Understanding Neural Networks through Deep Visualization

Team 25: Pixel Ninjas

Karthik Prasanna N
2020115007

Sreya Garapati
2020102055

Shruti Kolachana
2020102053

Neural Networks

01

seen as "black boxes" : lack interpretability

02

difficult to understand how they arrive at decisions

03

limits ability to diagnose and correct errors

Proposed solutions

Visualizing activations produced on each layer of a trained convnet

Visualizing features at each layer of a DNN via regularized optimization in image space

Implementation Details

Visualisation of feature maps of intermediate layers of standard models like VGG

Visualisation of activations of a given layer using Regularisation methods

---

Novel Implementations:

1. Visualisation of activations of a given layer using gradient ascent and comparison with the regularised outputs
2. Visualisation using inverted representation of an input image
3. Visualizing problems caused by simple gradient-based methods

Visualizing Activations

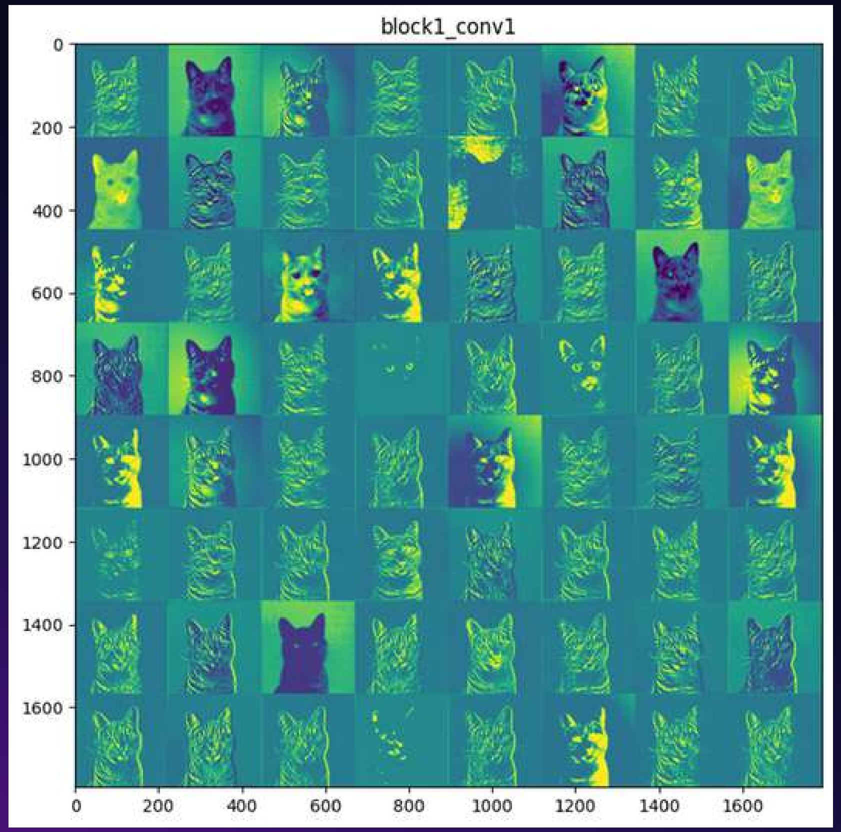
01

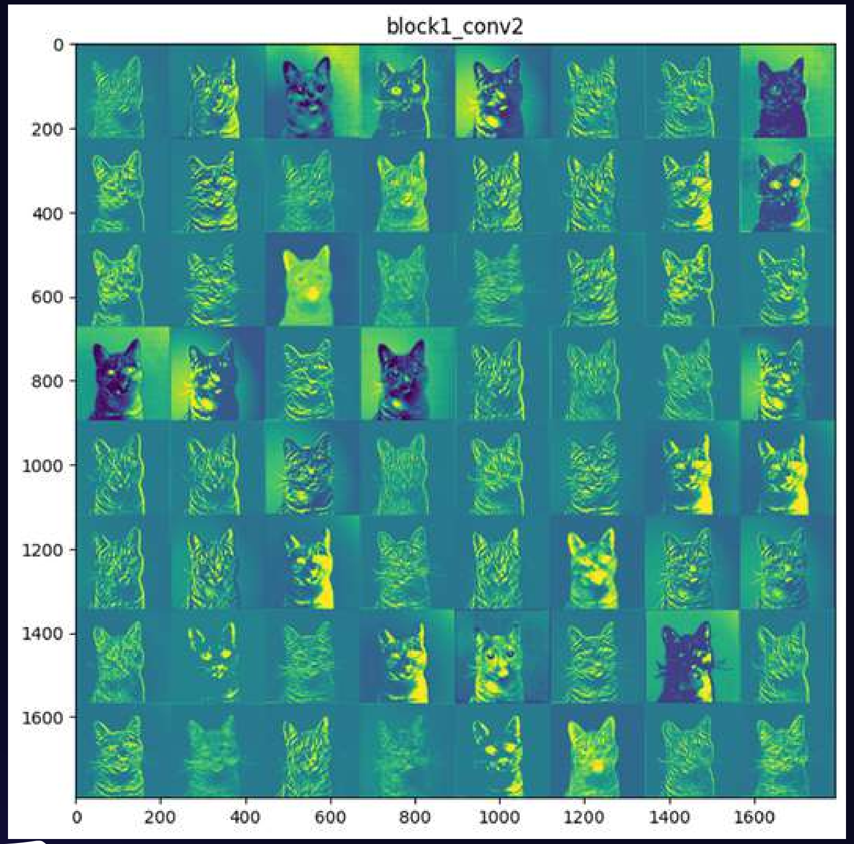
Retrive activation values for neurons in each layer of a convnet in response to an input image

