FACL-Attack: Frequency-Aware Contrastive Learning for Transferable Adversarial Attacks

Hunmin Yang[1,2,*], Jongoh Jeong[1,*], and Kuk-Jin Yoon[1]
[1]Visual Intelligence Lab., KAIST, [2]Agency for Defense Development
{hmyang, jeong2, kjyoon}@kaist.ac.kr

Abstract

*Deep neural networks are known to be vulnerable to security risks due to the inherent transferable nature of adversarial examples. Despite the success of recent generative model-based attacks demonstrating strong transferability, it still remains a challenge to design an efficient attack strategy in a real-world strict black-box setting, where both the target domain and model architectures are unknown. In this paper, we seek to explore a feature contrastive approach in the frequency domain to generate adversarial examples that are robust in both cross-domain and cross-model settings. With that goal in mind, we propose two modules that are only employed during the training phase: a **F**requency-**A**ware **D**omain **R**andomization (FADR) module to randomize domain-variant low- and high-range frequency components and a **F**requency-**A**ugmented **C**ontrastive **L**earning (FACL) module to effectively separate domain-invariant mid-frequency features of clean and perturbed image. We demonstrate strong transferability of our generated adversarial perturbations through extensive cross-domain and cross-model experiments, while keeping the inference time complexity.*

1. Introduction

Deep neural networks have brought forth tremendous improvements in visual recognition tasks. However, the inherent transferable nature of adversarial examples still exposes the security vulnerability to malicious attackers targeting such susceptible classifiers, causing serious threats and undesirable outcomes in real-world applications. The majority of current attack methods can be primarily classified into two main categories: iterative or optimization-based approaches, and generative model-based approaches. Over the past years, iterative attack methods [8, 21, 20, 4, 17, 5, 37, 18, 25] have been the standard attack protocol for its simplicity and effectiveness. However, this iterative approach is frequently constrained by inefficient time complexity and the potential risk of over-fitting to the training data and models. Moreover, it has shown limited applicability in practical situations due to the low transferability to unknown models and domains.

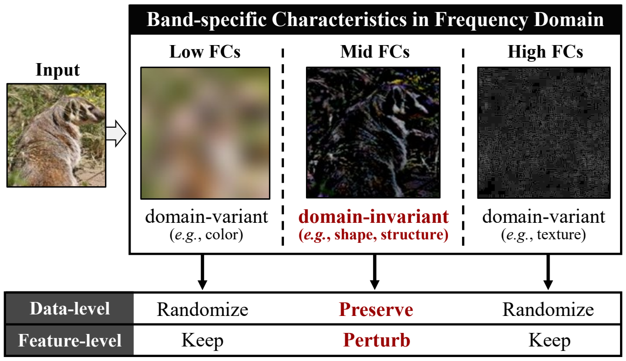

Figure 1: To boost the transferability of adversarial examples, we exploit band-specific characteristics of natural images in the frequency domain. Our method randomizes *domain-variant* low- and high-band frequency components (FCs) in data space, and contrasts *domain-invariant* mid-range clean and perturbed feature pairs in feature space.

In this light, generative attacks [27, 26, 29, 24, 41] have recently gained attention, demonstrating the high transferability across unknown models and domains. Moreover, generator-based attacks yield lower time complexity than iterative or optimization-based methods in the inference stage, which is also a crucial part for real-world attacks. While the current chain of generative attack methods [27, 26, 24, 29, 41, 36] are time-efficient and effective against various black-box settings, we remark that their methods do not actively leverage domain-related characteristics for more transferable attacks.

In that sense, our work is inspired by frequency domain manipulations [40, 34, 35] in domain adaptation (DA) [39] and generalization (DG) [13, 38], demonstrating the superior generalization capabilities of the trained model. As we target transferable attack on unknown target domains and victim models to boost the transferability in a similar set-

[*]Hunmin Yang and Jongoh Jeong contributed equally to this work.

ting, we seek to exploit domain-related characteristics from simple yet effective frequency manipulations.

