Fake Image Detectors & Adversarial Attacks

Trends & Applications of Computer Vision

2024/2025

Martina D'Angelo, Thomas Lazzerini, Samuele Cappelletti

Introduction: Fake Images

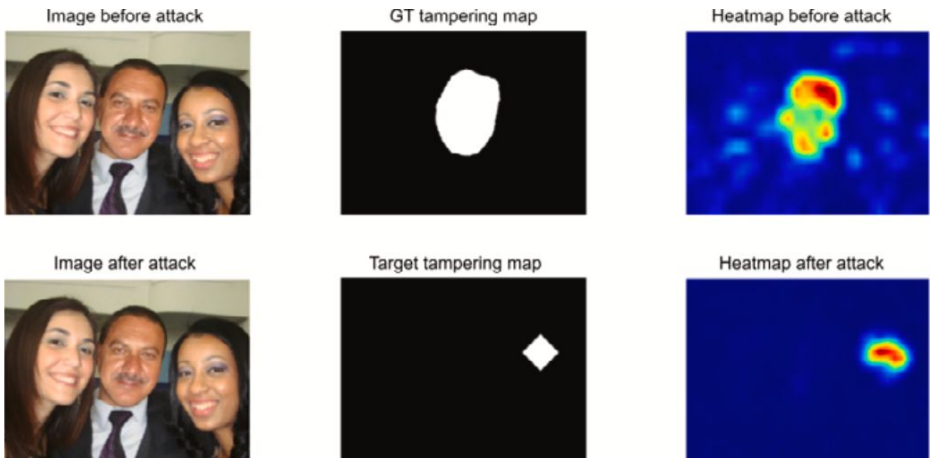
- Synthetic images are now flooding the real world
- From online dating sites to social media, fake profiles and scams are everywhere
- Some images are funny and harmless
- While others could be exploited by malicious users

Andrew Walz, was a congressional candidate running for office in Rhode Island. He called himself “a proven business leader” with the slogan “Let’s make change in Washington together.”

Carlini, N., & Farid, H. (2020). Evading deepfake-image detectors with white-and black-box attacks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 658-659).

Introduction: Fake Images

- Continuous fight between:
  - **detectors** that try to distinguish fake images from real ones
  - **attacks** that try to trick the detector by manipulating the images

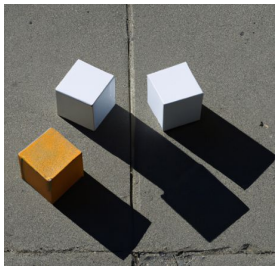


Boato, G., De Natale, F. G., De Stefano, G., Pasquini, C., & Roli, F. (2024). Adversarial mimicry attacks against image splicing forensics: An approach for jointly hiding manipulations and creating false detections. Pattern Recognition Letters, 179, 73-79.

Introduction: How to Detect Fake Images

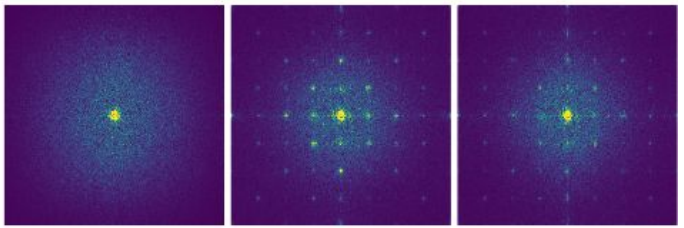
- The generation process introduces in the images some artifacts that we can exploit

High-level forensic techniques



- semantic features
- physical inconsistencies
- mostly visible

Low-level forensic techniques

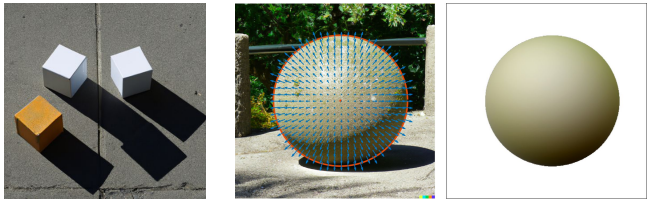


- pixel-level artifacts
- not apparently visible
- struggle to generalize
- sensitive to laundering

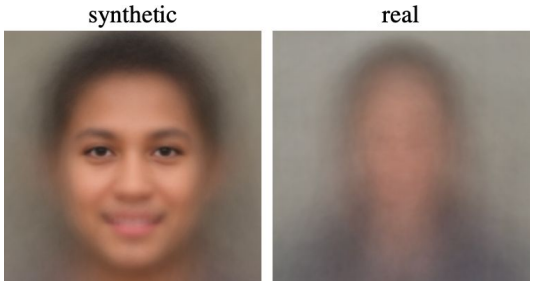
Farid, H. (2022). Perspective (in) consistency of paint by text. arXiv preprint arXiv:2206.14617.
Mandelli, S., Bestagini, P., & Tubaro, S. (2024). When Synthetic Traces Hide Real Content: Analysis of Stable Diffusion Image Laundering. arXiv preprint arXiv:2407.10736.

High-level Forensic Techniques

Shadows, lighting and reflections inconsistencies



Repeated & Uniform patterns



Farid, H. (2022). Lighting (in) consistency of paint by text. arXiv preprint arXiv:2207.13744.
Farid, H. (2022). Perspective (in) consistency of paint by text. arXiv preprint arXiv:2206.14617.
Mundra, S., Porcile, G. J. A., Marvaniya, S., Verbus, J. R., & Farid, H. (2023). Exposing gan-generated profile photos from compact embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 884-892).

Fake Image Detectors

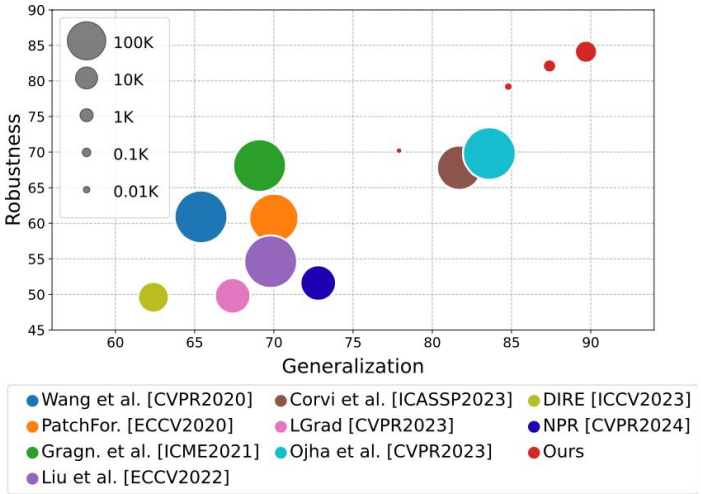
Raising the Bar of AI-generated Image Detection with CLIP

Davide Cozzolino[1] Giovanni Poggi[1] Riccardo Corvi[1] Matthias Nießner[2] Luisa Verdoliva[1,2]

[1]University Federico II of Naples [2]Technical University of Munich

Image Detection with CLIP: Introduction

- The method:
  - based on CLIP features
  - good generalization ability and robustness with just a few examples



- The performance:
  - similar as SoTA methods on in-distribution data
  - significant improvement on out-of-distribution data

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., & Verdoliva, L. (2024). Raising the Bar of AI-generated Image Detection with CLIP. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4356-4366).

