CSCI 1430 Final Project Report: Your project title

Pixel Pioneers: Yuyang Luo, Zhihao Li, Jiachen Wang.
TA name: Hannah. Brown University

Abstract

Deep neural networks (DNNs) have achieved remarkable performance in various computer vision tasks, such as image classification, object detection, face recognition, and autonomous driving. However, recent studies have shown that DNNs are vulnerable to adversarial examples, which can cause misclassification through human-imperceptible perturbations. These adversarial examples exhibit a property called transferability, allowing attacks on one model to mislead other models, posing significant security risks. Although several attack methods have demonstrated impressive performance in white-box settings, their effectiveness diminishes in black-box scenarios, particularly against models with advanced defenses.

To address the gap between white-box and transfer-based black-box attacks, we reproduced several representative methods, including VMI-FGSM, FIA, RPA, and SSA, within a unified framework to facilitate direct comparison under consistent experimental conditions. Motivated by SSA's approach of random scaling and noise addition in the frequency domain, we explored perturbing images' amplitude and phase components to enhance adversarial transferability. By transforming images to the frequency domain and adding Gaussian noise to both amplitude and phase to get more diverse inputs, our method generates more transferable adversarial examples. Experimental results show that our approach significantly outperforms baseline methods, demonstrating its superiority and effectiveness.

1. Introduction

Deep neural networks (DNNs) have achieved great performance in various computer vision tasks, such as image classification [29, 15, 30], objective detection [11, 27, 14], face recognition [32, 5, 1], and autonomous driving [50, 45, 46]. However, recent works have shown that DNNs are vulnerable to adversarial examples [31, 12], in which applying human-imperceptible perturbations on clean input can result in misclassification. Furthermore, adversarial examples have an intriguing property of transferability [6, 21], *i.e.*, the adversarial example generated on the surrogate model can also mislead other victim models. The existence of transferability makes adversarial attacks practical to real-world applications because hackers do not need to know any information about the target model, which brings serious security threats to security-sensitive applications.

Though several existing attack methods [18, 25] have exhibited impressive attack performance in the white-box setting, their effectiveness notably diminishes in black-box scenarios, particularly against models equipped with advanced defenses [19, 33]. Recently, numerous methods have been proposed to improve adversarial transferability, such as introducing momentum in gradient iterations [6, 21, 38, 40, 10], adopting various input transformations [48, 39, 51, 42, 37], and integrating multiple models for attacks [22, 2]. However, there is still a distinct gap between the performance of white-box attacks and transfer-based black-box attacks.

Numerous efforts have been made to bridge the gap between white-box attacks and transfer-based black-box attacks. However, the diversity in frameworks, datasets, and experimental settings has complicated the comparison of these methods. Motivated by this challenge, we reproduced several representative methods, including VMI-FGSM, FIA, RPA and SSA, within a unified framework. By testing these methods under consistent experimental conditions, we facilitated a clear and direct comparison of their effectiveness and performance.

SSA randomly scales images and adds noise in the frequency domain, inspiring us to explore more processes in this domain. Given that images are composed of amplitude and phase components, we designed a method to perturb both components for data augmentation, thereby generating more transferable adversarial examples. Specifically, we transform images to the frequency domain and add Gaussian noise to both amplitude and phase, enhancing the perturbations. Experimental results demonstrate that our method significantly outperforms baseline methods, showing a clear margin of improvement and highlighting the superiority and effectiveness of our approach.

2. Related Work

In general, adversarial attacks can be divided into two categories, *i.e.*, white-box attacks and black-box attacks. In the white-box setting, the attacker has all the information about the architecture and parameters of the target model [12, 18]. By contrast, black-box attacks are more practical since they only access limited or no information about the target model. There are two types of black-box adversarial attacks [9, 36, 3]: query-based and transfer-based attacks. Query-based attacks [17, 28] often take hundreds or even thousands of queries to generate adversarial examples, making them inefficient. On the other hand, transfer-based attacks [7, 21] generate adversaries on the surrogate model without accessing the target model, leading to superior practical applicability and attracting increasing attention.