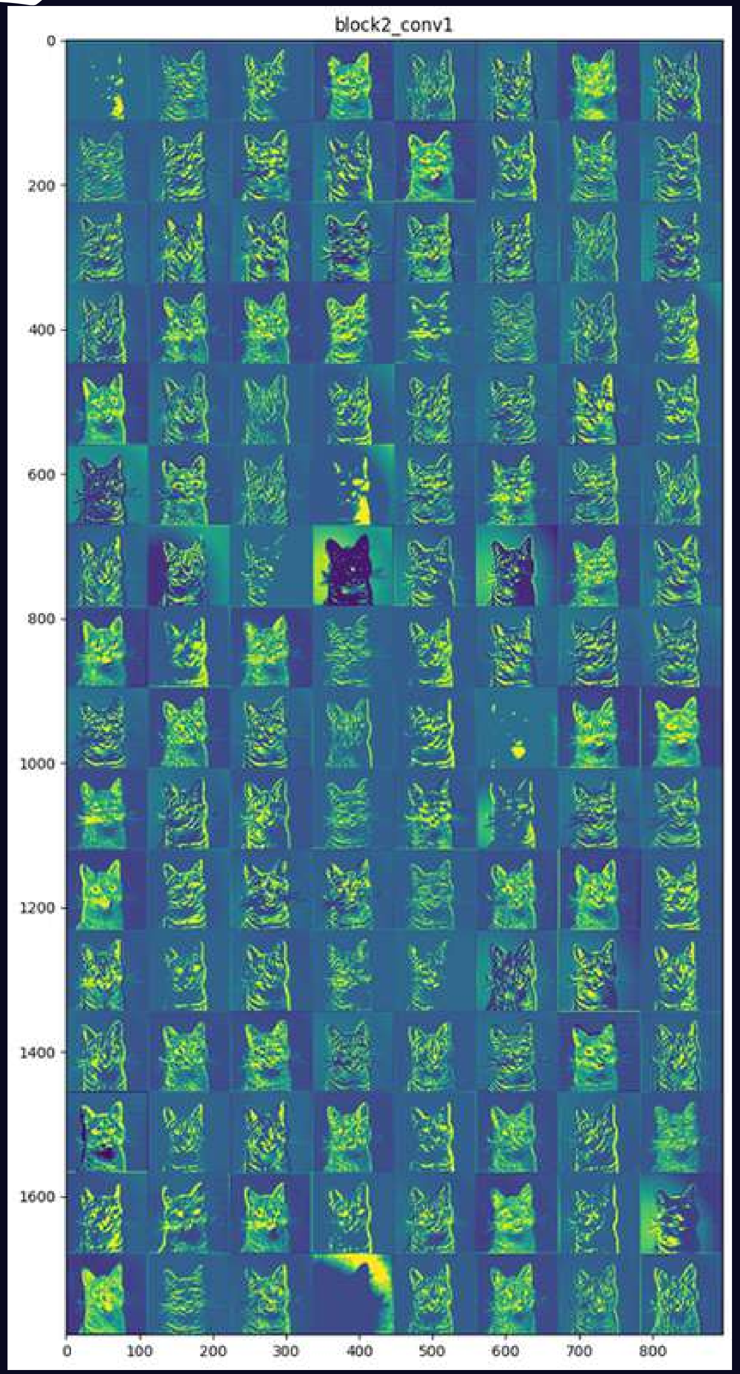
02

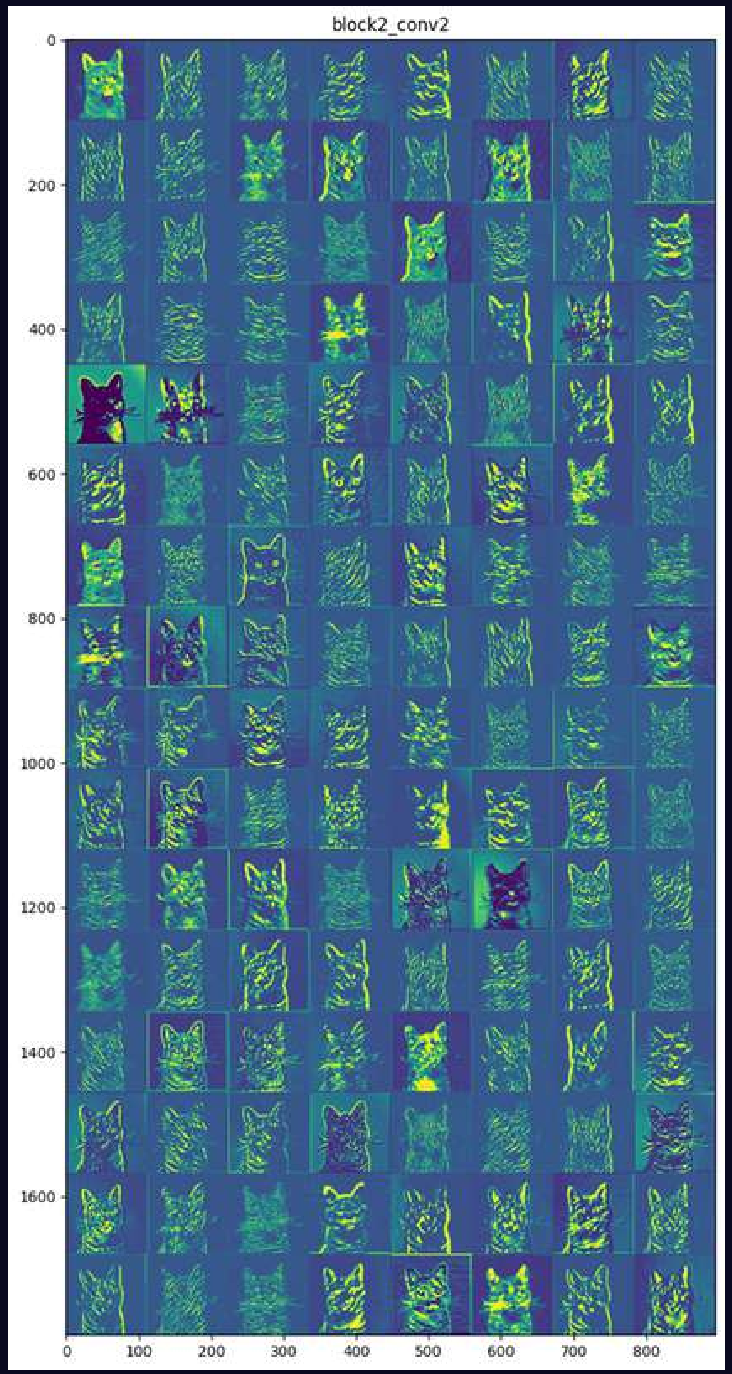
Arrange plots spatially

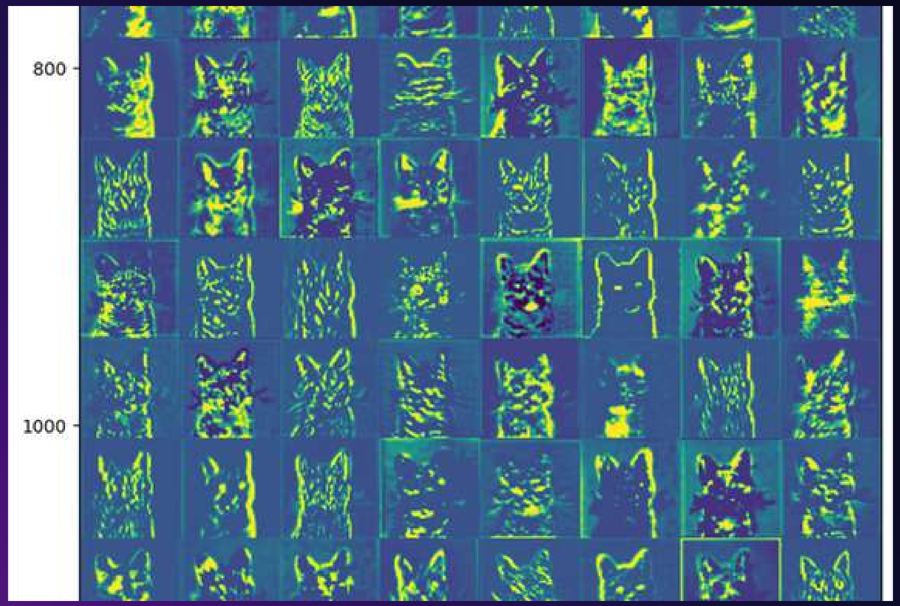
Results: VGG16

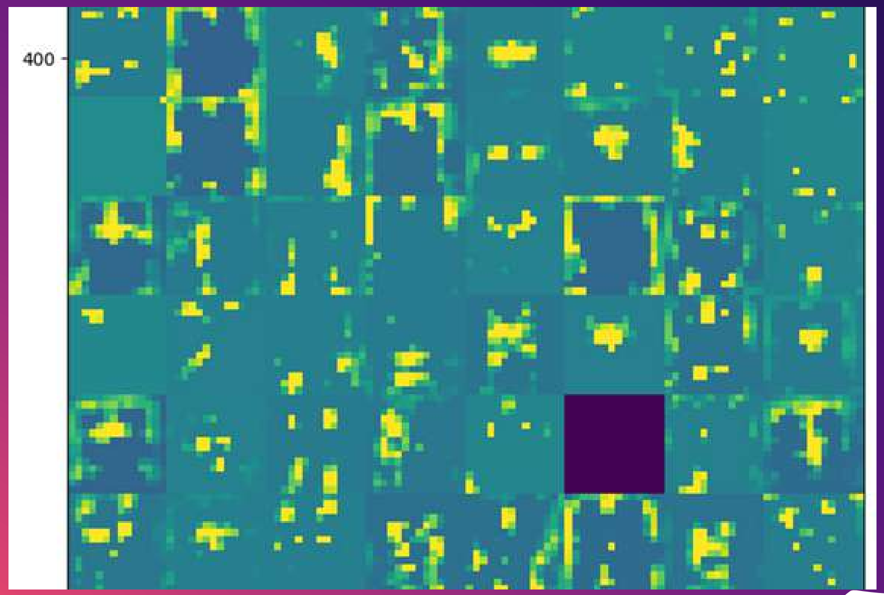




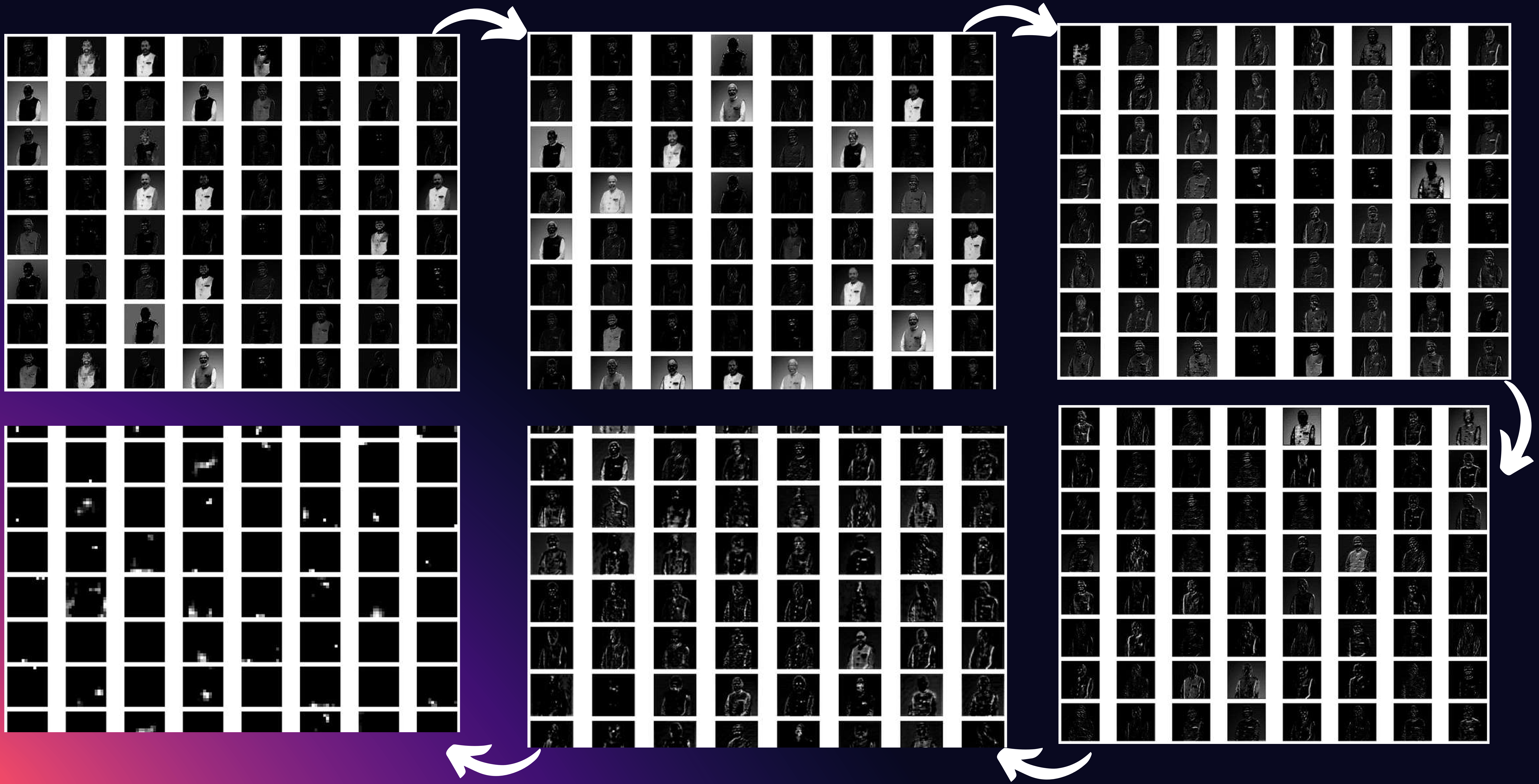




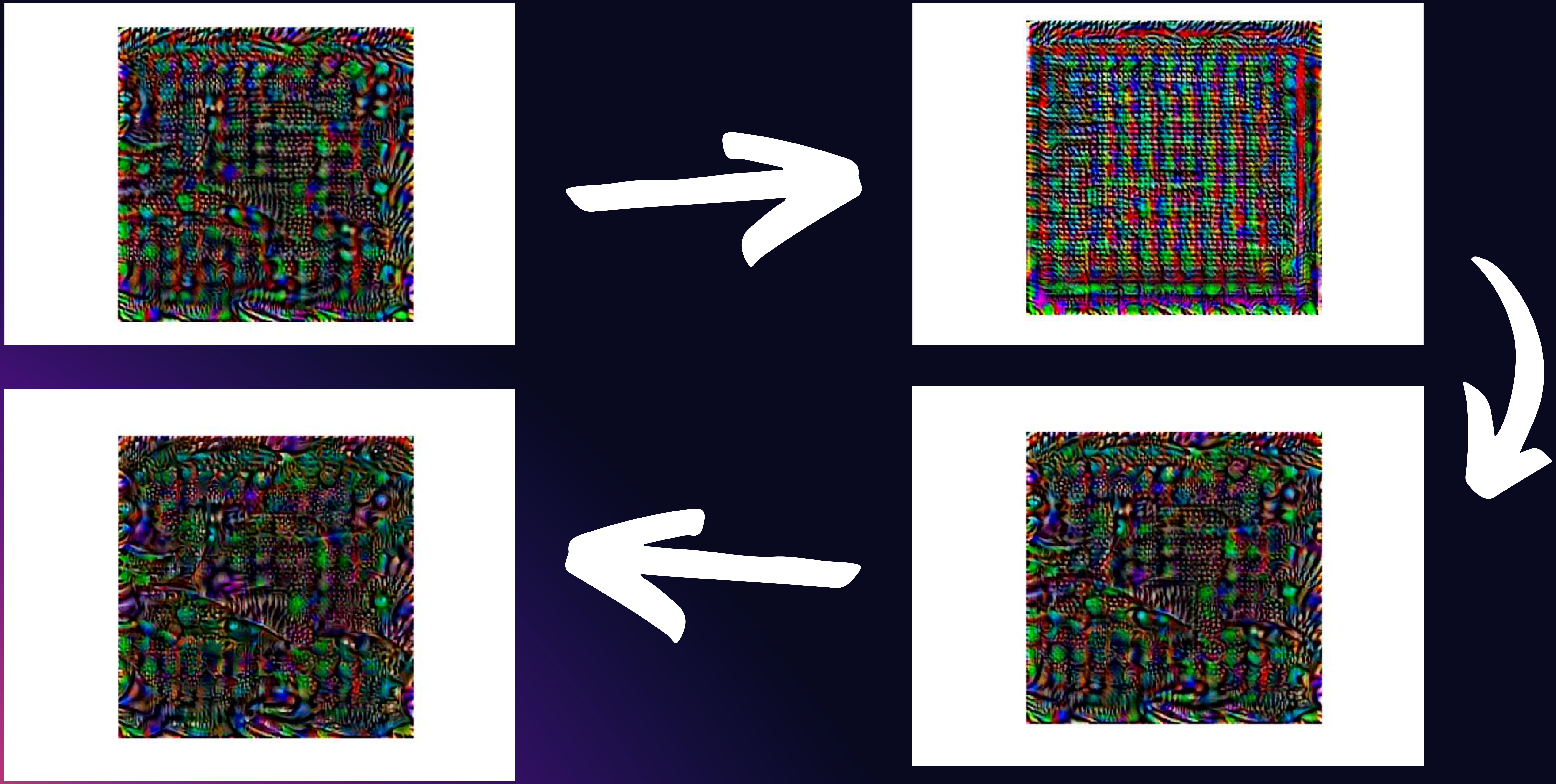




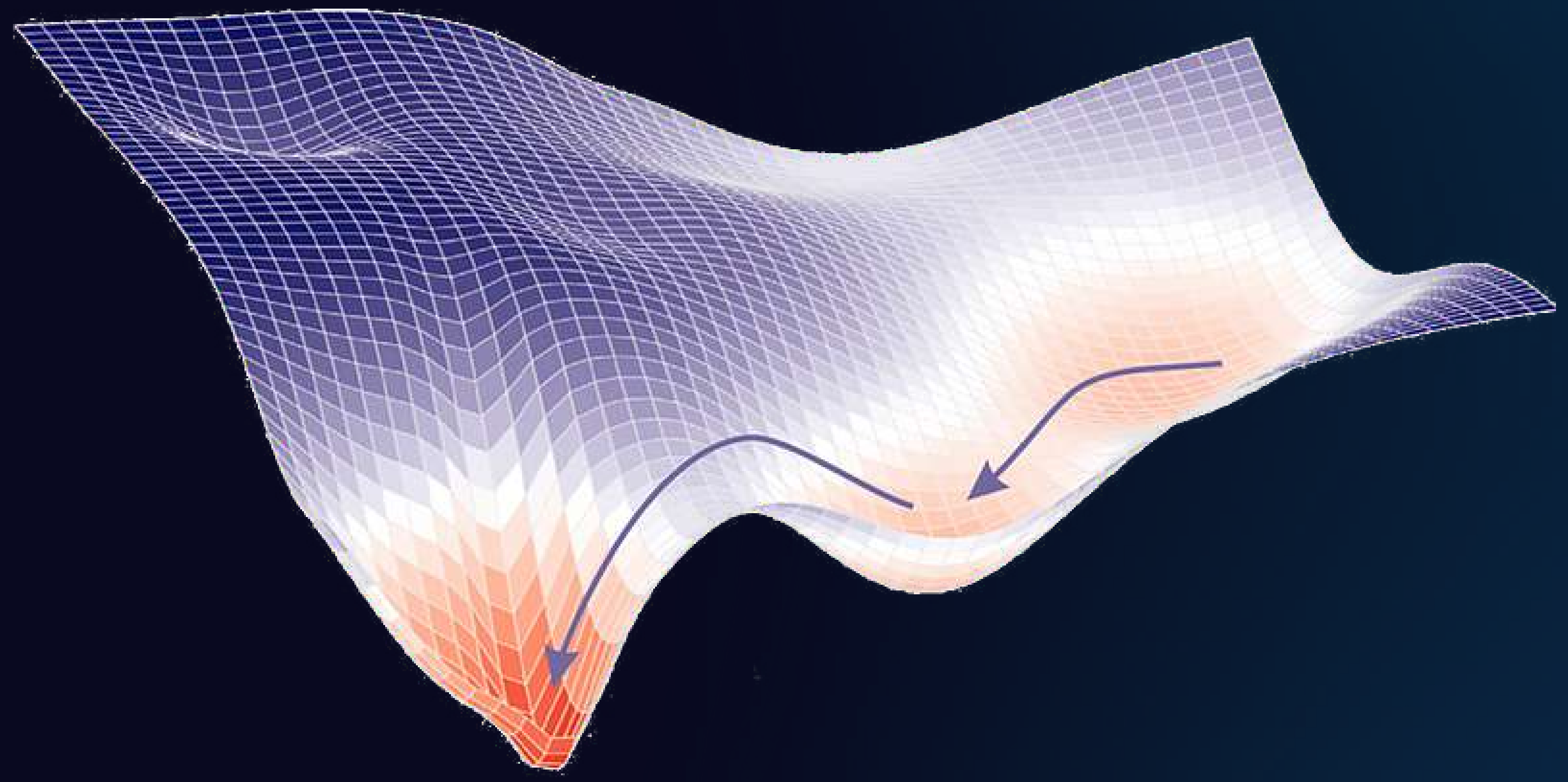
Results: VGG19



Results: Gradient Ascent



Gradient-based Approaches



High frequency patterns

Extreme pixel vaues

Copies of common motifs without global structure

