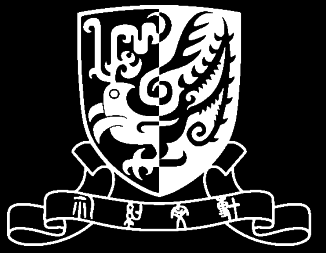
Structure and Interpretation of Deep Networks in Vision

David Bau, Jun-Yan Zhu, Hendrik Strobelt,
Bill Peebles, Jonas Wulff, Bolei Zhou, Antonio Torralba

MIT-IBM Watson AI Lab



What does a Deep Net Learn?

Top5 acc 85%

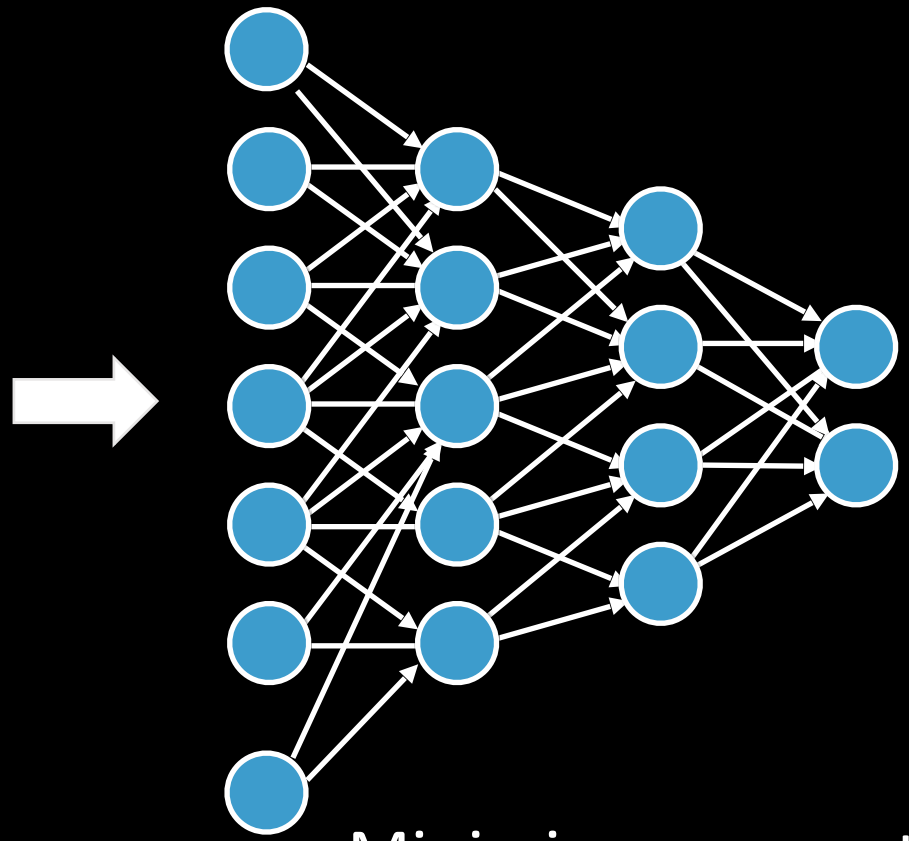
✅ ✅ ✅

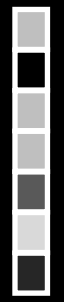
Prediction vector

“indoor market”

“baseball field”

“tree farm”



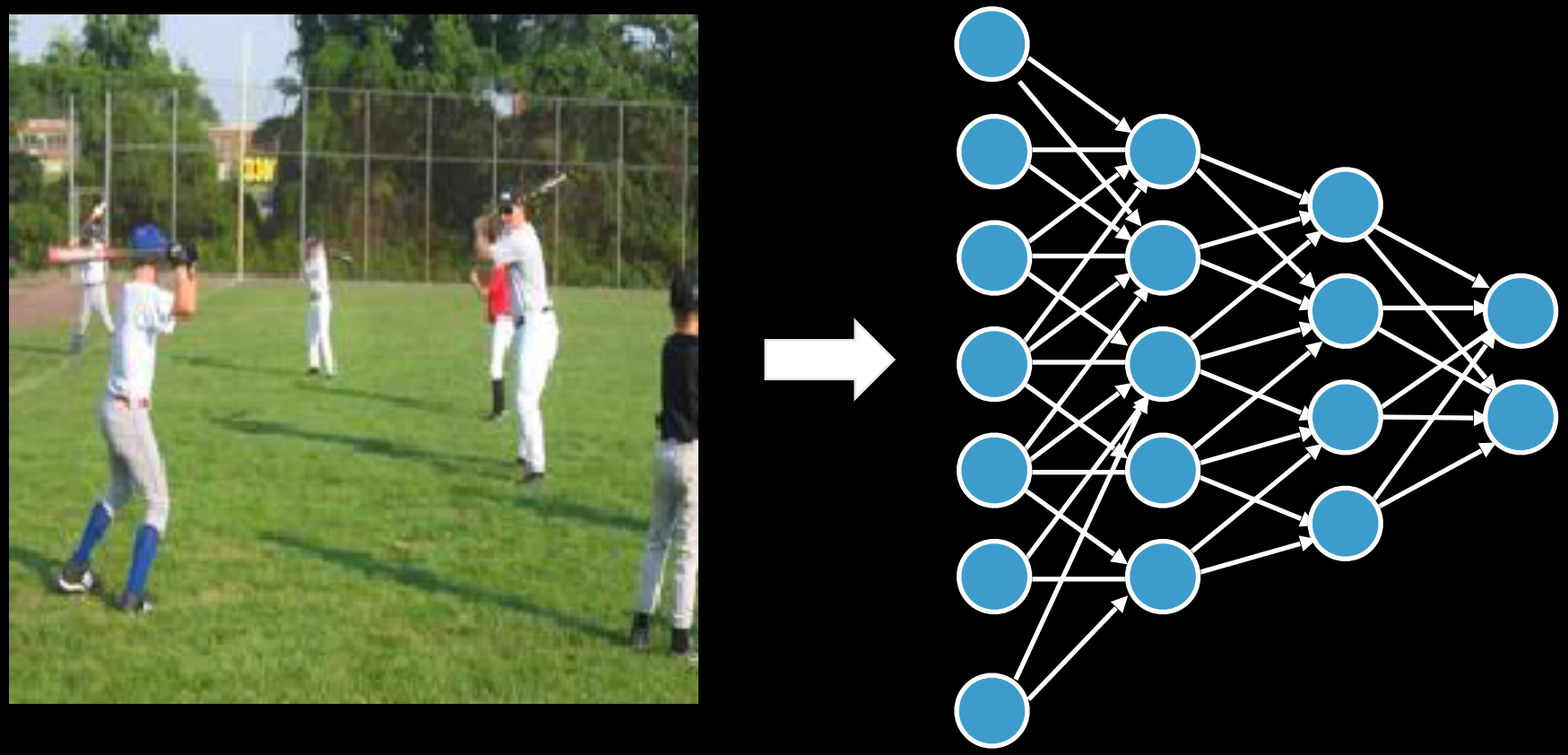


365 dimensions

Minimize cross-entropy loss

[Places dataset, Zhou 2016]

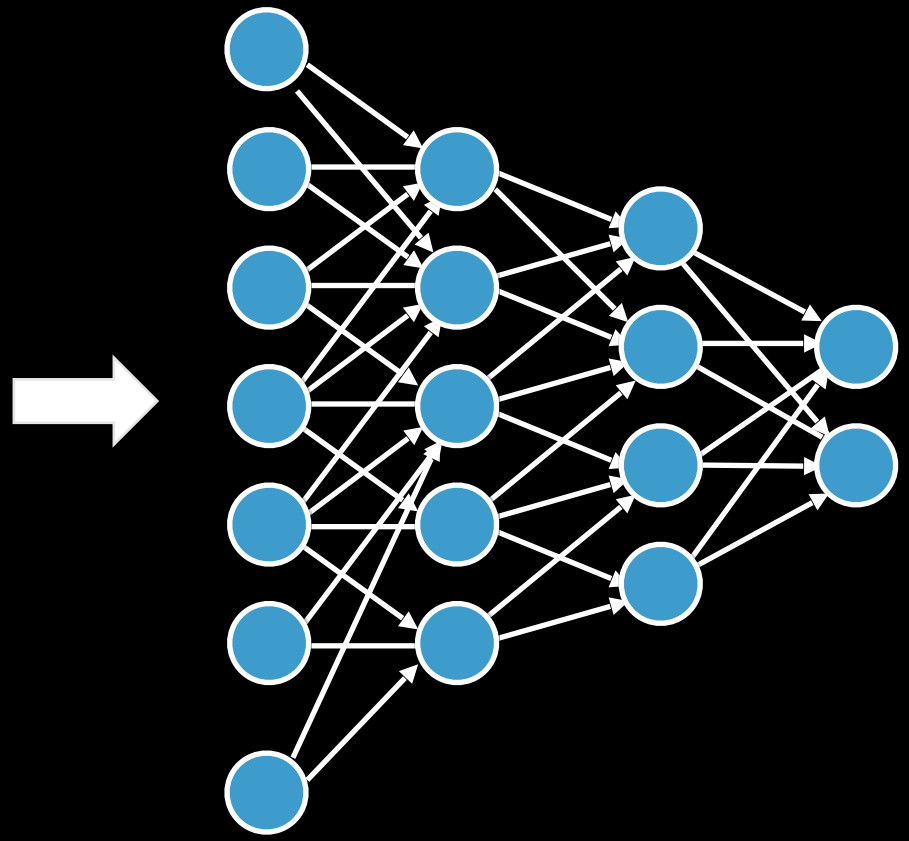
What does it really learn?



[VGG-16, Simonyan 2014]

What does it really learn?

Where did the network look?



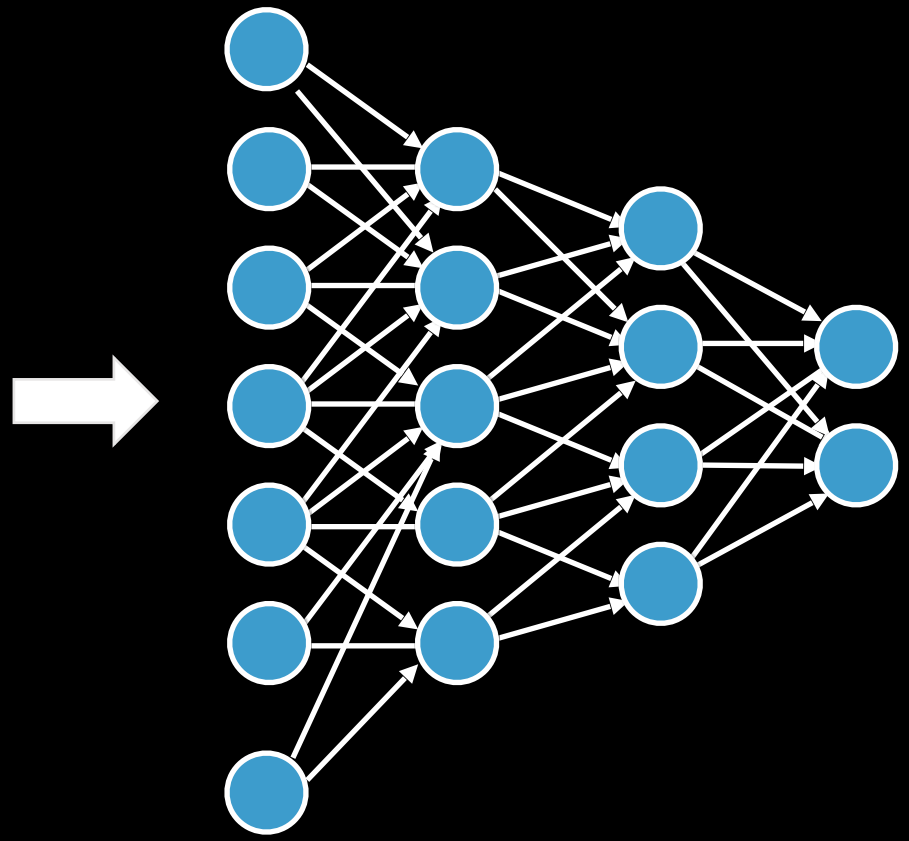
Prediction vector

“baseball field”

Grad-CAM [Selvaraju CVPR 2017]

What does it really learn?

Why is the network looking there?



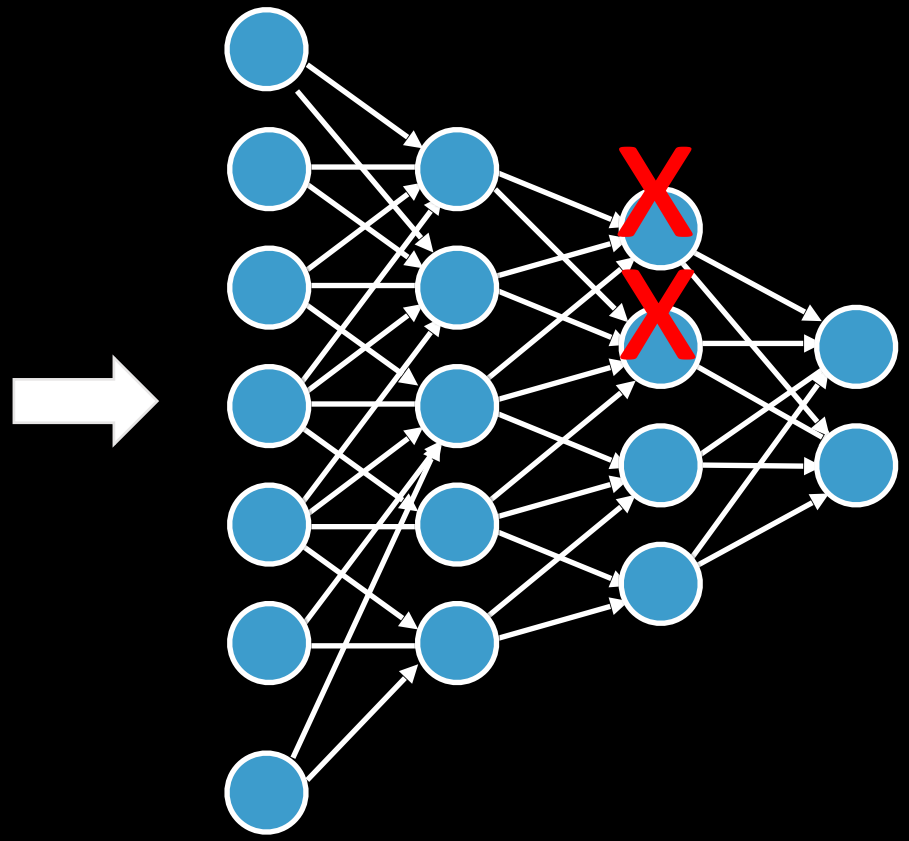
Prediction vector

“baseball field”

Where did the network look?