Though existing methods (*e.g.* I-FGSM) have achieved great effectiveness in the white-box setting, they exhibit low transferability when attacking black-box models. To improve adversarial transferability, many works have been proposed from different perspectives. Gradient-based attacks use better optimization methods to make adversarial examples more transferable. For instance, MIM [6] integrates momentum into I-FGSM to stabilize the update direction and escape from poor local maxima at each iteration. VMI [38] enhances the momentum by accumulating the gradient of several data points in the direction of the previous gradient for better transferability. Inspired by the data augmentation strategies [53, 35, 52], various input transformation-based attacks have been proposed to effectively boost adversarial transferability. [48] adopt diverse input patterns by randomly resizing and padding to generate transferable adversarial examples. [39] mix up a set of images randomly sampled from other categories while maintaining the original label of the input to craft more transferable adversaries. [24] propose a novel spectrum simulation attack by transforming the input image in the frequency domain. Besides, architecture-related works try to modify the architecture of the source model to improve transferability. SGM [44] adjusts the decay factor to increase gradient backpropagation from skip connections in ResNet. LinBP [13] alternates the gradient of ReLU to 1 and rescales the gradients in each block. BPA [41] recovers the gradient truncated by non-linear layers using non-zero function.

Several works have been proposed to enhance adversarial transferability by perturbing the intermediate features. TAP [56] maximizes the distance among the intermediate features and smooths the adversarial perturbations with a regularizer. ILA [16] fine-tunes an adversarial example crafted from another method (*e.g.*, MIM) by increasing the feature difference similarity between the original/current adversarial example and the benign example at a specific layer. FIA [43] disrupts object-aware features by minimizing a weighted feature spectrum in the intermediate layer. The weight is determined by computing the average gradient with respect to the feature across several randomly pixel-wise masked input images. RPA [55] computes the average gradient of randomly patch-wise masked images with different patch sizes, which serves as the weight in FIA to highlight important intrinsic object-related features effectively. NAA [54] uses neuron attribution for accurate neuron importance estimation, which develops an approximation scheme to reduce computation time and generates adversarial samples by minimizing a weighted combination of positive and negative neuron attribution values.

2.1. Adversarial Defenses

To mitigate the threat of adversarial examples, many methods have been proposed recently. A notable and efficacious defense mechanism is adversarial training [12, 25], which injects adversarial examples into training data to improve the network robustness. In particular, ensemble adversarial training [34] fortifies model robustness by training with adversarial examples derived from other pre-trained models, showing effectiveness against transfer-based attacks. Despite its efficacy, adversarial training is often hampered by significant computational demands and scalability issues, especially with expansive datasets and intricate neural networks. Additionally, several defense methods purify the adversarial examples before feeding into the model. For example, JPEG [8] indicates that adversarial perturbations can be partly removed via JPEG compression. HGD [20] trains a high-level representation guided denoiser to suppress the influence of adversarial perturbation. NIPS-r3 sends the transformed (*e.g.*, rotation, sheer, shift) input images to an ensemble of adversarially trained models to get the final output. R&P [47] utilizes random resizing and padding to mitigate adversarial effects. Bit-Red [49] reduces image color depth and employs smoothing to decrease pixel variations. FD [23] introduces a JPEG-based defensive compression framework to rectify adversarial examples while preserving classification accuracy on benign data. NRP [26] trains a Purifier Network in a self-supervised manner to purify the input. RS [4] utilizes randomized smoothing to train a certifiably ℓ_2 robust classifier.

3. Method

3.1. VNI-FGSM

Traditional MI-FGSM (and NI-FGSM) techniques have sought to enhance attack robustness and transferability by stabilizing update directions, building upon the foundation of I-FGSM. However, these methods solely consider points along the optimization path throughout all iterations. VNI-FGSM represents a notable advancement over MI-FGSM (and NI-FGSM) by incorporating the neighbors of a point x in perturbation calculation. It defines the variance at point x

as (Eq. 1)

$$V(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla_{x^i} J(x^i, y; \theta) - \nabla_x J(x, y; \theta), \quad (1)$$

where N denotes the number of examples sampled within x's neighborhood, and x^i represents one such example sampled from a Uniform distribution controlled by hyperparameters β and ϵ.

