Deepfake Detection

CSIT375/975 AI and Cybersecurity

Dr Wei Zong

SCIT University of Wollongong

Disclaimer: The presentation materials come from various sources. For further information, check the references section

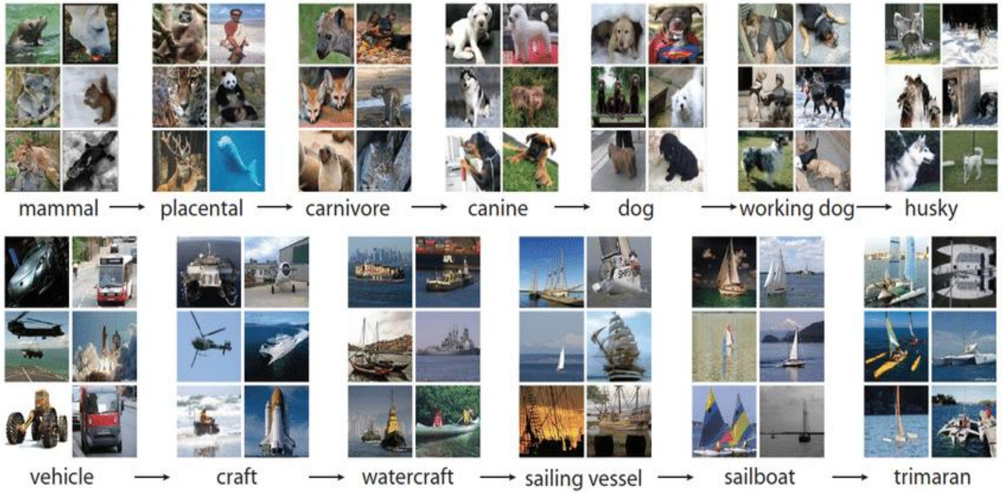
Outline

- Introduction to Generative Models
- Deepfake image detection
  - Reactive detection
    - Detecting artifacts.
  - Proactive detection
    - Watermarking techniques.

Introduction

Deep learning:

- Discriminative models
  - Classifies input



- Generative models
  - Learns distribution of data

Introduction

GAN

- GAN is a hot research topic.

Generative adversarial nets A* CCF A

I Goodfellow, J Pouget-Abadie… - Advances in neural …, 2014 - proceedings.neurips.cc

… We propose a new framework for estimating **generative** models via **adversarial nets**, in which we simultaneously train two models: a **generative** model G that captures the data …

Save Cite Cited by 75796 Related articles All 67 versions

  - Over 75k citations so far.
  - In 2020, approximately 28,500 papers related to GANs were published
    - Approximately 78 papers every day or more than three per hour

GAN

Generative Adversarial Nets

Ian J. Goodfellow,* Jean Pouget-Abadie,† Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,‡ Aaron Courville, Yoshua Bengio§
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

- Published in Neurips 2014
  - Top AI conference
- Two versions online
  - arXiv vs. conference paper
  - Related work

Ian Goodfellow

DeepMind
Verified email at deepmind.com - Homepage
Deep Learning

FOLLOW

Cited by

	All	Since 2018
Citations	262417	245133
h-index	87	85
i10-index	144	140

Yoshua Bengio

Professor of computer science, University of Montreal, Mila, IVADO, CIFAR
Verified email at umontreal.ca - Homepage
Machine learning deep learning artificial intelligence

FOLLOW

Cited by VIEW ALL

	All	Since 2018
Citations	709126	594552
h-index	225	200
i10-index	791	708

GAN

Why need GAN? (motivation)

- "Deep generative models have had less of an impact, due to the difficulty of approximating many intractable probabilistic computations that arise in maximum likelihood estimation and related strategies, and due to difficulty of leveraging the benefits of piecewise linear units in the generative context."

Variational Autoencoder:

- Optimize evidence lower bound (ELBO)

GAN

What is GAN?

- Adversarial nets is a **framework**, the generative model is pitted against an adversary: a discriminative model that learns to determine whether a sample is from the model distribution or the data distribution.
- The two-player minimax game -> Nash Equilibrium

Generative Model

Discriminative Model

a team of counterfeiters

police

GAN

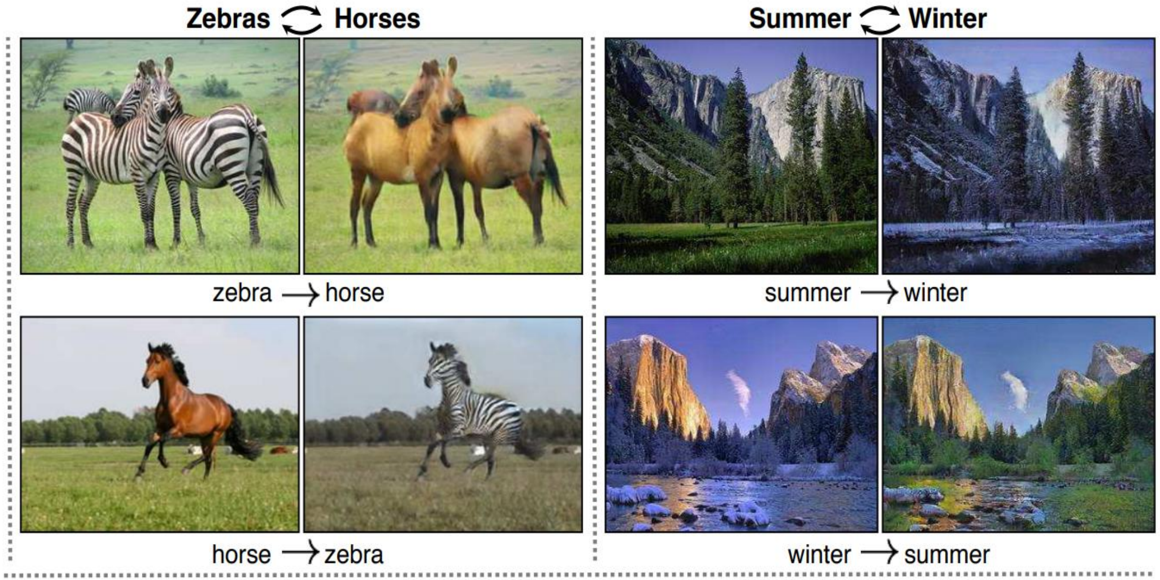
- Results
  - a) MNIST; b) TFD c) CIFAR-10 d) CIFAR-10.
  - Rightmost column shows the nearest training example of the neighboring sample.
    - Demonstrate that the model has not memorized the training set.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. Advances in neural information processing systems, 27.

Rapid Advance in GAN

CycleGAN

- Given any two unordered image collections X and Y
- CycleGAN learns to automatically "translate" an image from one into the other.
- "Translate" horses into zebra and vice versa.
- "Translate "summer scene into winter scene and vice versa.

