NoiseScope: Detecting Deepfake Images in a Blind Setting

Jiameng Pu
Virginia Tech

Neal Mangaokar, Bolun Wang, Chandan K. Reddy, Bimal Viswanath
Virginia Tech, Facebook

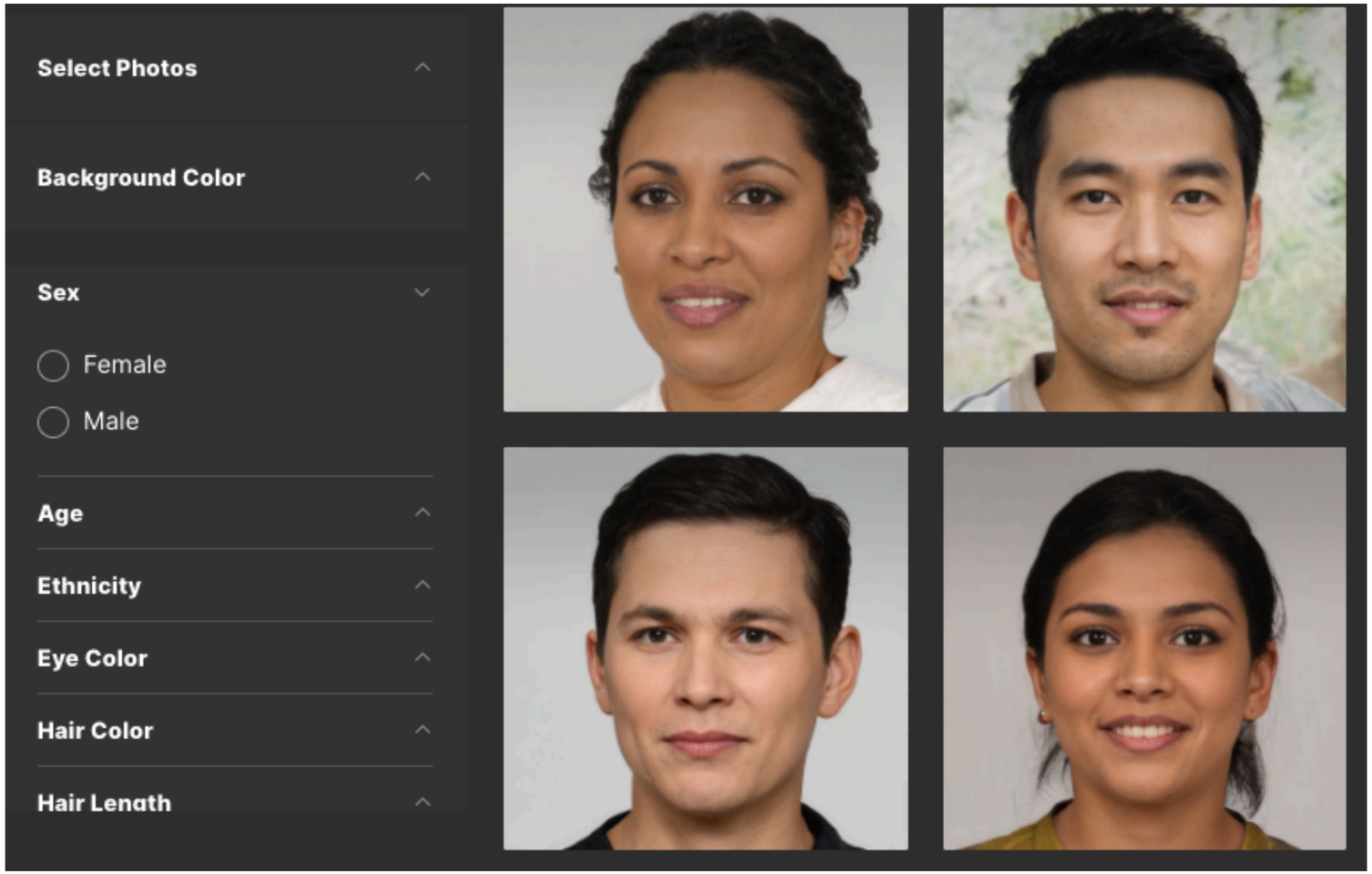
What are deepfakes?

