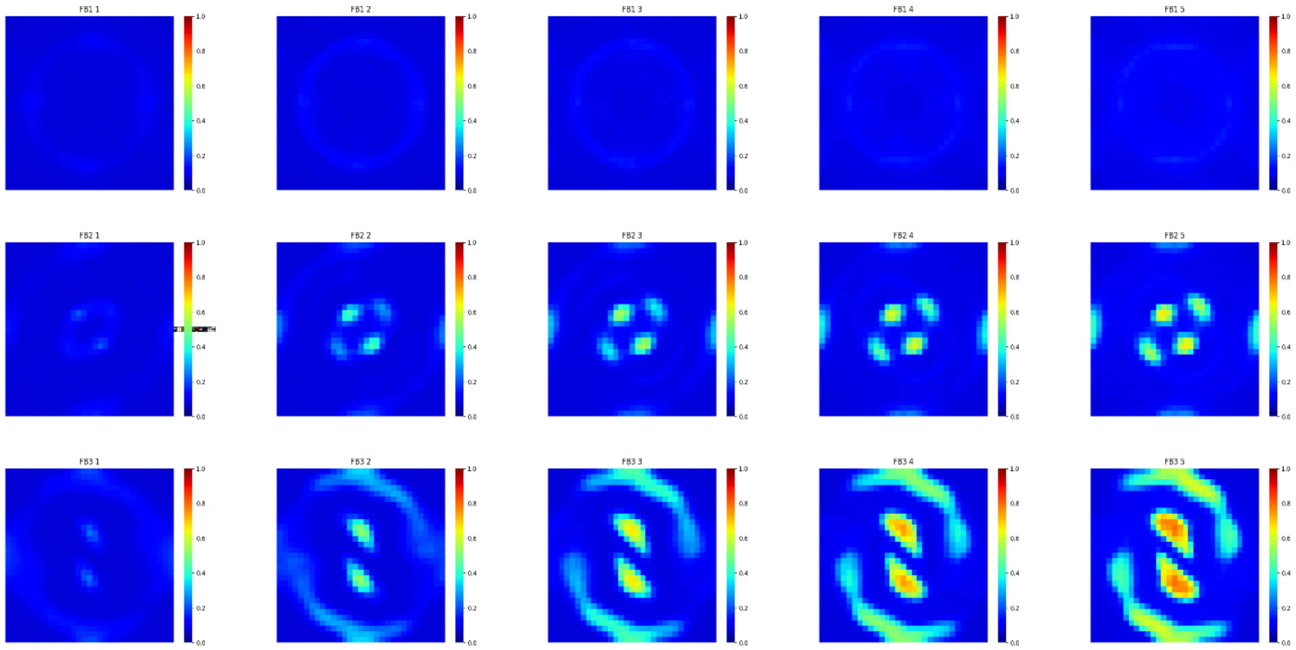


Fig. 6. Fourier heat map. **FB1** is robust to Fourier-basis noise in general, while it is a bit weak against middle frequencies. **FB2** has a poor performance against low frequencies. **FB3** performs much weaker.

7 CONCLUSION

This study focuses on evaluating the effectiveness of Fourier-basis noise in improving adversarial robustness. The result of Experiment 1 indicates that the Fourier-basis noise can help the baseline model to improve its robustness against adversarial attacks. Especially, the policy of prioritizing the mid-to-high frequency range has been identified as the most effective in addressing adversarial examples. Furthermore, we compared the performance with other defense methods specifically for adversarial training in Experiment 2. The result shows that Fourier-basis noise augmentation is superior to other methods only when the perturbation is sufficiently small.

8 FUTURE WORK

This study only evaluates the robustness of the model against FGSM and PGD attacks. There are many other types of adversarial attacks such as deep fool [18], C&W [2], one-pixel [25], etc. that can be assessed in future research. Despite not performing as impressively compared to other methods, we believe that this approach has potential. It may be worth exploring the possibility of combining Fourier-basis noise augmentation with other defense mechanisms such as adversarial training to enhance its effectiveness in improving adversarial robustness in the future.

REFERENCES

[1] Unknown Author. Unknown Year. A Bayesian Data Augmentation Approach for Learning Deep Models. *arXiv preprint arXiv:1710.10564* (Unknown Year). https://arxiv.org/abs/1710.10564

[2] Nicholas Carlini and David Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. *arXiv preprint arXiv:1608.04644* (2017). https://doi.org/10.48550/arXiv.1608.04644

[3] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv preprint arXiv:1406.1078* (2014). https://doi.org/10.48550/arXiv.1406.1078

[4] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. 2019. AutoAugment: Learning Augmentation Strategies From Data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 113–123. https://doi.org/10.1109/CVPR.2019.00020

[5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting Adversarial Attacks with Momentum. *arXiv preprint arXiv:1710.06081* (2018).