Image Detection with CLIP: Introduction

- SoTA methods:
  - good in detecting images generated by same family generator
  - hard to generalize to new and unseen generative methods

- CLIP-based method:
  - good in detecting any type of generated image (with and without augmentations)
  - CLIP features achieve excellent generalization

Method
Wang et al.
PatchFor.
Grag. et al.
Mand. et al.
Liu et al.
Corvi et al.
LGrad
Ojha et al.
DIRE-1
DIRE-2
NPR

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., & Verdoliva, L. (2024). Raising the Bar of AI-generated Image Detection with CLIP. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4356-4366).

Image Detection with CLIP: Method

- Collect N real images
  - $\{R_1,...,R_N\}$
- Collect corresponding N real images captions
  - $\{t_1,...,t_N\}$
- Generate N images using the captions
  - $\{F_1,...,F_N\}$, $F_i = G(t_i)$
- Extract feature vectors using CLIP
  - $\{\mathbf{r}_1,...,\mathbf{r}_N\}$ and $\{\mathbf{f}_1,...,\mathbf{f}_N\}$, $\mathbf{r}_i = \text{CLIP}(R_i)$ and $\mathbf{f}_i = \text{CLIP}(F_i)$
- Feed feature vectors to the linear SVM classifier

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., & Verdoliva, L. (2024). Raising the Bar of AI-generated Image Detection with CLIP. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4356-4366).

Image Detection with CLIP: Results (AUC)

No post-processing

Method	GAN family					Diffusion family									Commercial tools				AVG
	Pro GAN	Style GAN2	Style GAN3	Style GANT	Giga GAN	Score SDE	ADM	GLIDE	Latent Diff.	Stable Diff.	DeepFl. IF	Ediff-I	DiT	SDXL	DALL·E2	DALL·E3	Midj.	Adobe Firef.	
Wang et al.	100.	96.5	98.5	98.9	66.6	32.9	64.3	48.5	59.2	41.5	78.0	64.9	58.6	54.3	64.8	10.9	40.2	84.8	64.6
PatchFor.	92.3	84.5	91.8	91.2	64.7	83.3	74.8	96.2	78.1	62.4	62.7	78.7	83.1	68.4	41.9	52.7	57.8	49.4	73.0
Grag. et al.	100.	**99.8**	97.5	98.8	82.8	92.1	74.7	62.8	91.9	52.5	69.9	69.6	65.3	58.0	58.3	2.4	43.1	63.5	71.3
Mand. et al.	96.2	93.8	**100.**	92.6	61.8	99.8	56.5	40.5	70.0	36.8	47.2	65.0	59.1	27.0	14.5	14.7	24.3	36.7	57.6
Liu et al.	100.	**99.8**	98.4	98.5	**98.2**	**95.4**	82.5	76.5	**97.6**	77.4	72.2	**98.7**	88.0	31.1	70.4	0.2	40.7	11.8	74.3
Corvi et al.	79.4	73.7	50.0	97.1	63.4	65.0	80.7	91.9	100.	**100.**	**99.9**	85.7	**100.**	**100.**	69.4	60.8	**100.**	**98.0**	84.2
LGrad	100.	91.2	83.8	81.8	82.2	80.6	76.9	66.1	81.1	61.5	68.8	74.1	56.2	57.2	58.6	37.9	56.3	40.6	69.7
Ojha et al.	100.	93.9	92.3	98.2	96.0	58.4	**86.7**	80.8	85.7	89.5	92.9	80.6	77.8	85.1	**95.2**	36.4	66.2	**97.5**	84.1
DIRE-1	50.6	56.9	47.8	**99.9**	74.1	44.3	75.7	71.4	68.7	39.4	**98.9**	**99.1**	**99.6**	47.1	44.7	47.6	51.0	57.4	65.2
DIRE-2	54.2	52.5	43.0	**99.6**	76.0	41.0	70.1	70.1	69.3	46.9	97.0	98.2	98.3	42.8	41.0	49.6	47.8	43.0	63.3
NPR	100.	85.6	77.0	96.4	88.7	91.1	**86.3**	79.3	90.2	64.5	91.6	80.1	78.4	76.7	39.5	48.7	77.0	32.1	76.8
Ours 1k	**98.9**	90.5	85.5	**100.**	81.3	89.1	81.1	**99.9**	94.1	87.6	96.5	**98.5**	94.1	87.8	89.0	70.0	73.0	74.4	88.4
Ours 1k+	91.4	80.9	84.0	**99.8**	74.7	84.3	75.2	**99.6**	81.6	89.8	98.0	**99.1**	92.5	88.9	83.6	**93.6**	78.7	85.1	87.8
Ours 10k	**99.8**	91.8	86.8	**100.**	83.6	89.0	81.4	**99.9**	94.2	90.7	97.0	**98.7**	95.0	87.4	89.2	77.6	75.3	80.1	**89.8**
Ours 10k+	93.4	87.1	87.6	**99.9**	78.5	89.2	79.9	**99.7**	84.7	91.3	97.9	**99.4**	94.0	90.1	86.3	**92.9**	81.7	87.2	**90.0**

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., & Verdoliva, L. (2024). Raising the Bar of AI-generated Image Detection with CLIP. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4356-4366).

Image Detection with CLIP: Results (AUC)