Adversarial Attacks: Fast Gradient Sign Method

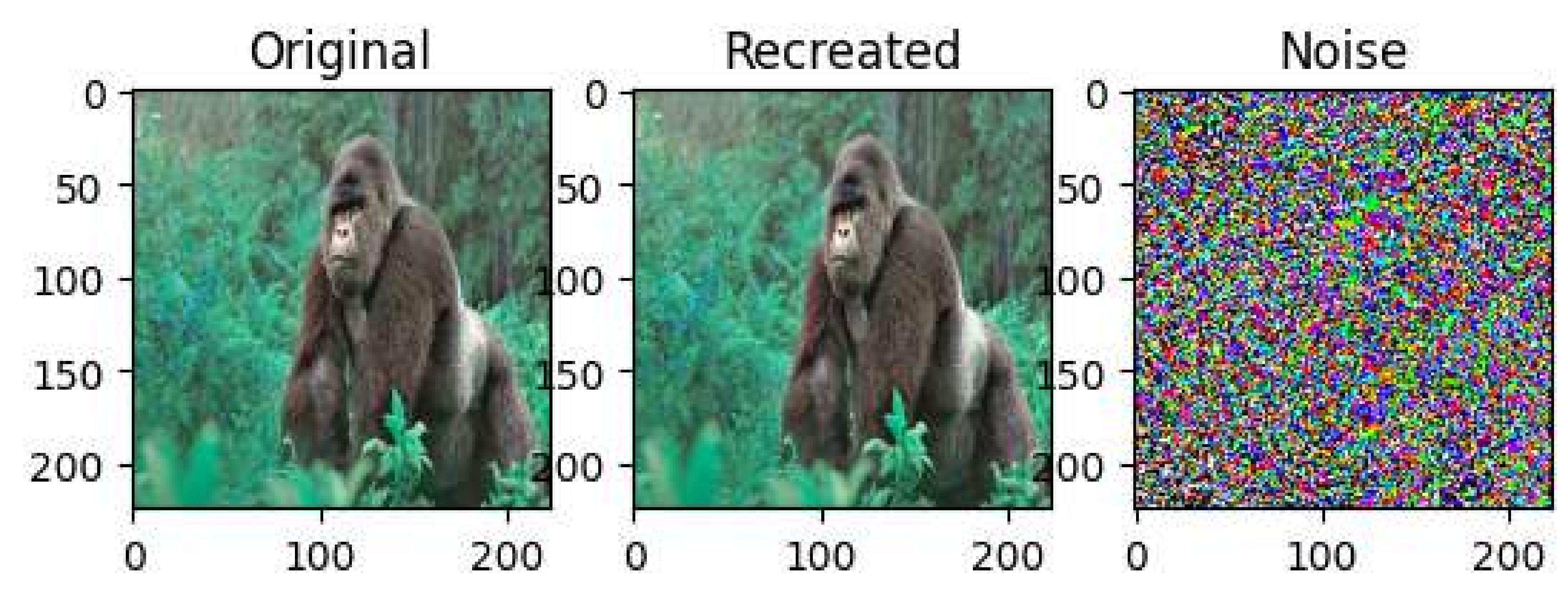
01

Computes the gradients of a loss function with respect to an input image

02

Uses the sign of the gradients to create a new image that maximizes the loss

Results: VGG16




```
No. of Iters: 0 Real Label: 366 Predicted Label: 109 Probability: 0.49386105
No. of Iters: 1 Real Label: 366 Predicted Label: 109 Probability: 0.62703884
No. of Iters: 2 Real Label: 366 Predicted Label: 109 Probability: 0.6902871
No. of Iters: 3 Real Label: 366 Predicted Label: 109 Probability: 0.8680306