VNI-FGSM leverages the variance v_t at iteration t to adjust the gradient x^{adv} at iteration $t+1$, introducing non-deterministic noise in each iteration. This strategic integration of variance facilitates superior robustness and transferability in adversarial attacks, paving the way for enhanced security in machine learning systems.

3.2. FIA

The Feature Importance-aware Attack (FIA) method innovates by leveraging model sensitivity to specific image regions. It focuses perturbations on important areas to maximize impact while preserving the image's overall naturalness. FIA uses aggregated gradients from transformed image versions to identify key features, employs random pixel dropout to maintain object integrity, and guides adversarial sample creation to disrupt critical features for enhanced transferability. The aggregated gradient can be represented as follows, with the probability of random pixel dropout denoted as p_d.

$$\hat{x}_k = \frac{1}{C}\sum_{n=1}^{N} x^{\circ M_{p_d}^n}, \quad M_{p_d} \sim \text{Bernoulli}(1-p_d), \quad (2)$$

where the M_{p_d} is a binary matrix with the same size as x, and $\circ$ denotes the element-wise product. The normalizer C is obtained by ℓ_2-norm on the corresponding summation term. The ensemble number N indicates the number of random masks applied to the input x.

3.3. RPA

Random Patch Attack (RPA) effectively captures the intrinsic key features of objects through random patch transformations, significantly enhancing the transferability of adversarial examples. We introduced random patch transformations into benign images, where important object-related feature areas are highlighted by computing and aggregating the feature maps' gradients at intermediate layers. In contrast, model-specific features are suppressed. The aggregated gradients guide the adversarial perturbations, distorting important features and thus increasing the transferability of adversarial samples. The model attention areas on adversarial examples generated by FIA overlap with those on clean images, whereas RPA-produced adversarial examples significantly disperse the model's attention, causing the model to shift its focus from areas on clean images to completely different areas.

3.4. SSA

Traditional adversarial methods often struggle with a large transferability gap when the attack examples crafted using one model (substitute model) fail to deceive another model (victim model). To overcome this, SSA proposes enhancing the robustness and transferability of adversarial examples by manipulating them in the frequency domain rather than the traditional spatial domain. The focus on the frequency domain stems from the hypothesis that different models may focus on different frequency components of the input data, and by simulating this variability, the adversarial examples can be made more universally effective.

To be specific, SSA proposes a random spectrum transformation $T(\cdot)$ which decomposes matrix multiplication into matrix addition and Hadamard product to get diverse spectrums. Specifically, in combination with the DCT/IDCT, the transformation can be expressed as:

$$\mathcal{T}(\boldsymbol{x}) = \mathcal{D}_{\mathcal{I}}((\mathcal{D}(\boldsymbol{x}) + \mathcal{D}(\boldsymbol{\xi})) \odot \boldsymbol{M}) \quad (3)$$

$$= \mathcal{D}_{\mathcal{I}}(\mathcal{D}(\boldsymbol{x} + \boldsymbol{\xi}) \odot \boldsymbol{M}) \quad (4)$$

where $\odot$ denotes Hadamard product, $\xi \sim \mathcal{N}(0, \sigma^2 I)$ and each element of $M \sim \mathcal{U}(1-\rho, 1+\rho)$ are random variants sampled from Gaussian distribution and Uniform distribution, respectively.

3.5. Our Method

Inspired by the SSA perturbing frequency domain, we propose a method to perturb the amplitude and phase components of an image by transforming it into the frequency domain using the Discrete Cosine Transform (DCT). First, we apply DCT to convert the image from the spatial domain to the frequency domain. Next, we decompose the transformed image into its amplitude and phase components. We then add Gaussian noise to the amplitude component to create a perturbed amplitude. Finally, we recombine the perturbed amplitude with the original phase and apply the inverse DCT (IDCT) to transform the image back to the spatial domain. This process can be mathematically formulated as follows: transform the image using DCT, $(I_f = \text{DCT}(I))$; extract amplitude and phase, $(A = |I_f|)$ and $(\Phi = \angle I_f)$; perturb the amplitude and the phase, $(A' = A + \mu * N(0, \sigma^2))$ and $\Phi' = \Phi + N(0, \sigma^2)$; and recombine and apply IDCT, $(I'_f = A' \cdot e^{i\Phi'})$ and $(I' = \text{IDCT}(I'_f))$. This technique allows for effective perturbation of the image in the frequency domain for various image processing and machine learning applications.

