Neutralizing Proactive Defense using Diffusion-based Upsampling

Giuseppe Daidone, F. Bartolucci, M. R. Briglia, M. H. Mirza, G. Lisanti, I. Masi



Giuseppe Daidone
Computer Science Department
OmnAI Lab, Sapienza, University of Rome, Italy
https://omnai.di.uniroma1.it

Neutralizing Proactive Defense using Diffusion-based Upsampling

Introduction

The Evolution of Generative Media



GAN



Stable Diffusion



Midjourney



Flux



Nano Banana

The Evolution of Generative Media



GAN



Stable Diffusion



Midjourney



Flux



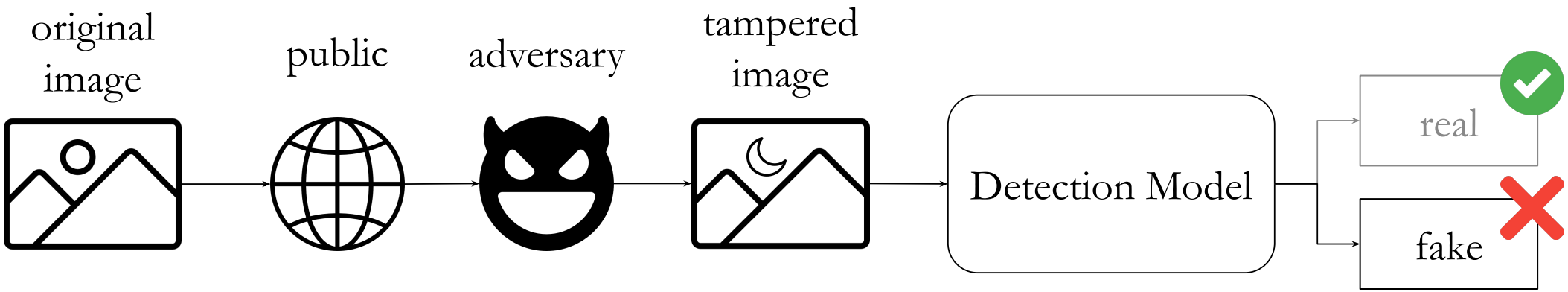
Nano Banana

- Passive detection is **no longer sufficient**.
- The paradigm is shifting towards **proactive defense mechanisms** (e.g., watermarks, immunizations).

