# Responsible Disclosure of Generative Models Using Scalable Fingerprinting

Ning Yu*[1,2,3] Vladislav Skripniuk*[4] Dingfan Chen[4] Larry Davis[2] Mario Fritz[4]
*Equal contribution
[1]Salesforce Research [2]University of Maryland [3]Max Planck Institute for Informatics
[4]CISPA Helmholtz Center for Information Security
https://github.com/ningyu1991/ScalableGANFingerprints

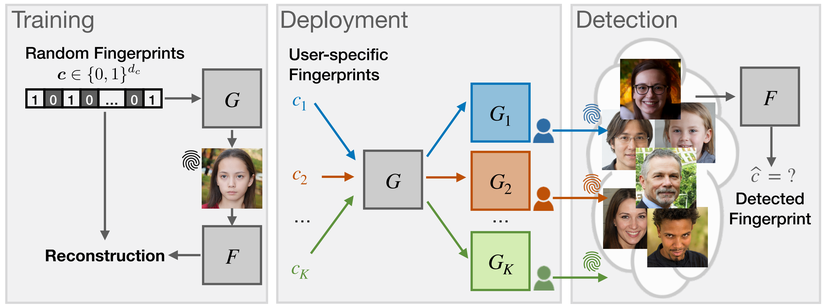
ICLR International Conference On Learning Representations

## Motivations

- Generative Adversarial Networks (GANs), evolve fast in the past 8 years for photorealistic generation, which raise significant concerns about visual misinformation.
- Move from passive to **proactive** defense against DeepFake misuses, so as to be **independent** of the arms race between DeepFake generation and detection.
- Enable **responsible** release and regulation of DeepFake models.
- Train a meta GAN model once and can instantiate a **large population** of fingerprinted GAN versions **efficiently** during deployment stage.

## Pipeline

- Jointly train a **meta GAN** with a **fingerprint auto-encoder**.
- During deployment, **ad-hoc** instantiate different GAN versions with unique fingerprints for different user downloads.
- When a misuse happens, use the decoder to **detect** the fingerprint in the generated image and **attribute** whose generator instance made it.

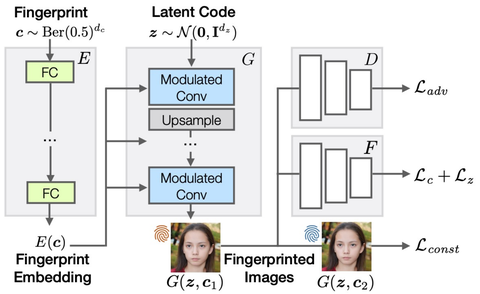


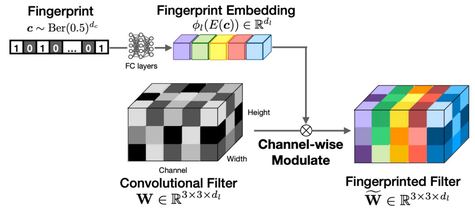

## Goals

| Fingerprinting efficiency & scalability | Generation fidelity | Fingerprint robustness | DeepFake Detection/attribution |
|---|---|---|---|
| After **one-time training**, fingerprinting should **scale up** to a **large population** of generator instances efficiently. | Encoded fingerprints do **not hurt** the original **generation quality** and keep **invisible** to human eyes. | Encoded fingerprints should **persist** within a reasonably wide range of image/model **perturbations.** | The effectiveness to convert the complex classification problem to the **simple fingerprint verification** problem. |

## Method

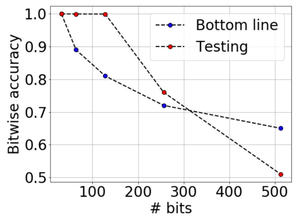
- Train to **encode** 128-bit fingerprints to their embeddings.
- **Modulate** the convolutional kernels of the generator using the fingerprint embeddings.
- Input random latent codes through the beginning of the generator to **generate** images.
- Three loss terms:
  - **Adversarial loss** preserves generation fidelity.
  - Fingerprint **reconstruction loss** enables fingerprint detection.
  - Consistency loss encourages generators with different fingerprints to have the same generation appearances.