No. of Iters: 4 Real Label: 366 Predicted Label: 109 Probability: 0.9437923
No. of Iters: 5 Real Label: 366 Predicted Label: 109 Probability: 0.98842514
No. of Iters: 6 Real Label: 366 Predicted Label: 109 Probability: 0.997454
No. of Iters: 7 Real Label: 366 Predicted Label: 109 Probability: 0.9995827
No. of Iters: 8 Real Label: 366 Predicted Label: 109 Probability: 0.99994504
No. of Iters: 9 Real Label: 366 Predicted Label: 109 Probability: 0.9999893
```

Visualizing via Regularized Optimization

01 **Generate random image**

02 **Get Activation**

03 **Gradient ascent**

+

Regularization in image space

$$\mathbf{x} \leftarrow r_\theta\left(\mathbf{x} + \eta \frac{\partial a_i}{\partial \mathbf{x}}\right)$$

Regularization Methods

- **L2 Decay :** penalizes large values

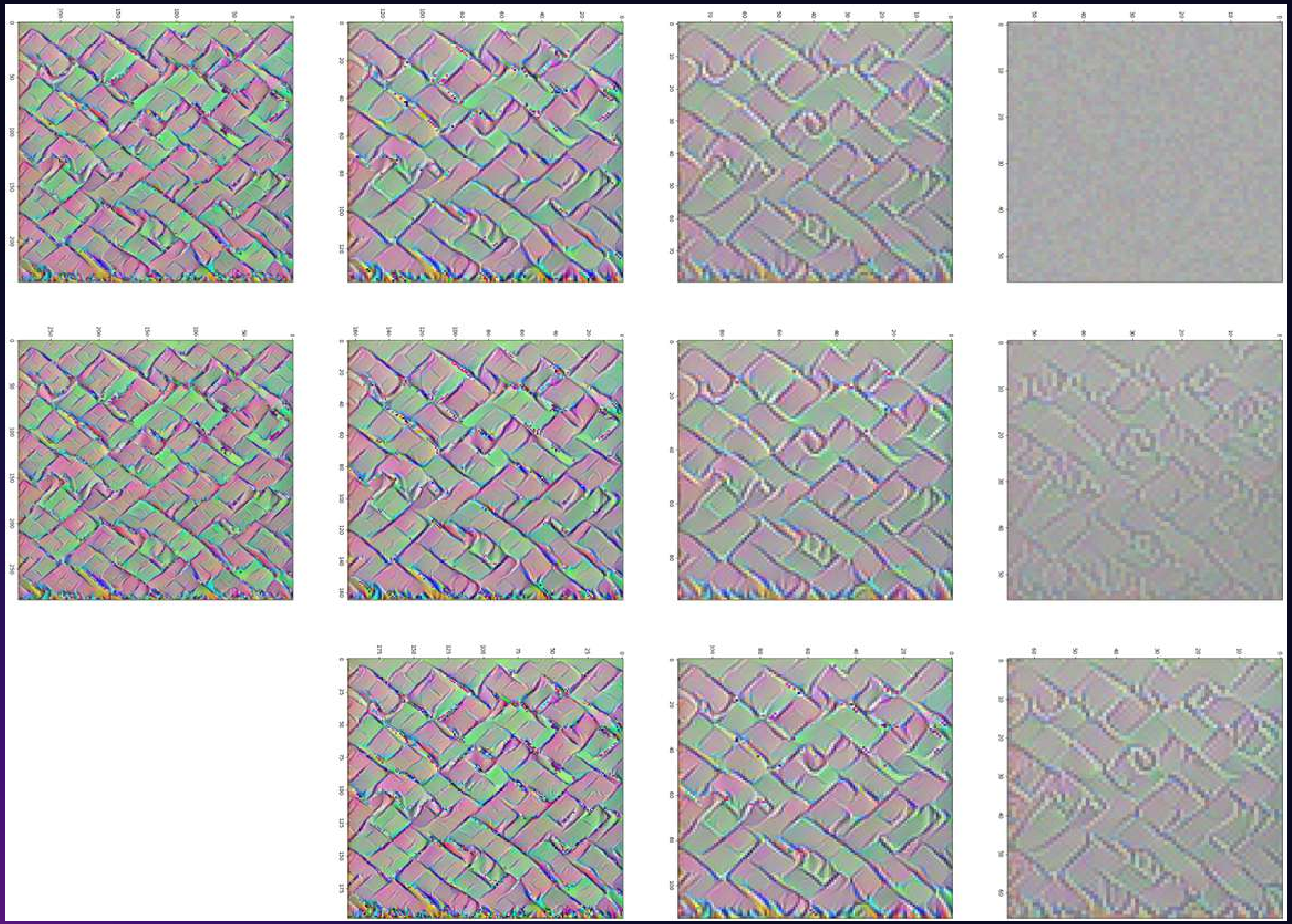
$$r_\theta(x) = (1-\theta_{decay}) \cdot x$$

- **Gaussian Blur:** penalizes high frequency information

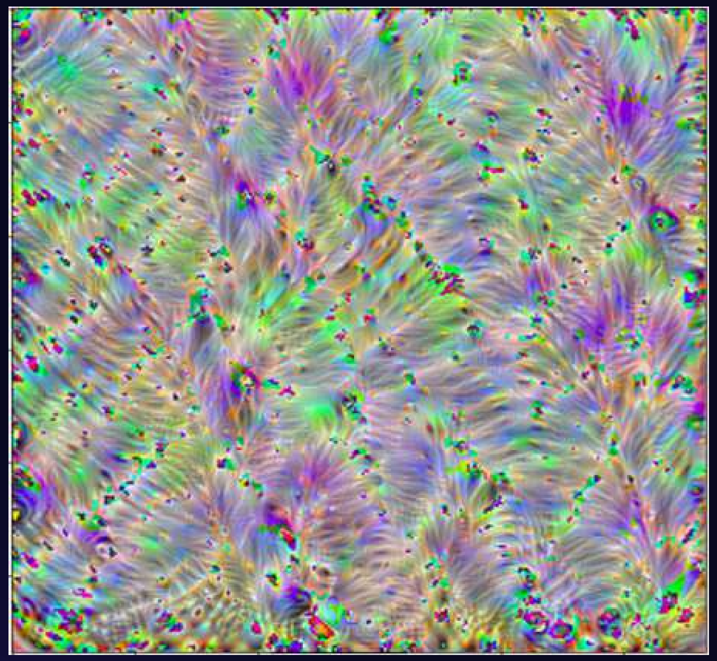
$$r_\theta(x) = \mathrm{Blur}(x, \theta_{b\_width})$$