Deepfakes are synthetic media produced using deep generative models.

- This includes multiple modalities: images, videos, audio, text, etc.
- In this work, we focus on deepfake images.

Deepfakes are now easily generated

- https://generated.photos



Deepfakes: An AI-powered threat

Deepfakes on the web are a serious threat:

- Fake social media profiles [1]
- Fake pornography [2]
- Disinformation campaigns [3]

[1] Facebook removes bogus accounts that used ai to create fake profile pictures.

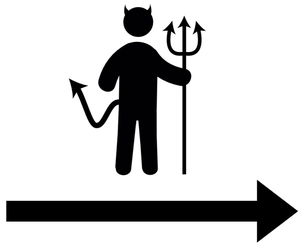
[2] Deepfake Porn Nearly Ruined My Life.

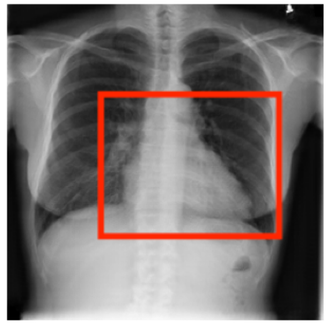
[3] Deepfake videos could 'spark' violent social unrest.

Deepfakes can be a threat beyond the web

In the healthcare domain, deepfake images can be used by attackers to trigger misdiagnosis by both doctors and machine learning algorithms [1].

Non-disease X-ray



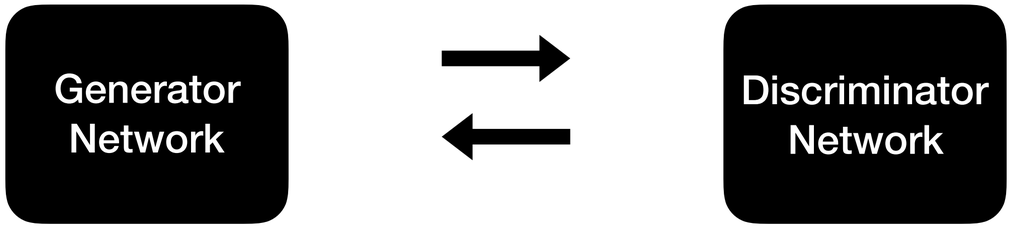


X-ray with disease ‘cardiomegaly’ (deepfake)

[1] Jekyll: Attacking Medical Image Diagnostics using Deep Generative Models. In Proc. of Euro S&P, 2020.

Deepfakes enabled by Generative Adversarial Networks (GANs)

Generator creates an image



Discriminator gives it feedback for improvement

GAN 101

Advances in GANs over the years

GAN
2014

DCGAN
2016

CoGAN
2016

PGGAN
2018

StyleGAN
2019

Photorealistic images produced by GANs bring new threats.

Our goal is to detect deepfake images.

Limitations of prior detection schemes

- Train a supervised DNN classifier to distinguish between fake and real images [1-3].
  - Hard to get access to a large amount of fake content.
  - Poor generalizability.
- Anomaly detection: Analyze high-level image content for semantic inconsistencies.
  - However, the quality of GAN-generated images has been significantly improved.
- Both above approaches are prone to bias issues as they analyze high level image content.
  - E.g., detection schemes can lead to racial bias [4].

[1] Mesonet: a Compact Facial Video Forgery Detection Network. In Proc. of WIFS, 2018.

[2] Detection Of GAN-generated Fake Images Over Social Networks. In Proc. of MIPR, 2018.

[3] Detecting Both Machine and Human Created Fake Face Images in the Wild. In Proc. of MPS, 2018.