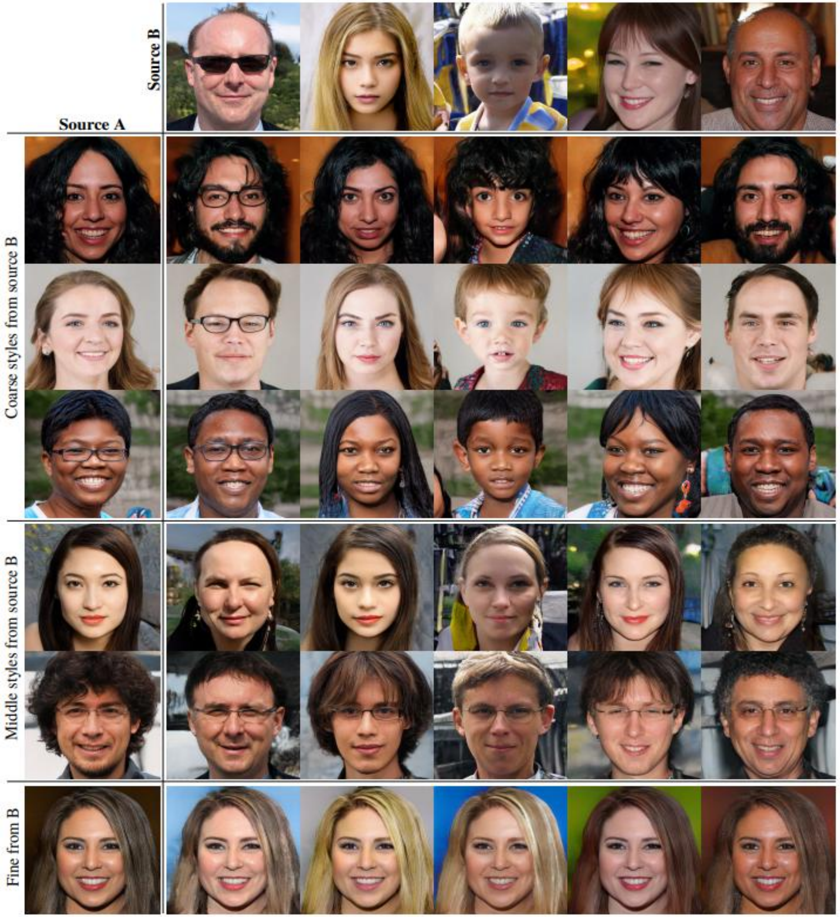


Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

Rapid Advance in GAN

StyleGan

- Images generated by copying a specified subset of styles from source B and taking the rest from source A.
- Able to control the level of details copied from B.
    - Coarse styles
    - Middle styles
    - Fine styles



Karras, T., Laine, S. and Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).

Beyond GAN

Variational Autoencoders (VAE)

- Following variational bayes inference, VAEs are generative models that attempt to reflect data to a probabilistic distribution and learn reconstruction that is close to its original input.

Flow

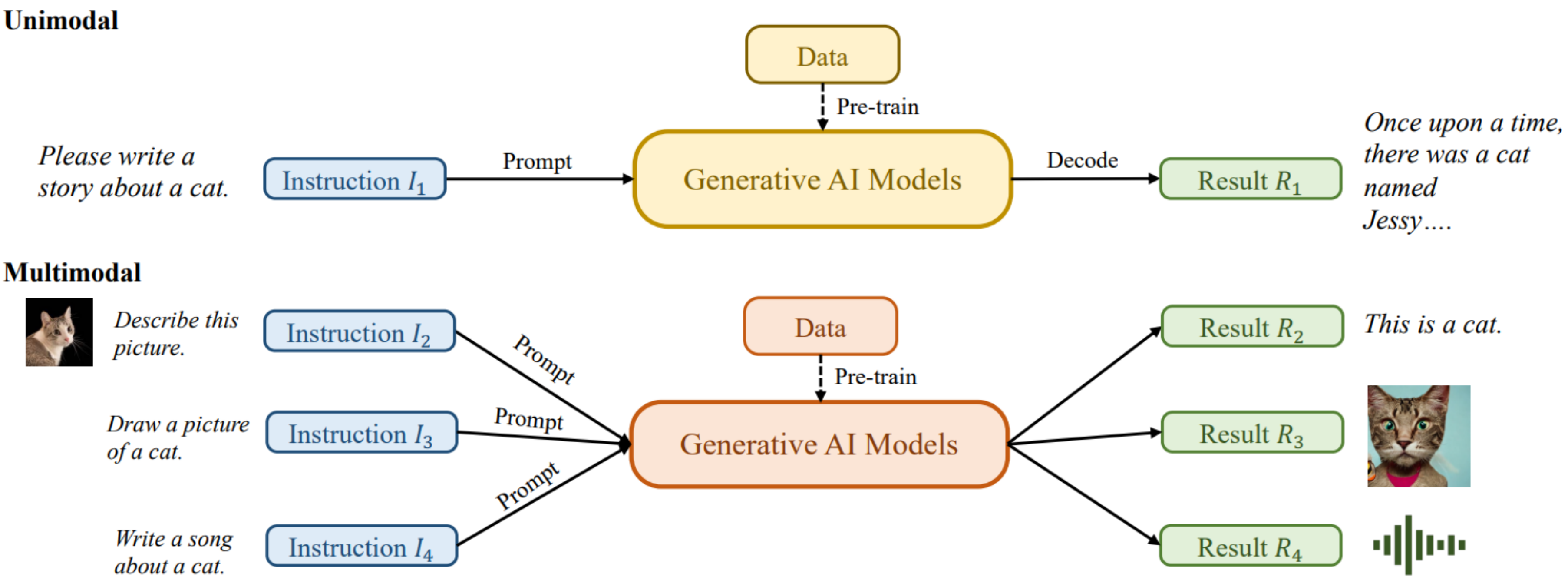
- A Flow is a distribution transformation from simple to complex by a sequence of **invertible** and differentiable mappings.

Diffusion

- The Generative Diffusion Model (GDM) is a cutting-edge class of generative models based on probability, which demonstrates state-of-the-art results in the field of computer vision.
- It works by progressively corrupting data with multiple-level noise perturbations and then learning to reverse this process for sample generation.

Beyond GAN

From Unimodal to MultiModel

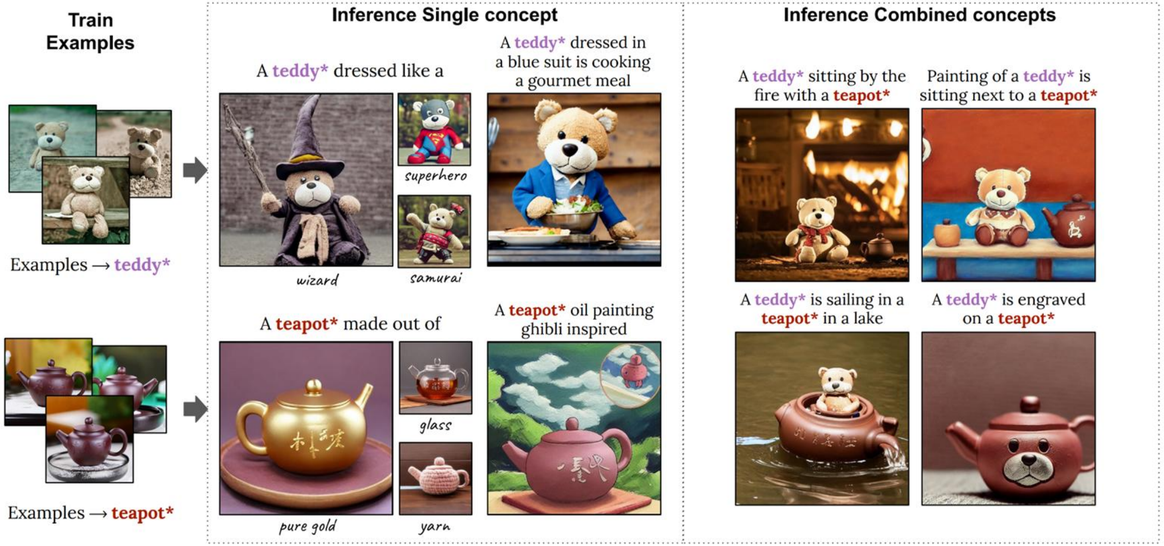


Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S. and Sun, L., 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*.

