Current Trends in AI for Cybersecurity (Deepfake Detection)

CSIT375 AI for Cybersecurity

Dr Wei Zong

SCIT University of Wollongong

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

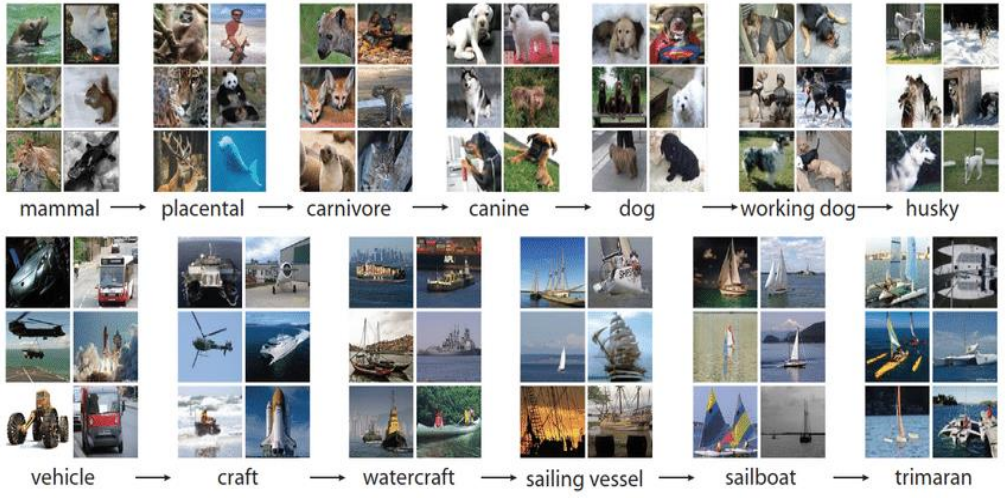
Outline

- GAN
- Security threats posed by GAN and countermeasures
- Beyond GAN

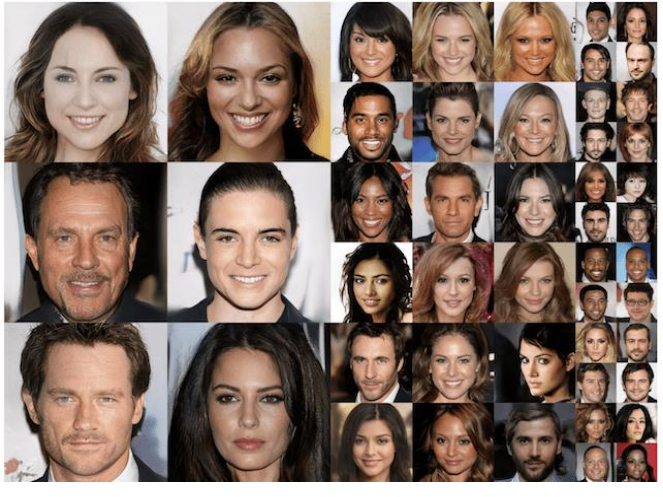
Intro

Deep learning:

- Discriminative models
  - Classifies input



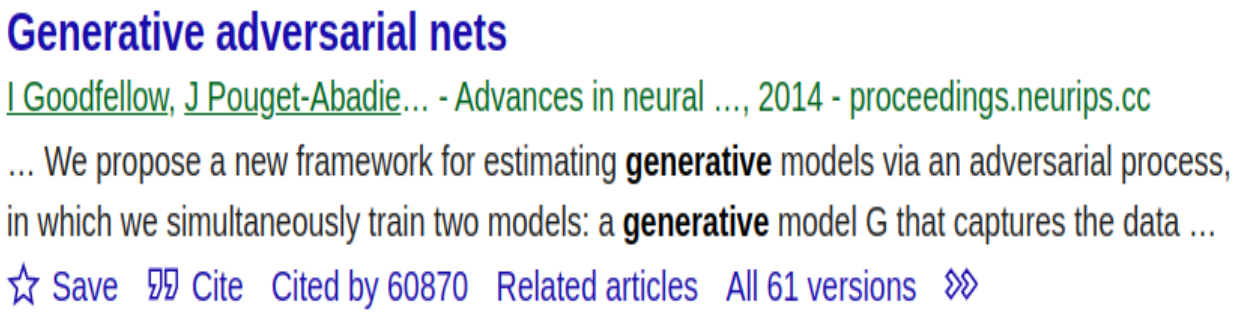
- Generative models
  - Learns distribution of data



Intro

GAN

- has recently become a hot research topic



  - 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

FOLLOW

DeepMind
Verified email at deepmind.com - Homepage
Deep Learning

Cited by

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

Yoshua Bengio

FOLLOW

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

Cited by VIEW ALL

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

GAN

A gossip about GAN

Juergen Schmidhuber

FOLLOW

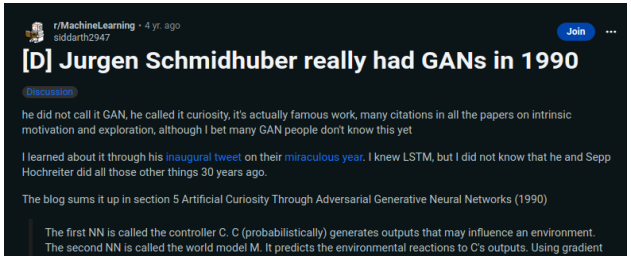
King Abdullah University of Science and Technology / The Swiss AI Lab, IDSIA / University of Lugano

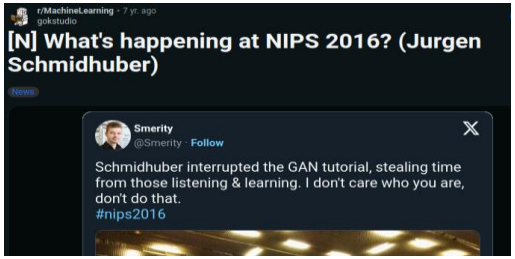
Verified email at kaust.edu.sa - Homepage

computer science artificial intelligence reinforcement learning neural networks physics

Cited by		VIEW ALL
	All	Since 2018
Citations	218083	174204
h-index	116	91
i10-index	417	272

Inventor of Long Short Term Memory (LSTM) model.





Mu Li

FOLLOW

Amazon

Verified email at amazon.com - Homepage

Machine Learning Distributed Systems

Cited by		VIEW ALL
	All	Since 2018
Citations	17708	16010
h-index	37	36
i10-index	64	60

“Techniques are re-invented all over the time. People give credit to those who make them popular.”

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

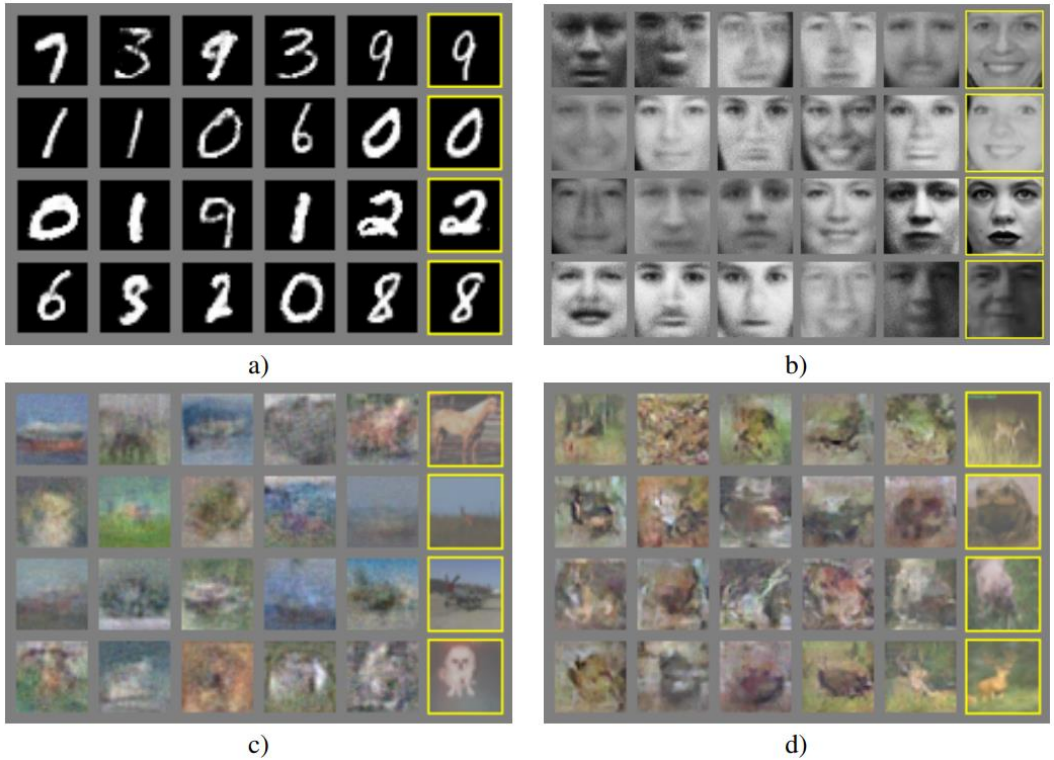
a team of counterfeiters

Discriminative Model

police

GAN

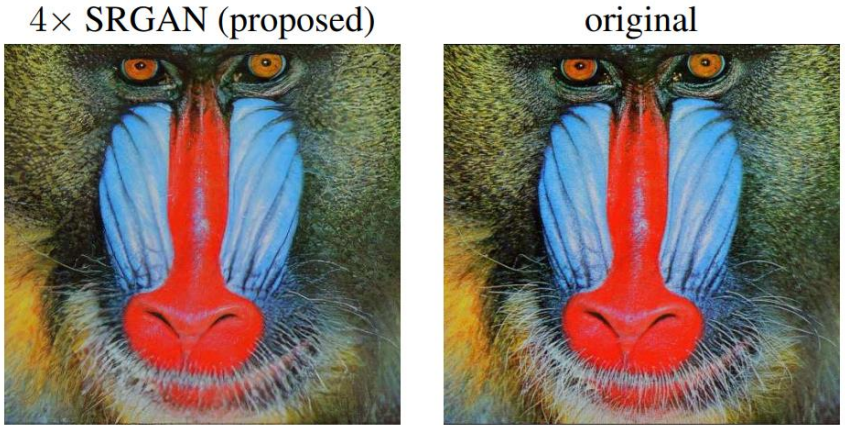
Results



Rapid Advance in GAN

Super-Resolution GAN

- a variant of GAN for image superresolution
- infer photo-realistic natural images for 4× upscaling factors
- is able to recover photo-realistic textures from heavily downsampled images



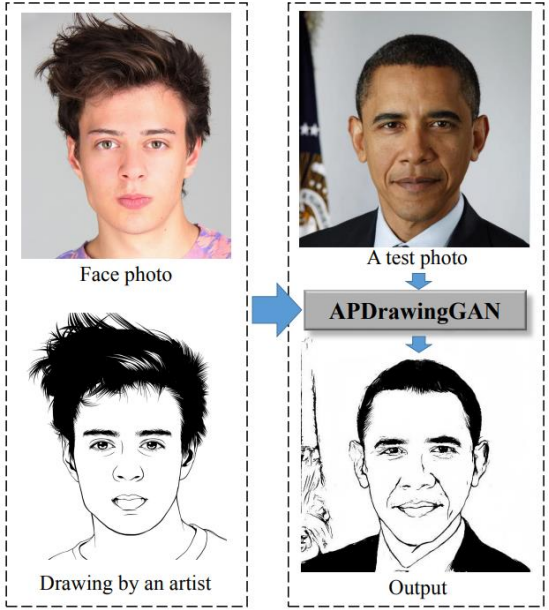
Super-resolved image (left) is almost indistinguishable from original (right). (4× upscaling)

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. and Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4681-4690).

Rapid Advance in GAN

APDrawingGAN

- Artistic portrait drawings have a highly abstract style, containing a sparse set of continuous graphical elements such as lines.
- Artists tend to use different strategies to draw different facial features and the lines drawn are only loosely related to obvious image features



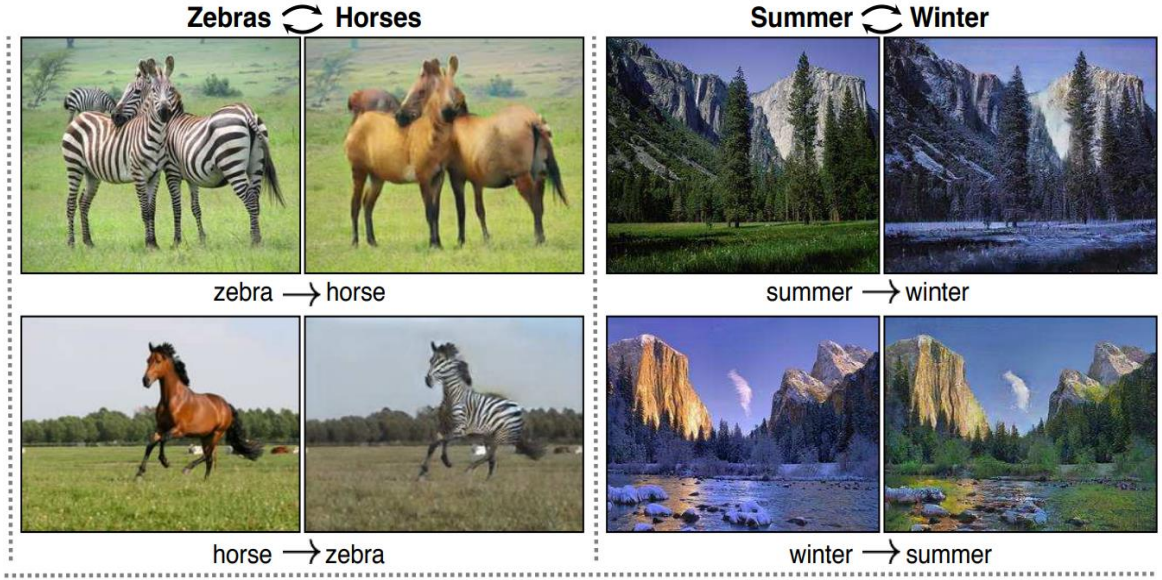
(left) An artist draws a portrait drawing using a sparse set of lines and very few shaded regions to capture the distinctive appearance of a given face photo.
(right) APDrawingGAN learns this artistic drawing style and automatically transforms a face photo into a high-quality artistic portrait drawing.

Yi, R., Liu, Y.J., Lai, Y.K. and Rosin, P.L., 2019. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10743-10752).

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.

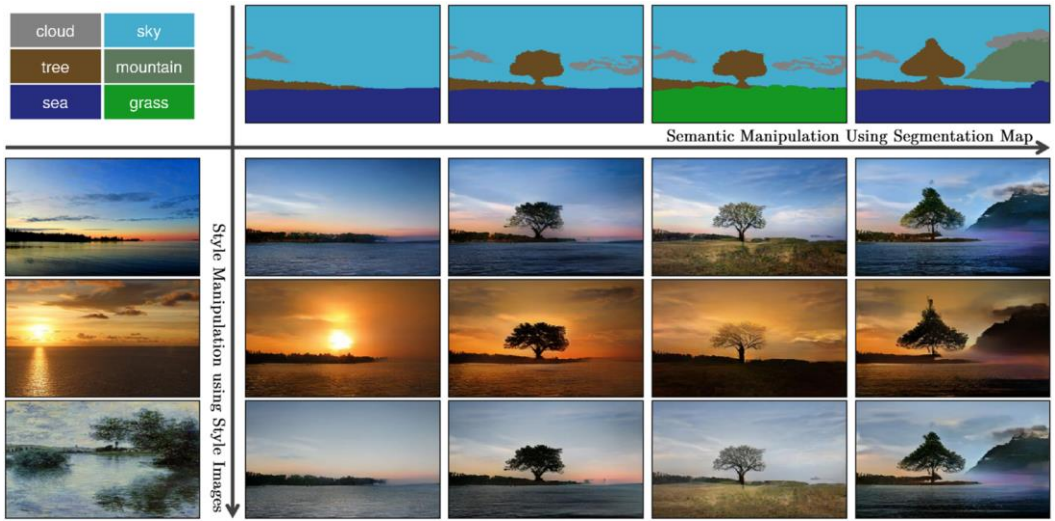


Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

Rapid Advance in GAN

GauGAN

- Allows user to control over both semantic and style as synthesizing an image
- The smentic (e.g., existence of a tree) is controlled via a label map (visualized in the top row).
- The style is controlled via the reference style image (visualized in the leftmost column)

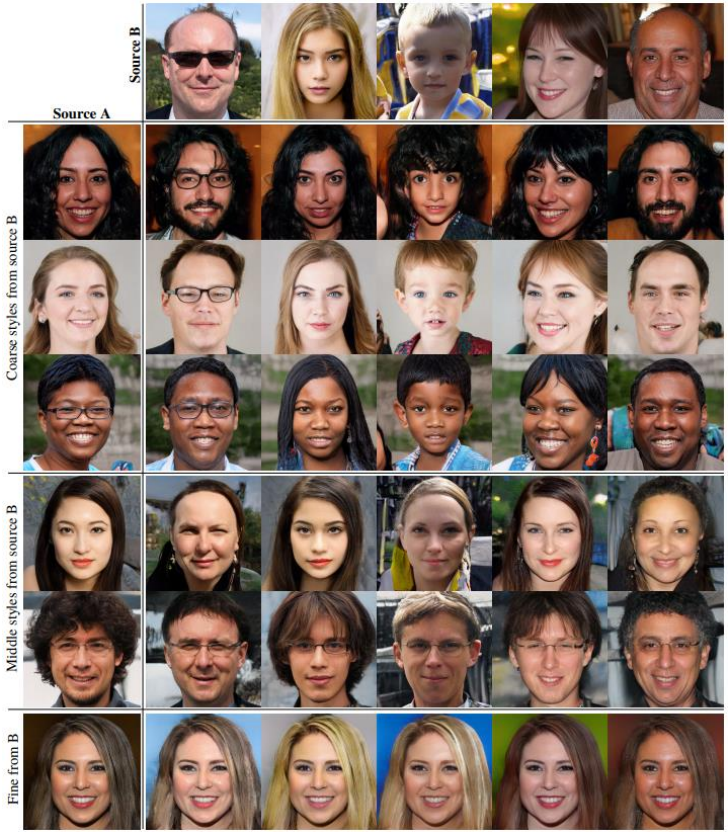


Park, T., Liu, M.Y., Wang, T.C. and Zhu, J.Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2337-2346).

Rapid Advance in GAN

StyleGan

- images generated by copying a specified subset of styles from source B and taking the rest from source A.
- Copying the styles corresponding to coarse spatial resolutions brings high-level aspects such as pose, general hair style, face shape, and eyeglasses from source B, while all colors (eyes, hair, lighting) and finer facial features resemble A.
- If instead copying the styles of middle resolutions from B, smaller scale facial features are inherited, hair style, eyes open/closed from B, while the pose, general face shape, and eyeglasses from A are preserved.
- Finally, copying the fine styles from B brings mainly the color scheme and microstructure.



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).

Fake Image Detection

GAN is double-edged sword

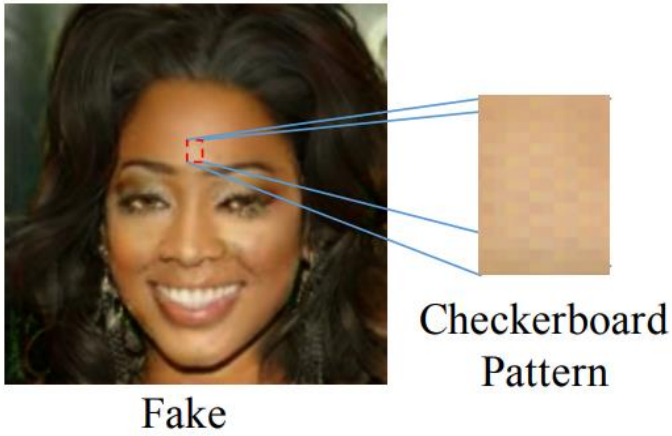
- Humans cannot distinguish between real or fake images anymore.
- Adversaries can use this technique to spread fake information or commit crimes.