Grad-CAM [Selvaraju CVPR 2017]

What is the network looking for?



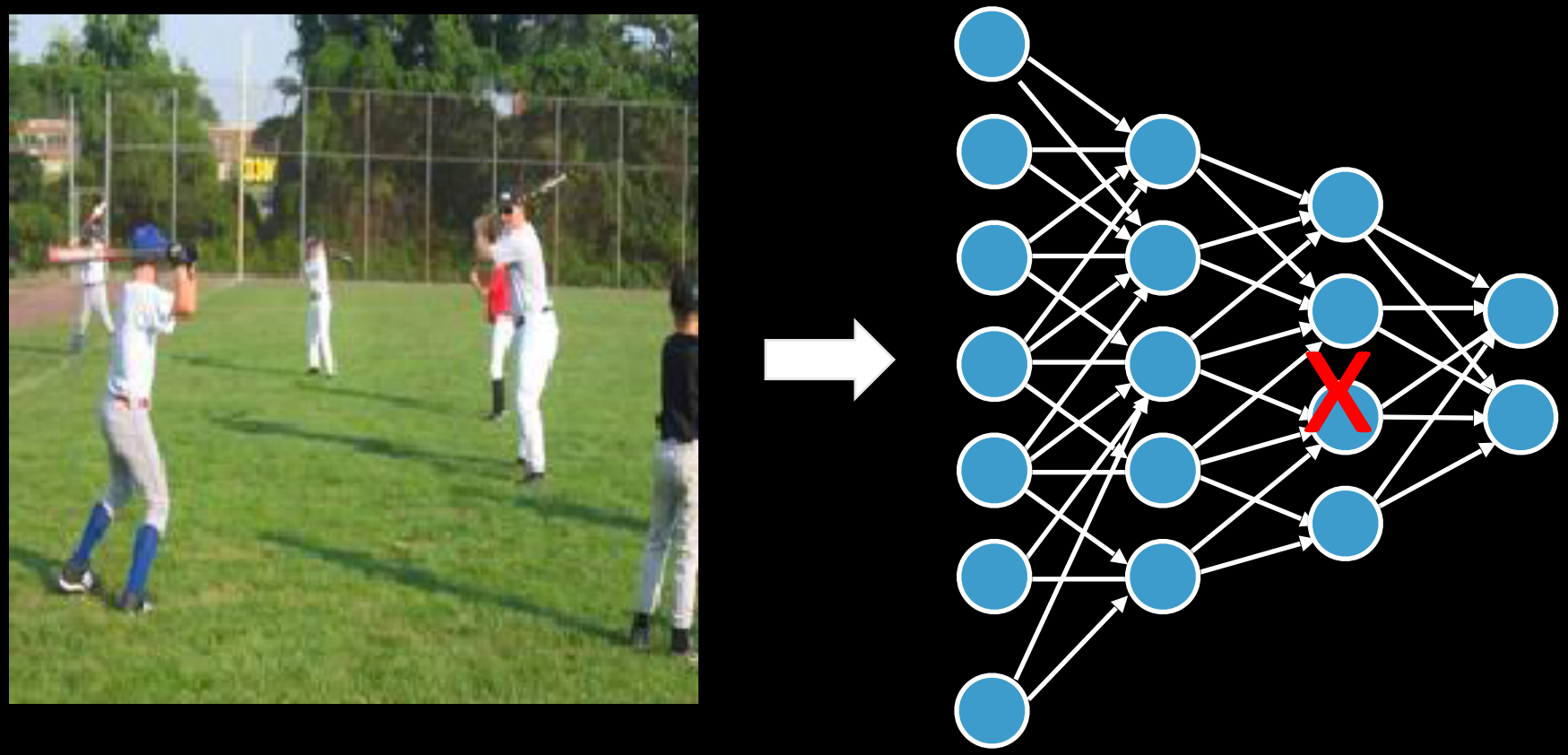
Prediction vector

“baseball field”

unchanged

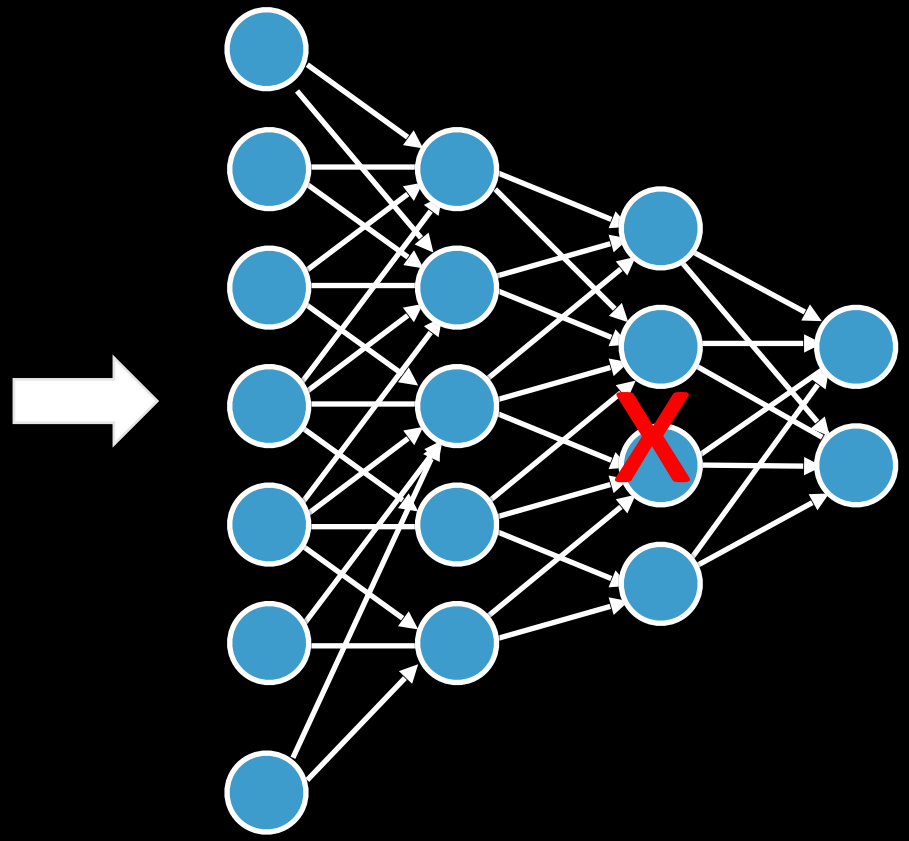
509/512 units do not change prediction when removed

What the ~~network~~ looking for?



Units 208, 467, and 161 do change this prediction when removed

What are the units looking for?



Prediction vector

"SOCCER field"

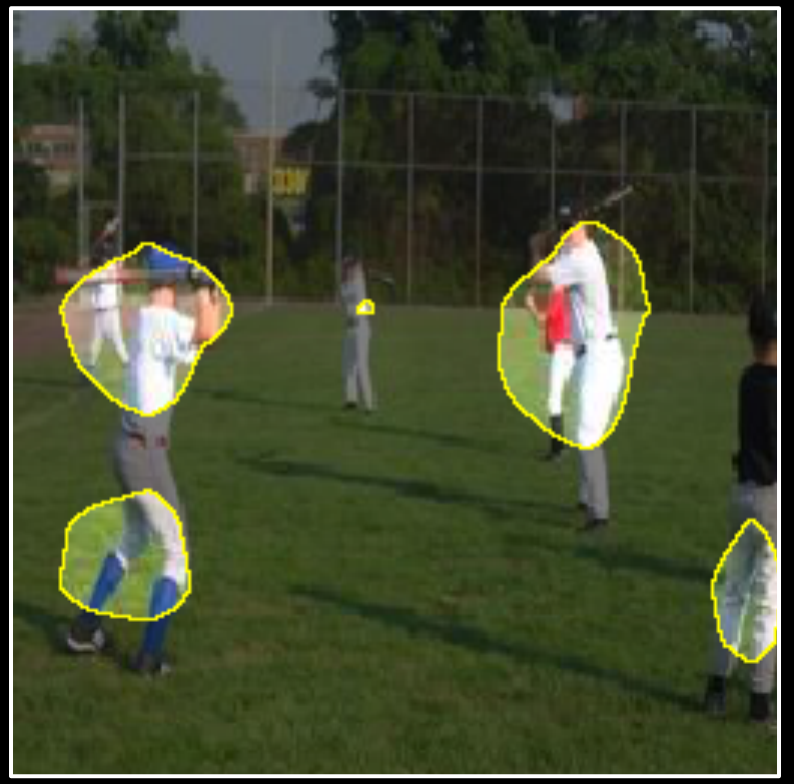
WRONG

Units 208, 467, and 161 do change this prediction when removed

What are the units looking for?

Top 1% activations of unit 208

Unit 467

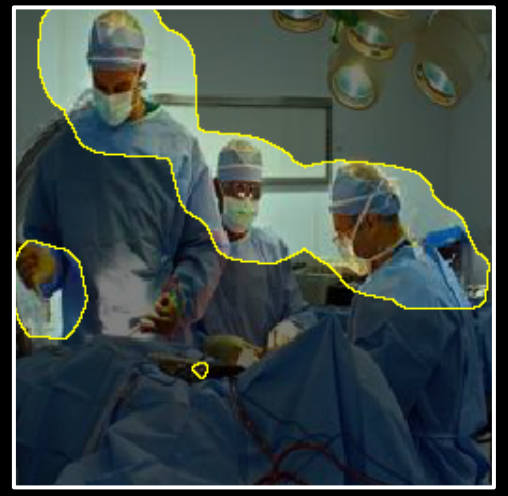


Unit 161

What is unit 208 looking for?

Top 1% activations of unit 208





Top 1% activations of unit 208 across whole data set

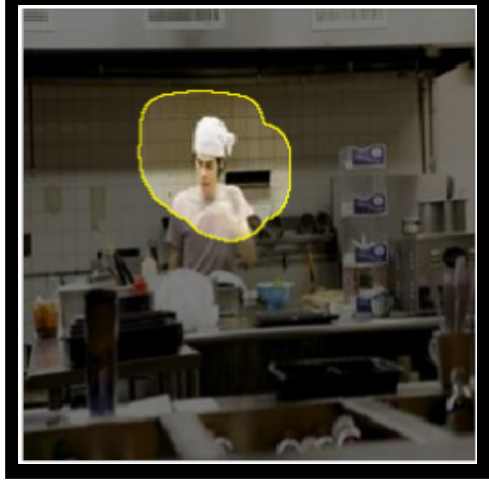
The role of a unit

Unit 208

What it can detect

What impact it has

Detects: hats



Affects output: baseball field

Why do interpretable units emerge?

Supervision hypothesis:

Each concept unit emerges due to the supervision of an output class.

The role of a unit

Output class:
ski resort

Accuracy -11%

Accuracy -3.5%

Accuracy -2.5%

The role of a unit

Output class:
ski resort

Accuracy -11%

Accuracy -3.5%

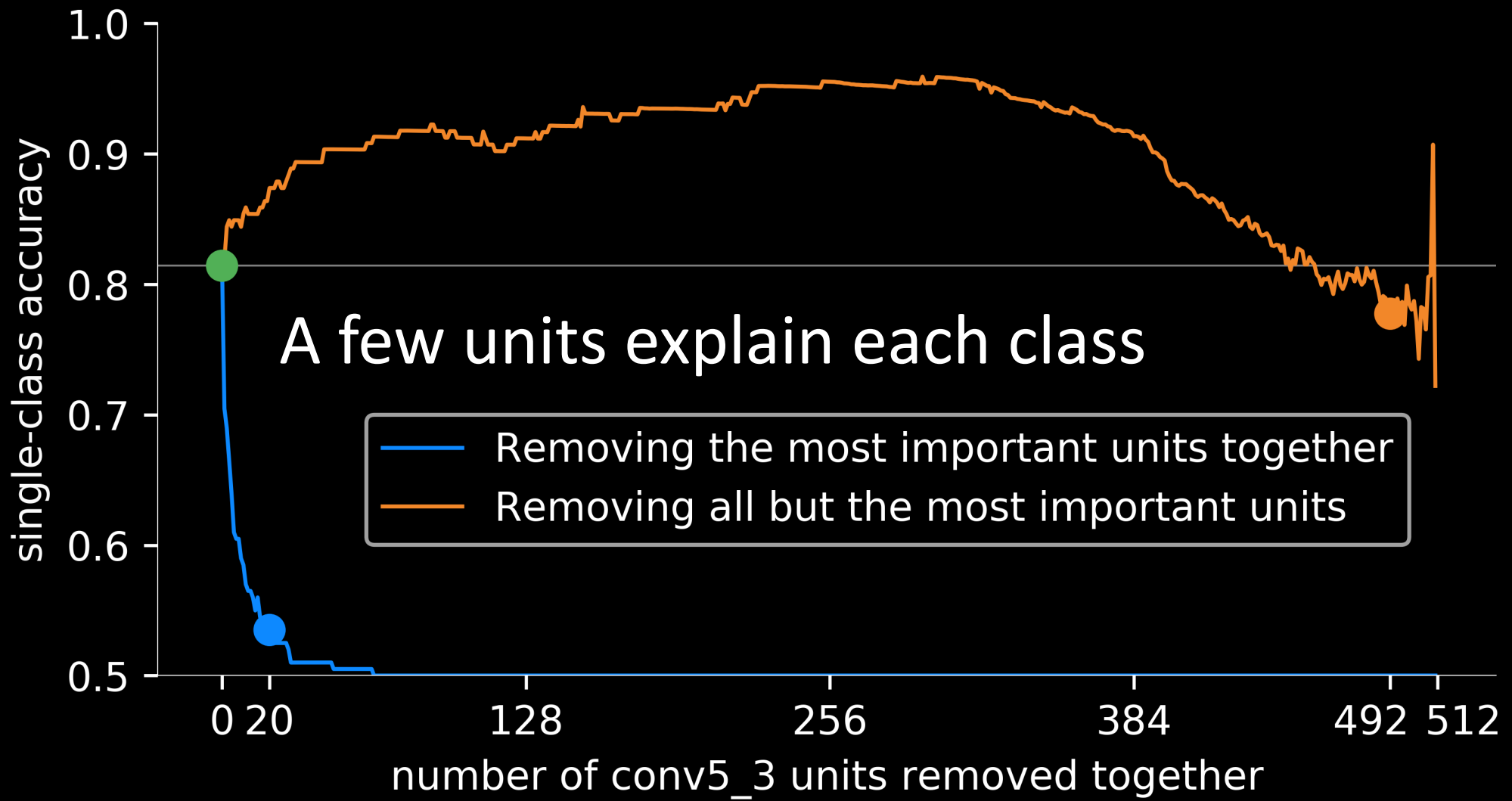
Accuracy -2.5%

Remove all three: accuracy -15%

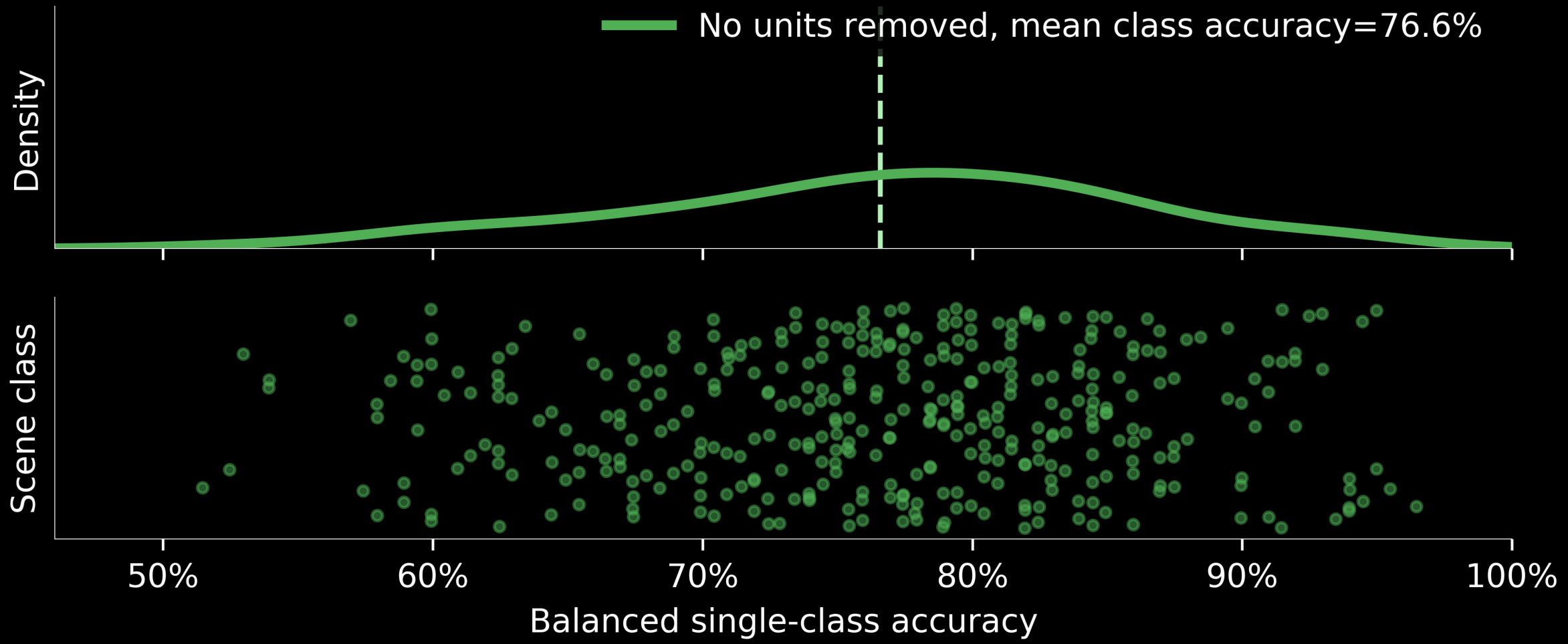
The role of a unit

A few units explain each class

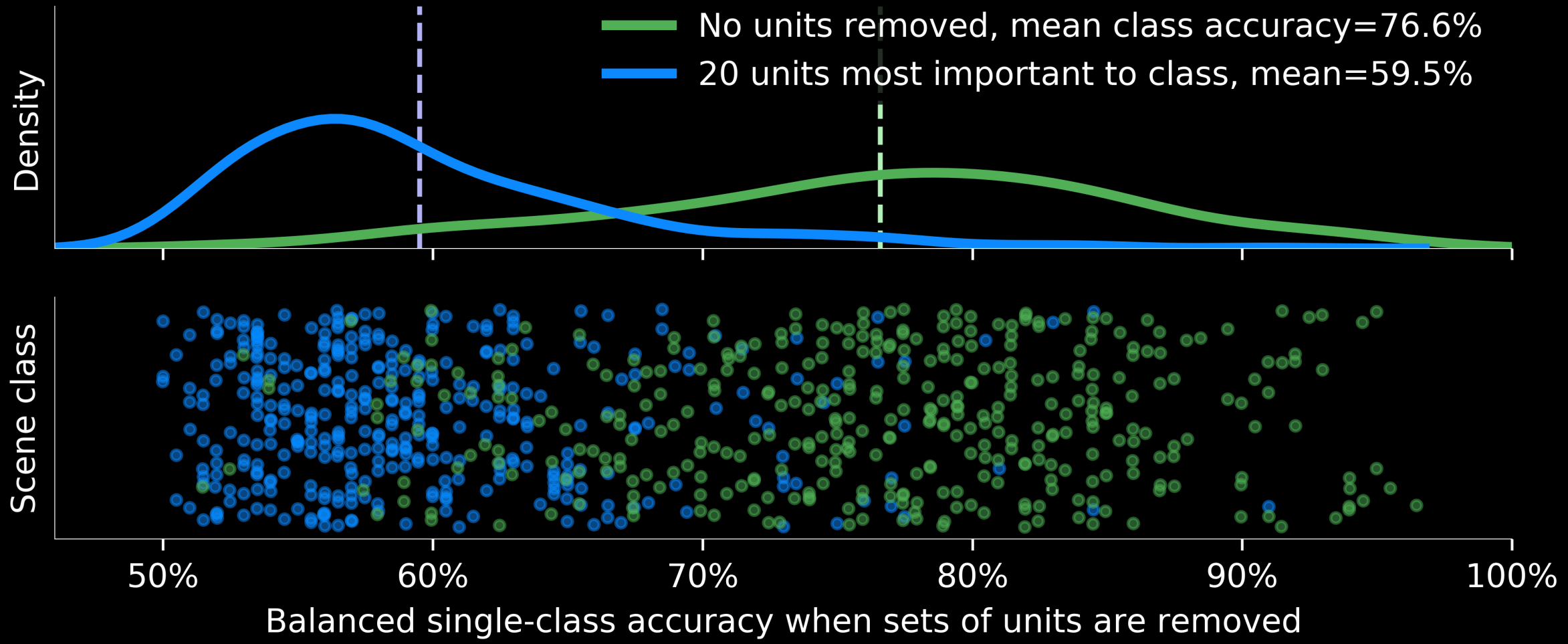
The role of a unit



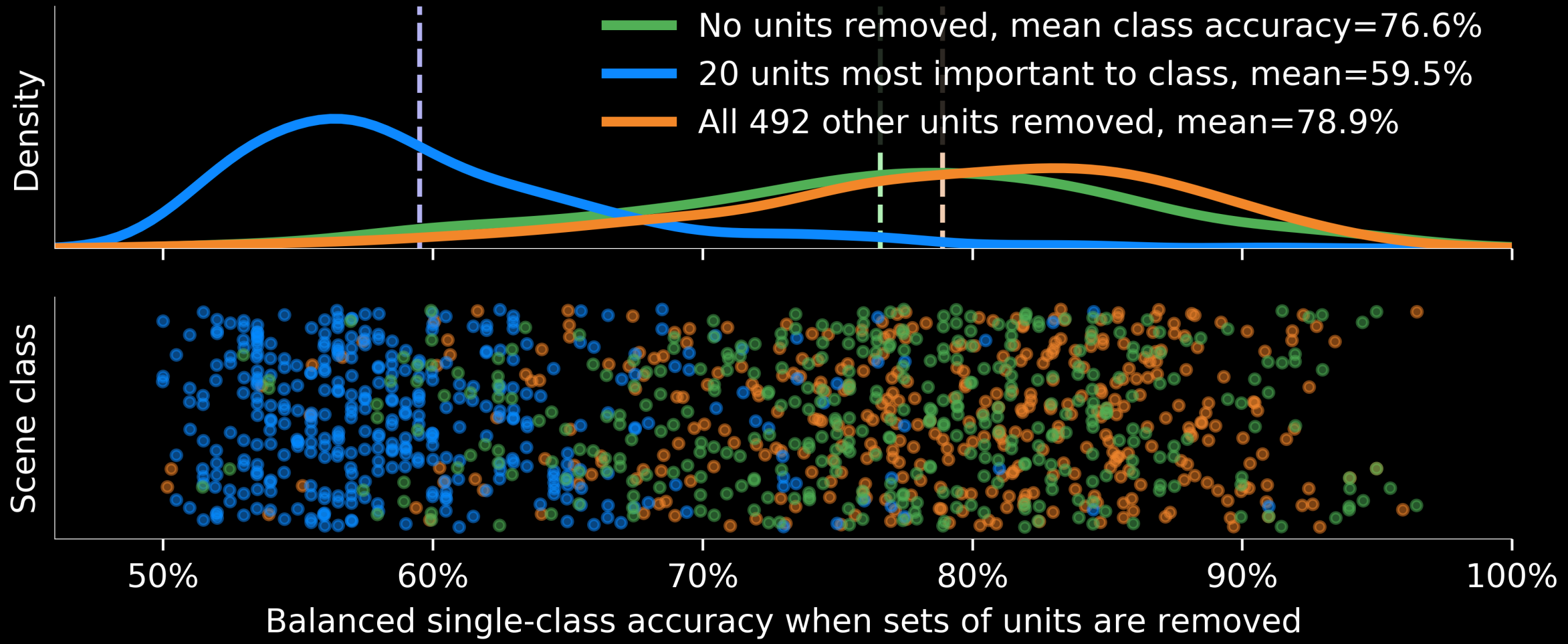
The learned model is sparse



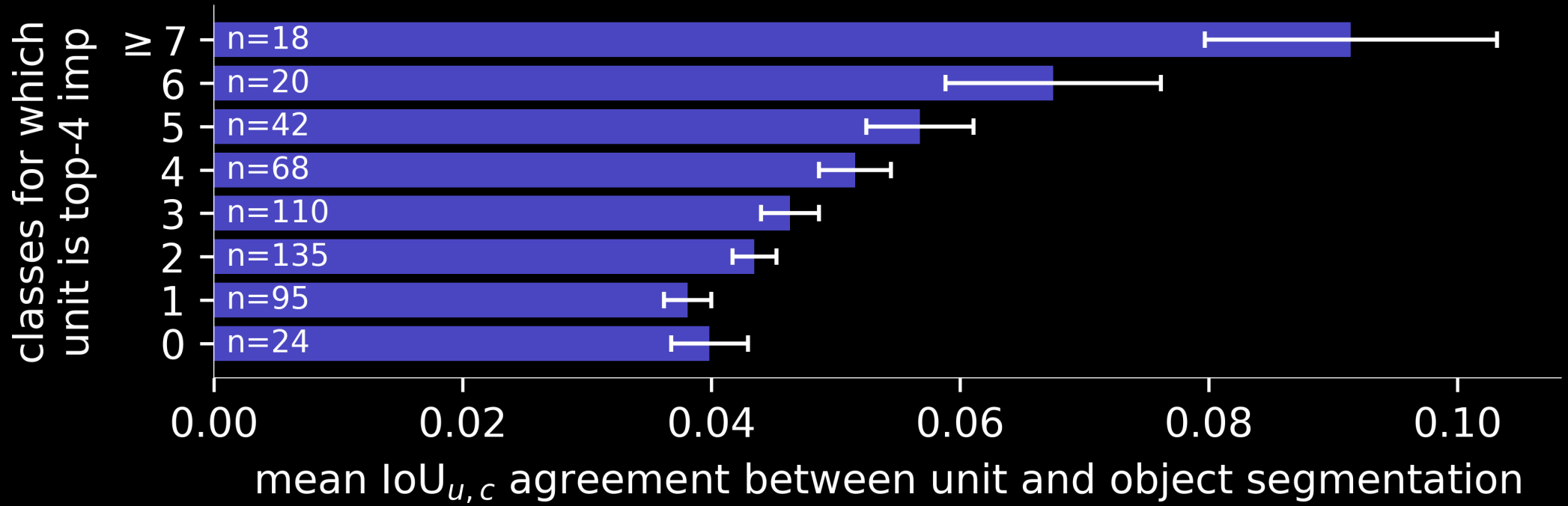
The learned model is sparse



The learned model is sparse



Which units match concepts best?



Unit 184: road (10)

Unit 191: grass (8)

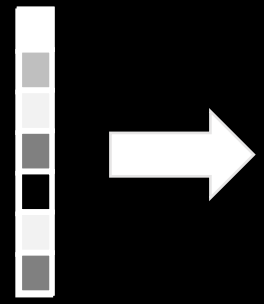
Unit 437: horse (6)

Is Label Supervision Necessary for Concept Units to Emerge?

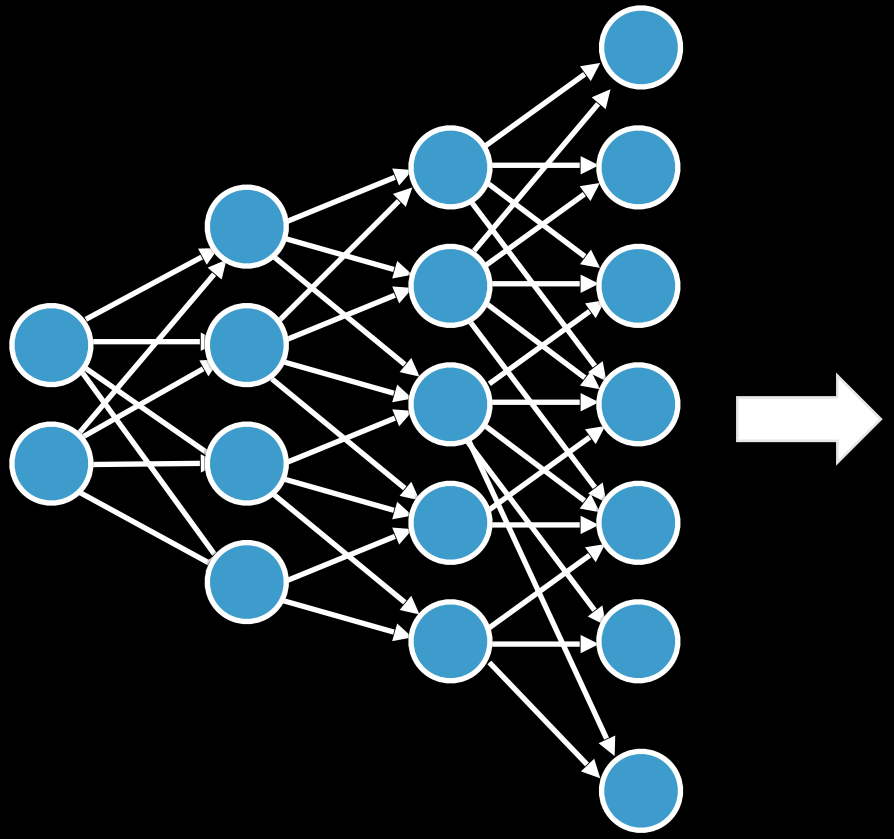
Part 2:
What does a GAN learn?

Generative Adversarial Network (GAN)

Random vector

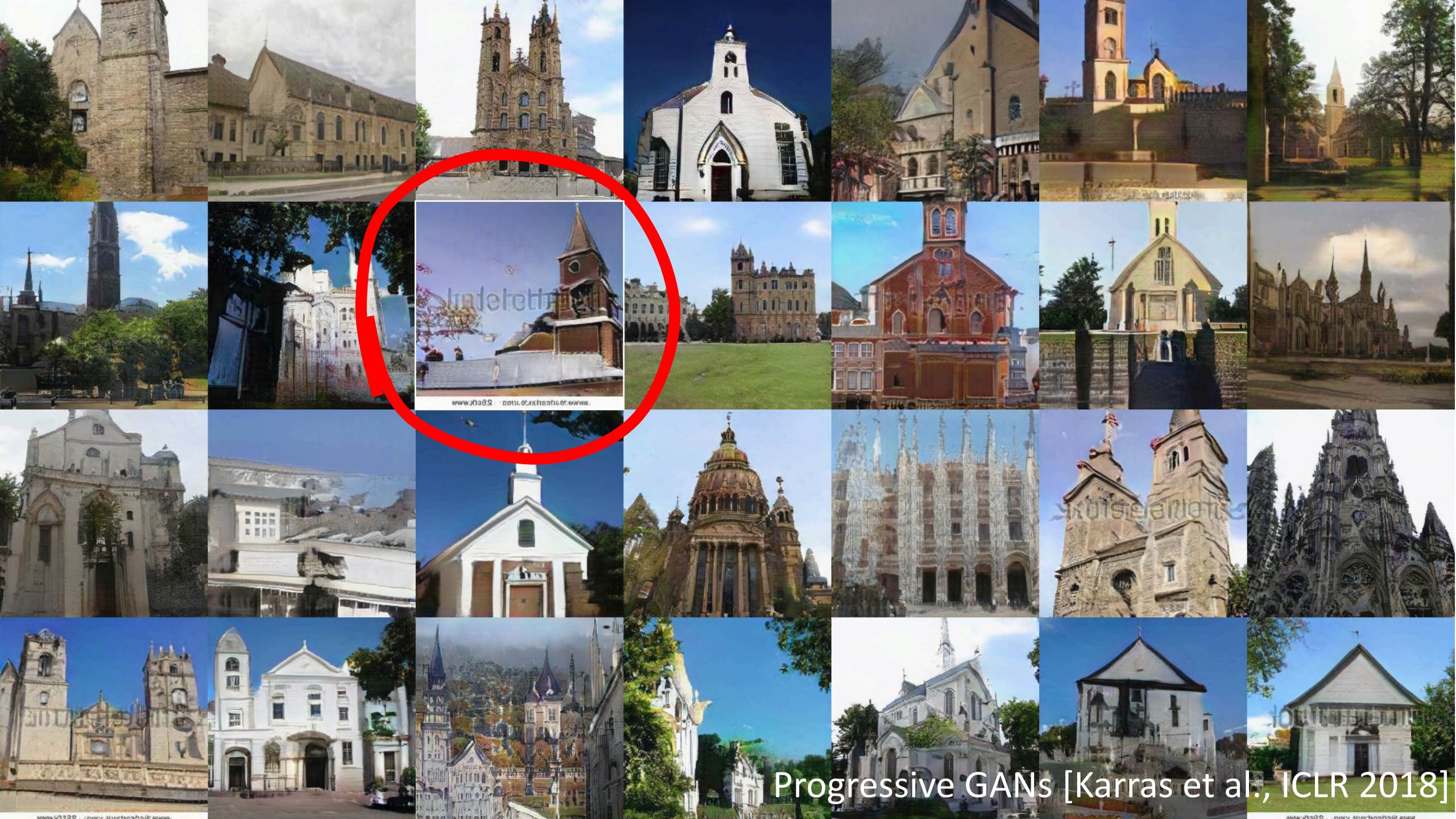


512 dimensions



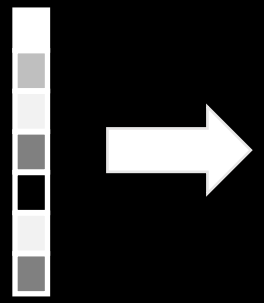
Randomly generated image


Progressive GANs [Karras et al., ICLR 2018]

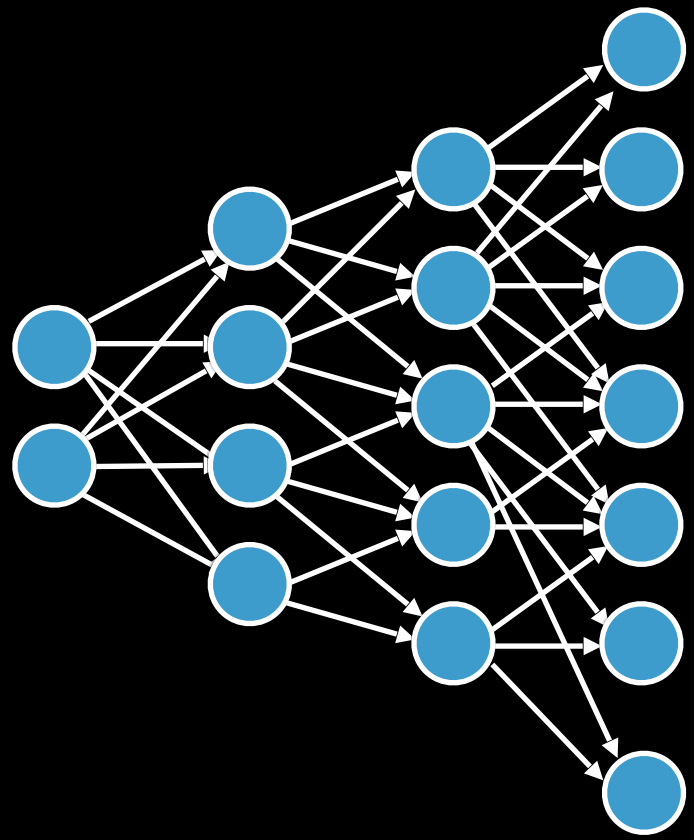

Progressive GANs [Karras et al., ICLR 2018]

Are there watermark neurons?

Random vector



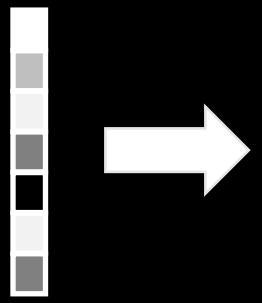
512 dimensions



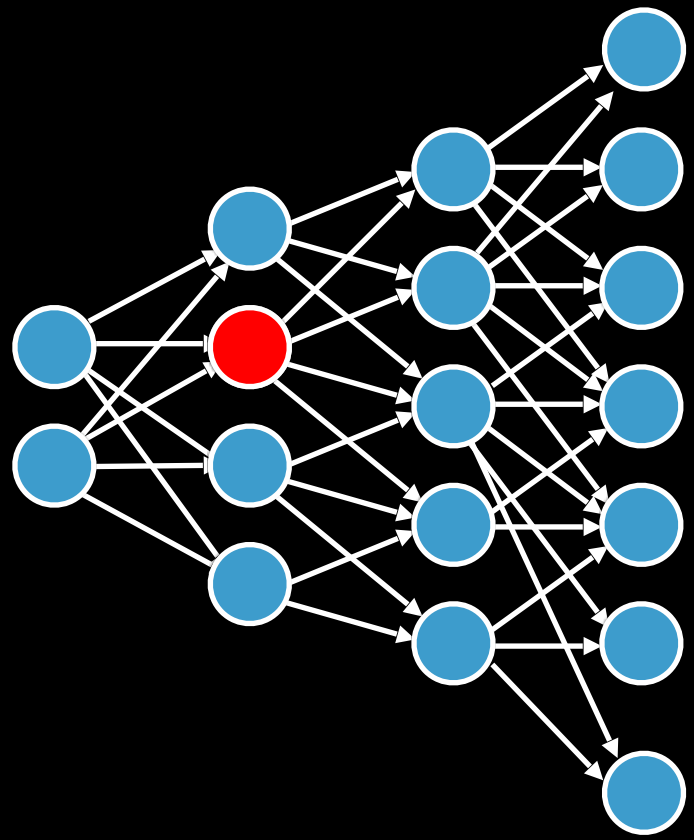
Randomly generated image

Are there watermark neurons?

Random vector



512 dimensions



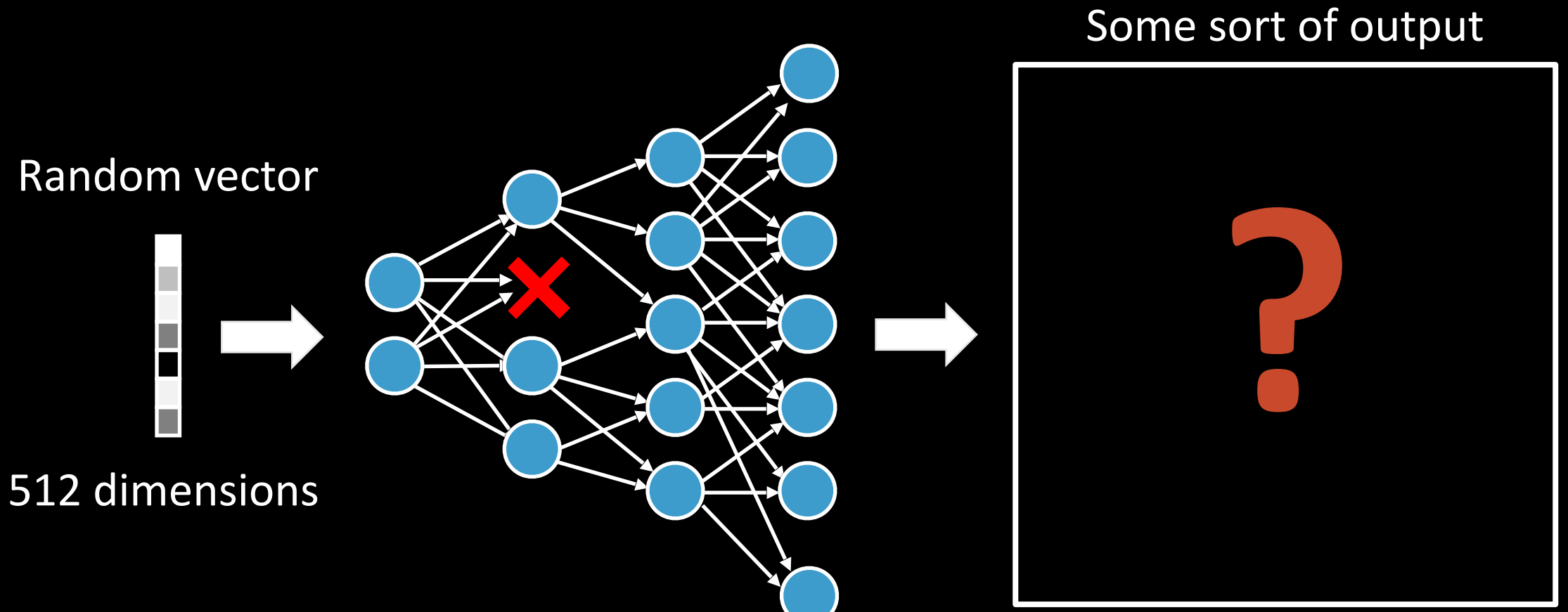
Randomly generated image

Layer 4, Neuron 201

Layer 4, Neuron 445

… and five more neurons

What if we turn off these neurons?



Deactivating banner neurons in layer 4

Deactivating watermak neurons





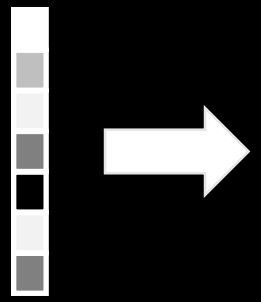



Seeing behind the banner

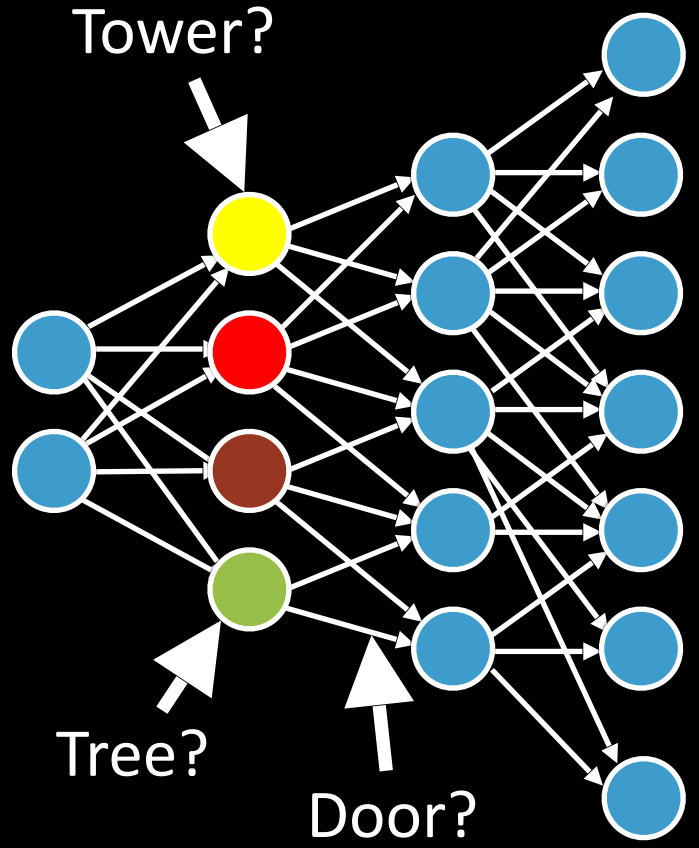
What must the GAN know to erase the banner?

Are there other objects?

Random vector



512 dimensions

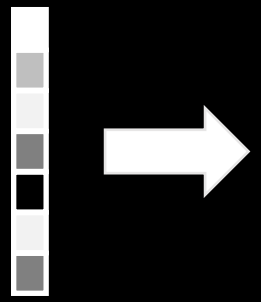


Randomly generated image

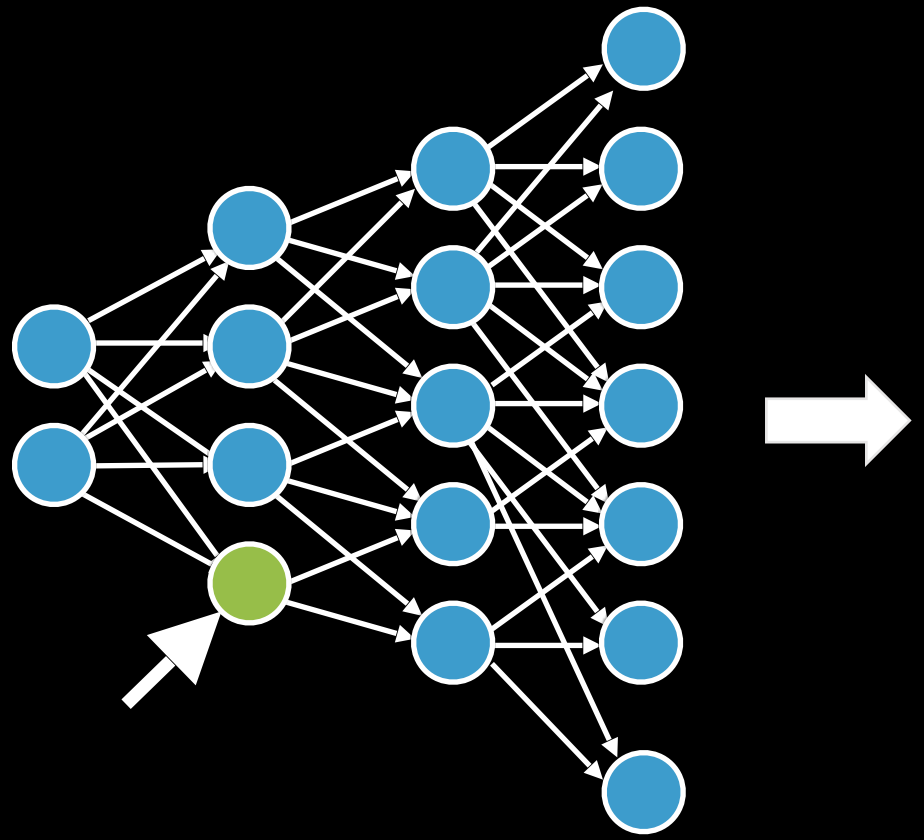
GAN Dissection [Bau et al., ICLR 2019]

Are there other objects?

Random vector



512 dimensions



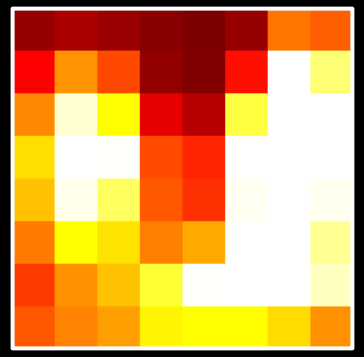
Randomly generated image

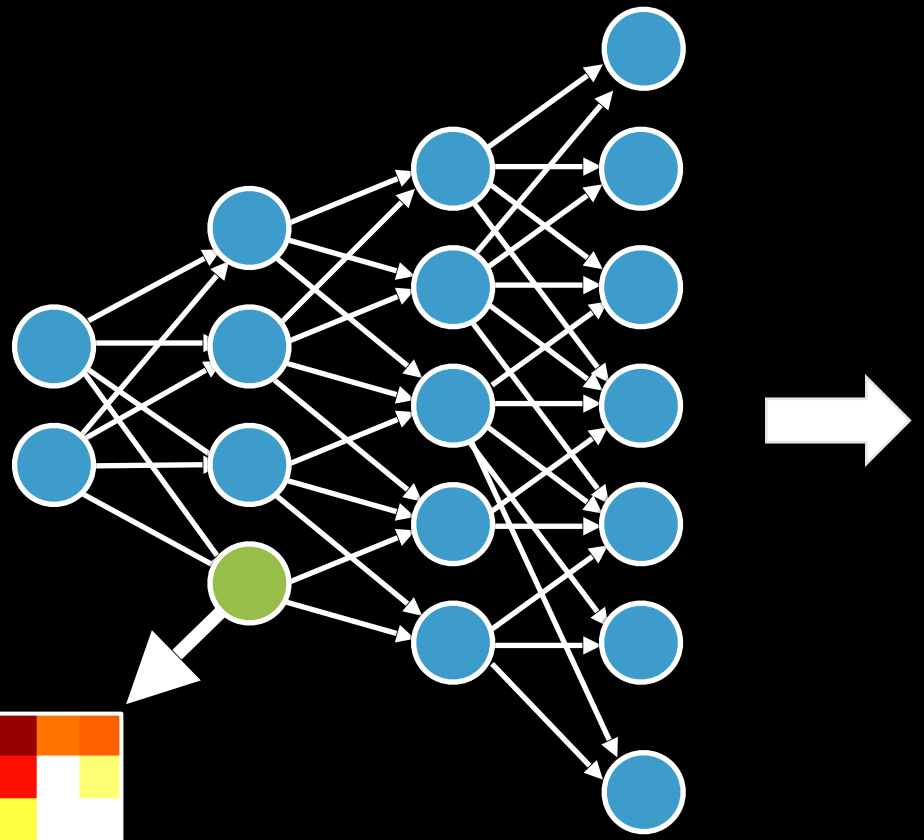
GAN Dissection [Bau et al., ICLR 2019]

Are there other objects?

Random vector

512 dimensions



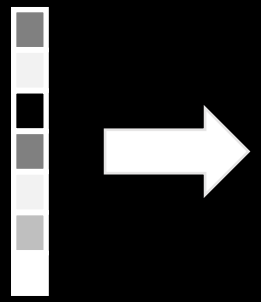


Randomly generated image

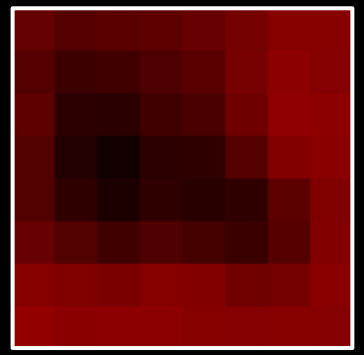
GAN Dissection [Bau et al., ICLR 2019]

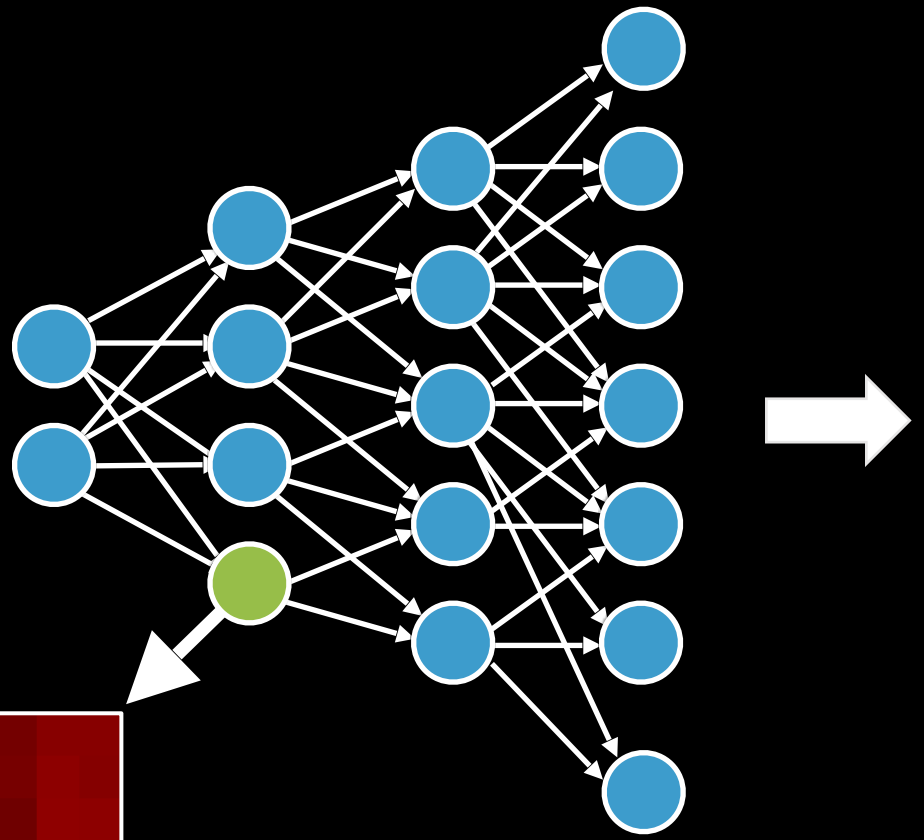
Are there other objects?

Random vector



512 dimensions



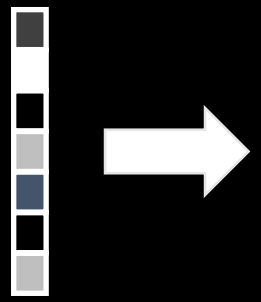


Randomly generated image

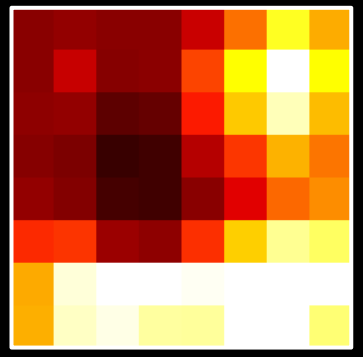
GAN Dissection [Bau et al., ICLR 2019]

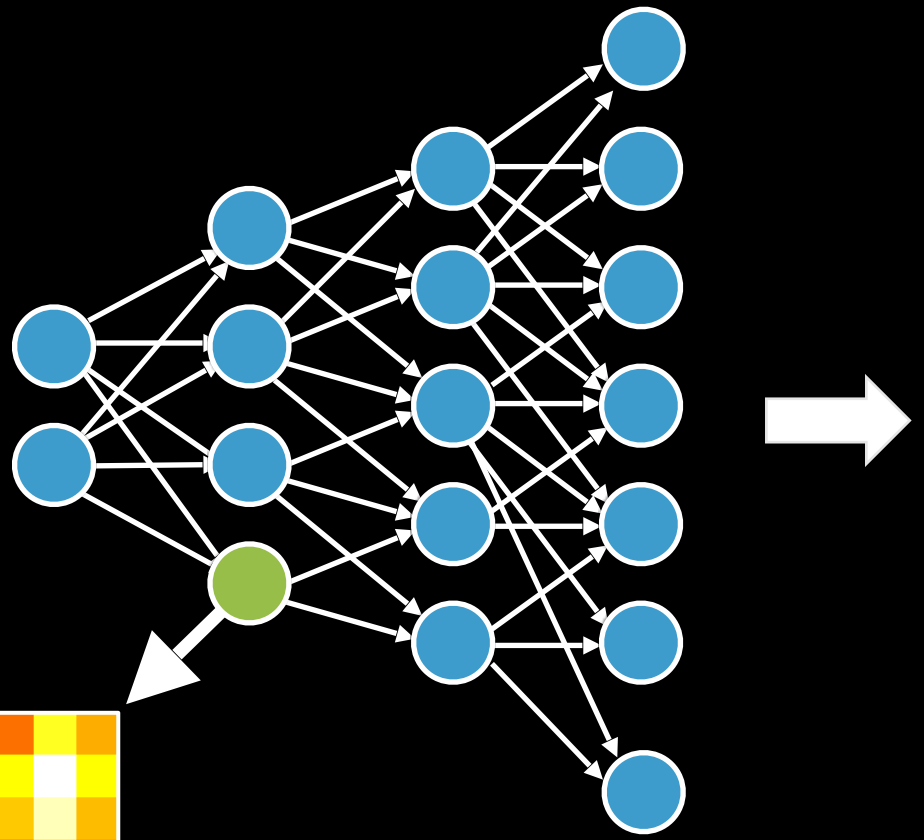
Are there other objects?