4. Results

4.1. Evaluation

To further evaluate the transferability of the adversarial examples generated by each method, we conducted tests across

Attack	Res-18	Res-101	ResNeXt	Dense-101
MIM	100.0*	42.5	46.5	75.1
FIA	99.2	30.1	35.8	65.8
RPA	100*	65.3	69.1	92.3
VNI-FGSM	100.0*	62.3	64.8	89.0
SSA	100.0*	71.2	73.3	94.0
Ours	100.0*	74.0	78.1	95.1

Table 1. Attack success rates (%) on four CNN models. The adversarial examples are crafted on Res-18. * indicates the white-box model.

Attack	ViT	PiT	Visformer	Swin
MIM	17.2	23.8	33.2	40.3
FIA	10.7	16	23.7	35.4
RPA	24.3	33.6	50.9	54.6
VNI-FGSM	27.0	35.9	52.4	55.5
SSA	29.9	39.3	55.1	61.9
Ours	35.4	45.2	61.9	66.9

Table 2. Attack success rates (%) on four advanced ViT models. The adversarial examples are crafted on Res-18.

a range of models. We assessed four conventionally trained CNNs, specifically ResNet-18, ResNet-101, ResNeXt-50, and DenseNet-101. Additionally, we tested the examples on four Vision Transformer (ViT) models: ViT, PiT, Visformer, and Swin. We also evaluated their effectiveness against four defensive methods, namely Adversarial Training (AT), High-gradient Denoising (HGD), Random Smoothing (RS), and Neural Representation Purifier (NRP).

The outcomes for both CNNs and ViTs models are presented in Tab. 1 and Tab. 2. The findings indicate that under the white-box setting, all baseline methods achieve a 100% success rate in attacks. However, in the context of black-box attacks, the FIA exhibits suboptimal performance. Among the baseline methods evaluated, SSA stands out, achieving the highest attack success rates on seven targeted models. Our proposed method delivers superior results on CNN-based target models and significantly outperforms SSA on ViT-based models, demonstrating a clear margin of superiority. These results underscore the effectiveness and enhanced performance of our proposed method.

To further validate the effectiveness of our proposed methods, we evaluated the adversarial examples generated on ResNet-18 against four robust defense mechanisms. The results, presented in Tab. 3, show a noticeable decline in attack performance, confirming the efficacy of these defense strategies. Among the baseline methods, SSA records the best performance against all four defenses. However, our method consistently surpasses SSA across all defenses, demonstrating superior versatility and effectiveness in counteracting a range of advanced and sophisticated defense mechanisms.

Attack	AT	HGD	RS	NRP
MIM	33.1	32.0	22.4	26.5
FIA	31.4	18.9	21.1	19.8
RPA	35.8	56.3	26.7	39.1
VNI-FGSM	34.6	50.2	25.0	38.2
SSA	37.2	62.1	29.2	50.9
Ours	39.2	64.1	29.4	53.9

Table 3. Attack success rates (%) on four advanced defense methods. The adversarial examples are generated on the Res-18 model. The best results are in bold.

4.2. Technical Discussion

Given our goal to enhance the transferability of adversarial examples, it is crucial to evaluate the changes based on their effectiveness in fooling multiple models. This means assessing the success rate of the perturbed images not only on the model they were generated on but also across different architectures and datasets. By introducing variations in the amplitude and phase components in the frequency domain, we aim to create adversarial examples that are universally challenging for diverse models.

Key considerations include maintaining a balance between the level of perturbation and the preservation of essential image features. While effective perturbations are necessary for improving transferability, it is also important to retain enough of the original image structure to ensure the examples remain realistic and semantically meaningful. This balance is critical for practical applications in adversarial training and robustness evaluation. In summary, the changes should focus on maximizing transferability while maintaining image quality.