[6] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. 2019. Efficient Decision-Based Black-Box Adversarial Attacks on Face Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 7706–7714. https://doi.org/10.1109/CVPR.2019.00790

[7] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2022. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231* (2022).

[8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572* (2015). https://doi.org/10.48550/arXiv.1412.6572

[9] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. 2019. Low Frequency Adversarial Perturbation. *arXiv preprint arXiv:1809.08758* (2019). http://arxiv.org/abs/1809.08758

[10] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2020. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. *arXiv preprint arXiv:1912.02781* (2020). https://doi.org/10.48550/arXiv.1912.02781

[11] Wei Jia, Zhaojun Lu, Haichun Zhang, Zhenglin Liu, Jie Wang, and Gang Qu. 2022. Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems. In *Proceedings of the 2022 Network and Distributed System Security Symposium*. Internet Society, San Diego, CA, USA.

[12] Shiyin Kang, Xiaojun Qian, and Helen Meng. 2013. Multi-distribution deep belief network for speech synthesis. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 8012–8016. https://doi.org/10.1109/ICASSP.2013.6639225

[13] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* (2017).

[14] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. 2019. Fast AutoAugment. *arXiv preprint arXiv:1905.00397* (2019). https://doi.org/10.48550/arXiv.1905.00397

[15] Peter Lorenz, Paula Harder, Dominik Strassel, Margret Keuper, and Janis Keuper. 2021. Detecting AutoAttack Perturbations in the Frequency Domain. *arXiv preprint arXiv:2111.08785* (2021). https://doi.org/10.48550/arXiv.2111.08785

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv preprint arXiv:1706.06083* (2019). https://doi.org/10.48550/arXiv.1706.06083

[17] Shishira R. Maiya, Max Ehrlich, Vatsal Agarwal, Ser-Nam Lim, Tom Goldstein, and Abhinav Shrivastava. 2021. A Frequency Perspective of Adversarial Robustness. *arXiv preprint arXiv:2111.00861* (2021). http://arxiv.org/abs/2111.00861

[18] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: a simple and accurate method to fool deep neural networks. *arXiv preprint arXiv:1511.04599* (2016).

[19] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. 2018. Adversarial Robustness Toolbox v1.2.0. *CoRR* 1807.01069 (2018). https://arxiv.org/pdf/1807.01069

[20] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. *arXiv preprint arXiv:1511.04508* (2016). http://arxiv.org/abs/1511.04508

[21] Andy Phung and Mark Stamp. 2021. *Universal Adversarial Perturbations and Image Spam Classifiers*. Springer, 633–651.

[22] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017). https://doi.org/10.48550/arXiv.1707.06347

[23] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial Training for Free! *arXiv preprint arXiv:1904.12843* (2019).

[24] Ryan Soklaski, Michael Yee, and Theodoros Tsiligkaridis. 2022. Fourier-Based Augmentations for Improved Robustness and Uncertainty Calibration. *arXiv preprint arXiv:2202.12412* (2022).

[25] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation* 23, 5 (October 2019), 828–841. https://doi.org/10.1109/TEVC.2019.2890858

[26] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2014). https://doi.org/10.48550/arXiv.1312.6199

[27] Yusuke Tsuzuku and Issei Sato. 2019. On the Structural Sensitivity of Deep Convolutional Networks to the Directions of Fourier Basis Functions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 51–60. https://doi.org/10.1109/CVPR.2019.00014

[28] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. 2020. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 8681–8691. https://doi.org/10.1109/CVPR42600.2020.00871

[29] Boxi Wu, Heng Pan, Li Shen, Jindong Gu, Shuai Zhao, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. 2021. Attacking Adversarial Attacks as A Defense. *arXiv preprint arXiv:2106.04938* (2021). https://doi.org/10.48550/arXiv.2106.04938

[30] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. 2020. A Fourier Perspective on Model Robustness in Computer Vision. *arXiv preprint arXiv:1906.08988* (2020).

[31] Y. Zeng. 2023. Learning data augmentation policies for computer vision using additive Fourier-basis noise. http://essay.utwente.nl/94515/