## Fingerprint detection accuracy and generation fidelity

| Method | CelebA Bit acc ⇑ | CelebA $p$-value ⇓ | CelebA FID ⇓ | LSUN Bedroom Bit acc ⇑ | LSUN Bedroom $p$-value ⇓ | LSUN Bedroom FID ⇓ | LSUN Cat Bit acc ⇑ | LSUN Cat $p$-value ⇓ | LSUN Cat FID ⇓ |
|---|---|---|---|---|---|---|---|---|---|
| StyleGAN2 | - | - | 9.37 | - | - | 19.24 | - | - | 31.01 |
| outguess | 0.533 | 0.268 | 10.02 | 0.526 | 0.329 | 20.15 | 0.523 | 0.329 | 32.30 |
| steghide | 0.535 | 0.268 | 9.48 | 0.530 | 0.268 | 19.77 | 0.541 | 0.213 | 31.67 |
| (Yu et al., 2021) | 0.989 | $< 10^{-36}$ | 14.13 | 0.983 | $< 10^{-34}$ | 21.31 | 0.990 | $< 10^{-36}$ | 32.60 |
| Ours | 0.991 | $< 10^{-36}$ | 11.50 | 0.993 | $< 10^{-36}$ | 20.50 | 0.996 | $< 10^{-36}$ | 33.94 |
| Ours Variant I | 0.999 | $< 10^{-38}$ | 12.98 | 0.999 | $< 10^{-38}$ | 20.68 | 0.500 | 0.535 | 34.23 |
| Ours Variant II | 0.987 | $< 10^{-36}$ | 13.86 | 0.927 | $< 10^{-25}$ | 21.70 | 0.869 | $< 10^{-17}$ | 34.33 |
| Ours Variant III | 0.990 | $< 10^{-36}$ | 22.59 | 0.896 | $< 10^{-21}$ | 64.91 | 0.901 | $< 10^{-23}$ | 51.74 |

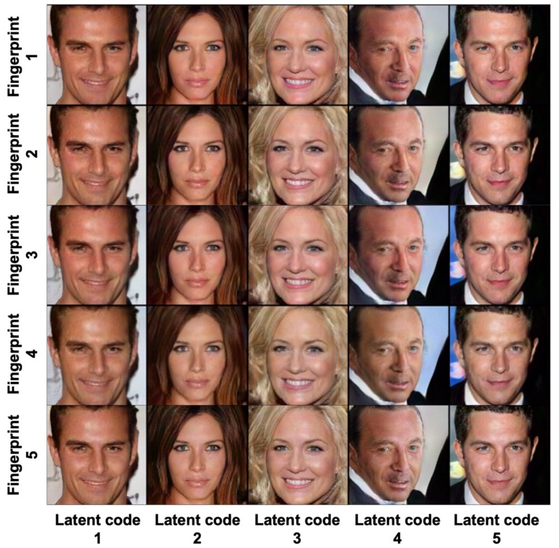



Capacity

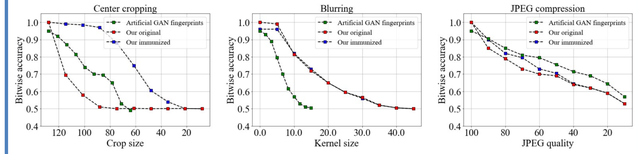
| Fingerprint set size | Training acc ⇑ | Testing acc ⇑ |
|---|---|---|
| 10 | 1.000 | 0.512 |
| 100 | 1.000 | 0.537 |
| 1k | 1.000 | 0.752 |
| 10k | 0.990 | 0.988 |
| 100k | 0.983 | 0.981 |
| Sampling on the fly | 0.991 | 0.991 |

Scalability

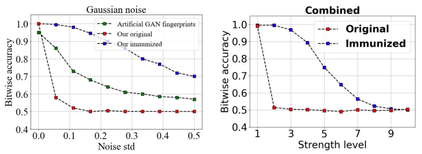
## Fingerprint & latent code disentanglement




## Fingerprint Robustness




Center cropping Gaussian blur JPEG compression




Gaussian noise Combined perturbations

## DeepFake Detection and Attribution

| Method | Closed world #GANs 1 | Closed world #GANs 10 | Closed world #GANs 100 | Open world #GANs 1 | Open world #GANs 10 | Open world #GANs 100 |
|---|---|---|---|---|---|---|
| (Yu et al., 2019) | 0.997 | 0.998 | 0.955 | 0.893 | 0.102 | N/A |
| (Wang et al., 2020) | 0.890 | N/A | N/A | 0.883 | N/A | N/A |
| (Yu et al., 2021) | 1.000 | 1.000 | N/A | 1.000 | 1.000 | N/A |
| Ours | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |