ImageNet Supplementary Material

Description

This document contains many examples of context-sensitive ImageNet feature perturbations of different granularities. The example given are the same as those used in our experiments. Since our experiments evaluate four classifiers, each example consists of a row of images per classifier. Each row of five images consists of the original, unperturbed, generated image, followed by generated images for which the generators latent activations were perturbed. These perturbations are optimised to ensure that the classifier (whose name is at the top-left of the row) predicts a randomly pre-chosen incorrect target label for the perturbed image. The top-right of the row shows the original (correct) label and the target (incorrect) label.

By performing the activation perturbations to different sets of layers in the generator, the granularity of the features affected is influenced. The set of layers whose activations are perturbed is noted underneath each image. Above each image is the pixel-wise difference with the original image. In some cases, the magnitude of this pixel-wise difference is increased by a multiplicative factor, so that the changes are more easily visible; any such scale factors are indicated above the pixel-wise difference in question.

To Note

Note that for the 'robust' classifiers—those adversarially trained against pixel-space perturbations—the pixel-wise differences are not magnified by a scale factor, whereas this scaling is necessary for the standard ResNet50 and the state-of-the-art EfficientNet, especially at the finer-grained feature perturbations. This is likely because the pixel-robust models are also more robust to finer-grained feature perturbations, requiring larger (and so more visible) perturbations.

Note also that the examples here are not filtered to remove incorrect original images. Conversely, in our experiments, the first step was to filter out any images for which either the classifier or the human judges do not assign the desired initial label.

These results are mainly interesting to get a feel for the differences between classifiers and between perturbation types. In short, the robust classifiers adversarially trained against pixel-space perturbations are noticeably less robust to coarser-grained feature perturbations, but more robust to the finer-grained changes. And the difference in downstream feature granularity between perturbations to different generator layers is quite easily seen.

These results are less useful for getting a feel for the kinds of features affected by the perturbations to latent generator activations. An intuition for this is best built by viewing our animations that show the effect of gradually introducing these perturbations; these can be found in the animations supplementary material.

ResNet50

‘Komodo dragon’ → ‘overskirt’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

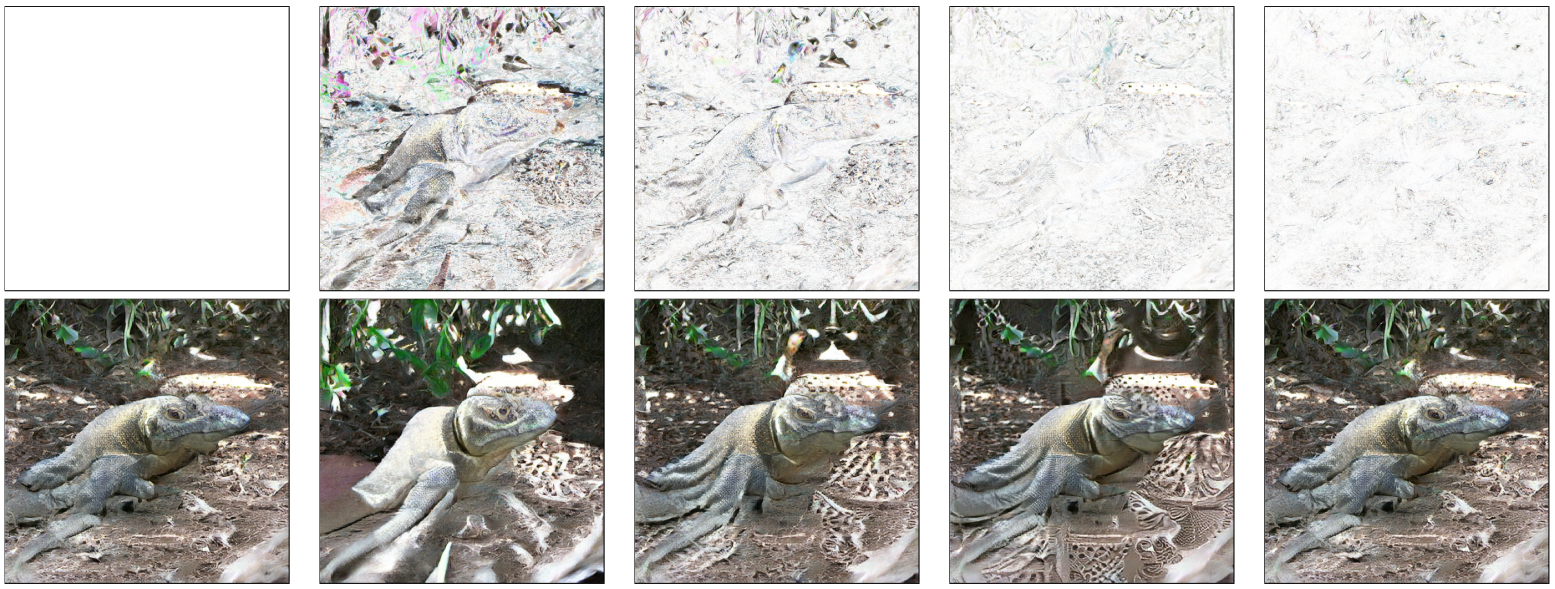
EfficientNet-B4NS

‘Komodo dragon’ → ‘overskirt’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'Komodo dragon' → 'overskirt'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Fast") 'Komodo dragon' → 'overskirt'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

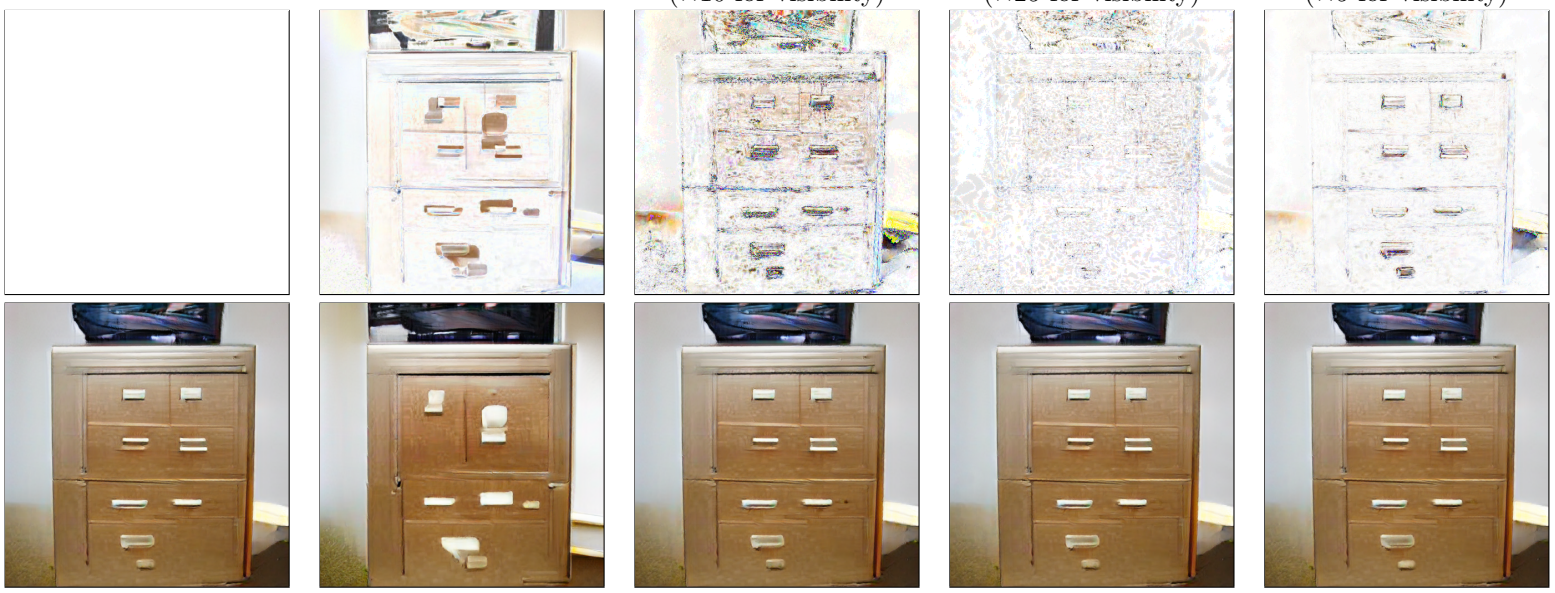
‘file’ → ‘doormat’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

EfficientNet-B4NS

‘file’ → ‘doormat’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'file' → 'doormat'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

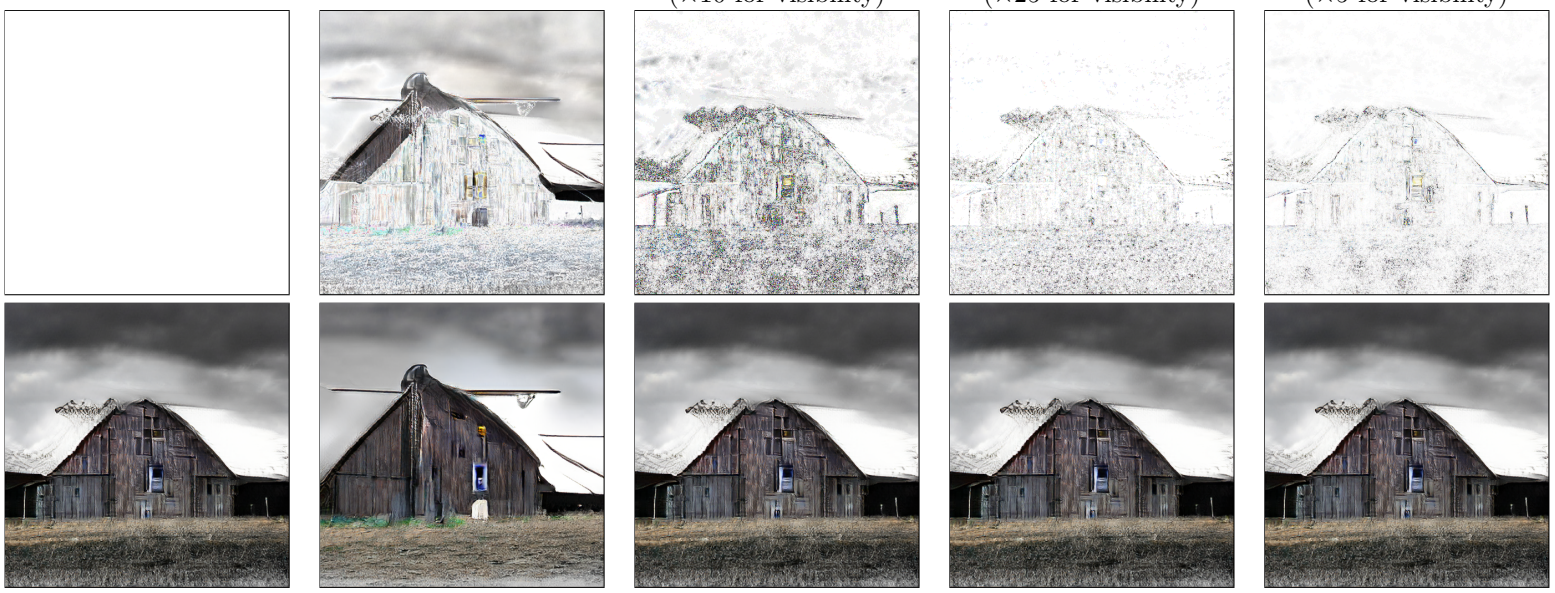
ResNet50 Robust ("Fast") 'file' → 'doormat'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

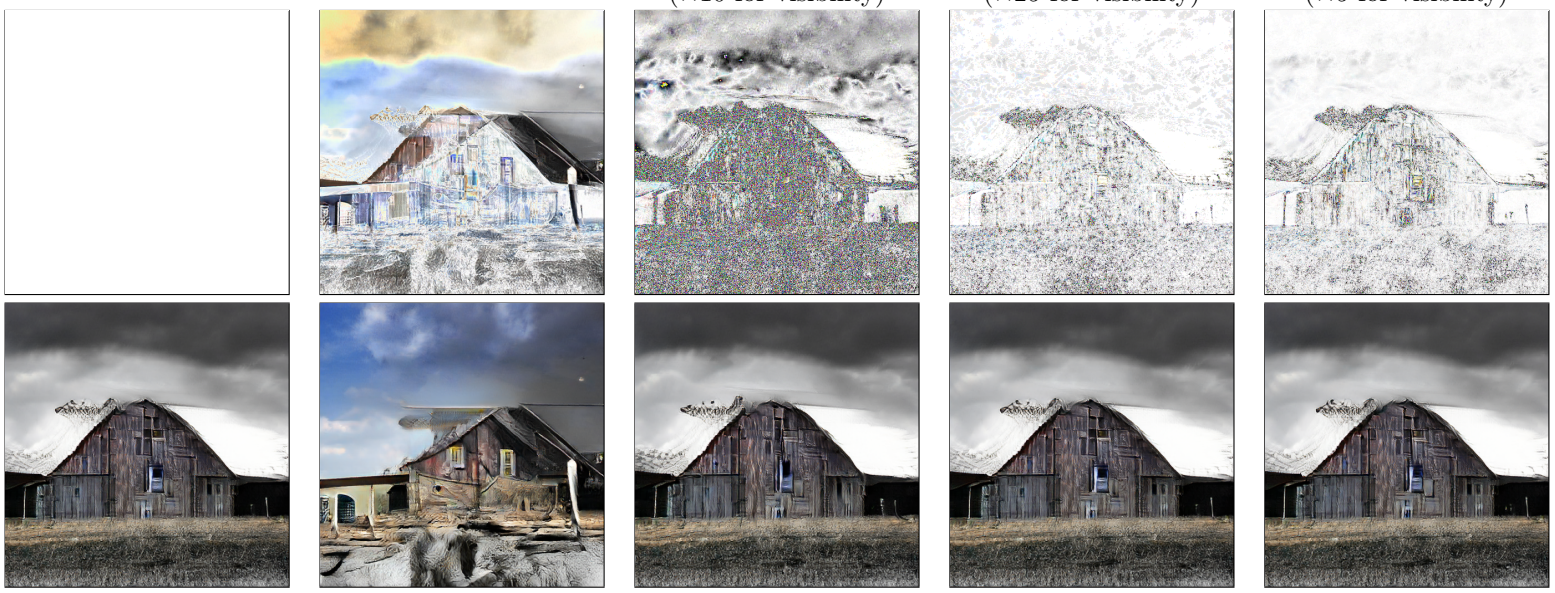
ResNet50

‘barn’ → ‘Afghan hound’

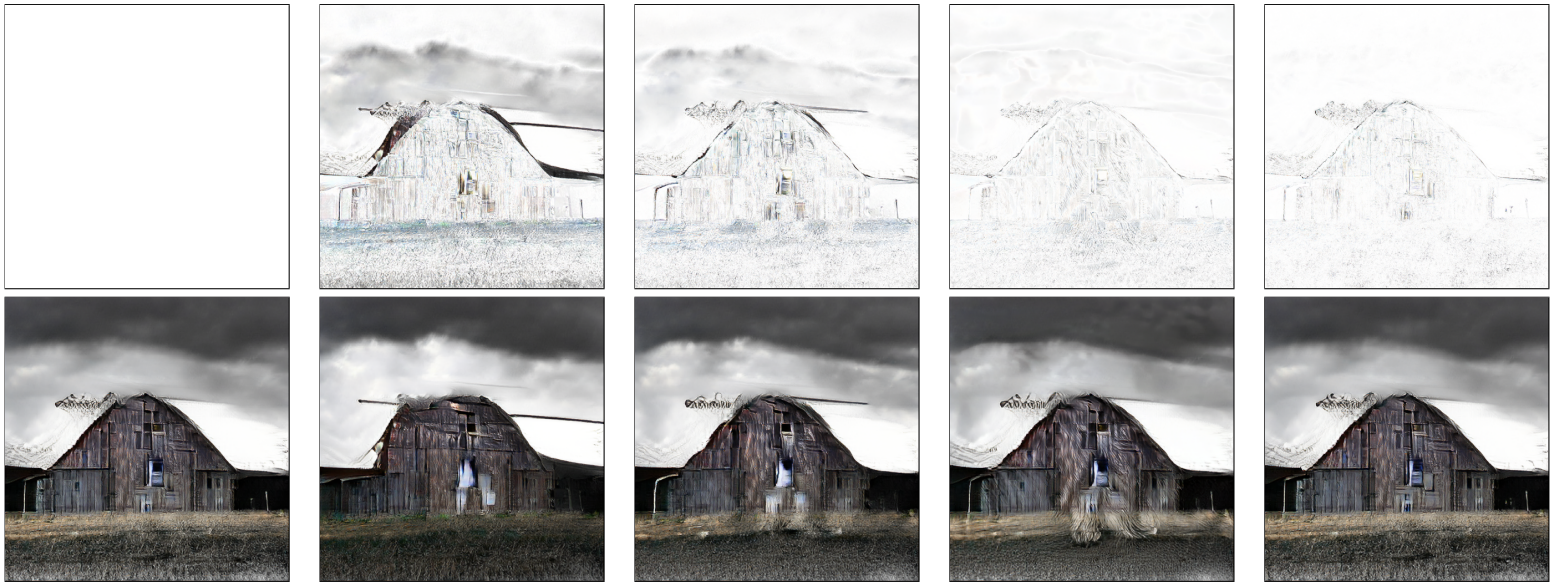


EfficientNet-B4NS

‘barn’ → ‘Afghan hound’

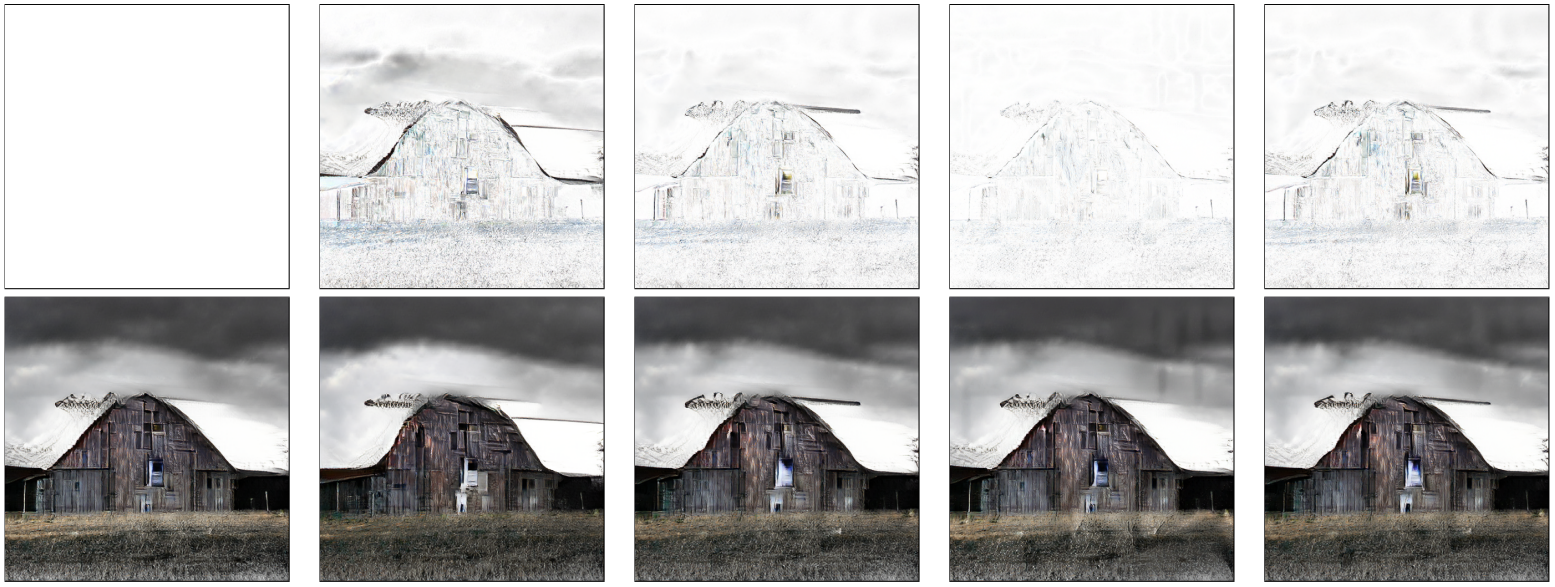


ResNet50 Robust ("Engstrom") 'barn' → 'Afghan hound'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Fast") 'barn' → 'Afghan hound'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

‘palace’ → ‘throne’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

EfficientNet-B4NS

‘palace’ → ‘throne’



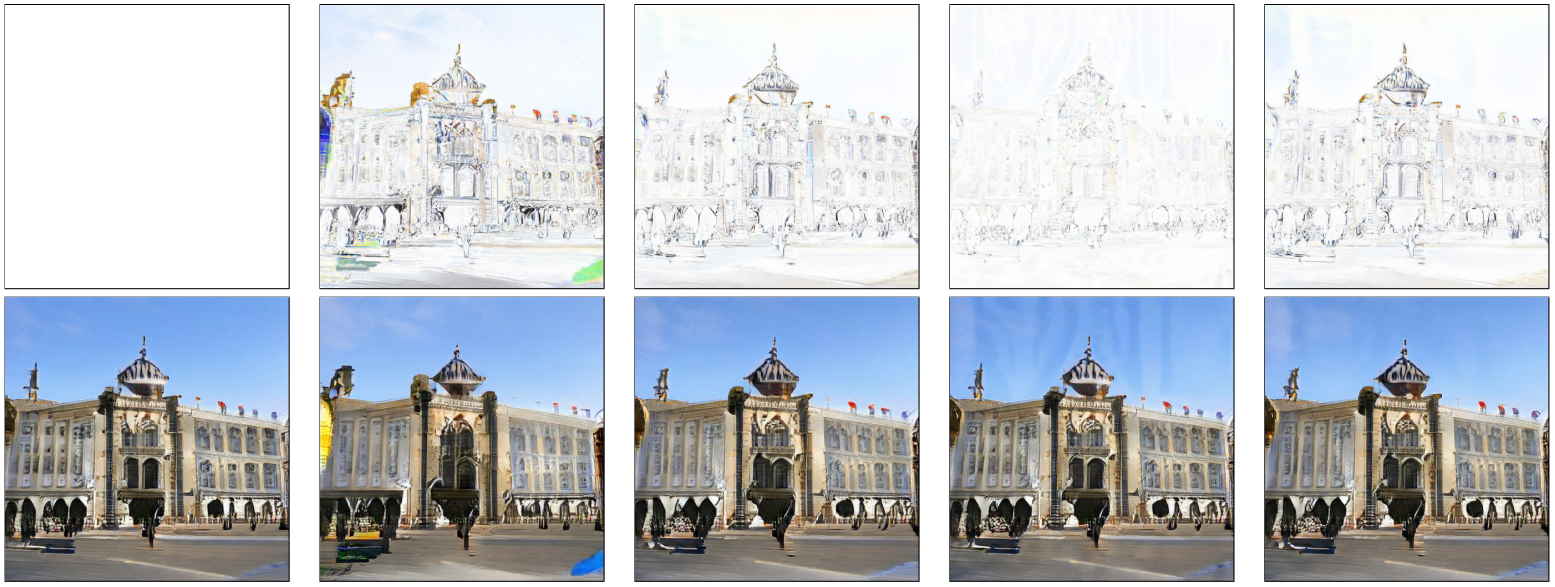
original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'palace' → 'throne'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

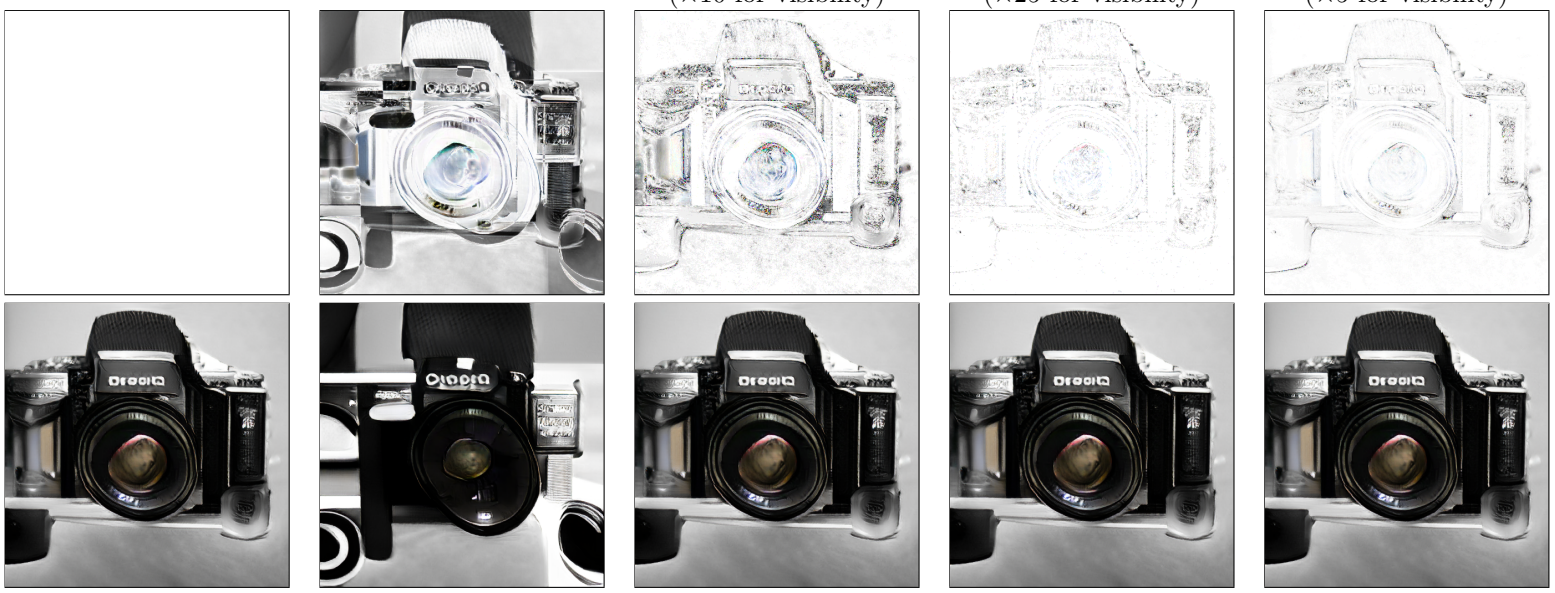
ResNet50 Robust ("Fast") 'palace' → 'throne'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