Several recent studies have focused on frequency-based adversarial attacks to manipulate adversarial examples, aimed at deeper understanding of their dataset dependency [22], adversarial robustness [6], and the security vulnerability [7]. With a slightly different motive, SSAH [19] aims to improve the perceptual quality, whereas [9] designs low-frequency perturbations to enhance the efficiency of black-box queries. Although low-frequency perturbations are efficient, they are known to provide less effective transfer between models [30]. As such, we delve deeper into frequency-driven approaches that effectively enhance the transferability of adversarial examples, especially crafted in a generative framework.

To this end, we propose a novel generative attack method, **FACL-Attack**, to facilitate transferable attacks across various domains and models from the frequency domain perspective. In our training, we introduce frequency-aware domain randomization and feature contrastive learning, explicitly leveraging band-specific characteristics of image attributes such as color, shape, and texture, as shown in Figure 1. We highlight our contributions as follows:

- We propose two modules to boost the adversarial transferability, *FADR* and *FACL*, in which FADR randomizes *domain-variant* data components while FACL contrasts *domain-invariant* feature pairs in the frequency domain.
- We achieve the state-of-the-art attack transferability across various domains and model architectures, demonstrating the effectiveness of our method.
- Our plug-and-play modules can be easily integrated into existing generative attack frameworks, further boosting the transferability while keeping the time complexity.

2. Method

Overview of FACL-Attack. Our method aims to train a robust perturbation generator that yields effective adversarial examples given arbitrary images from black-box domains to induce the unknown victim model to output misclassification. It consists of two key modular operations in the frequency domain, each applied to the input image data and features extracted from the surrogate model only during the training stage, as illustrated in Figure 2.

2.1. Frequency-Aware Domain Randomization

This subsection describes our FADR module designed to boost the robustness of perturbation generator $G_\theta(\cdot)$ against arbitrary domain shifts in practical real-world scenarios. Inspired by recent works [13, 38], we decompose the input training images into multiple-range FCs by leveraging DCT, and apply random masked filtering operation on *domain-specific* image attributes. While FSDR [13] and FACT [38] each employs histogram matching and Fourier-based amplitude mix-up, our FADR module explicitly manipulates the DCT coefficients to diversify input images, aligning with a recent work [16] that narrows the gap between the surrogate model and possible victim models via spectrum transformation. We remark that our approach applies domain randomization exclusively to *domain-specific* FCs, whereas the existing work [16] applies spectral transformation over the whole frequency bands.

In converting the input images into the frequency domain, we apply DCT to each channel separately. We then apply random masked filtering to diversify the input images for boosting the cross-domain transferability. Our spectral transformation operation $\mathcal{T}_{\text{FADR}}(\cdot)$ for source images $\boldsymbol{x}_s$ can be mathematically expressed as follows:

$$\mathcal{T}_{\text{FADR}}(\boldsymbol{x}_s) = \phi^{-1}\Big((\phi(\boldsymbol{x}_s + \boldsymbol{\xi}) \odot \boldsymbol{M}\Big), \tag{1}$$

with the mask $\boldsymbol{M}$ defined as follows:

$$\boldsymbol{M} = \begin{cases} \mathcal{U}(1-\rho, 1+\rho), & \text{if } f < f_l, \\ 1, & \text{if } f_l \leq f < f_h, \\ \mathcal{U}(1-\rho, 1+\rho), & \text{if } f \geq f_h, \end{cases} \tag{2}$$

where $\odot$, ϕ, ϕ^{-1} denote Hadamard product, DCT, and inverse DCT (IDCT) operation, respectively. The random noise $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is sampled from a Gaussian distribution, and the mask values are randomly sampled from Uniform distribution, denoted as $\mathcal{U}$. For the random mask matrix $\boldsymbol{M}$ which has same dimension with the DCT output, we assign its matrix component values as defined in Equation 2, where we set the low and high thresholds as f_l, and f_h, respectively, to distinguish low-, mid-, and high-frequency bands. Note that we can parameterize our FADR module with hyperparameters ρ and σ.

The augmented image output from FADR is then fed as input to the generator $G_\theta(\cdot)$ to yield the adversarial image $\boldsymbol{x}_s' = P(G_\theta(\mathcal{T}_{\text{FADR}}(\boldsymbol{x_s})))$, after the perturbation projection within the pre-defined budget of $\|\delta\|_\infty \leq \epsilon$.

2.2. Frequency-Augmented Contrastive Learning

Recent works on multi-object scene attacks have highlighted the importance of feature-level contrast for transferable generative attacks. In a similar approach to their ideas of exploiting local patch differences [2] or CLIP features [1], our FACL module seeks to apply feature contrast specifically in the *domain-agnostic* mid-frequency range for improving the generalization capability of the trained perturbation generator $G_\theta(\cdot)$.

Spectral decomposition. According to the training pipeline of our FACL-Attack in Figure 2, the generated ad-

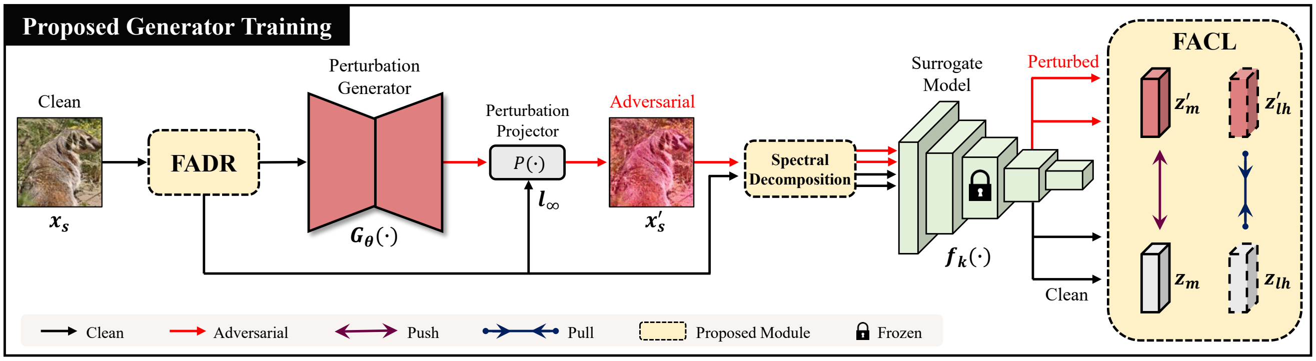

Figure 2: **Overview of FACL-Attack**. From the clean input image, our FADR module outputs the augmented image after spectral transformation, which is targeted to randomize only the domain-variant low/high FCs. The perturbation generator $G_\theta(\cdot)$ then produces the l_∞-budget bounded adversarial image $\boldsymbol{x}'_s$ with perturbation projector $P(\cdot)$ from the randomized image. The resulting clean and adversarial image pairs are decomposed into mid-band (*domain-agnostic*) and low/high-band (*domain-specific*) FCs, whose features $f_k(\cdot)$ extracted from the k-th layer of the surrogate model are contrasted in our FACL module to boost the transferability. The adversarial image $\boldsymbol{x}'_s$ is colorized only for visualization.

versarial image $\boldsymbol{x}'_s$ undergoes spectral decomposition before feature extraction from the surrogate model. This process is carried out by using a band-pass filter $\boldsymbol{M}_{\text{bp}}$ and a band-reject filter $\boldsymbol{M}_{\text{br}}$, to decompose the surrogate model inputs into mid- and low- and high-band FCs, respectively. The spectral decomposition operator is defined as follows:

$$\boldsymbol{M}_{\text{bp}} = \begin{cases} 1, & \text{if } f_l \leq f < f_h, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $\boldsymbol{M}_{\text{br}}$ is the opposite of $\boldsymbol{M}_{\text{bp}}$, holding its values in reverse. Then the spectrally decomposed features from the surrogate model $\boldsymbol{f}$ are defined as:

$$\mathbf{z}_{\text{band}} = f_k\Big(\phi^{-1}\Big(\phi(\boldsymbol{x}_{\text{input}}) \odot \boldsymbol{M}_{\text{band}}\Big)\Big), \tag{4}$$

where $\boldsymbol{M}_{\text{band}}$ is set to either $\boldsymbol{M}_{\text{bp}}$ or $\boldsymbol{M}_{\text{br}}$, and $f_k(\cdot)$ denotes the k-th layer of f. Given $\boldsymbol{x}_s$ and $\boldsymbol{x}'_s$ as input, we finally obtain two pairs of band-specific frequency-augmented features to contrast, *i.e.*, $(\mathbf{z}_m, \mathbf{z}'_m)$ for repelling, and $(\mathbf{z}_{lh}, \mathbf{z}'_{lh})$ for attracting each other.

Loss function. The baseline loss $\mathcal{L}_{\text{orig}}$ for attacking the surrogate model via contrasting clean and adversarial feature pairs is defined as follows:

$$\mathcal{L}_{\text{orig}} = \text{sim}(f_k(\boldsymbol{x}_s), f_k(\boldsymbol{x}'_s)), \tag{5}$$

where sim refers to the standard cosine similarity metric. To boost the attack transferability, our FACL module effectively exploits the spectrally decomposed feature pairs in our proposed FACL loss function defined as follows:

$$\mathcal{L}_{\text{FACL}} = \text{sim}(\mathbf{z}_m, \mathbf{z}'_m) - \text{sim}(\mathbf{z}_{lh}, \mathbf{z}'_{lh}), \tag{6}$$

where the goal of $\mathcal{L}_{\text{FACL}}$ is to reinforce the effectiveness of *domain-agnostic* mid-band feature contrast $(\mathbf{z}_m, \mathbf{z}'_m)$, while minimizing the importance of *domain-specific* low- and high-band feature difference $(\mathbf{z}_{lh}, \mathbf{z}'_{lh})$. In this approach, our $\mathcal{L}_{\text{FACL}}$ facilitates the push-pull action among the band-specific feature pairs, further guiding the perturbation generation towards more robust regime.

Final learning objective. We train our perturbation generator by minimizing the total loss function as follows:

$$\min_\theta \; (\lambda_{\text{orig}} \cdot \mathcal{L}_{\text{orig}} + \lambda_{\text{FACL}} \cdot \mathcal{L}_{\text{FACL}}), \tag{7}$$

where λ_{orig} and λ_{FACL} are loss coefficients. The objective guides our generator $G_\theta(\cdot)$ to generate more robust perturbations to domain shifts as well as model variances.

3. Experiments

Cross-domain transferability. We compare our method with the state-of-the-art generative attacks on various domains such as CUB-200-2011 [33], Stanford Cars [15], and FGVC Aircraft [23]. Our perturbation generator is trained by only leveraging ImageNet-1K [28] and pretrained surrogate model (*i.e.*, VGG-16 [31]). Then, it is evaluated on black-box victim models trained with DCL framework [3] using different backbones (*i.e.*, Res-50 [10], SENet154 [11], SE-Res101 [11]). We closely follow recent works [26, 29, 41, 1] for implementation of perturbation generator for fair comparison. We train with an Adam [14] optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$) with a learning rate of 2×10^{-4}, and a batch size of 16. For FADR hyperparameters, we follow the insights from [13] to set the low and high frequency thresholds to 7 and 112, respectively, with $\rho = 0.01$ and $\sigma = 8$ for spectral transformation.

