



# Redes Neurais Adversárias Generativas

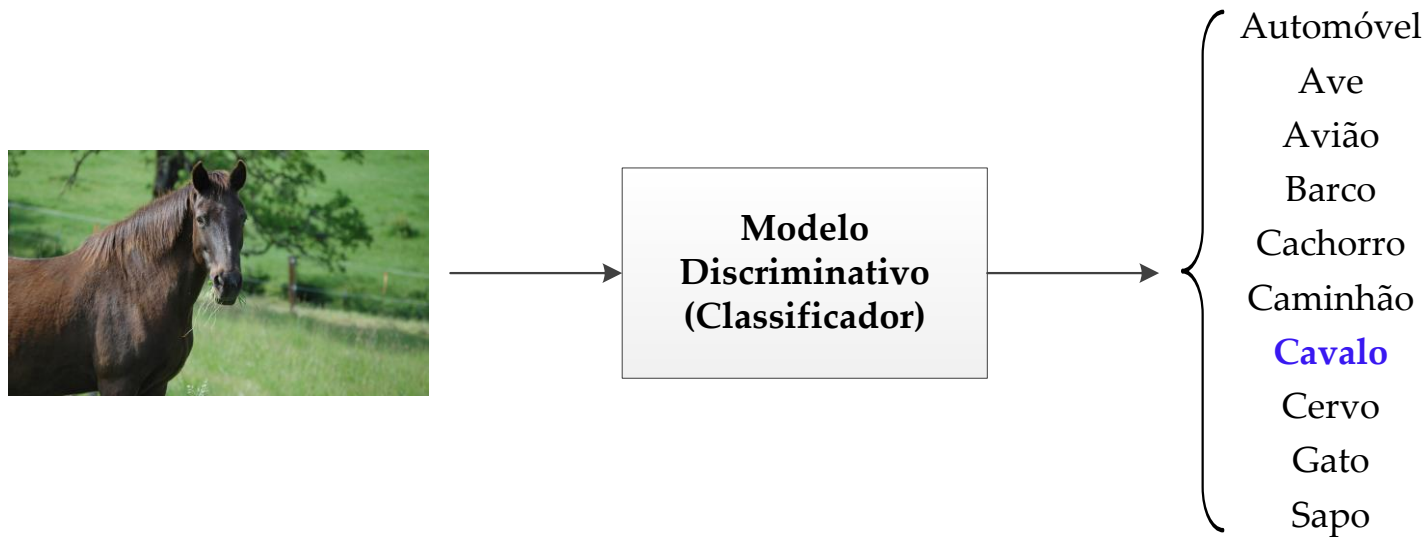
---

Prof. Levy Boccato  
Faculdade de Engenharia Elétrica e de Computação  
Universidade Estadual de Campinas (UNICAMP)

# Modelo discriminativo

---

- Um modelo discriminativo recebe um padrão completo em sua entrada e realiza um mapeamento para uma determinada classe ao reconhecer características típicas daquela classe.

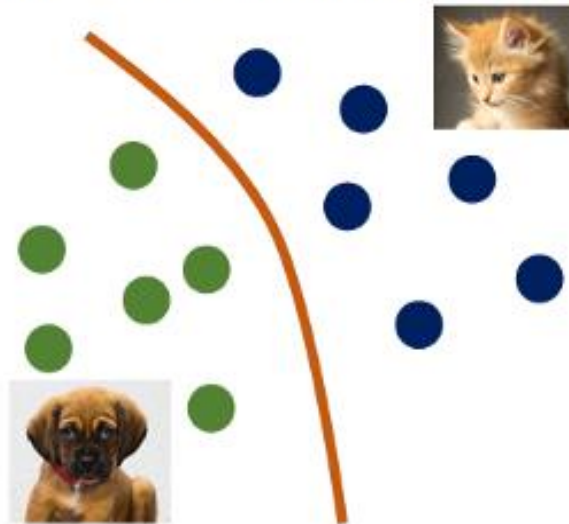


# Modelo discriminativo

---

- Considerando o espaço de possíveis padrões, o modelo discriminativo identifica regiões associadas a cada classe, as quais são delimitadas por fronteiras.

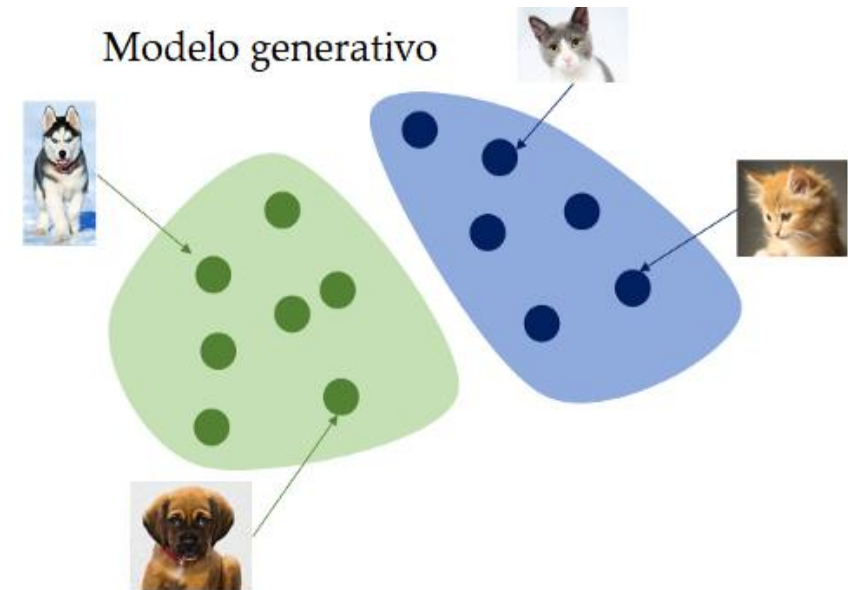
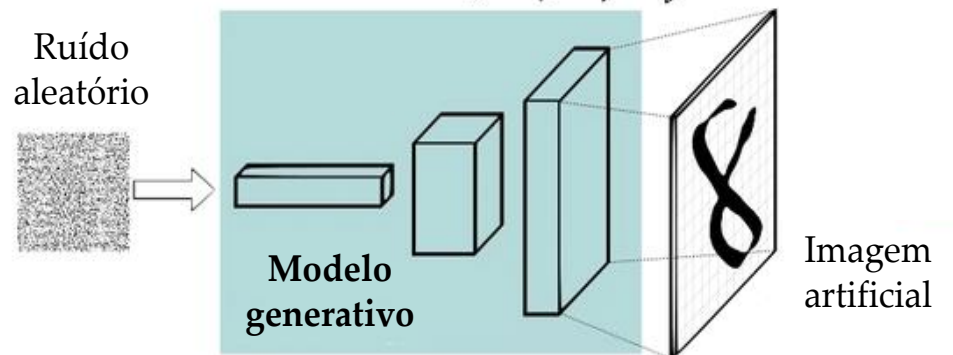
Modelo discriminativo



# Modelo generativo

---

- Os modelos generativos, em contrapartida, operam no sentido oposto: a partir de um descritor de uma classe, eles devem gerar padrões completos que apresentem as características típicas daquela classe.



# Modelo generativo

---

► A partir de seu aprendizado, um modelo generativo deve ser capaz de gerar um conjunto ilimitado de padrões sintéticos de algum domínio de interesse.

► **Desafios:**

- ❑ Criar padrões artificiais que não sejam facilmente distinguíveis de padrões reais;
- ❑ Oferecer algum grau de controle sobre o tipo de padrão gerado.

# Redes Adversárias Generativas

---

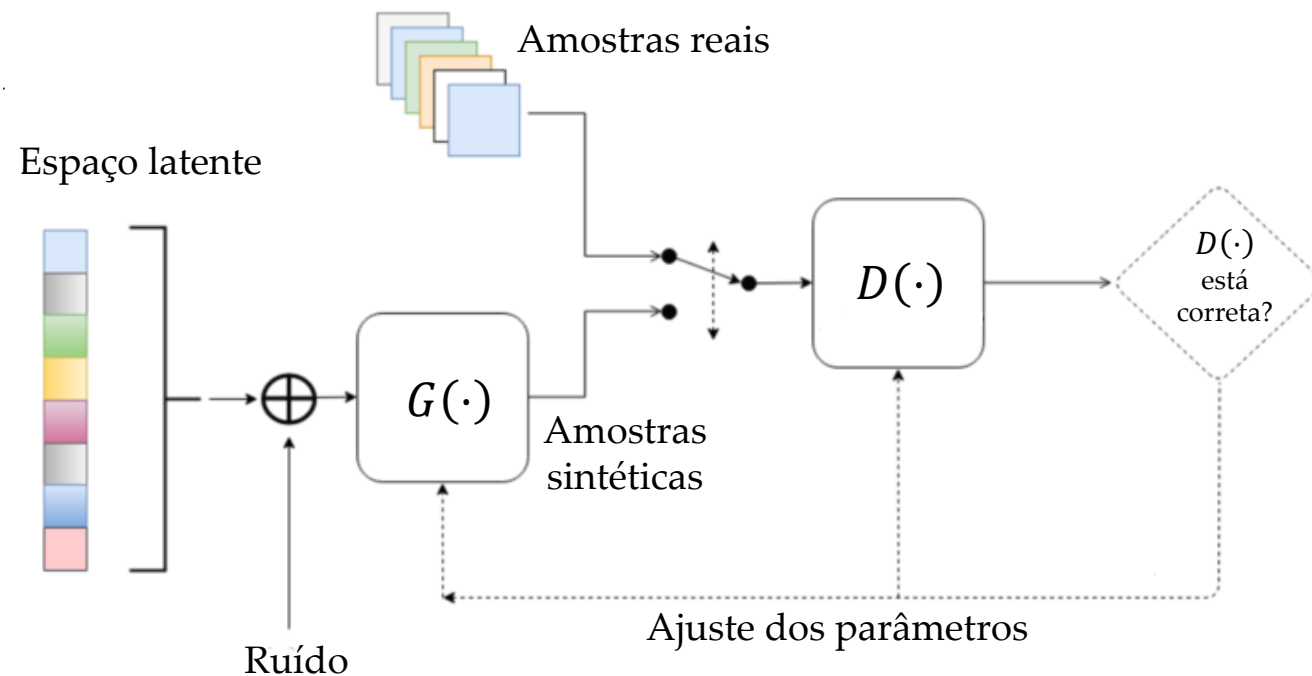
- ▶ É justamente neste contexto que surgiram as redes neurais adversárias generativas (GANs, do inglês *generative adversarial networks*) (Goodfellow et al., 2014; Gui et al., 2020).



- ▶ Yann LeCun sobre GANs: *"a ideia mais interessante nos últimos 10 anos em aprendizado de máquina"*.

# Redes Adversárias Generativas

- As GANs trazem a ideia de se estabelecer um jogo entre duas redes neurais, denominadas de *geradora* e *discriminadora*.



