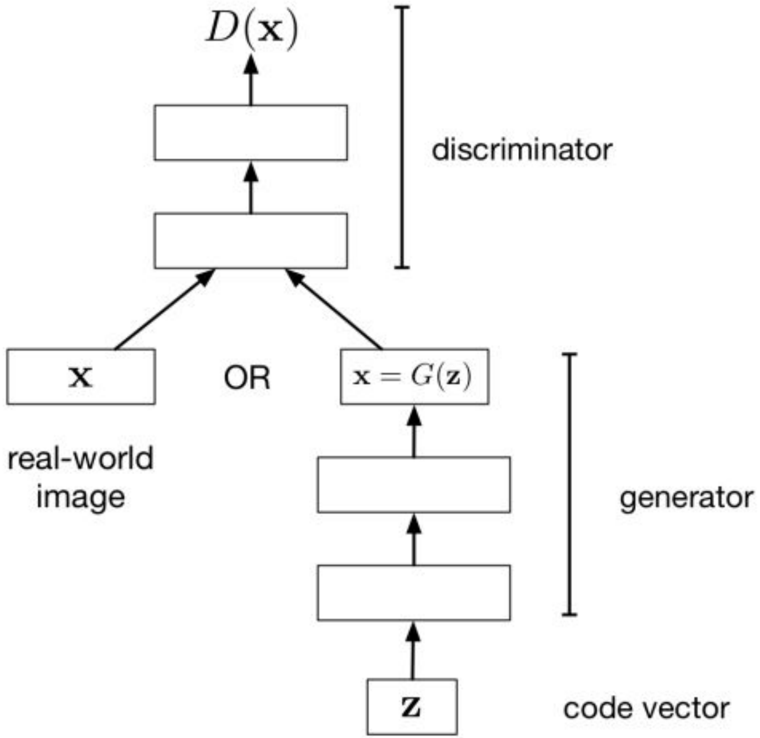
Uses of Generative Adversarial Networks

Tutorial 10 - March 22, 2021
Led by Caroline Malin-Mayor

Reminder: What is a GAN?

- Generative: generate an image (or other data) from a random code vector
- Adversarial: generator and discriminator are competing during training



Examples

[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation”

Drawbacks of GANs

- Not known if they truly model the underlying distribution of the data
- Can be tricky to train properly
- Hard to evaluate quantitatively
- No control over the generated image content

Example #1: Manipulating image content

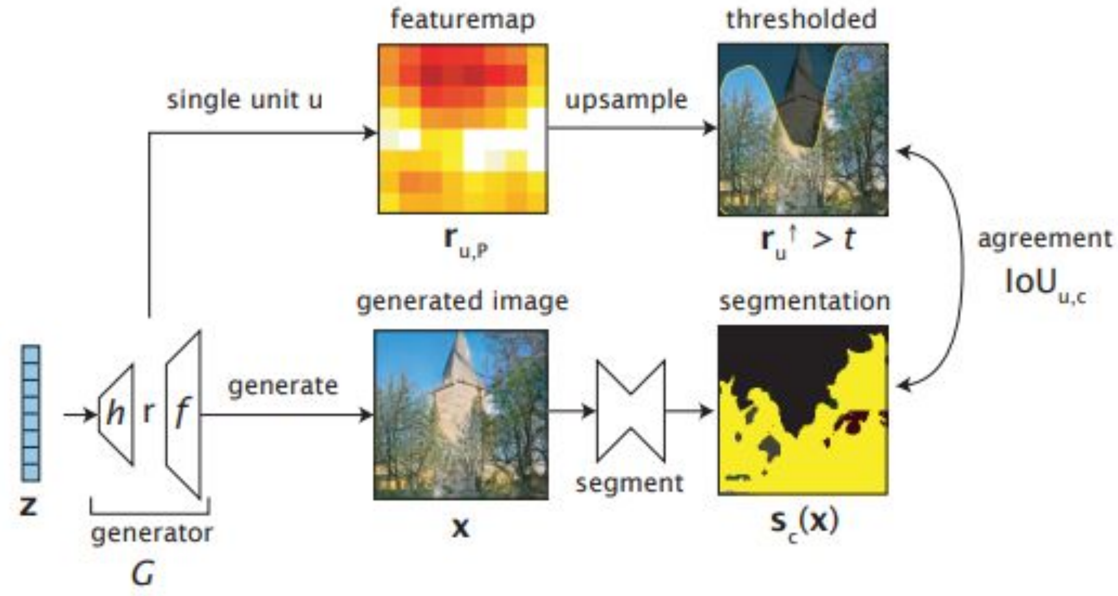
Published as a conference paper at ICLR 2019