With post-processing

Method	GAN family					Diffusion family									Commercial tools				AVG
	Pro GAN	Style GAN2	Style GAN3	Style GANT	Giga GAN	Score SDE	ADM	GLIDE	Latent Diff.	Stable Diff.	DeepFl. IF	Ediff-I	DiT	SDXL	DALL·E 2	DALL·E 3	Midj.	Adobe Firef.	
Wang et al.	100.	86.6	88.4	61.7	59.2	68.0	65.0	60.6	67.1	55.2	50.3	48.3	55.1	64.5	46.2	27.7	46.7	55.9	61.5
PatchFor.	57.9	51.8	57.0	50.6	53.8	69.0	66.2	83.3	58.8	48.3	61.3	65.0	68.1	63.3	64.3	63.3	59.0	65.1	61.4
Grag. et al.	100.	**95.4**	**90.9**	94.4	64.4	77.1	77.1	79.8	**84.0**	53.5	50.6	55.6	66.7	66.6	55.2	25.1	48.5	60.2	69.2
Mand. et al.	81.1	79.3	87.2	49.1	49.3	64.0	54.8	42.6	52.9	39.4	55.7	54.5	49.8	42.2	47.9	42.3	35.2	53.4	54.5
Liu et al.	64.3	55.1	50.1	57.3	45.4	62.6	51.1	58.6	50.7	58.6	50.9	64.2	53.9	56.0	44.4	61.8	52.6	53.1	55.0
Corvi et al.	77.5	74.7	69.4	82.1	66.6	70.4	79.0	93.5	99.3	69.9	60.7	72.1	89.2	61.8	65.9	32.4	51.9	58.1	70.8
LGrad	56.3	58.3	49.8	52.3	43.5	45.9	49.2	42.3	50.4	54.8	40.7	46.4	49.4	53.2	41.8	53.5	50.4	51.8	49.4
Ojha et al.	99.8	75.5	75.4	91.1	**88.5**	79.3	**83.7**	83.3	81.8	75.0	59.9	68.7	70.1	61.8	63.2	41.7	40.6	52.9	71.8
DIRE-1	48.4	42.5	39.1	53.5	54.3	44.1	48.0	44.7	46.1	47.0	66.2	62.8	53.2	47.1	44.6	47.6	51.0	57.4	49.9
DIRE-2	49.3	41.6	38.6	53.8	55.0	44.3	45.1	40.2	45.9	56.4	70.7	72.2	53.0	42.8	40.9	49.7	47.8	43.0	49.5
NPR	54.5	48.5	41.9	54.0	44.8	44.7	46.9	47.2	47.7	55.4	49.6	54.6	50.9	52.8	50.0	67.5	50.8	55.5	51.0
Ours 1k	**85.0**	64.0	66.6	90.2	75.2	74.7	78.1	97.2	77.1	77.6	80.1	86.6	77.5	76.5	77.9	77.4	63.1	70.5	77.5
Ours 1k+	78.7	62.5	68.4	**97.5**	67.9	84.0	74.3	**99.6**	78.2	83.7	94.5	97.1	88.9	**89.6**	81.2	**90.9**	77.6	83.7	83.2
Ours 10k	**85.7**	65.5	68.1	90.5	74.7	75.8	78.4	97.7	77.8	78.1	81.2	87.1	77.2	76.4	78.2	76.4	65.0	72.2	78.1
Ours 10k+	82.8	67.4	70.7	**98.4**	71.9	**85.4**	77.3	**99.7**	80.2	**85.8**	**95.9**	**98.2**	**91.1**	**89.9**	**83.8**	**90.1**	**79.4**	**85.5**	**85.2**

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., & Verdoliva, L. (2024). Raising the Bar of AI-generated Image Detection with CLIP. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4356-4366).

ON THE DETECTION OF SYNTHETIC IMAGES GENERATED BY DIFFUSION MODELS

Riccardo Corvi⋆, *Davide Cozzolino* ⋆, *Giada Zingarini*⋆, *Giovanni Poggi*⋆, *Koki Nagano*†, *Luisa Verdoliva*⋆

⋆ University Federico II of Naples
† NVIDIA

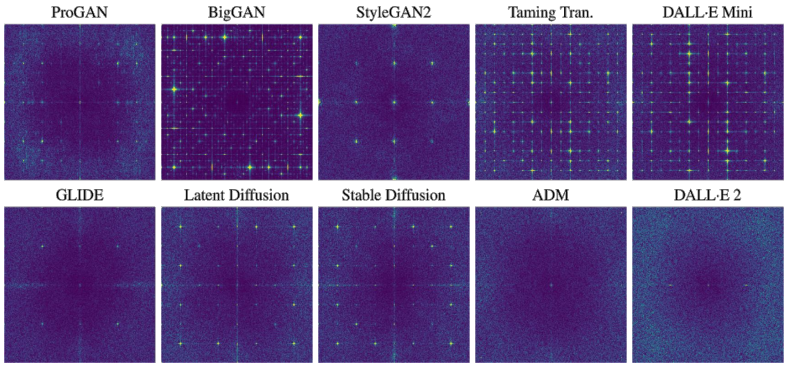
SoTA Detectors: Are They Good for Diffusion Models?

- Low level traces:
  - many SoTA detectors try to identify low level traces introduced in the generation phase
  - such traces are weak and can be (intentionally) removed
  - SoTA detectors work primarily on GAN, as DM are a relatively new architecture
- Questions:
  - do DM introduce low level artifacts, similar to GAN?
  - do SoTA detectors work also on DM?

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023, June). On the detection of synthetic images generated by diffusion models. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

SoTA Detectors: Artifact Analysis on DM

- Use the Fourier Domain:
  - DM presents some characteristic patterns
  - possible to perform also source identification
- Extraction process:
  - denoise the generated image
  - obtain the residual noise
  - average a large number of residuals
  - apply a Fourier transform
- Considerations:
  - large peaks in GAN generators
  - similar for some DM, suggesting good performances for fingerprint detectors
  - some other DM present weaker artifacts



Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023, June). On the detection of synthetic images generated by diffusion models. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

SoTA Detectors: Performances with GAN Training

- Uncompressed images:
  - good results in terms of AUC
  - accuracy often bad due to fixed threshold
  - almost perfect on ProGAN
  - good results on the other architectures
- Compressed/resized images:
  - reduction in performances
  - acceptable AUC
  - almost random accuracy
  - worst results on DALL-E 2 and ADM due to weak artifacts

Acc. threshold fixed at 0.5

	Trained on ProGAN							
	Uncompressed				Resized and Compressed			
Acc./AUC%	Spec	PatchFor.	Wang2020	Grag2021	Spec	PatchFor.	Wang2020	Grag2021
ProGAN	83.5/ 99.2	64.9/ 97.6	99.9/100	99.9/100	49.7/ 48.5	50.4/ 65.3	99.7/100	99.9/100
StyleGAN2	65.3/ 72.0	50.2/ 88.3	74.0/ 97.3	98.1/ 99.9	51.8/ 50.5	50.8/ 73.6	54.8/ 85.0	63.3/ 94.8
StyleGAN3	33.8/ 4.4	50.0/ 91.8	58.3/ 95.1	91.2/ 99.5	52.9/ 51.9	50.2/ 76.7	54.3/ 86.4	58.3/ 94.4
BigGAN	73.3/ 80.5	52.5/ 85.7	66.3/ 94.4	95.6/ 99.1	52.1/ 52.2	50.5/ 58.8	55.4/ 85.9	79.0/ 99.1
EG3D	80.3/ 89.6	50.0/ 78.4	59.2/ 96.7	99.4/100	58.9/ 60.6	49.8/ 81.9	52.1/ 85.1	56.8/ 96.6
Taming Tran.	79.6/ 86.6	50.5/ 69.4	51.2/ 66.5	73.5/ 96.6	49.0/ 49.1	50.0/ 64.1	50.5/ 71.0	56.2/ 94.3
DALL·E Mini	80.1/ 88.1	51.5/ 82.2	51.7/ 60.6	70.4/ 95.6	59.1/ 61.9	50.1/ 68.7	51.1/ 66.2	62.3/ 95.4
DALL·E 2	82.1/ 93.3	50.0/ 52.5	50.3/ 85.8	51.9/ 94.9	62.0/ 65.0	49.7/ 58.4	50.0/ 44.8	50.0/ 64.4
GLIDE	73.4/ 81.9	50.3/ 96.6	51.1/ 62.6	58.6/ 86.4	53.1/ 52.5	51.0/ 71.5	50.3/ 65.9	51.8/ 90.0
Latent Diff.	72.1/ 78.5	51.8/ 84.3	51.0/ 62.5	58.2/ 91.5	47.9/ 46.3	50.6/ 65.2	50.7/ 69.1	52.4/ 89.4
Stable Diff.	66.8/ 74.7	50.8/ 85.0	50.9/ 65.9	62.1/ 92.9	46.5/ 44.5	51.1/ 77.2	50.7/ 72.9	58.1/ 93.7
ADM	55.1/ 53.3	50.4/ 87.1	50.6/ 56.3	51.2/ 57.4	49.1/ 49.1	51.0/ 69.1	50.3/ 68.1	50.6/ 77.2
AVG	70.5/ 75.2	51.9/ 83.2	59.5/ 78.6	75.8/ 92.8	52.7/ 52.7	50.4/ 69.2	55.8/ 75.0	61.5/ 90.8

