Detecting images generated by ML models

Tomáš Pevný

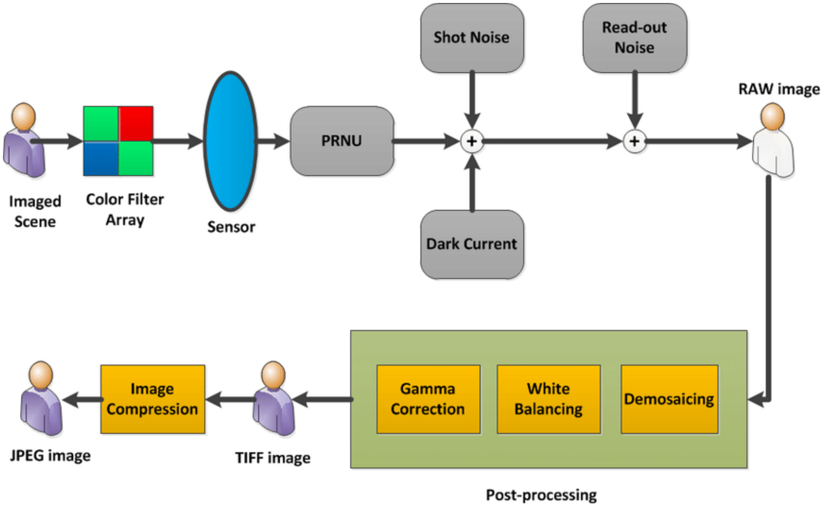
Czech Technical University in Prague
Gen Digital Inc.

March 27, 2024

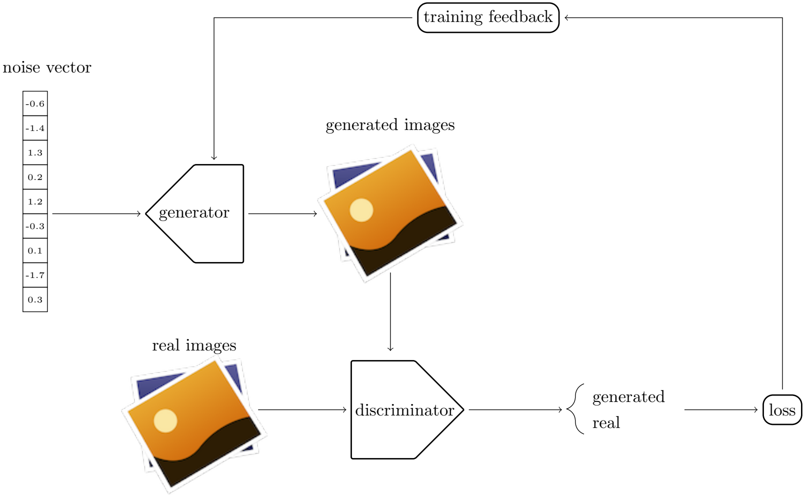
Who am I?

- 1997–2003, Ing. at FJFI, CTU in Prague
- 2004–2008, Ph.D. at SUNY at Binghamton, USA
- 2008–2009, Postdoc at INPG, Grenoble, France
- 2009–onwards Research at FEE, CTU in Prague
- 2013–2019, Consulting scientist at Cisco
- 2019–onwards, Consulting scientist at Gen digital, Inc (Avast)

Inside digital camera



Inside generative models



training feedback

noise vector

-0.6
-1.4
1.3
0.2
1.2
-0.3
0.1
-1.7
0.3

generator

generated images

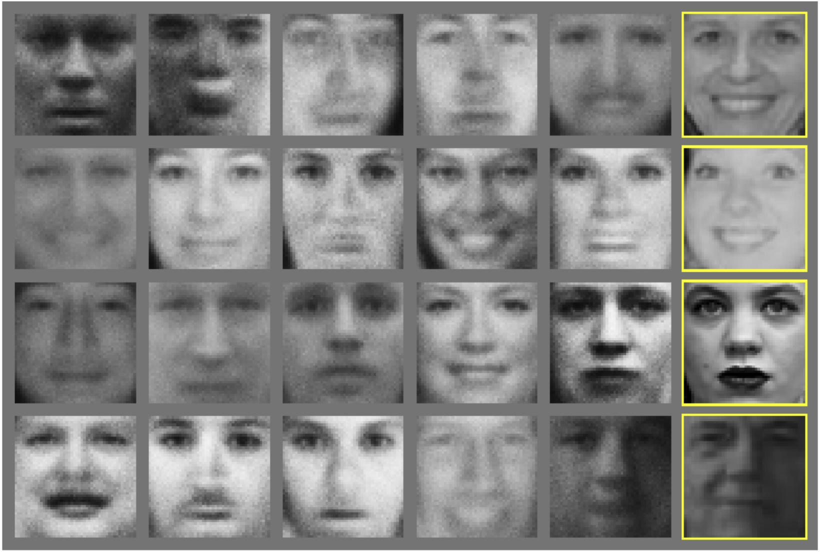
real images

discriminator

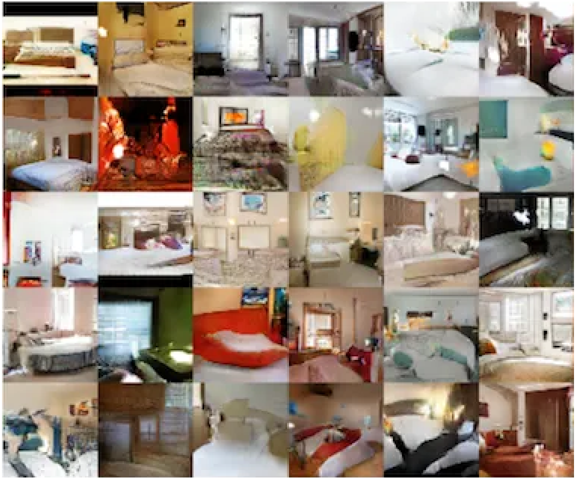
generated
real

loss

2014: Generative adversarial networks (GAN), 1st paper

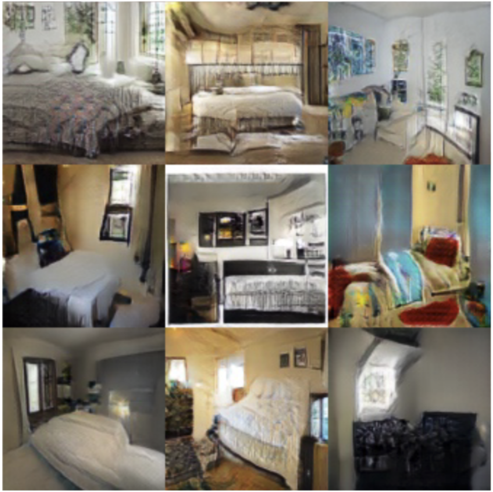


2015: DC-GAN



Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

2016: GAN with L2 loss

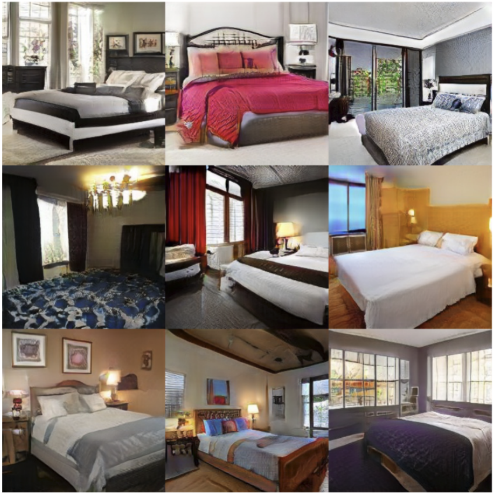


Mao, Xudong, et al. "Multi-class generative adversarial networks with the L2 loss function." arXiv preprint arXiv:1611.04076 5 (2016): 1057-7149.

2017: Wassertein GAN

[1] Gulrajani, Ishaan, et al. ”Improved training of wasserstein gans.” Advances in neural information processing systems 30 (2017).

2017: Progressive GAN



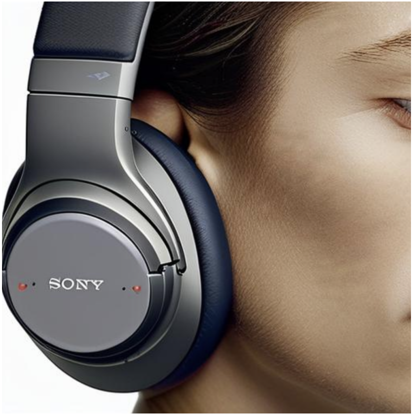
Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." arXiv preprint arXiv:1710.10196 (2017).

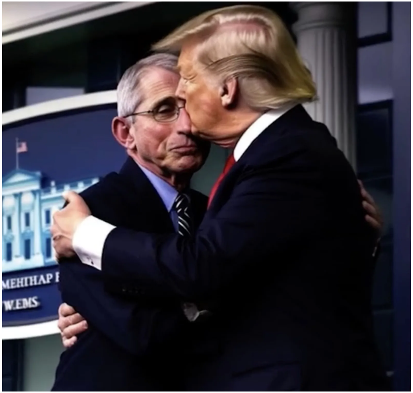
2022: Diffusion models

Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

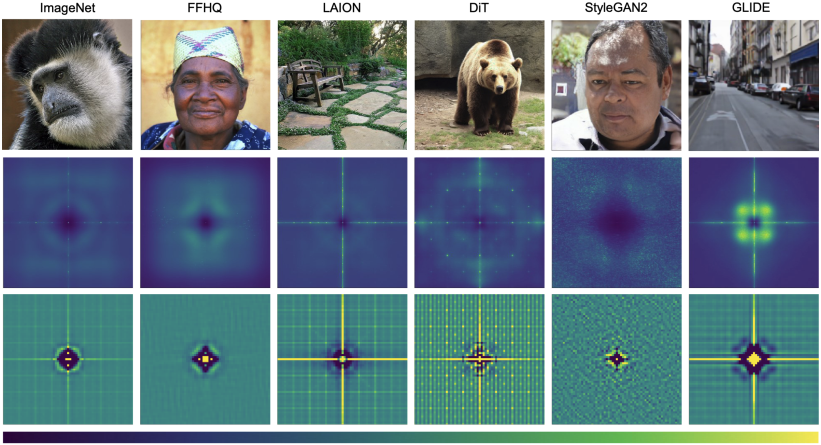
Passive detection

Detecting visible artifacts



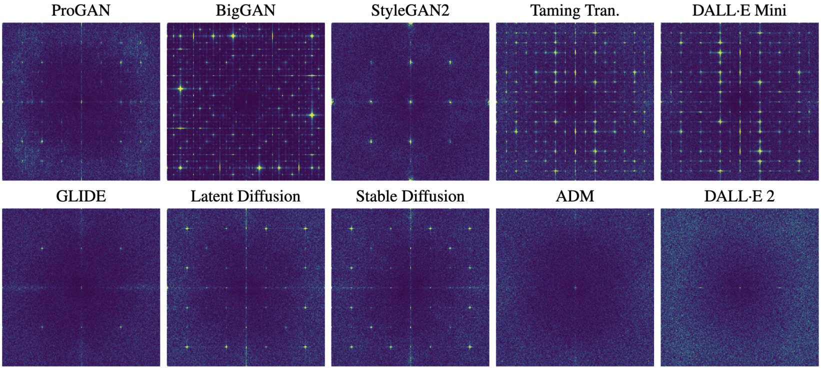


Detecting invisible artifacts



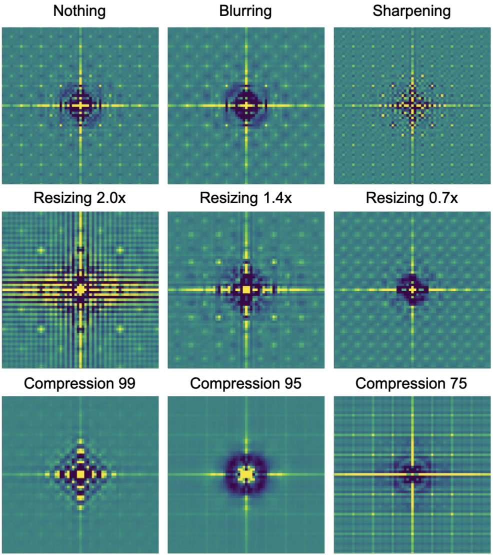
Corvi, Riccardo, et al. "Intriguing properties of synthetic images: from generative adversarial networks to diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Detecting invisible artifacts



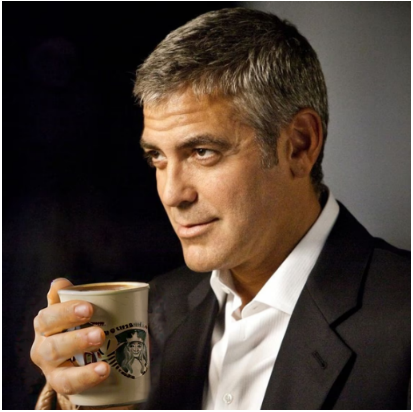
Corvi, Riccardo, et al. "On the detection of synthetic images generated by diffusion models." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

The effects of image processing



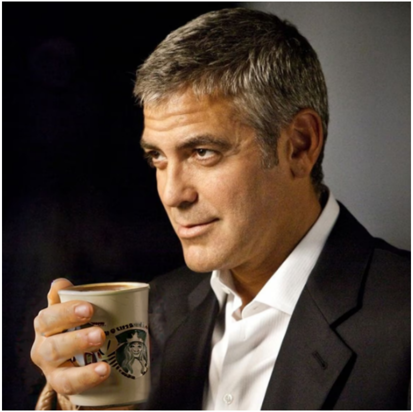
Corvi, Riccardo, et al. "Intriguing properties of synthetic images: from generative adversarial networks to diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Dubious correlations



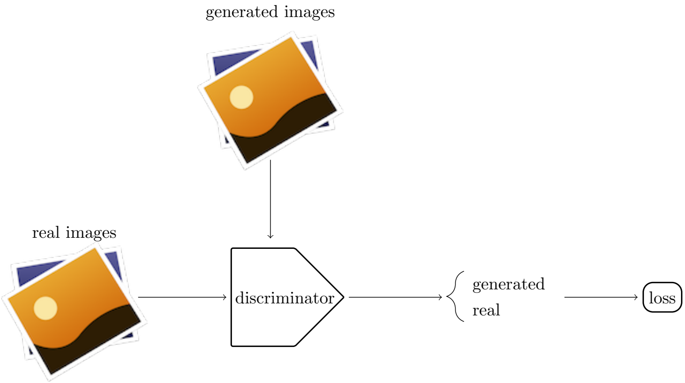
Credit goes to Matěj Račinský

Dubious correlations



Credit goes to Matěj Račinský

How automated detection works?



generated images

real images

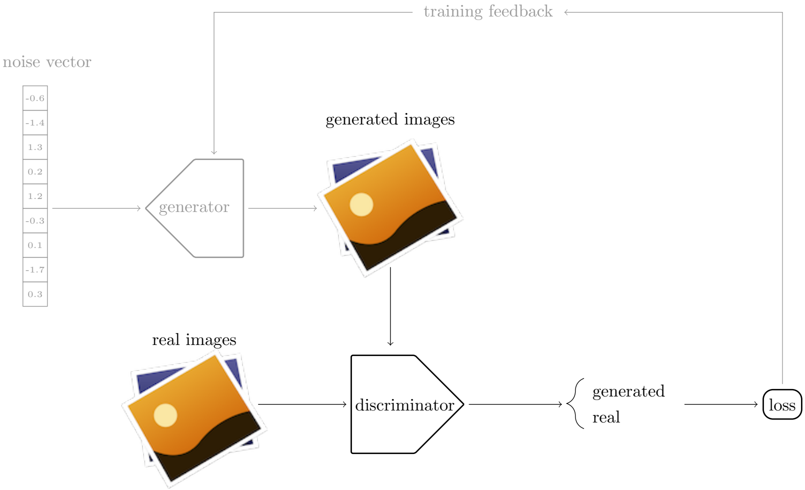
discriminator

generated

real

loss

Why is it a problem?



training feedback

noise vector

-0.6
-1.4
1.3
0.2
1.2
-0.3
0.1
-1.7
0.3

generator

generated images

real images

discriminator

generated
real

loss

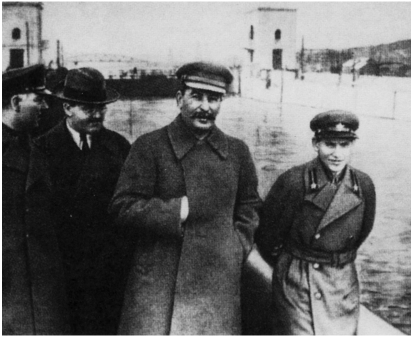
Can generated images be detected?

Can generated images be detected?

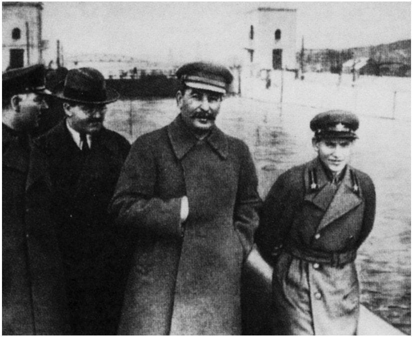
- Do we know the image history?
- What are the incentives of parties?
- How sophisticated is the attacker?
- Are we detecting single or multiple images?

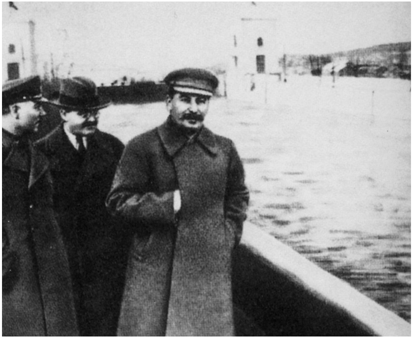
Do we really need to detect generated images?

Is this image genuine?

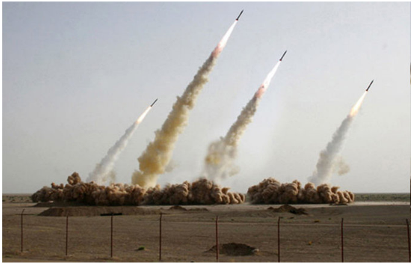


Is this image genuine?

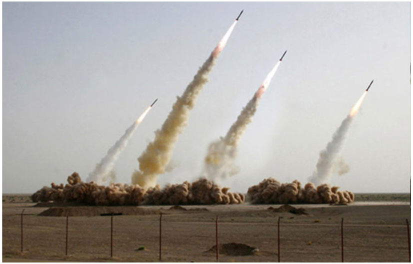


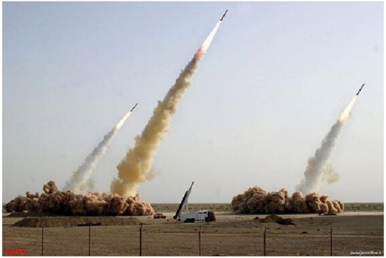


Is this image genuine?