‘reflex camera’ → ‘sunglass’



(×10 for visibility) (×25 for visibility) (×5 for visibility)

original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

EfficientNet-B4NS

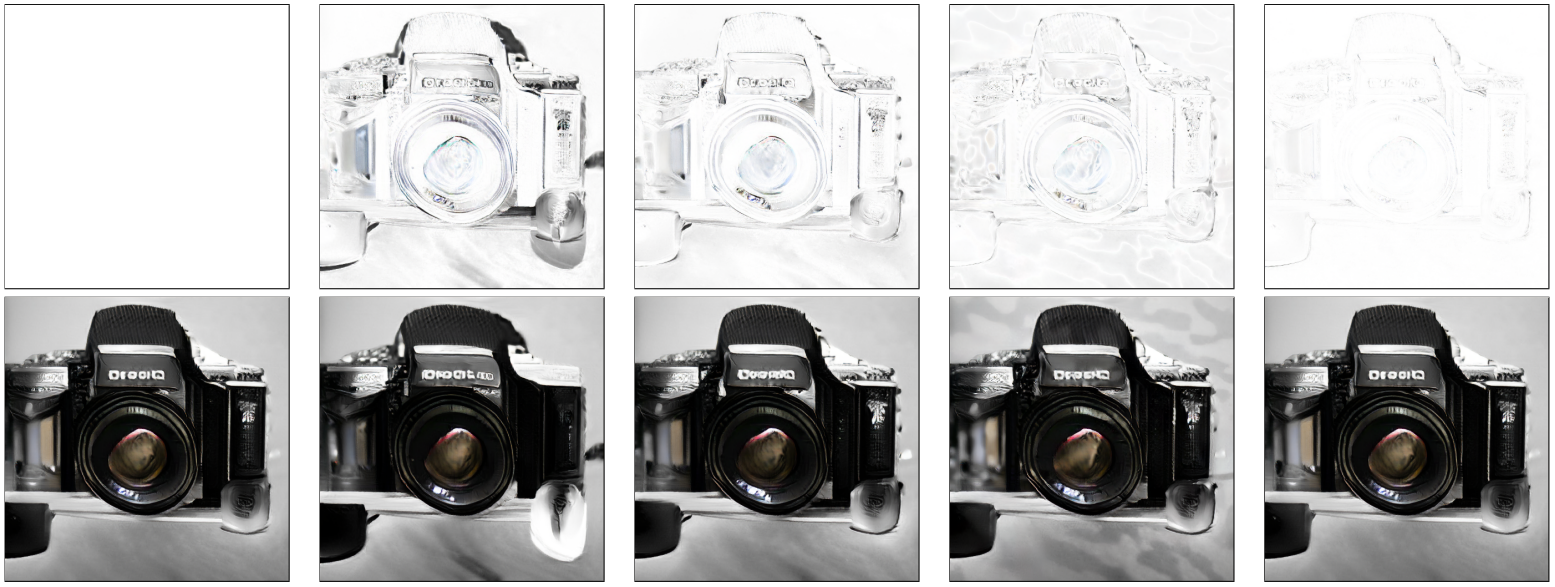
‘reflex camera’ → ‘sunglass’



(×10 for visibility) (×25 for visibility) (×5 for visibility)

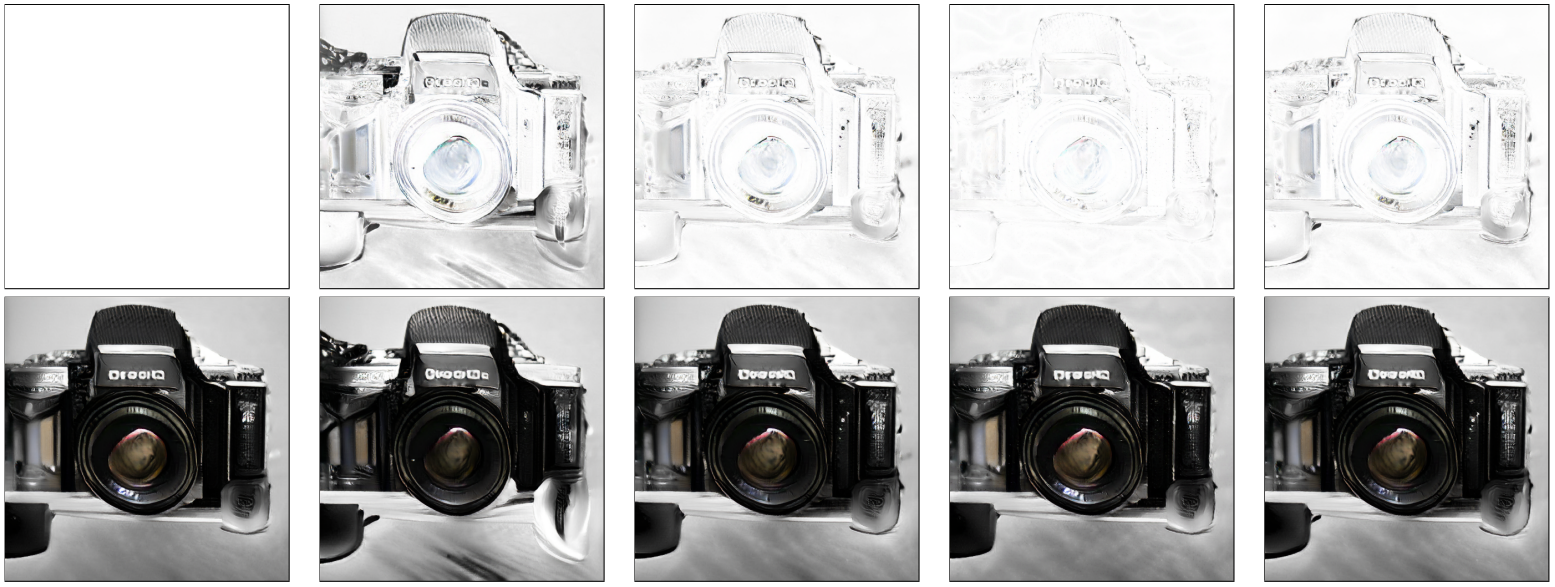
original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'reflex camera' → 'sunglass'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

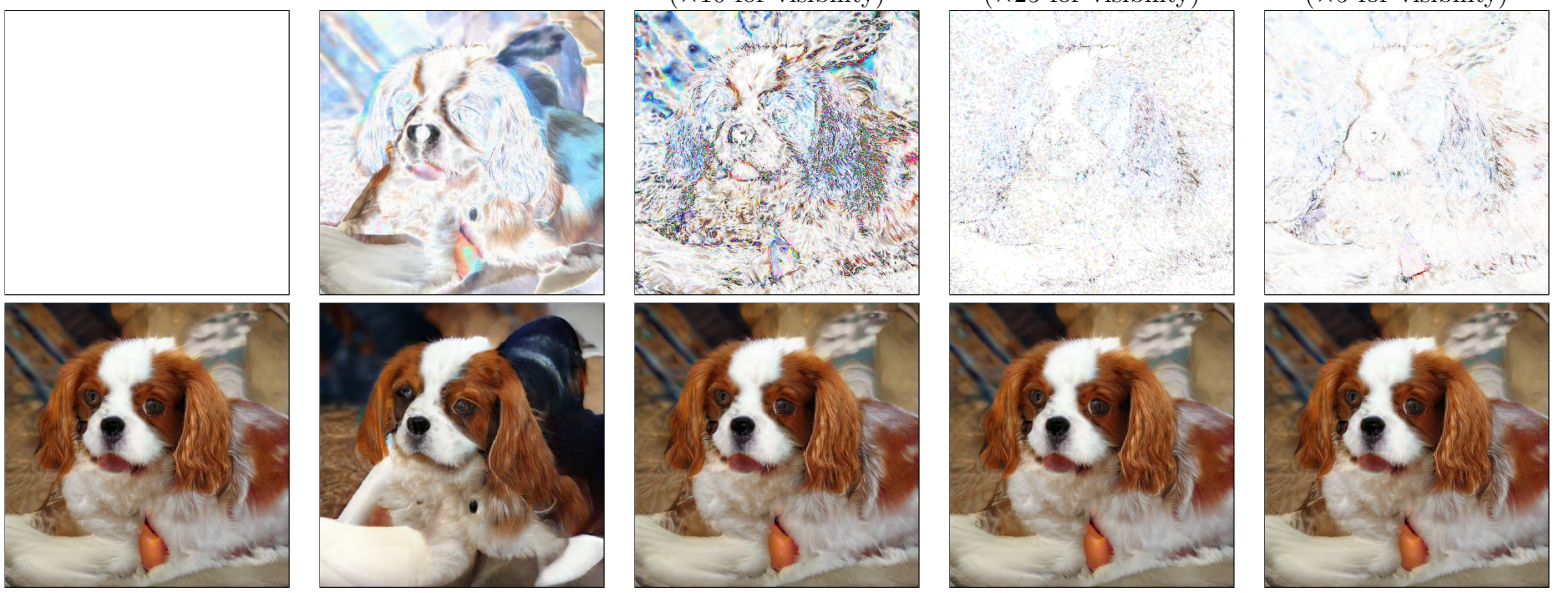
ResNet50 Robust ("Fast") 'reflex camera' → 'sunglass'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

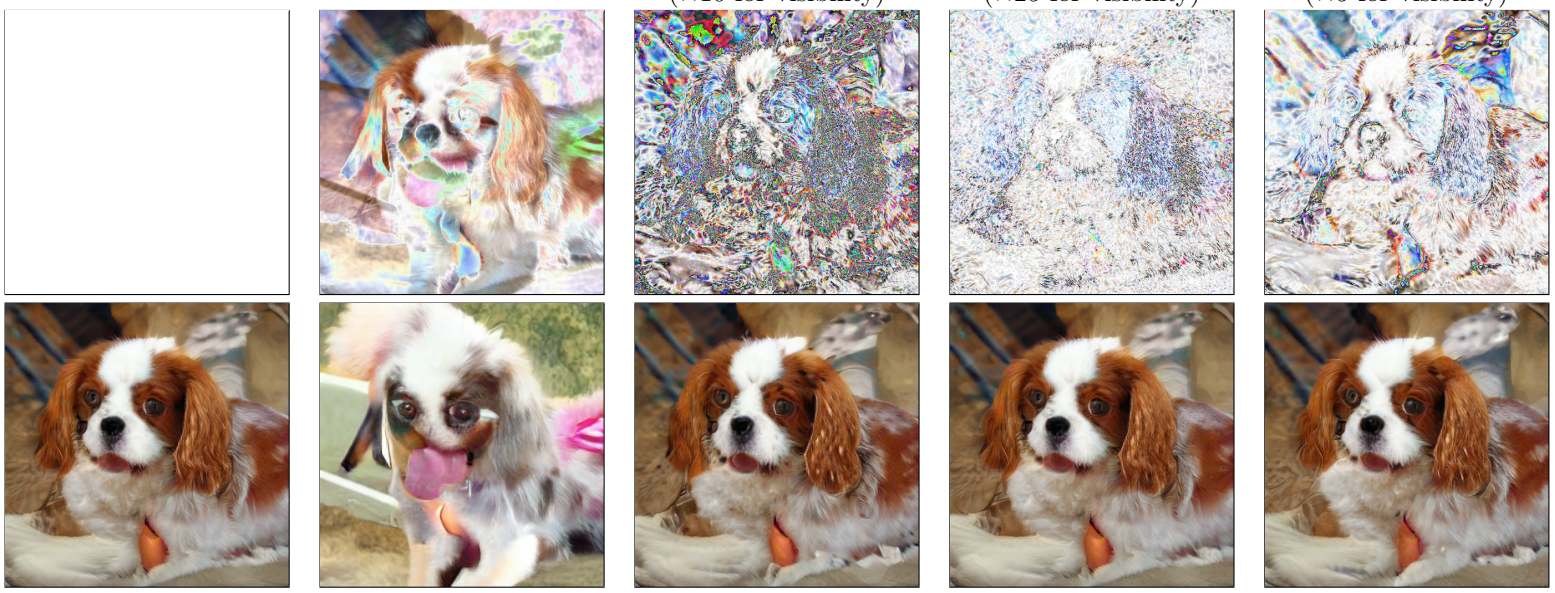
‘Blenheim spaniel’ → ‘black-and-tan coonhound’



original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

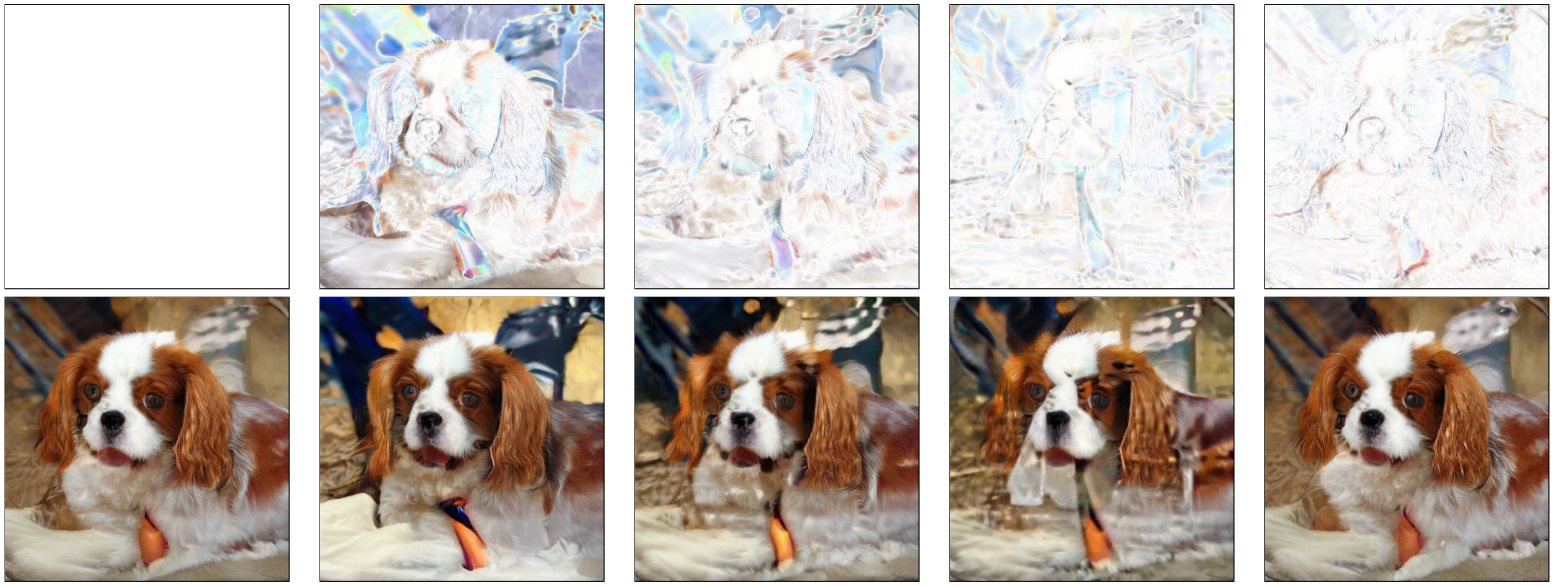
EfficientNet-B4NS

‘Blenheim spaniel’ → ‘black-and-tan coonhound’



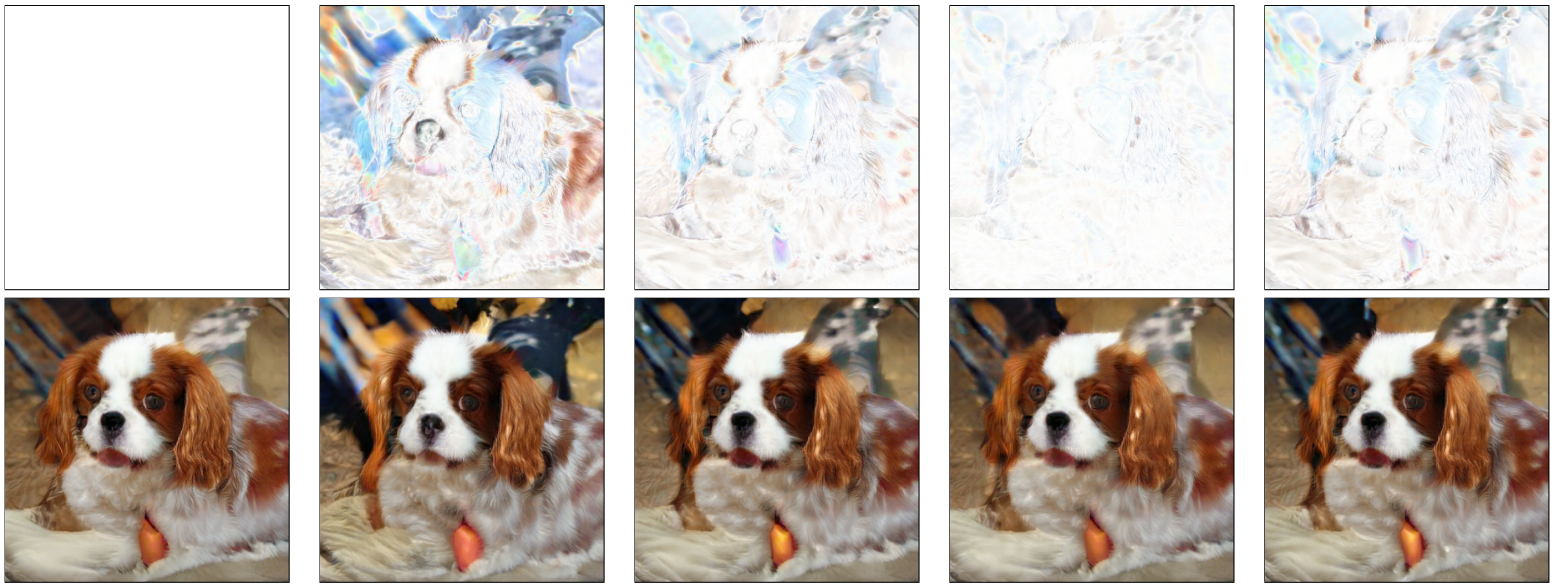
original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

ResNet50 Robust ("Engstrom") 'Blenheim spaniel' → 'black-and-tan coonhound'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

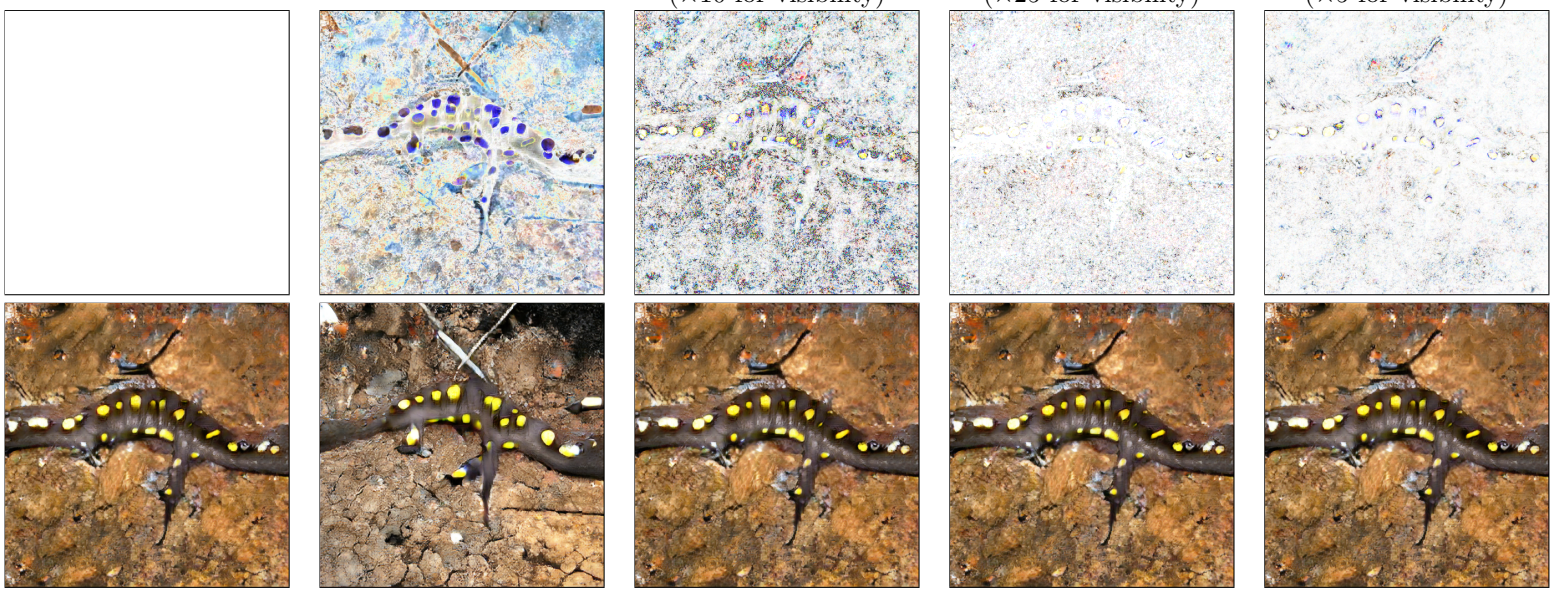
ResNet50 Robust ("Fast") 'Blenheim spaniel' → 'black-and-tan coonhound'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

‘spotted salamander’ → ‘digital clock’

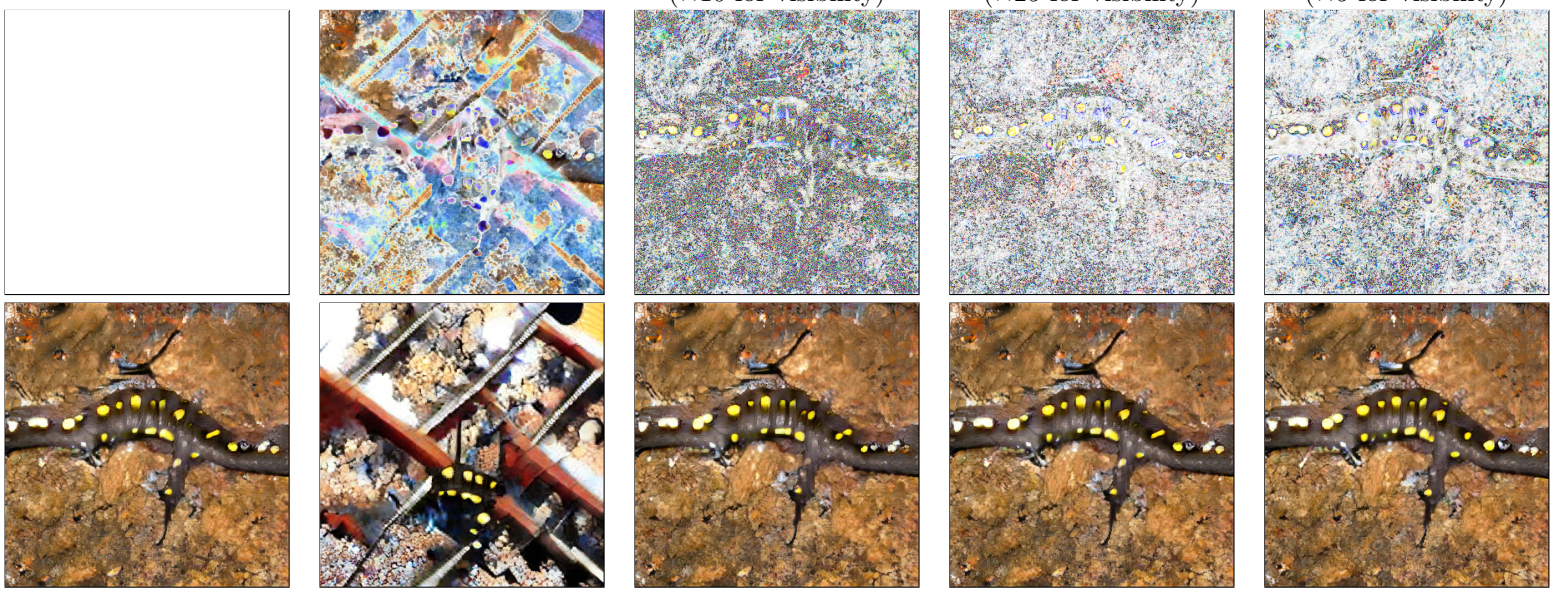


(×10 for visibility) (×25 for visibility) (×5 for visibility)

original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

EfficientNet-B4NS

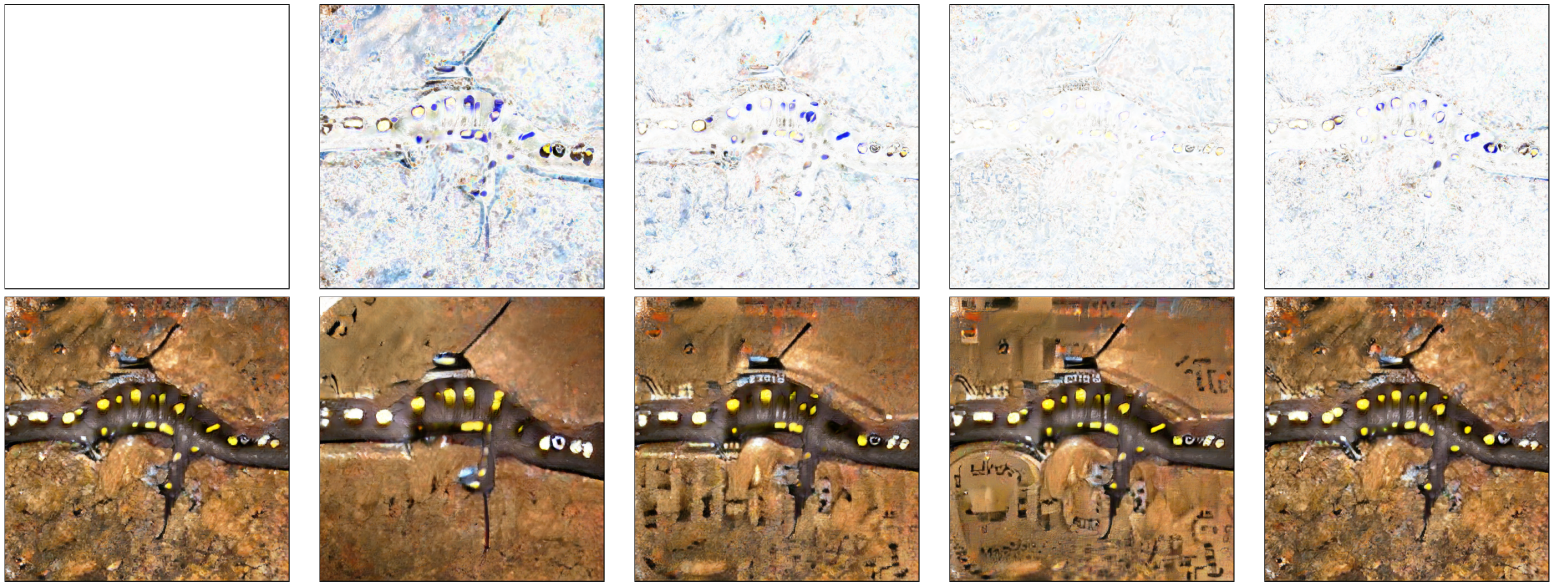
‘spotted salamander’ → ‘digital clock’



