以 DPM-Solver++ 加速之 GDMP 對抗淨化：在交通號誌分類上的評估與分析

作者 : **黃文良**，沈婉瑛，劉冠廷，黃意婷

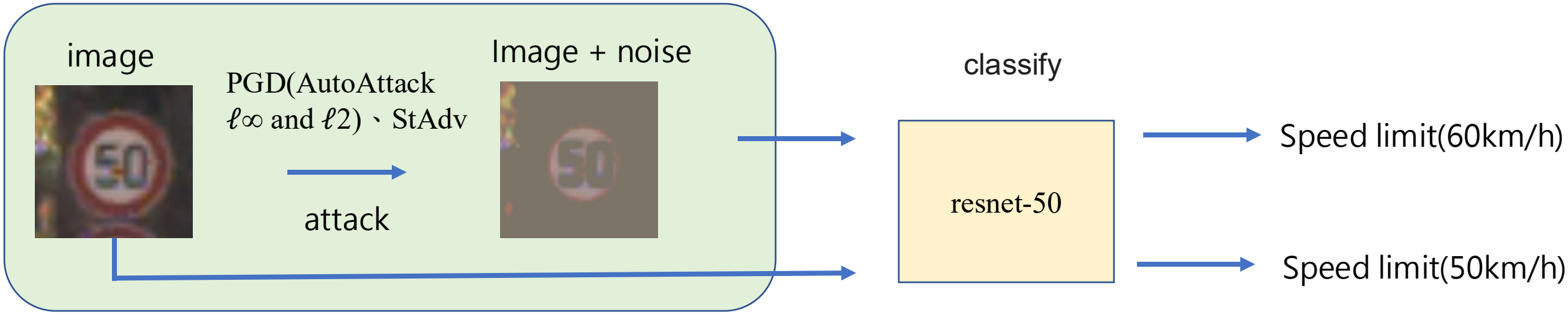
單位 : 國立臺灣科技大學電機工程學系

聯絡信箱 : {B11107152, M11307508, B11207111, ythuang}@mail.ntust.edu.tw

Paper ID : 168

介紹

自駕感知易受**對抗攻擊**影響而誤判，本研究針對真實交通號誌在對抗攻擊下的安全風險，先以淨化修正影像，使模型能正確辨識，以確保自駕車安全。



動機與問題

- 自駕感知易受對抗攻擊(PGD $\ell\infty/\ell2$、StAdv)。
- 既有擴散淨化步數多、延遲高，難以接近即時。

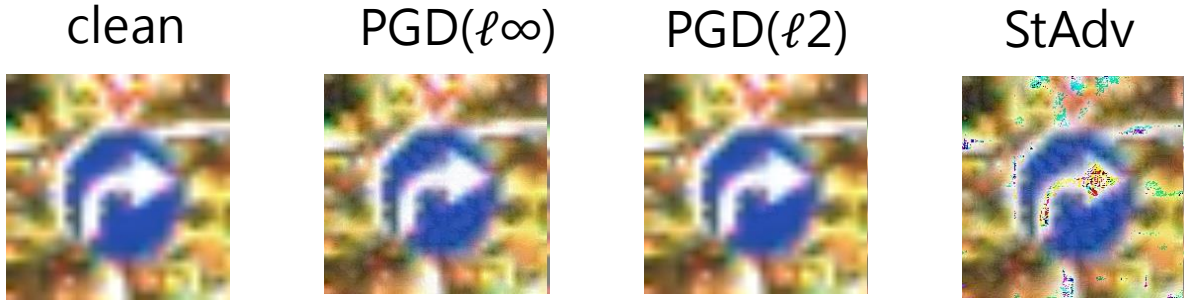
- 能否在較少步數下，保留關鍵特徵同時去除擾動？

研究聚焦在在**較少步數、較低延遲**條件下，檢驗擴散式淨化是否能兼顧**關鍵特徵與去擾動效果**。

研究貢獻

- 結合 **GDMP** 與 **DPM-Solver++**之加速採樣框架。
- 在 **GTSRB**[10]上系統性評估 **Standard/Robust** 指標，並與傳統前處理對比，同時量測**Purification latency**。
- 平均淨化延遲可控制在秒級 (約 5.9 秒)。