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023, June). On the detection of synthetic images generated by diffusion models. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

SoTA Detectors: Performances with DM Training

- Results of *Grag2021* trained on compressed/resized images:
  - good results on LD and SD due to similar architecture
  - laking performances on other DM
- Detector ensemble:
  - better performances by averaging the results with the network of the previous table
  - still low accuracy due to fixed threshold
- Calibration:
  - apply the "*Platt scaling method*"[A]
  - general increase in performances
  - not reliable with artifacts not seen during training

Acc./AUC%	Trained on Latent Diffusion	Fusion	Calibration
ProGAN	52.0/ 78.3	90.2/100	97.0/100
StyleGAN2	58.0/ 85.0	56.6/ 94.6	86.1/ 94.6
StyleGAN3	59.5/ 87.6	55.4/ 93.9	85.5/ 93.9
BigGAN	52.9/ 80.6	59.3/ 98.5	92.1/ 98.5
EG3D	65.4/ 91.8	54.4/ 97.7	92.3/ 97.7
Taming Tran.	78.2/ 97.3	61.5/ 98.2	91.2/ 98.2
DALL·E Mini	73.9/ 97.3	65.9/ 97.7	88.4/ 97.7
DALL·E 2	50.4/ 74.2	50.0/ 72.5	66.9/ 72.5
GLIDE	62.5/ 96.2	52.5/ 95.9	89.2/ 95.9
Latent Diff.	97.1/ 99.9	84.9/ 99.8	96.4/ 99.8
Stable Diff.	99.7/100	92.5/100	97.2/100
ADM	52.9/ 81.9	50.8/ 80.6	70.8/ 80.6
AVG	66.9/ 89.2	64.5/ 94.1	87.8/ 94.1

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023, June). On the detection of synthetic images generated by diffusion models. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.
[A] Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers, 10(3), 61-74.

SoTA Detectors: Conclusions on DM

- DM images are characterized by distinctive fingerprints
- Generalization is still problematic
- Inclusion of DM during training may help with similar architectures

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023, June). On the detection of synthetic images generated by diffusion models. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Adversarial Attacks

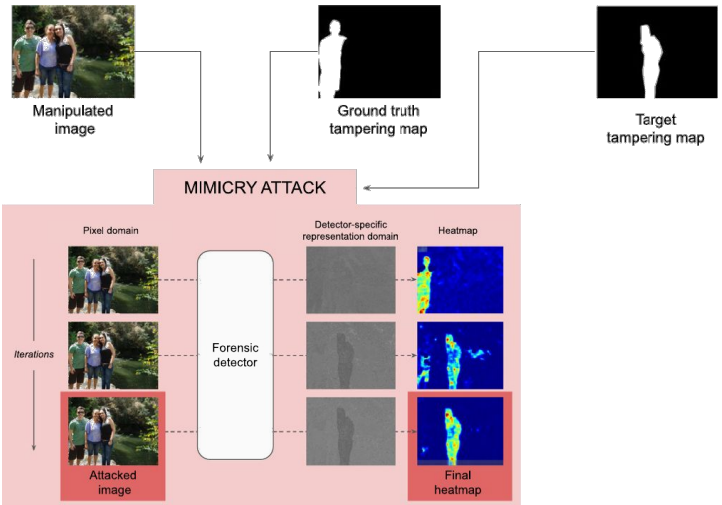
Adversarial mimicry attacks against image splicing forensics: An approach for jointly hiding manipulations and creating false detections

Giulia Boato [a,*], Francesco G.B. De Natale [a,e], Gianluca De Stefano [a,d], Cecilia Pasquini [a,c], Fabio Roli [b]

[a] *Department of Information Engineering and Computer Science, University of Trento, Italy*
[b] *Department of Informatics, Bioengineering, Robotics, and Systems Engineering, University of Genova, Italy*
[c] *Center for Cybersecurity, Fondazione Bruno Kessler, Italy*
[d] *CISPA Helmholtz Center for Information Security, Germany*
[e] *CNIT (Consorzio Nazionale Interuniversitario per le Telecomunicazioni), Italy*

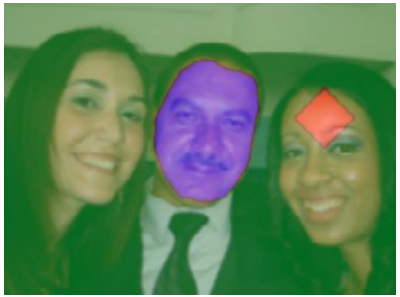
Mimicry: General Idea

- Modify a manipulated image such that:
  - the forged portion appears pristine to the detector
  - a decoy area (which can be either forged or pristine) appears as forged to the detector
- What do we have as input:
  - the manipulated image
  - the “ground truth” (aka. real) tampering map
  - a “target” tampering map
- The attack process:
  - split the image in patches
  - compute target representations
  - iteratively adjust the patches resemble the target tampering map



Boato, G., De Natale, F. G., De Stefano, G., Pasquini, C., & Roli, F. (2024). Adversarial mimicry attacks against image splicing forensics: An approach for jointly hiding manipulations and creating false detections. Pattern Recognition Letters, 179, 73-79.