Beyond GAN

Perfusion (2023 NVIDIA)

- A new text-to-image personalization method.
- With only a 100KB model size per concept (excluding the pretrained model, which is a few GBs), trained for roughly 4 minutes, Perfusion can creatively portray personalized objects.
- It allows significant changes in their appearance, while maintaining their identity.
- Perfusion can also combine individually learned concepts into a single generated image.



Fake Image Detection

- The generative technique is a double-edged sword.
  - Humans cannot distinguish between real or fake images anymore.
  - The traditional perspective of treating visual media as trustworthy content is not longer valid.
    - Adversaries can use this technique to spread fake information or commit crimes.

- Detecting fake images becomes an emerging research trend:
  - Reactive detection
    - Detect artefacts in generated images.
  - Proactive detection
    - Embed watermarks into generated images.

Is this a real or synthesized?

https://en.wikipedia.org/wiki/StyleGAN

Detecting Artefacts

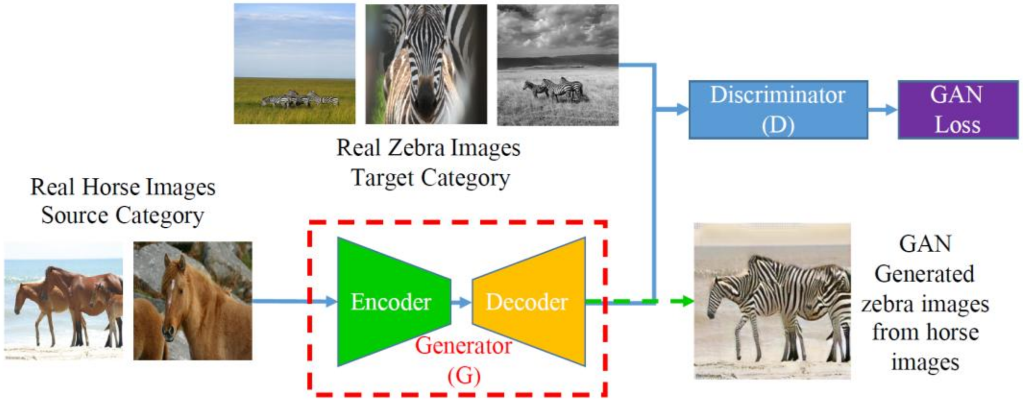
- A typical way to design a real vs. GAN fake image classifier:
  - Collect a large number of GAN generated images from one or multiple pre-trained GAN models.
  - Train a binary classifier.

- Limitation
  - Generalization can be a problem.
    - Commonly do not have access to the specific model used by the attacker.
    - The attack may design a unique and secret architecture.

- To improve generalization
  - Identify the key artifacts commonly shared by GAN.
  - Encourage classifiers to learn these artifacts.

Detecting and Simulating Artifacts in GAN Fake Images

- A typical generator contains an encoder and a decoder

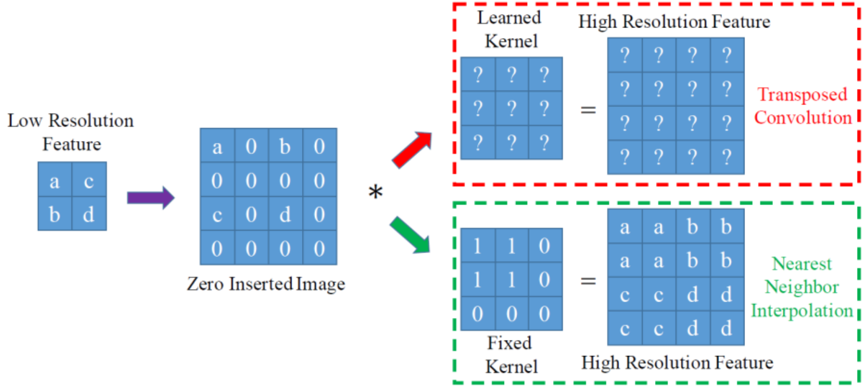


- The encoder contains a few down-sampling layers.
  - Extract high-level information from the input image and generate a low-resolution feature.
- The decoder contains a few up-sampling layers.
  - Take the low-resolution feature as input and output a high-resolution image.
  - A discriminator is trained to distinguish between real and fake images.

Zhang, X., Karaman, S. and Chang, S.F., 2019, December. Detecting and simulating artifacts in gan fake images. In 2019 IEEE international workshop on information forensics and security (WIFS) (pp. 1-6). IEEE.

Detecting and Simulating Artifacts in GAN Fake Images

- Up-sampling artifacts are introduced by generators in GAN

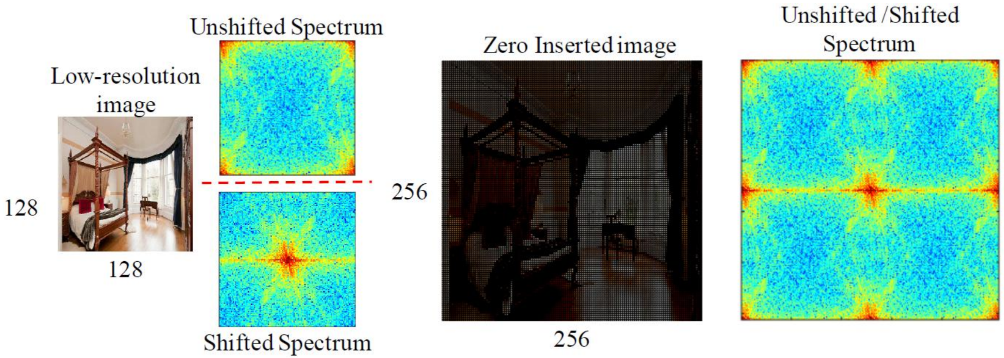


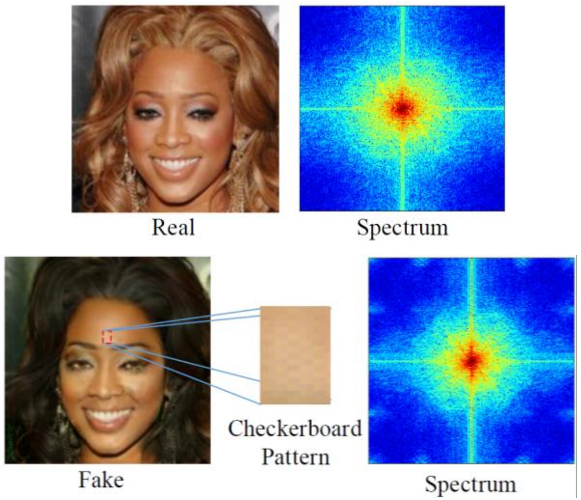
- Up-sampling modules used in different GAN models are consistent.
  - Transposed convolution (a.k.a deconvolution)
  - Nearest neighbor interpolation
- Both up-samplers can be formulated as a simple pipeline.
  - The up-sampler increases both the horizontal and vertical resolutions by a factor of m, e.g., 2.
    - The up-sampler inserts one zero row/column after each row/column in the low-resolution feature tensor.
  - Applies a convolution operation to assign appropriate values to the “zero-inserted” locations.
    - The convolution kernel in transposed convolution is learnable.
    - It is fixed in the nearest neighbor interpolation.

Zhang, X., Karaman, S. and Chang, S.F., 2019, December. Detecting and simulating artifacts in gan fake images. In 2019 IEEE international workshop on information forensics and security (WIFS) (pp. 1-6). IEEE.

Detecting and Simulating Artifacts in GAN Fake Images

- Up-sampling artifacts



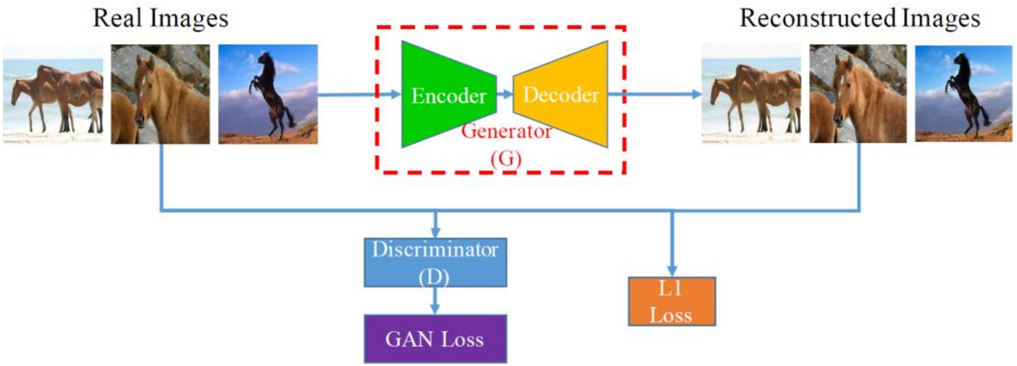


- In the frequency domain
  - Low frequency components are shifted to the center of the spectrum for better visualization.
  - Inserting zeros in the low-resolution image is equivalent to replicating multiple copies of the spectrum of the original low-resolution image over the high frequency part of the spectrum of the final high-resolution image.
    - Can be mathematically proved.
  - Such artifacts can be identified in the generated images.
    - Bright "blobs" at ¼ and ¾ of the width/height.

Zhang, X., Karaman, S. and Chang, S.F., 2019, December. Detecting and simulating artifacts in gan fake images. In 2019 IEEE international workshop on information forensics and security (WIFS) (pp. 1-6). IEEE.

Detecting and Simulating Artifacts in GAN Fake Images

- Exploit the GAN artifact for deepfake detection: **AutoGAN**

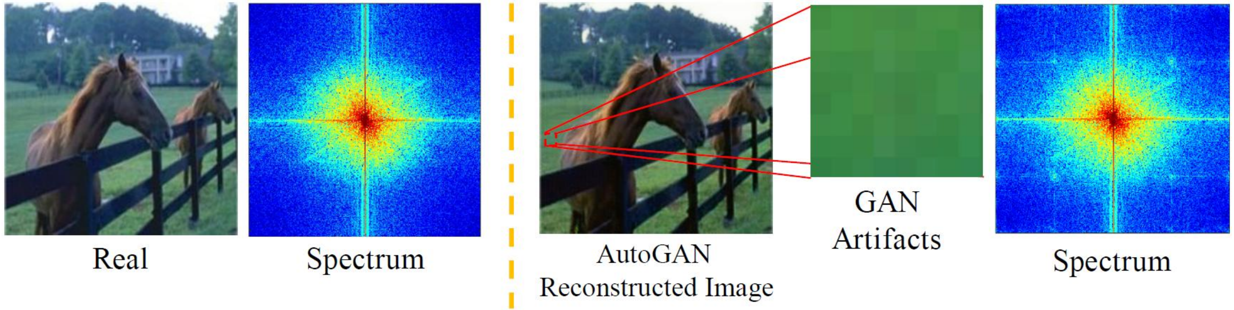


- Simulate the common generation pipeline shared by popular GAN models.
  - Synthesize GAN artifacts in any image via a **dummy generator**.
    - No need to access any pre-trained GAN model.
  - The decoder contains a widely used up-sampling module, such as **transposed convolution** or **nearest neighbor interpolation**.
  - The output of the generator aims to be the same as the original input.
    - Fooling the discriminator.
    - Minimizing the L_1 norm between input and output.

Zhang, X., Karaman, S. and Chang, S.F., 2019, December. Detecting and simulating artifacts in gan fake images. In 2019 IEEE international workshop on information forensics and security (WIFS) (pp. 1-6). IEEE.

Detecting and Simulating Artifacts in GAN Fake Images

- Exploit the GAN artifact for deepfake detection: **AutoGAN**



- The **dummy generator** successfully injects GAN artifacts into any image.
  - The reconstructed images are used to train classifiers.

Zhang, X., Karaman, S. and Chang, S.F., 2019, December. Detecting and simulating artifacts in gan fake images. In 2019 IEEE international workshop on information forensics and security (WIFS) (pp. 1-6). IEEE.

Detecting and Simulating Artifacts in GAN Fake Images

- Experiment setup
  - Img
    - Learned with real images and fake images generated by cycleGAN.
    - For example, real horse images and fake horse images generated from zebra images.
  - Spec
    - The training data is the same as Img.
    - The classifier is trained with the spectrum input.
  - A-Img
    - Learned with real image and fake image generated by injecting GAN artifacts.
    - For example, real horse images and reconstructed horse images.
  - A-Spec
    - The training data is the same as A-Img.
    - The classifier is trained with the spectrum input.

Zhang, X., Karaman, S. and Chang, S.F., 2019, December. Detecting and simulating artifacts in gan fake images. In 2019 IEEE international workshop on information forensics and security (WIFS) (pp. 1-6). IEEE.