As shown in Table 1, our FACL-Attack outperforms on most *cross-domain* benchmarks, among which are also

Method	CUB-200-2011			Stanford Cars			FGVC Aircraft			AVG.
	Res-50	SENet154	SE-Res101	Res-50	SENet154	SE-Res101	Res-50	SENet154	SE-Res101	
Clean	87.35	86.81	86.56	94.35	93.36	92.97	92.23	92.08	91.90	90.85
GAP [27]	68.85	74.11	72.73	85.64	84.34	87.84	81.40	81.88	76.90	79.30
CDA [26]	69.69	62.51	71.00	75.94	72.45	84.64	71.53	58.33	63.39	69.94
LTP [29]	30.86	52.50	62.86	34.54	**65.53**	73.88	**15.90**	60.37	52.75	49.91
BIA [41]	32.74	52.99	58.04	39.61	69.90	70.17	28.92	60.31	**46.92**	51.07
FACL-Attack (Ours)	**24.74**	**44.06**	**53.75**	**26.58**	65.71	**61.40**	19.72	**52.01**	48.51	**44.05**

Table 1: **Cross-domain evaluation results.** The perturbation generator is trained on ImageNet-1K with VGG-16 as the surrogate model and evaluated on black-box domains with black-box models. We compare the top-1 classification accuracy after attacks with the perturbation budget of $l_\infty \leq 10$ (the lower, the better). **Best** and second best.

Method	Venue	VGG-16	VGG-19	Res-50	Res-152	Dense-121	Dense-169	Inc-v3	AVG.
Clean	-	70.14	70.95	74.61	77.34	74.22	75.75	76.19	74.17
GAP [27]	CVPR'18	23.63	28.56	57.87	65.50	57.94	61.37	63.30	55.76
CDA [26]	NeurIPS'19	**0.40**	**0.77**	36.27	51.05	38.89	42.67	54.02	32.01
LTP [29]	NeurIPS'21	1.61	2.74	21.70	39.88	23.42	**25.46**	41.27	22.30
BIA [41]	ICLR'22	1.55	3.61	25.36	42.98	26.97	32.35	41.20	24.86
FACL-Attack (Ours)	-	1.45	2.92	**19.72**	**36.61**	**21.34**	25.61	**29.97**	**19.66**

Table 2: **Cross-model evaluation results**. The perturbation generator is trained on ImageNet-1K with VGG-16 as the surrogate model and evaluated on black-box models including white-box model (*i.e.*, VGG-16). We compare the top-1 classification accuracy after attacks with the perturbation budget of $l_\infty \leq 10$ (the lower, the better). **Best** and second best.

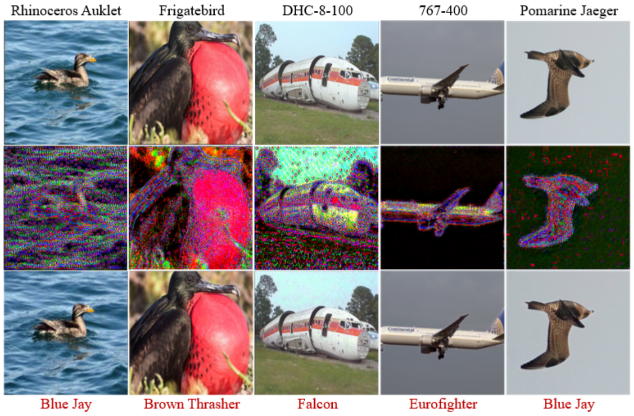

Figure 3: **Qualitative results**. Clean images (*row 1*), unbounded adversarial images (*row 2*), and bounded adversarial images (*row 3*; actual inputs to the classifier) are shown for various domains. The ground truth and each mispredicted class label are shown on the *top* and *bottom*.

cross-model, by significant margins. We posit that the remarkable generalization ability of FACL-Attack owes to the synergy between our two proposed modules that effectively guide feature-level separation in the *domain-agnostic* mid-frequency band (*i.e.*, FACL), complemented by data-level randomization only applied to the *domain-specific* frequency components (*i.e.*, FADR). As shown in Figure 3, FACL-Attack can craft effective and high-quality adversarial images with imperceptible perturbations.