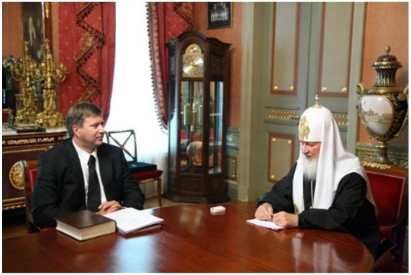


Is this image genuine?

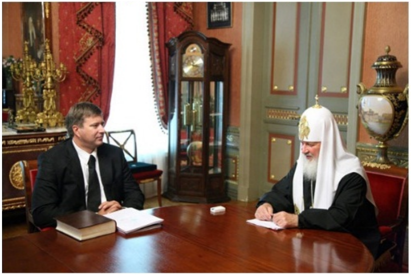


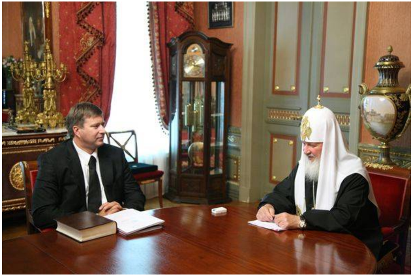


Is this image genuine?



Is this image genuine?





May-be, we care about authenticity.

Ensuring authenticity

- Audit trail — Coalition for Content Provenance and Authenticity (`c2pa.org`)
- (Fragile) watermarking

Watermarking

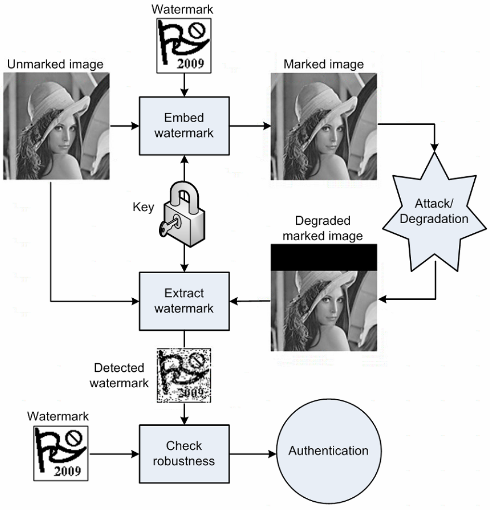
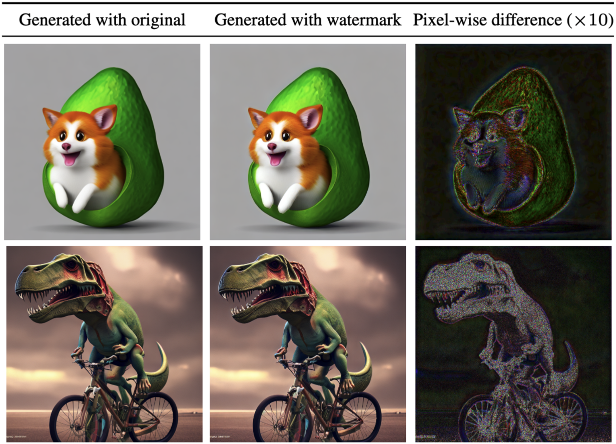


Image taken from:https://external-content.duckduckgo.com/iu/?u=https%3A%2F%2Fraw.githubusercontent.com%2Fetfovac%2Fwatermark%2Fmaster%2Fgraphics%2FFlowchart_ENG.png&f=1&nofb=1&ipt=b07e17ea48956052b6c2f4233aae11a7d47b1c566d02ff99eb7de07b4b4af47b&ipo=images

Deep-learning based watermarking



Fernandez, Pierre, et al. ”The stable signature: Rooting watermarks in latent diffusion models.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Summary

- Generated images can be sometimetimes detected.
- Incentives of parties
- Sophistication of attacks.
- Detecting images vs. users (accounts).
- Using semantic information / intentions.
- Required audit trails.
- Educating users.