Mimicry: Evaluation



- Split the heatmap in 3 regions:
  - the **GT area**, representing the real forgery we want to hide
  - the **D area**, representing the decoy forgery we want to highlight
  - the **BG area**, the background we are not interested in modifying

- Threshold-less indicators:
  - statistics extracted interpreting the detector heatmap
  - before the attack, the forgery is correctly detected
  - after the attack, the decoy forgery is detected instead
  - the real forgery is mixed with the background

Performance of EXIF-SC on DSO-1 dataset

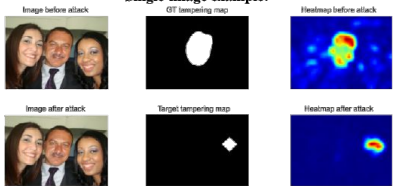
Threshold-less indicators

	med_{BG}	med_{D}	med_{GT}	V_{D}	V_{GT}	AUC_{D}	AUC_{GT}
No attack	0,13	0,15	0,47	0,02	0,34	0,48	0,82
After attack	0,04	0,46	0,06	0,39	0,02	0,93	0,66

Threshold-based indicators

	$F1_{D}$	$F1_{GT}$	MCC_{D}	MCC_{GT}	dr_{BG}	dr_{D}	dr_{GT}
τ_{GT}	0,08	0,29	0,12	0,12	0,60	0,96	0,85
τ_{D}	0,49	0,05	0,50	-0,03	0,06	0,72	0,07
τ_{OTSU}	0,38	0,07	0,40	-0,02	0,05	0,69	0,06
τ_{MP}	0,36	0,02	0,36	-0,03	0,01	0,42	0,01
τ_{d}	0,41	0,01	0,40,	-0,03	0,01	0,41	0,01
$\tau_{0.2}$	0,16	0,24	0,25	0,13	0,14	0,89	0,36

Single-image example:



Performance of Noiseprint on DSO-1 dataset

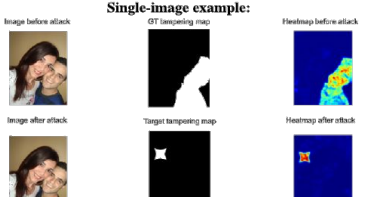
Threshold-less indicators

	med_{BG}	med_{D}	med_{GT}	V_{D}	V_{GT}	AUC_{D}	AUC_{GT}
No attack	0,03	0,04	0,33	0,01	0,29	0,46	0,91
After attack	0,02	0,64	0,06	0,61	0,04	0,97	0,68

Threshold-based indicators

	$F1_{D}$	$F1_{GT}$	MCC_{D}	MCC_{GT}	dr_{BG}	dr_{D}	dr_{GT}
τ_{GT}	0,10	0,37	0,14	0,21	0,60	0,96	0,92
τ_{D}	0,83	0,07	0,83	0,01	0,01	0,90	0,07
τ_{OTSU}	0,74	0,12	0,75	0,05	0,03	0,88	0,12
τ_{MP}	0,73	0,04	0,74	0,00	0,00	0,68	0,03
τ_{d}	0,81	0,03	0,80	-0,01	0,00	0,81	0,02
$\tau_{0.2}$	0,17	0,33	0,27	0,24	0,13	0,96	0,49

Single-image example:



Boato, G., De Natale, F. G., De Stefano, G., Pasquini, C., & Roli, F. (2024). Adversarial mimicry attacks against image splicing forensics: An approach for jointly hiding manipulations and creating false detections. Pattern Recognition Letters, 179, 73-79.

Mimicry: Evaluation

- Threshold-based indicators:
  - statistics extracted binarizing the heatmap using a certain threshold value
  - threshold value chosen by the forensic analyst, unknown to the attacker
  - generally, the real forgery is missed while the decoy is largely detected
  - with the best possible threshold (computed considering the GT TM), real forgery is largely detected, however also the majority of the background is detected as fake, resulting anyway in a high false alarm rate

Performance of EXIF-SC on DSO-1 dataset

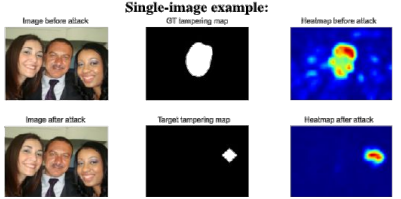
Threshold-less indicators

	med_{BG}	med_D	med_{GT}	V_D	V_{GT}	AUC_D	AUC_{GT}
No attack	0,13	0,15	0,47	0,02	0,34	0,48	0,82
After attack	0,04	0,46	0,06	0,39	0,02	0,93	0,66

Threshold-based indicators

	$F1_D$	$F1_{GT}$	MCC_D	MCC_{GT}	dr_{BG}	dr_D	dr_{GT}
τ_{GT}	0,08	0,29	0,12	0,12	0,60	0,96	0,85
τ_D	0,49	0,05	0,50	-0,03	0,06	0,72	0,07
τ_{OTSU}	0,38	0,07	0,40	-0,02	0,05	0,69	0,06
τ_{MP}	0,36	0,02	0,36	-0,03	0,01	0,42	0,01
τ_d	0,41	0,01	0,40,	-0,03	0,01	0,41	0,01
$\tau_{0.2}$	0,16	0,24	0,25	0,13	0,14	0,89	0,36

Single-image example:



Performance of Noiseprint on DSO-1 dataset

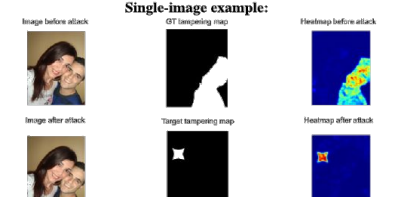
Threshold-less indicators

	med_{BG}	med_D	med_{GT}	V_D	V_{GT}	AUC_D	AUC_{GT}
No attack	0,03	0,04	0,33	0,01	0,29	0,46	0,91
After attack	0,02	0,64	0,06	0,61	0,04	0,97	0,68

Threshold-based indicators

	$F1_D$	$F1_{GT}$	MCC_D	MCC_{GT}	dr_{BG}	dr_D	dr_{GT}
τ_{GT}	0,10	0,37	0,14	0,21	0,60	0,96	0,92
τ_D	0,83	0,07	0,83	0,01	0,01	0,90	0,07
τ_{OTSU}	0,74	0,12	0,75	0,05	0,03	0,88	0,12
τ_{MP}	0,73	0,04	0,74	0,00	0,00	0,68	0,03
τ_d	0,81	0,03	0,80	-0,01	0,00	0,81	0,02
$\tau_{0.2}$	0,17	0,33	0,27	0,24	0,13	0,96	0,49

Single-image example:



Boato, G., De Natale, F. G., De Stefano, G., Pasquini, C., & Roli, F. (2024). Adversarial mimicry attacks against image splicing forensics: An approach for jointly hiding manipulations and creating false detections. Pattern Recognition Letters, 179, 73-79.