Random vector



512 dimensions



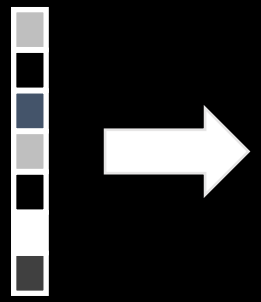


Randomly generated image

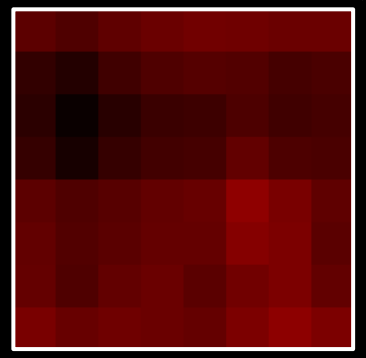
GAN Dissection [Bau et al., ICLR 2019]

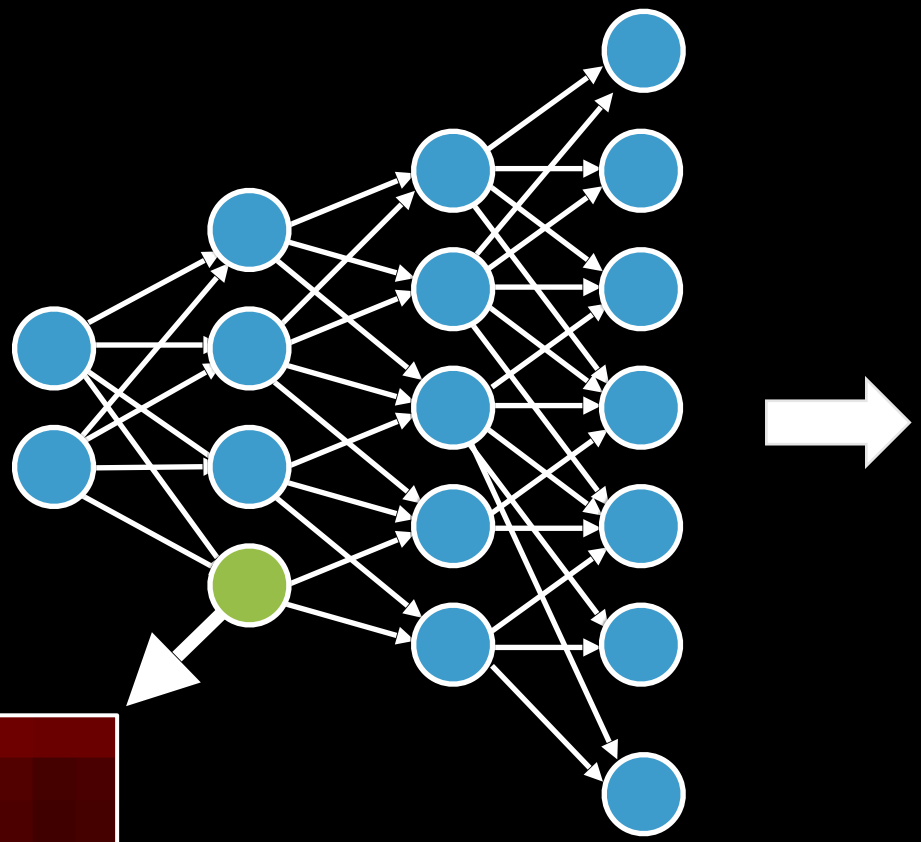
Are there other objects?

Random vector



512 dimensions



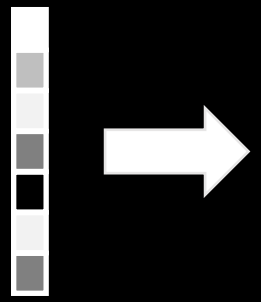


Randomly generated image

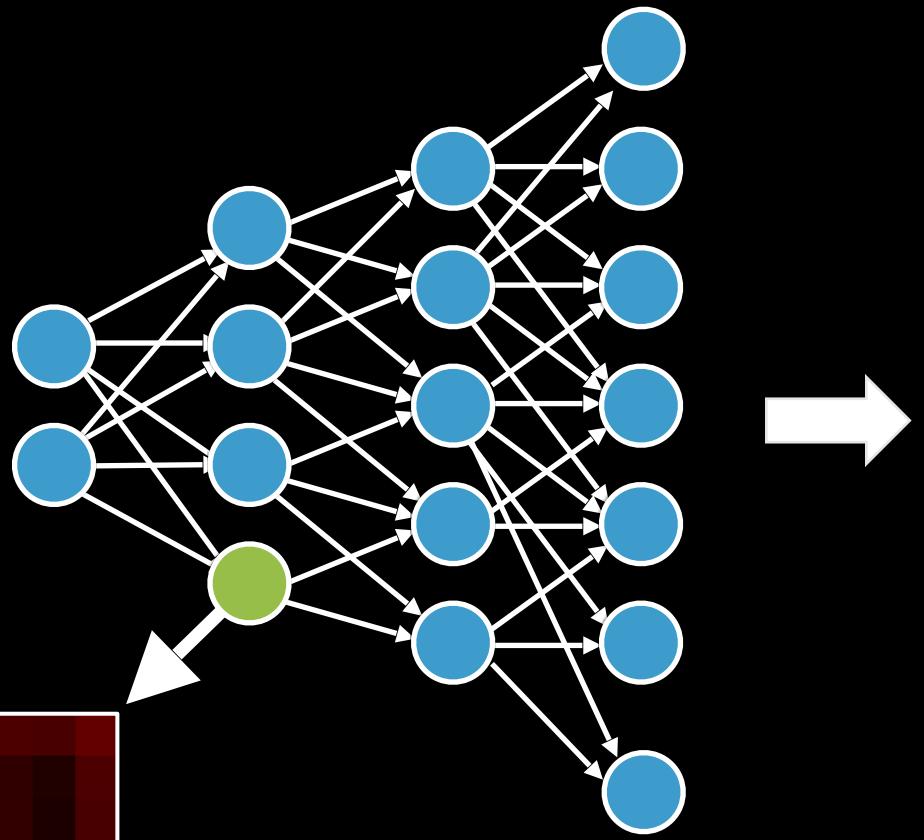
GAN Dissection [Bau et al., ICLR 2019]

Are there other objects?

Random vector



512 dimensions

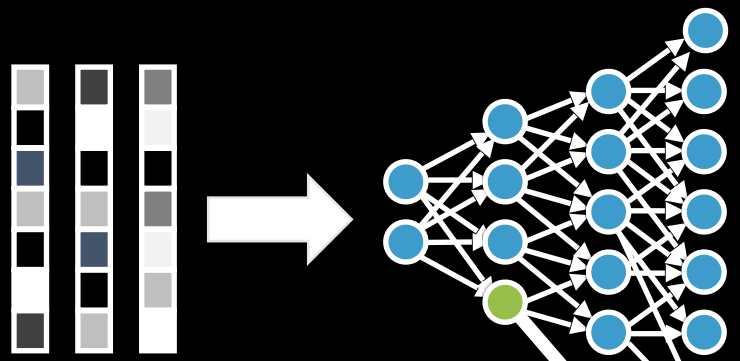


Randomly generated image

GAN Dissection [Bau et al., ICLR 2019]

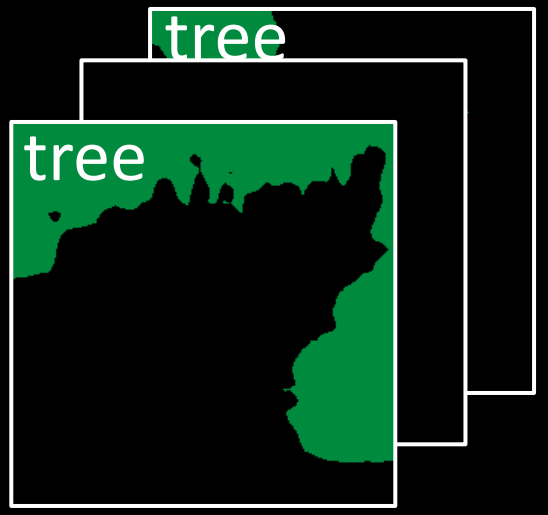
Dissecting a GAN

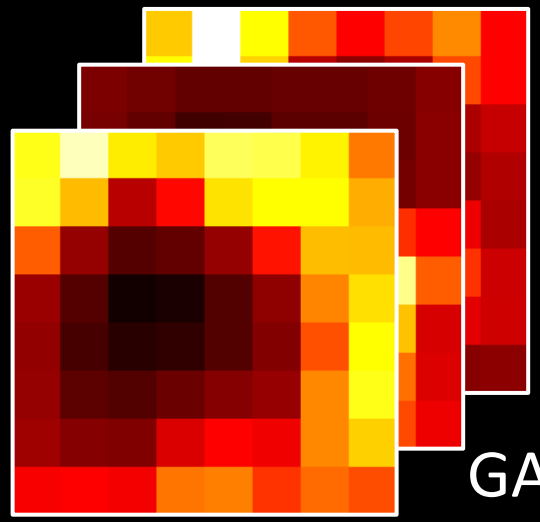
Random vectors



Generate lots of images

Semantic segmentation





Agreement

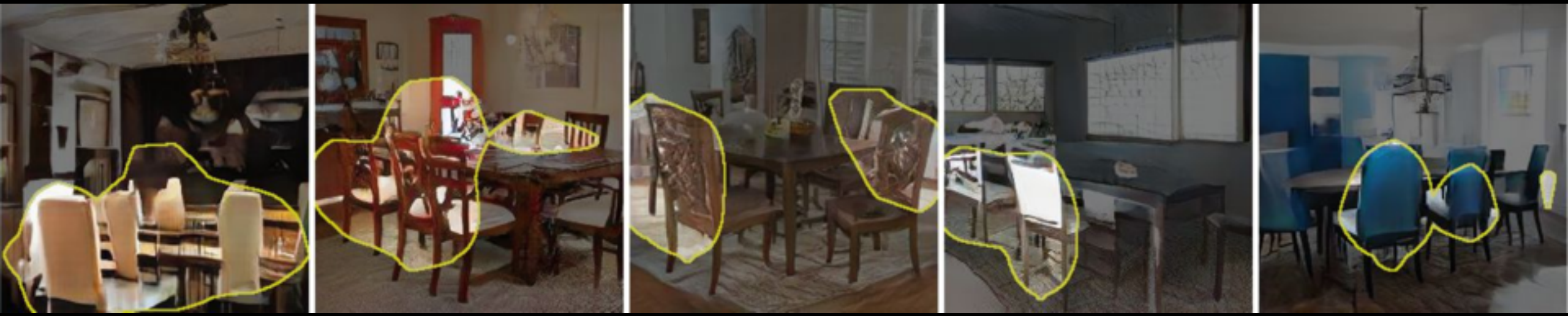
GAN Dissection [Bau et al., ICLR 2019]