- **Clipping small norm:** sets pixels with small norm (over RGB channels) to zero
- **Clipping small contribution:** computes $|\, x \cdot \nabla_x a_i(x) \,|$ (over RGB channels)

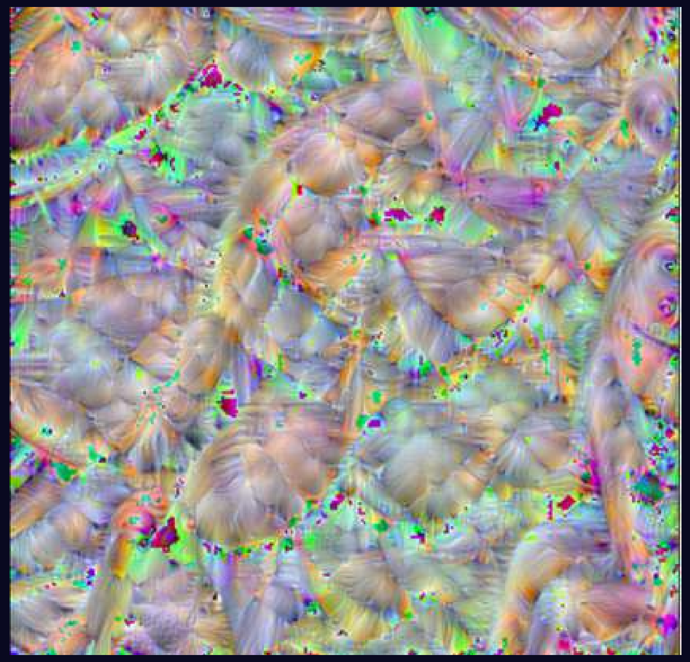
Results: VGG16-11



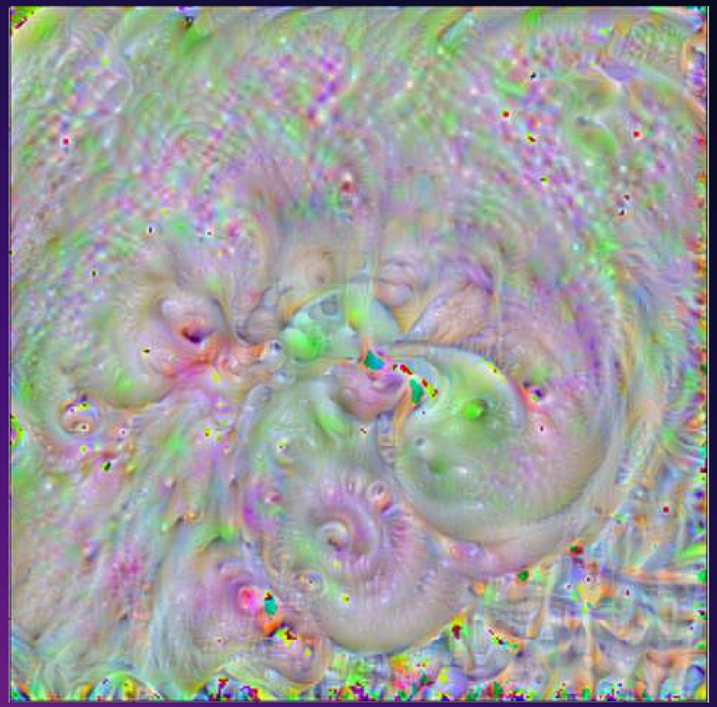
Results: VGG16



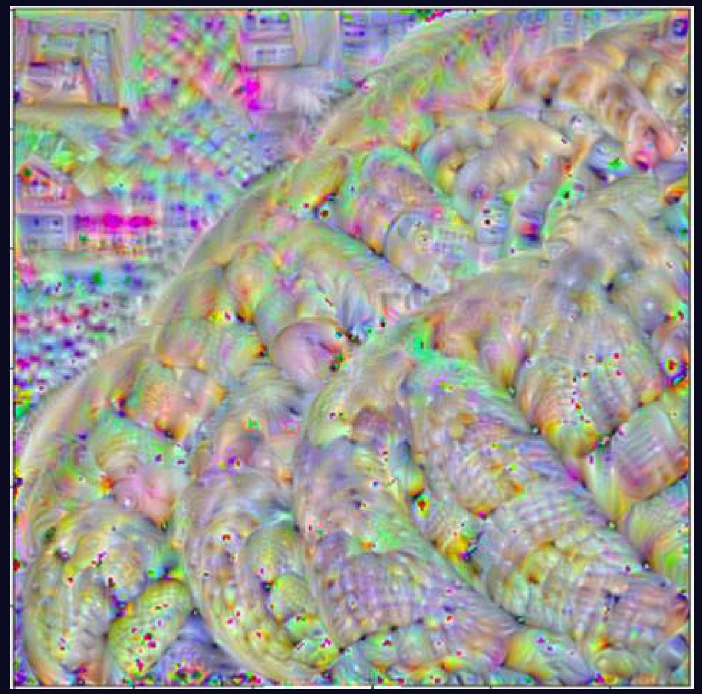
20



25

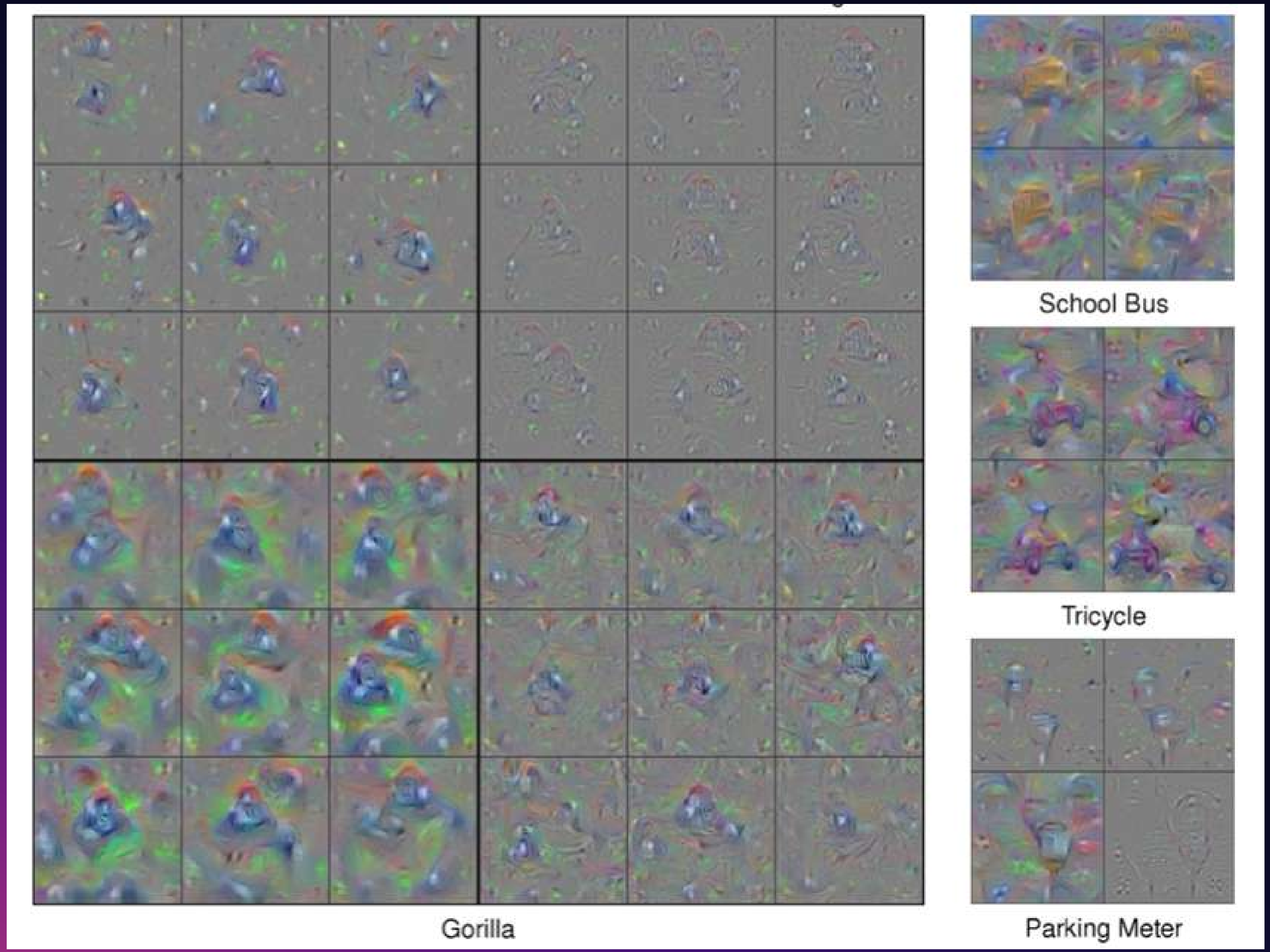


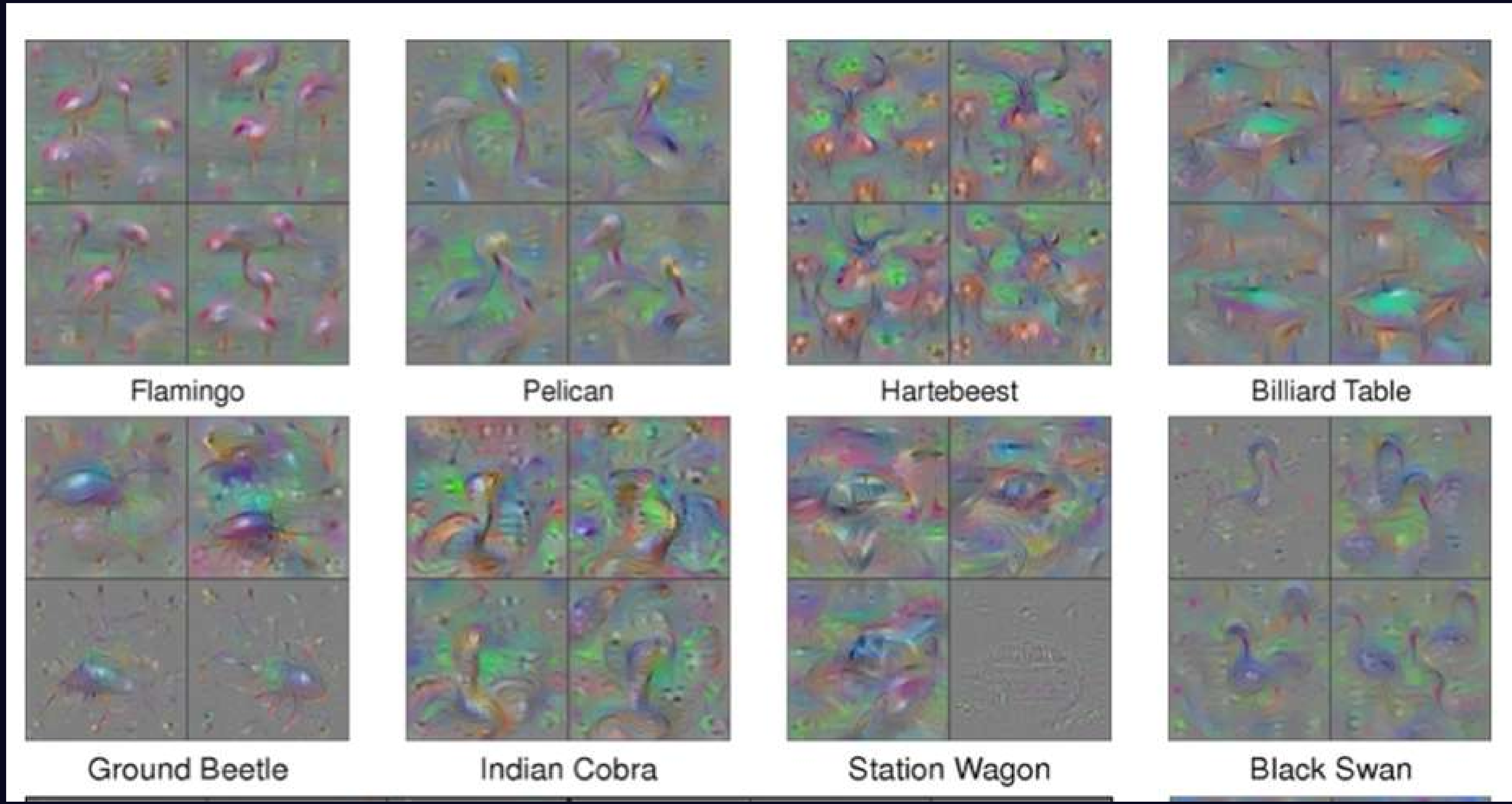
38



31

Results: Paper





