Supplementary Material

In the supplementary material section, the paper first presents the ResPA pseudocode. Subsequently, it presents the visualization results of more adversarial examples generated by various attacks, aiming to present the relevant characteristics more intuitively. Finally, relevant work content is supplemented to further enrich the discussion and support of the research.

A. ResPA Pseudo Code

Algorithm 1 Residual Perturbation Attack (ResPA)

Input: A clean image x with ground-truth label y, surrogate model f and the loss function J.

Input: The magnitude of perturbation ϵ; the iteration number T; the decay factor μ; the balanced coefficient γ; the exponential decay rates θ; the upper bound factor β and the sample quantity N.

Output: An adversarial example x^{adv}.

$g_0 = 0; e_0 = 0; x_0^{adv} = x; \rho = \alpha = \epsilon/T$
for $t = 0 \rightarrow T-1$ **do**
  Set $\bar{g} = 0$
  for $i = 0 \rightarrow N-1$ **do**
    Randomly sample an example $x_t^i = x_t^{adv} + \lambda_t^i$
    Calculate the gradient at x_t^i: $g' = \nabla_{x_t^i} J\left(x_t^i, y\right)$
    Compute the EMA of the gradient by:
    $M_{t+1} = \theta \cdot e_t + (1-\theta) \cdot g'$
    Compute the residual gradient g_{t+1}^{res} by:
    $g_{t+1}^{res} = g' - M_{t+1}$
    Compute the predicted point x^*:
    $x^* = x_t^i - \alpha \cdot \frac{g_{t+1}^{res}}{\left\|g_{t+1}^{res}\right\|_1}$
    Calculate the gradient at x^*: $g^* = \nabla_{x^*} J\left(x^*, y\right)$
    Update gradient $\bar{g}$ by:
    $\bar{g} = \bar{g} + \frac{1}{N} \cdot [(1-\gamma) \cdot g' + \gamma \cdot g^*]$
  end for
  Compute the EMA of $\bar{g}$ by:
  $e_{t+1} = \theta \cdot e_t + (1-\theta) \cdot \bar{g}$
  Update the momentum by $g_{t+1} = \mu \cdot g_t + \frac{\bar{g}}{\|\bar{g}\|_1}$
  Update adversarial example x_{t+1}^{adv} by:
  $x_{t+1}^{adv} = \mathrm{Clip}_{\mathrm{x}}^{\epsilon}\left\{x_t^{adv} + \alpha \cdot \operatorname{sign}\left(g_{t+1}\right)\right\}$
end for
$x^{adv} = x_T^{adv}$
return x^{adv}

The algorithm of ResPA is summarized in Algorithm 1.

B. Visualizations on Adversarial Examples

In Figure 1, this paper presents four randomly chosen benign images and their corresponding adversarial examples generated by various attacks. These adversarial examples are generated on the Den-121 model, leveraging MI, VMI, GRA, PGN, AdaMSI, TPA, and ResPA respectively. Notably, these generated adversarial examples are imperceptible to the human eye.

C. Relate work

C.1. Adversarial Attacks

Typically, adversarial attacks can be classified into two categories: white-box attacks and black-box attacks. In the white-box setting, the attacker has full access to the target model. For example, Goodfellow et al. [8] proposed the Fast Gradient Sign Method (FGSM) for generating adversarial examples via one-step gradient update. Subsequently, Kurakin et al. [13] further extended FGSM into an iterative form with a smaller step size, named I-FGSM. Moreover, Madry et al. [20] extended I-FGSM with a random starting point to generate diverse adversarial examples. The existing white-box attacks have achieved remarkable performance by exploiting the knowledge of the target model. Conversely, black-box attacks are more practical as they can only obtain limited or no information regarding the target model. There are two types of black-box adversarial attacks: query-based attacks [1, 10] and transfer-based attacks [3, 24, 25]. Query-based attacks generally require hundreds or even thousands of queries to generate adversarial examples, which renders them inefficient in real-world applications. In contrast, transfer-based attacks generate adversarial examples on the surrogate model. These examples are also able to attack other models without accessing the target model, thereby resulting in high practical applicability and attracting increasing attention.

Regrettably, adversarial examples generated through white-box attacks typically exhibit limited transferability. To enhance adversarial transferability, a variety of momentum-based attacks have been proposed, such as MI [3], NI [16], VMI [25], GRA [33]. Moreover, several in-

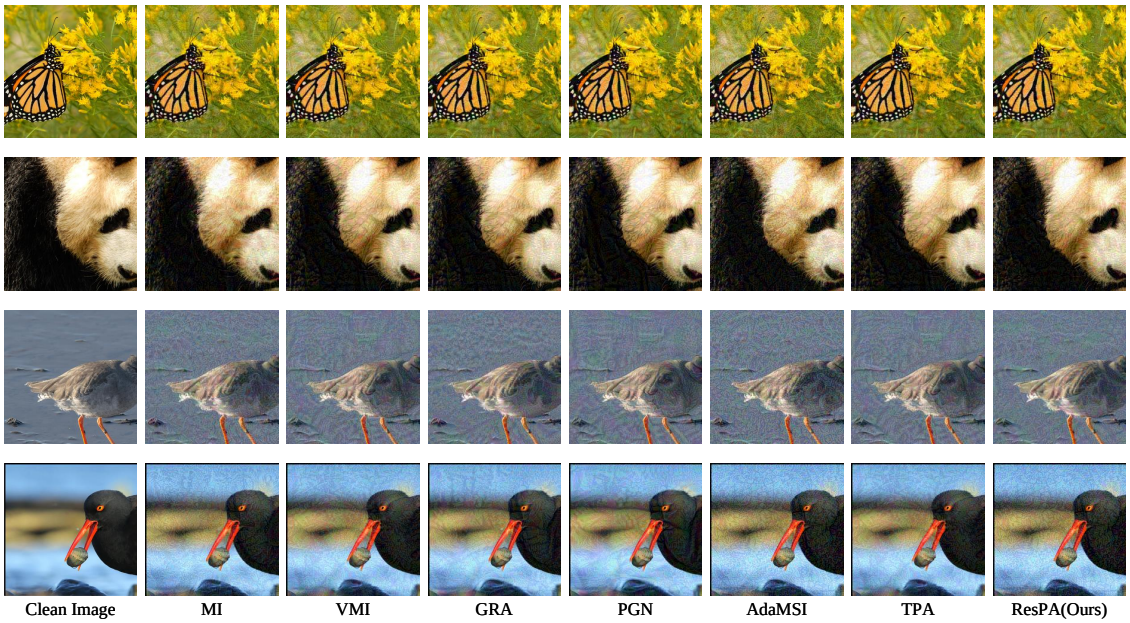

Figure 1. Visualize the original samples and adversarial examples. The adversarial examples are generated on the Den-121 model by various attack methods, with a maximum perturbation value of $\epsilon = 16$.

put transformation methods have also been put forward, such as DIM [30], TIM [4], SIM [16], Admix [26], SSA [19], BSR [24], etc., which augment images utilized for adversarial perturbation computation to boost transferability. In addition, some approaches improve adversarial transferability from different perspectives. For instance, Liu et al. [17] proposed an ensemble attack, which simultaneously attacks multiple surrogate models. Wu et al. [28] employed an adversarial transformation network that can capture the most harmful deformations to adversarial noises. References [5, 7, 27] search for adversarial examples in flat regions to achieve better transferability.

C.2. Adversarial Defense

The presence of adversarial examples presents a significant security risk to deep neural networks (DNNs). To mitigate this risk, researchers have put forward a range of methods, among which adversarial training [15, 20] has emerged as a widely-utilized and effective approach. By supplementing the training data with adversarial examples, this method enhances the robustness of trained models against adversarial assaults. However, while adversarial training is effective, it entails high training costs, particularly when dealing with large-scale datasets and complex networks. Consequently, researchers have proposed innovative defense methods as alternatives. Guo et al. [9] use various input transformations, such as JPEG compression and total variance minimization, to eliminate adversarial perturbations from input images. Liao et al. [15] train a denoising autoencoder, known as the High-level representation guided denoiser (HGD), to purify the adversarial perturbations. Xie et al. [29] suggest randomly resizing the image and adding padding to lessen the adversarial impact, which is named Randomized resizing and padding (RP). Xu et al. [31] propose the Bit depth reduction (Bit-Red) method, which decreases the number of bits per pixel to restrain the perturbation. Liu et al. [18] proposed Feature Distillation (FD) to safeguard against adversarial attacks by applying a JPEG-based compression method to adversarial images. Cohen et al. [2] utilize randomized smoothing (RS) to train a certifiably robust classifier. Naseer et al. [21])propose a Neural Representation Purifier (NRP) to get rid of the perturbation.

C.3. Flat Minima

Since Hochreiter et al. [11] pointed out that models with good generalization ability might have flat minima, the academic community [23, 32] has conducted in-depth research on the relationship between the flatness of minima and the model generalization ability from both empirical and theoretical perspectives. Li et al. [14] noticed that skip connections can facilitate the formation of flat minima. This finding provides strong evidence for explaining the crucial role of skip connections in training extremely deep networks. Similarly, Santurkar et al. [22] discovered that Batch Normalization (BatchNorm) can significantly smooth the optimization surface during the training process. The "Sharpness-Aware Minimization" (SAM) [6] method enhances the model's generalization ability by minimizing the loss value and sharpness simultaneously and searching for parameters within the neighborhood where the loss value

remains consistently low. Jiang et al. [12] studied 40 complexity metrics, and the results showed that the sharpness-based metric has the most significant correlation with generalization ability. Zhao et al. [32] also confirmed that adding the gradient norm of the loss function helps the optimizer find flat local minima.

References

[1] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *Proceedings of the European conference on computer vision (ECCV)*, pages 154–169, 2018. 1

[2] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019. 2

[3] 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 Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018. 1

[4] 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*, pages 4312–4321, 2019. 2

[5] Mingyuan Fan, Xiaodan Li, Cen Chen, Wenmeng Zhou, and Yaliang Li. Transferability bound theory: Exploring relationship between adversarial transferability and flatness. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024. 2

[6] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. 2

[7] Zhijin Ge, Fanhua Shang, Hongying Liu, Yuanyuan Liu, and Xiaosen Wang. Boosting adversarial transferability by achieving flat local maxima. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023. 2

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

[9] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018. 2

[10] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International conference on machine learning*, pages 2484–2493. PMLR, 2019. 1

[11] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997. 2

[12] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*. 3

[13] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 1

[14] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018. 2

[15] 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 conference on computer vision and pattern recognition*, pages 1778–1787, 2018. 2

[16] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and J. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2019. 1, 2

[17] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017. 2

[18] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868. IEEE, 2019. 2

[19] 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

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 2

[21] Muzammal Naseer, Salman 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 262–271, 2020. 2

[22] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018. 2

[23] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *ArXiv e-prints*, pages arXiv–1609, 2016. 2

[24] Kunyu Wang, Xuanran He, Wenxuan Wang, and Xiaosen Wang. Boosting adversarial transferability by block shuffle and rotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24336–24346, 2024. 1, 2

[25] 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

[26] 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 16158–16167, 2021. 2

[27] Tao Wu, Tie Luo, and Donald C Wunsch. Gnp attack: Transferable adversarial examples via gradient norm penalty. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3110–3114. IEEE, 2023. 2

[28] Weibin Wu, Yuxin Su, Michael R Lyu, and Irwin King. Improving the transferability of adversarial samples with adversarial transformations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9024–9033, 2021. 2

[29] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018. 2

[30] 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*, pages 2730–2739, 2019. 2

[31] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society, 2018. 2

[32] Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *International Conference on Machine Learning*, pages 26982–26992. PMLR, 2022. 2, 3

[33] Hegui Zhu, Yuchen Ren, Xiaoyan Sui, Lianping Yang, and Wuming Jiang. Boosting adversarial transferability via gradient relevance attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4741–4750, 2023. 1