Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Sister Conferences Best Papers Track

PAR-AdvGAN: Improving Adversarial Attack Capability with Progressive Auto-Regression AdvGAN (Extended Abstract)*

Jiayu Zhang[1], **Zhiyu Zhu**[2], **Xinyi Wang**[3], **Silin Liao**[4], **Zhibo Jin**[2], **Flora Salim**[5] and **Huaming Chen**[6]
[1]Suzhou University of Technology, China
[2]University of Technology Sydney, Australia
[3]University of Malaya, Malaysia
[4]Nanning Normal University, China
[5]University of New South Wales, Australia
[6]The University of Sydney, Australia
zjy@szut.edu.cn, xinyiwangnoctis@outlook.com, anivie@email.nnnu.edu.cn, {zhiyu.zhu, zhibo.jin}@student.uts.edu.au, flora.salim@unsw.edu.au, huaming.chen@sydney.edu.au

Abstract

Deep neural networks are vulnerable to adversarial examples, posing serious security risks. While GAN-based attacks such as AdvGAN enable fast adversarial example generation, they typically operate in a single step, limiting attack effectiveness and transferability. In this work, we propose PAR-AdvGAN, a progressive auto-regressive GAN framework that iteratively refines adversarial perturbations. By conditioning each iteration on the previous output, PAR-AdvGAN produces stronger, more transferable attacks while maintaining low visual distortion and high generation speed. Experiments on multiple models and datasets demonstrate that PAR-AdvGAN significantly outperforms baseline methods in attack success rate, achieving up to 335.5 frames per second, highlighting its practical efficiency for black-box attack scenarios.

1 Introduction

Deep neural networks are widely deployed in safety-critical systems such as image classification [Li *et al.*, 2020], emotional analysis [Qiang *et al.*, 2020], and item recommendations [Pan *et al.*, 2020]. However, their vulnerability to adversarial examples [Szegedy *et al.*, 2013; Ma *et al.*, 2021; Deng *et al.*, 2020] poses severe security risks. Existing adversarial attack methods fall into two main categories: gradient-based methods (e.g., FGSM [Goodfellow *et al.*, 2014b], PGD [Madry *et al.*, 2017]) and GAN-based methods. Gradient-based methods require repeated gradient computation for each input, leading to high computational cost. In contrast, GAN-based methods (e.g., AdvGAN [Goodfellow *et al.*, 2014a]) train a generator once and can generate adversarial examples in a single forward pass, making them suitable for large-scale security evaluation.

Despite their efficiency, GAN-based attacks show limitations. AdvGAN generates perturbations in a single iteration, which can cause perturbation escalation. The magnitude of perturbations will increase rapidly with training iterations, sacrificing visual imperceptibility. Additionally, single-shot generation fails to capture fine-grained input structures, resulting in poor transferability to unseen models, which is critical for real-world black-box attack scenarios.

To address the challenges, we propose PAR-AdvGAN, introducing three key innovations: (1) A progressive generator network that takes both the current adversarial example and the original clean image as input, enabling incremental perturbation refinement. (2) An auto-regressive iterative mechanism that generates perturbations sequentially, leveraging historical generation results to improve attack effectiveness. (3) A multi-stage joint loss function that combines adversarial loss, perturbation norm loss, and distance loss to balance attack success rate, imperceptibility, and training stability.

2 Motivation

Although AdvGAN and other GAN-based attacks can efficiently generate adversarial examples, their single-shot generation strategy has intrinsic limitations. First, perturbations tend to escalate during training, which can compromise visual imperceptibility and reduce stealthiness. Second, treating each generated sample independently prevents the attack from leveraging historical perturbation information, limiting the strength and transferability of adversarial examples, especially on unseen or defense-trained models. Third, single-step generation fails to capture fine-grained structures of the input, which is critical for realistic black-box attack scenarios.