Visualizing Inverted Image Representations

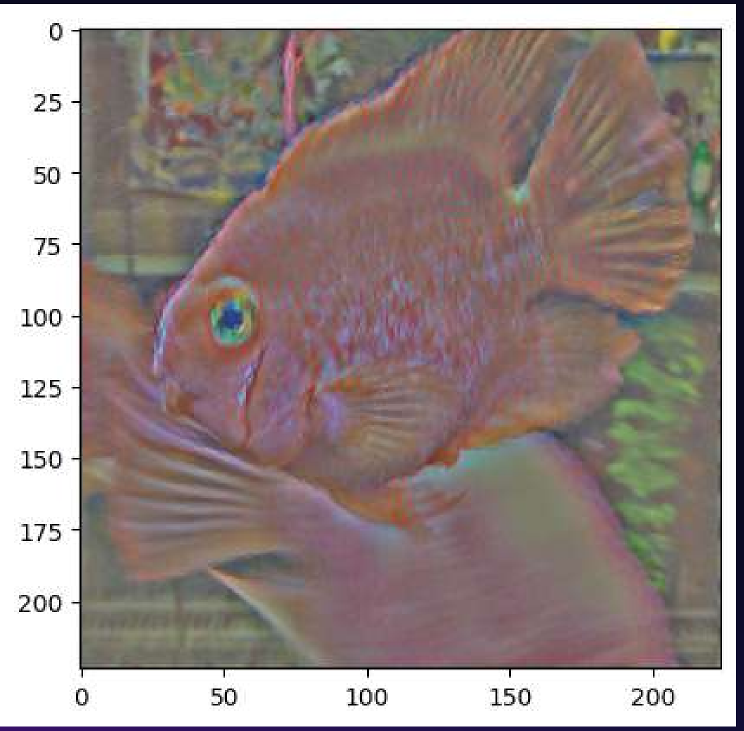
01 Generate random image

02 Get Image Representation from target layer

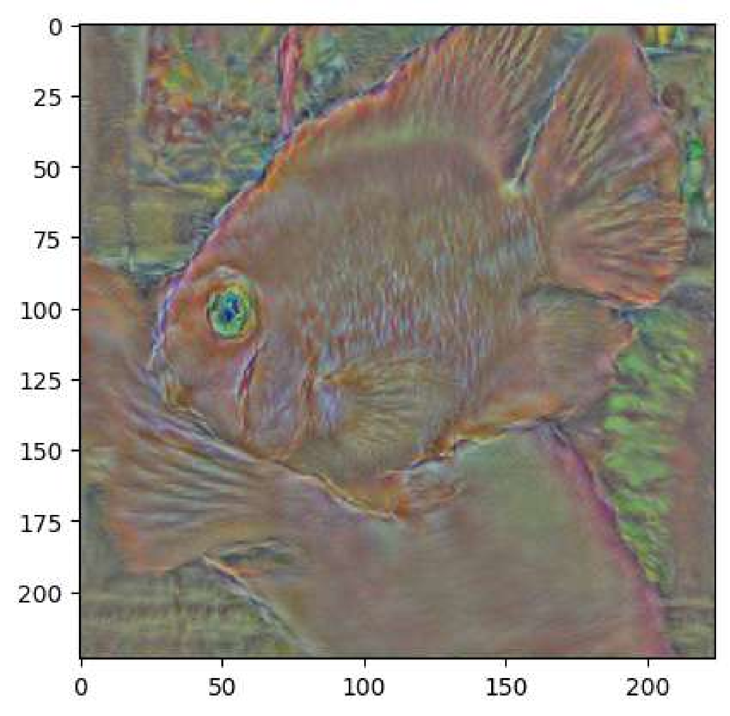
03 Gradient descent + Regularization in image space

$$\|\Phi(\sigma\mathbf{x}) - \Phi_0\|_2^2/\|\Phi_0\|_2^2 + \lambda_\alpha \mathcal{R}_\alpha(\mathbf{x}) + \lambda_{V^\beta}\mathcal{R}_{V^\beta}(\mathbf{x})$$

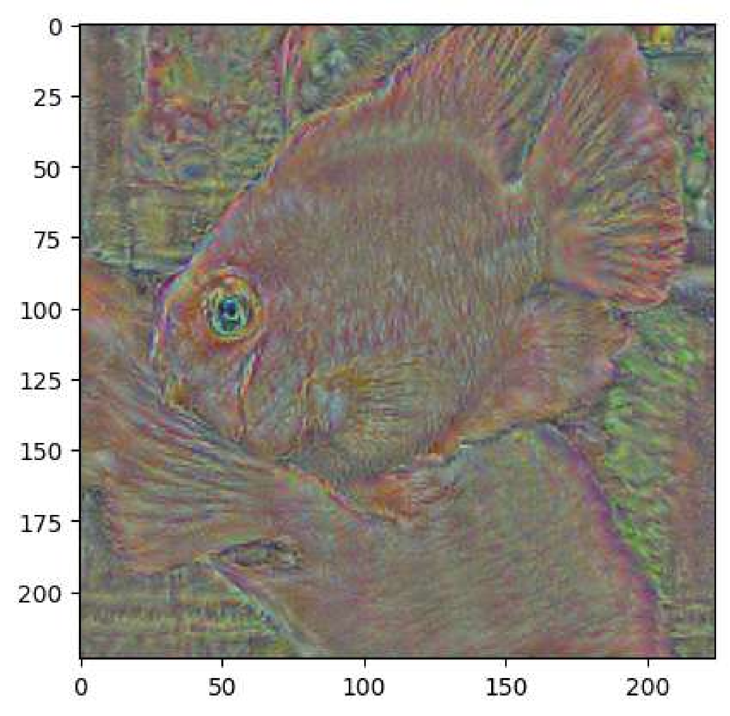
Results: VGG16