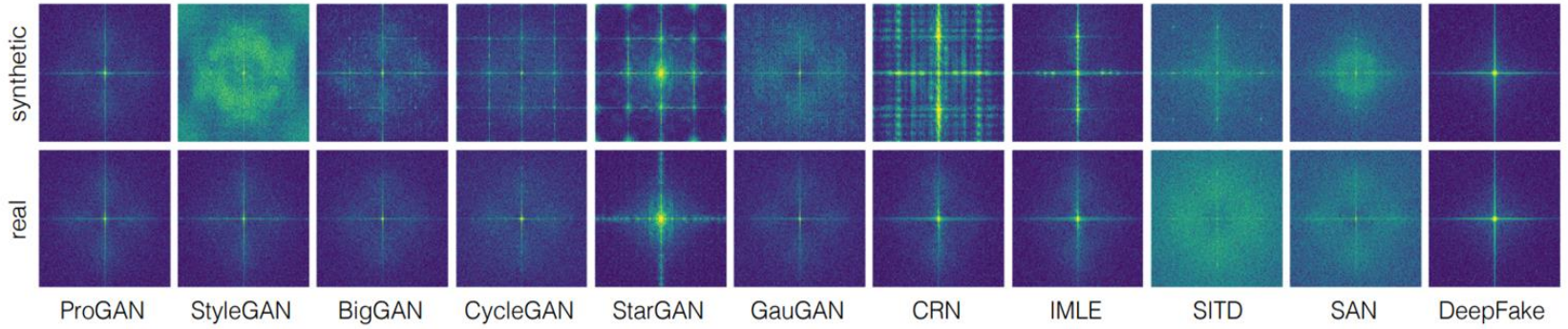
Is this a real or synthesized photo?

Fake Image Detection

Detecting fake images becomes an emerging research trend:

- Detect fake images based on artifacts introduced by deep generative models
- Checkerboard patterns identified in fake images.
- The figure below shows the average spectra of each high-pass filtered image, for both the real and fake images. There are periodic patterns (dots or lines) in most of the synthetic 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.
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).

Fake Image Detection

The first is the original fake image. The remainings are modified by adversaries to escape from detection.

However

- Adversaries can modify fake images to escape from detection.

Two limitations:

- The rapid advance in GAN will eventually invalidate detection methods.
  - Nash Equilibrium
- lack of explainability in model predictions
  - Prevent being used in law

The judicial demand for **explainable artificial intelligence**
A Deeks - Columbia **Law** Review, 2019 - JSTOR
... in shaping the nature and form of "**explainable AI**" (xAI). Using the tools of the common **law**, courts can develop what xAI should mean in different **legal** contexts. There are advantages to ...
☆ Save 99 Cite Cited by 198 Related articles All 10 versions

Explainable artificial intelligence, **lawyer's** perspective
Ł Górski, S Ramakrishna - ... Conference on **Artificial Intelligence** and **Law**, 2021 - dl.acm.org
... This work was thought as the first step towards the identification of requirements of **explainable AI**-based systems that would involve **legal** perspective to a greater extent. For the ...
☆ Save 99 Cite Cited by 25 Related articles

[PDF] **Explainable artificial intelligence** the new frontier in **legal** informatics
B Waltl, R Vogl - Jusletter IT, 2018 - wwwmatthes.in.tum.de
... This article explores the increasingly important topic of «**explainable AI**» and ... of **explainability** as a property inherent to machine learning algorithms. It highlights that **explainability** can ...
☆ Save 99 Cite Cited by 57 Related articles »

Regulating **Explainable AI** in the European Union. An Overview of the Current **Legal** Framework (s)
M Ebers - ... Overview of the Current **Legal** Framework (s)(August 9 ..., 2020 - papers.ssrn.com
...) arguments against full transparency of **Artificial Intelligence** (**AI**) systems, especially in the ... EU **law**, entitled to a right to explanation of automated decision-making, especially when **AI** ...
☆ Save 99 Cite Cited by 22 Related articles »

[HTML] **Explainable AI** under contract and tort **law**: **legal** incentives and technical challenges
P Hacker, R Krestel, S Grundmann... - **Artificial Intelligence** and ..., 2020 - Springer
... **explainability** is an important **legal** category not only in data protection **law**, but also in contract

Fake Image Detection

Watermarking

- overcome the limitations of detecting artifacts in fake images
- Proactively detect fake images instead of reactively
- Goal embedding watermarks into generative models
  - such that fake images still contain predefined watermarks
  - training set

(a) (b) (c)