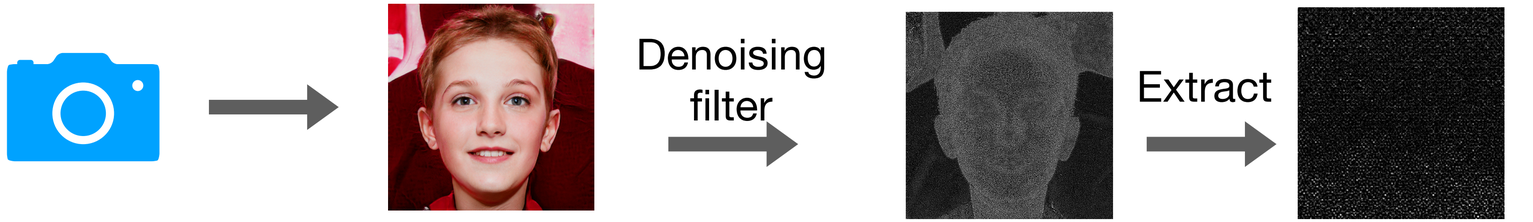
[4] Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In Proc. Of CVPR, 2018.

NoiseScope: Blind detection of deepfake images

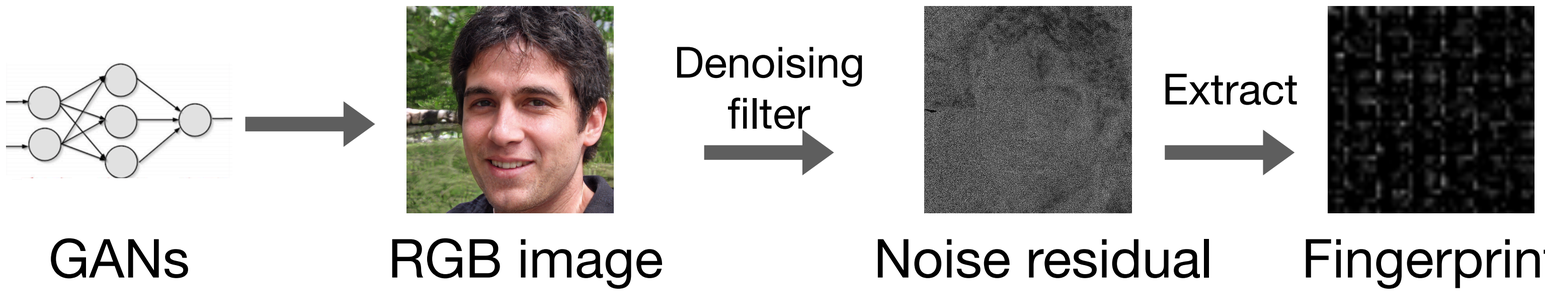
- Given a test set of fake and real images, NoiseScope identifies fake images.
- Blind detection scheme: Defender has no a priori access to fake images, or knowledge of the generative model used by the attacker.
- Defender has access to a set of real images.
- NoiseScope is agnostic to the type of GAN used and works for any type of high-level image content.

Key idea: Leverage noise pattern

Our work is inspired by prior work in camera fingerprinting [1-2].



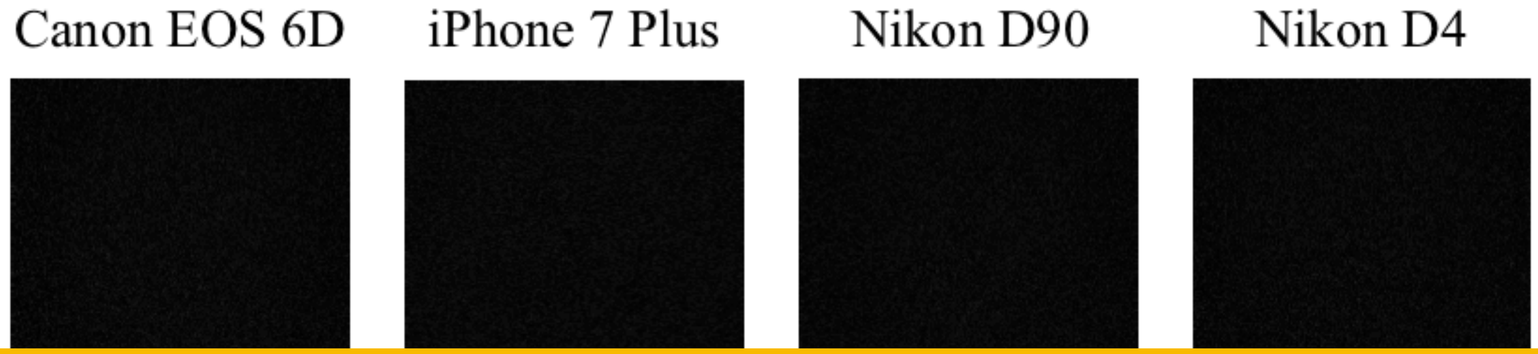
Do GANs leave fingerprints in generated images as well?