Neutralizing Proactive Defense using Diffusion-based Upsampling

Proactive Defense Frameworks

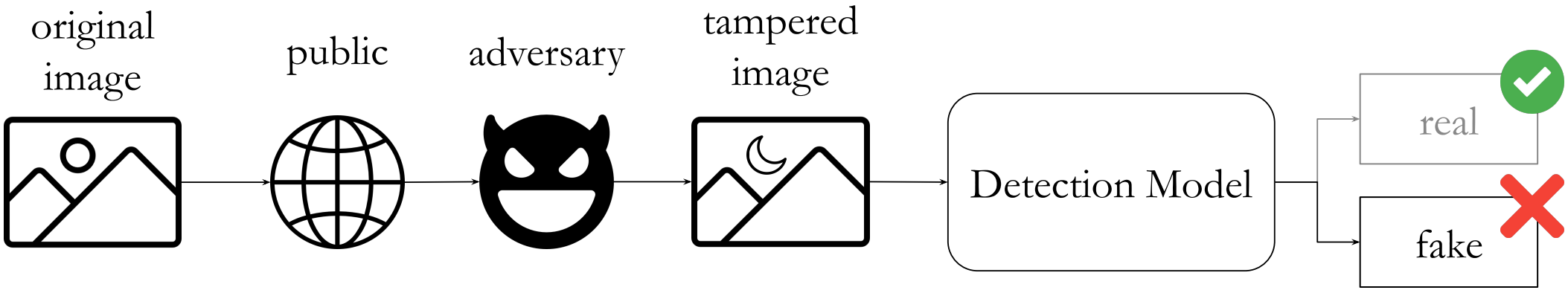
Passive Detection vs. Proactive Defense



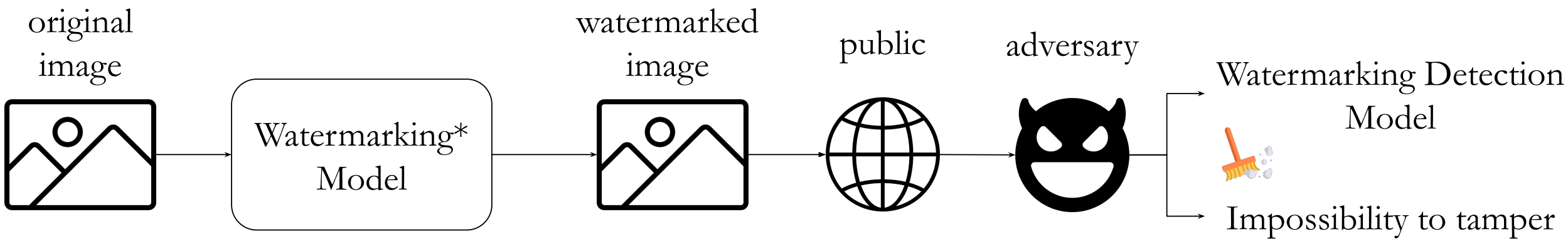
Passive Detection Pipeline

Proactive Defense Pipeline

Passive Detection vs. Proactive Defense

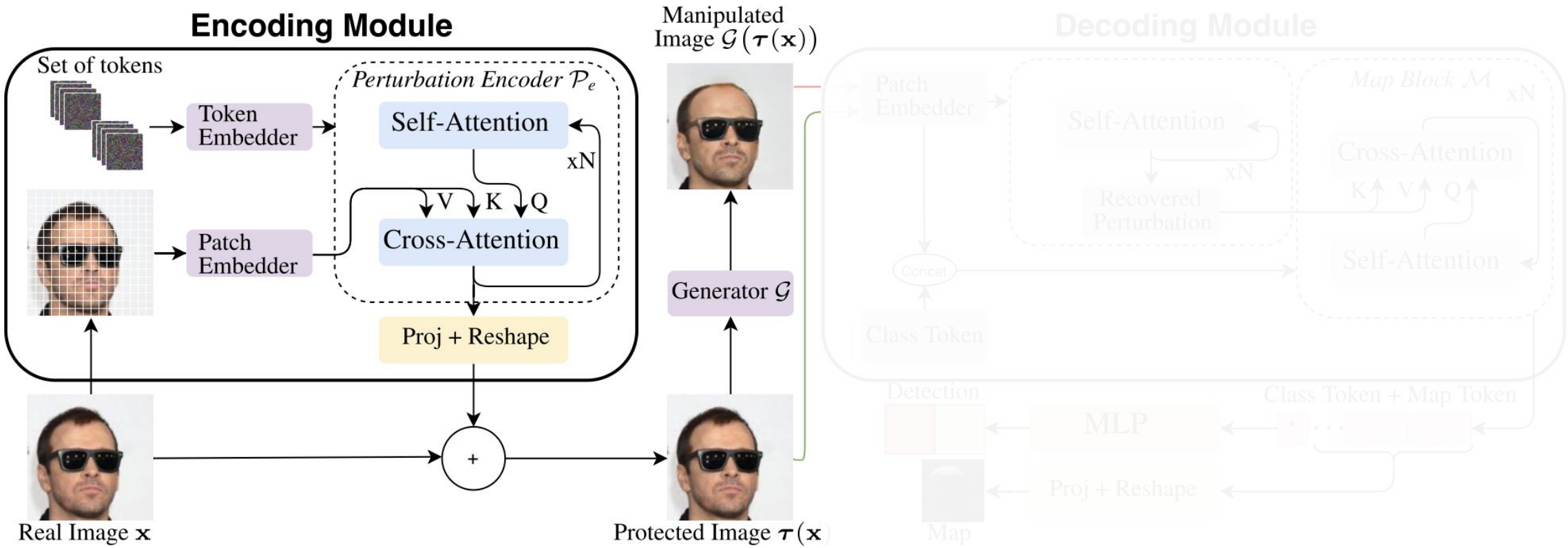


Passive Detection Pipeline



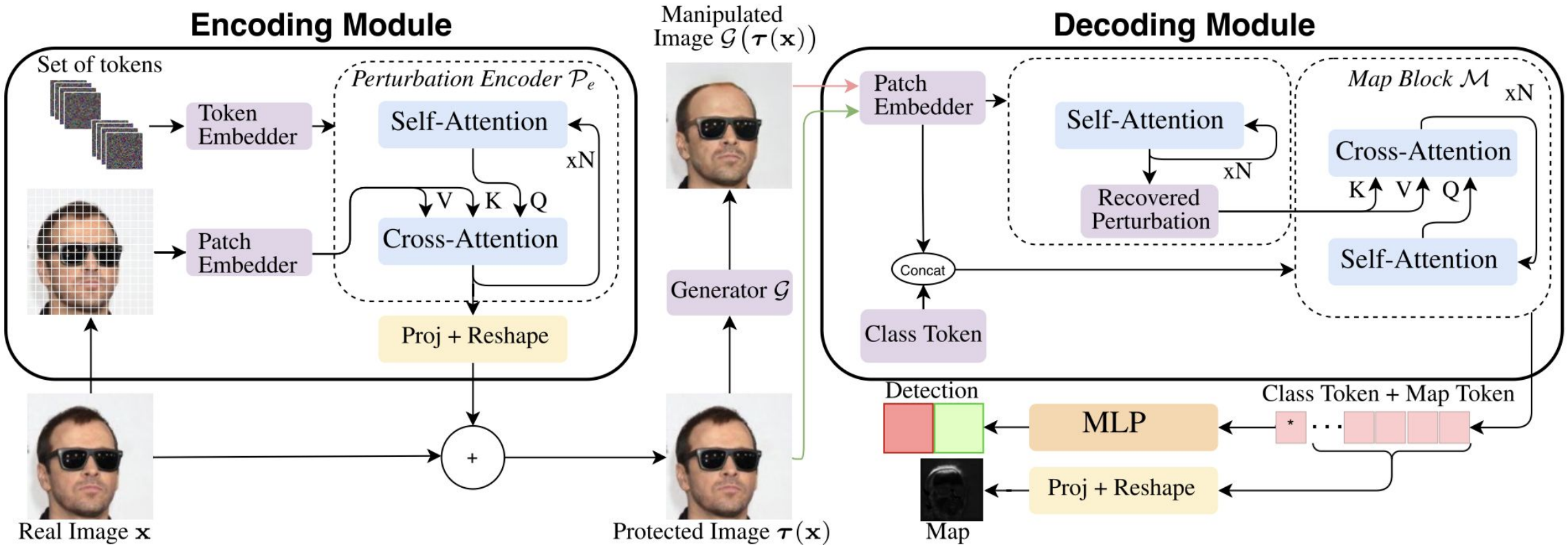
Proactive Defense Pipeline

PADL



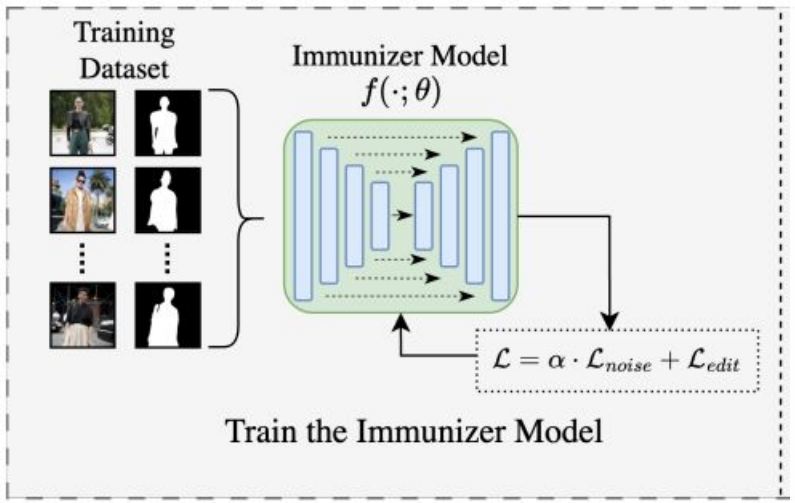
F. Bartolucci, et al. Perturb, Attend, Detect, and Localize (PADL): Robust Proactive Image Defense. *IEEE Access* 2025.

PADL



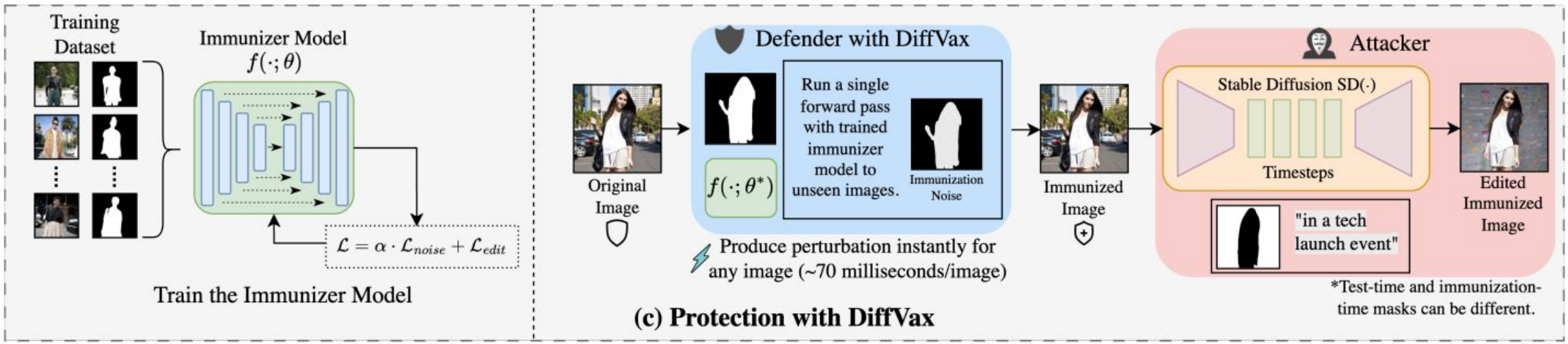
F. Bartolucci, et al. Perturb, Attend, Detect, and Localize (PADL): Robust Proactive Image Defense. *IEEE Access* 2025.

DiffVax



T. C. Ozden, et al. DiffVax: Optimization-Free Image Immunization Against Diffusion-Based Editing. In *ICLR 2026*.

DiffVax



T. C. Ozden, et al. DiffVax: Optimization-Free Image Immunization Against Diffusion-Based Editing. In *ICLR 2026*.

Neutralizing Proactive Defense using Diffusion-based Upsampling

Method

Research Question

- We have been inspired by the work of R. Hönig et al., Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI from ICLR 2025.
- The authors applied **Gaussian noise + Stable Diffusion Upscaling** to reconstruct and purify the images against anti-style mimicking watermarking (e.g., Glaze).

Research Question

- We have been inspired by the work of R. Hönig et al., Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI from ICLR 2025.
- The authors applied **Gaussian noise + Stable Diffusion Upscaling** to reconstruct and purify the images against anti-style mimicking watermarking (e.g., Glaze).

Question: Are reconstruction attacks also effective against proactive defenses for deepfake/generative AI?

Threat Model

Assumption: The adversary has access to the protected images but no knowledge of the specific defense algorithm or access to unprotected version (black-box + model-agnostic).

Threat Model

Assumption: The adversary has access to the protected images but no knowledge of the specific defense algorithm or access to unprotected version (black-box + model-agnostic).

Objective: Neutralize the proactive defense while maintaining high visual fidelity.

- PADL: Flip prediction from *protected* to *unprotected*.
- DiffVax: Output *meaningful* edited images.

Threat Model

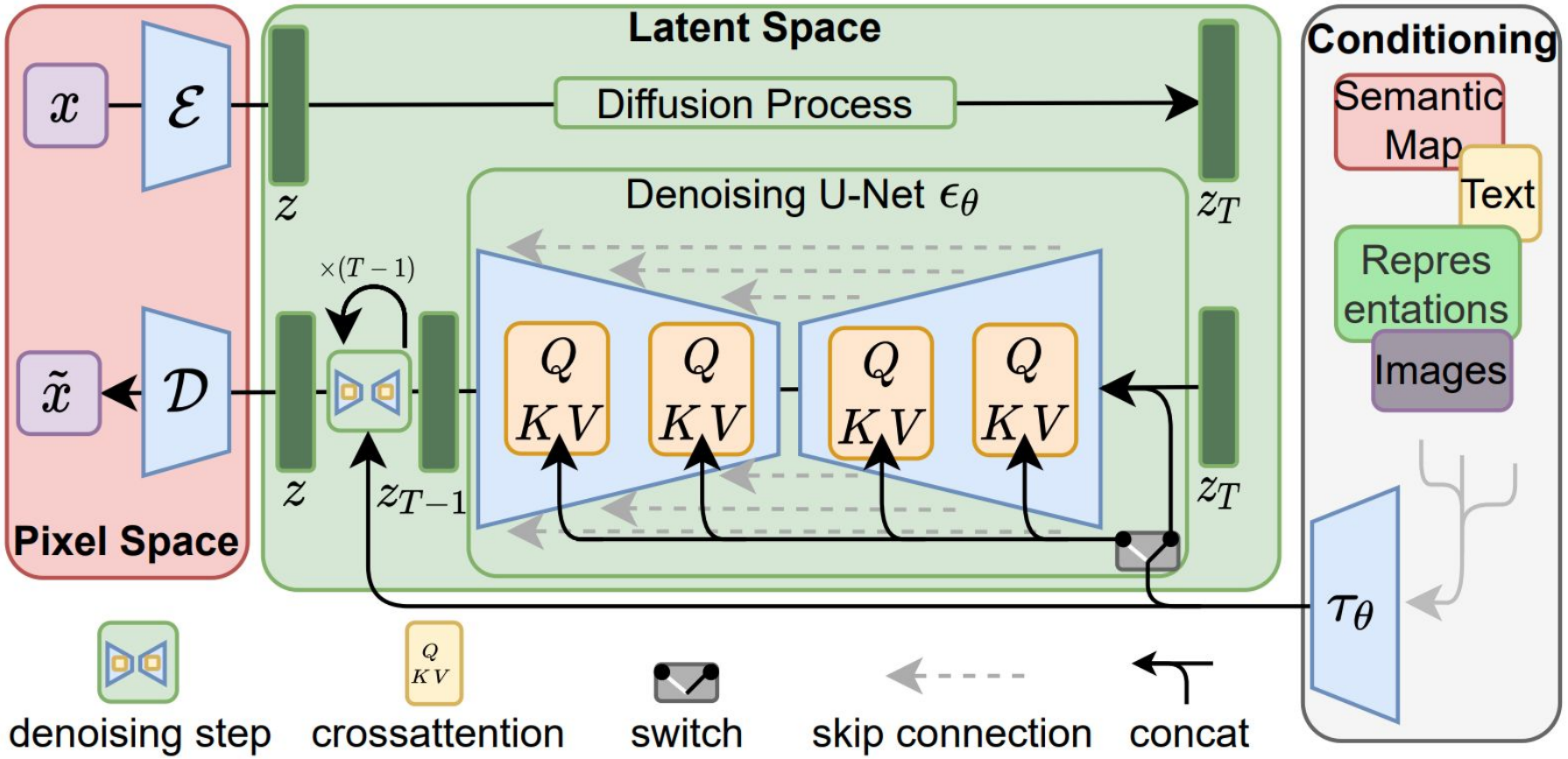
Assumption: The adversary has access to the protected images but no knowledge of the specific defense algorithm or access to unprotected version (black-box + model-agnostic).

Objective: Neutralize the proactive defense while maintaining high visual fidelity.

- PADL: Flip prediction from *protected* to *unprotected*.
- DiffVax: Output *meaningful* edited images.

Computational Capability: The adversary is capable of applying any image transformation and has access to Stable Diffusion Upscaler API.

Stable Diffusion X4 Upscaling



R. Rombach, et al. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR 2022*.

Adversarial Setup



Classical

Interpolation (Bicubic), Salt & Pepper noise, and Shear transformations.

Adversarial Setup



Classical

Interpolation (Bicubic), Salt & Pepper noise, and Shear transformations.



Pure Upscaling

Standard Diffusion-based Upscaling using pre-trained models.

Adversarial Setup



Classical

Interpolation (Bicubic), Salt & Pepper noise, and Shear transformations.



Pure Upscaling

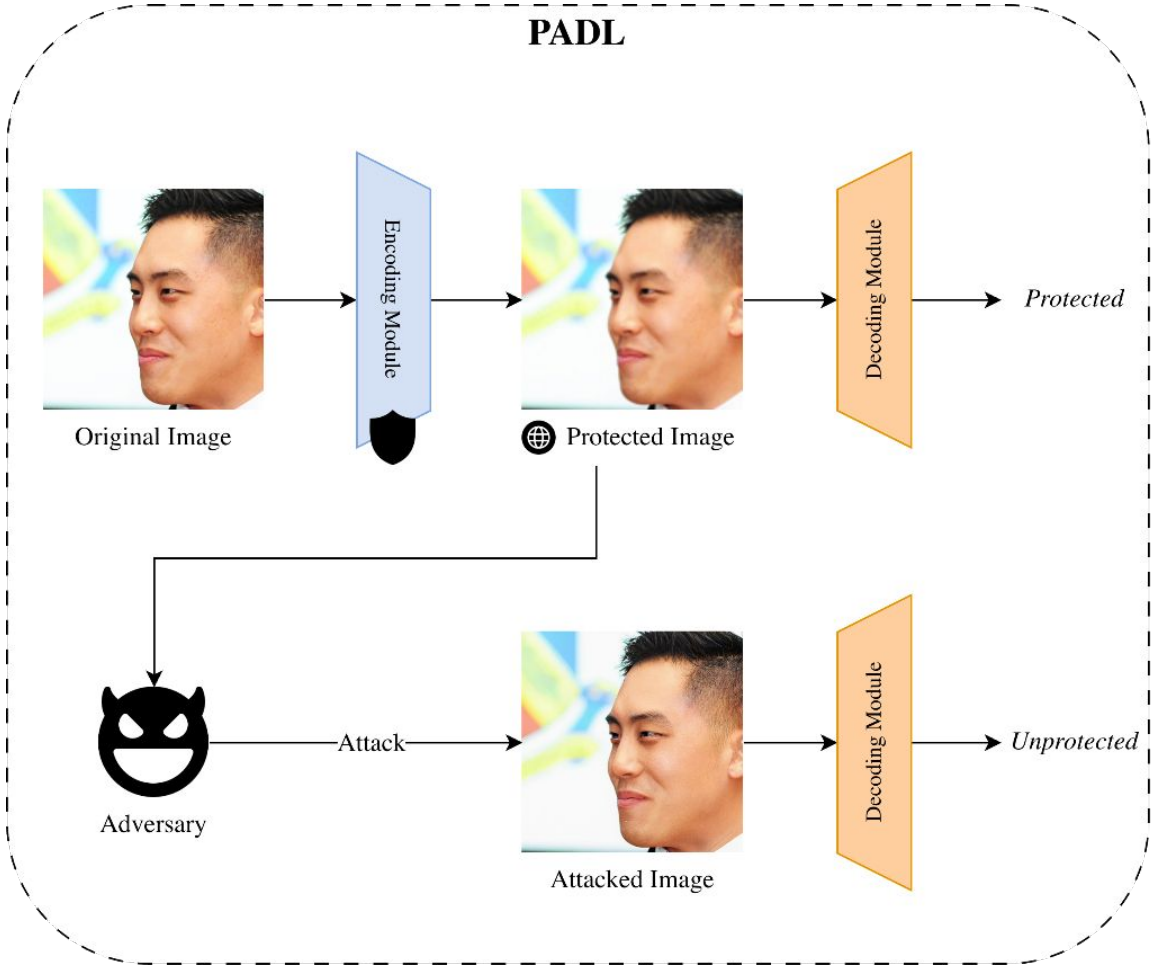
Standard Diffusion-based Upscaling using pre-trained models.