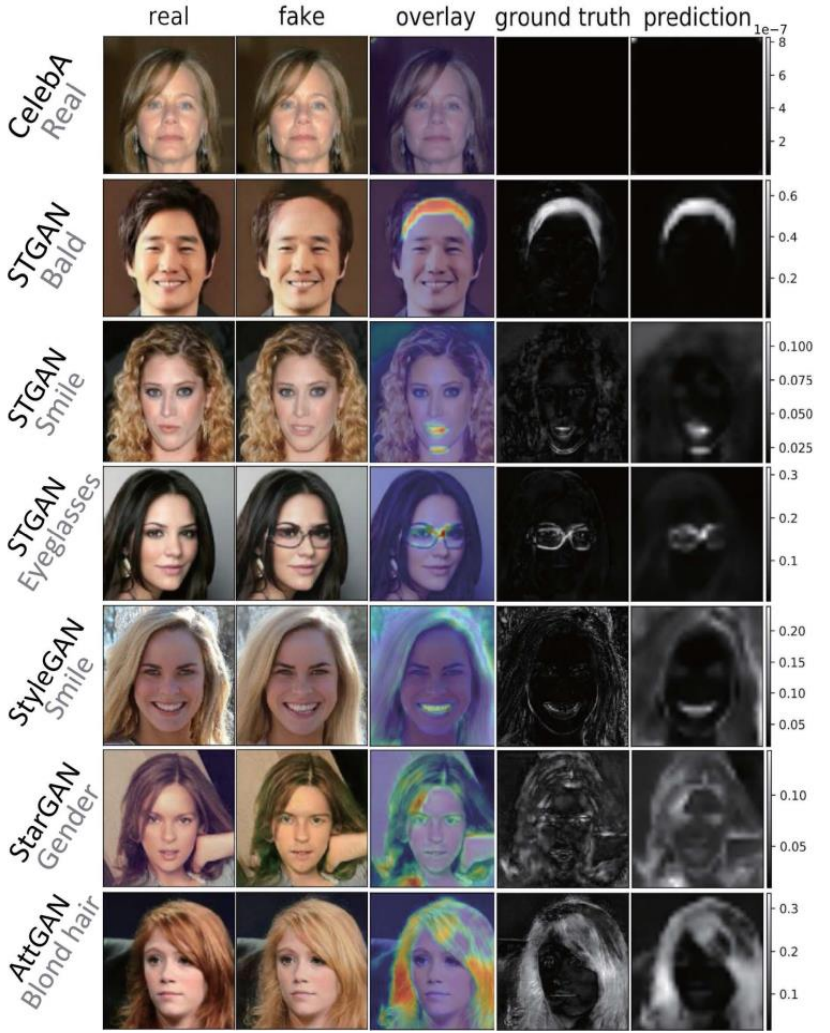
(a) Original real training samples. (b) Fingerprinted real training samples. (c) The difference between (a) and (b), 10× magnified for easier visualization.

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).

Fake Image Detection

Watermarking

- Embedding fragile watermarks into real images
  - If an adversary modifies parts of a watermarked image, the corresponding watermarks would be destroyed
  - defenders can easily locate modified parts.



Fake Image Detection

Watermarking

- The role of defenders and adversaries exchanges
  - Nash Equilibrium is beneficial for defenders
  - In the end, watermarks will be unremovable
- Explainable

The arms race between defenders and adversaries continues...