These challenges motivate our PAR-AdvGAN, which introduces a progressive, auto-regressive framework. This de-

*Original paper: Jiayu Zhang, Zhiyu Zhu, Xinyi Wang, Silin Liao, Zhibo Jin, Flora Salim, and Huaming Chen. PAR-AdvGAN: Improving Adversarial Attack Capability with Progressive Auto-regression AdvGAN. Research Track: European Conference, ECML PKDD 2025, Porto, Portugal, September 15–19, 2025, Proceedings, Part VII, Springer-Verlag, Berlin, Heidelberg, 164–180. https://doi.org/10.1007/978-3-032-06109-6_10

sign improves transferability across both convolutional networks and transformer-based architectures.

3 The Proposed PAR-AdvGAN Framework

3.1 Problem definition

Given a clean data distribution p_{data}, benign samples $X \subseteq p_{\text{data}}$, and a target network f with true label m for sample $x \in X$, the goal of an untargeted adversarial attack is:

$$f(x_{\text{adv}}) \neq m, \quad \|x_{\text{adv}} - x\|_n \leq \epsilon,$$

where $\|\cdot\|_n$ denotes the n-norm (e.g., L_2), and ϵ is the maximum allowed perturbation.

3.2 Progressive Generator Network

PAR-AdvGAN consists of a progressive generator G and a discriminator D. To maintain information from the original sample x_0, the generator G is modified to accept both the current adversarial example x_{adv}^{t-1} and x_0:

$$P_t = G\left(x_{\text{adv}}^{t-1}, x_0\right),$$

where P_t is the perturbation at iteration t. The adversarial example is then updated as:

$$x_{\text{adv}}^{t} = \text{clip}\left(x_{\text{adv}}^{t-1} + P_t\right),$$

ensuring incremental refinement. The generator leverages both x_{adv}^{t-1} and x_0 to produce perturbations that remain close to the original input while gradually enhancing attack results.

3.3 Auto-Regression Iterative Method

The generator iteratively updates adversarial examples using an auto-regressive mechanism. Let the network N map input x to label m. At iteration t, the update satisfies:

$$N(x_{\text{adv}}^{t}) \neq m.$$

The gradient of the adversarial loss with respect to generator parameters θ_G can be expressed as:

$$\nabla_{\theta_G} = \frac{\partial L_{\text{adv}}}{\partial x_{\text{adv}}^{t} + G(x_{\text{adv}}^{t})} \cdot \frac{\partial G(x_{\text{adv}}^{t})}{\partial \theta_G},$$

which guides the generator to iteratively refine perturbations in a self-correcting manner.

3.4 Generator Constraints

To prevent perturbation explosion and maintain visual fidelity, PAR-AdvGAN employs three complementary losses:

1. Adversarial loss (encouraging misclassification):

$$L_{\text{adv}} = -\text{CrossEntropy}\left(x_{\text{adv}}^{t}, m\right)$$

2. Perturbation norm loss (controlling magnitude):

$$L_p = \|P_t\|_2 = \left\|G\left(x_{\text{adv}}^{t-1}, x_0\right)\right\|_2$$

3. Distance loss (enforcing proximity to original sample):

$$L_d = \left\|x_{\text{adv}}^{t} - x_0\right\|_2$$

The generator is trained to minimize a combined loss:

$$g_{\theta_G} = \nabla_{\theta_G}\Big(L_G + \lambda_1 L_p + \lambda_2 L_d + \lambda_3 L_{\text{adv}}\Big),$$

where

$$L_G = \left(1 - D(x_{\text{adv}}^{t})\right)^2$$

is the GAN loss to deceive the discriminator D, and $\lambda_1, \lambda_2, \lambda_3$ balance the contributions of each term.

The parameters of generator G are updated via gradient descent:

$$\theta_G = \theta_G - \eta_2 \cdot g_{\theta_G},$$

and the parameters of discriminator D are updated similarly:

$$\theta_D = \theta_D - \eta_1 \cdot \nabla_{\theta_D} L_D,$$

where

$$L_D = \left(1 - D(x)\right)^2 + D(x_{\text{adv}}^{t})^2.$$

Here η_1 and η_2 denote the learning rates of the discriminator and the generator, respectively.

4 Experiment

4.1 Experimental Setup

Datasets and Models: Experiments were conducted on an ImageNet-compatible dataset of 1,000 images (299×299×3) [Papernot *et al.*, 2018]. We evaluated attacks on both normally-trained models—Inception-v3 (Inc-v3) [Szegedy *et al.*, 2016], Inception-v4 (Inc-v4), Inception-ResNet-v2 (IncRes-v2) [Szegedy *et al.*, 2017], ResNet-v2-50 (Res-50), ResNet-v2-101 (Res-101), ResNet-v2-152 (Res-152) [He *et al.*, 2016a; He *et al.*, 2016b] and defense-trained models [Tramèr *et al.*, 2017] (Inc-v3ens3, Inc-v3ens4, IncRes-v2ens) using ensemble adversarial training. Additionally, experiments included ViT-B/16 [Dosovitskiy *et al.*, 2021] to test performance on transformer-based architectures.

Baseline Methods: We compared PAR-AdvGAN with the original AdvGAN and seven state-of-the-art transferable attacks: FGSM [Goodfellow *et al.*, 2014b], BIM [Kurakin *et al.*, 2018], PGD [Madry *et al.*, 2017], DI-FGSM [Xie *et al.*, 2019], TI-FGSM [Dong *et al.*, 2019], MI-FGSM [Dong *et al.*, 2018], and SINI-FGSM [Lin *et al.*, 2019]

Parameter Settings: PAR-AdvGAN and AdvGAN were trained for 60 epochs with initial learning rates of 0.001, reduced to 0.0001 at epoch 50. For iterative gradient-based attacks, we followed standard parameter settings from prior literature. Attack performance was measured using Attack Success Rate (ASR) and Frames Per Second (FPS), with perturbation rate ensuring visual fidelity.

To evaluate the effectiveness of PAR-AdvGAN, we designed experiments to answer three key questions:**RQ1:** How does PAR-AdvGAN improve attack success rate compared to AdvGAN? **RQ2:** How does PAR-AdvGAN perform in attack transferability and speed compared to state-of-the-art methods? **RQ3:** Why does PAR-AdvGAN work effectively?

4.2 RQ1: Attacking Performance

As shown in Table. 3, we compare the attack success rates of the original AdvGAN and our proposed PAR-AdvGAN at three different perturbation rates of 8, 9, and 10. The comparisons are conducted using Inc-v3 and Inc-v4 as surrogate models and attacks are lunched on IncRes-v2. The results indicate that in most cases, our algorithm outperforms AdvGAN in terms of attack success rate.

Model	Attack	Perturbation	Inc-v4	Res-50	Res-101	Res-152	Inc-v3ens3	Inc-v3ens4	IncResv2ens	mASR
Inc-v3	AdvGAN	8.41/9.88/10.26	32.9/61.9/65.9	42.6/77.7/82.8	47.8/75.3/83.1	38.8/66.6/72.8	15.1/44.0/52.5	30.2/54.5/63.0	9.9/25.7/26.5	31.04/57.95/63.80
	PAR-AdvGAN	**8.29/9.27/9.95**	53.7/63.1/68.4	**74.8/85.0/89.8**	**79.3/86.2/89.5**	**68.2/78.0/82.5**	**39.4/53.5/63.8**	**45.7/60.8/69.4**	**28.0/39.0/46.5**	**55.58/66.51/72.84**
	FGSM	8.79/9.74/10.69	26.2/28.0/30.3	26.2/29.0/30.6	23.3/25.8/27.6	22.8/24.1/27.0	13.8/14.5/15.5	14.0/14.3/14.3	6.0/6.1/6.1	18.9/20.25/21.62
	BIM	8.46/9.50/9.96	47.6/52.2/56.6	42.1/45.8/48.5	36.7/40.6/42.9	35.6/39.2/39.3	14.4/16.0/15.8	14.1/14.6/14.5	8.5/8.1/8.4	28.42/30.92/32.28
	PGD	8.76/9.79/10.35	44.6/46.2/50.2	38.6/40.7/43.5	33.1/35.5/37.3	28.9/34.5/35.8	12.4/14.2/14.4	12.4/13.3/13.1	6.8/7.4/7.7	25.25/27.40/28.85
	DI-FGSM	8.51/9.56/10.02	**66.0/70.0/70.9**	54.0/60.0/61.4	50.7/55.1/55.2	49.3/53.7/52.7	19.1/19.6/20.0	18.4/20.4/19.1	10.5/11.0/11.0	38.28/41.40/41.47
	TI-FGSM	8.60/9.65/10.10	52.4/55.3/56.5	39.7/45.2/45.2	34.4/39.9/41.1	34.8/39.0/43.1	31.6/34.8/35.3	34.1/37.7/39.2	21.4/25.9/25.8	35.48/39.68/40.88
	MI-FGSM	8.98/9.77/10.55	44.5/47.9/51.4	39.9/41.7/45.1	36.5/38.8/41.0	34.3/37.8/38.9	16.3/17.1/16.8	15.6/14.9/16.2	6.5/7.7/7.4	27.65/29.41/30.97
	SINI-FGSM	8.99/9.77/10.55	56.0/59.7/64.2	53.8/58.0/62.5	47.2/51.1/55.2	45.6/50.7/53.8	24.6/24.4/26.4	23.9/23.8/25.9	11.0/12.0/12.6	37.44/39.95/42.94

Table 1: The attack success rates (%) on four undefended models and three adversarial trained models by various transferable adversarial attacks. The adversarial examples are crafted on Inc-v3 with different perturbations. The best results are in bold.

Model	Attack	Perturbation	Inc-v3	Inc-v4	Res-50	Res-101	Res-152	Inc-v3ens3	Inc-v3ens4	IncResv2ens	mASR
ViT-B/16	AdvGAN	10.59/11.30/12.25	75.6/72/80.3	70.3/60.5/70.2	72.5/75.7/84.3	70.8/75.2/84.9	68.9/71.8/77.1	61.2/74.9/83.4	61.8/69.5/79.1	53.8/50/60.5	66.86/68.7/77.46
	PAR-AdvGAN	**10.21/11.12/12.01**	**76.1/81.8/84.5**	63.6/67/70.9	**80.3/84.9/87.3**	**79.9/84.8/87.3**	**75.7/80.4/83**	**82.3/87.6/90.6**	**73.9/79.5/83.1**	**55.2/62.5/66.3**	**73.38/78.56/81.63**
	FGSM	10.76/11.71/12.65	35/36.3/38.6	31.9/34.6/36.4	33.6/35.1/37.6	32.2/34/36	31.9/34/36	27.9/30.1/31.7	29.9/30.3/31.5	23.8/25.3/26.5	30.78/32.46/34.29
	BIM	10.90/11.47/12.37	55.5/58.7/60.7	50.5/49.7/54.1	54.2/55.6/59.2	48.5/51.4/56.8	46.6/47.6/54.4	39.2/38.5/42.4	39.7/41/42.8	30.5/32.4/35.1	45.59/46.86/50.69
	PGD	10.73/11.23/12.24	46.6/48/53	40.7/42.3/44	44.2/47/49.6	40.3/42.7/45.7	37.9/39.6/42.4	28.2/29.7/31.7	31.8/31.5/33.9	21.8/22.9/25.2	36.44/37.96/40.69
	DI-FGSM	10.59/11.58/12.04	75.6/77.8/78.9	**70.3/74.1/75.1**	72.5/77/74.9	70.8/74.5/73.1	68.9/71.8/71.9	61.2/67.7/67.4	61.8/65.8/67.5	53.8/58.5/57.9	66.86/70.9/70.84
	TI-FGSM	10.77/11.21/12.23	60.5/61/65.4	54.1/56.2/59.9	53.1/55.4/60.2	52.6/54.6/58.3	50.7/54.8/59.3	56.4/58.6/61.2	59.4/62.4/63.6	52.1/53/57.7	54.86/57/60.7
	MI-FGSM	10.75/11.58/12.36	53.7/55.6/59.1	47.4/49/52.8	50.9/54.3/57.7	47.7/50.5/53.2	45.5/48.9/50.9	38.6/40.5/43.7	40.5/42.6/43.9	30.7/33.7/36.7	44.38/46.89/49.75
	SINI-FGSM	10.83/11.66/12.45	58.2/61.8/65.5	52.7/56.3/60.9	55.7/61.2/64.7	51.1/56.1/59.6	49.8/54.1/57.9	44.5/47.5/49.4	44.1/46.7/50.4	37.9/39.7/43.9	49.25/52.93/56.54

Table 2: The attack success rates (%) on four undefended models and three adversarial trained models by various transferable adversarial attacks. The adversarial examples are crafted on ViT-B/16 with different perturbations. The best results are in bold.

Model	Attack	IncRes-v2
Inc-v3	AdvGAN	13.9/38.9/43.2
	PAR-AdvGAN	**27.1/35.2/41.4**
Inc-v4	AdvGAN	6.1/9.6/15.9
	PAR-AdvGAN	**21.4/30.8/34.7**

Table 3: ASR (%) of AdvGAN and PAR-AdvGAN on IncRes-v2. The adversarial examples are crafted on Inc-v3 and Inc-v4.

We can see that compared to the most representative AdvGAN algorithm, PAR-AdvGAN has made significant improvements for attacking performance at a low perturbation rate. Specifically, similar to AdvGAN, PAR-AdvGAN, as a generative model, does not require additional gradient calculations based on different input data after training the generator. Compared to traditional gradient-based black-box transferable attack methods, it possesses faster attack speed. Therefore, we consider the PAR-AdvGAN algorithm feasible and suitable for attack scenarios that demand high transferability and fast generation of adversarial samples.

4.3 Effectiveness Experiment for RQ2: Transferability and Attack Speed

To validate the transferability and attack speed of PAR-AdvGAN compared to other SOTA methods, we conduct the experiments using various attack methods on Inc-v3, Inc-v4, IncRes-v2 and ViT-B/16 as source models to generate adversarial samples. Due to space limitations of the extended abstract, we will use the Inc-v3 and ViT-B/16 models as an example here. We then conduct transferable attacks on different target models and use ASR and FPS as the main metrics, to validate the effectiveness of our algorithm.

Experiments on Inc-v3

As shown in Table. 1, we conduct attacks using Inc-v3 as the source model with three different perturbation rates on target models of Inc-v4, Res-50, Res-101, Res-152, Inc-v3ens3, Inc-v3ens4, and IncRes-v2. We can see that our PAR-AdvGAN algorithm has achieved an average increase of 30.3% in attack success rate compared to other baselines. Moreover, despite DI-FGSM achieving better performance than PAR-AdvGAN on Inc-v4, which may be attributed to the randomness in model training, a comprehensive comparison across all models reveals that the attack success rate of PAR-AdvGAN is elevated by 24.6% compared to the best-performing competing baseline, DI-FGSM.

Experiments on ViT-B/16

In Table 2, we present the results of attacks using ViT-B/16 as the source model with three different perturbation rates on several target models, including Inc-v3, Inc-v4, Res-50, Res-101, Res-152, Inc-v3ens3, Inc-v3ens4, and IncRes-v2. Unlike traditional convolutional neural networks (CNNs), Vision Transformers (ViTs) adopt a fundamentally different architecture for image classification tasks. Our experimental findings show that the proposed PAR-AdvGAN algorithm performs exceptionally well when transferred to ViT-based models, achieving an average increase of 6.85% in attack success rate (ASR) compared to AdvGAN across all target models. Specifically, PAR-AdvGAN consistently delivers the highest mean ASR values of 73.38%, 78.56%, and 81.63% across the three perturbation rates, surpassing all baseline methods, including DI-FGSM, which was the best performer in certain cases. These results underscore the robustness and transferability of PAR-AdvGAN, demonstrating its ability to maintain high effectiveness not only with traditional CNNs but also with more recent transformer-based models like ViT, thus proving its versatility and reliability across different model architectures.

Attack Speed Analysis

As shown in Table. 4, we evaluated the computational efficiency of PAR-AdvGAN and seven competitive baselines using Inc-v3, Inc-v4, and IncRes-v2 as source models. We

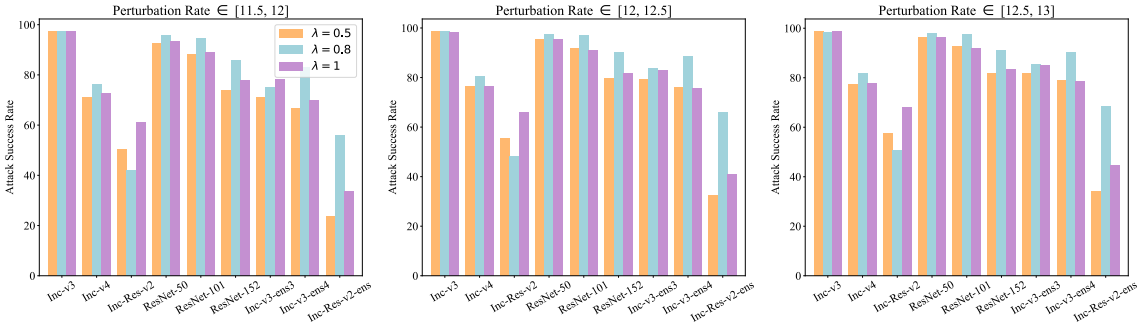

Figure 1: The performance of PAR-AdvGAN at different high perturbation rate intervals

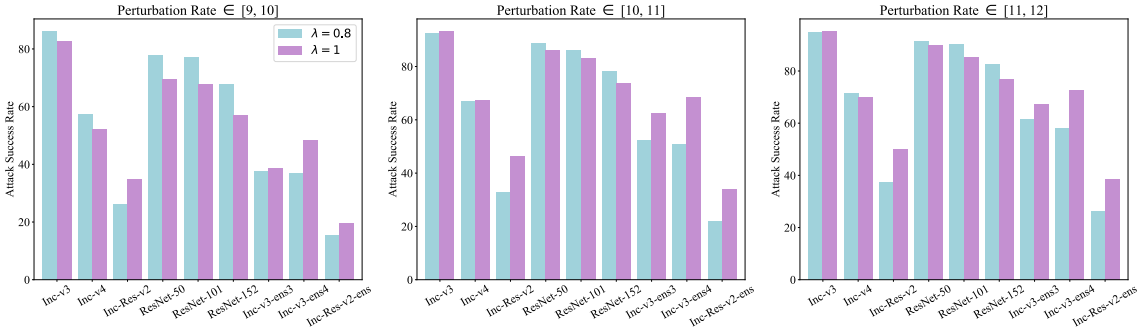

Figure 2: The performance of PAR-AdvGAN at different low perturbation rate intervals

Method	Inc-v3	Inc-v4	IncRes-v2
FGSM	176.5	116.7	76.1
BIM	10.6	6.5	4.1
PGD	5.5	3.3	2.1
DI-FGSM	10.8	6.6	4.1
TI-FGSM	10.6	6.5	4.2
MI-FGSM	42.7	26	16.4
SINI-FGSM	8.6	5.3	3.3
PAR-AdvGAN	**335.5**	**291.3**	**332.8**