Cross-model transferability. We further investigated on the black-box model scenario in a controlled white-box domain (*i.e.*, ImageNet-1K). The generator is trained against a surrogate model (*i.e.*, VGG-16) and evaluated on various victim models which include VGG-16 (white-box), VGG-19, Res-50, Res-152 [10], Dense-121, Dense-169 [12], and Inc-v3 [32]. As shown in Table 2, ours also outperforms on most benchmarks where they seem to partially overfit to the white-box model (*i.e.*, VGG-16). We posit that the frequency-augmented feature learning could help the perturbation generator craft more robust perturbations, which exhibit better generalization capability to unknown feature space. This aligns with a recent finding [16] that spectral data randomization contributes to enhance the transferability via simulating diverse victim models.

4. Conclusion

In this paper, we proposed a novel generator-based transferable attack framework called FACL-Attack, leveraging spectral transformation and feature contrast in the frequency domain. Our method targets spectral randomization on *domain-specific* image components, and *domain-agnostic* feature contrast for training a more robust perturbation generator. Our extensive evaluation results validate the effectiveness in practical black-box scenarios with domain shifts and model variances. It can also be easily integrated into existing attack frameworks, further boosting the transferability while keeping the inference time complexity.

References

[1] Abhishek Aich, Calvin Khang-Ta, Akash Gupta, Chengyu Song, Srikanth V Krishnamurthy, M Salman Asif, and Amit K Roy-Chowdhury. Gama: Generative adversarial multi-object scene attacks. *arXiv preprint arXiv:2209.09502*, 2022. 2, 3

[2] Abhishek Aich, Shasha Li, Chengyu Song, M Salman Asif, Srikanth V Krishnamurthy, and Amit K Roy-Chowdhury. Leveraging local patch differences in multi-object scenes for generative adversarial attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1308–1318, 2023. 2

[3] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *CVPR*, 2019. 3

[4] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 1

[5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 1

[6] Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, A Kai Qin, and Yuan He. Advdrop: Adversarial attack to dnns by dropping information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7506–7515, 2021. 2

[7] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. 2

[8] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1

[9] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pages 1127–1137. AUAI Press, 2019. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 4

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3

[12] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 4

[13] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6891–6902, 2021. 1, 2, 3

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3

[15] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013. 3

[16] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 549–566. Springer, 2022. 2, 4

[17] Peter Lorenz, Paula Harder, Dominik Straßel, Margret Keuper, and Janis Keuper. Detecting autoattack perturbations in the frequency domain. *arXiv preprint arXiv:2111.08785*, 2021. 1

[18] Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *CVPR*, 2020. 1

[19] Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15315–15324, 2022. 2

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1

[22] Shishira R Maiya, Max Ehrlich, Vatsal Agarwal, Ser-Nam Lim, Tom Goldstein, and Abhinav Shrivastava. A frequency perspective of adversarial robustness. *arXiv preprint arXiv:2111.00861*, 2021. 2

[23] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. volume abs/1306.5151, 2013. 3

[24] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. On generating transferable targeted perturbations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2021. 1

[25] Muzammal Naseer, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *CVPR*, 2020. 1

[26] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3, 4

[27] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4422–4431, 2018. 1, 4

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 3

[29] Mathieu Salzmann et al. Learning transferable adversarial perturbations. *Advances in Neural Information Processing Systems*, 34:13950–13962, 2021. 1, 3, 4

[30] Yash Sharma, Gavin Weiguang Ding, and Marcus A. Brubaker. On the Effectiveness of Low Frequency Perturbations. 2019. 2

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 3

[32] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 4

[33] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011. 3

[34] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8684–8694, 2020. 1

[35] Zifan Wang, Yilin Yang, Ankit Shrivastava, Varun Rawal, and Zihao Ding. Towards frequency-based explanation for robust cnn. *arXiv preprint arXiv:2005.03141*, 2020. 1

[36] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1161–1170, 2020. 1

[37] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019. 1

[38] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 1, 2

[39] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 1

[40] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[41] Qilong Zhang, Xiaodan Li, Yuefeng Chen, Jingkuan Song, Lianli Gao, Yuan He, and Hui Xue. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. *arXiv preprint arXiv:2201.11528*, 2022. 1, 3, 4