Layer 4, Neuron 119: tree

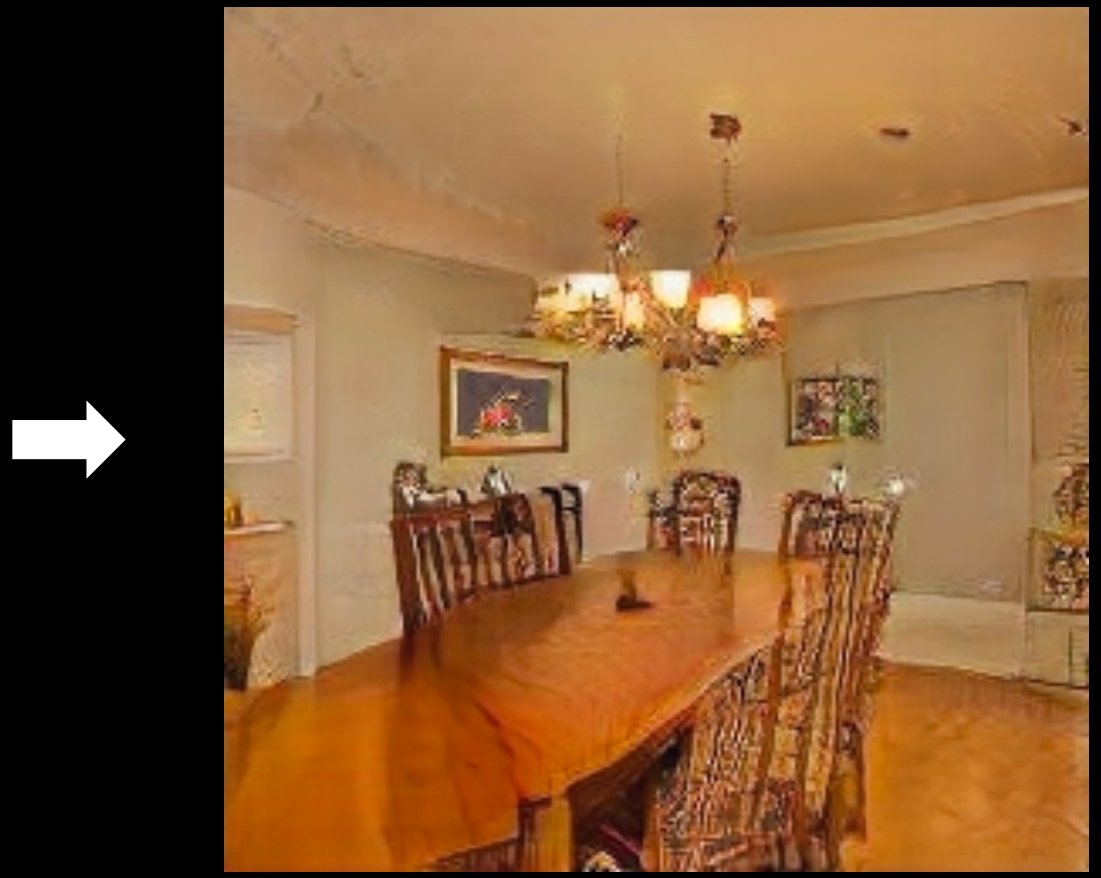
Layer 4, Neuron 43 : dome

Layer 4, Neuron 84: window

Layer 4, Neuron 315: chair



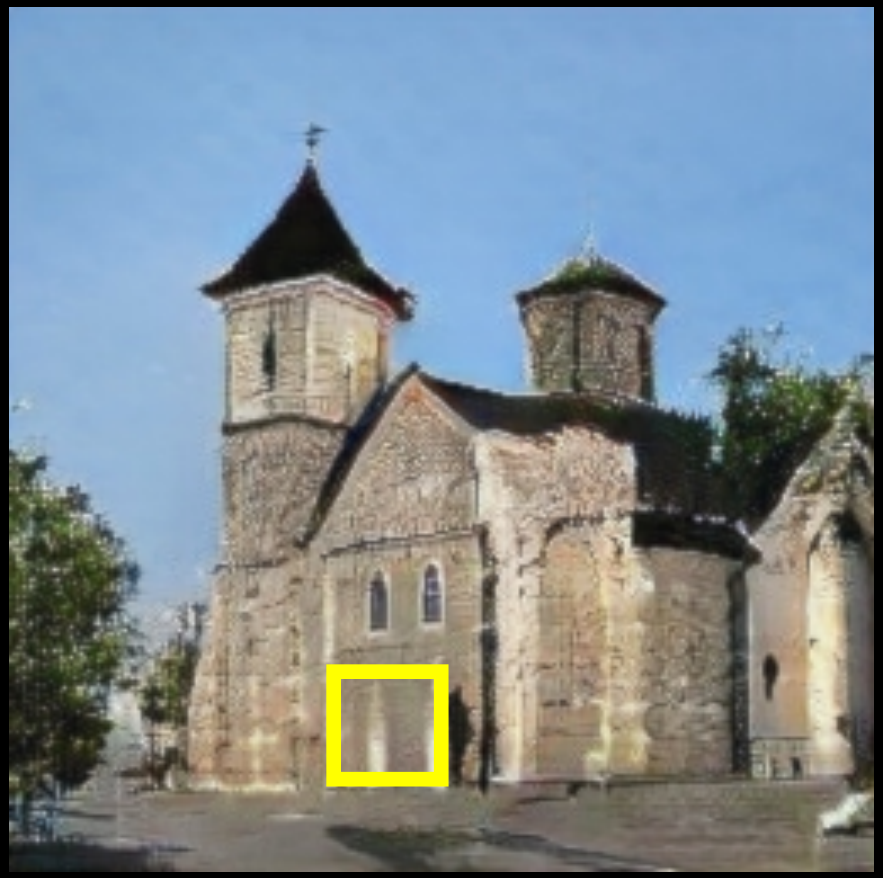
Turning off window neurons



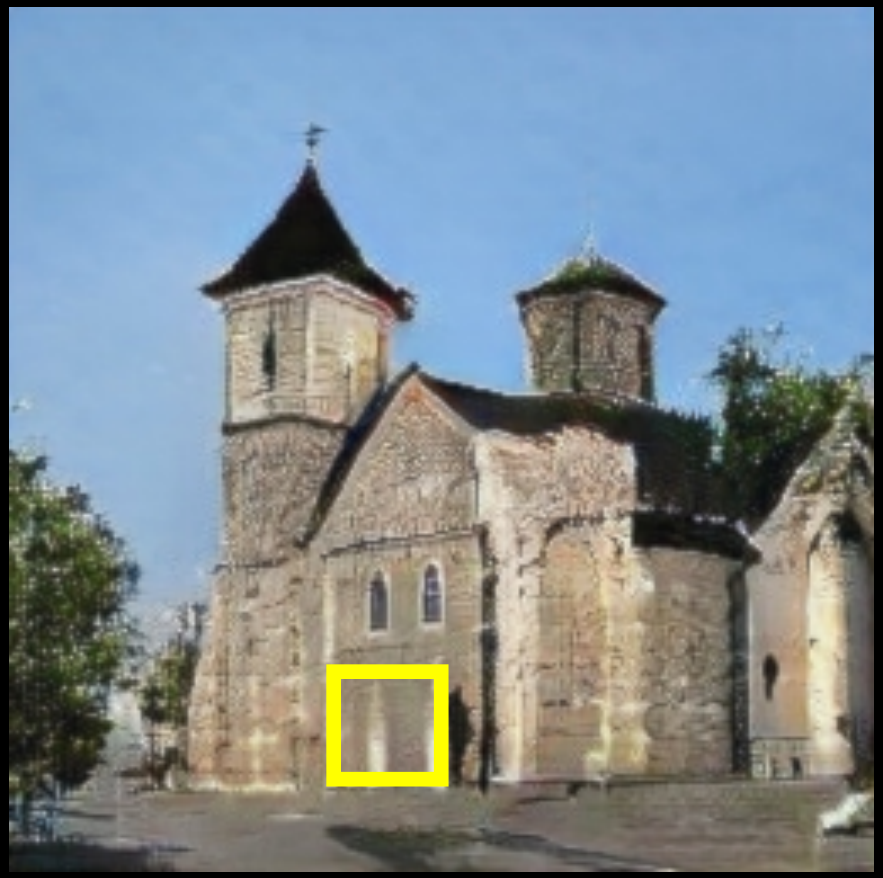
Turning off tree neurons

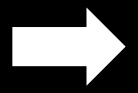
Occluded buildings are now visible

Turning on door neurons

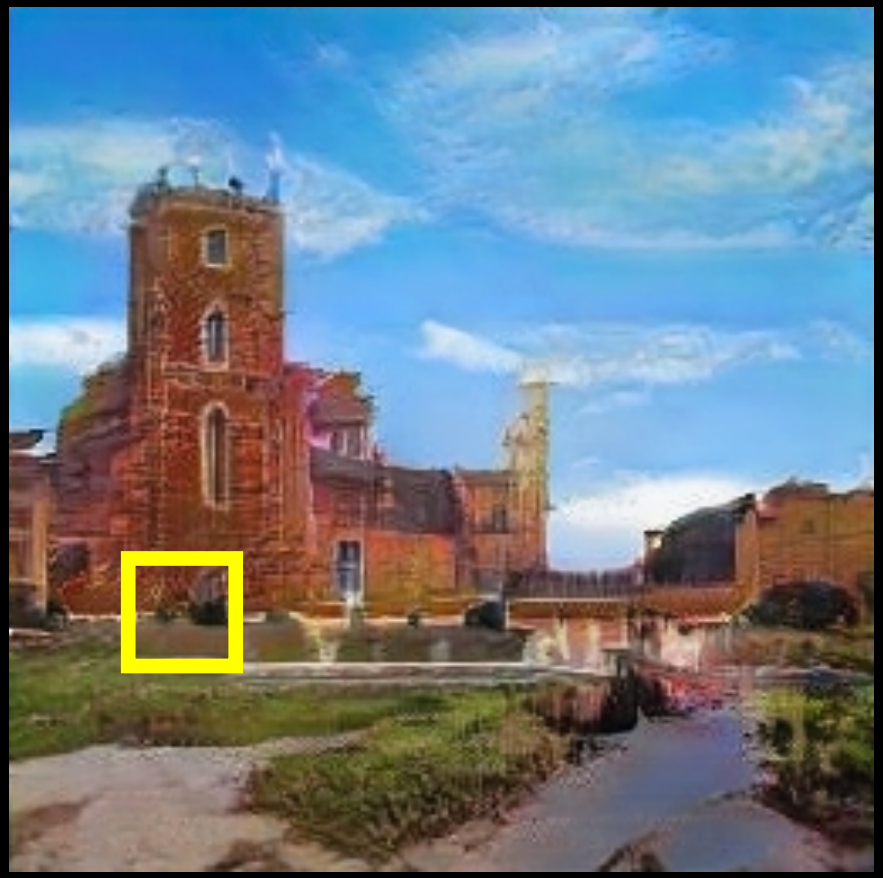


Turning on door neurons

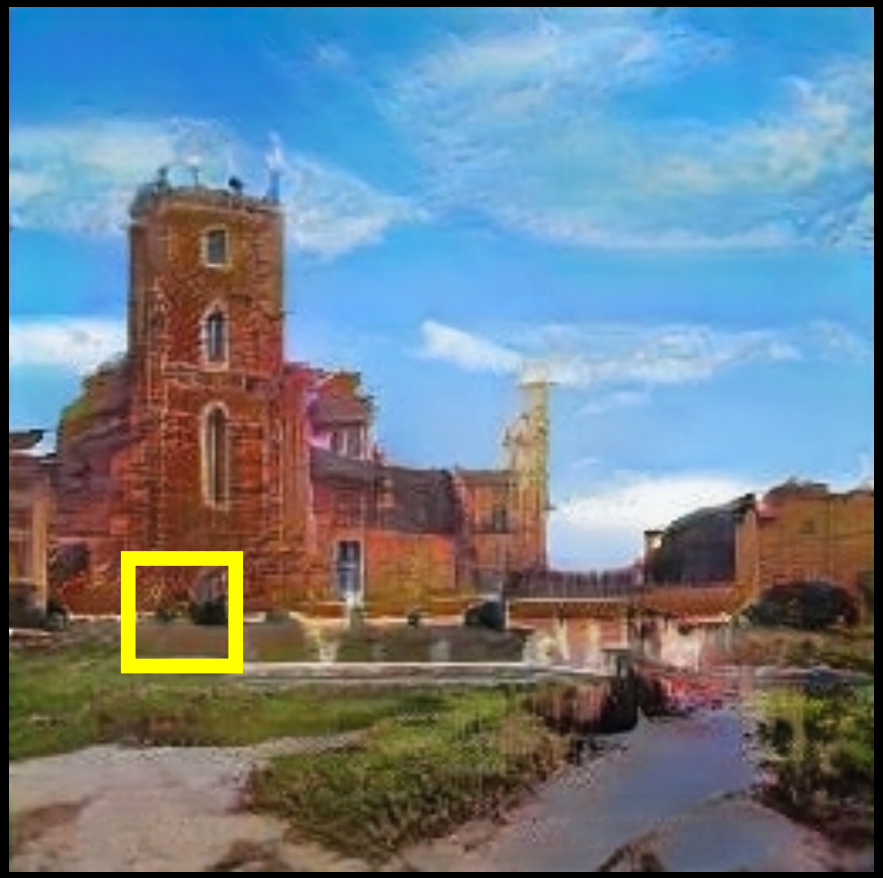


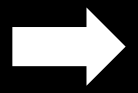


Turning on door neurons



Turning on door neurons

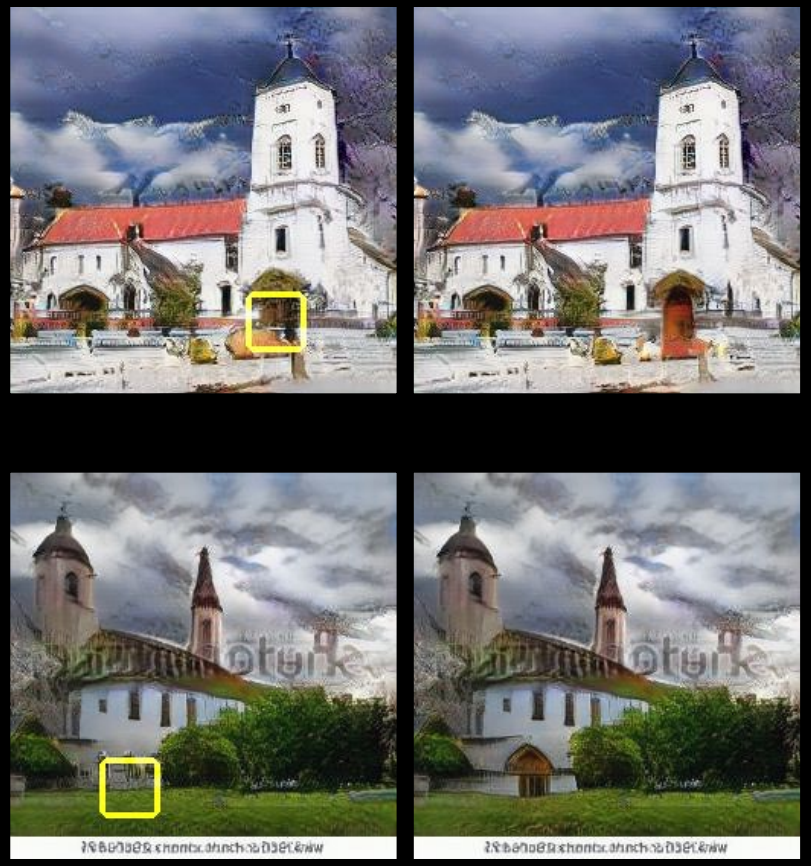


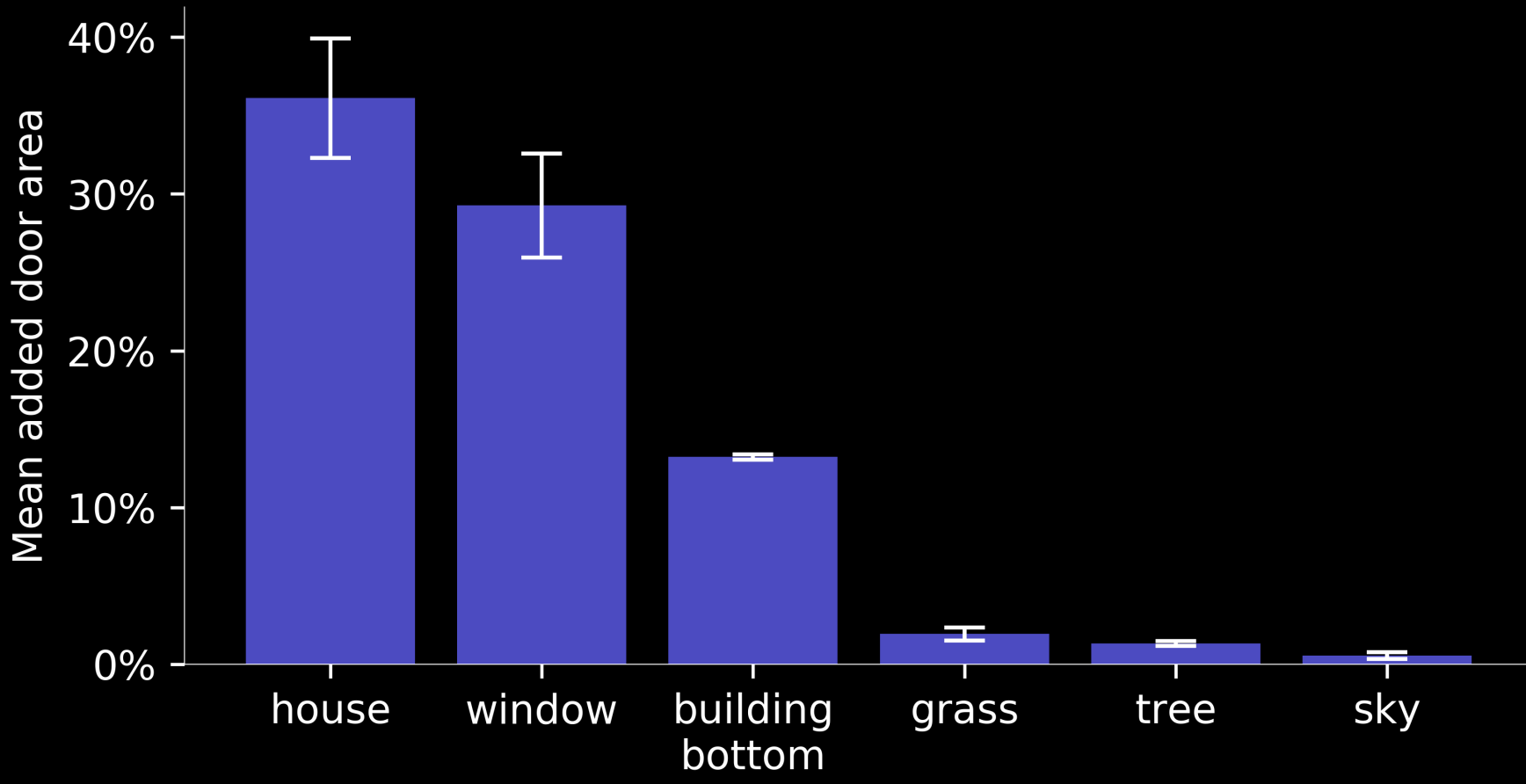


Turning on door neurons

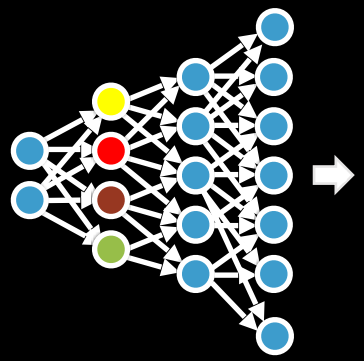
Turning on door neurons

Where can a door go?





Randomly generated image



