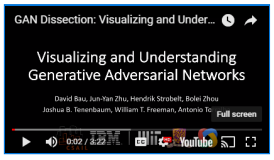
GAN Dissection: Visualizing and Understanding Generative Adversarial Networks

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]MIT CSAIL, [2]MIT-IBM Watson AI Lab, [3]IBM Research, [4]The Chinese University of Hong Kong

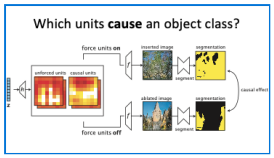
New: In Proceedings of the National Academy of Sciences, Sep 2020, we update the methods and unify analysis with classifiers.

Code and Data Github

ICLR 2019 Paper



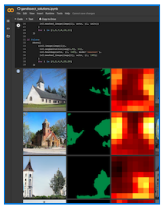
Video Demo



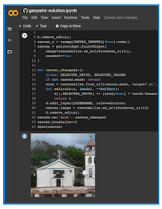
Tutorial Slides

Interactive Demo GAN Paint

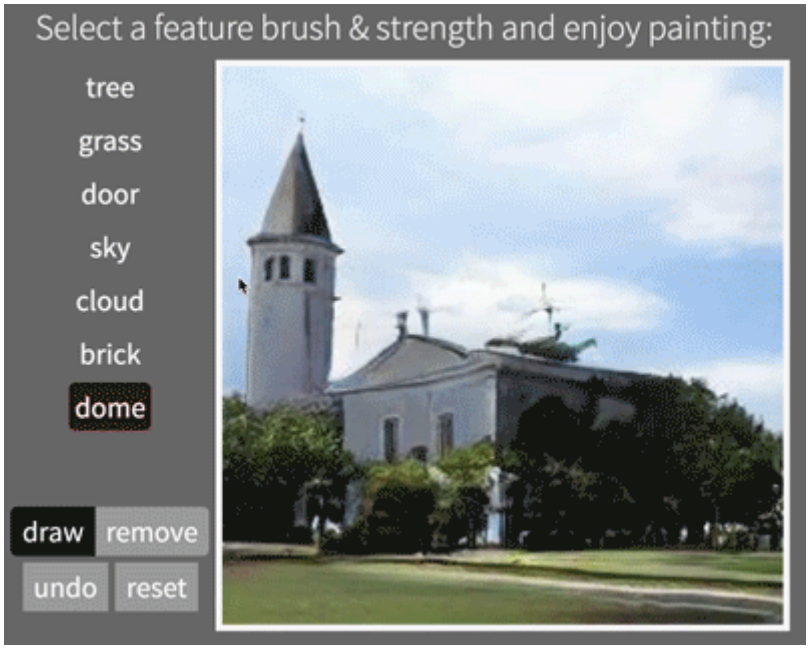
Colab Playground



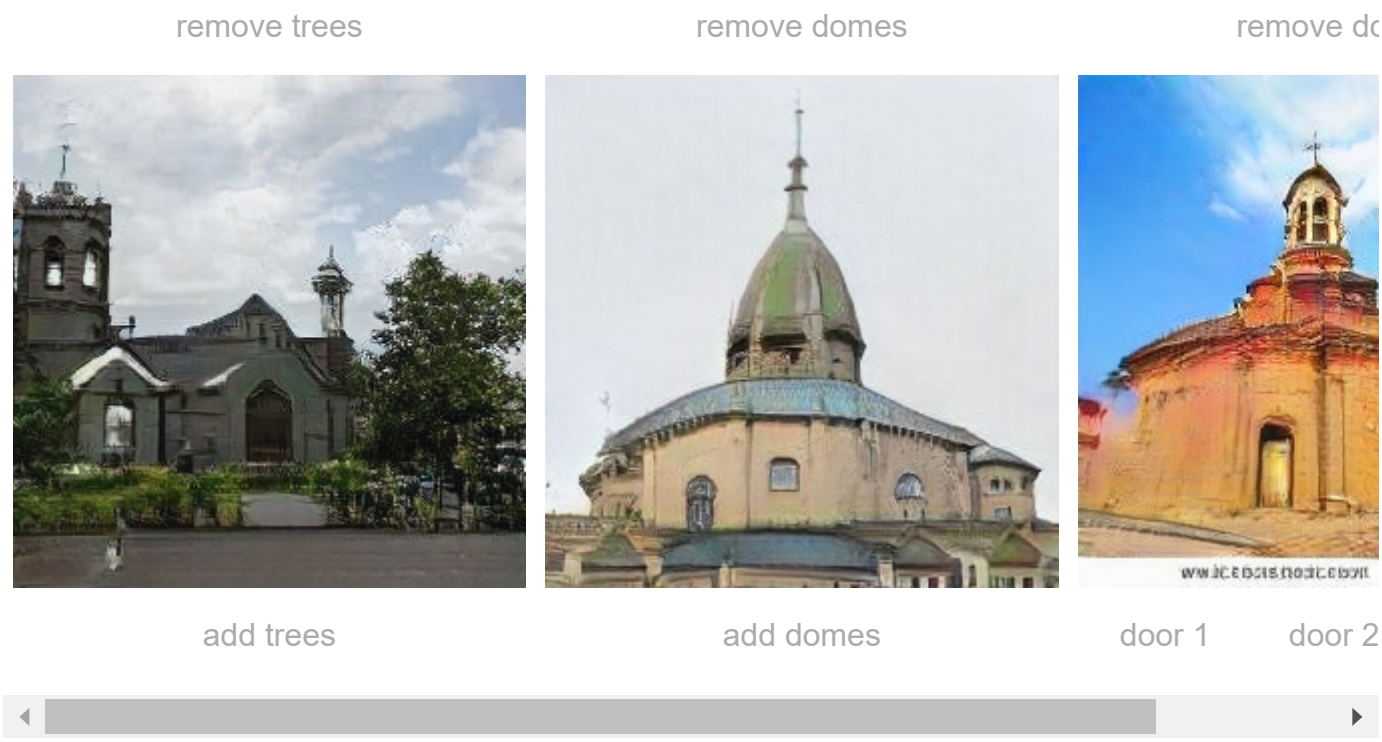
Colab Tutorial #1



Colab Tutorial #2



The #GANpaint app works by directly activating and deactivating sets of neurons in a deep network trained to generate images. Each button on the left ("door", "brick", etc) corresponds to a set of 20 neurons. The app demonstrates that, by learning to draw, the network also learns about objects such as trees and doors and rooftops. By switching neurons directly, you can observe the structure of the visual world that the network has learned to model. (Try it here.)



Why Painting with a GAN is Interesting

A computer could draw a scene in two ways:

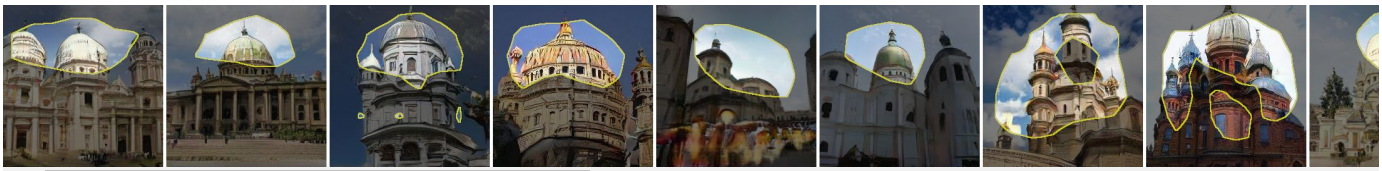
1. It could *compose* the scene out of objects it knows.

2. Or it could *memorize* an image and replay one just like it.

In recent years, innovative Generative Adversarial Networks (GANs, I. Goodfellow, et al, 2014) have demonstrated a remarkable ability to create nearly photorealistic images. However, it has been unknown whether these networks learn *composition* or if they operate purely through *memorization* of pixel patterns.

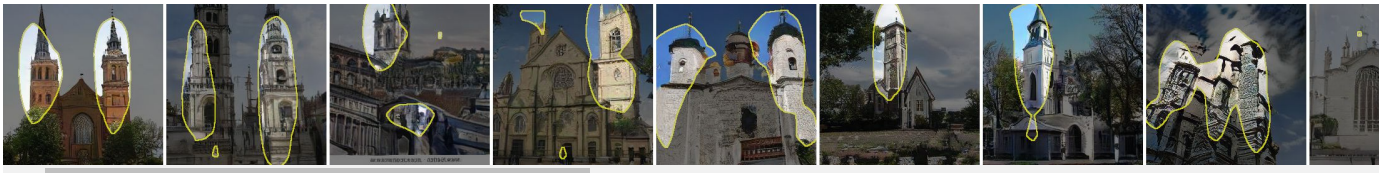
Our GAN Paint demo and our GAN Dissection method provide evidence that the networks have learned some aspects of composition.

Unit 365 of layer4 of a church Progressive GAN (T. Karras, et al, 2018) draws trees.



Unit 43 draws domes.

Unit 14 draws grass.

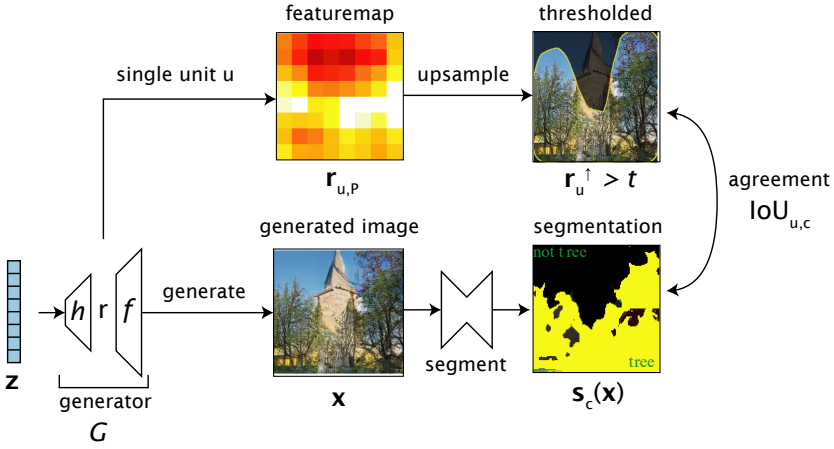


Unit 276 draws towers.

What is GAN Dissection?

Our paper describes a framework for visualizing and understanding the structure learned by a generative network. GAN dissection allows us to ask:

1. Does the network learn internal neurons that match meaningful concepts?
2. Do these sets of neurons merely correlate with objects, or does the GAN use those neurons to reason about objects?
3. Can causal neurons be manipulated to improve the output of a GAN?

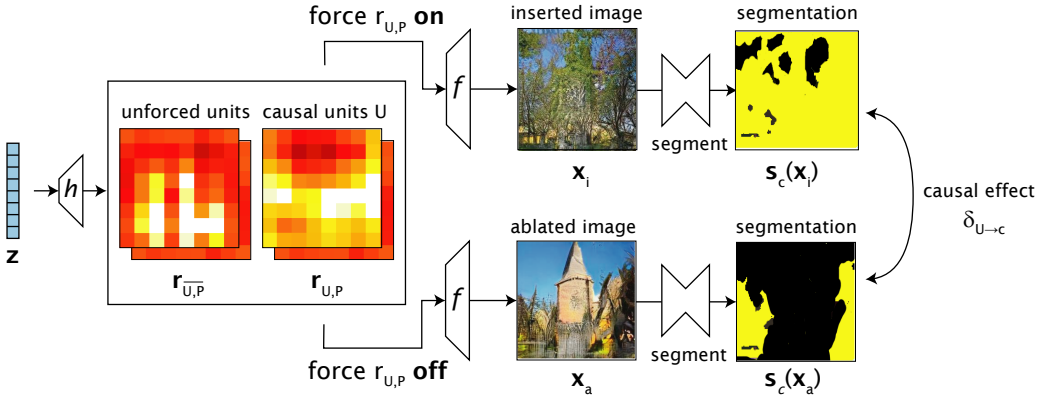


Dissection uses a segmentation network (T. Xiao, et al, 2018) along with a dissection method (D. Bau, et al, 2017) to find individual units of the generator that match meaningful object classes, like trees.

The neurons that a GAN learns depend on the type of scene it learns to draw:
A business suit neuron appears when learning conference rooms, and a stove neuron appears when drawing kitchens.

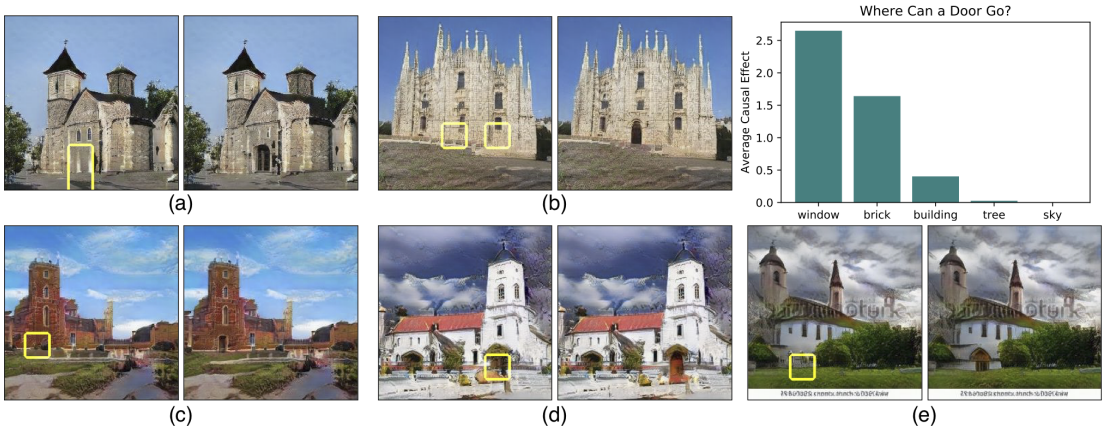
What Does Each Neuron Control?

To verify that sets of neurons control the drawing of objects rather than merely correlating, we *intervene* in the network and activate and deactivate neurons directly.

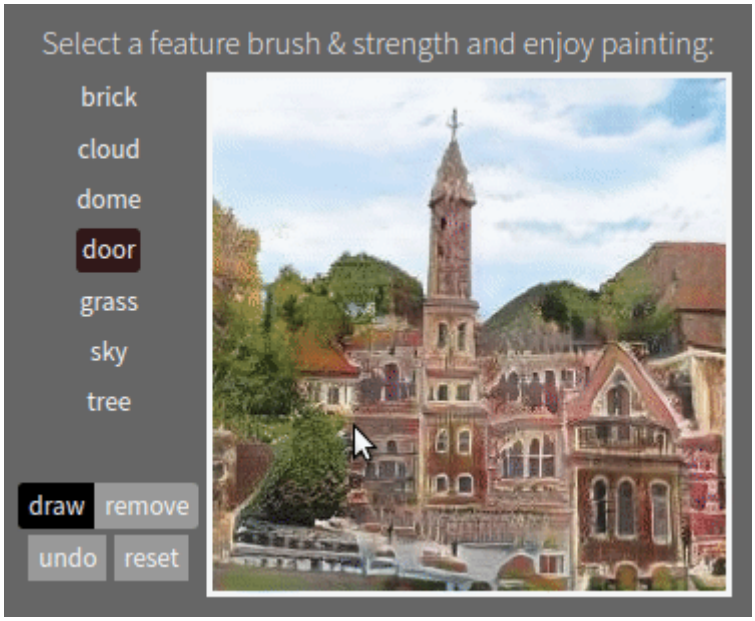


One surprising finding is that the same neurons control a specific object class in a variety of contexts, even if the final appearance of the object varies widely. The same neurons can switch on the concept of a "door" even if a big stone wall requires a big heavy door facing to the left, or a little hut requires a small curtain door facing to the right.

The network also understands when it can and cannot compose objects. For example, turning on neurons for a door in the proper location of a building will add a door. But doing the same in the sky or on a tree will typically have no effect. This structure can be quantified.



Above: yellow boxes highlight the locations of neurons that can be switched on to add a door. One way to make a big door as in (d) is to emphasize a smaller door, but there are many places where the GAN will refuse to draw a door. Below: this is why GAN Paint is not like an ordinary paint program. It does not always do what you want - it wants objects to go in the right place.



Can GAN Mistakes be Debugged and Fixed?

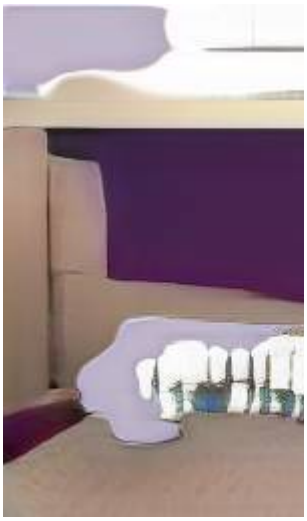
One reason it is important to understand the internal concepts of a network is that the insights can help us improve the network's behavior.

For example, a GAN will sometimes generate terribly unrealistic images, and the cause of these mistakes has been previously unknown. We have identified that these mistakes can be triggered by specific sets of neurons that cause the visual artifacts.

By identifying and silencing those neurons, we can improve the the quality of the output of a GAN.

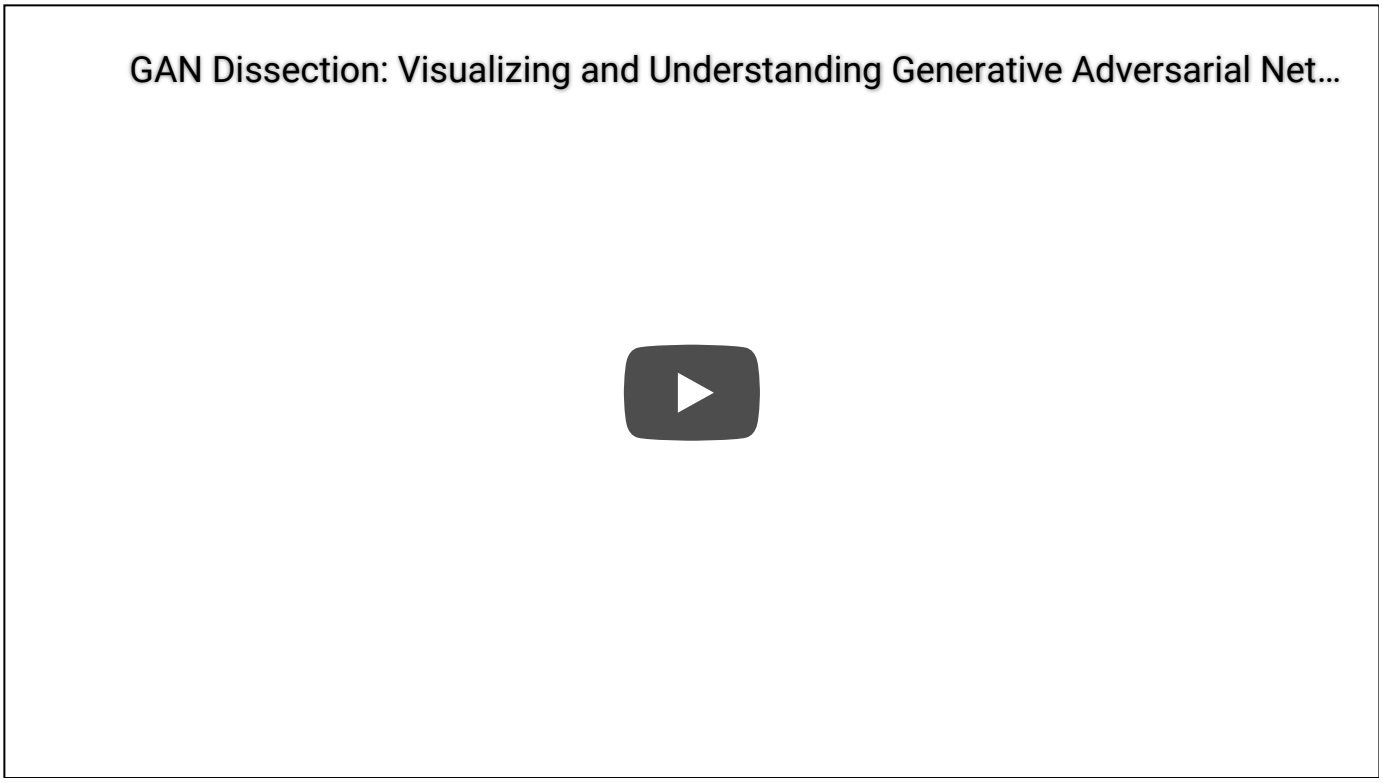
fix artifacts

fix artifacts



fix artifac

Video Demo



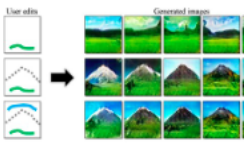
Preprints and Workshops

- GAN Dissection, ArXiv Preprint
- Visualizing and Understanding GANs, AAAI-19 Workshop on Network Interpertability for Deep Learning
- Visualizing and Understanding GANs, New England Computer Vision Workshop 2018

Related Work

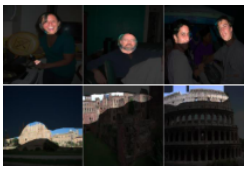
I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *Generative Adversarial Networks*. NIPS 2014. The original GAN paper: explains and establishes soundness of training with an adversarial discriminator; demonstrates the method with MNIST, Faces, and CIFAR.

A. Radford, L. Metz, S. Chintala, *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. ICLR 2016. DCGAN: develops methods for stable training of deep GANs for images, and shows several approaches for visualization of GANs, with filter visualizations of discriminators, and vector arithmetic on latent spaces.



J.Y. Zhu, P. Krähenbühl, E. Shechtman, A. Efros, *Generative Visual Manipulation on the Natural Image Manifold*. ECCV 2016. iGAN: develops a method and system for interactive drawing using a GAN, by optimizing

within the latent space to match user drawings. Our method enables a new approach, drawing with neurons directly rather than solving for each drawing.



D. Bau*, B. Zhou*, A. Khosla, A. Oliva, A. Torralba, *Network Dissection: Quantifying Interpretability of Deep Visual Representations*. CVPR 2017 Net Dissection: quantifies and visualizes the emergence of single-neuron semantic object detectors in classifier CNNs. We build upon these methods and apply them to GANs.

T. Karras, T. Aila, S. Laine, J. Lehtinen, *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. ICLR 2018. Progressive GANS: introduces several training improvements, of which progressive refinement is just one; they also describe minibatch statistics, learning rate equalization, and pixelwise normalization. We analyze several Progressive GANs.

T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, *Unified Perceptual Parsing for Scene Understanding*, ICCV 2018. Semantic Segmentation: state-of-the-art semantic scene segmentation by unified training on scene, object, part, material, and texture labels. We use these pretrained models for labeling the contents of GAN output.

How to cite

Citation

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, Antonio Torralba. *GAN Dissection: Visualizing and Understanding Generative Adversarial Networks,* Proceedings of the International Conference on Learning Representations (ICLR), 2019.

Bibtex

```
@inproceedings{bau2019gandissect,
 title={GAN Dissection: Visualizing and Understanding Generative Adversarial
Networks},
 author={Bau, David and Zhu, Jun-Yan and Strobelt, Hendrik and Zhou, Bolei and
Tenenbaum, Joshua B. and Freeman, William T. and Torralba, Antonio},
 booktitle={Proceedings of the International Conference on Learning
Representations (ICLR)},
```

```
 year={2019}
}
```

Also Cite

We more recently published a journal article in the Proceedings of the National Academy of Sciences:

Citation

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. *Understanding the role of individual units in a deep neural network.* Proceedings of the National Academy of Sciences (2020).

Bibtex

```
@article{bau2020units,
  author = {Bau, David and Zhu, Jun-Yan and Strobelt, Hendrik and Lapedriza,
Agata and Zhou, Bolei and Torralba, Antonio},
  title = {Understanding the role of individual units in a deep neural network},
  elocation-id = {201907375},
  year = {2020},
  doi = {10.1073/pnas.1907375117},
  publisher = {National Academy of Sciences},
  issn = {0027-8424},
  URL = {https://www.pnas.org/content/early/2020/08/31/1907375117},
  journal = {Proceedings of the National Academy of Sciences}
}
```

Acknowledgments: We would like to thank Zhoutong Zhang, Guha Balakrishnan, Didac Suris, Adrià Recasens and Zhuang Liu for valuable discussions. We are grateful for the support of the MIT-IBM Watson AI Lab, the DARPA XAI program FA8750-18-C000, NSF 1524817 on Advancing Visual Recognition with Feature Visualizations, NSF BIGDATA 1447476, and a hardware donation from NVIDIA.