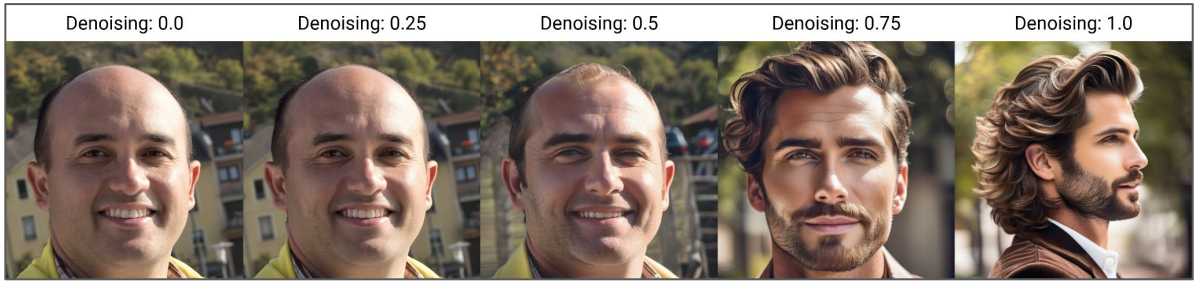
When Synthetic Traces Hide Real Content: Analysis of Stable Diffusion Image Laundering

Sara Mandelli, Paolo Bestagini, Stefano Tubaro
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy.
Emails: name.surname@polimi.it

Mandelli, S., Bestagini, P., & Tubaro, S. (2024). When Synthetic Traces Hide Real Content: Analysis of Stable Diffusion Image Laundering. arXiv preprint arXiv:2407.10736.

Image Laundering: Fast Introduction to Stable Diffusion

- Diffusion models! How do they work:
  - get a text prompt and some random noise
  - **encode** them in the latent space
  - perform some iterative denoising in the latent space
  - **decode** the final image
- Stable diffusion! Like usual diffusion models, but cooler:
  - they also accept **images** as input (img2img), with the addition of some random noise
  - a “strength” parameter [0, 1] allows the trade off between generated content and original image
- A simple example of a img2img translation:



PROMPT: Portrait of a man with a thick, voluminous hairstyle, realistic brown hair, neat and clean haircut, medium length, highly detailed, photorealistic, natural outdoor lighting, soft shadows, subtle background blur.

MODEL: sd_xl_base_1.0[B]

Mandelli, S., Bestagini, P., & Tubaro, S. (2024). When Synthetic Traces Hide Real Content: Analysis of Stable Diffusion Image Laundering. arXiv preprint arXiv:2407.10736.
[B] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0/tree/main

Image Laundering: The Problem

====>

strength = 0

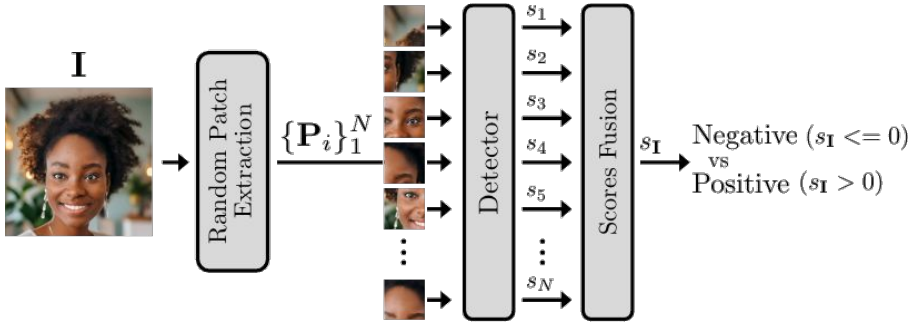
The (highlighted) difference between the 2 images

- What's happening:
  - The real image is encoded in the latent space using the SD encoder
  - Given the strength = 0, no noise is added and no denoising steps are performed
  - Then the the image is decoded using the SD decoder, introducing some **invisible artifacts**
- This is a problem:
  - the laundered image has no real forgery
  - despite this, it is classified as fake by many SoTA detectors
  - this approach could conceal nasty stuff as fake, despite being highly real

Mandelli, S., Bestagini, P., & Tubaro, S. (2024). When Synthetic Traces Hide Real Content: Analysis of Stable Diffusion Image Laundering. arXiv preprint arXiv:2407.10736.

Image Laundering: How To Address the Problem?

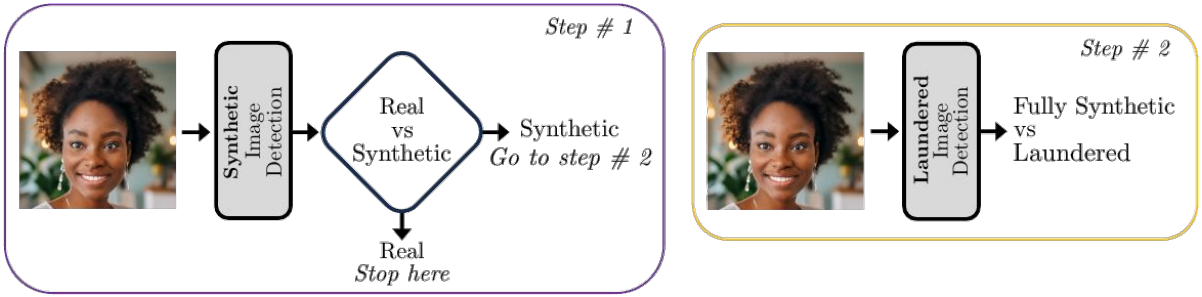
- Backbone architecture[C]:
  - split image in N random patches
  - assign a “fake” score to each patch
  - average the highest scores
    - positive score → synthetic image
    - negative score → real image



- Problems with backbone:
  - architecture perform only real vs. synthetic
  - laundered images detected as fully synthetic
  - trained only on real and fully synthetic images

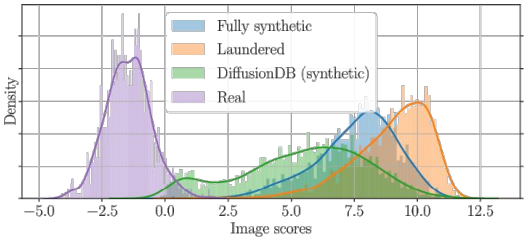
Mandelli, S., Bestagini, P., & Tubaro, S. (2024). When Synthetic Traces Hide Real Content: Analysis of Stable Diffusion Image Laundering. arXiv preprint arXiv:2407.10736.
[C] Mandelli, S., Bonettini, N., Bestagini, P., & Tubaro, S. (2022, October). Detecting gan-generated images by orthogonal training of multiple cnns. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 3091-3095). IEEE.

Image Laundering: Proposed Adjustments



- A two-stage architecture:
  - a first stage to discriminate real images from all the others (both synthetic and laundered)
  - a second stage to discriminate fully synthetic from laundered images

- Result evaluation:
  - optimal results at threshold 0 for the first stage
  - optimal results in the second stage for multiple SD models



	All data	SD-1.5	SD-2.1	SD-XL	SD-XL-turbo
AUC	0.994	0.993	0.994	0.993	**0.999**
B-ACC$_{max}$	96.1%	95.8%	96.6%	96.9%	**99.3%**
B-ACC$_{@thr=0}$	95.9%	93.9%	96.2%	94.8%	**98.6%**
TPR$_{@thr=0}$	94.8%	89.8%	94.9%	97.1%	**97.1%**
FPR$_{@thr=0}$	2.9%	2.0%	2.5%	7.4%	**0.0%**

Mandelli, S., Bestagini, P., & Tubaro, S. (2024). When Synthetic Traces Hide Real Content: Analysis of Stable Diffusion Image Laundering. arXiv preprint arXiv:2407.10736.

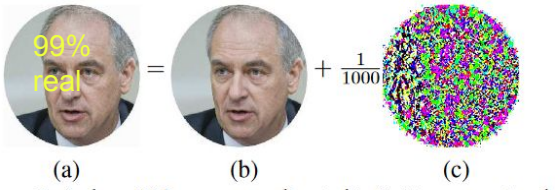
Evading Deepfake-Image Detectors with White- and Black-Box Attacks

Nicholas Carlini
Google Brain
Mountain View, CA
ncarlini@google.com

Hany Farid
University of California, Berkeley
Berkeley, CA
hfarid@berkeley.edu

Vulnerabilities in SoTA Forensic Detectors

- SoTA forensic classifiers achieve an AUC of 0.95, effectively identifying synthetic images from different generators



- Both whitebox and blackbox attacks lead to drastic reductions in AUC, exposing major weakness

White-Box Attacks

Full access to classifier's parameters

Minor pixel modifications can dramatically decrease the classifier's effectiveness

Black-Box Attacks

No access to classifier's parameters

Aware of **classifier's type**

Adversarial examples transferability: an input misclassified by one model is likely to be misclassified by another

Vulnerable even in this more restrictive setting

Carlini, N., & Farid, H. (2020). Evading deepfake-image detectors with white-and black-box attacks.

Exploited White & Black Box Attacks

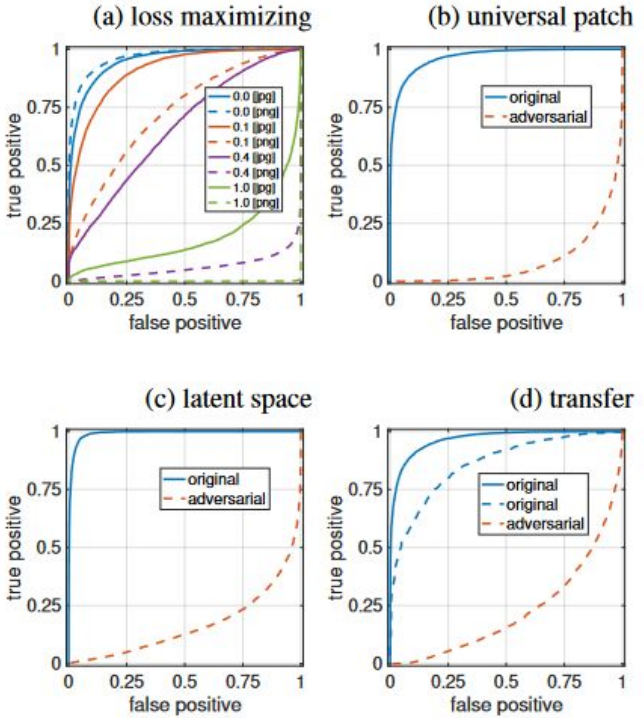
A. **Loss-Maximizing Attack**:
   Increases classification error by *maximizing loss*, reducing TPR

B. **Universal Adversarial-Patch Attack**:
   Generate a *single universal patch* that can be overlaid onto any fake image that then leads to misclassification

C. **Universal Latent-Space Attack**:
   Do not perturb input images → alters the generative model's *low-level attributes universally*, creating adversarial versions of fake images

D. **Black-Box Transfer Attack**:
   Uses adversarial examples from a surrogate model to degrade a more robust classifier's performance

ROC curves for forensic classifier (*Wang et al.* CNN based detector) before and after four distinct attacks.

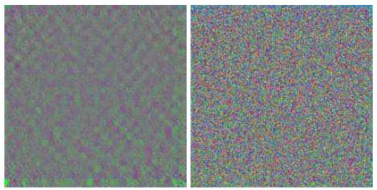


Carlini, N., & Farid, H. (2020). Evading deepfake-image detectors with white-and black-box attacks.

Findings

Classifier Sensitivity

- To fool **ImageNet** classifier detectors the perturbations are 10 times smaller



Mean perturbation for the forensics classifier (*Wang et al*) (left) and an ImageNet classifier (right) needed to lead to misclassification.

- **Reverse Attack:**
  - it is also possible to generate adversarial perturbation that cause real images to be misclassifier as fake → harder

Counter-Defenses

- Attacks are very **powerful**!
- Forensic classifiers need to build an adversarial model into their defenses
- This model must go beyond the standard laundering attacks of recompression, resizing, blurring, or adding white noise
- Effective defenses: **adversarial training** and **randomized smoothing**

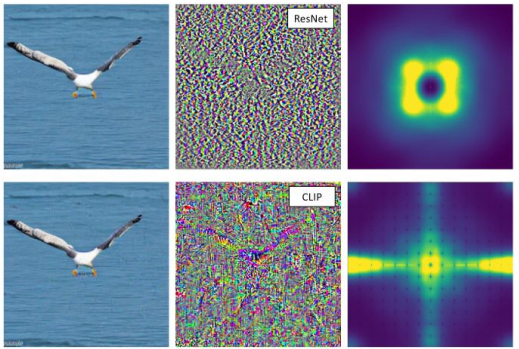
Carlini, N., & Farid, H. (2020). Evading deepfake-image detectors with white-and black-box attacks.

Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection

Vincenzo De Rosa, Fabrizio Guillaro, Giovanni Poggi, Davide Cozzolino and Luisa Verdoliva
University Federico II of Naples, Italy
Email: {vincenzo.derosa3, fabrizio.guillaro, poggi, davide.cozzolino, verdoliv}@unina.it

Challenges faced by CNN-based forensic detectors

- CNN-based:
  - perform well in both white and black-box scenarios but bad generalization
  - difficulties against unseen models and transformations
- CLIP-based:
  - promising alternative due to their Transformer backbones



- **Attack Transferability**:
  - Transferability of adversarial attacks examines whether an attack designed for one model is effective against another
  - 💡 the goal is to fool all (or most) of them!

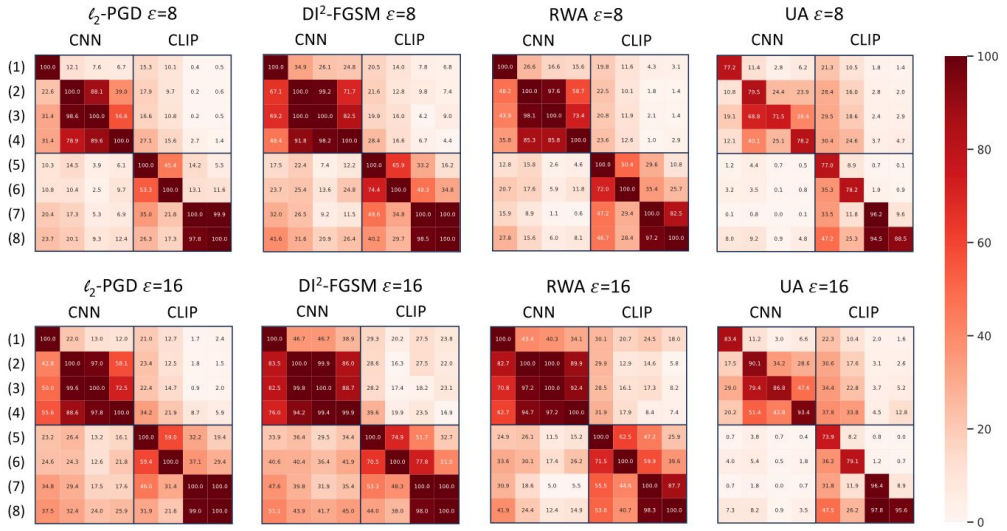
De Rosa, V., Guillaro, F., Poggi, G., Cozzolino, D., & Verdoliva, L. (2024). Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection. arXiv preprint arXiv:2407.19553.

Types of Attacks

- **Image-Specific Attacks:**
  - Modify input images x by adding perturbation δ to mislead the detector: *PGD*, *DI2-FGSM*
- **Universal Attacks (*UA*)**:
  - Develops one single adversarial perturbation that misleads classifiers across different images (fewer computations)
- **Latent-Space Attacks:**
  - Manipulate latent representations to induce errors without obvious changes to the image

De Rosa, V., Guillaro, F., Poggi, G., Cozzolino, D., & Verdoliva, L. (2024). Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection. arXiv preprint arXiv:2407.19553.

Architectural differences impact on attack transferability



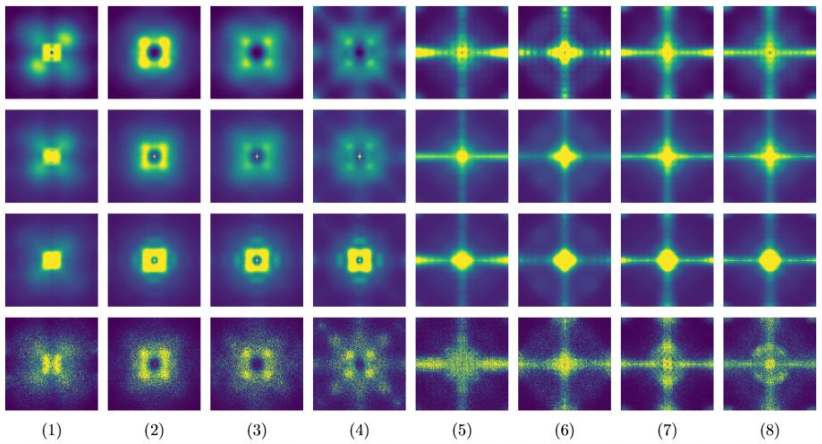
Successful Attack Rate (SAR) of four attacks.
- White-box → diagonal
- Transferred attacks → off-diagonal cells

- Transferability is highly successful within the same family of detectors (CNN → CNN or CLIP → CLIP)
- Mostly non-transferable between CNN and CLIP models
  - CLIP based detectors are found to be vulnerable to white-box attacks just like CNN-based detectors
  - CLIP models are more robust → better resist fake-to-real attacks

De Rosa, V., Guillaro, F., Poggi, G., Cozzolino, D., & Verdoliva, L. (2024). Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection. arXiv preprint arXiv:2407.19553.

Spatial and Frequency Domains

- **Low transferability** of adversarial attacks. This indicates that each model type may process images in fundamentally different ways



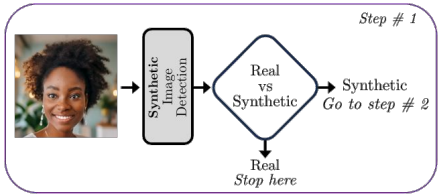
- CNN-based detectors focus on **medium-to-high frequencies**, **isotropic** spectra
- CLIP-based detectors rely on **low-frequency** patterns, **cross-shaped** spectra

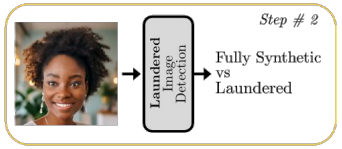
De Rosa, V., Guillaro, F., Poggi, G., Cozzolino, D., & Verdoliva, L. (2024). Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection. arXiv preprint arXiv:2407.19553.

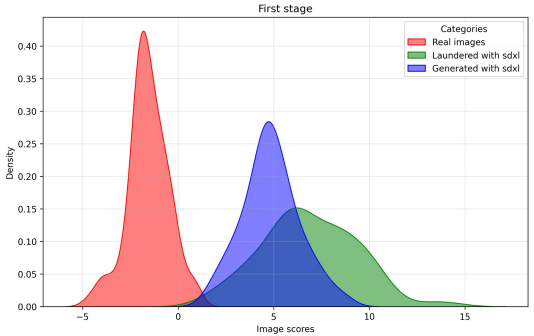
Conclusions

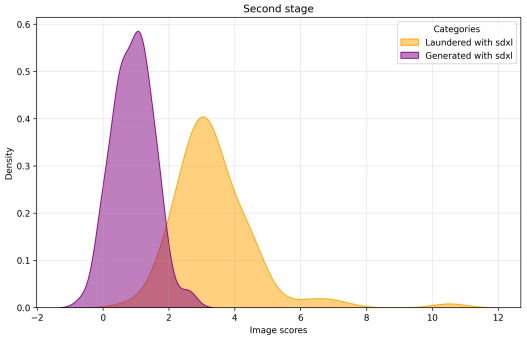
- We presented a general overview of detectors and attacks
- We performed some initial testing with the dataset provided by MMLAB using the code provided by the previous papers, obtaining promising results

===>









Our Next Work

- We will focus our future work on DM:
  - such architectures are relatively new
  - they can be executed on cheap hardware…
  - … allowing more users to use them …
  - … hence the necessity to assess their real capabilities

- The next steps:
  - continue the previous experiments with the MMLAB dataset
  - assess the performance of the described approaches on the MMLAB dataset
  - assess the capabilities of the attacks proposed on the MMLAB dataset
  - research, if feasible, some possible solutions to such attacks

Thanks for the attention