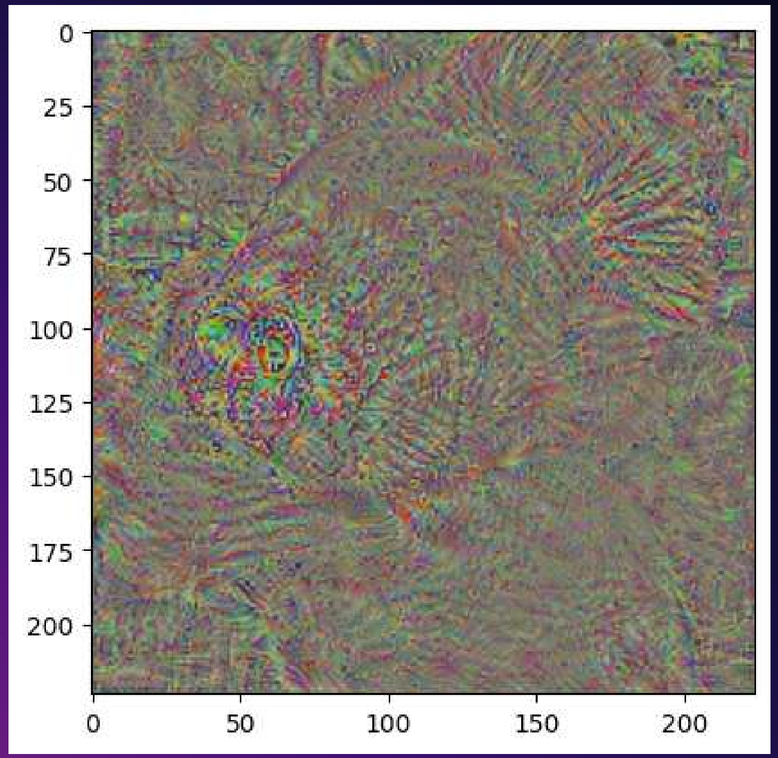
Layer 5



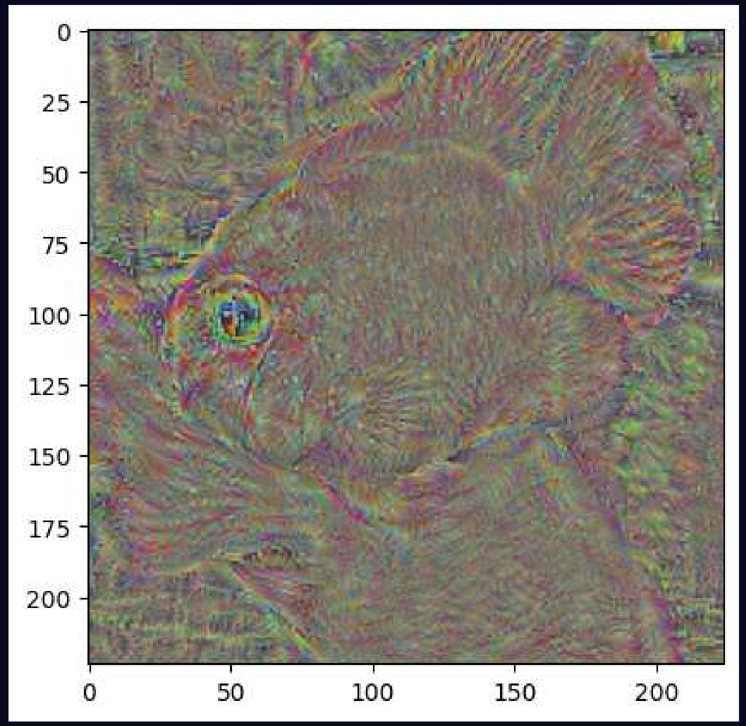
Layer 10



Layer 15



Layer 25



Layer 20

Main References

Understanding Neural Networks Through Deep Visualization
Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, Hod Lipson

---

Understanding Deep Image Representations by Inverting Them
Aravindh Mahendran, Andrea Vedaldi

Explaining and Harnessing Adversarial Examples
Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy

THANK YOU!