# Redes Adversárias Generativas

---

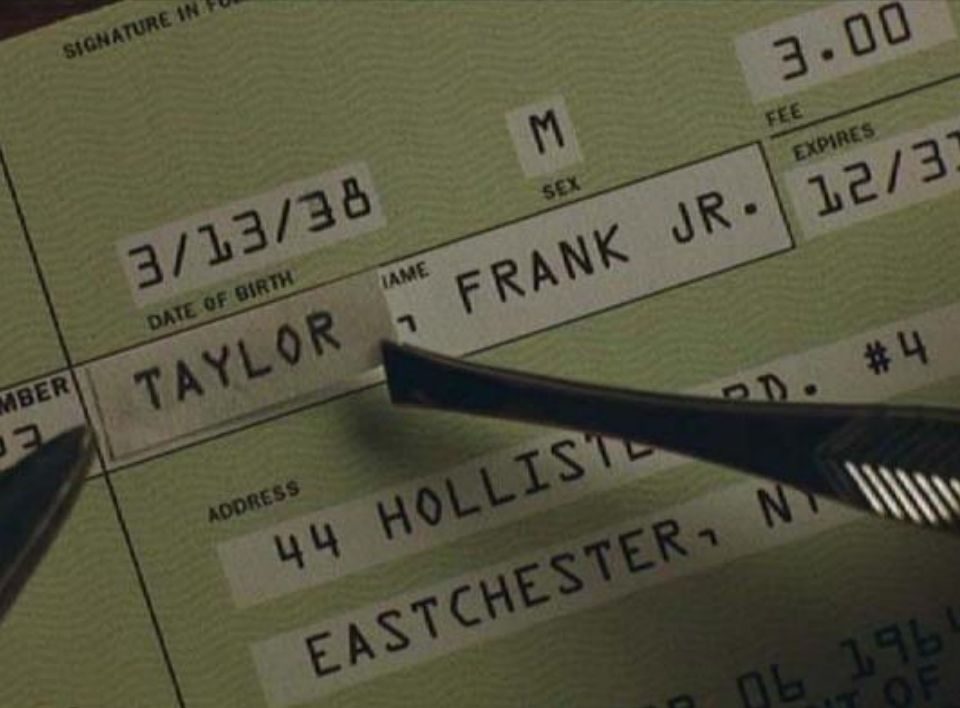


**Rede geradora:** deve criar dados sintéticos que sejam fidedignos em relação aos dados verdadeiros, *i.e.*, cujas características sejam plausíveis a ponto de estas amostras se confundirem com os dados verdadeiros.



**Rede discriminadora:** atua como um classificador binário, tentando descobrir quando o dado recebido em sua entrada é real ou fictício.





# Redes Adversárias Generativas

---

## ► Por que as redes são adversárias?

- ❑ A rede discriminadora busca se tornar cada vez mais hábil em discriminar imagens sintéticas de imagens reais.
- ❑ Concomitantemente, a rede geradora tenta criar imagens sintéticas que sejam cada vez mais verossímeis, ou, em outras palavras, menos distinguíveis de imagens reais, a fim de enganar a discriminadora.
- ❑ O aumento de desempenho de uma rede acarreta uma perda de desempenho da outra.



# Redes Adversárias Generativas

---

- ▶ Note que a rede geradora nunca recebe como entrada dados retirados do conjunto de treinamento. Na verdade, ela deve aprender a gerar amostras válidas apenas por meio de sua interação com a rede discriminadora.
- ▶ Quanto mais a rede discriminadora consegue perceber que os dados criados pela geradora são falsos, mais a rede geradora é levada pelo treinamento a aprimorar seus parâmetros de modo a conseguir construir amostras que se confundam com os dados reais.

# Redes Adversárias Generativas

---

- ▶ **Equilíbrio:** a rede geradora se torna tão hábil em criar dados plausíveis que a rede discriminadora, embora muito competente em reconhecer dados falsos, já não consegue mais perceber a diferença, convergindo para uma taxa de acerto de 50%.
- ▶ **Cuidado:** é preciso “sintonizar” o aprendizado das duas redes no sentido de manter um equilíbrio entre os progressos.
  - ❑ Caso uma rede avance muito rapidamente em relação à outra, isso pode prejudicar o aprendizado como um todo, pois a rede geradora precisa ser constantemente desafiada por um bom discriminador.

# Redes Adversárias Generativas

---

## ► Treinamento:

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{dados}}} \{\log D(\mathbf{x})\} + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} \{\log(1 - D(G(\mathbf{z})))\} \quad (1)$$

## ► Rede discriminadora:

- ❑ Quando sua entrada é um padrão do conjunto de treinamento, a saída esperada é  $D(\mathbf{x}) = 1$ .
- ❑ Quando a entrada é um padrão sintetizado pela rede geradora, denotado como  $G(\mathbf{z})$ , a saída deve ser zero:  $D(G(\mathbf{z})) = 0$ .
- ❑ Portanto, maximizar a função em (1) com respeito aos parâmetros da rede discriminadora corresponde a melhorar sua capacidade de distinguir os padrões reais dos sintéticos.

# Redes Adversárias Generativas

---

## ► Treinamento:

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{dados}}} \{\log D(\mathbf{x})\} + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} \{\log(1 - D(G(\mathbf{z})))\} \quad (1)$$

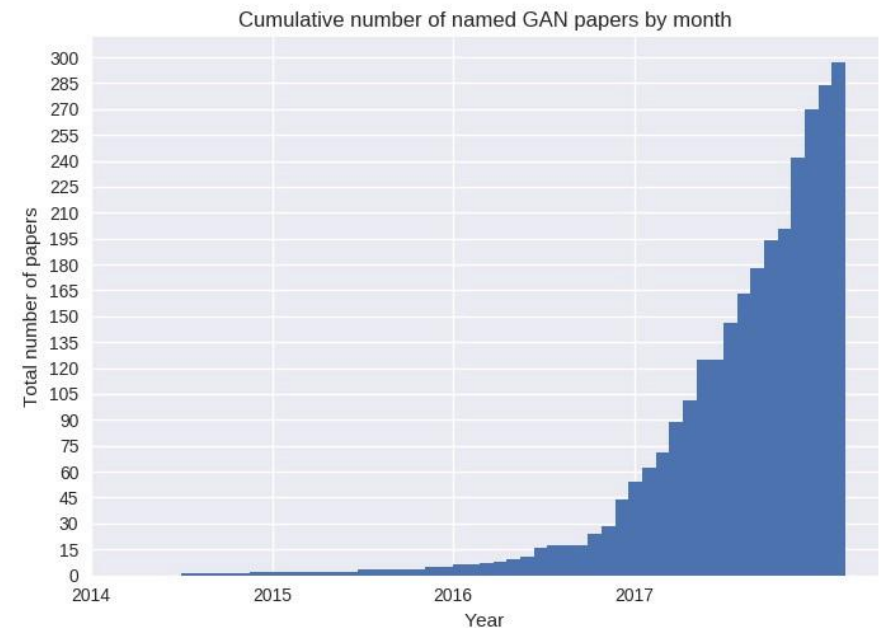
## ► Rede geradora:

- O alvo da rede geradora é fazer com que suas saídas,  $G(\mathbf{z})$ , sejam classificadas como padrões reais, *i.e.*,  $D(G(\mathbf{z})) = 1$ .

# Redes Adversárias Generativas

---

- ▶ A partir do trabalho pioneiro de Goodfellow et al. (2014), houve uma explosão de trabalhos relacionados a GANs e muitos esforços na comunidade para aprimorar os resultados.
- ▶ Por exemplo, foram propostas novas funções custo, além de estruturas mais sofisticadas para a rede geradora, como, por exemplo, as StyleGANs (Karras et al., 2019; Karras et al., 2020).



# Aplicações de GANs

---

## ► Síntese de imagens de faces humanas



[Goodfellow et al., 2014]  
University of Montreal



[Radford et al., 2015]  
Facebook AI Research



[Roth et al., 2017]  
Microsoft and ETHZ



[Karras et al., 2018]  
NVIDIA



# Aplicações de GANs

---

## ► Síntese de imagens de faces humanas



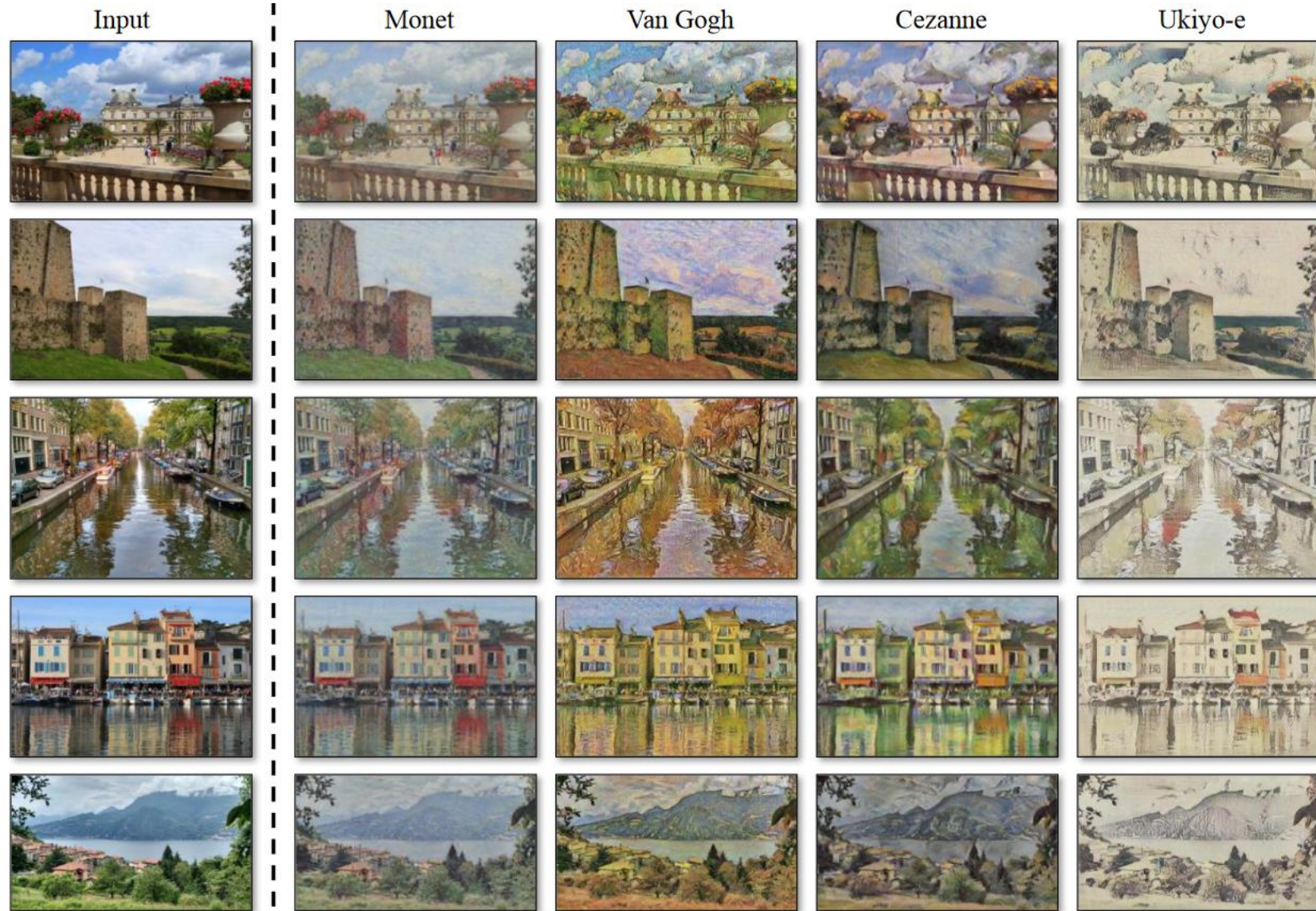
StyleGAN2 - <https://thispersondoesnotexist.com/>

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T., "Analyzing and Improving the Image Quality of StyleGAN", arXiv:1912.04958v2, 2020.



# Aplicações de GANs

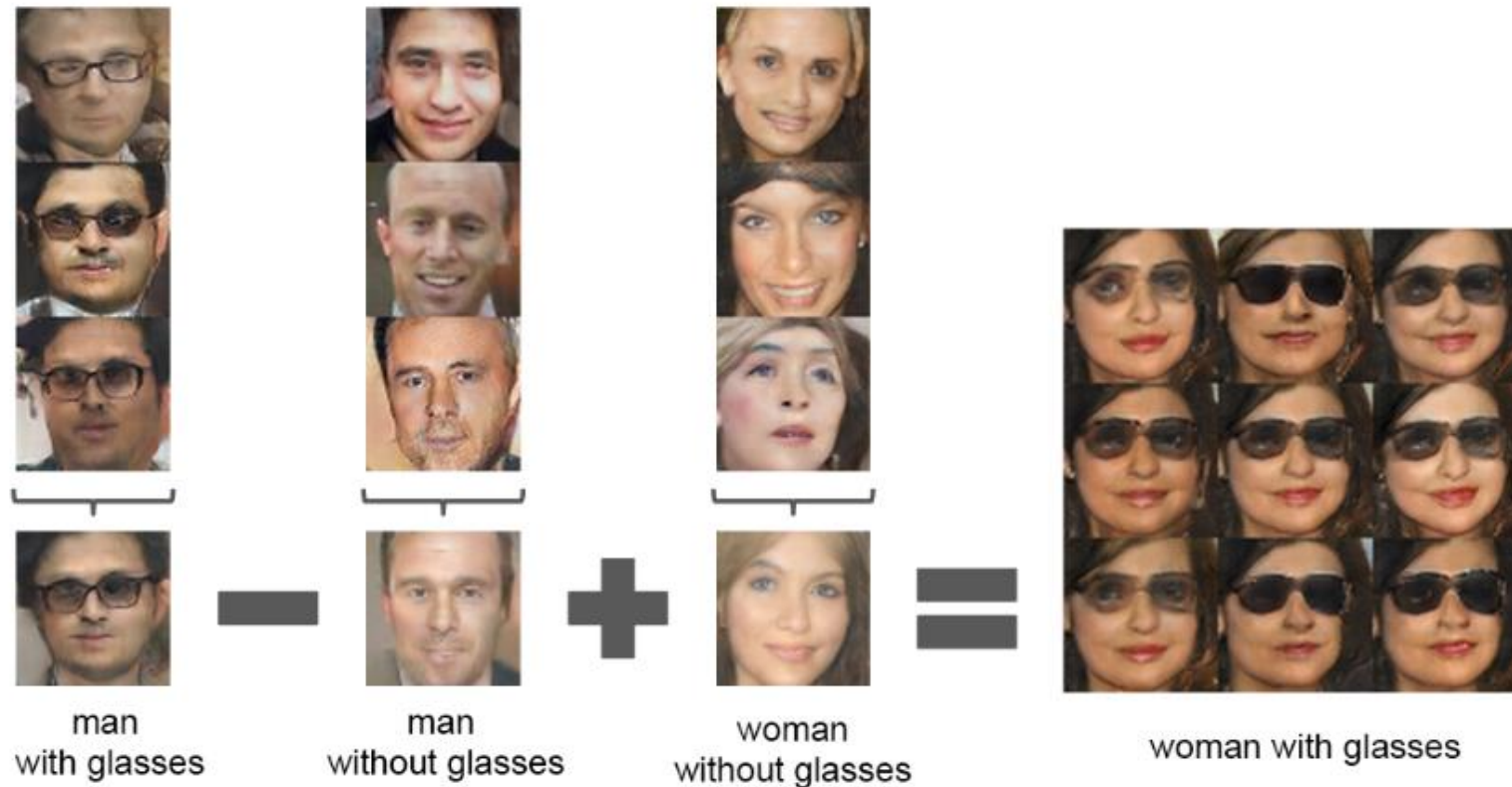
## ► Transferência de estilo



Zhu, J.-Y., Park, T., Isola, P., Efros, A. A., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", arXiv:1703.10593v7 [cs.CV], 2017.

# Aplicações de GANs

## ► Aritmética no espaço latente



Radford, A., Metz, L., Chintala, S., "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", arXiv:1511.06434v2, 2016.

# Aplicações de GANs

---

## ► Síntese de áudio / música



















Lee, S.-G., Hwang, U., Min, S., Yoon, S., "Polyphonic Music Generation with Sequence Generative Adversarial Networks", arXiv:1710.11418v2, 2017.



# Aplicações de GANs

## ► Síntese de imagem a partir de texto

Qiao, T., Zhang, J., Xu, D., Tao, D., "MirrorGAN: Learning Text-to-image Generation by Redescription", arXiv:1903.05854v1, 2019.

Input	a yellow bird with brown and white wings and a pointed bill	this bird is blue and black in color, with a sharp black beak	a small bird with a red belly, and a small bill and red wings	this small blue bird has a white underbelly
(a) AttnGAN				
(b) MirrorGAN Baseline				
(c) MirrorGAN				
(d) Ground Truth				

# Aplicações de GANs

---

## ► DeepFakes

- Este tipo uso de modelos de aprendizado de máquina, incluindo GANs, desperta muitas preocupações morais, éticas e legais.
- Exemplos:
  - ❑ Aplicativos como *DeepNude* e outras iniciativas ligadas à pornografia (e.g., por vingança).
  - ❑ Roubo de identidade através da adulteração de vídeos ou gravações falsas de voz.
  - ❑ Difamação de pessoas.
  - ❑ Desinformação.

Mirsky, Y., Lee, W., "The Creation and Detection of Deepfakes: A Survey", arXiv:2004.11138v3, 2020.

# Referências

---

Géron, A., "Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow", O'Reilly Media, 2ª ed., 2019.

Goodfellow, I., Bengio, Y., Courville, A., "Deep Learning", MIT Press, 2016.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., "Generative Adversarial Networks", Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), pp. 2672-2680, 2014.

Gui, J., Sun, Z., Wen, Y., Tao, D., Ye, J., "A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications", arXiv:2001.06937v1, 2020.

Karras, T., Laine, S., Aila, T., "A Style-Based Generator Architecture for Generative Adversarial Networks", arXiv:1812.04948v3, 2019.