(×10 for visibility) (×25 for visibility) (×5 for visibility)

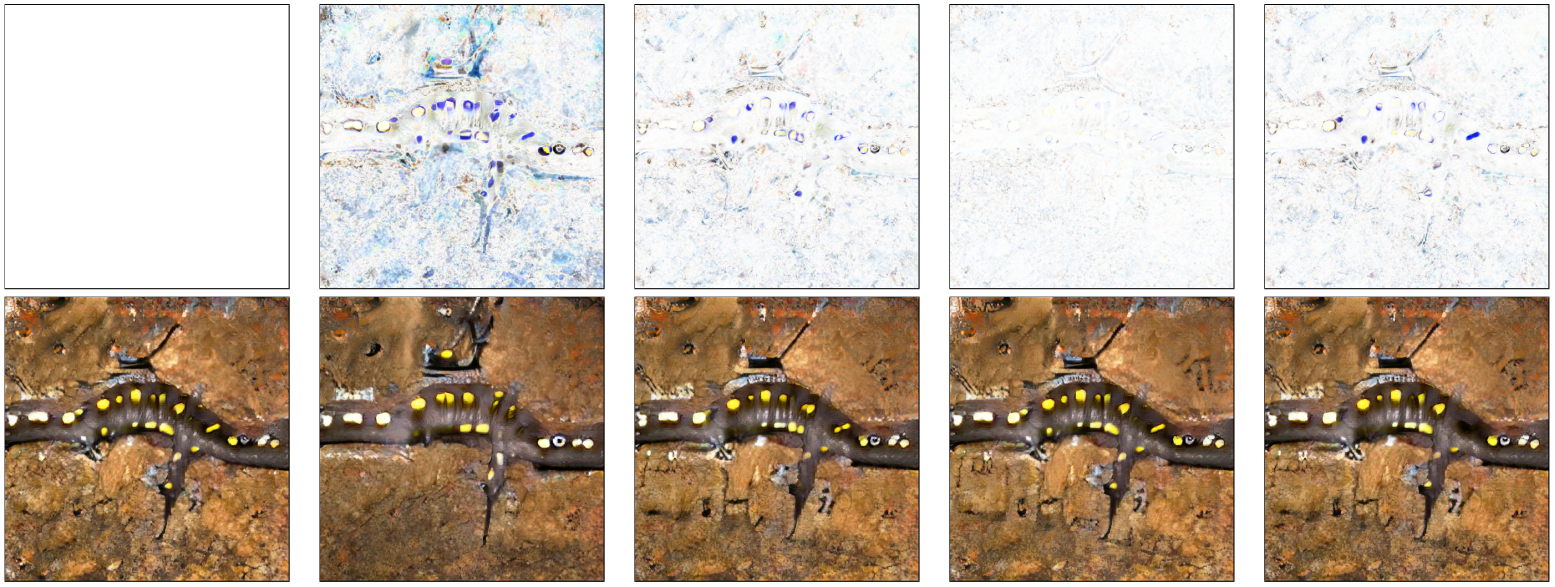
original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'spotted salamander' → 'digital clock'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Fast") 'spotted salamander' → 'digital clock'

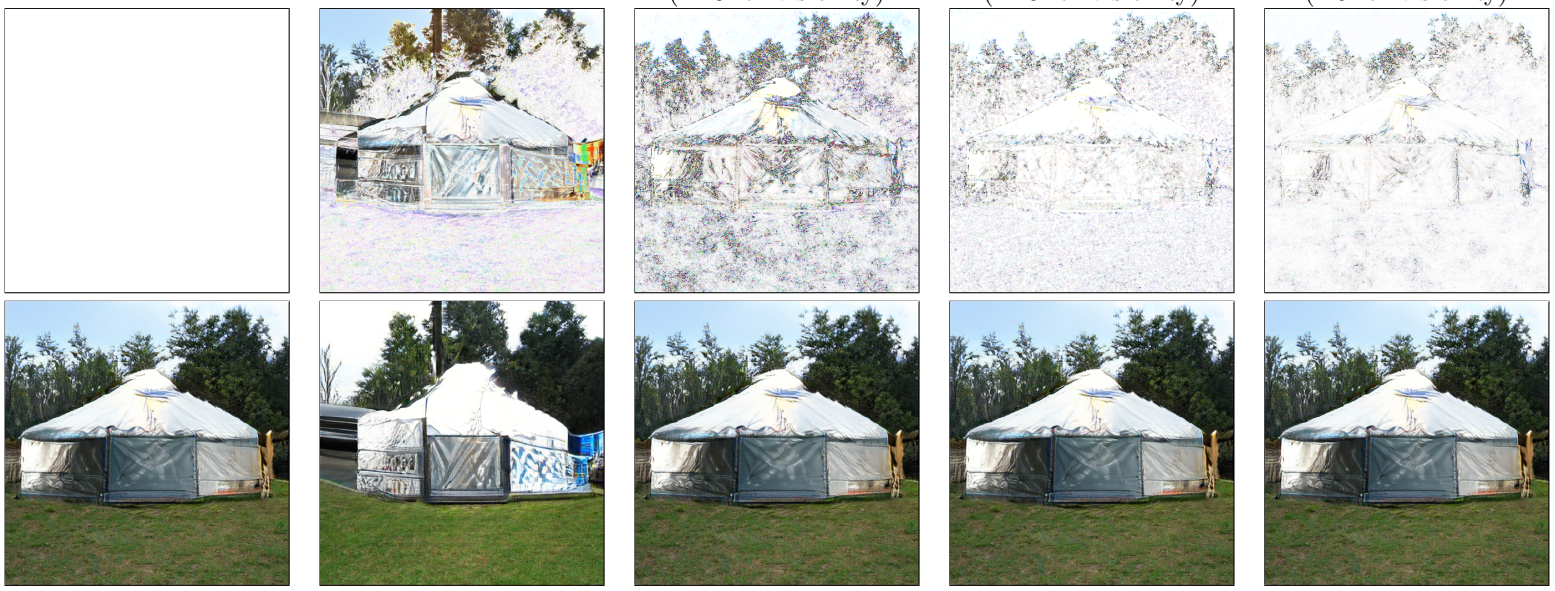


original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

‘yurt’ → ‘library’

(×10 for visibility) (×25 for visibility) (×5 for visibility)

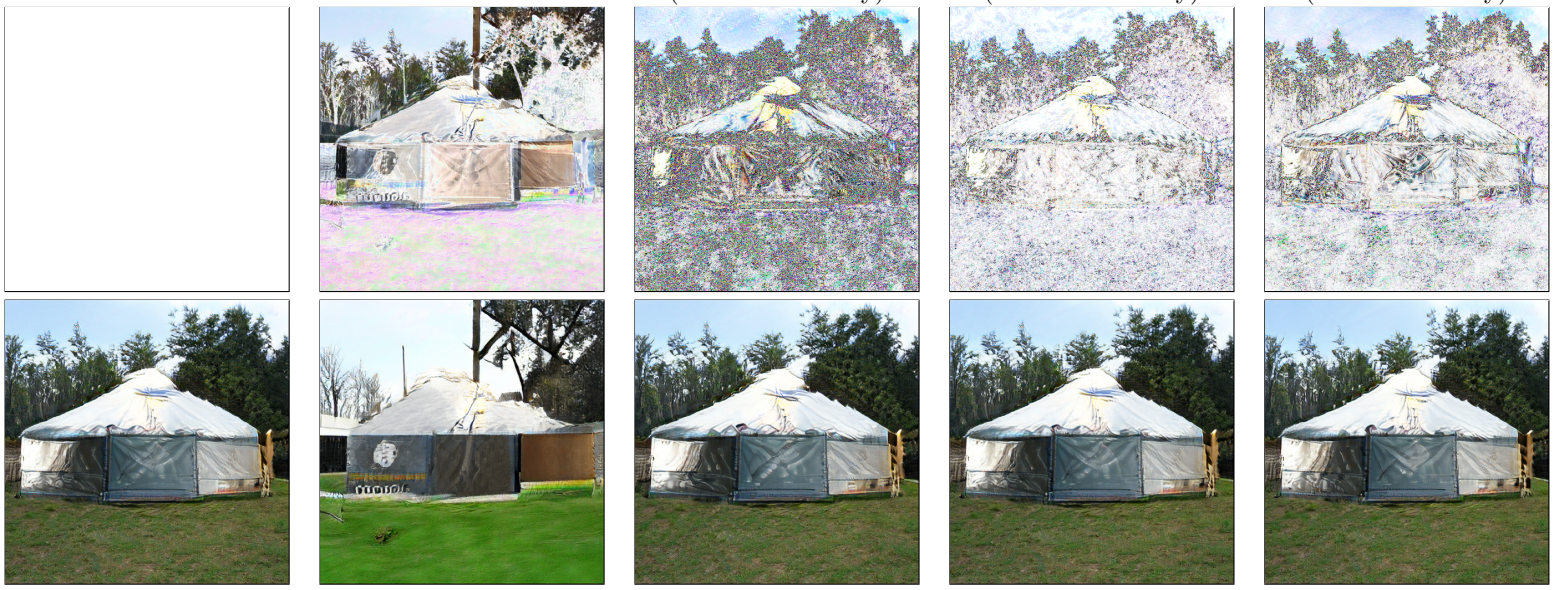


original first 6 layers middle 6 layers last 6 layers all 18 layers

EfficientNet-B4NS

‘yurt’ → ‘library’

(×10 for visibility) (×25 for visibility) (×5 for visibility)



original first 6 layers middle 6 layers last 6 layers all 18 layers

ResNet50 Robust ("Engstrom") 'yurt' → 'library'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

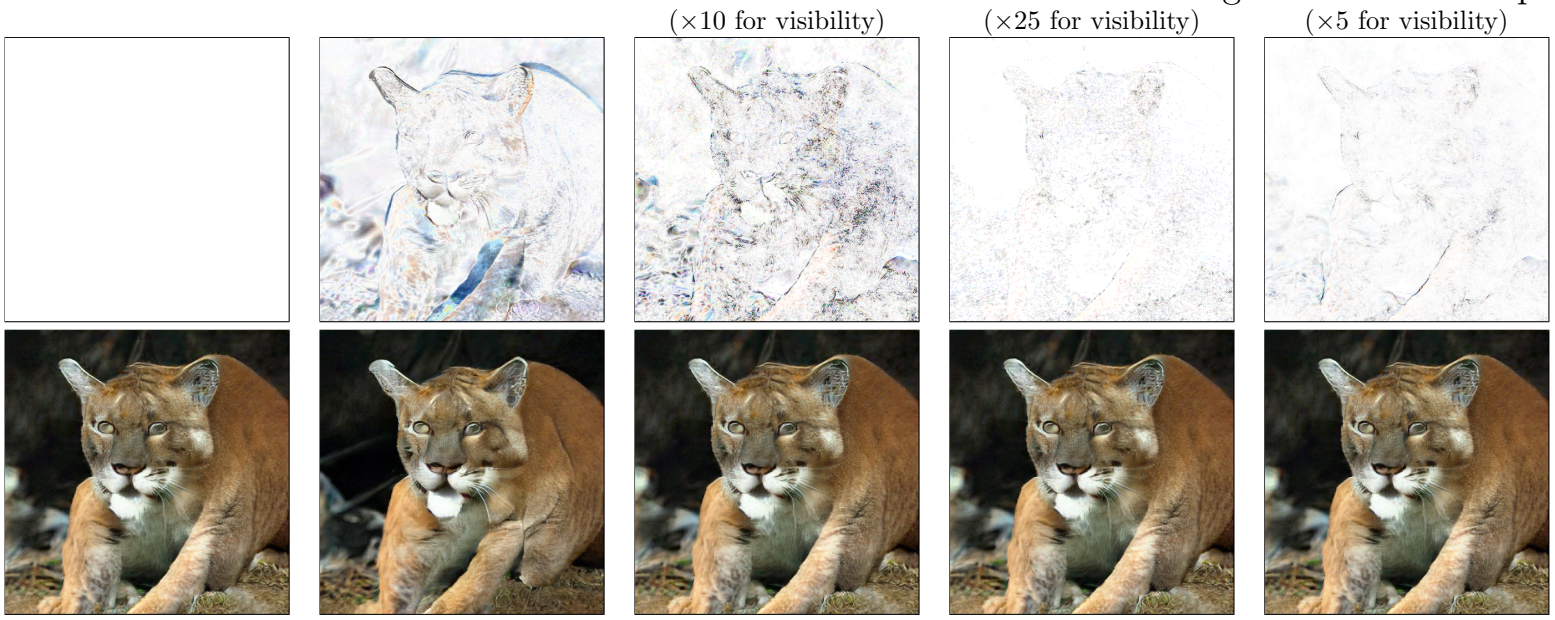
ResNet50 Robust ("Fast") 'yurt' → 'library'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

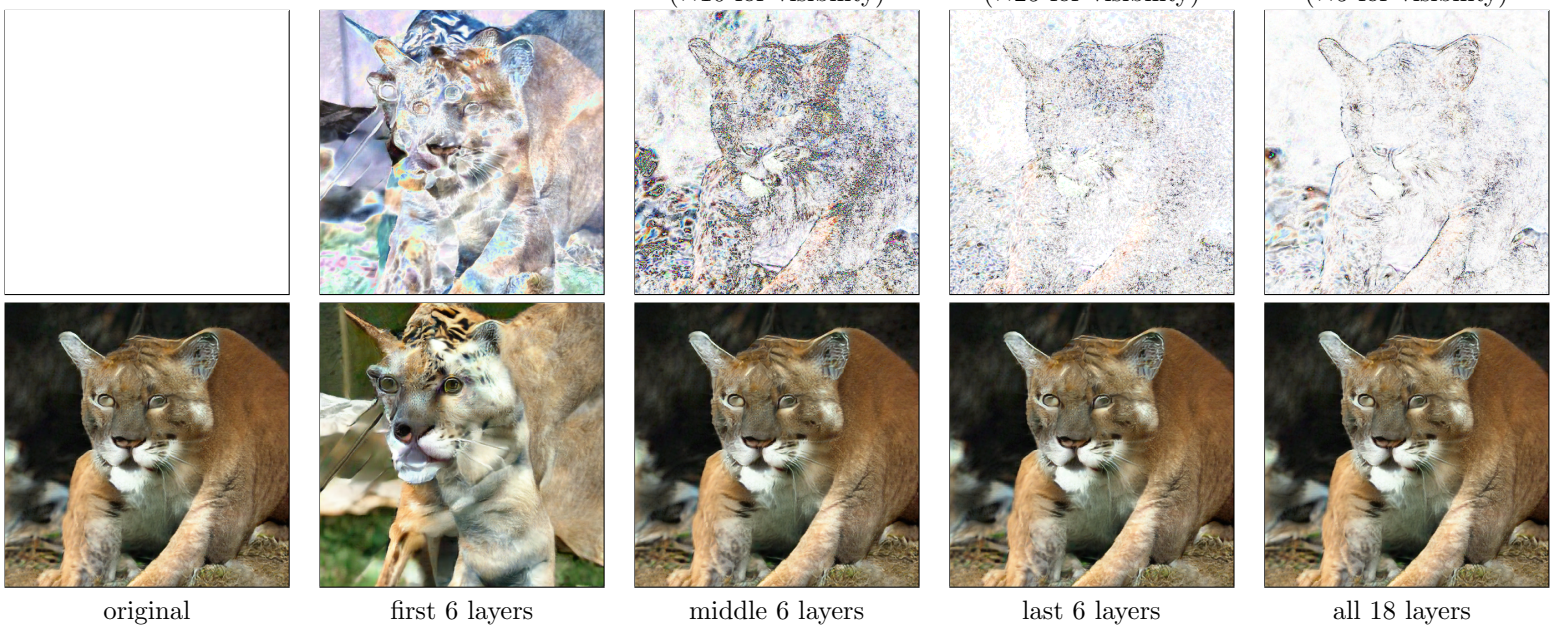
‘cougar’ → ‘shower cap’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

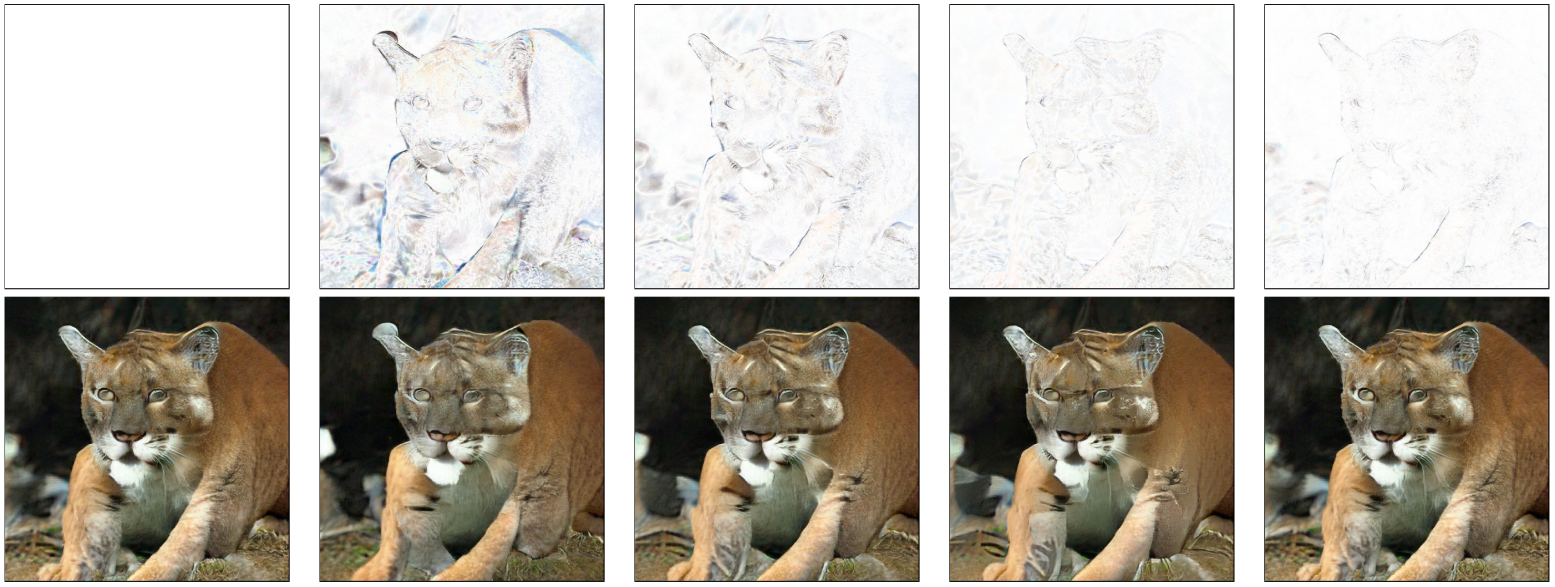
EfficientNet-B4NS

‘cougar’ → ‘shower cap’



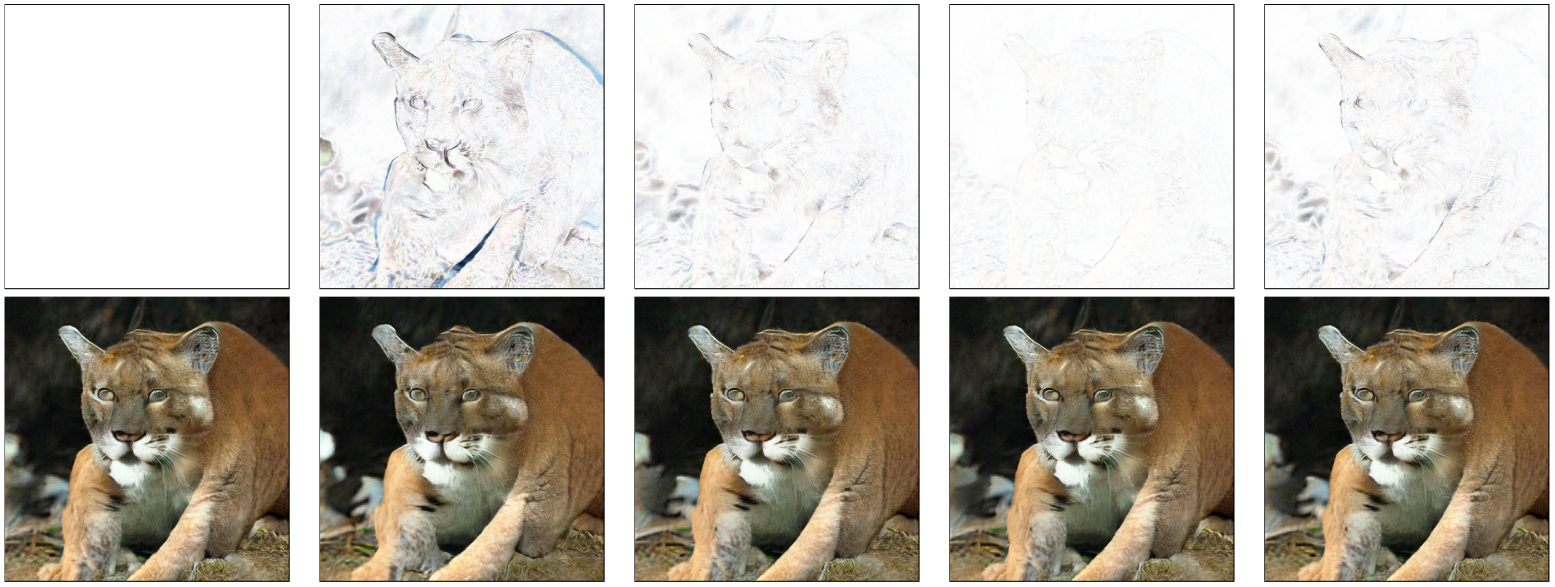
original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'cougar' → 'shower cap'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

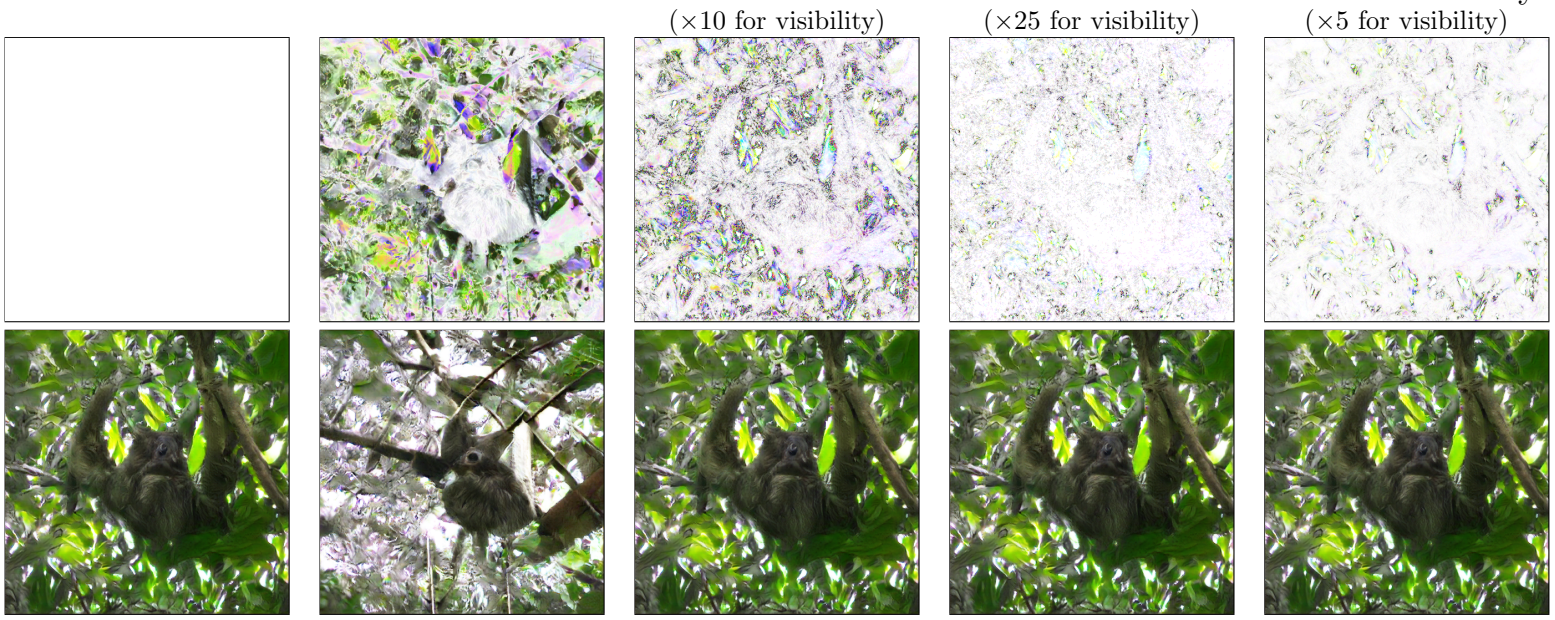
ResNet50 Robust ("Fast") 'cougar' → 'shower cap'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