5. Conclusion

In conclusion, our project successfully addressed the challenge of comparing adversarial attack methods by standardizing the experimental framework and conditions. By reproducing and evaluating key techniques such as VMI-FGSM, FIA, RPA and SSA within a consistent setting, we provided a clearer understanding of their relative strengths and weaknesses. This approach not only enhances the reliability of comparative analyses in the field of adversarial machine learning but also sets a precedent for future research to follow a more standardized methodology for assessing and developing adversarial techniques. Our findings contribute significantly to the ongoing discourse on improving the robustness and efficacy of adversarial attack strategies.

References

[1] Xiang An, Jiangkang Deng, Jia Guo, Ziyong Feng, Xuhan Zhu, Yang Jing, and Liu Tongliang. Killing two birds with one stone: Efficient and robust training of face recognition cnns

by partial fc. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[2] Huanran Chen, Yichi Zhang, Yinpeng Dong, and Jun Zhu. Rethinking Model Ensemble in Transfer-based Adversarial Attacks. *arXiv preprint arXiv:2303.09105*, 2023. 1

[3] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2020. 2

[4] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. In *Proceedings of the International Conference on Machine Learning*, 2019. 2

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1

[6] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting Adversarial Attacks With Momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018. 1, 2

[7] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[8] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of JPG compression on adversarial images. *CoRR*, abs/1608.00853, 2016. 2

[9] Lianli Gao, Qilong Zhang, Xiaosu Zhu, Jingkuan Song, and Heng Tao Shen. Staircase Sign Method for Boosting Adversarial Attacks. *arXiv preprint arXiv:2104.09722*, 2021. 2

[10] Zhijin Ge, Xiaosen Wang, Fanhua Shang, Hongying Liu, and Yuanyuan Liu. Boosting Adversarial Transferability by Achieving Flat Local Maxima. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023. 1

[11] Ross B. Girshick. Fast R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015. 1

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *Proceedings of the International Conference on Learning Representations*, 2015. 1, 2

[13] Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating Linearly Improves Transferability of Adversarial Examples. In *Proceedings of the Advances in Neural Information Processing Systems*, 2020. 2

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[16] Qian Huang, Isay Katsman, Zeqi Gu, Horace He, Serge J. Belongie, and Ser-Nam Lim. Enhancing Adversarial Example Transferability With an Intermediate Level Attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4732–4741, 2019. 2

[17] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box Adversarial Attacks with Limited Queries and Information. In *Proceedings of the International Conference on Machine Learning*, 2018. 2

[18] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Proceedings of the International Conference on Learning Representations*, 2017. 1, 2

[19] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial Machine Learning at Scale. In *Proceedings of the International Conference on Learning Representations*, 2017. 1

[20] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018. 2

[21] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In *ICLR*, 2020. 1, 2

[22] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of the International Conference on Learning Representations*, 2017. 1

[23] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature Distillation: DNN-Oriented JPEG Compression Against Adversarial Examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[24] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency Domain Model Augmentation for Adversarial Attack. In *Proceedings of the European Conference on Computer Vision*, pages 549–566, 2022. 2

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proceedings of the International Conference on Learning Representations*, 2018. 1, 2

[26] Muzammal Naseer, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A Self-supervised Approach for Adversarial Robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 259–268, 2020. 2

[27] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 91–99, 2015. 1

[28] Yucheng Shi, Siyu Wang, and nYahong Han. Curls & Whey: Boosting Black-Box Adversarial Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6519–6527, 2019. 2

[29] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the International Conference on Learning Representations*, 2015. 1

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 1

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, 2014. 1

[32] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014. 1

[33] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble Adversarial Training: Attacks and Defenses. In *Proceedings of the International Conference on Learning Representations*, 2018. 1

[34] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble Adversarial Training: Attacks and Defenses. In *Proceedings of the International Conference on Learning Representations*, 2018. 2

[35] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold Mixup: Better Representations by Interpolating Hidden States. In *Proceedings of the International Conference on Machine Learning*, 2019. 2

[36] Jiafeng Wang, Zhaoyu Chen, Kaixun Jiang, Dingkang Yang, Lingyi Hong, Yan Wang, and Wenqiang Zhang. Boosting the Transferability of Adversarial Attacks with Global Momentum Initialization. *arXiv preprint arXiv:2211.11236*, 2022. 2