Detecting and Simulating Artifacts in GAN Fake Images

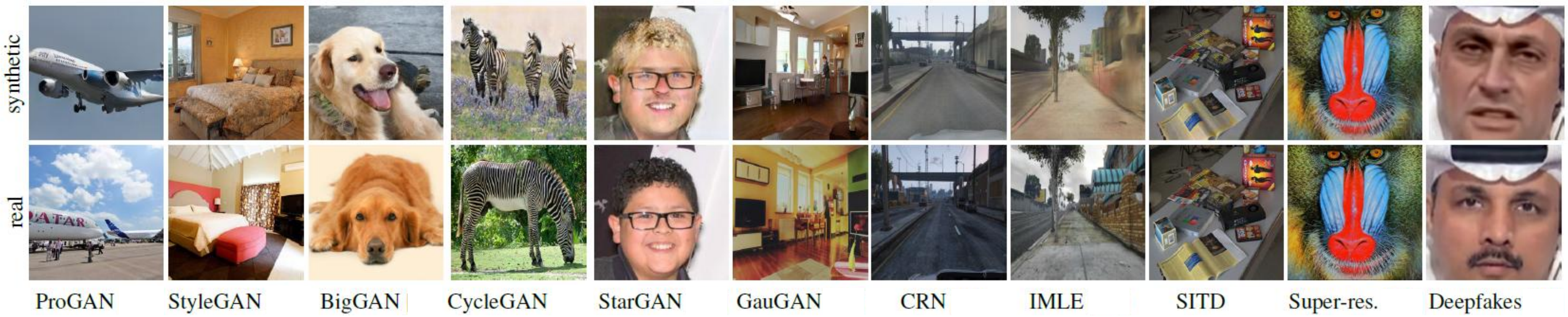
- Results

Training	Feature	H	Z	S	W	A	O	F	City	Map	U	V	C	M	P	Ave.
Horse	Img	99.2	78.5	96.2	86.3	78.0	70.6	75.5	72.2	55.9	61.5	95.0	87.0	87.7	93.6	81.2
	Spec	100	99.6	99.8	85.0	99.4	99.8	98.6	96.7	50.0	96.3	83.1	99.4	93.1	99.2	92.9
	A Img	91.9	72.7	87.9	77.3	83.7	89.1	52.4	50.4	57.7	29.0	61.9	34.9	36.9	89.0	65.3
	A Spec	98.1	98.1	99.3	88.7	99.6	100	100	96.0	63.5	99.2	86.2	99.1	88.1	100	**94.0**
Zebra	Img	68.8	96.5	78.8	63.4	54.1	50.2	50.0	50.0	50.0	39.5	87.2	45.4	80.3	87.6	64.4
	Spec	98.1	100	92.5	74.6	97.5	97.1	100	93.9	50.0	91.1	53.4	91.0	55.5	98.1	**85.2**
	A Img	93.8	92.7	82.6	84.1	79.4	82.1	50.0	50.0	51.4	38.5	75.9	49.4	57.5	87.0	69.6
	A Spec	76.9	88.8	94.7	52.1	81.5	77.6	99.5	80.4	55.9	97.2	60.6	97.8	61.2	99.0	80.2
COCO	A Img	77.3	78.8	58.7	75.3	64.8	69.5	100.0	99.9	73.4	88.9	89.0	97.4	91.5	37.7	78.7
	A Spec	90.4	90.4	83.7	85.2	94.0	93.8	99.5	100.0	88.9	97.8	91.6	98.6	96.4	81.8	**92.3**

- The classifier trained with images (Img & A-Img) struggles to generalize well.
  - It achieves good performance in the same category.
- The spectrum-based classifier (Spec & A Spec) greatly improves the generalization.
  - The “A Spec” classifier has never seen any cycleGAN images during training.
- Using a diverse image dataset, e.g., MSCOCO, can result in overall good performance.

Zhang, X., Karaman, S. and Chang, S.F., 2019, December. Detecting and simulating artifacts in gan fake images. In 2019 IEEE international workshop on information forensics and security (WIFS) (pp. 1-6). IEEE.

Spotting Artifacts in CNN-generated Images



- Are there common features or artifacts shared across diverse CNN generators?
    - A broad set of generative techniques use convolutional neural networks (CNNs).
    - Detecting whether an image was generated by a specific synthesis technique is relatively simple.
        - High accuracy is easily achieved.
    - Existence of common artifacts allows a classifier to generalize to an entire family of generation methods.
        - Rather than a single one, e.g., CycleGAN

Wang, S.Y., Wang, O., Zhang, R., Owens, A. and Efros, A.A., 2020. CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8695-8704).

Spotting Artifacts in CNN-generated Images

- Use one specific model, ProGAN, to train the detector on.
  - Evaluate generalization to other CNN models.
  - ProGAN generates high quality images with a simple CNN structure.
  - Create a large-scale dataset with only ProGAN-generated images and real images.
    - 720K images for training + 4K images for validation.
    - Equal numbers of real and fake images.
  - Data augmentation is applied during training.
    - E.g., Gaussian blur, JPEG, and horizontal flipping.
  - Metric: average precision (AP)
    - Use it as a black-box metric.
    - Range between 0 and 1.
      - The larger the better.
      - A perfect model has an AP score of 1.

Wang, S.Y., Wang, O., Zhang, R., Owens, A. and Efros, A.A., 2020. CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8695-8704).

Spotting Artifacts in CNN-generated Images

- Results (Symbols ✓ and † mean the augmentation is applied with 50% or 10% probability)

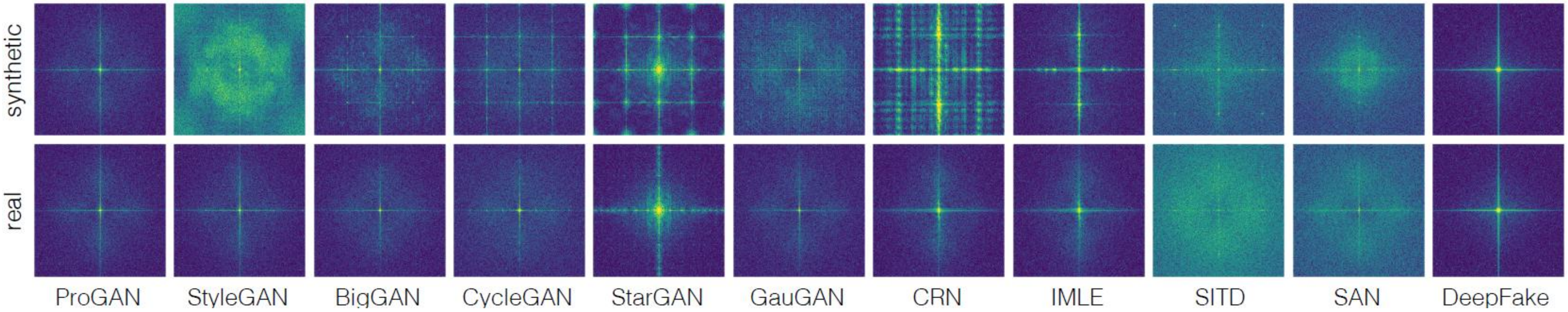
Family	Name	Training settings: Train	Input	No. Class	Augments: Blur	Augments: JPEG	Individual test generators: Pro-GAN	Style-GAN	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN	CRN	IMLE	SITD	SAN	Deep-Fake	Total: mAP
Simulating Artifacts	Cyc-Im	CycleGAN	RGB	–			84.3	65.7	55.1	100.	99.2	79.9	74.5	90.6	67.8	82.9	53.2	77.6
	Cyc-Spec	CycleGAN	Spec	–			51.4	52.7	79.6	100.	**100.**	70.8	64.7	71.3	92.2	78.5	44.5	73.2
	Auto-Im	AutoGAN	RGB	–			73.8	60.1	46.1	99.9	**100.**	49.0	82.5	71.0	80.1	86.7	80.8	75.5
	Auto-Spec	AutoGAN	Spec	–			75.6	68.6	84.9	**100.**	**100.**	61.0	80.8	75.3	89.9	66.1	39.0	76.5
Common Artifacts	2-class	ProGAN	RGB	2	✓	✓	98.8	78.3	66.4	88.7	87.3	87.4	94.0	97.3	85.2	52.9	58.1	81.3
	4-class	ProGAN	RGB	4	✓	✓	99.8	87.0	74.0	93.2	92.3	94.1	95.8	97.5	87.8	58.5	59.6	85.4
	8-class	ProGAN	RGB	8	✓	✓	99.9	94.2	78.9	94.3	91.9	95.4	98.9	99.4	91.2	58.6	63.8	87.9
	16-class	ProGAN	RGB	16	✓	✓	100.	98.2	87.7	96.4	95.5	**98.1**	**99.0**	**99.7**	95.3	63.1	71.9	91.4
	No aug	ProGAN	RGB	20			100.	96.3	72.2	84.0	**100.**	67.0	93.5	90.3	96.2	**93.6**	**98.2**	90.1
	Blur only	ProGAN	RGB	20	✓		100.	99.0	82.5	90.1	**100.**	74.7	66.6	66.7	**99.6**	53.7	95.1	84.4
	JPEG only	ProGAN	RGB	20		✓	100.	99.0	87.8	93.2	91.8	97.5	**99.0**	99.5	88.7	78.1	88.1	**93.0**
	Blur+JPEG (0.5)	ProGAN	RGB	20	✓	✓	100.	98.5	**88.2**	96.8	95.4	**98.1**	98.9	99.5	92.7	63.9	66.3	90.8
	Blur+JPEG (0.1)	ProGAN	RGB	20	†	†	100.	**99.6**	84.5	93.5	98.2	89.5	98.2	98.4	97.2	70.5	89.0	92.6

- Better performance than AutoGAN.
- Including more classes for training improves performance
- Augmentation improves generalization.
  - Except for "Blur only"

Wang, S.Y., Wang, O., Zhang, R., Owens, A. and Efros, A.A., 2020. CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8695-8704).

Spotting Artifacts in CNN-generated Images

- Visualize the average frequency spectra from each dataset



- The real image spectra generally look alike.
  - With minor variations due to differences in the datasets.
- Periodic patterns (dots or lines) in most of the synthetic images.
  - BigGAN and ProGAN contain relatively few such artifacts.
  - DeepFake images do not contain obvious artifacts.
    - DeepFake images have gone through various pre- and post-processing.
    - Synthesized face region is resized, blended, and compressed with MPEG.

Wang, S.Y., Wang, O., Zhang, R., Owens, A. and Efros, A.A., 2020. CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8695-8704).

Spotting Artifacts in CNN-generated Images

- Although artifacts can be exploited for detection, there are limitations.
  - Adversaries can modify fake images to bypass detection.
    - There are always novel modifications, e.g., Photoshop, that can bypass detection.
  - The rapid advance in GAN will eventually invalidate detection methods.
    - Nash Equilibrium
  - Detection results tend to lack explainability.
    - Prevent classifiers from being supported by the law.

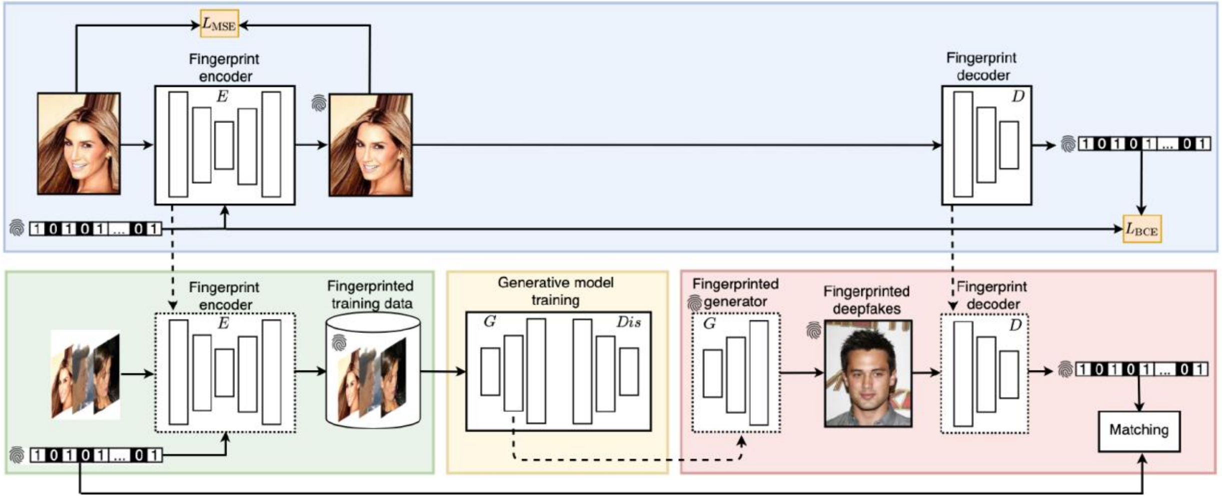
Wang, S.Y., Wang, O., Zhang, R., Owens, A. and Efros, A.A., 2020. CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8695-8704).

Proactive Detection - Watermarking

- Embedding watermarks into generated images.
  - Proactively detect fake images instead of reactively.
    - Embed watermarks into generative models.
      - Generated images still contain predefined watermarks.
    - Watermarks are artifacts deliberately introduced to deepfake images.
      - Different from artifacts that naturally exist because of the generative techniques.
      - Analogous to adversarial examples and backdoor attacks.
  - Overcome the limitations of detecting artifacts in fake images.
    - Robustness
      - Defenders can make watermarks robust
        - Introduce transformations during the training of watermarking models.
        - Defenders have no control over natural artifacts.
    - Nash Equilibrium is beneficial for defenders
      - The roles of defenders and adversaries exchange.
      - Nash Equilibrium is beneficial for defenders
        - In the end, watermarks will be unremovable.
    - Explainability
      - The existence of watermarks explains detection results.

Watermarking Generative Models

- Embedding watermarks into generative models



  - Train an encoder and a decoder to embed watermarks.
    - The encoder embeds watermarks into the training data.
    - The decoder recovers watermarks from the input.
  - Train a generative model on the watermarked data.
    - Generated images will also contain the same watermark.
    - Model inventors are encouraged to embed watermarks into their models before releasing models to the public.
    - Cannot work if adversaries train their own models.

Yu, N., Skripniuk, V., Abdelnabi, S. and Fritz, M., 2021. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In Proceedings of the IEEE/CVF International conference on computer vision (pp. 14448-14457).

Watermarking Generative Models

- Detecting deepfake images is then a trivial task.
  - Images with the same or very similar watermarks are detected as fake.
  - Images with random watermarks are detected as real.
- Training loss:

$$\min_{E,D} \mathbb{E}_{\tilde{\mathbf{x}} \sim \tilde{\mathbb{X}}, \mathbf{w} \sim \{0,1\}^n} L_{\text{BCE}}(\tilde{\mathbf{x}}, \mathbf{w}; E, D) + \lambda L_{\text{MSE}}(\tilde{\mathbf{x}}, \mathbf{w}; E)$$

$$\mathrm{L}_{\text{BCE}}(\tilde{\mathbf{x}}, \mathbf{w}; E, D) = \frac{1}{n} \sum_{k=1}^{n} \left(\mathbf{w}_k \log \hat{\mathbf{w}}_k + (1 - \mathbf{w}_k) \log(1 - \hat{\mathbf{w}}_k) \right)$$

$$L_{\text{MSE}}(\tilde{\mathbf{x}}, \mathbf{w}; E) = ||E(\tilde{\mathbf{x}}, \mathbf{w}) - \tilde{\mathbf{x}}||_2^2$$

$$\hat{\mathbf{w}} = D\big(E(\tilde{\mathbf{x}}, \mathbf{w})\big)$$

  - E and D are the encoder and decoder, respectively.
  - w represents the watermark to be embedded into the training data $\tilde{X}$.
  - BCE: binary cross entropy; MSE: measures squared error.

Yu, N., Skripniuk, V., Abdelnabi, S. and Fritz, M., 2021. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In Proceedings of the IEEE/CVF International conference on computer vision (pp. 14448-14457).

Watermarking Generative Models

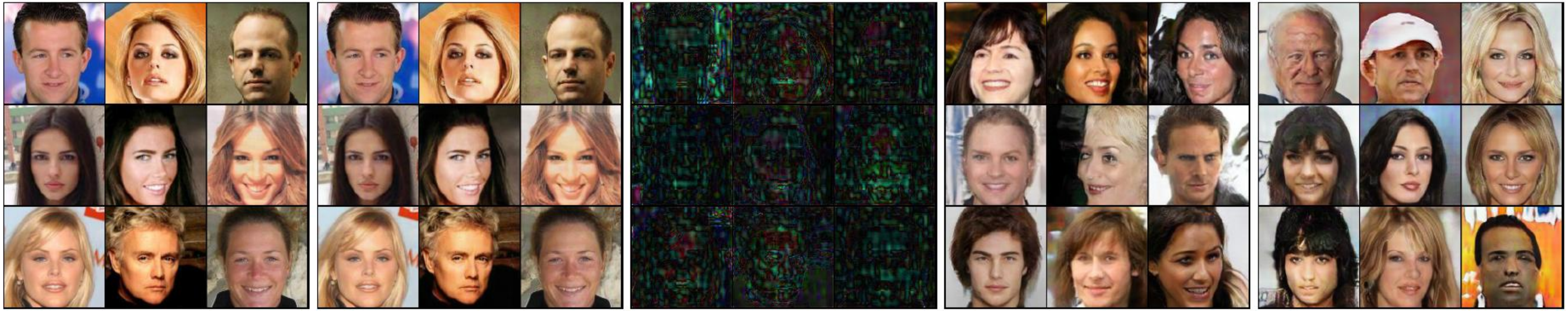
- Results
  - Conventional watermarking methods (denoted as "non-deep watermarks") do not transfer hidden information into generative models
    - Indicated by the random guess performance during decoding.
    - The transferability watermarks is non-trivial.
  - Deep learning-based watermarks can be almost perfectly detected from generated images
    - Results are consistent across a variety of models and datasets.
  - Quality of watermarked images
    - Frechet Inception Distance (FID) evaluates the generation quality: the lower, the more realistic.
    - Watermarked generative models are similar with the original baselines in terms of FID.

Dataset		Model	Bit acc $\Uparrow$	p-value	Orig FID	Fgpt FID $\Downarrow$
CelebA	non-deep watermarks	StyleGAN2	0.51	0.46	6.41	6.93
		StyleGAN2	0.53	0.31	6.41	6.82
		Data	1.00	-	-	1.15
		ProGAN	0.98	$< 10^{-26}$	14.09	14.38
		StyleGAN	0.99	$< 10^{-28}$	8.98	9.72
		StyleGAN2	0.99	$< 10^{-28}$	6.41	6.23
LSUN *Bedroom*		ProGAN	0.93	$< 10^{-19}$	29.16	32.58
		StyleGAN	0.98	$< 10^{-26}$	24.95	25.71
		StyleGAN2	0.99	$< 10^{-28}$	13.92	14.71
LSUN *Cat*		ProGAN	0.98	$< 10^{-26}$	45.22	48.97
		StyleGAN	0.99	$< 10^{-28}$	33.45	34.01
		StyleGAN2	0.99	$< 10^{-28}$	31.01	32.60
CIFAR-10		BigGAN	0.99	$< 10^{-28}$	6.25	6.80
Horse→Zebra		CUT	0.99	$< 10^{-28}$	22.98	23.43
Cat→Dog		CUT	0.99	$< 10^{-28}$	55.78	56.09

Yu, N., Skripniuk, V., Abdelnabi, S. and Fritz, M., 2021. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In Proceedings of the IEEE/CVF International conference on computer vision (pp. 14448-14457).

Watermarking Generative Models

- Visualization



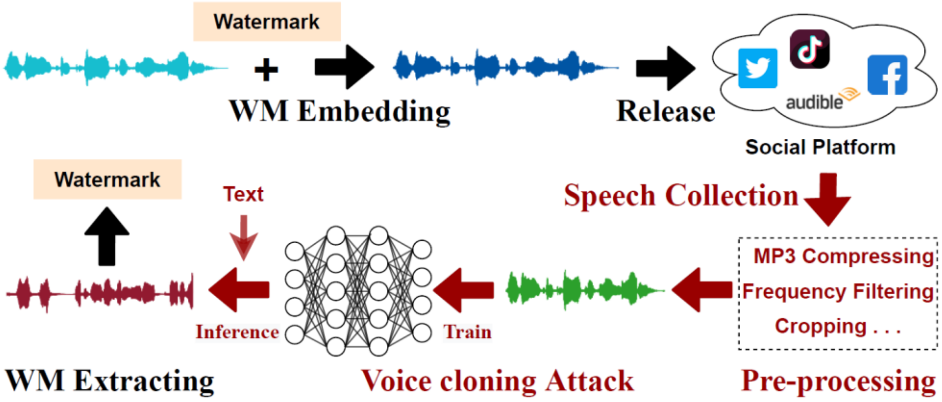
(a) (b) (c) (d) (e)

- (a) Original real training samples.
- (b) Fingerprinted real training samples.
- (c) The difference between (a) and (b)
    - 10× magnified for easier visualization.
- (d) Samples from the non-watermarked ProGAN.
- (e) Samples from the watermarked ProGAN.