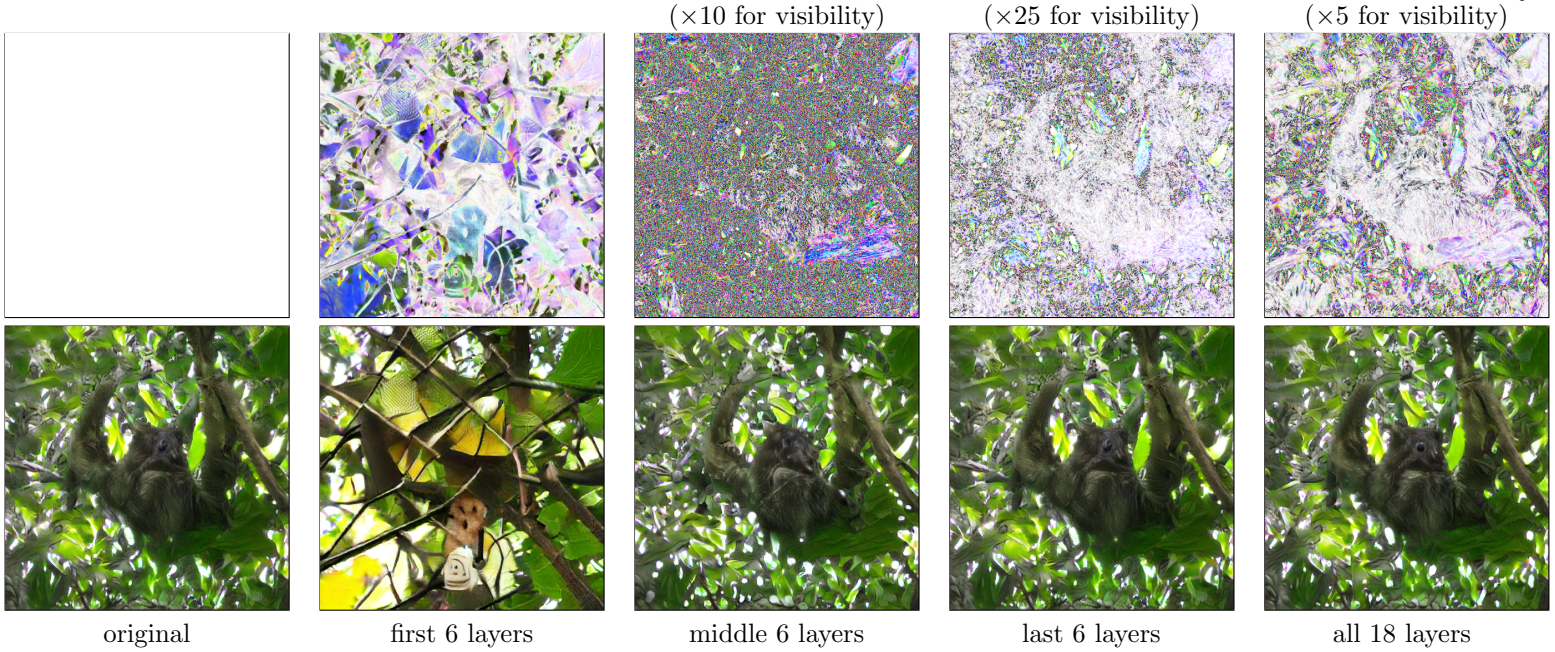
‘three-toed sloth’ → ‘bakery’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

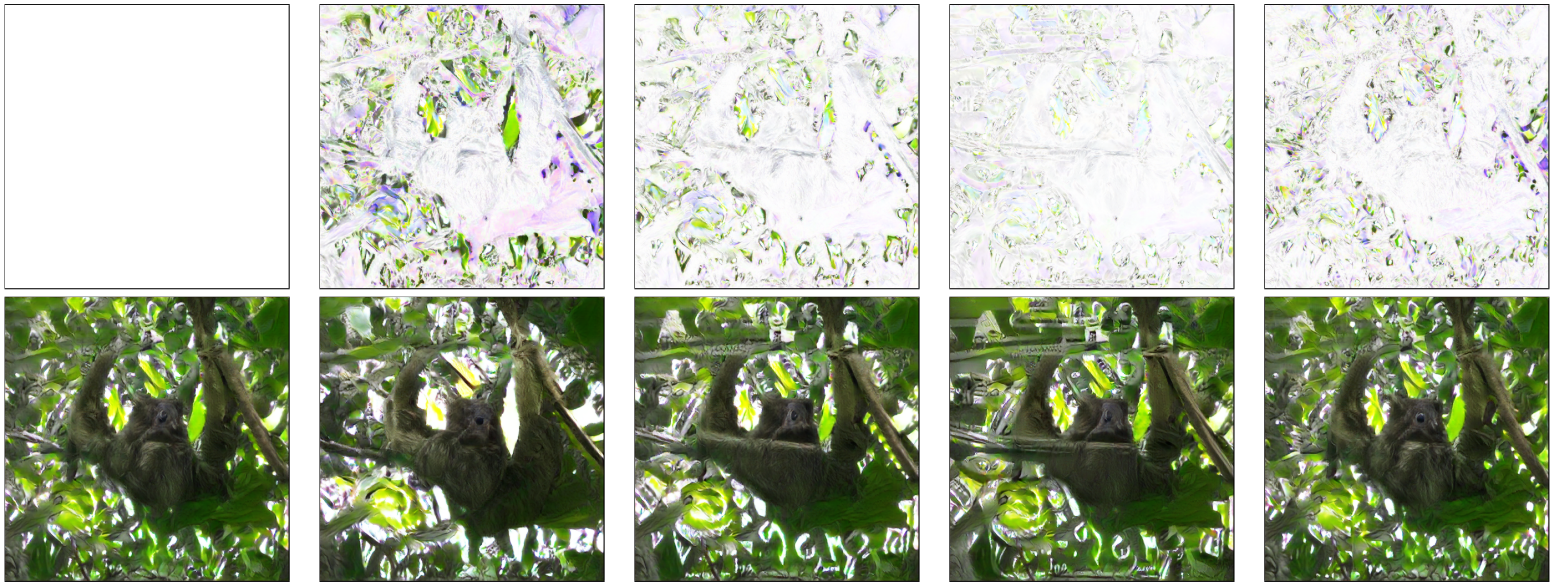
EfficientNet-B4NS

‘three-toed sloth’ → ‘bakery’



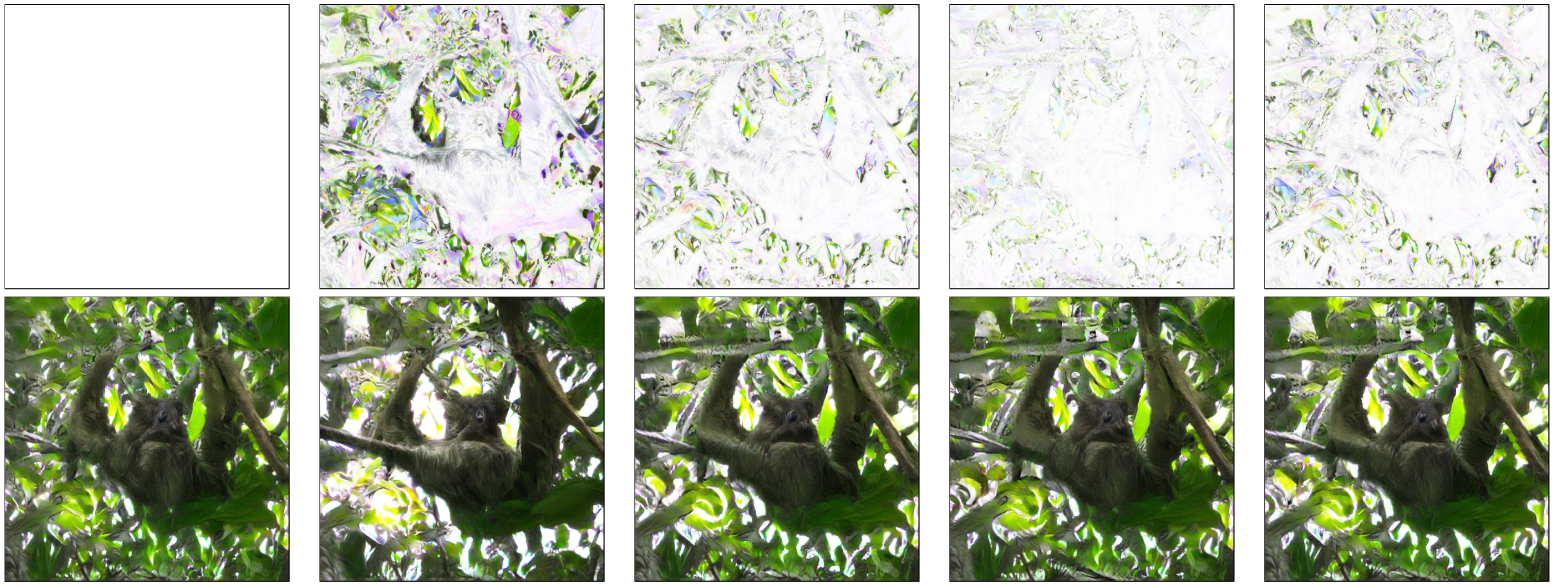
original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'three-toed sloth' → 'bakery'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

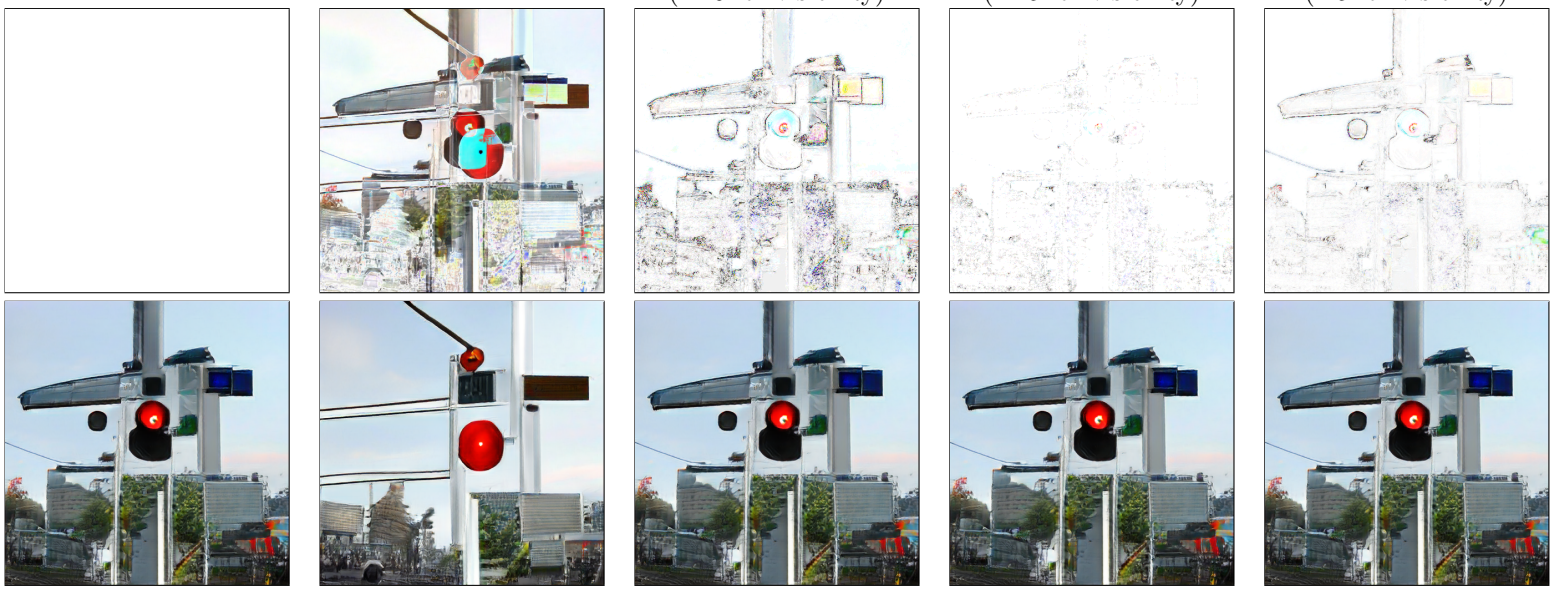
ResNet50 Robust ("Fast") 'three-toed sloth' → 'bakery'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

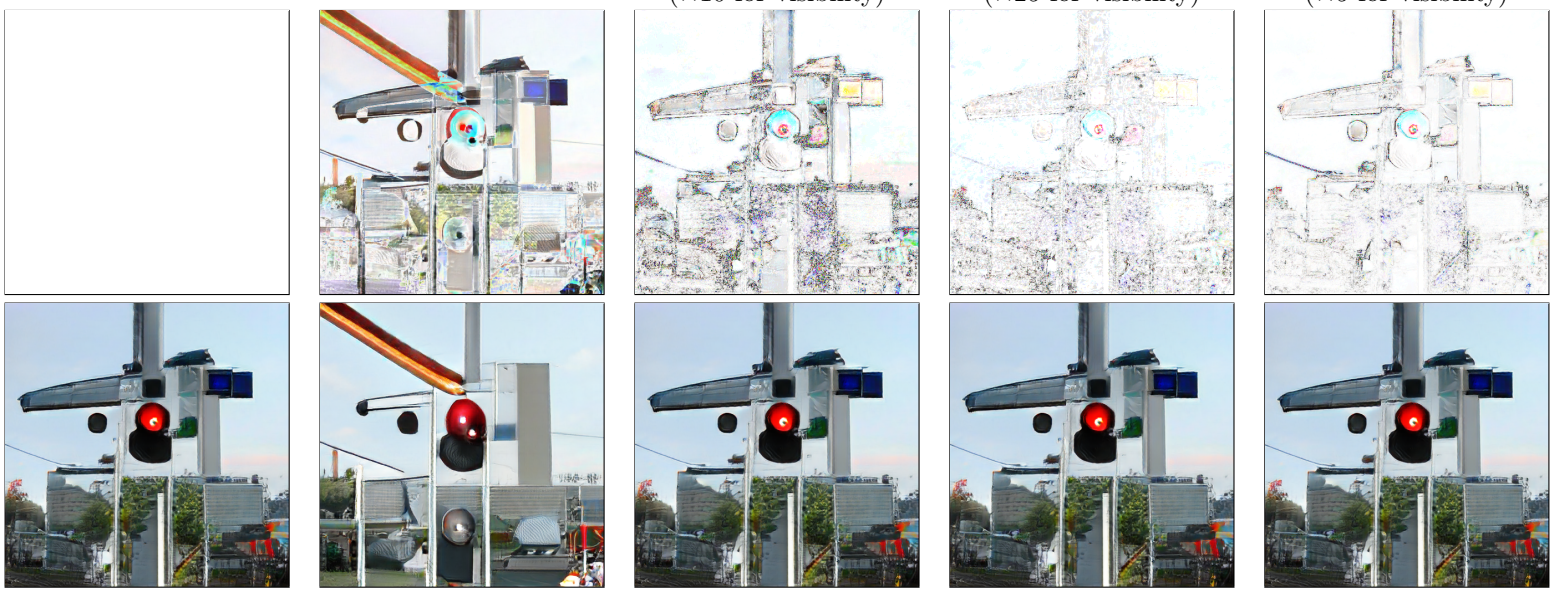
ResNet50

‘traffic light’ → ‘horizontal bar’

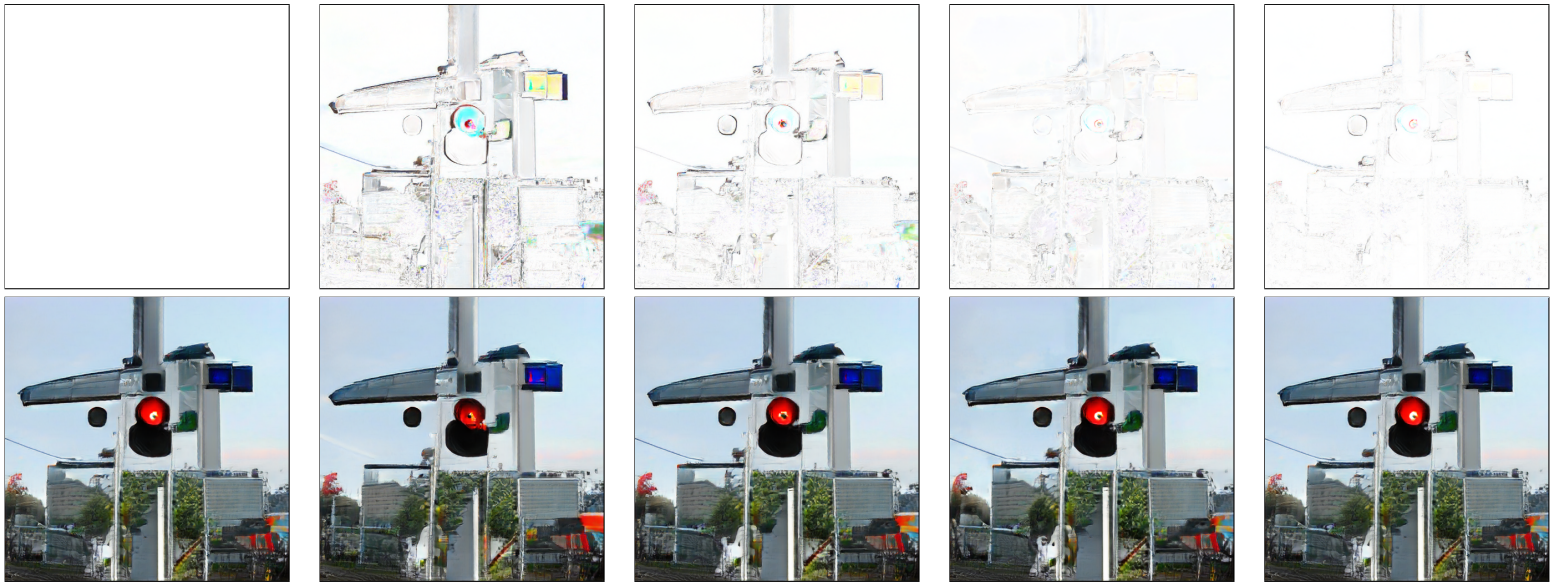


EfficientNet-B4NS

‘traffic light’ → ‘horizontal bar’

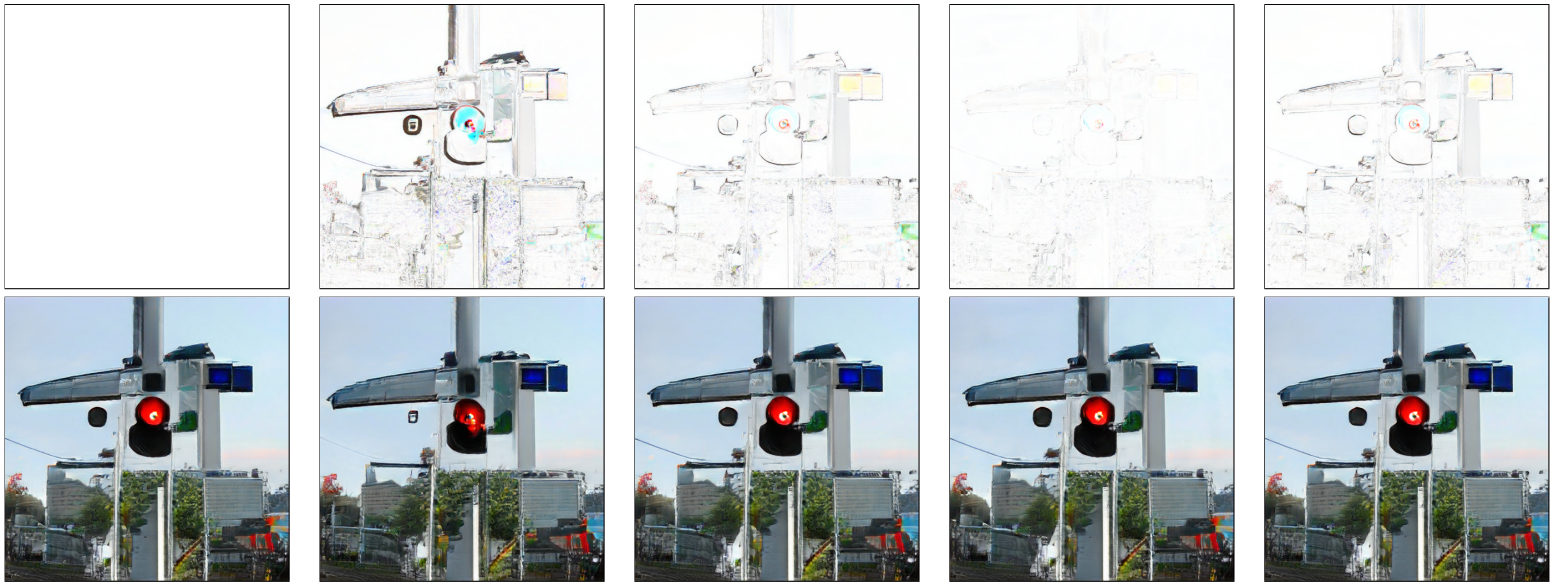


ResNet50 Robust ("Engstrom") 'traffic light' → 'horizontal bar'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

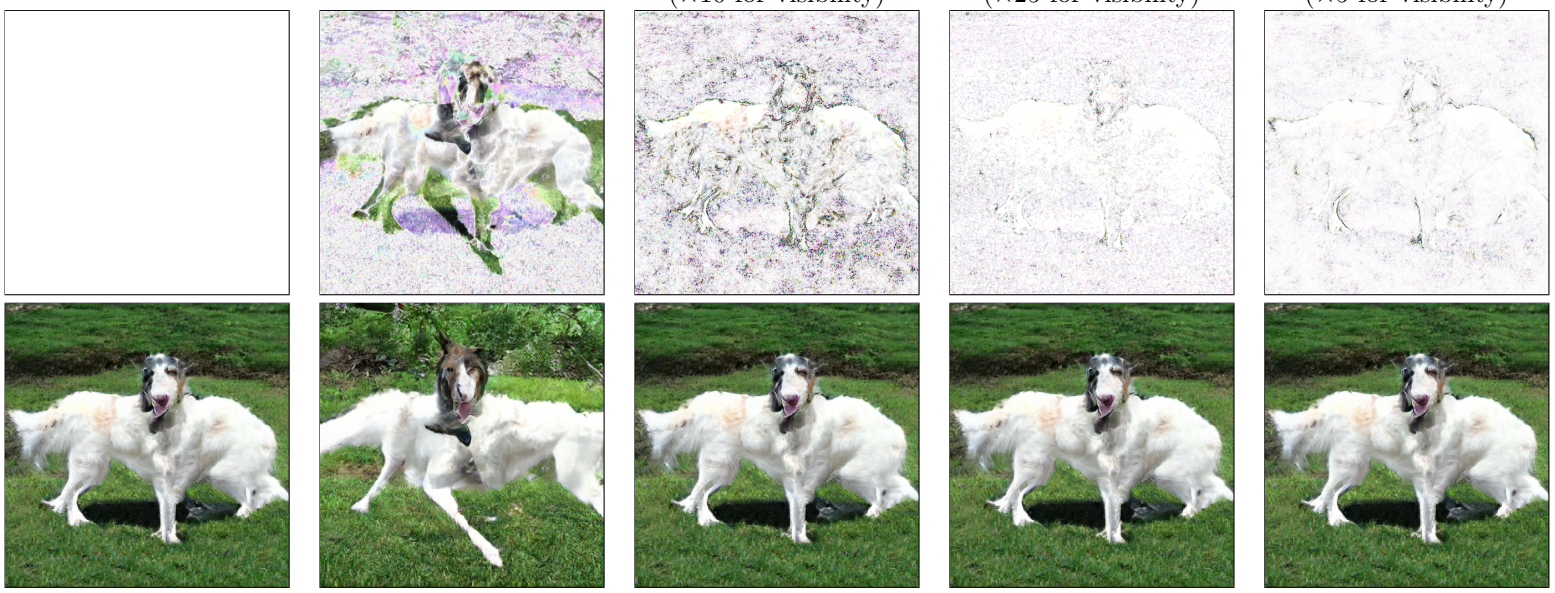
ResNet50 Robust ("Fast") 'traffic light' → 'horizontal bar'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

‘borzoi’ → ‘thatch’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

EfficientNet-B4NS

‘borzoi’ → ‘thatch’



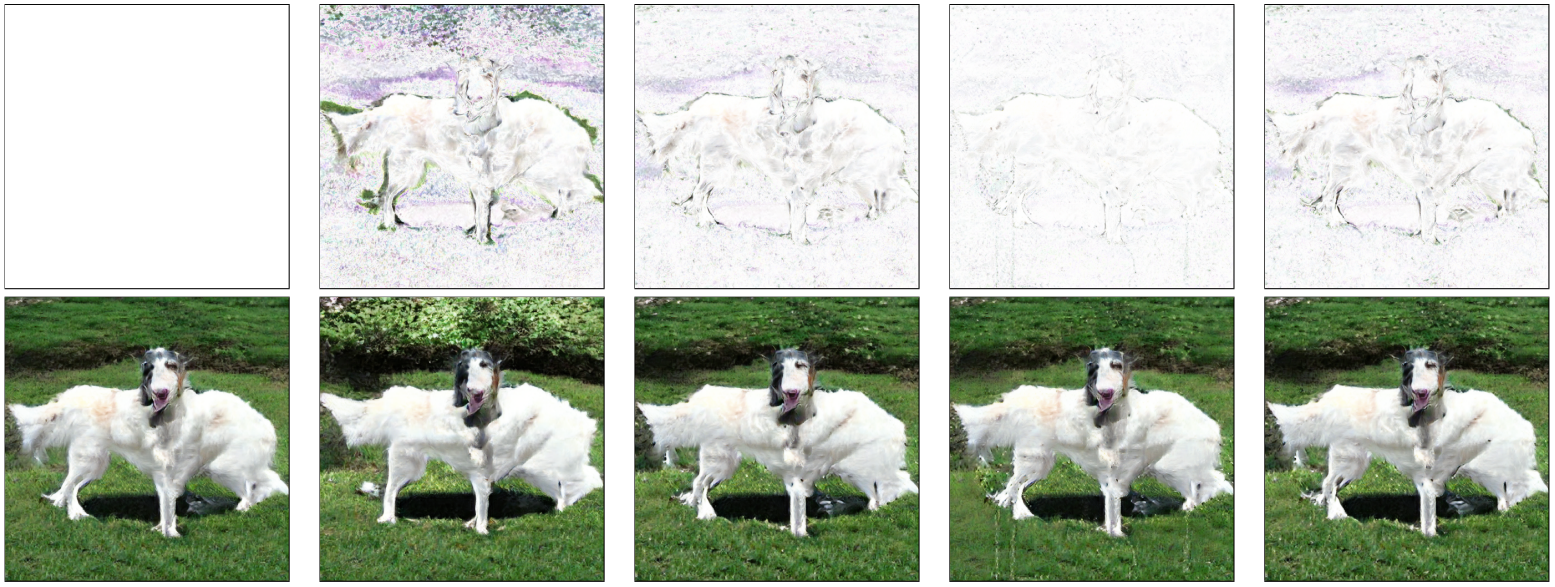
original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'borzoi' → 'thatch'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

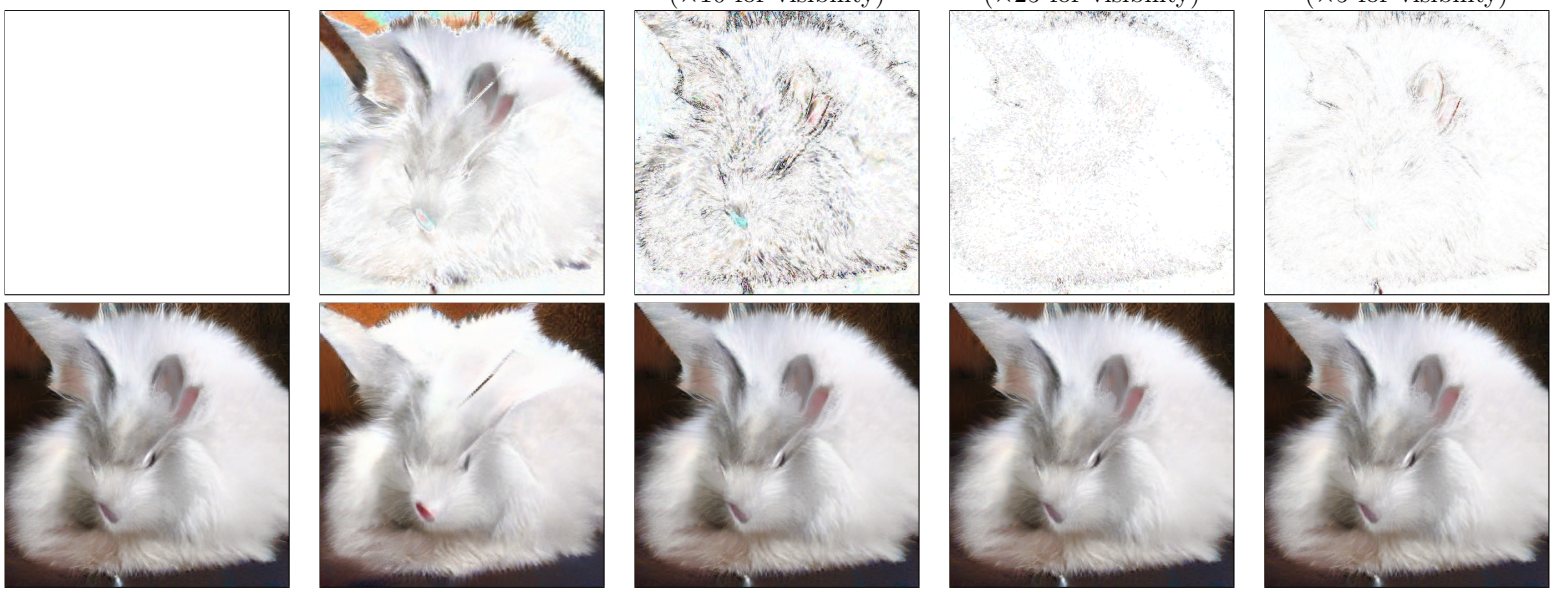
ResNet50 Robust ("Fast") 'borzoi' → 'thatch'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

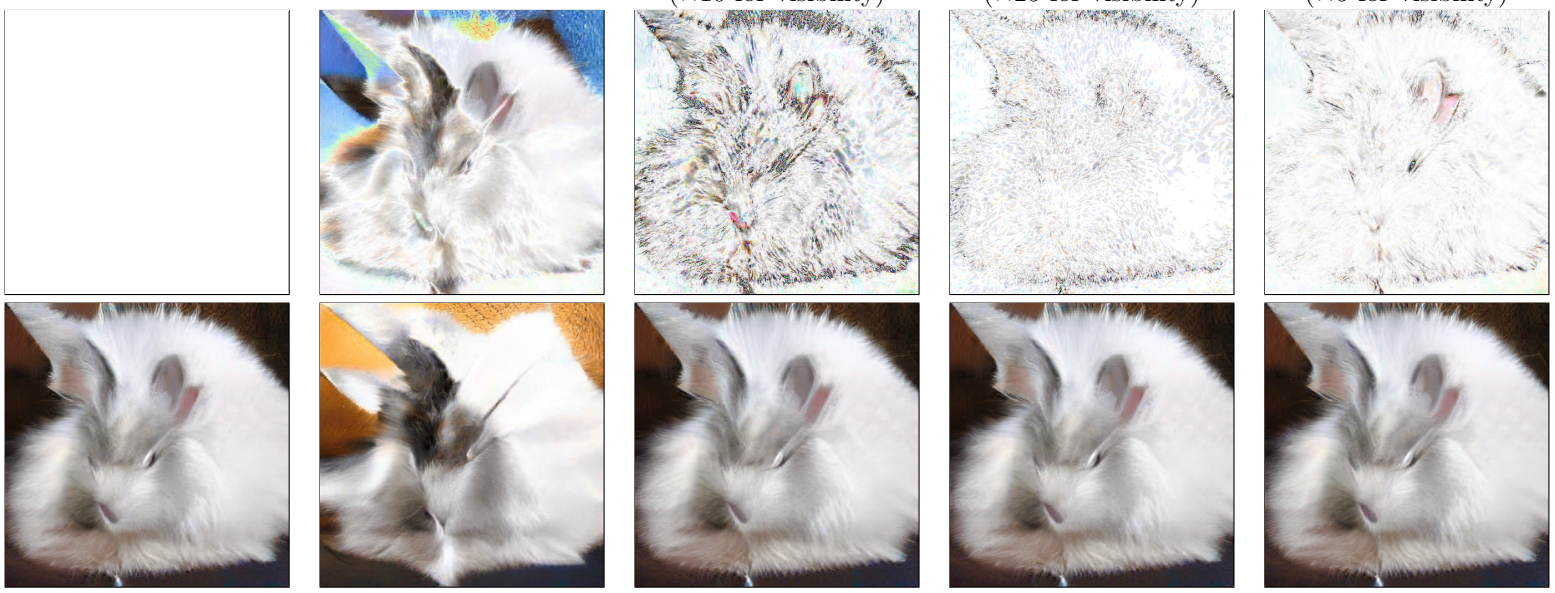
ResNet50

‘Angora’ → ‘envelope’

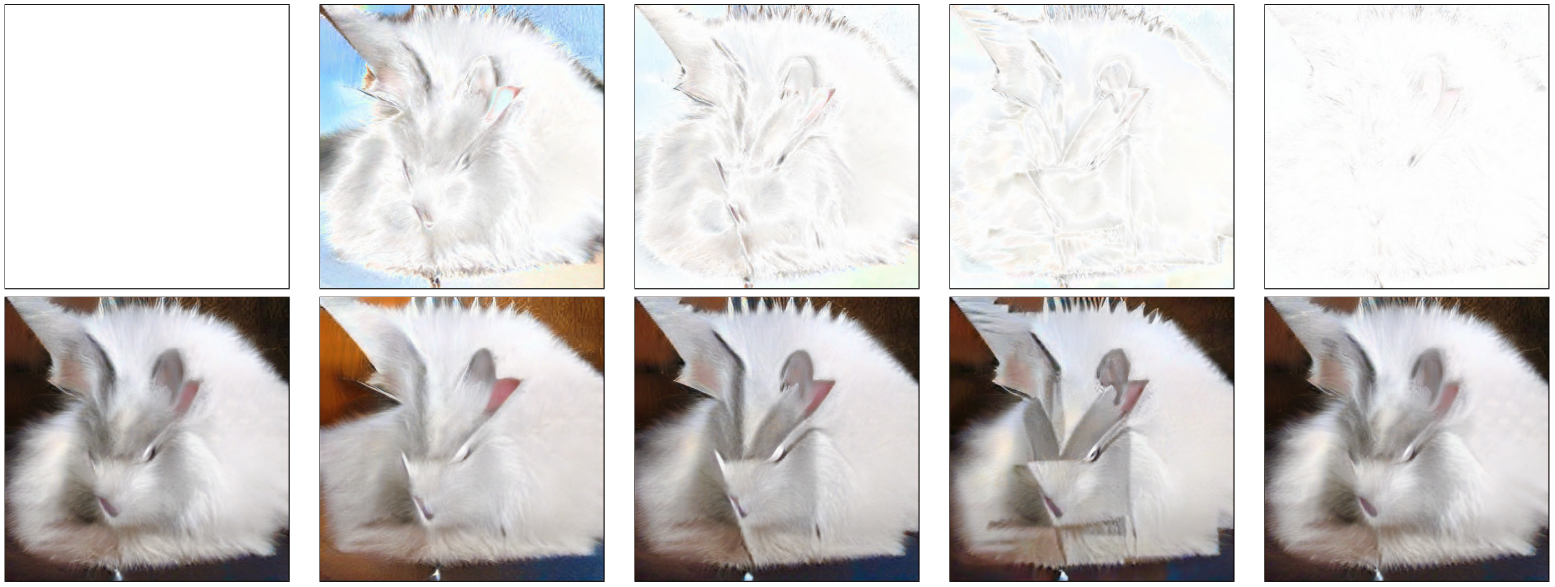


EfficientNet-B4NS

‘Angora’ → ‘envelope’

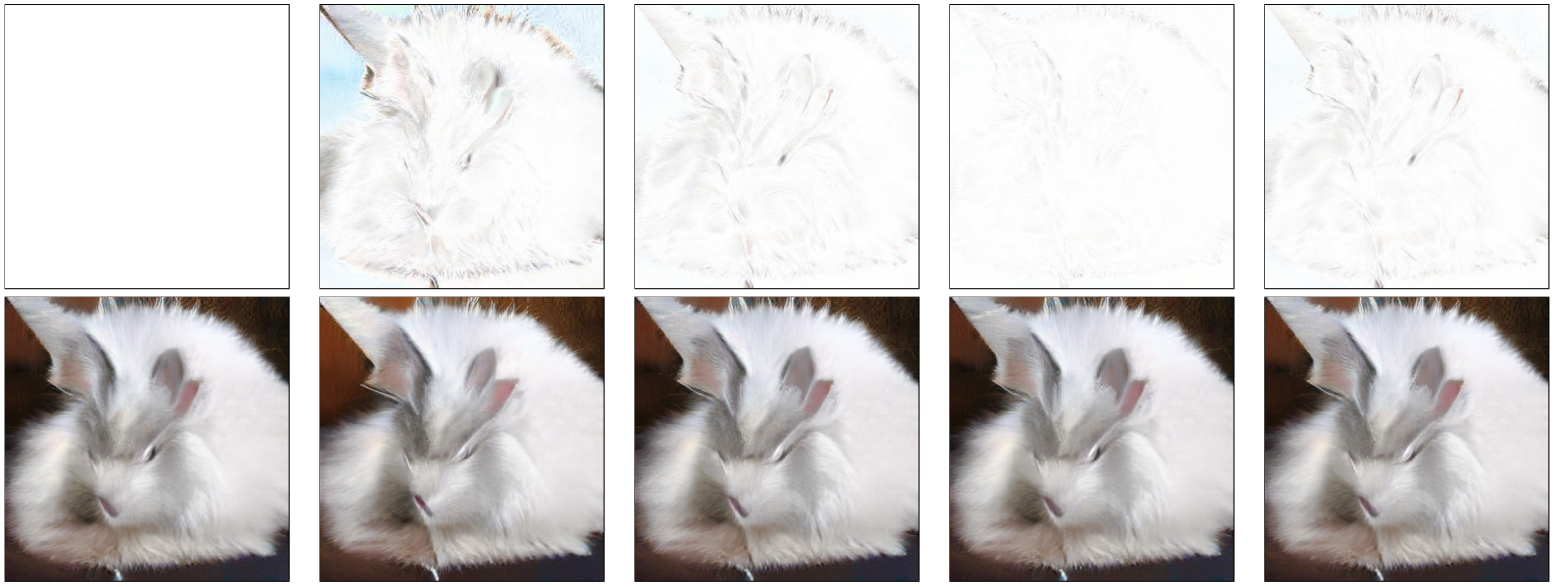


ResNet50 Robust ("Engstrom") 'Angora' → 'envelope'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Fast") 'Angora' → 'envelope'



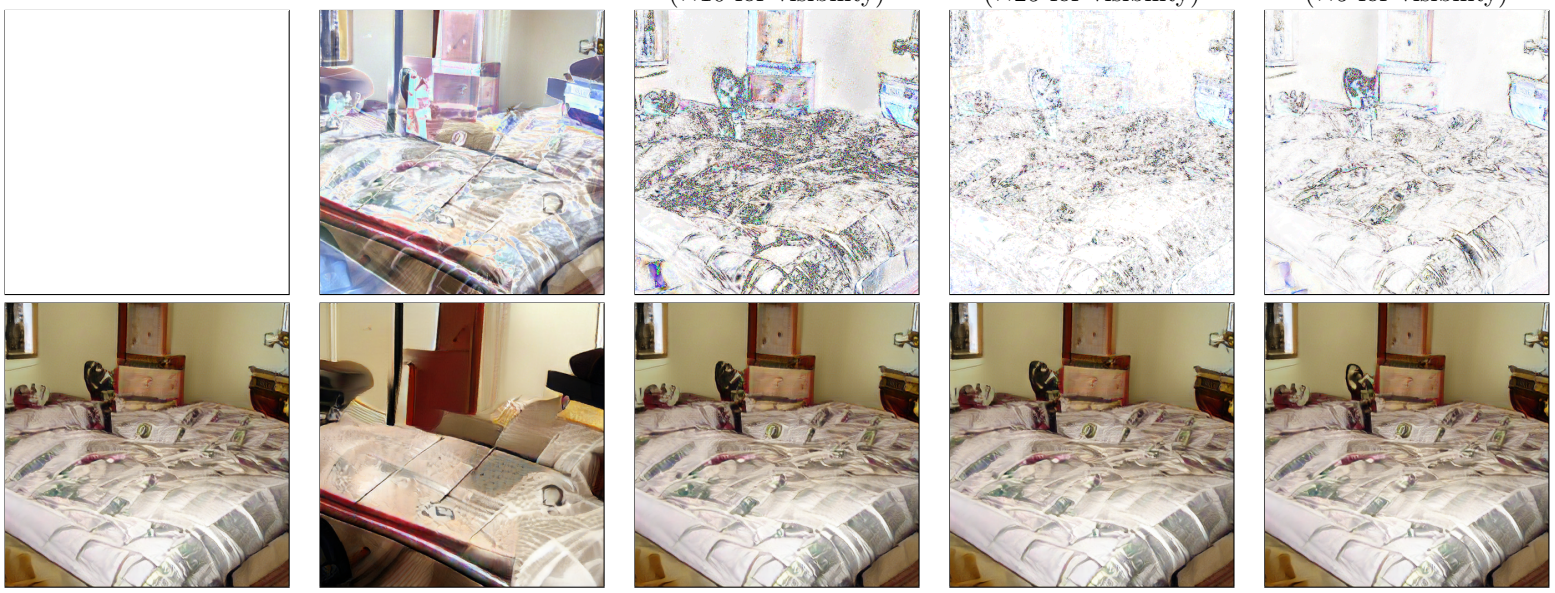
original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 ‘quilt’ → ‘disk brake’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

EfficientNet-B4NS ‘quilt’ → ‘disk brake’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'quilt' → 'disk brake'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

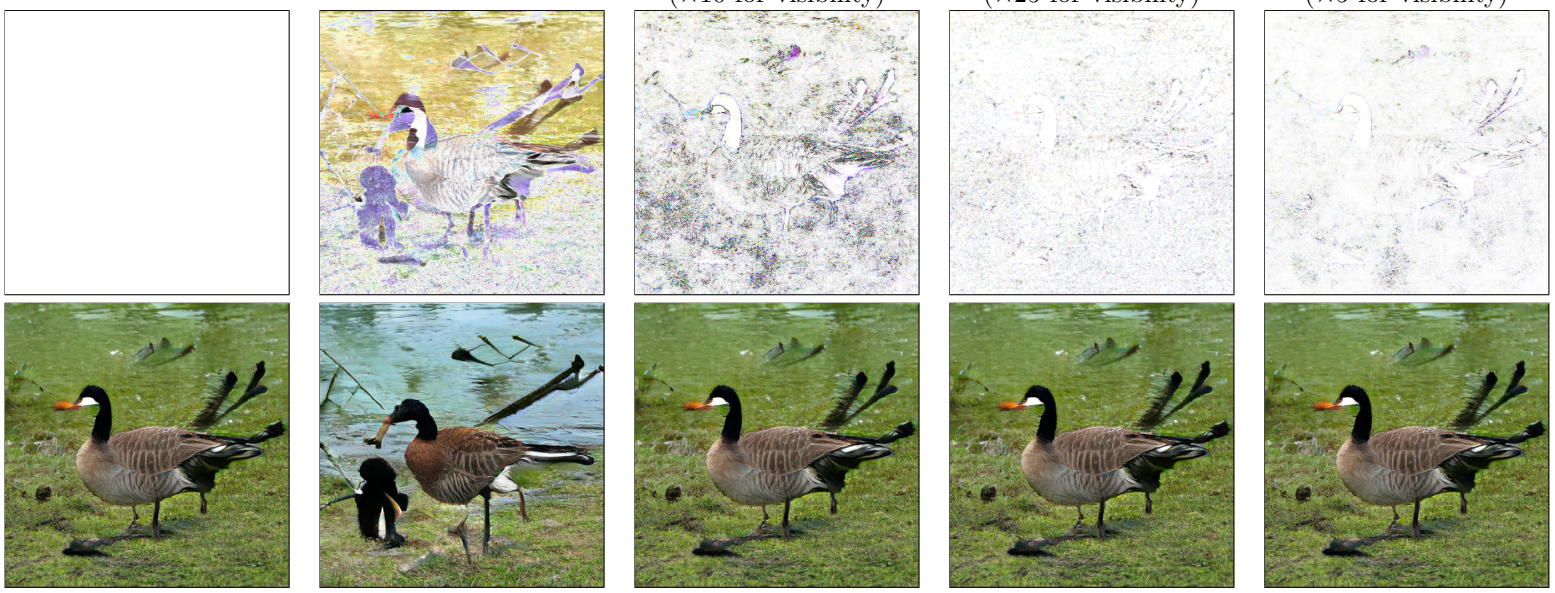
ResNet50 Robust ("Fast") 'quilt' → 'disk brake'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

‘goose’ → ‘crutch’

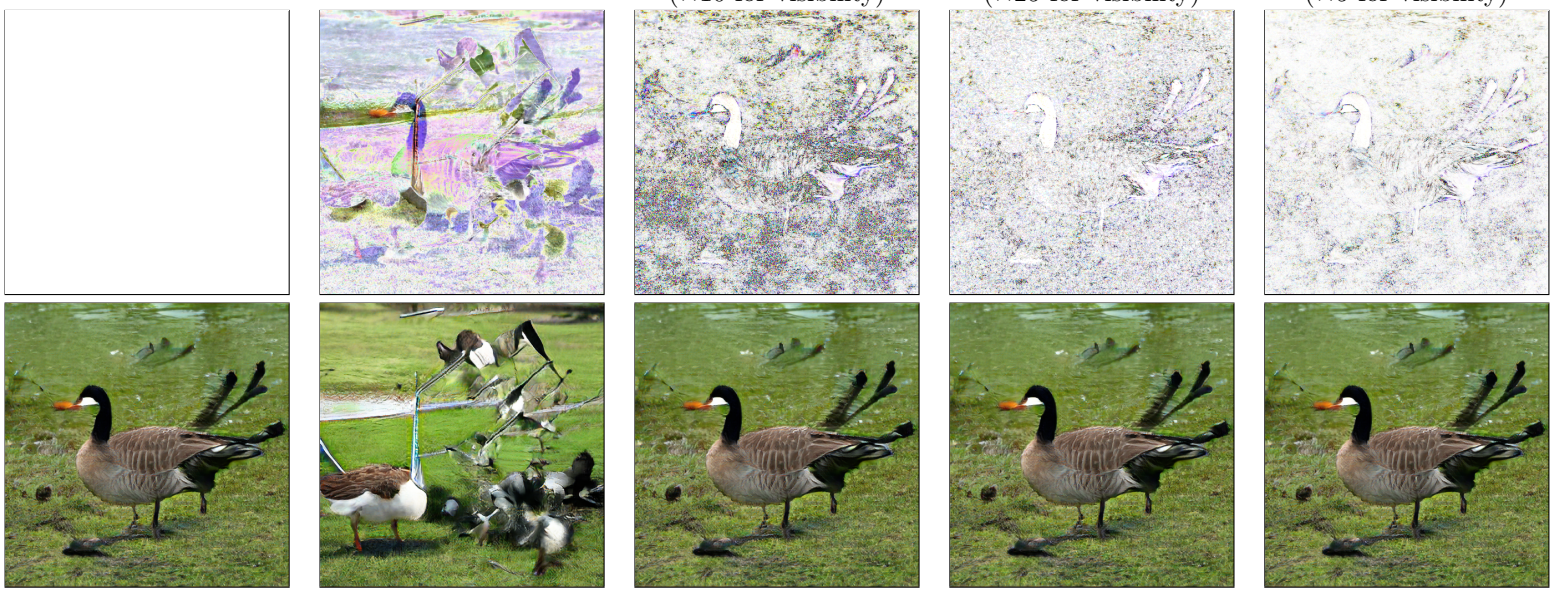


(×10 for visibility) (×25 for visibility) (×5 for visibility)

original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

EfficientNet-B4NS

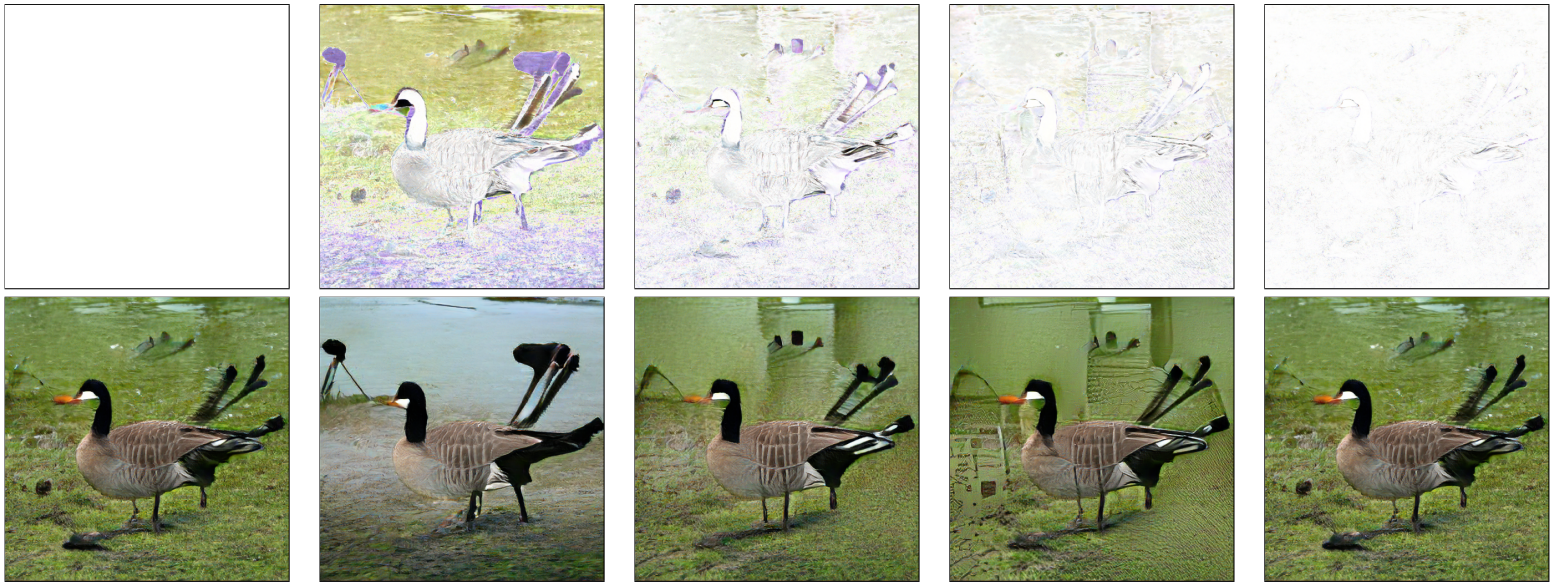
‘goose’ → ‘crutch’



(×10 for visibility) (×25 for visibility) (×5 for visibility)

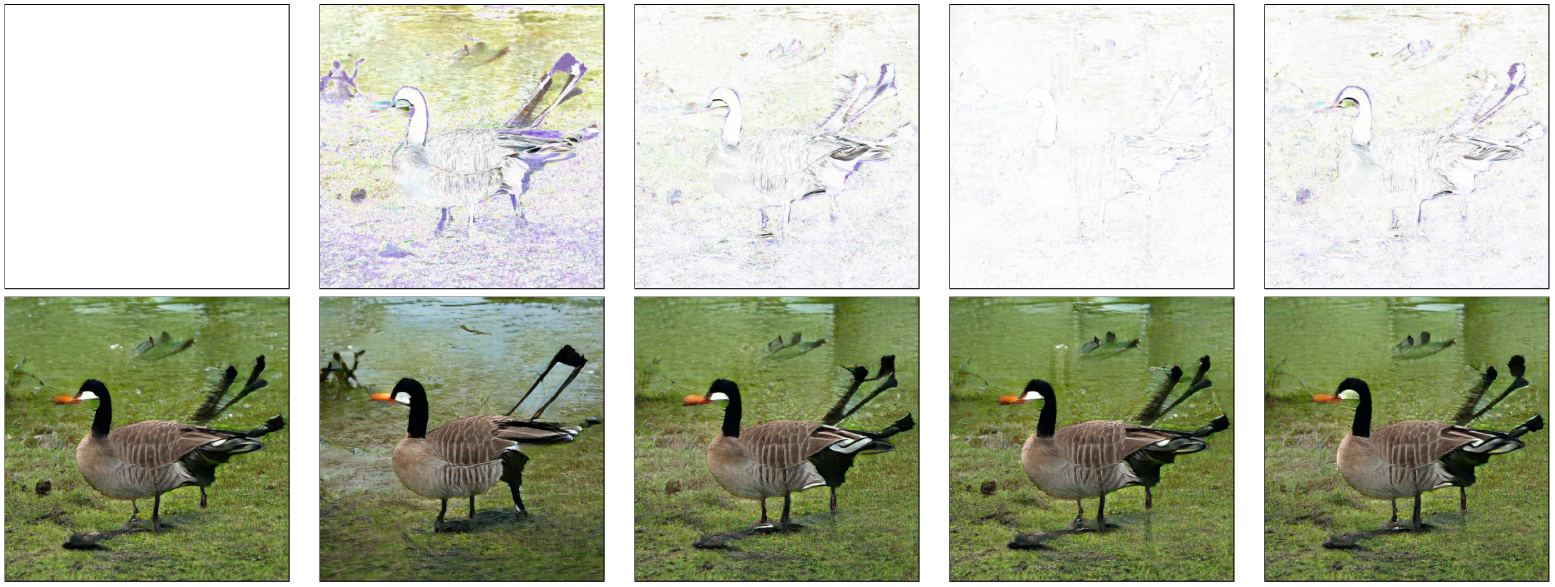
original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'goose' → 'crutch'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Fast") 'goose' → 'crutch'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

‘dogsled’ → ‘syringe’



original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

EfficientNet-B4NS

‘dogsled’ → ‘syringe’



original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

ResNet50 Robust ("Engstrom") 'dogsled' → 'syringe'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

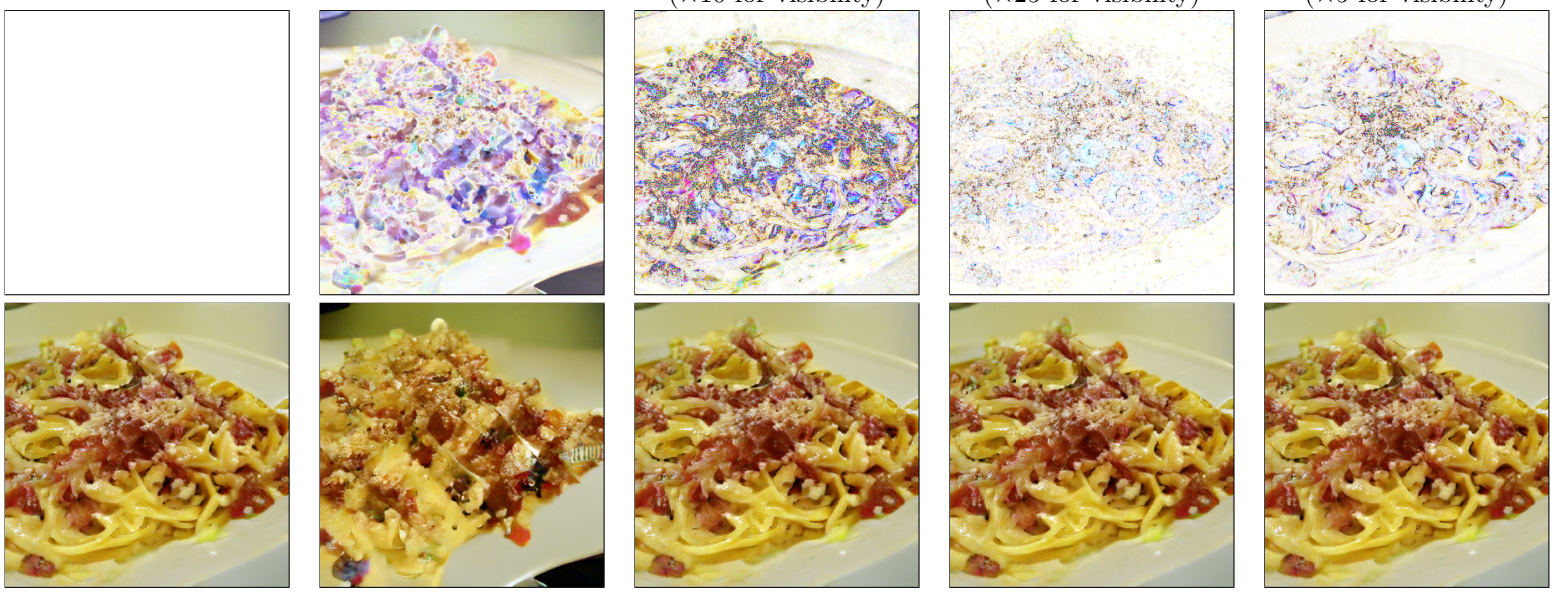
ResNet50 Robust ("Fast") 'dogsled' → 'syringe'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

‘carbonara’ → ‘electric guitar’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

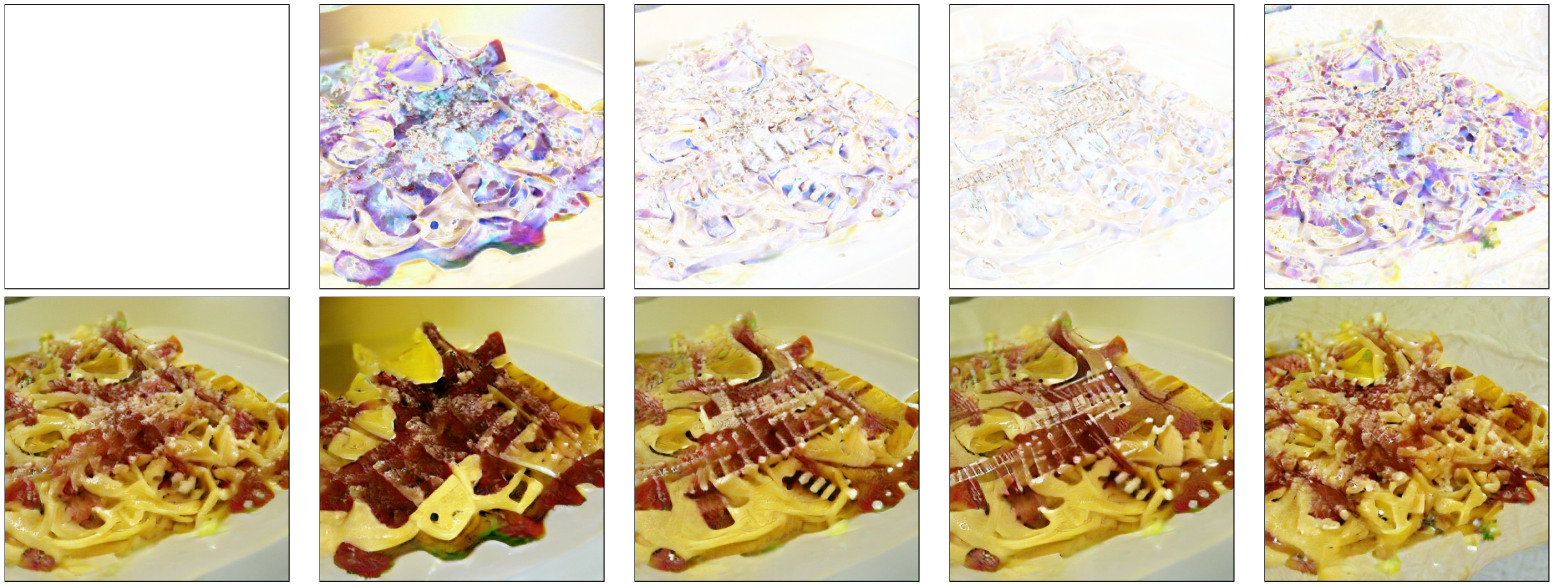
EfficientNet-B4NS

‘carbonara’ → ‘electric guitar’



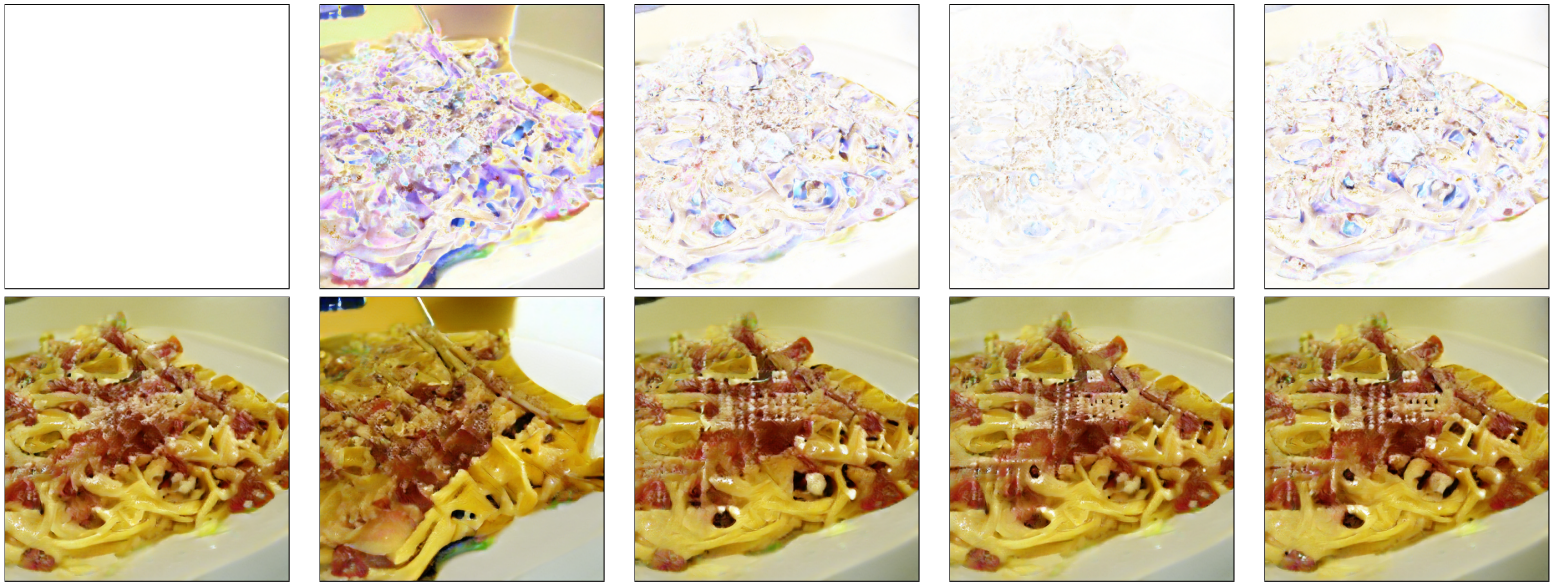
original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'carbonara' → 'electric guitar'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

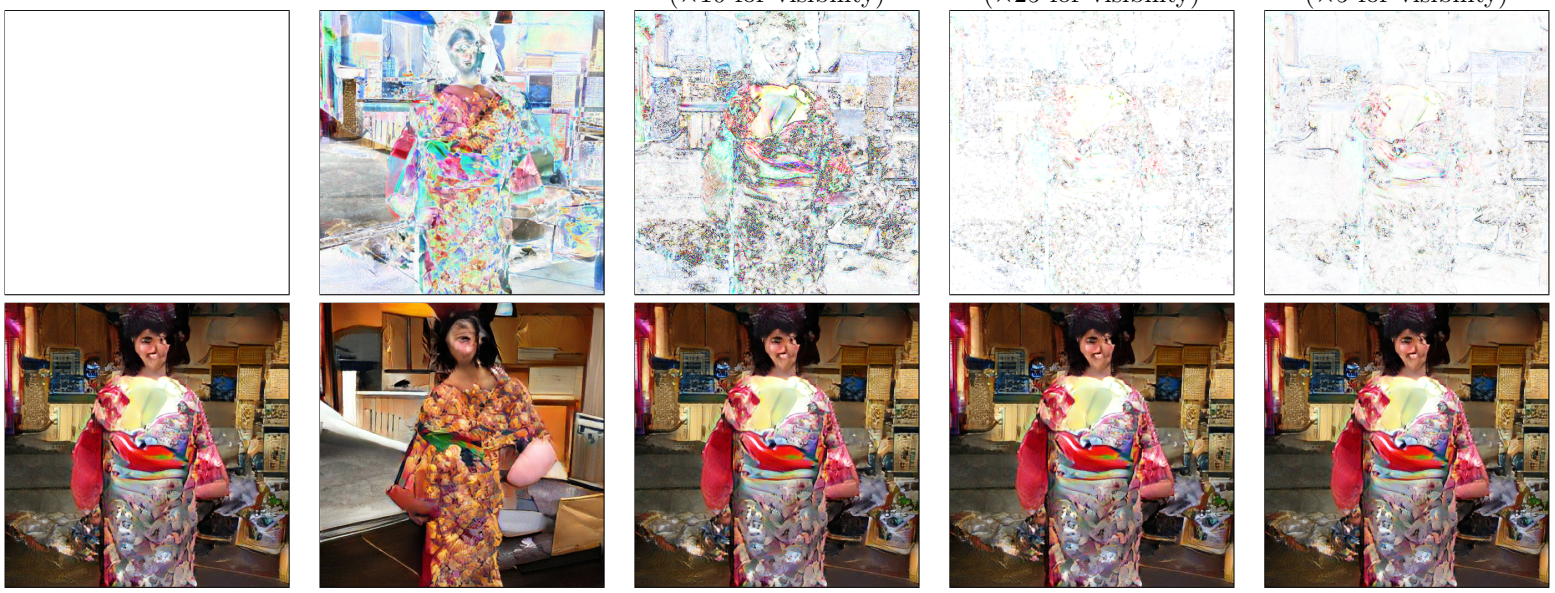
ResNet50 Robust ("Fast") 'carbonara' → 'electric guitar'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

‘kimono’ → ‘Norwegian elkhound’



original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

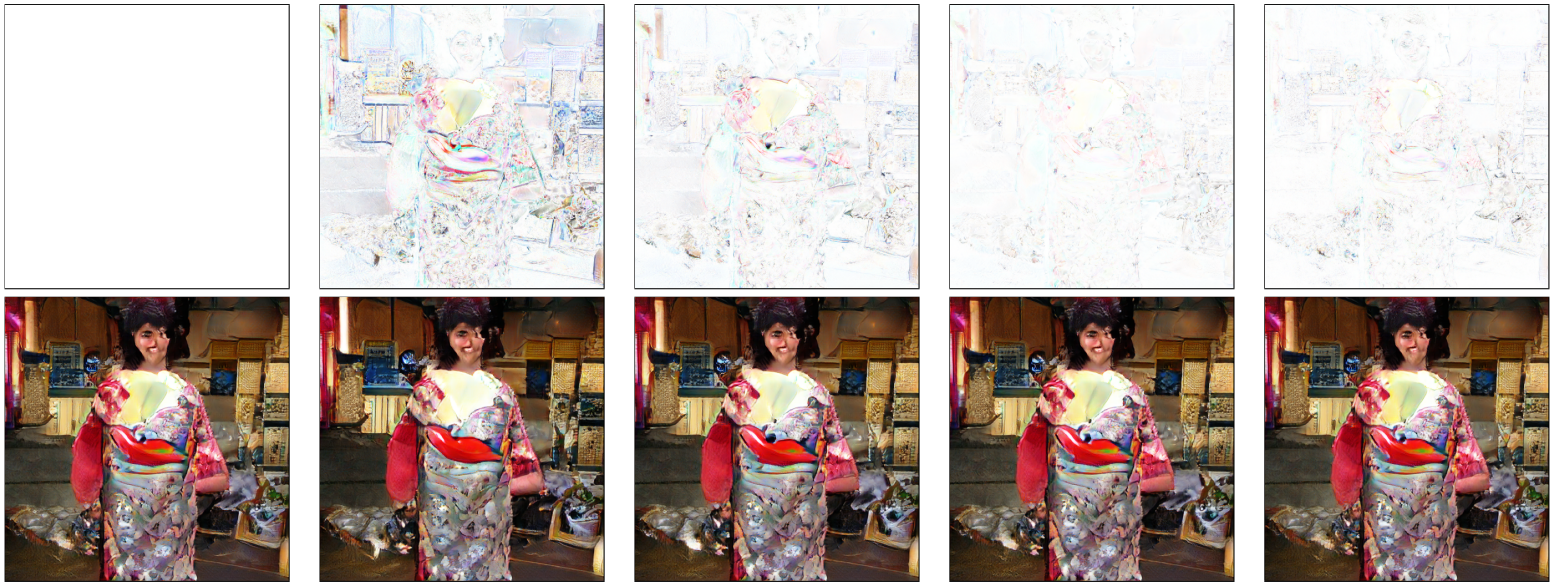
EfficientNet-B4NS

‘kimono’ → ‘Norwegian elkhound’



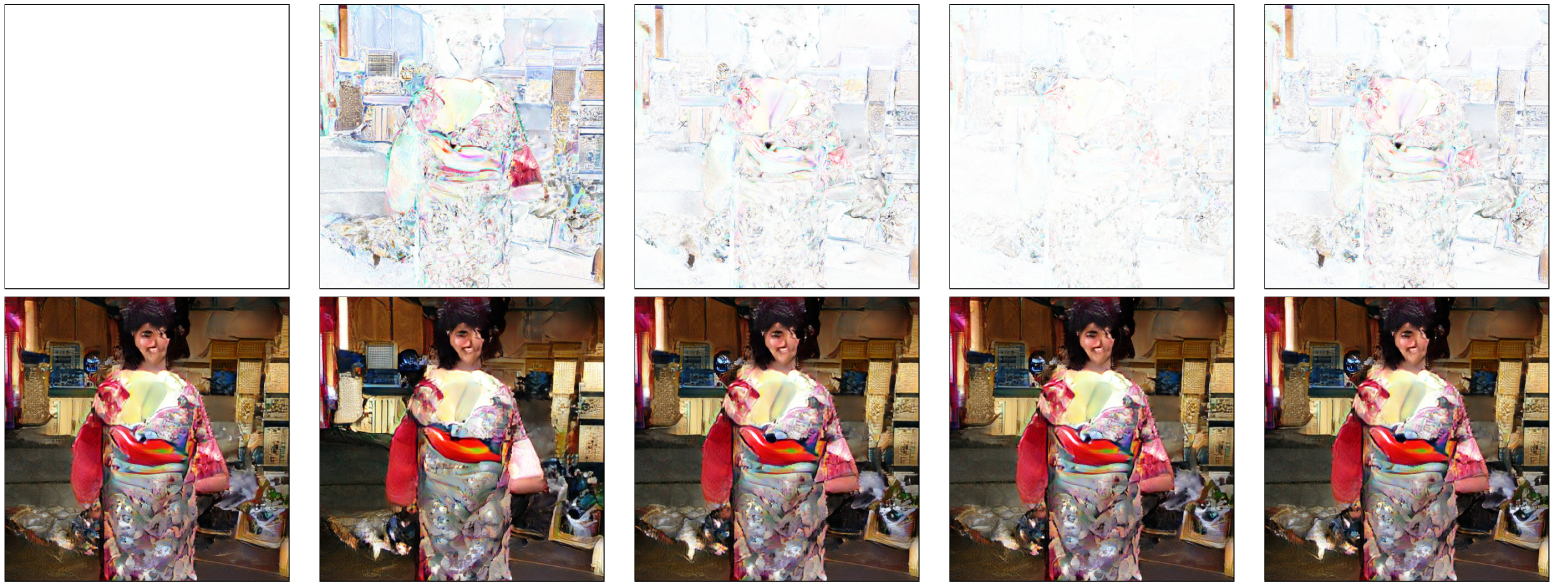
original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

ResNet50 Robust ("Engstrom") 'kimono' → 'Norwegian elkhound'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Fast") 'kimono' → 'Norwegian elkhound'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

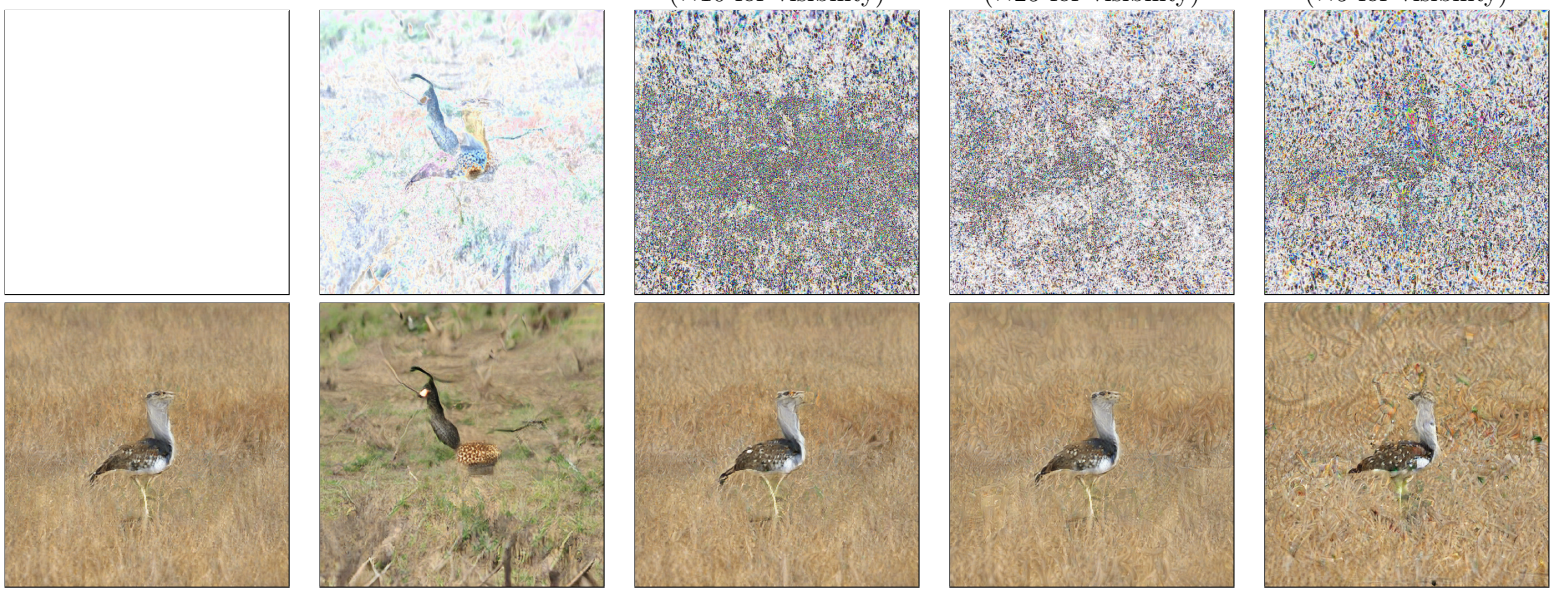
‘bustard’ → ‘cheeseburger’



original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

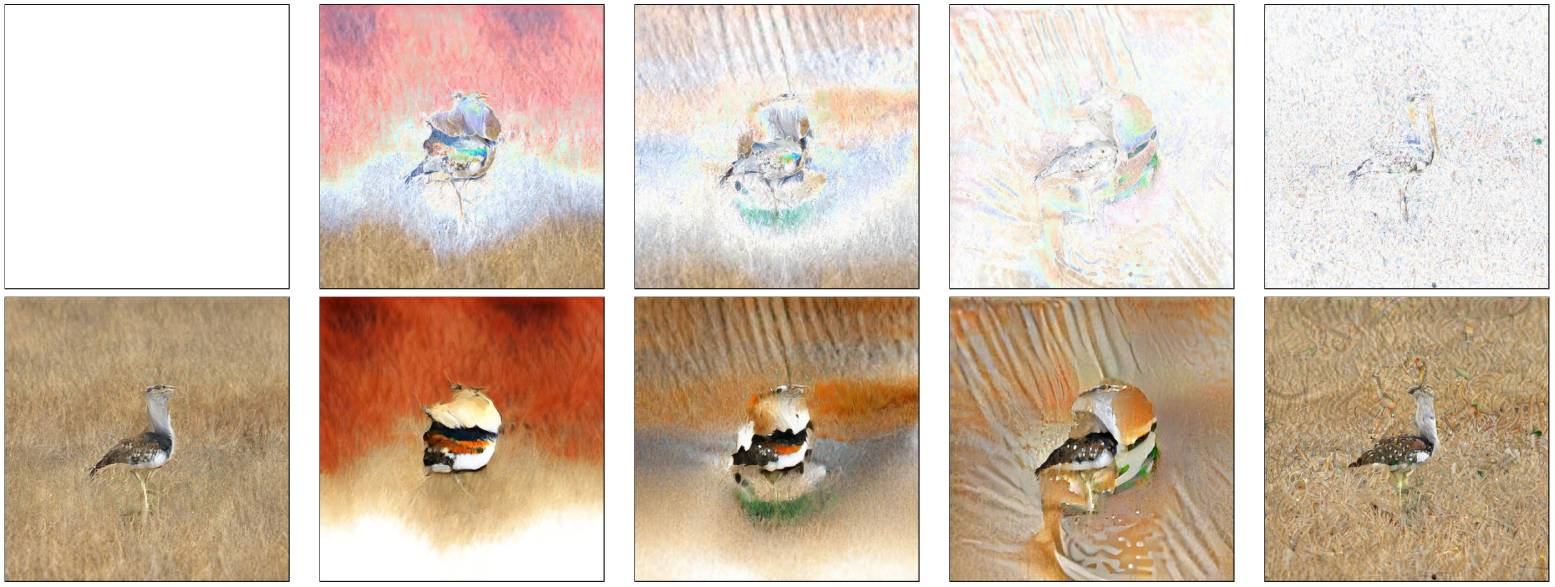
EfficientNet-B4NS

‘bustard’ → ‘cheeseburger’



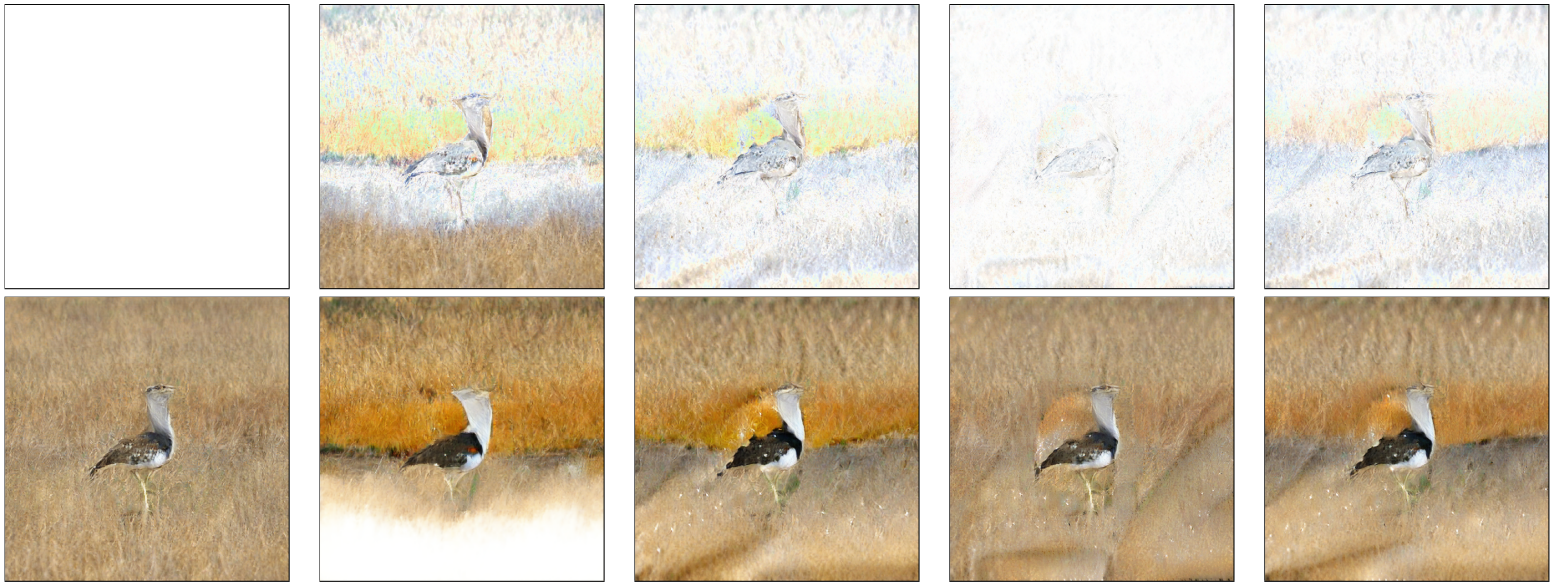
original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

ResNet50 Robust ("Engstrom") 'bustard' → 'cheeseburger'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

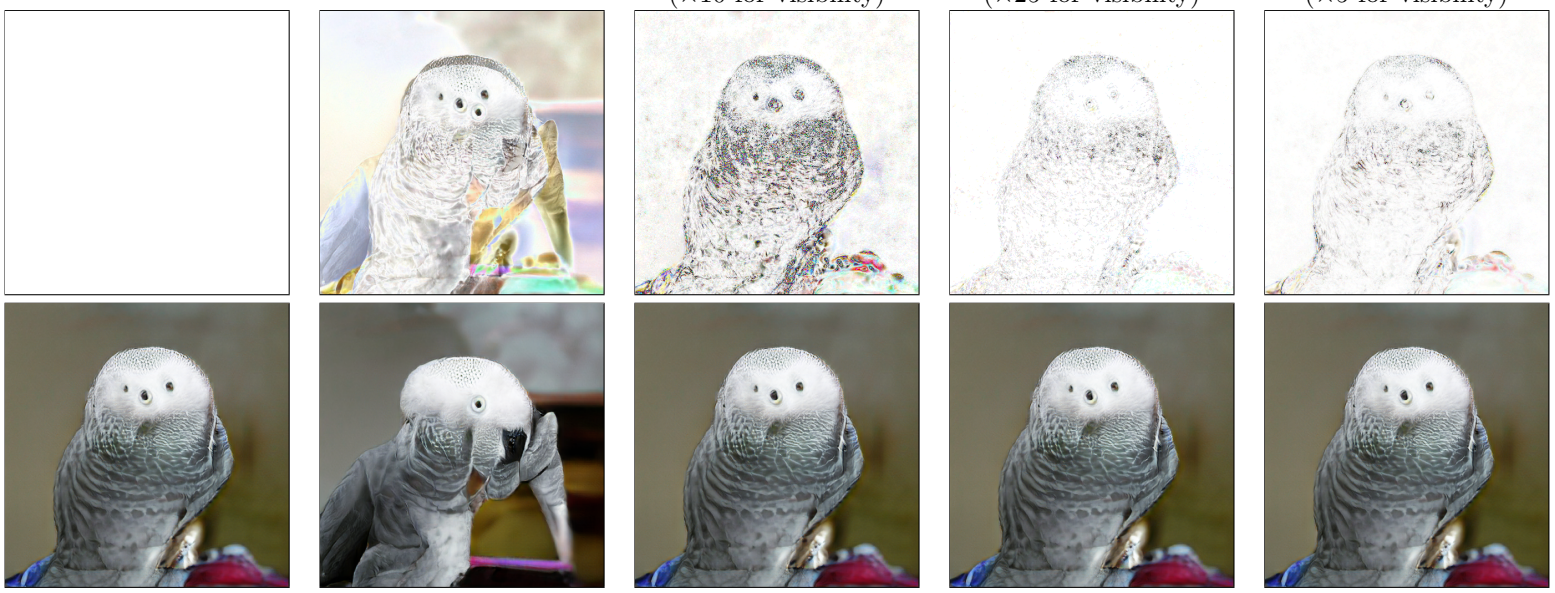
ResNet50 Robust ("Fast") 'bustard' → 'cheeseburger'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

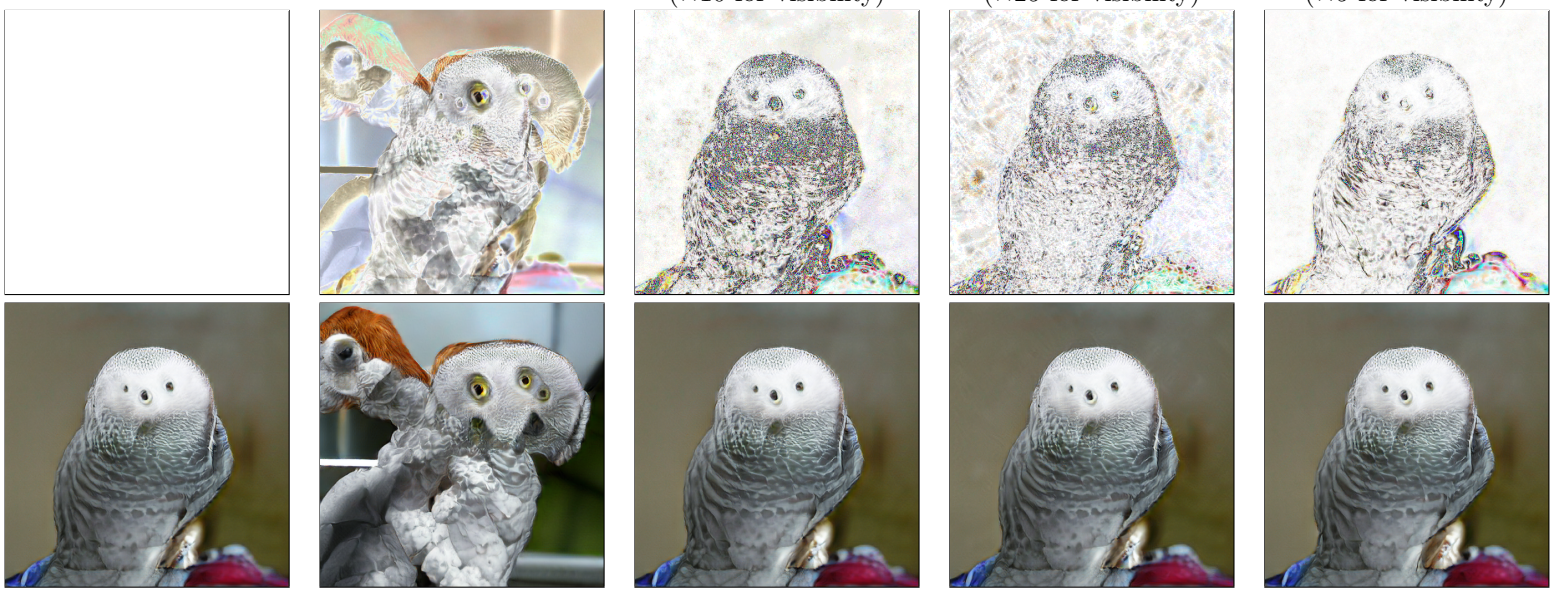
‘African grey’ → ‘cauliflower’



original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

EfficientNet-B4NS

‘African grey’ → ‘cauliflower’



original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

ResNet50 Robust ("Engstrom") 'African grey' → 'cauliflower'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

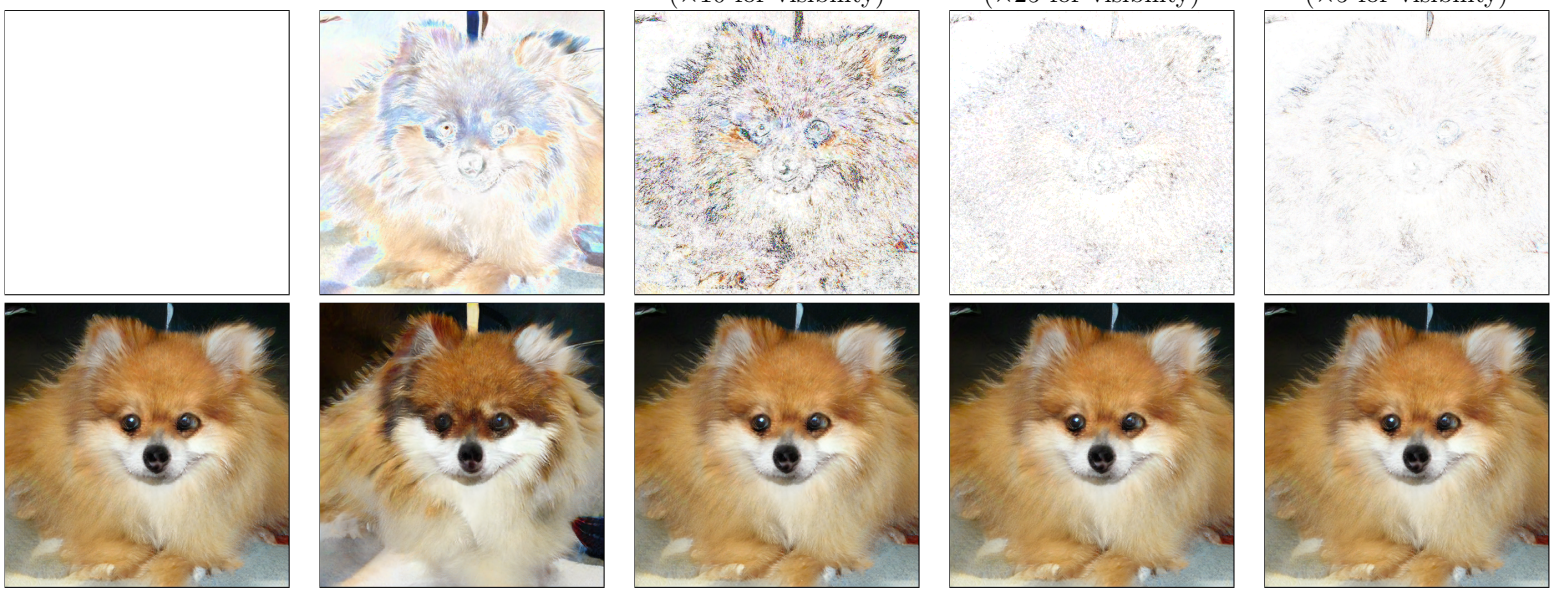
ResNet50 Robust ("Fast") 'African grey' → 'cauliflower'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

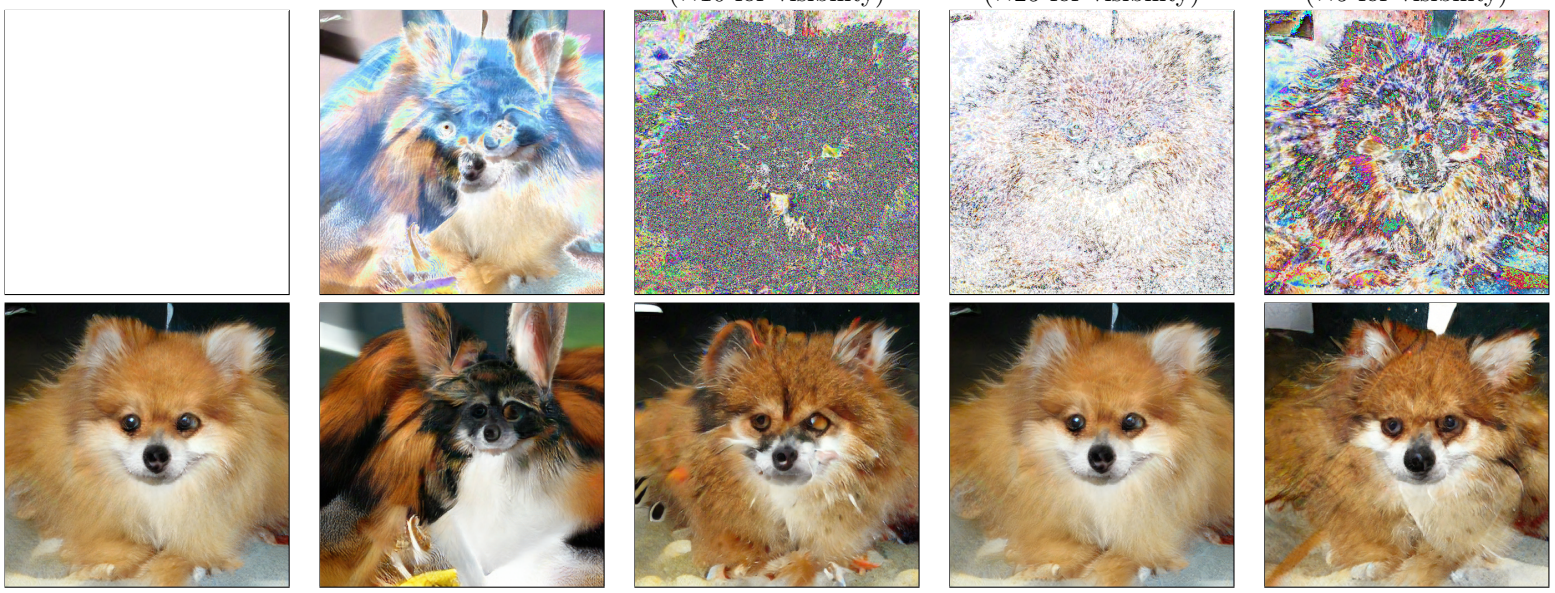
‘Pomeranian’ → ‘American lobster’



original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

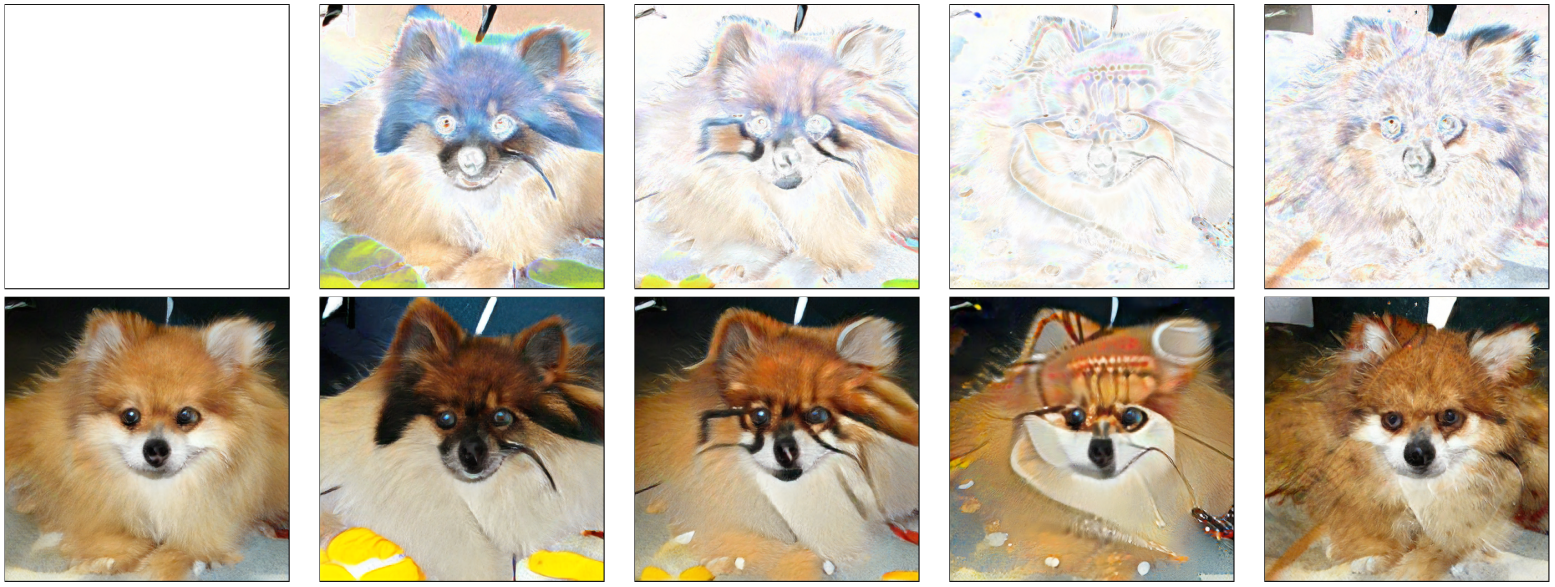
EfficientNet-B4NS

‘Pomeranian’ → ‘American lobster’



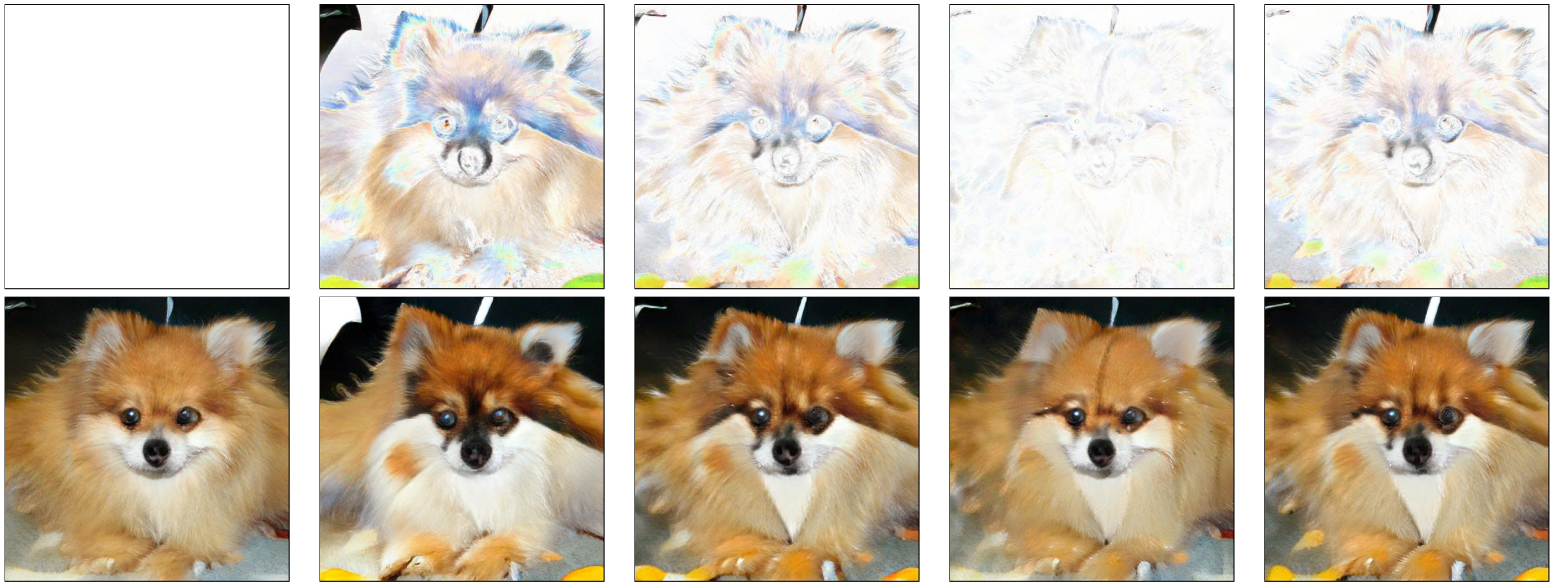
original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

ResNet50 Robust ("Engstrom") 'Pomeranian' → 'American lobster'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

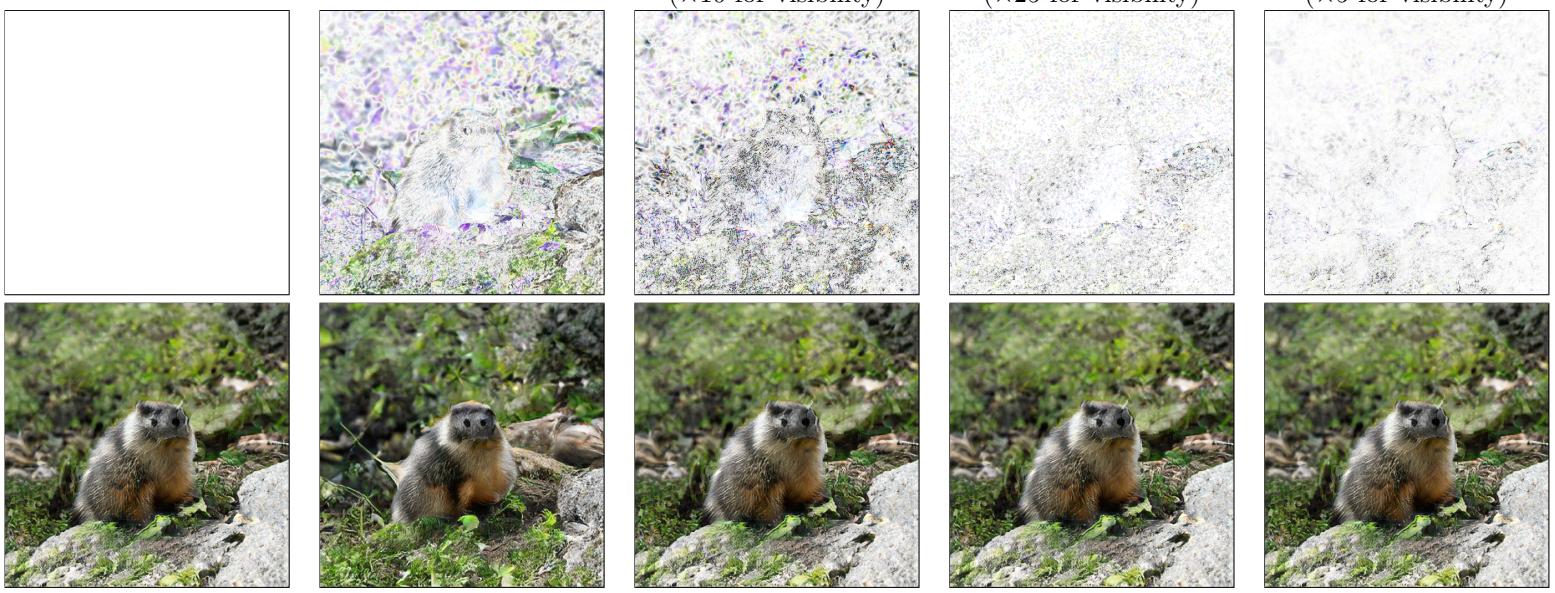
ResNet50 Robust ("Fast") 'Pomeranian' → 'American lobster'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