[37] Kunyu Wang, Xuanran He, Wenxuan Wang, and Xiaosen Wang. Boosting Adversarial Transferability by Block Shuffle and Rotation. *arXiv preprint arXiv:2308.10299*, 2023. 1

[38] Xiaosen Wang and Kun He. Enhancing the Transferability of Adversarial Attacks Through Variance Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021. 1, 2

[39] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the Transferability of Adversarial Attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16138–16147, 2021. 1, 2

[40] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting Adversarial Transferability through Enhanced Momentum. In *Proceedings of the British Machine Vision Conference*, page 272, 2021. 1

[41] Xiaosen Wang, Kangheng Tong, and Kun He. Rethinking the Backward Propagation for Adversarial Transferability. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023. 2

[42] Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure Invariant Transformation for better Adversarial Transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4607–4619, 2023. 1

[43] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature Importance-aware Transferable Adversarial Attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[44] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets. In *Proceedings of the International Conference on Learning Representations*, 2020. 2

[45] Hai Wu, Chenglu Wen, Wei Li, Xin Li, Ruigang Yang, and Cheng Wang. Transformation-Equivariant 3D Object Detection for Autonomous Driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2795–2802, 2023. 1

[46] Penghao Wu, Li Chen, Hongyang Li, Xiaosong Jia, Junchi Yan, and Yu Qiao. Policy Pre-training for Autonomous Driving via Self-supervised Geometric Modeling. In *Proceedings of the International Conference on Learning Representations*, 2023. 1

[47] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating Adversarial Effects Through Randomization. In *Proceedings of the International Conference on Learning Representations*, 2018. 2

[48] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[49] Weilin Xu, David Evans, and Yanjun Qi. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Network and Distributed System Security Symposium*, 2018. 2

[50] Longhui Yu, Yifan Zhang, Lanqing Hong, Fei Chen, and Zhenguo Li. Dual-Curriculum Teacher for Domain-Inconsistent Object Detection in Autonomous Driving. In *Proceedings of the British Machine Vision Conference*, 2022. 1

[51] Zheng Yuan, Jie Zhang, and Shiguang Shan. Adaptive Image Transformations for Transfer-Based Adversarial Attack. In *Proceedings of the European Conference on Computer Vision*, 2022. 1

[52] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2

[53] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *Proceedings of the International Conference on Learning Representations*, 2018. 2

[54] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R. Lyu. Improving Adversarial Transferability via Neuron Attribution-based Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[55] Yaoyuan Zhang, Yu-an Tan, Tian Chen, Xinrui Liu, Quanxin Zhang, and Yuanzhang Li. Enhancing the Transferability of Adversarial Examples with Random Patch. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2022. 2

[56] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable Adversarial Perturbations. In *Proceedings of the European Conference on Computer Vision*, pages 471–486, 2018. 2

Social Impact

In the world of computer vision, technology is advancing rapidly. We have already realized many useful functions such as autonomous driving and facial recognition. However, as technology progresses, potential threats and impacts in the security domain are also increasing, along with a variety of social ethical issues. For instance, in the realm of autonomous driving, if adversarial attack technology causes an accident, accountability is extremely difficult to establish because responsibility is still not clearly defined.

The advancement of adversarial attack techniques, especially methods like VNI-FGSM, FIA, RPA, and SSA mentioned in our research, has enhanced the transferability of attacks, allowing attackers to launch attacks without precise knowledge of the target model. Traditional network security mechanisms are unable to cope with such adjustments. This enhancement of capabilities could lead to misuse of technology and endanger data security, particularly in areas of critical infrastructure and personal data protection.

Moreover, such adversarial attacks can effectively deceive image recognition systems without significantly altering visual effects. To the human eye, they are indistinguishable from the original images, because the model often opts to modify the high-frequency components of the image during an adversarial attack. This makes it more covert and easier to execute undetected.

In our research, we implemented and evaluated various types of Untargeted Attack methods, including Gradient-based, Input transformation-based, and Advanced objective. We assessed their effects in all aspects, especially in terms of enhancing transferability, and through experiments, we demonstrated the significant impact existing technologies can have on image classification tasks.