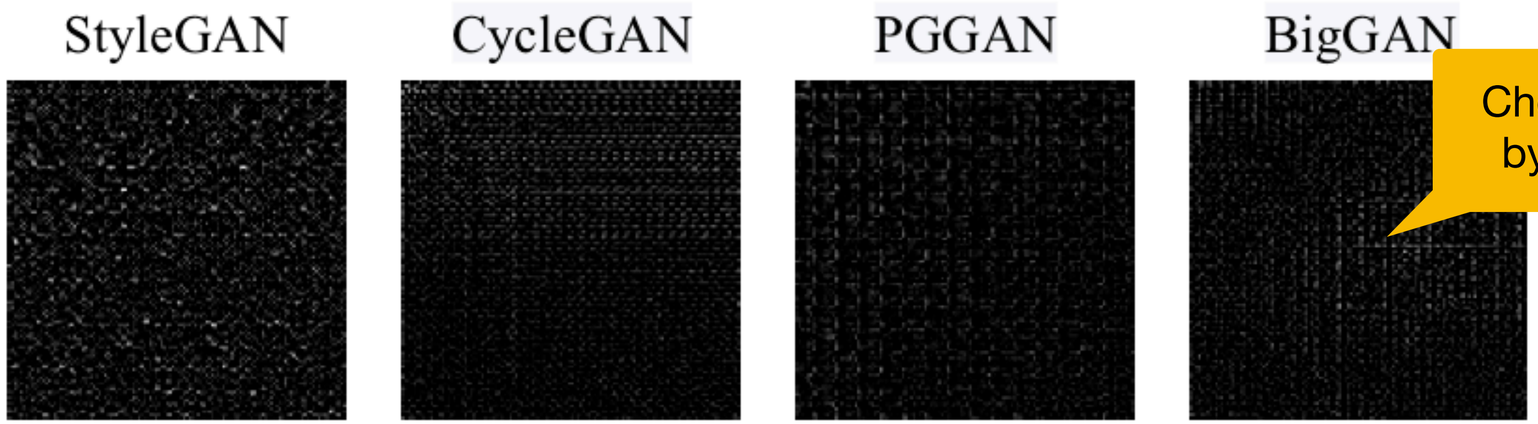
[1] Digital Imaging Sensor Identification. In Proc. of Security, Steganography, and Watermarking of Multimedia Contents, 2007.

[2] Determining Image Origin and Integrity Using Sensor Noise. IEEE Transactions on Information Forensics and Security, 2008.

Model vs device fingerprints



It is possible to differentiate between model and device fingerprints



Checkerboard patterns are caused by deconvolution layers of GANs

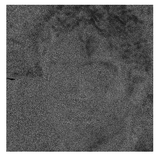
How do we extract a fingerprint?




...


Images from the same source
(device or model)

Denoising filter

...


Noise residuals

Averaging

The source fingerprint
(camera or model)

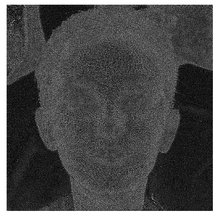
Cancel out random noise by averaging residuals

Using a model fingerprint to detect fake images

Once we have a model fingerprint, it is easy to detect any fake images in the test set!