‘marmot’ → ‘proboscis monkey’



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

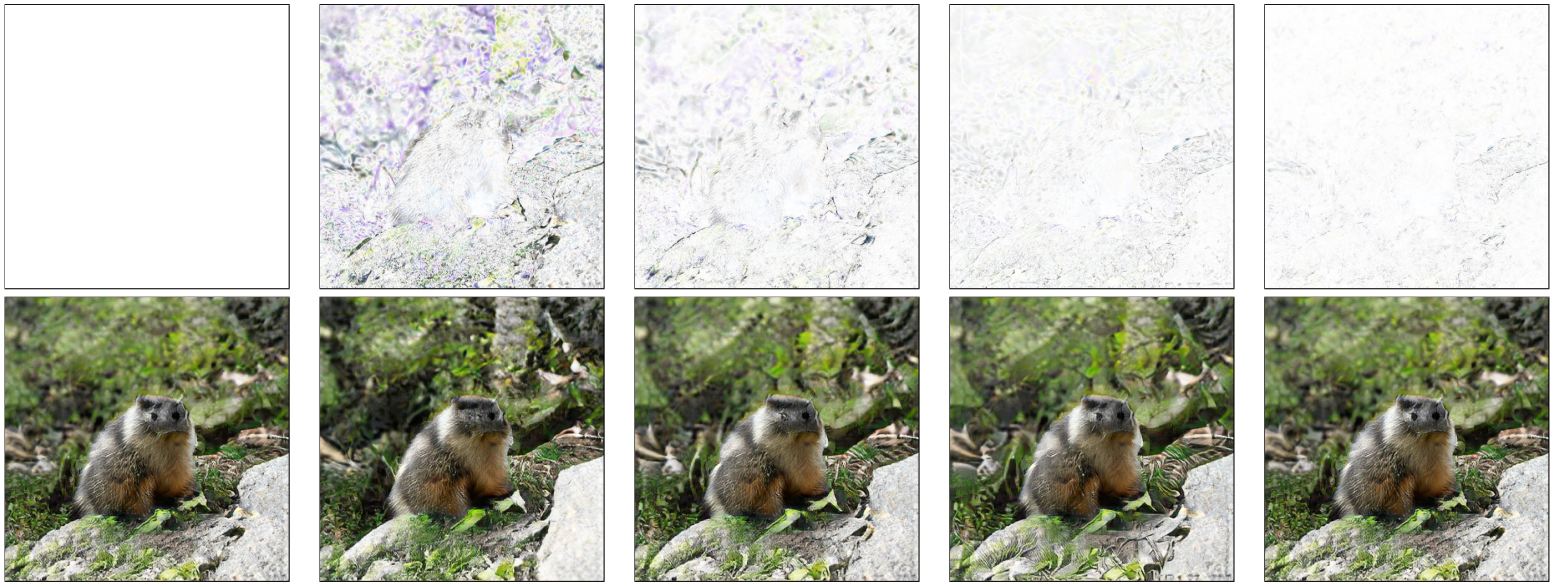
EfficientNet-B4NS

‘marmot’ → ‘proboscis monkey’



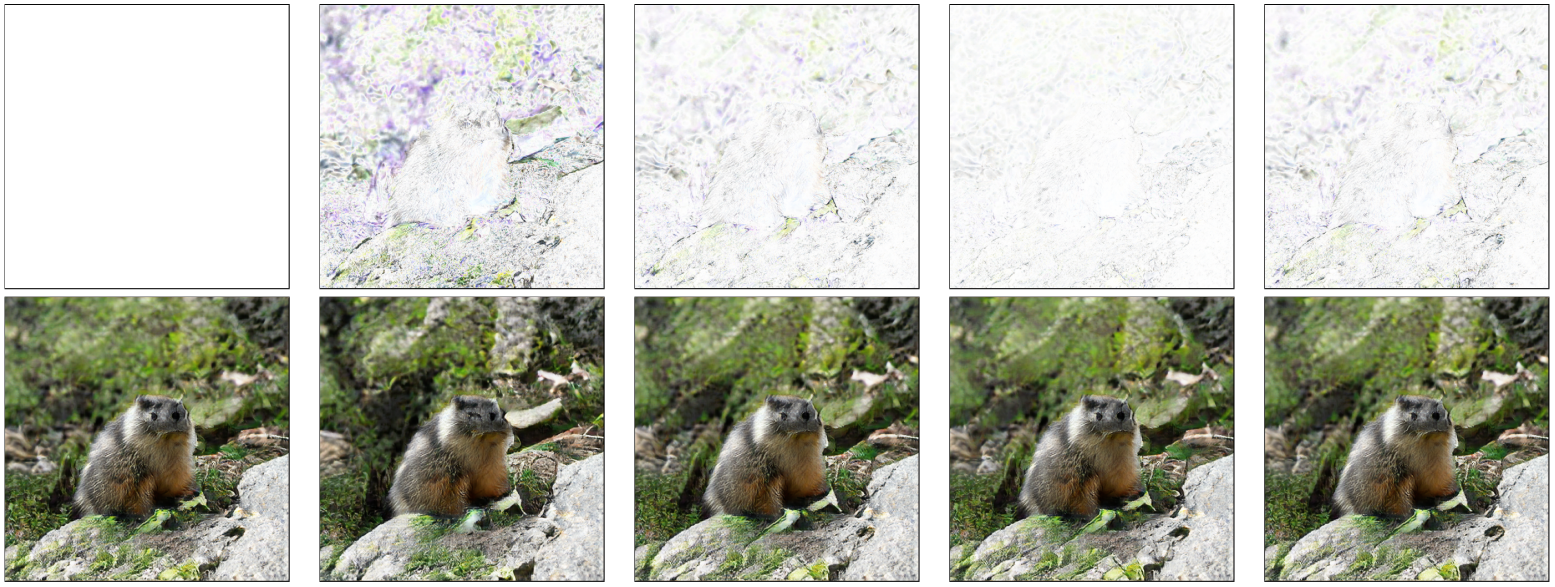
original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Engstrom") 'marmot' → 'proboscis monkey'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

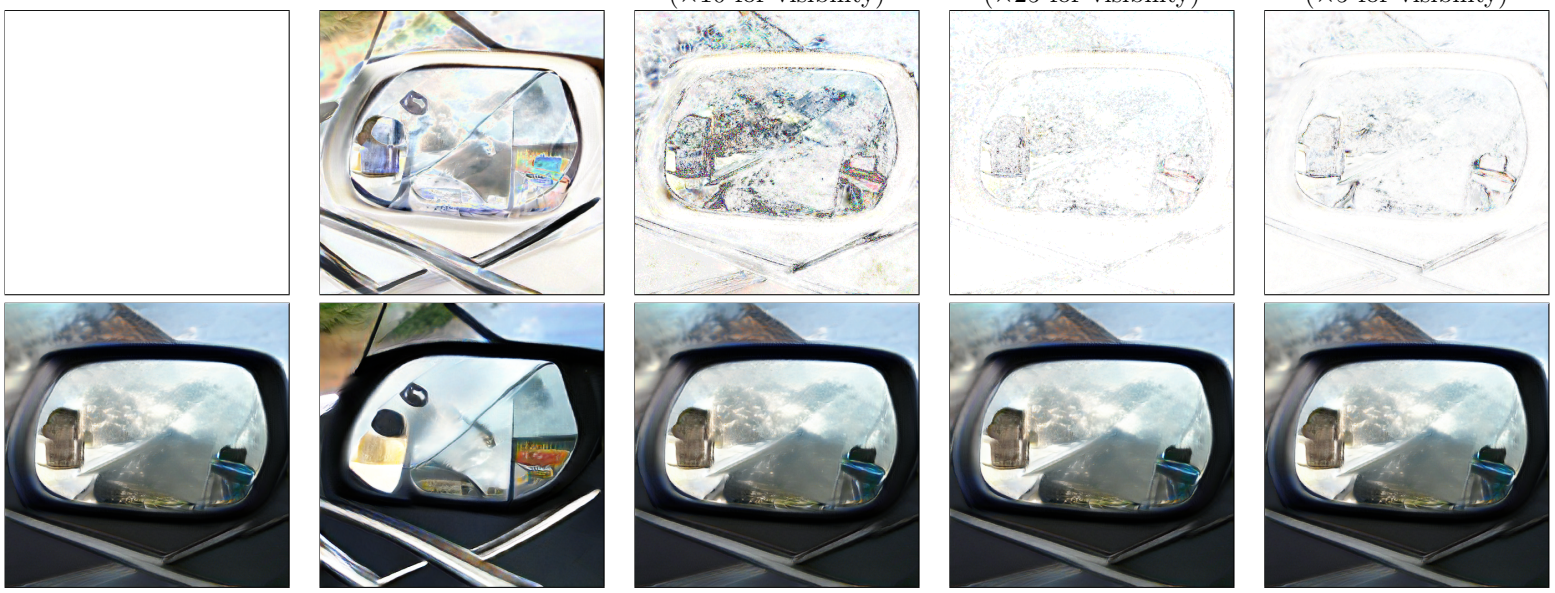
ResNet50 Robust ("Fast") 'marmot' → 'proboscis monkey'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50

‘car mirror’ → ‘timber wolf’



original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

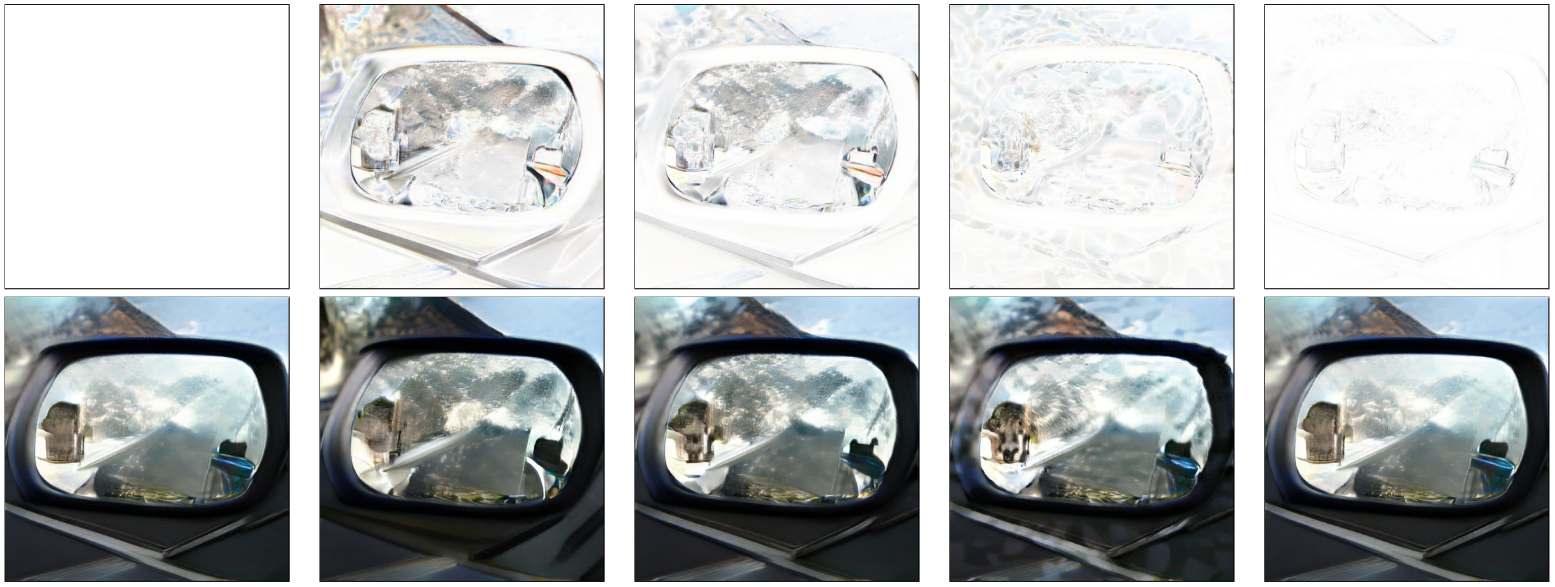
EfficientNet-B4NS

‘car mirror’ → ‘timber wolf’



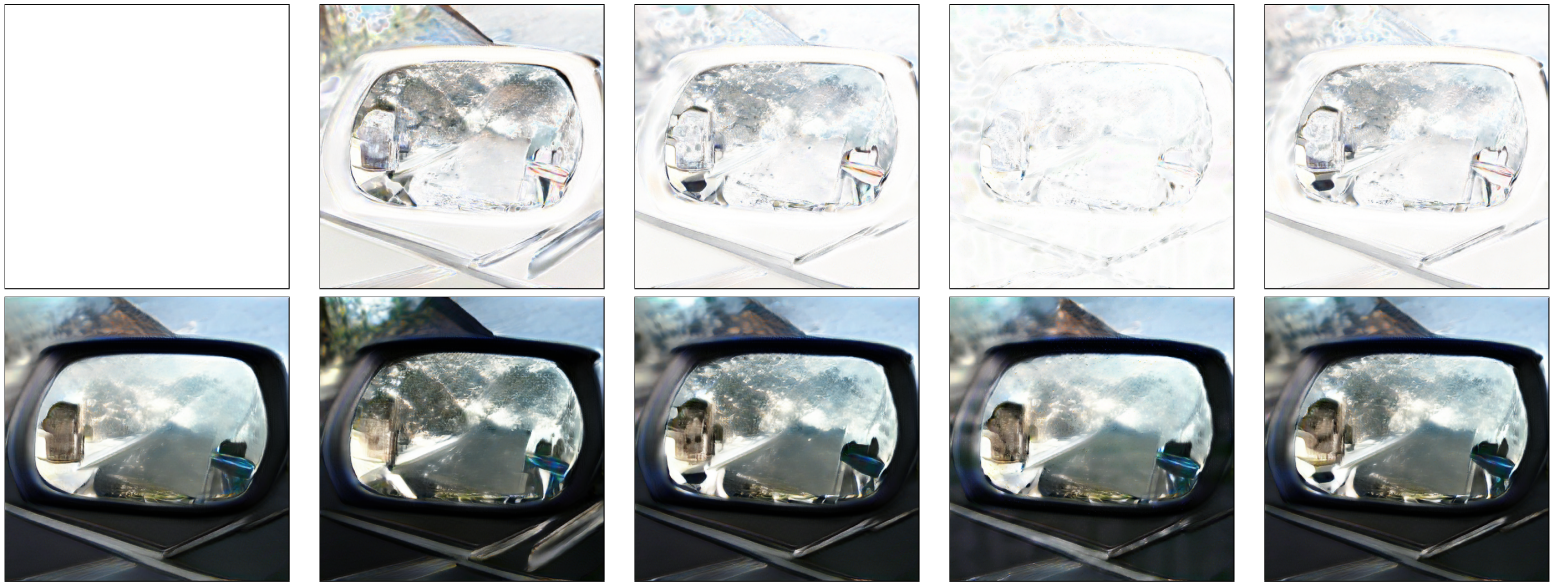
original | first 6 layers | middle 6 layers (×10 for visibility) | last 6 layers (×25 for visibility) | all 18 layers (×5 for visibility)

ResNet50 Robust ("Engstrom") 'car mirror' → 'timber wolf'



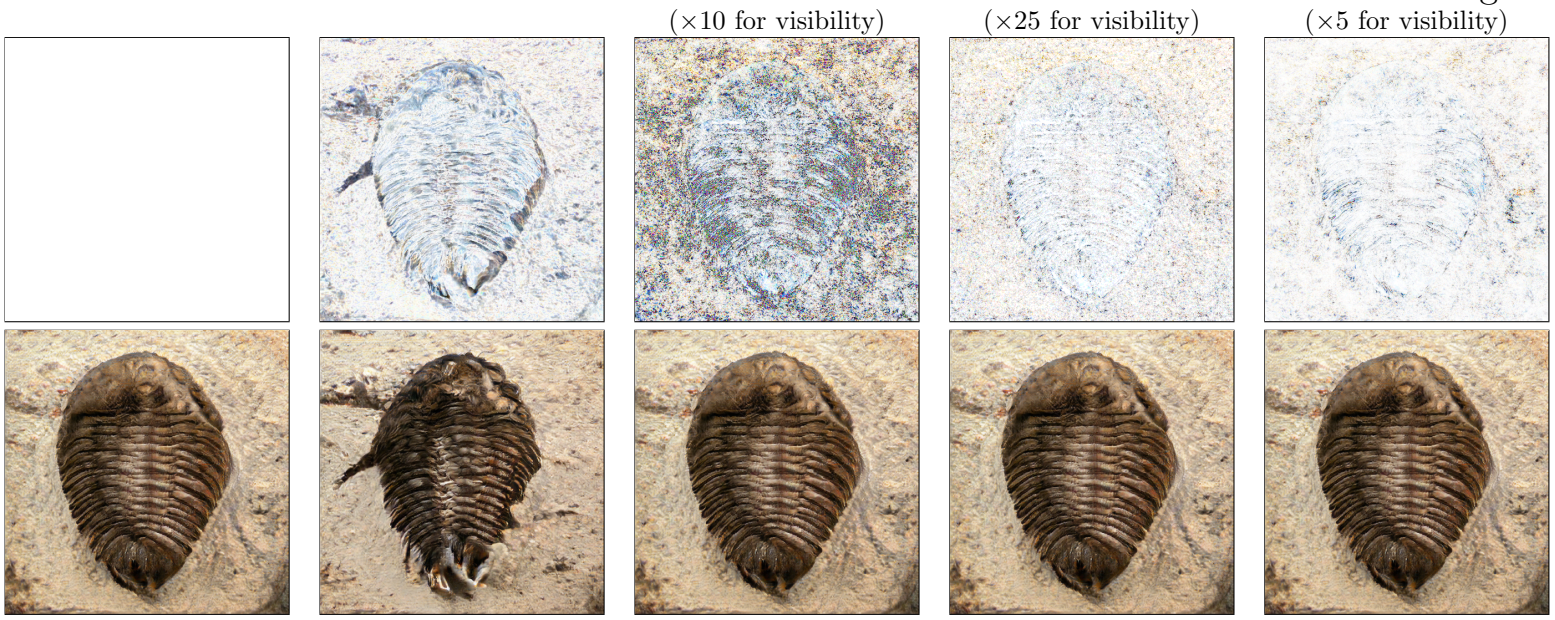
original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Fast") 'car mirror' → 'timber wolf'

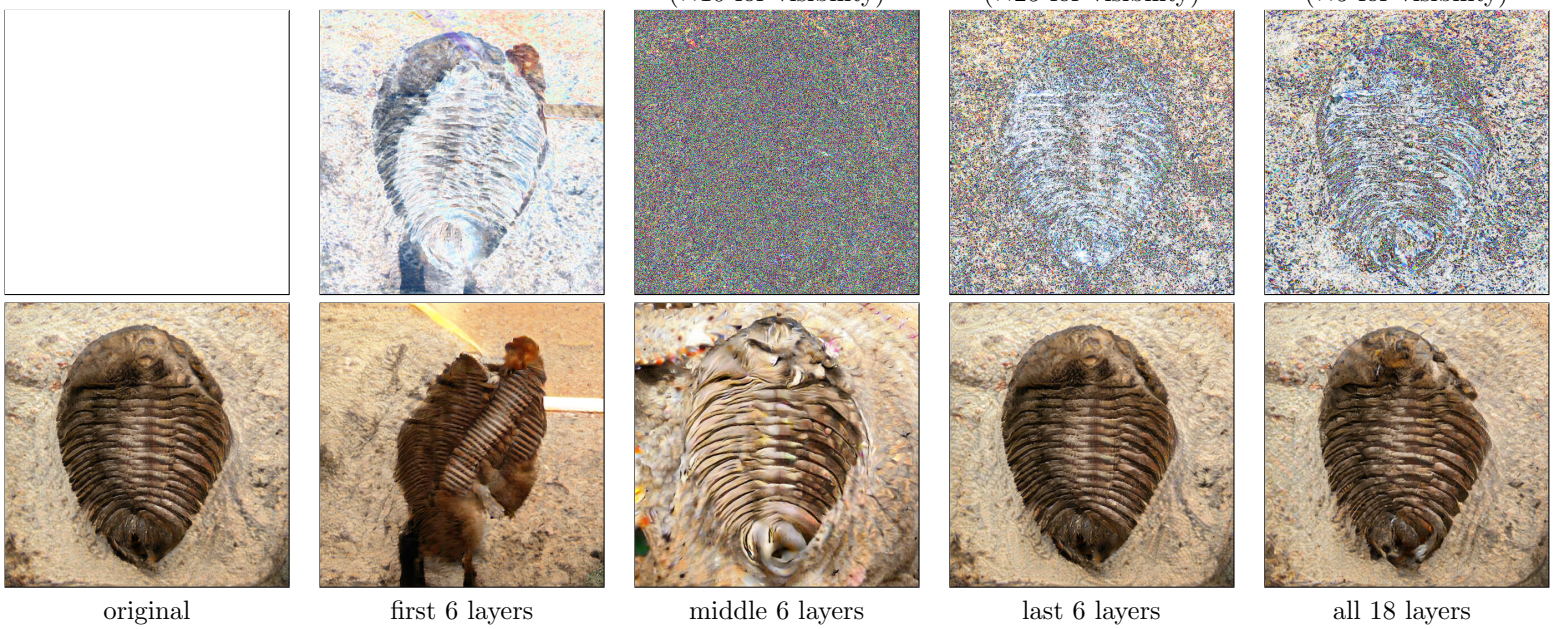


original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 ‘trilobite’ → ‘Cardigan’



EfficientNet-B4NS ‘trilobite’ → ‘Cardigan’

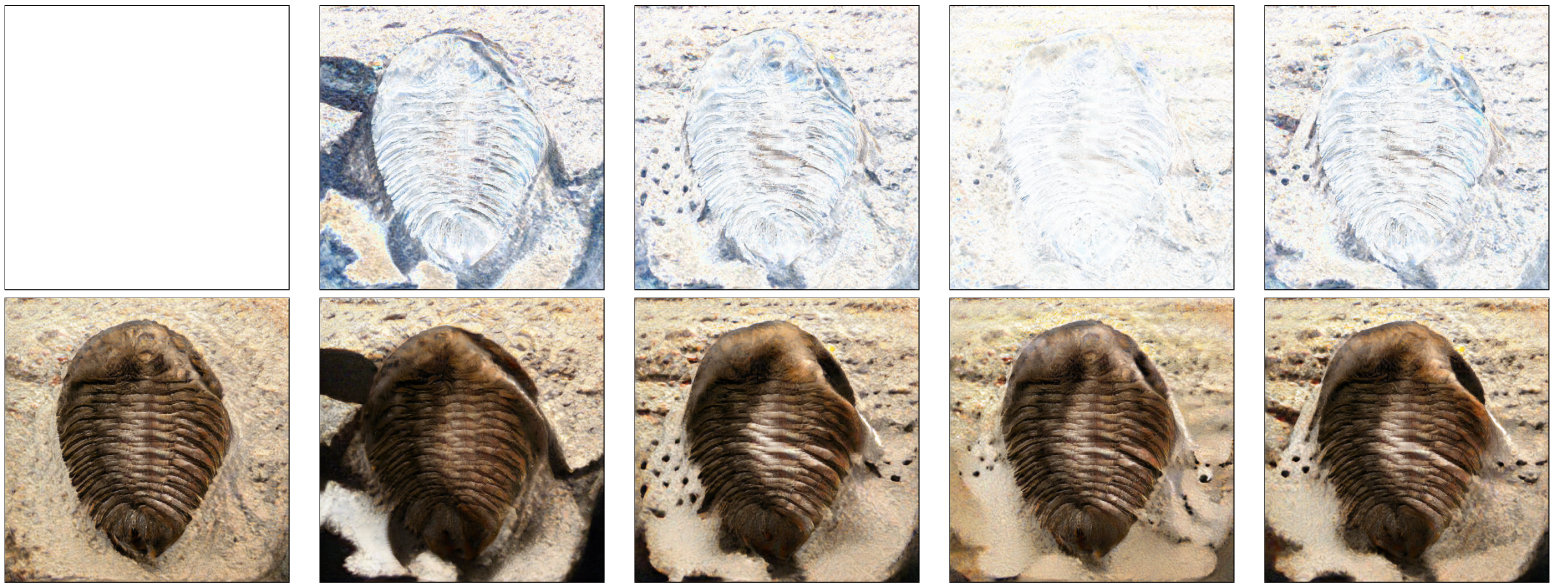


ResNet50 Robust ("Engstrom") 'trilobite' → 'Cardigan'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers

ResNet50 Robust ("Fast") 'trilobite' → 'Cardigan'



original | first 6 layers | middle 6 layers | last 6 layers | all 18 layers