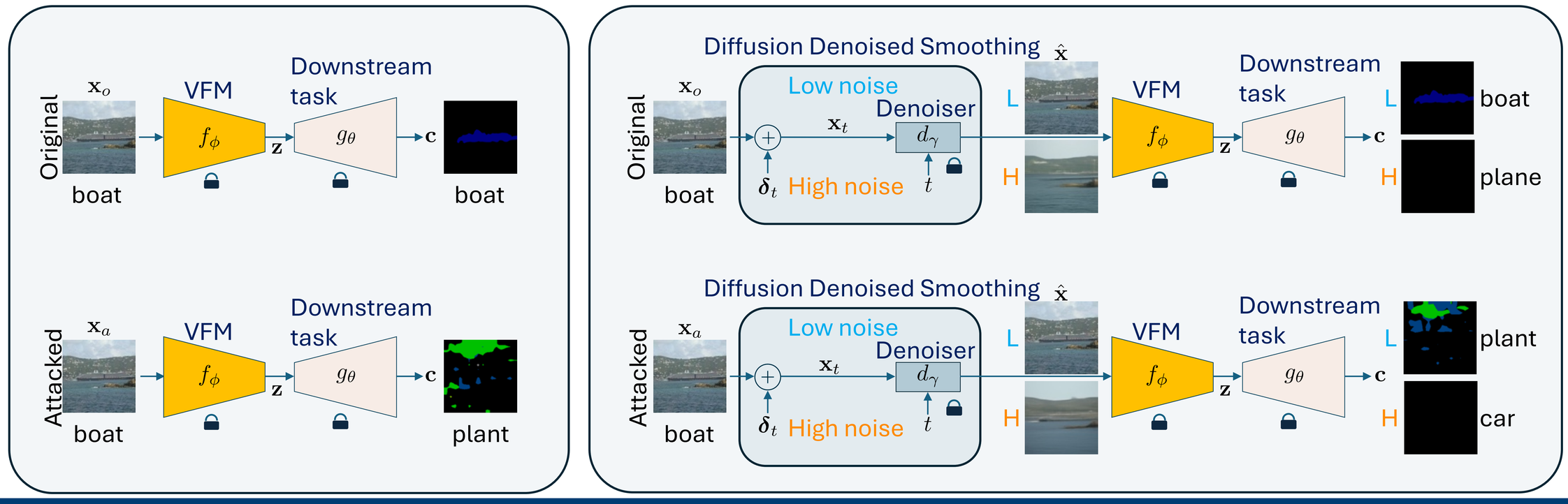
Beyond Classification: Evaluating Diffusion Denoised Smoothing for Security-Utility Trade off

Yury Belousov, Brian Pulfer, Vitaliy Kinakh, Slava Voloshynovskiy
University of Geneva, Switzerland



Vision Foundation Models (VFMs) show impressive performance but remain *highly vulnerable* to adversarial attacks. **Diffusion Denoised Smoothing** [1] emerges as a promising defense by preprocessing inputs with diffusion models. We evaluate beyond prior classification-only work and show both noise regimes have limitations:

- **Low noise** denoising preserves clean **performance** across all tasks but **fails** against complex adversarial attacks
- **High noise** denoising provides strong **protection** against attacks but significantly **degrades** clean performance by 14-57%

$\Longrightarrow$ We introduce a **novel attack** exploiting the diffusion process itself.

https://arxiv.org/abs/2505.15594

UNIVERSITÉ DE GENÈVE

Eusipco2025
SIGNAL PROCESSING IN THE LAND OF ART, CULTURE AND BEAUTY

Experimental Setup

Model: DINOv2 (ViT-S/B/L)
Tasks: Classification, Segmentation, Depth, Retrieval
Datasets: PascalVOC, NYU-Depth, rOxford
Attacks: PGD [2], MI-FGSM [3], SIA [4]

Key Findings

- **Robustness:** VFMs robust to image distortions but fail under adv. attacks
- **Low noise:** Preserves performance but vulnerable to complex attacks
- **High noise:** Strong protection, severe penalty (14-57% degradation)
- **Novel attack (*):** Targeting diffusion defenses by integrating diffusion model into the attack loop, reducing protection at all noise levels

Robustness and Security Results

Transform	PascalVOC		NYU	rOxford
	Classif.↑	Segment.↑	Depth↓	mAP↑
None (original images)	*94.9*	*79.3*	*0.6*	*84.5*
blur	94.5	76.1	0.7	84.9
jpeg	94.6	78.6	0.6	84.6
low noise diffusion	95.5	79.0	0.6	83.8
high noise diffusion	80.6	61.3	0.8	62.0
PGD Attack	0.0	10.5	6.4	0.5
PGD + low noise	46.3	54.6	1.4	81.7
PGD* w/ diffusion	0.0	11.7	6.2	0.5
PGD* w/ diffusion + low noise	0.0	12.4	6.1	0.5
SIA Attack	0.4	21.3	4.1	0.6
SIA + low noise	2.2	26.0	3.7	15.4
SIA + high noise	79.4	61.5	0.8	58.2
SIA* w/ diffusion + high noise	66.6	53.3	1.0	57.1

References

[1] Nicholas Carlini et al. "(Certified!!) Adversarial Robustness for Free!" In: *International Conference on Learning Representations (ICLR)* (2023).
[2] Aleksander Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018.
[3] Yinpeng Dong et al. "Boosting adversarial attacks with momentum". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
[4] 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*. 2023.