Test image

Denoising filter



Noise residual R

Model fingerprint F_{GAN}

$$\rho_{F_{GAN},R} = corr(F_{GAN}, R)$$

Fake image

Real image

Extracting model fingerprints in a blind setting

Key challenges:

- It is hard to extract a fingerprint from a single image.
- Defender has no a priori access to fake images, or the knowledge of generative models used.

NoiseScope extracts any available model fingerprints from the test itself.

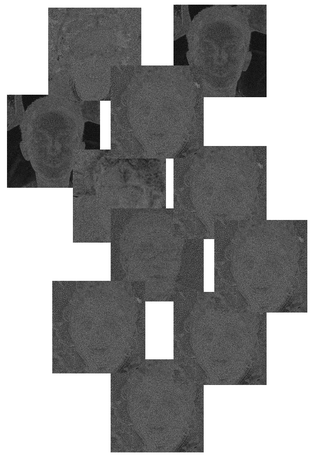
- In an unsupervised manner.

NoiseScope step 1: Extract noise residuals

RGB space

Noise space

Denoising filter

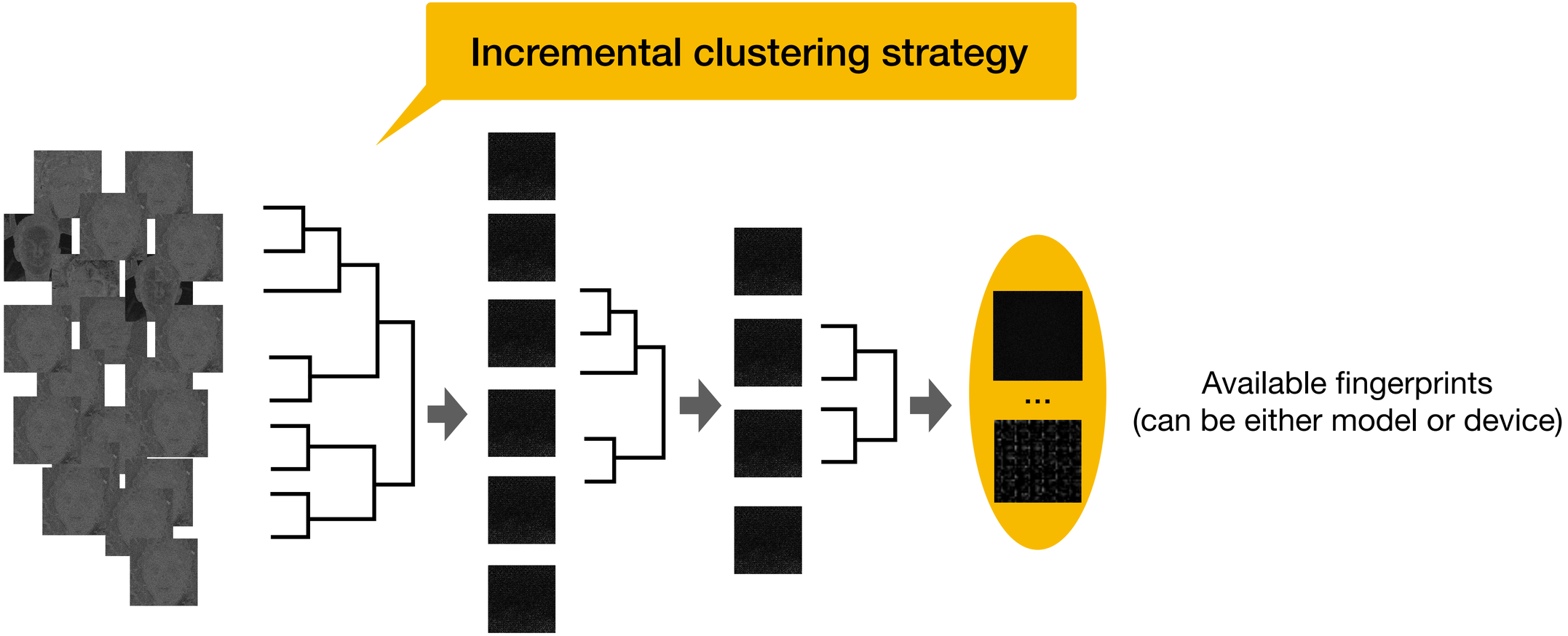


Images in the test set

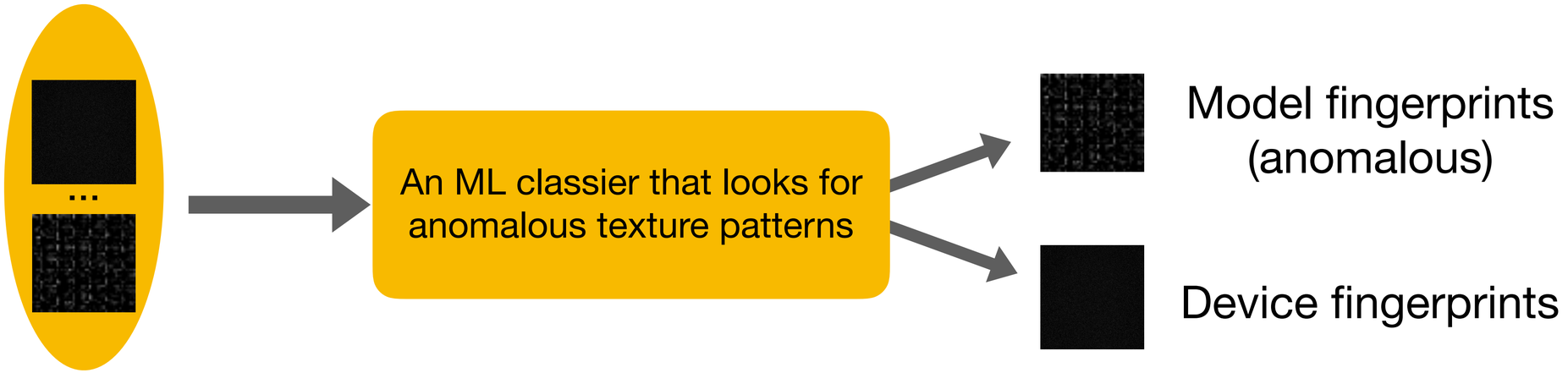
Noise residuals

Convert images to noise residuals by suppressing high-level content.

NoiseScope step 2: Extracting fingerprints via clustering



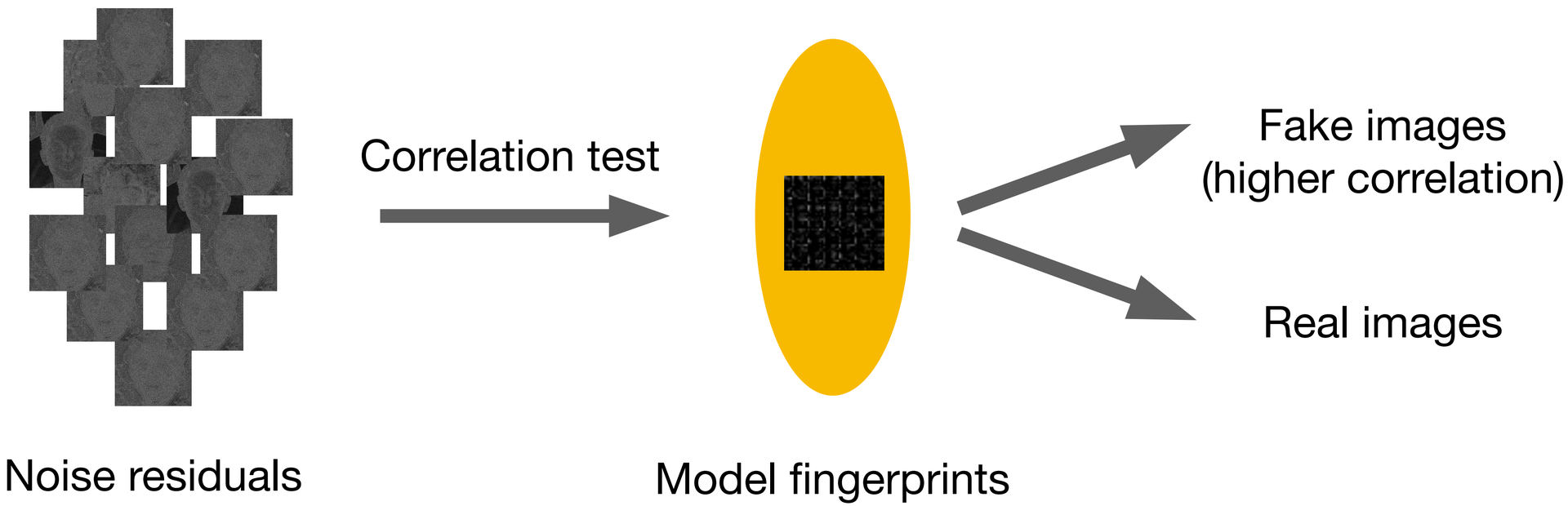
NoiseScope step 3: Fingerprint classification



Available fingerprints
(can be either model or device)

A highly performant outlier detector can be trained based on the texture features.

NoiseScope step 4: Fake image detector



If the correlation higher than a pre-calibrated threshold, then image is fake.

Detection performance of NoiseScope

- Evaluated on 11 datasets covering 4 GANs: StyleGAN, BigGAN, PGGAN, CycleGAN.
- NoiseScope outperforms existing CSD-SVM [1] on all the datasets.

F1 score

	NoiseScope	CSD-SVM
StyleGAN-Face1	**99.56%**	92.93%
StyleGAN-Face2	**90.14%**	67.53%
StyleGAN-Bed	**99.63%**	94.82%
BigGAN-DogLV	**99.38%**	86.94%
BigGAN-DogHV	**92.6%**	70.1%
BigGAN-BurgLV	**99.68%**	94.82%
BigGAN-BurgHV	**98.64%**	83.67%
PGGAN-Face	**99.09%**	64.07%
PGGAN-Tower	**95.93%**	91.61%
CycleGAN-Winter	**92.4%**	87.14%
CycleGAN-Zebra	**92.84%**	84.95%

Each dataset contains
500 real and 500 fake images.

[1] Detection of Deep Network Generated Images Using Disparities in Color Components. arXiv preprint, 2018

Comparison with supervised approach

We compared NoiseScope with a well known supervised approach, MesoNet [1]. MesoNet is trained on StyleGAN images, and tested on two testing datasets.

	F1 Score when tested on	
	StyleGAN	PGGAN
MesoNet (trained on StyleGAN)	94%	65%
NoiseScope	**99%**	**98%**

- MesoNet fails to generalize to unseen image distributions.
- NoiseScope consistently performs well on both datasets.

[1] Mesonet: a Compact Facial Video Forgery Detection Network. In Proc. of WIFS. 2018

Evaluate the robustness of NoiseScope

We evaluated 6 countermeasures targeting different components of NoiseScope's pipeline.

- Evade detection by disrupting fingerprints using image compression?
  - NoiseScope is resilient. NoiseScope can still capture any artifacts introduced by compression.

- Fingerprint spoofing: Disguise fake images to be from a specific camera device.
  - NoiseScope can be made resilient by using multiple filters while extracting the noise residuals.
  - Aggressive spoofing against multiple filters will significantly degrade the image quality.

A lot more evaluation in the paper…

- Detection performance on
  - Imbalanced test sets.
  - Test sets with too few fake images.
  - Test set contains fake images from multiple GAN models.
  - Test sets with images from multiple content domains.

- Other post-processing schemes to disrupt fingerprints.
  - Gamma correction, histogram equalization, adding noise, etc.

Conclusion and future work

- Blind detection is a promising direction.
  - It compensates for the poor generalization performance of supervised approaches.
- Insights from the mature field of digital image forensics are useful to detect deepfakes.
- We are working on extending NoiseScope to detect deepfake videos.

Our code is available at: https://github.com/jmpu/NoiseScope