Table 4: FPS comparison of PAR-AdvGAN with seven competitive baselines

use FPS as the metric for measuring attack speed, representing the number of images that can be processed by the attack per second. It can be observed that across Inc-v3, Inc-v4, and IncRes-v2, PAR-AdvGAN exhibits speed improvements of 61, 88.3, and 158.5 times over the slowest-performing PGD algorithm among the competitive baselines. Furthermore, in comparison to the fastest-performing FGSM algorithm among the competitive baselines, PAR-AdvGAN achieves speed enhancements of 1.9, 2.5, and 4.4 times, respectively. We assert that PAR-AdvGAN demonstrates significantly higher attack speed in comparison to traditional gradient-based transferable methods, while simultaneously achieving state-of-the-art transferability performance.

4.4 Ablation Experiment for RQ3

We investigate the effects of parameter λ on the attack transferability as it is an important parameter to control the perturbation range. Fig. 1 shows the performance of PAR-AdvGAN with Inc-v3 as the source model, with a fixed ϵ of 20, and λ set to 0.5, 0.8, and 1 for different target models. At λ of 0.5, the specific perturbation rates are 11.75, 12.56, and 12.89. At λ of 0.8, the specific perturbation rates are 11.55, 12.59, and 12.90. At λ of 1, the specific perturbation rates are 11.66, 12.45, and 12.81. We can see that at higher perturbation rate intervals, setting λ to 0.8 achieves best transferability performance. Fig. 2 compares the results with fixed ϵ of 16 and λ set to 0.8 and 1. At λ of 0.8, the specific perturbation rates are 9.40, 10.60, and 11.20. At λ of 1, the corresponding perturbation rates are 8.95, 10.88, and 11.40. For lower perturbation rates, setting λ to 1 achieves the best transferability.

5 Conclusion

In this paper, we introduce PAR-AdvGAN, which enhances adversarial attacks via a progressive generator, auto-regressive iterative training, and sample-distance constraints. Extensive experiments demonstrate that PAR-AdvGAN achieves superior attack transferability and significantly faster attack efficiency compared to state-of-the-art gradient-based transferable attacks.

References

[Deng *et al.*, 2020] Yao Deng, Xi Zheng, Tianyi Zhang, Chen Chen, Guannan Lou, and Miryung Kim. An analysis of adversarial attacks and defenses on autonomous driving models. In *2020 IEEE international conference on pervasive computing and communications (PerCom)*, pages 1–10. IEEE, 2020.

[Dong *et al.*, 2018] 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.

[Dong *et al.*, 2019] 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.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[Goodfellow *et al.*, 2014a] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Goodfellow *et al.*, 2014b] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[He *et al.*, 2016a] 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.

[He *et al.*, 2016b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[Kurakin *et al.*, 2018] 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.

[Li *et al.*, 2020] Xiangrui Li, Xin Li, Deng Pan, and Dongxiao Zhu. On the learning property of logistic and softmax losses for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4739–4746, 2020.

[Lin *et al.*, 2019] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

[Ma *et al.*, 2021] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021.

[Madry *et al.*, 2017] 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.

[Pan *et al.*, 2020] Deng Pan, Xiangrui Li, Xin Li, and Dongxiao Zhu. Explainable recommendation via interpretable feature mapping and evaluation of explainability. *arXiv preprint arXiv:2007.06133*, 2020.

[Papernot *et al.*, 2018] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.

[Qiang *et al.*, 2020] Yao Qiang, Xin Li, and Dongxiao Zhu. Toward tag-free aspect based sentiment analysis: A multiple attention network approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Szegedy *et al.*, 2016] 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.

[Szegedy *et al.*, 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[Tramèr *et al.*, 2017] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[Xie *et al.*, 2019] 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.