GAN DISSECTION:
VISUALIZING AND UNDERSTANDING
GENERATIVE ADVERSARIAL NETWORKS

David Bau[1,2], Jun-Yan Zhu[1], Hendrik Strobelt[2,3], Bolei Zhou[4], Joshua B. Tenenbaum[1], William T. Freeman[1], Antonio Torralba[1,2]
[1]Massachusetts Institute of Technology, [2]MIT-IBM Watson AI Lab,
[3]IBM Research, [4]The Chinese University of Hong Kong

[2]

Finding the meaning of hidden units



[2]

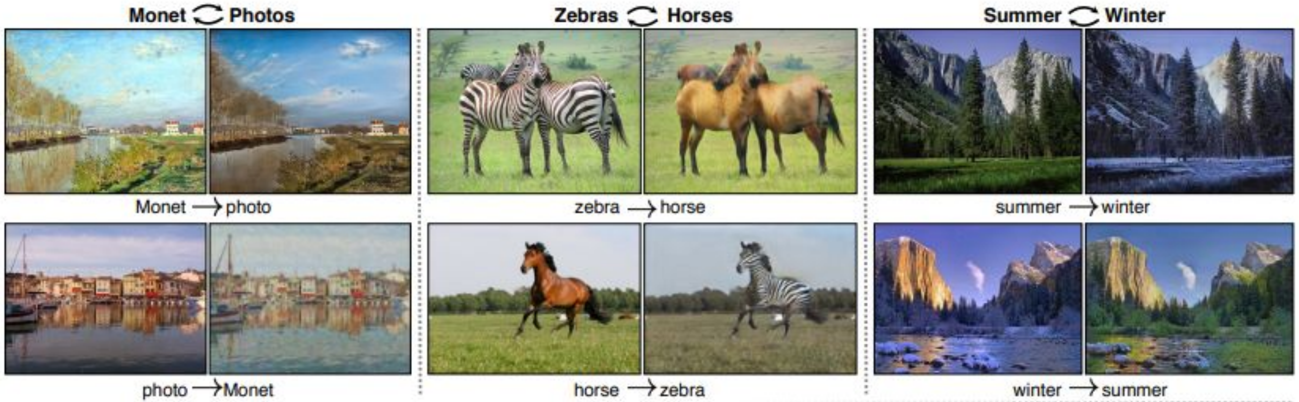
Demo: Manipulate generated images by activating hidden units

https://gandissect.csail.mit.edu/

Image-to-Image translation with CycleGANs

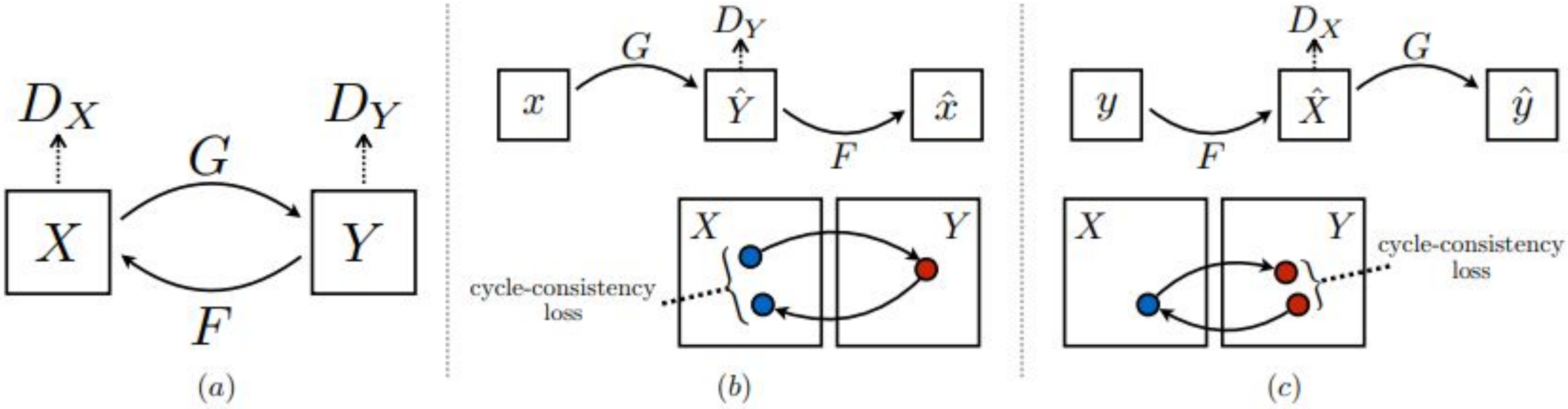
Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Jun-Yan Zhu* Taesung Park* Phillip Isola Alexei A. Efros
Berkeley AI Research (BAIR) laboratory, UC Berkeley



[10]

Reminder: CycleGAN Architecture



[10]

Example #2: Image-to-image translation in microscopy

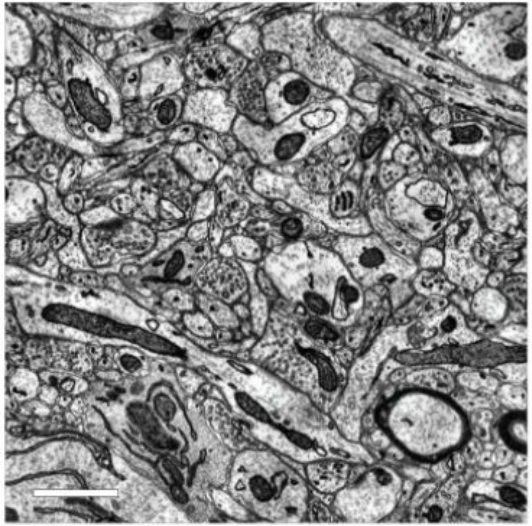
- There are **many** kinds of microscopes, that vary over:
  - speed
  - resolution
  - types of structures you can see
  - amount of energy used
  - if the sample can be re-imaged
  - cost
- Translating from one kind of image to another using deep learning can save cost and allow easier downstream processing of information

Example #2: Neuron reconstruction

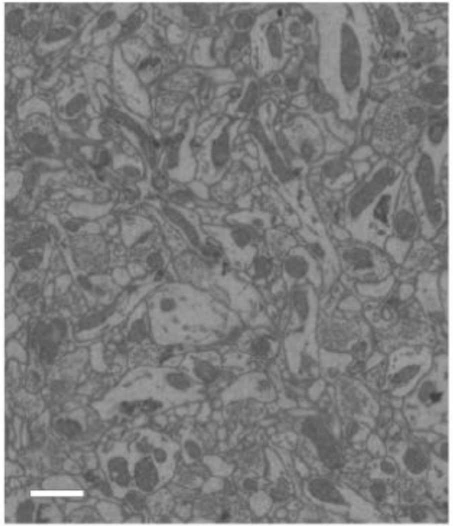
- https://www.youtube.com/watch?v=h6dF0htsTFc

Segmentation-Enhanced CycleGAN

Michał Januszewski, Viren Jain - Google AI Research, 2019



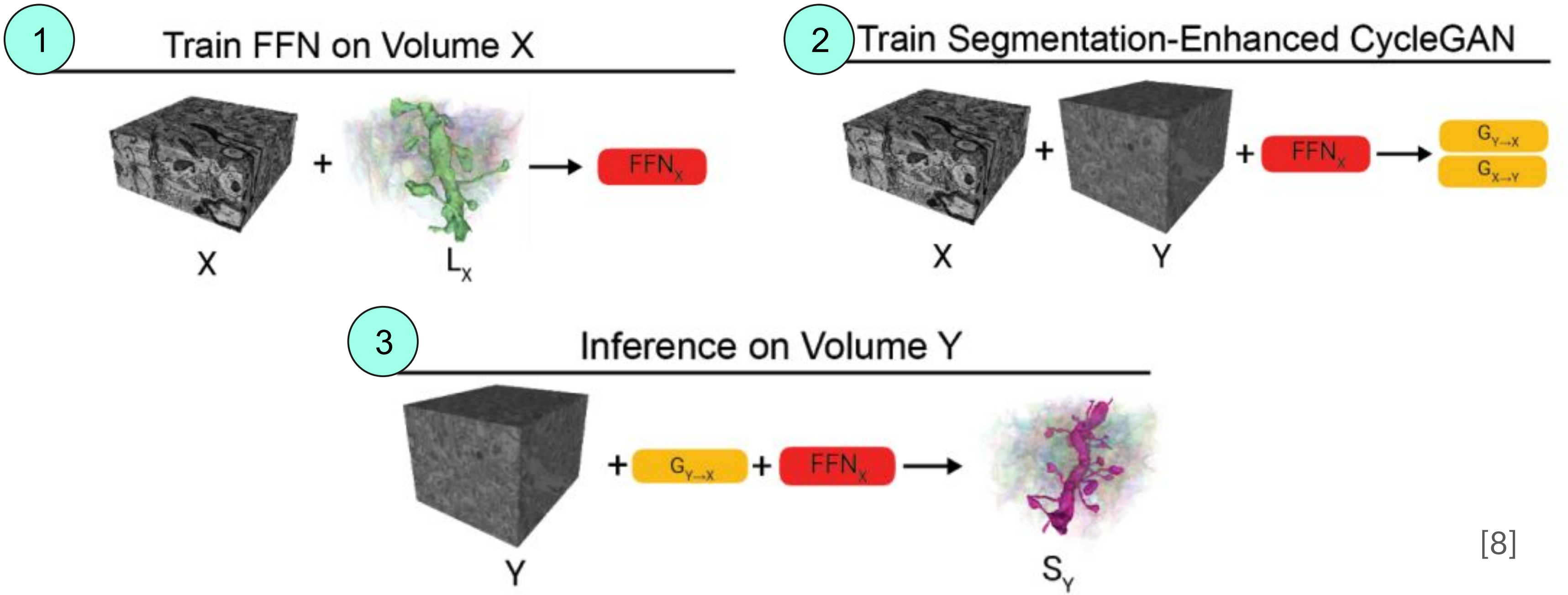
Original Data (X)



New Data (Y)

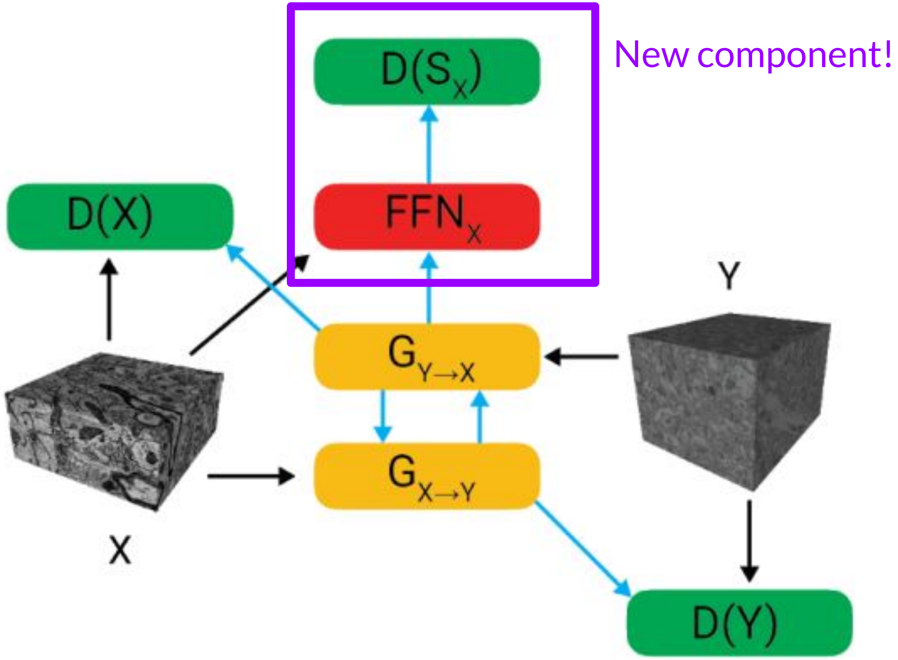
[8]

Segmentation-Enhanced CycleGAN



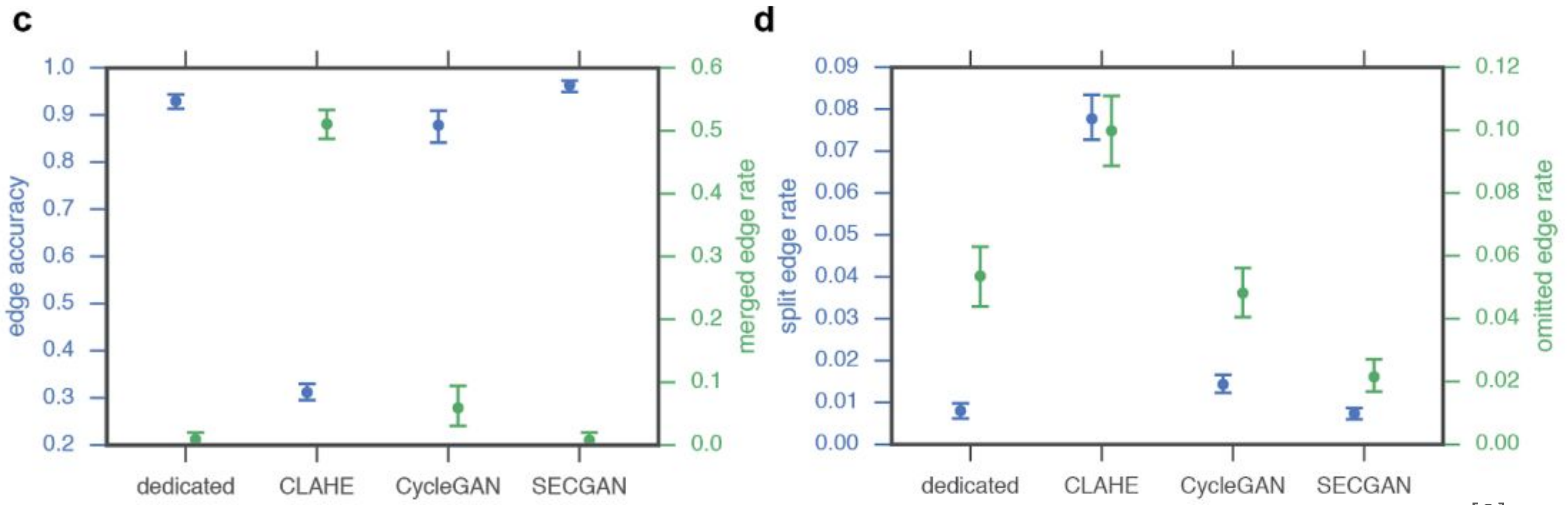
[8]

Segmentation-Enhanced CycleGAN



[8]

Evaluation



[8]

Caveats

- Again, no theoretical guarantee that we are modeling the underlying data distribution
- Be careful generalizing to different kinds of data
  - If structures are present in dataset Y that are not present in dataset X, these may be erased
  - If realistic neurons look different in Y than X, the discriminator will encourage them to be similar

Example #3: Interpreting a deep classifier

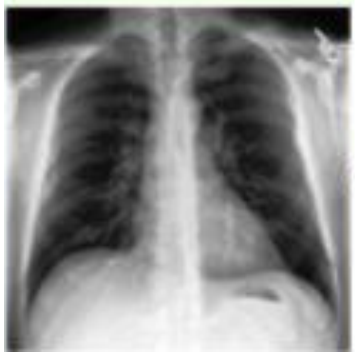
- Interpreting deep neural networks is an extremely popular area of research - why?
    - Show trustworthiness
    - Improve the model
    - Learn something about the data

Detecting Racial Features in Medical Images

- Neural networks can detect the race of patients from a variety of medical images (chest X-ray, chest CT scan, hand X-ray, and mammogram) with high accuracy, while human experts cannot [1]

White patient

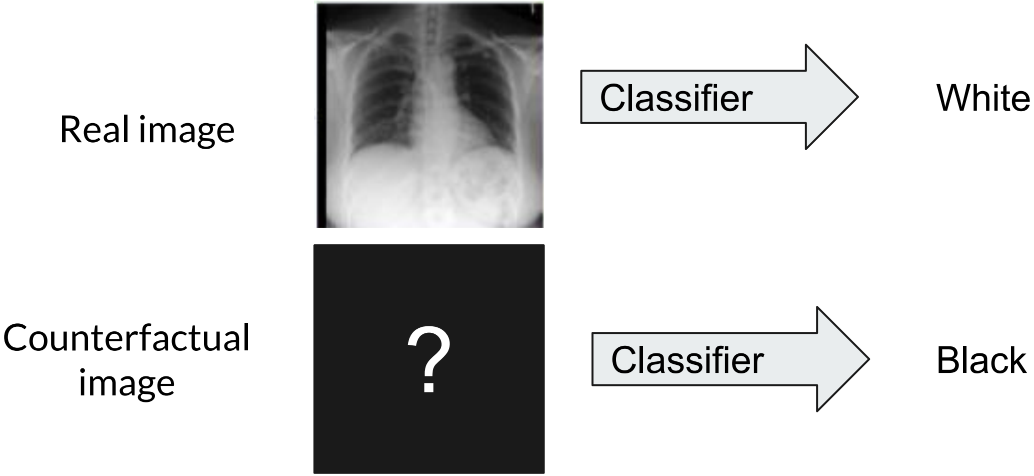
Black patient



MIMIC-CXR Dataset [4, 9]

Interpreting a classifier

- One possible question: What would have to change about the input to change the output class?



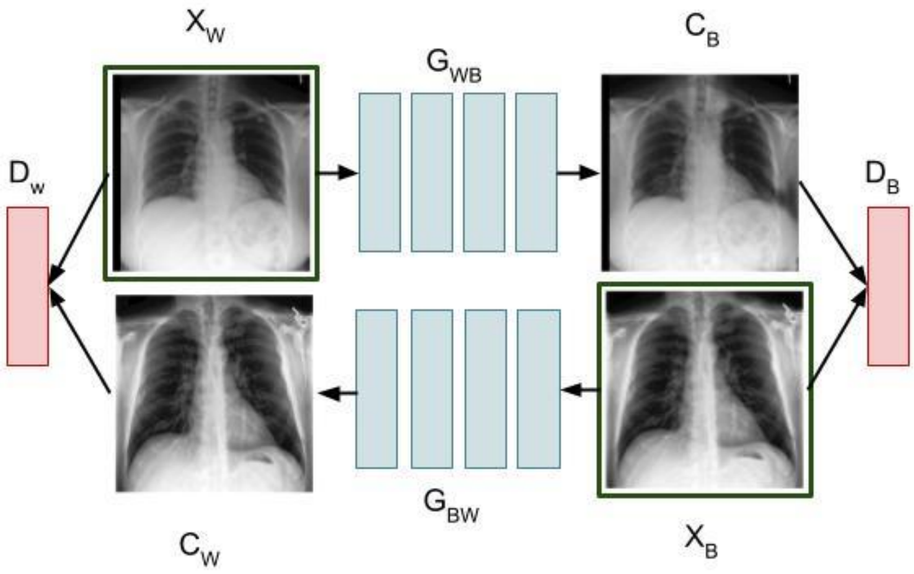
Generating counterfactuals

- Ways to generate counterfactuals
    - Find "closest" neighbors in the training set from the other class
    - Mutate the input to optimize another class by examining the gradients of the classifier
    - Create paired images with CycleGAN that translate from one class to another! [1]

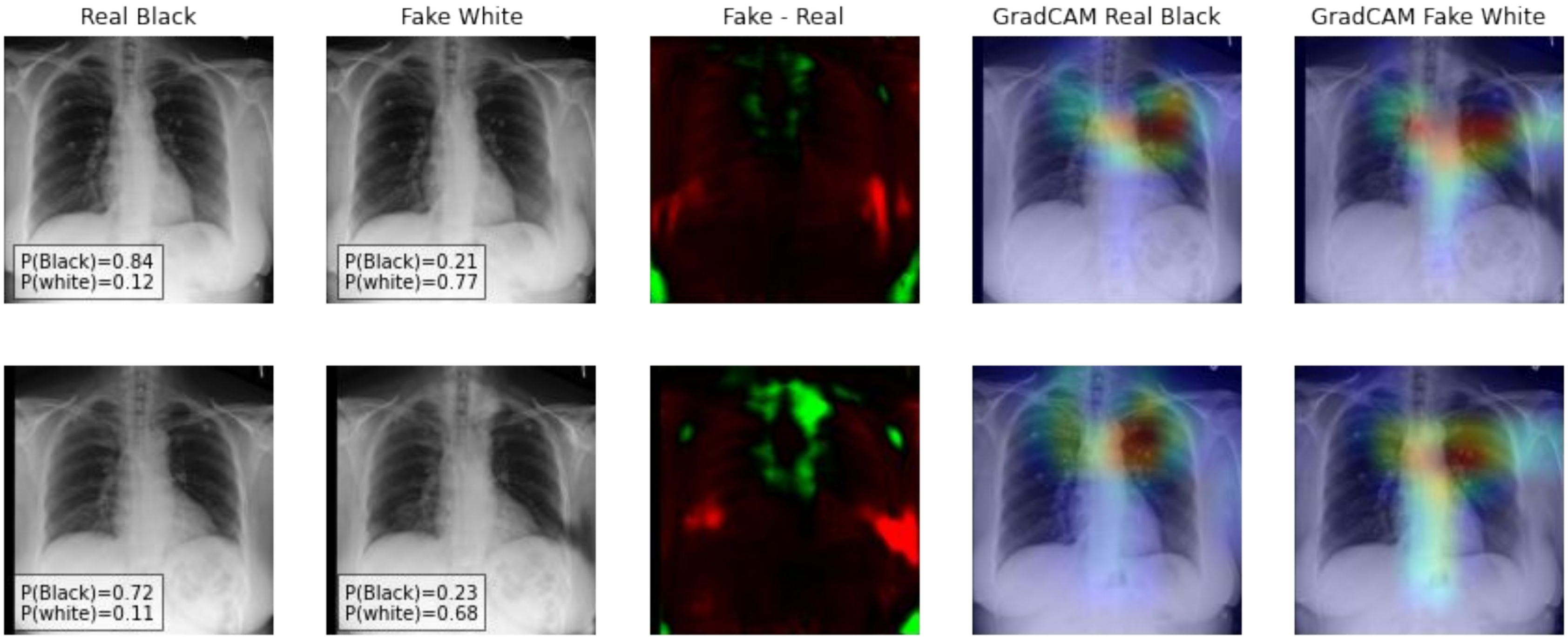
[1] N. Eckstein, A. S. Bates, G. S. X. E. Jefferis, and J. Funke, "Discriminative Attribution from Counterfactuals"

Our CycleGAN Architecture

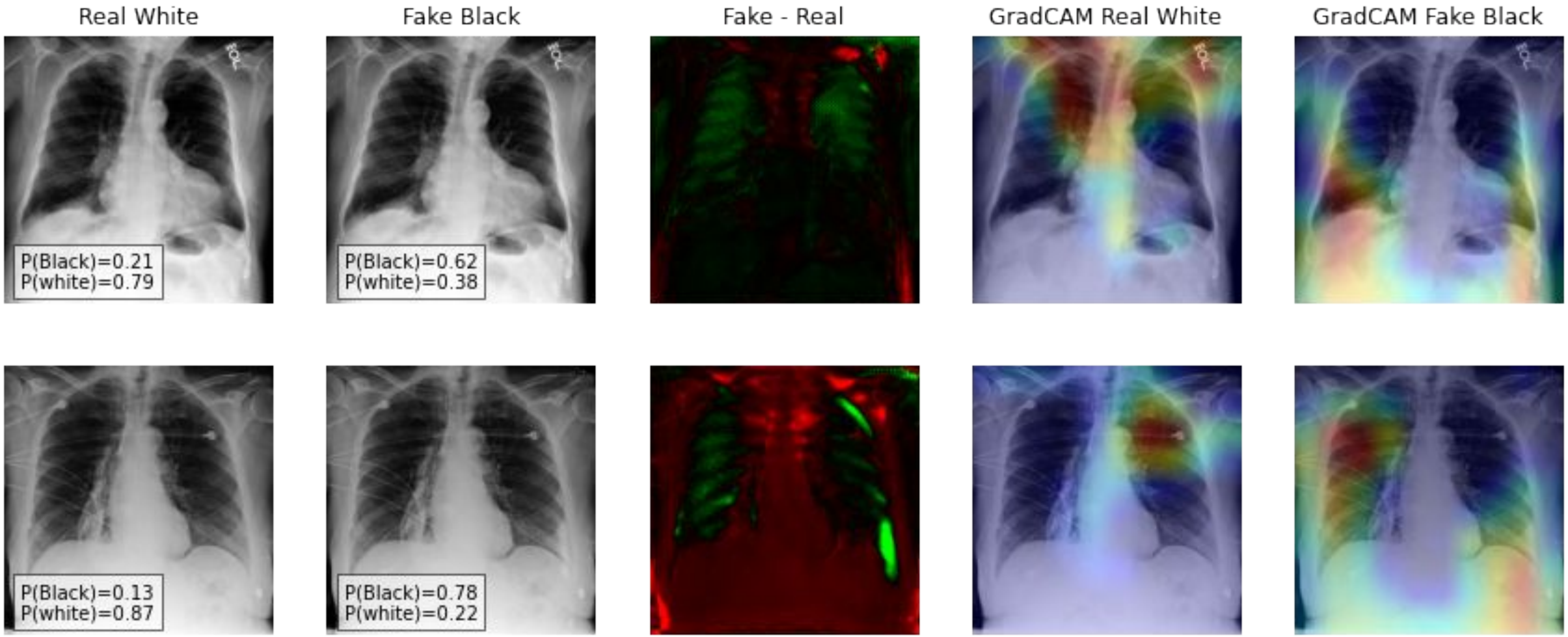
- G is a CNN with residual connections
- D is a PatchGAN
- Completely separate from previously trained racial classifier
    - Can use race classifier to evaluate "success"
    - 10-20% success rate at fooling the race classifier



Results


Real Black
Fake White
Fake - Real
GradCAM Real Black
GradCAM Fake White
P(Black)=0.84
P(white)=0.12
P(Black)=0.21
P(white)=0.77
P(Black)=0.72
P(white)=0.11
P(Black)=0.23
P(white)=0.68

Results


Real White
Fake Black
Fake - Real
GradCAM Real White
GradCAM Fake Black
P(Black)=0.21
P(white)=0.79
P(Black)=0.62
P(white)=0.38
P(Black)=0.13
P(white)=0.87
P(Black)=0.78
P(white)=0.22

Caveats and Open Questions

- As always, we are not guaranteed to model the underlying data distribution
    - We just want to interpret the classifier! Cannot claim these are all the differences, or the only differences between the x-rays of black and white patients
- Can we quantify this difference in intensity between X-rays from black and white patients?
- Do diagnoses remain constant across translated X-rays?

Conclusion

- GANs can be used in a wide variety of applications: computer graphics, image translation for transfer learning, network interpretability, and more!
- Always be careful about what you can and cannot claim about the output of your GAN

References

[1] N. Eckstein, A. S. Bates, G. S. X. E. Jefferis, and J. Funke, “Discriminative Attribution from Counterfactuals,” *arXiv:2109.13412 [cs]*, Sep. 2021, Accessed: Mar. 22, 2022. [Online]. Available: http://arxiv.org/abs/2109.13412

[2] D. Bau *et al.*, “GAN Dissection: Visualizing and Understanding Generative Adversarial Networks,” *arXiv:1811.10597 [cs]*, Dec. 2018, Accessed: Mar. 22, 2022. [Online]. Available: http://arxiv.org/abs/1811.10597

[3] I. Goodfellow *et al.*, “Generative Adversarial Nets,” in *Advances in Neural Information Processing Systems*, 2014, vol. 27. Accessed: Mar. 22, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[5] A. E. W. Johnson *et al.*, “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” *Sci Data*, vol. 6, no. 1, p. 317, Dec. 2019, doi: 10.1038/s41597-019-0322-0.

[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” *arXiv:1710.10196 [cs, stat]*, Feb. 2018, Accessed: Mar. 22, 2022. [Online]. Available: http://arxiv.org/abs/1710.10196

[7] I. Banerjee *et al.*, “Reading Race: AI Recognises Patient’s Racial Identity In Medical Images,” *arXiv:2107.10356 [cs, eess]*, Jul. 2021, Accessed: Mar. 22, 2022. [Online]. Available: http://arxiv.org/abs/2107.10356

[8] M. Januszewski and V. Jain, “Segmentation-Enhanced CycleGAN,” Neuroscience, preprint, Feb. 2019. doi: 10.1101/548081.

[9] A. E. W. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng, “The MIMIC-CXR Database.” physionet.org, 2019. doi: 10.13026/C2JT1Q.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” *arXiv:1703.10593 [cs]*, Aug. 2020, Accessed: Mar. 22, 2022. [Online]. Available: http://arxiv.org/abs/1703.10593