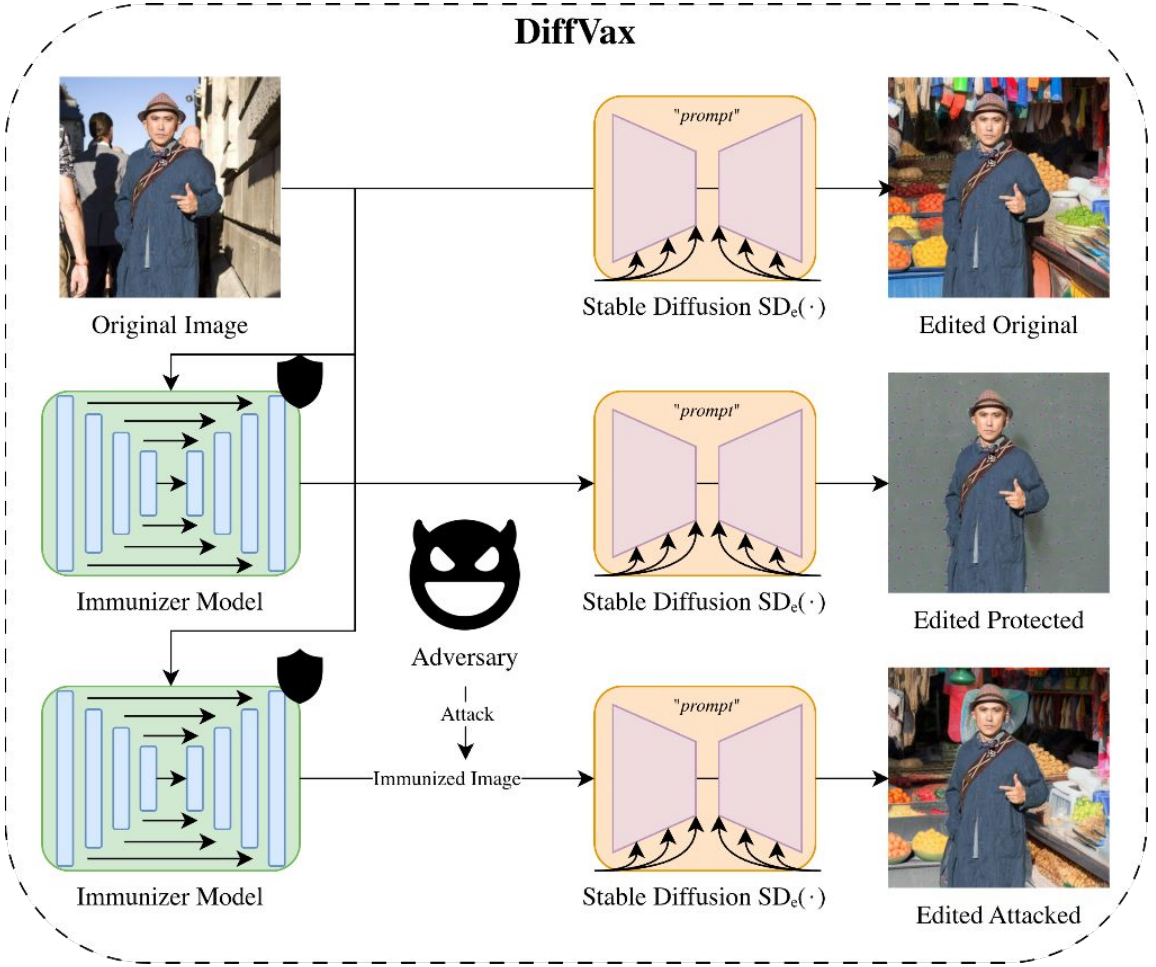
Reconstruction

Compressed and Noisy Upsampling variants using compression pre-SD.