Our research was conducted within a standard experimental framework, evaluating different attack methods to provide a standardized way of comparing different adversarial attack techniques. The models used for testing include CNN types, ViTs types, and defense models, comprehensively assessing the effects of adversarial samples and clearly demonstrating the strengths and weaknesses of different techniques. This framework serves as a valuable reference for researchers in this field.

Additionally, we developed a new method, which uses the Discrete Cosine Transform (DCT) to convert images into the frequency domain to perturb their amplitude and phase components. This innovative approach has shown excellent attack success rates across various models. Inspired by the SSA algorithm, this method involves converting images to the frequency domain using DCT and perturbing their amplitude and phase components. We added Gaussian noise to the amplitude components and then recombined them. This technique allows us to effectively perturb images in the frequency domain, surpassing most replicated attack methods.

This innovation expands the thinking on adversarial attacks, pointing the way for subsequent research and simultaneously promoting the study and upgrading of defense models.

For society, our research serves as an important warning, reminding the community of the importance of developing new defense mechanisms, such as Adversarial Training and High-gradient Denoising. Only by thoroughly analyzing and testing adversarial attacks and defense strategies can the security of existing machine learning systems be enhanced. Understanding and improving these technologies can help design more robust AI systems to withstand real-world adversarial attacks.

Future research should focus on improving the effectiveness and reliability of these defense technologies while exploring more methods that can effectively detect and neutralize adversarial attacks. Policymakers and regulatory bodies should also intervene, establishing appropriate regulations and standards to guide the healthy development of artificial intelligence technology and protect public interests.

However, we must also acknowledge that our research results could potentially be misused for nefarious purposes, posing harm and risk to society. This is something we do not wish to see, and we need to establish a secure mechanism and encourage governments to enact more comprehensive laws quickly. We also hope that more information security professionals will see our research and develop their defense technologies based on our findings. We believe this will form a positive cycle.

For the public, understanding these technologies is also a key factor in ensuring the healthy development of technology. Our research will make the public aware that attacks on original images are very covert and difficult to detect. This will make them more cautious and attentive to image attacks in the future. As researchers and technology developers, we hope that through these efforts, we can ensure the safety and harmony of society while safeguarding technological progress and innovation.

In summary, our research has its own value and significance to society, and the positive effects it brings will far outweigh its negative impacts. We believe that more people will benefit from it.

Appendix

Team contributions

Please describe in one paragraph per team member what each of you contributed to the project.

Person 1 (Jiachen Wang) I have replicated two adversarial sample generation techniques under the category of Advanced Objectives, specifically the Feature Importance-aware Attack (FIA) and Random Patch Attack (RPA). I gathered data on their attack success rates against various models, including CNNs, ViTs, and defensive models. After a thorough investigation of their code, I found that both techniques inherit from MI-FGSM, which led me to conduct a performance comparison among these three attack methods. Additionally, the adversarial samples generated by these methods exhibit distinct visual effects and characteristics, which I have also observed and analyzed.

Person 2 (Zhihao Li) I primarily focus on replicating the VMI-FGSM and VNI-FGSM attack methodologies, aiming at generating adversarial examples. These techniques involve manipulating input data to deceive machine learning models, resulting in erroneous predictions. As part of my exploration, I've delved into the intricacies of these methods, experimenting with replacing the conventional uniform distribution with a Gaussian distribution for sampling neighboring examples. Prior to initiating the hands-on implementation, I dedicated considerable time to conducting an extensive literature review. This comprehensive study encompassed previous research on gradient-based attack strategies, including renowned methods such as FGSM and I-FGSM. Through this thorough investigation, I cultivated a profound understanding of FGSM-based attack methodologies, laying a solid foundation for my subsequent research and experimentation.

Person 3 (Yuyang Luo) In this course project, I meticulously selected a relevant topic and conducted thorough preliminary research to ground our study. I also successfully replicated and compared two prominent methods for generating adversarial samples—MIFGSM and SSA—highlighting their distinct mechanisms and efficiencies. Further, I attempted modifications to the SSA method to enhance its effectiveness, thereby contributing to the existing knowledge base. The culmination of the efforts is documented in a comprehensive project report that not only details our methodologies and findings but also reflects our analytical and practical engagements with the subject matter throughout the course.