背景：對抗攻擊



像素級擾動 (Pixel-level Noise)

- **PGD(ℓ∞)**[13]：採多步迭代梯度攻擊，限制**單一像素**的最大改變量，產生人眼難以察覺的細微雜訊。
- **PGD(ℓ2)**[13]：採多步迭代梯度攻擊，限制**整張影像**的總擾動能量(歐氏距離)，控制整體的變異程度。
- **FGSM**[11]：利用梯度的**符號方向**進行**單步**攻擊，計算極快但攻擊力較弱，是早期最具代表性的基礎攻擊方法。

空間形變 (Spatial Deformation)

- **StAdv**[20](幾何位移): 透過最佳化流場產生幾何形變與像素位移，破壞影像結構語義而非直接修改像素數值。

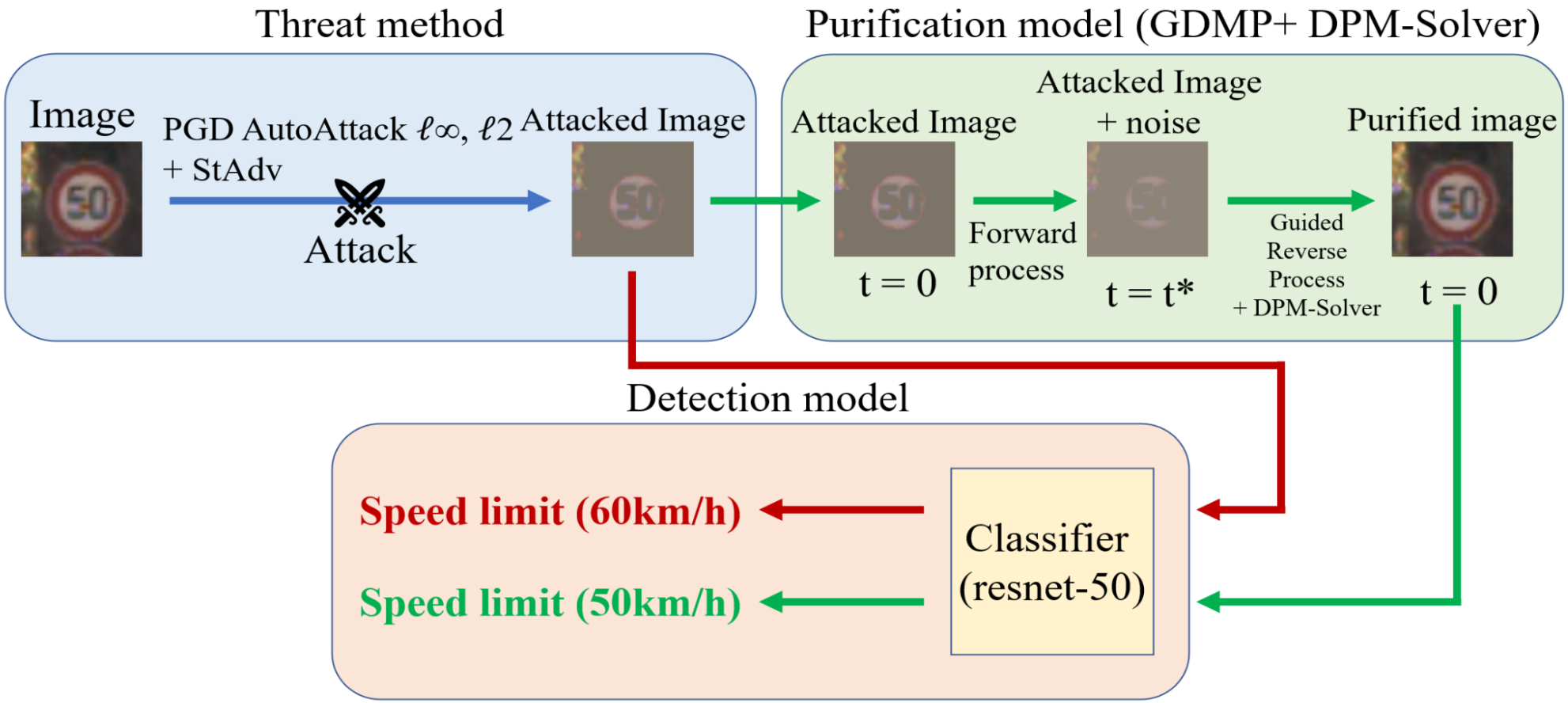
背景：擴散模型

- 擴散淨化：前向加噪稀釋對抗擾動，反向去噪恢復影像的關鍵特徵。

1. **DiffPure**[15]：把對抗樣本送到較高噪聲層後再進行反向採樣。
2. **IDC**[8]：把分類轉成「影像→標籤模板影像」的擴散生成，再用相似度評分決定類別。
3. **DDIM**[18]：以非馬可夫、確定性路徑加速反向採樣，能用遠少於 DDPM 的步數達到可接受品質。
4. **DPM- Solver++**[13]：把反向擴散視作 ODE，利用高階多步展開結合前幾步的模型輸出做二/三階修正。
5. **GDMP**[19]：在反向去噪期間加入「條件式距離引導」(如 MSE/SSIM)。

- 挑戰：少步數下的誤差累積與關鍵特徵保持。

方法總覽

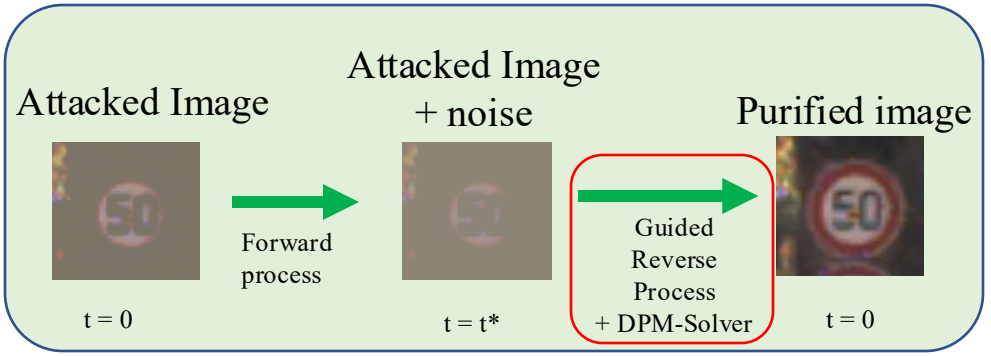

Threat method
Purification model (GDMP+ DPM-Solver)
Image
PGD AutoAttack ℓ∞, ℓ2
+ StAdv
Attack
Attacked Image
Attacked Image
Attacked Image