Evaluation Setup



Evaluation Setup



Neutralizing Proactive Defense using Diffusion-based Upsampling

Experiments

Experimental Setup

- **Classical Image Transformations:**

$$\boldsymbol{x} \xrightarrow{\text{PADL enc. } \mathcal{E}_\theta(\boldsymbol{x})} \tilde{\boldsymbol{x}} \xrightarrow{\text{Transformation } \mathcal{T}} \boldsymbol{x}' \xrightarrow{\text{PADL decoder}} \mathcal{D}_\theta(\boldsymbol{x}')$$

$$\boldsymbol{x} \xrightarrow{\text{DiffVax imm. } \mathcal{I}_\theta(\boldsymbol{x}', \mathcal{M})} \tilde{\boldsymbol{x}} \xrightarrow{\text{Transformation } \mathcal{T}} \boldsymbol{x}' \xrightarrow{\text{SD}_e\left(\boldsymbol{x}', \tilde{\mathcal{M}}, \mathcal{P}\right)} \boldsymbol{x}'_{\text{edit}}$$

- **Diffusion-based Upsampling:**

$$\boldsymbol{x} \xrightarrow{\text{PADL encoder } \mathcal{E}_\theta(\boldsymbol{x})} \tilde{\boldsymbol{x}} \xrightarrow{\mathcal{P}} \boldsymbol{x}' \xrightarrow{\text{SD } \times 4 \text{ upscaler}} \boldsymbol{x}'_{\uparrow} \xrightarrow{\text{PADL decoder}} \mathcal{D}_\theta(\boldsymbol{x}'_{\uparrow})$$

$$\boldsymbol{x} \xrightarrow{\text{DiffVax imm. } \mathcal{I}_\theta(\boldsymbol{x}, \mathcal{M})} \tilde{\boldsymbol{x}} \xrightarrow{\mathcal{P}} \boldsymbol{x}' \xrightarrow{\text{SD } \times 4 \text{ upscaler}} \boldsymbol{x}'_{\uparrow} \xrightarrow{\text{SD}_e\left(\boldsymbol{x}'_{\uparrow}, \tilde{\mathcal{M}}, \mathcal{P}\right)} \boldsymbol{x}'_{\uparrow,\text{edit}}$$

Measuring the Robustness

PADL

- **Robustness**:
  - Attack Success Rate (ASR)
  - Detection Logits Distribution
- **Baseline**:
  - False Negative Rate (FNR)
- **Perceptual Similarity**:
  - Learned Perceptual Image Patch Similarity (LPIPS)
  - Mean Squared Error (MSE)
  - Peak Signal-to-Noise Ratio (PSNR)
  - Structural Similarity Index Measure (SSIM)

Measuring the Robustness

PADL

- **Robustness**:
  - Attack Success Rate (ASR)
  - Detection Logits Distribution
- **Baseline**:
  - False Negative Rate (FNR)
- **Perceptual Similarity**:
  - Learned Perceptual Image Patch Similarity (LPIPS)
  - Mean Squared Error (MSE)
  - Peak Signal-to-Noise Ratio (PSNR)
  - Structural Similarity Index Measure (SSIM)

DiffVax

- **Robustness and Perceptual Similarity**:
  - Structural Similarity Index Measure (SSIM)
  - Peak Signal-to-Noise Ratio (PSNR)
  - Feature Similarity Index Method (FSIM)
  - CLIP-T Score
- **Baseline**:
  - Immunized images