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

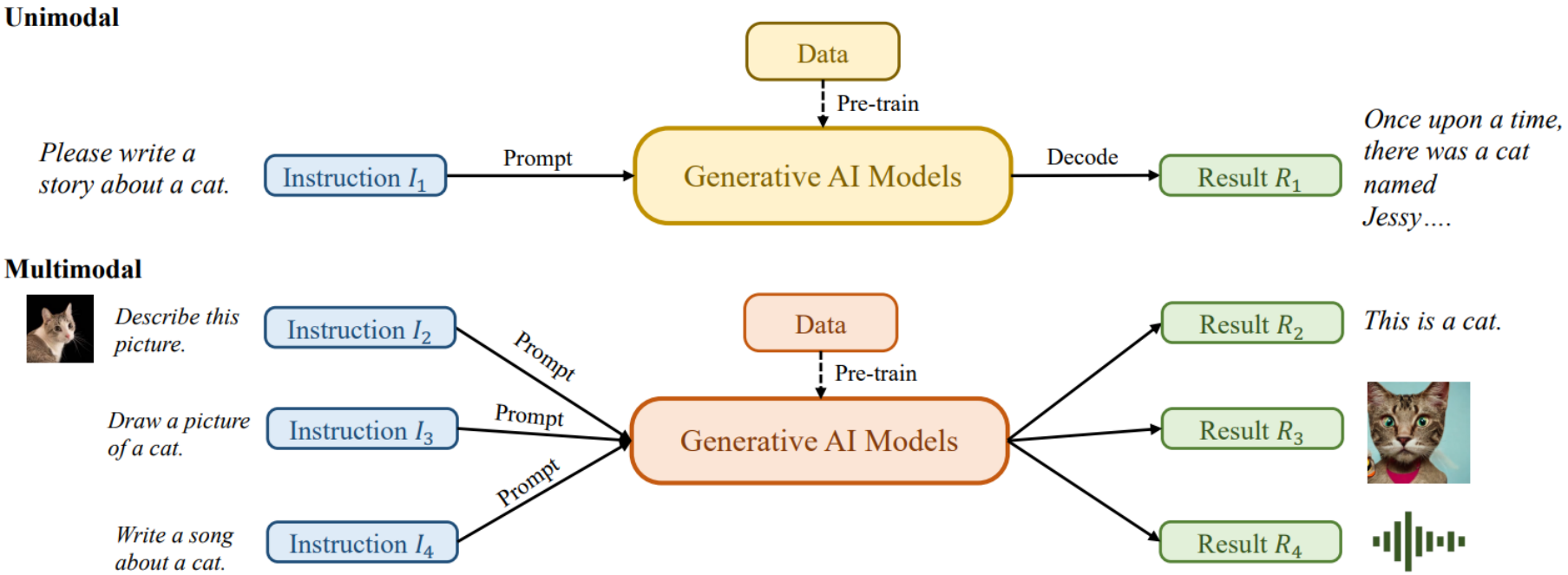
- A Flow is a distribution transformation from simple to complex by a sequence of **invertible** and differentiable mappings.

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

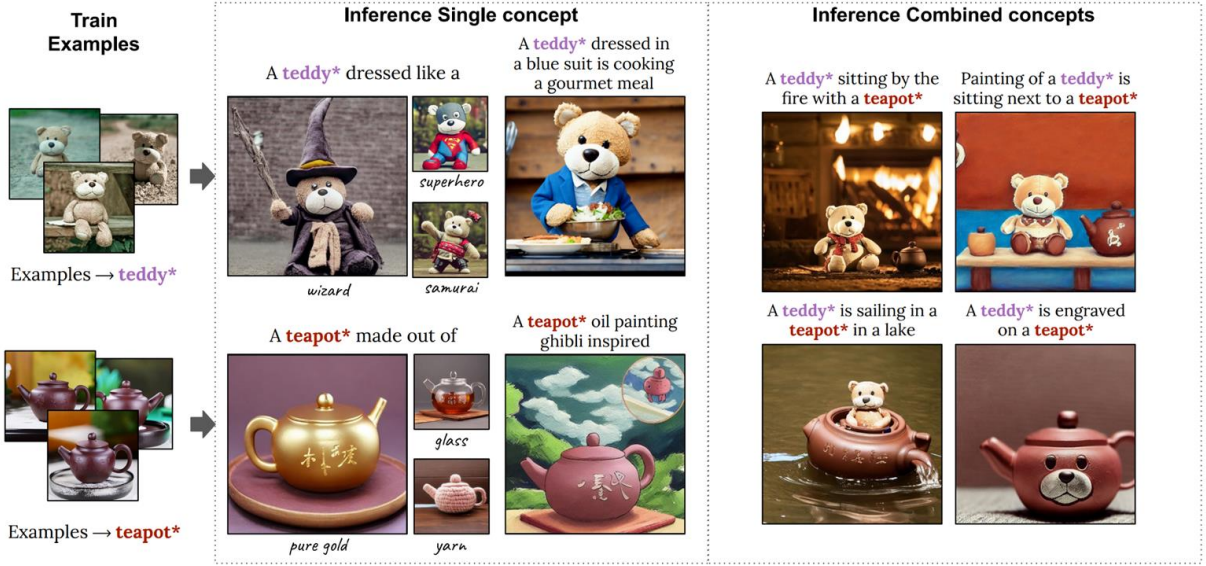
From Unimodal to MultiModel



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.



Questions?