Yu, N., Skripniuk, V., Abdelnabi, S. and Fritz, M., 2021. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In Proceedings of the IEEE/CVF International conference on computer vision (pp. 14448-14457).

Watermarking Voice Cloning Models

- Similar ideas also work for detecting **deepfake speech** generated by voice cloning.
  - Voice cloning refers to the process of creating a synthetic voice that closely resembles the voice of a target person.
    - Voice conversion and text-to-speech (TTS).
  - **Speaker adaptation** is a key component in voice cloning.
    - Fine-tuning a voice cloning model on target voice.
    - Collecting 1-minute voice is usually sufficient.
    - Provide chance to let voice cloning models learn watermarks.
      - Even if adversaries train their own voice cloning models, fine-tuning results in learning watermarks.

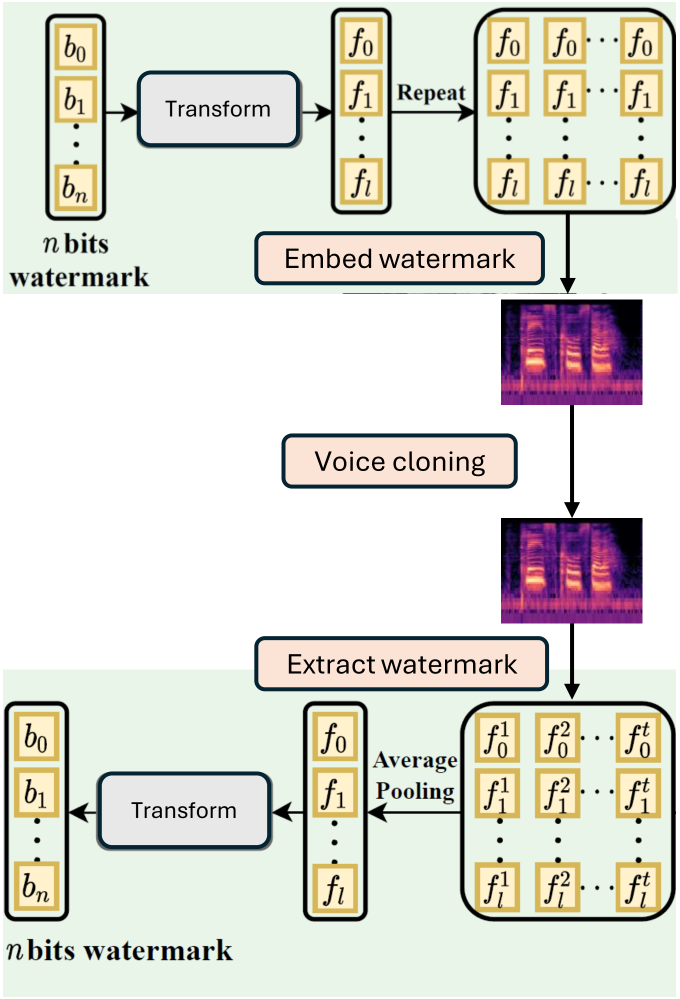


Liu, C., Zhang, J., Zhang, T., Yang, X., Zhang, W. and Yu, N., 2023. Detecting Voice Cloning Attacks via Timbre Watermarking. arXiv preprint arXiv:2312.03410.

Watermarking Voice Cloning Models

- Watermarks need to be **independent** on speech length.
  - Deepfake speech is different from deepfake images.
    - The length of deepfake speech varies.
    - Sizes of deepfake images are normally fixed.

  - One approach:
    - Repeat watermarks along the time axis.
    - Extract watermarks along the time axis and then do average pooling.
      - i.e., calculating averaged values.



Liu, C., Zhang, J., Zhang, T., Yang, X., Zhang, W. and Yu, N., 2023. Detecting Voice Cloning Attacks via Timbre Watermarking. arXiv preprint arXiv:2312.03410.

Watermarking Voice Cloning Models

- Transfer watermarks to voice cloning models when speaker adaptation is applied.

Service	Language	Metric	Speaker					
PaddleSpeech	English		P225	P226	P227	P228	P229	P230
		PESQ↑	2.5958	2.7235	2.3573	2.3235	2.7419	1.7095
		SECS↑	0.8611	0.8701	0.8552	0.8537	0.8592	0.8519
		ACC↑	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Chinese		D4	D6	D7	D8	D11	D12
		PESQ↑	1.7642	1.9851	2.6490	2.0223	2.3808	1.2313
		SECS↑	0.7836	0.8034	0.7622	0.8219	0.7304	0.7103
		ACC↑	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Voice-Cloning-App	English		P225	P226	P227	P228	P229	P230
		PESQ↑	0.7809	1.5610	1.1913	1.1684	1.2601	1.2694
		SECS↑	0.7576	0.8564	0.7324	0.8781	0.8495	0.8799
		ACC↑	0.9000	0.9100	0.9000	0.9000	0.9500	0.9200

- Metrics (the larger the better)
  - Perceptual Evaluation of Speech Quality (PESQ)
    - Measure the quality of speech.
  - Speaker Encoder Cosine Similarity (SECS)
    - Measure the identify of speech.
  - Watermark bit recovery accuracy (ACC)
    - Percentage of watermark bits recovered.

Liu, C., Zhang, J., Zhang, T., Yang, X., Zhang, W. and Yu, N., 2023. Detecting Voice Cloning Attacks via Timbre Watermarking. arXiv preprint arXiv:2312.03410.

Watermarking

- The arms race between defenders and adversaries continues.
  - Defenders want watermarks to be unremovable.
    - Adversaries want to remove watermarks while preserving the original signal.

  - Defenders want watermarks to be imperceptible.
    - This is feasible.
      - We have discussed a technique for images.
    - Robustness in this case is **arguably unachievable** if watermarks are *additive*.
      - Watermarks are unrelated to human perception.
      - Imperceptibility means watermarks can be theoretically removed without affecting the original signal.
        - E.g., noise reduction.
      - Novel non-additive watermarks need to be invented.

Liu, C., Zhang, J., Zhang, T., Yang, X., Zhang, W. and Yu, N., 2023. Detecting Voice Cloning Attacks via Timbre Watermarking. arXiv preprint arXiv:2312.03410.

References

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. Advances in neural information processing systems, 27.
- Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S. and Sun, L., 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. arXiv preprint arXiv:2303.04226.
- Zhang, X., Karaman, S. and Chang, S.F., 2019, December. Detecting and simulating artifacts in gan fake images. In 2019 IEEE international workshop on information forensics and security (WIFS) (pp. 1-6). IEEE.
- Wang, S.Y., Wang, O., Zhang, R., Owens, A. and Efros, A.A., 2020. CNN-generated images are surprisingly easy to spot… for now. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8695-8704).
- Yu, N., Skripniuk, V., Abdelnabi, S. and Fritz, M., 2021. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In Proceedings of the IEEE/CVF International conference on computer vision (pp. 14448-14457).
- Liu, C., Zhang, J., Zhang, T., Yang, X., Zhang, W. and Yu, N., 2023. Detecting Voice Cloning Attacks via Timbre Watermarking. arXiv preprint arXiv:2312.03410.