Neutralizing PADL Protection

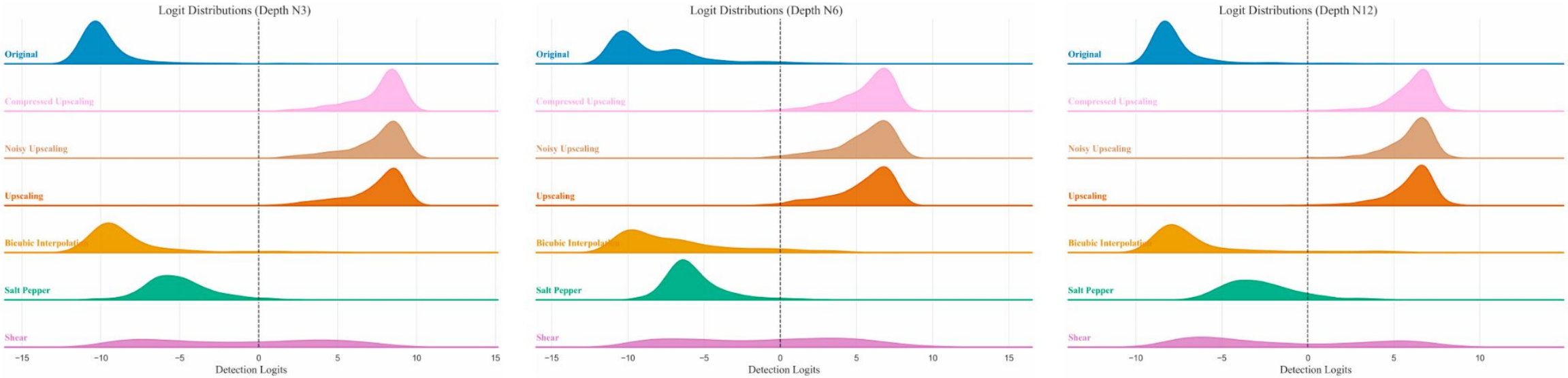
Original	PADL	Compressed Upsc. (Ours)	Noisy Upsc. (Ours)	Upscaling (Ours)
				

Neutralizing PADL Protection

	Attack	ASR ↑	LPIPS ↓	MSE ↓	PSNR ↑	SSIM ↑
Ours	Compressed Upscaling	**0.9950**	0.0648	0.0023	27.1354	0.8302
	Noisy Upscaling	0.9922	0.0615	**0.0020**	27.8002	0.8516
	Upscaling	0.9894	**0.0612**	**0.0020**	**27.8920**	**0.8550**
Classic	Bicubic	0.0689	<u>0.0006</u>	<u>0.0000</u>	<u>52.2473</u>	<u>0.9994</u>
	Salt & Pepper	0.0533	0.5073	0.0163	17.8924	0.4795
	Shear	<u>0.4317</u>	0.1158	0.0255	16.4496	0.3733

Results averaged across datasets and protection layers. The arrows ↑ and ↓ indicate the direction toward a better metric. ***Bold*** *and* <u>*underlined*</u> *values indicate the best similarity score for diffusion-based attacks and classical perturbation, respectively, where a high similarity score means the adversary does not get a corrupted image.*

Neutralizing PADL Protection



Neutralizing DiffVax Immunization

Original	Edited Original	Edited Protected	Compressed Upsc. (Ours)	Noisy Upsc. (Ours)	Upscaling (Ours)
					

Neutralizing DiffVax Immunization

	Attack	SSIM ↑	PSNR ↑	FSIM ↑	CLIP-T ↑
	DiffVax (baseline)	0.5218	14.2432	0.7134	17.9231
Ours	Compressed Upscaling	0.4933	14.8247	0.7220	22.0568
Ours	Noisy Upscaling	0.4940	**15.0524**	0.7290	22.4584
Ours	Upscaling	0.4896	14.7792	**0.7295**	21.6338
Classic	Bicubic	**0.5345**	13.9038	0.7067	21.6118
Classic	Salt & Pepper	0.2973	14.1111	0.6756	**25.8850**
Classic	Shear	0.2096	10.4120	0.5117	23.0509

The arrows ↑ and ↓ indicate the direction toward a better metric. ***Bold*** *values indicate the highest similarity scores, where higher similarity indicates less effective protection by DiffVax.*

Neutralizing Proactive Defense using Diffusion-based Upsampling

Conclusions

Conclusions

- State-of-the-art proactive defenses (PADL and DiffVax) can be effectively neutralized using Stable Diffusion Upsampling.
- The attack acts as a manifold projection, removing protective signatures while preserving exceptional visual quality.
- Preliminary experiments revealed that integrating reconstruction attacks into PADL adversarial training pipeline failed to increase robustness.

Takeaways

Future proactive defense mechanisms must evolve beyond robustness against basic image transformations and consider generative reconstruction attacks in their threat model.

Takeaways

Future proactive defense mechanisms must evolve beyond robustness against basic image transformations and consider generative reconstruction attacks in their threat model.

Media authenticity and provenance should be built on three pillars†:

- Cryptographic signatures in metadata;
- Robust perceptual fingerprinting;
- Watermarking in media content.

†J. Collomosse et al. To Authenticity, and Beyond! Building Safe and Fair Generative AI Upon the Three Pillars of Provenance. In *IEEE Computer Graphics and Applications*, 2024.

Question Time

Thank you for the attention!

Contacts:
Email: giuseppe.daidone@uniroma1.it
Linkedin: giuseppedaidone
Website: https://giuseppedaidone.github.io/

This work has been supported by the **Italian National Cybersecurity Agency (ACN)**