+ noise
Purified image
t = 0
Forward
process
t = t*
Guided
Reverse
Process
+ DPM-Solver
t = 0
Detection model
Speed limit (60km/h)
Speed limit (50km/h)
Classifier
(resnet-50)

關鍵公式與更新

Attacked Image
+ noise

t = t*

Purification model(GDMP+ DPM-Solver)



Purified image

t = 0

$$\widehat{x_0} = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t)}{\sqrt{\bar{\alpha}_t}}$$

預訓練
U-Net($\epsilon\theta(xt)$)

x_0-parameterization

DPM-Solver++

後續步

初始步

$$\tilde{x}_{t_i} \leftarrow \frac{\alpha_{t_i}}{\alpha_{t_{i-1}}}\tilde{x}_{t_{i-1}} - \alpha_{t_i}(e^{-h_i}-1)D_i$$

$$D_i \leftarrow \left(1+\frac{1}{2r_i}\right)x_\theta(\tilde{x}_{t_{i-1}}, t_{i-1}) - \frac{1}{2r_i}x_\theta(\tilde{x}_{t_{i-2}}, t_{i-2})$$

$$\tilde{x}_{t_1} \leftarrow \frac{\alpha_{t_1}}{\alpha_{t_0}}\tilde{x}_{t_0} - \alpha_{t_1}(e^{-h_1}-1)\widehat{x_0}(\tilde{x}_{t_0}, t_0)$$

$xt-\Delta t$

條件引導

$$x_{t-\Delta t} \leftarrow x_{t-\Delta t} - s \cdot \nabla_{x_t} l$$

$$l = MSE(x_t, x_{adv})$$

實作設定

- Dataset：GTSRB
- Detection model : ResNet-50
- 攻擊參數：PGD($\ell\infty$: ε=8/255、$\ell2$: ε=1.0)、StAdv ε=0.05
- 淨化參數 : 淨化步數設為 24、目標時間 t*=0.6、引導係數為 1.5
- 評估：Standard / Robust accuracy、延遲觀測

實驗結果

表1：對抗強自適應攻擊的標準準確率和穩健準確率

Method	Standard Acc. (%)	Robustness Acc. (%)		
		$\ell\infty$	$\ell2$	StAdv
無防禦	93.81	0.47	31.29	0.43
JPEG Compression[12]	66.43	20.97	71.62	5.79
Feature Squeezing[13]	75.38	28.38	56.19	10.32
Median smoothing[13]	65.36	4.48	58.60	1.47
ours	**68.76**	**19.50**	**19.50**	**13.00**

淨化範例(對比圖：x_adv / 淨化後)

clean

PGD($\ell\infty$)

PGD($\ell2$)

StAdv

影像品質與淨化延遲

表2 : Image Quality Metrics: $IQ(x, x_{our})$， x表示對抗樣本，x_{our}表示經我們的淨化方法後的影像

Attack	PSNR	SSIM	LPIPS
PGD-$\ell\infty$	10.7345	0.4366	0.7132
PGD-$\ell2$	10.7005	0.5411	0.5598
StAdv	10.1532	0.3506	0.7467

表3 : Purification Latency ， Latency(ms)為淨化的延遲

Attack	Latency (ms)
PGD-$\ell\infty$	5987.808
PGD-$\ell2$	5963.320
StAdv	5900.977

結論

- 以 DPM-Solver++ 加速 GDMP 淨化 。
- 實驗證實結合 DPM-Solver++ 後，平均淨化延遲可控制在 **秒級 (約 5.9 秒)**，為未來實現自駕車即時防禦奠定了基礎 。

未來工作

- 更完整的評估(擴充到 adaptive-attack、更多的資料集評估) 。
- 衡量引導權重 s 與目標步數t^*之權衡，使準確率近一步的提高。

Reference

[1] 葉臻. (2025, January 26). 國道1號楊梅休息站電動車超載自撞起火已4死4傷. Https://Www.Cna.Com.Tw/News/Asoc/202501260024.Aspx.https://www.cna.com.tw/

[2] Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp) (pp. 39-57). Ieee.

[3] Chen, X., Li, Y., Zhang, H., Wang, J., & Liu, M. (2025). DiT-Air: Revisiting the efficiency of diffusion model architecture design in text to image generation. arXiv preprint arXiv:2503.10618.

[4] Duan, R., Ma, X., Wang, Y., Bailey, J., Qin, A. K., & Yang, Y. (2020). Adversarial camouflage: Hiding physical-world attacks with natural styles. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1000-1008).

[5] Dziugaite, G. K., Ghahramani, Z., & Roy, D. M. (2016). A study of the effect of JPG compression on adversarial images. arXiv preprint arXiv:1608.00853.

[6] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1625-1634).

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

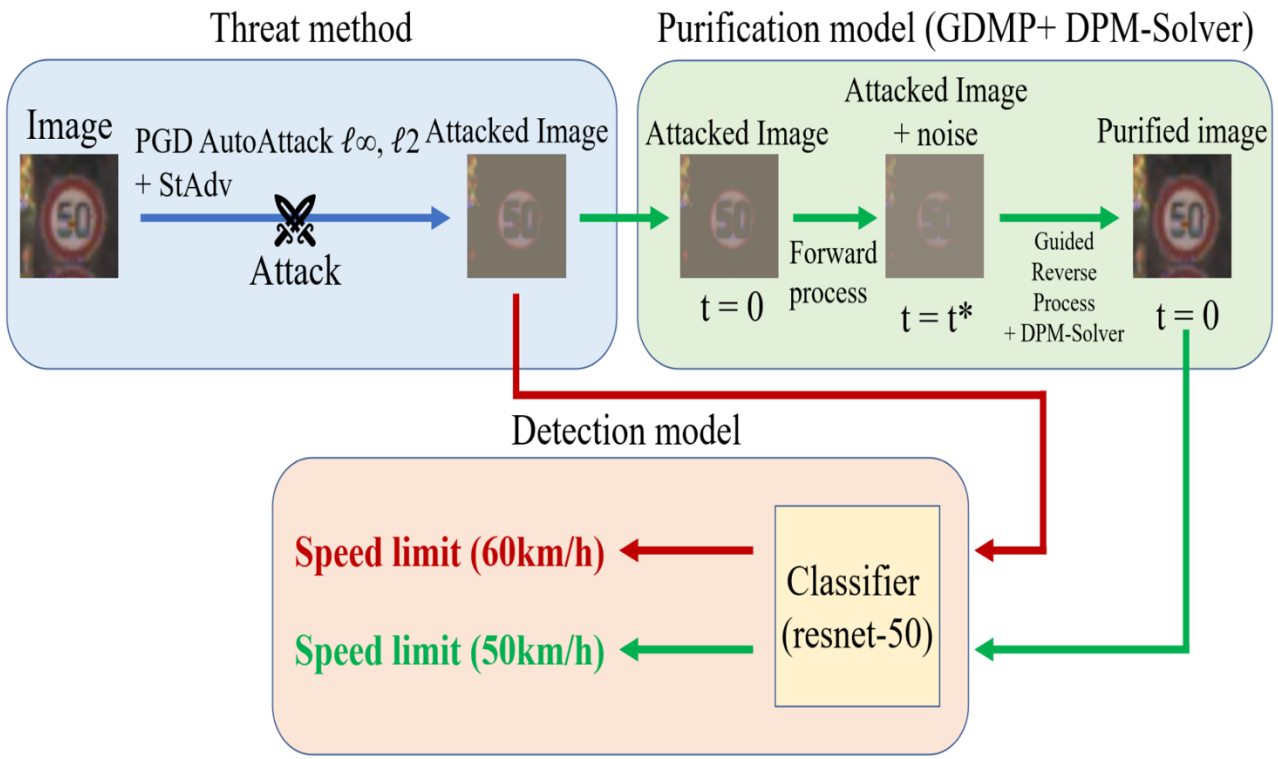
[8] Hefei Mei, Minjing Dong, Chang Xu. (2025). Efficient Image-to-Image Diffusion Classifier for Adversarial Robustness. arXiv preprint arXiv:2408.08502

[9] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851.
[10] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. (2015). EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES . arXiv preprint arXiv: 1412.6572
[11] Lowery , L. (2025, March 12). Researchers Find Low-Cost Malicious Attacks Can Affect AV Operations, Most U.S. Drivers Won’ t Ride Driverless. Https://Www.Repairerdrivennews.Com/2025/03/12/Researchers-Find-Low-Cost-Malicious-Attacks-Can-Affect-Av-Operations-Most-u-s-Drivers-Wont-Ride-Driverless/?Ut. https://www.repairerdrivennews.com/
[12] Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., & Zhu, J. (2022). DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. arXiv:2211.01095.
[13] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.
[14] Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., & Anandkumar, A. (2022). Diffusion models for adversarial purification. arXiv preprint arXiv:2205.07460.
[15] Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4195-4205).
[16] Song, D., Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., ... & Kohno, T. (2018). Physical adversarial examples for object detectors. In 12th USENIX workshop on offensive technologies (WOOT 18).
[17] Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.
[18] Wang, J., Lyu, Z., Lin, D., Dai, B., & Fu, H. (2022). Guided diffusion model for adversarial purification. arXiv preprint arXiv:2205.14969.

[19] Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., & Song,D. (2018). Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612.
[20] Xu, W., Evans, D., & Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155.
[21] Xue, M., Yuan, C., He, C., Wang, J., & Liu, W. (2021). NaturalAE: Natural and robust physical adversarial examples for object detectors. Journal of Information Security and Applications, 57, 102694.

Thanks for listening.

Q & A



Method	Standard Acc. (%)	Robustness Acc. (%)		
		ℓ∞	ℓ2	StAdv
無防禦	93.81	0.47	31.29	0.43
JPEG Compression	66.43	20.97	71.62	5.79
Feature Squeezingn	75.38	28.38	56.19	10.32
Median smoothing	65.36	4.48	58.60	1.47
ours	68.76	19.5	19.5	13