Add grass

Add trees

Add brick

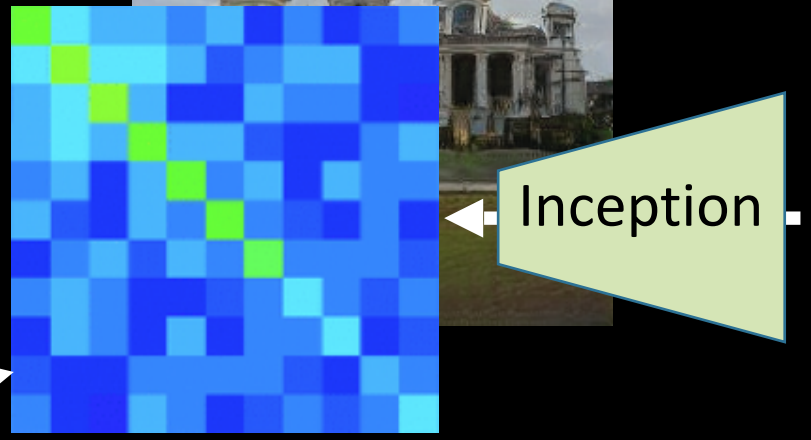
Change doors

Change roof

GAN Dissection [Bau et al., ICLR 2019]

Interlude:
What a GAN Cannot Generate

Fake Inception Statistics

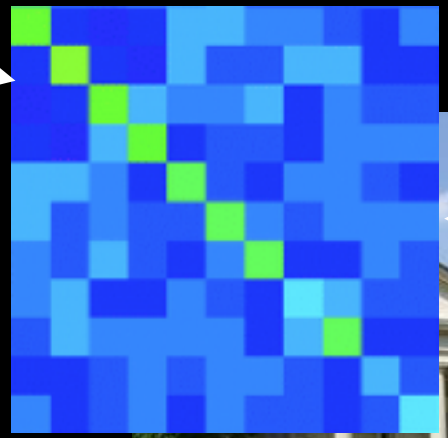


Inception

Fake Images

Measuring GAN Quality

FID = 18

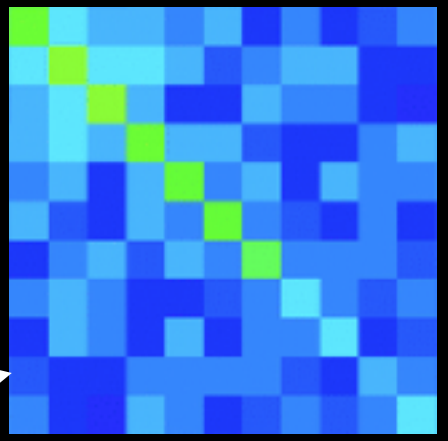


Inception

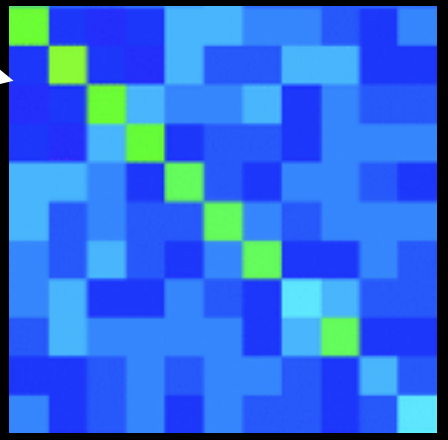
Real Inception Statistics

Real Images

Fake Inception Statistics



FID = 18



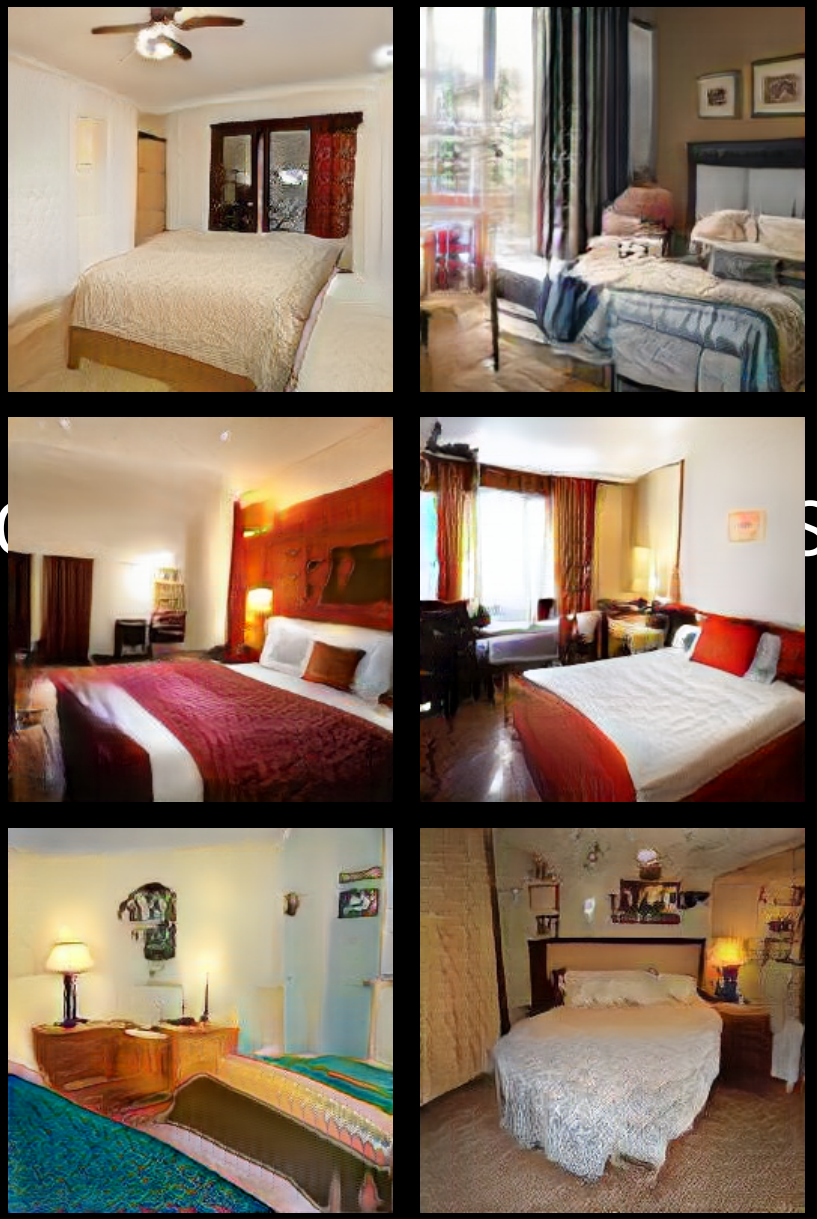
Real Inception Statistics

Two Questions abo s

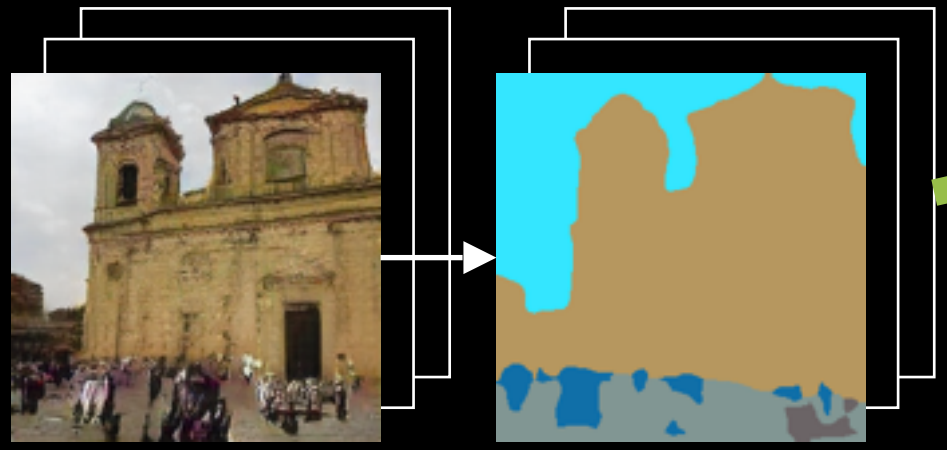
1. What is actually missing in the distribution?

2. What is actually missing in each image?

Generated Image Sample →

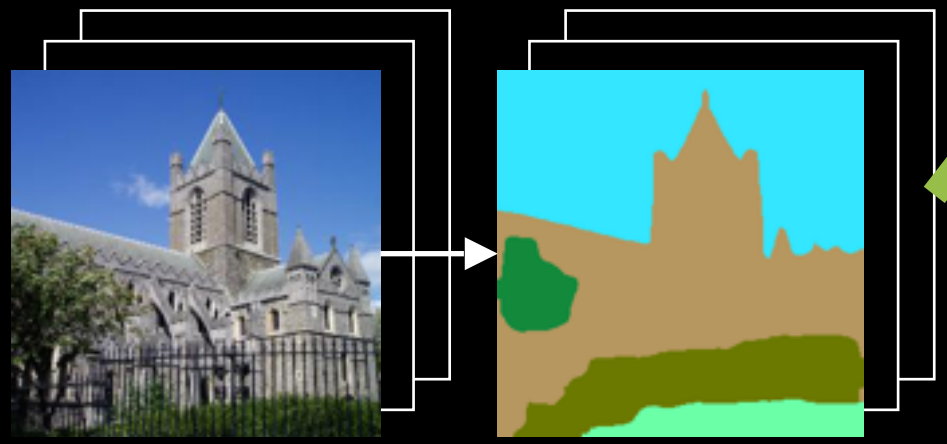


1. Understanding omissions in the distribution



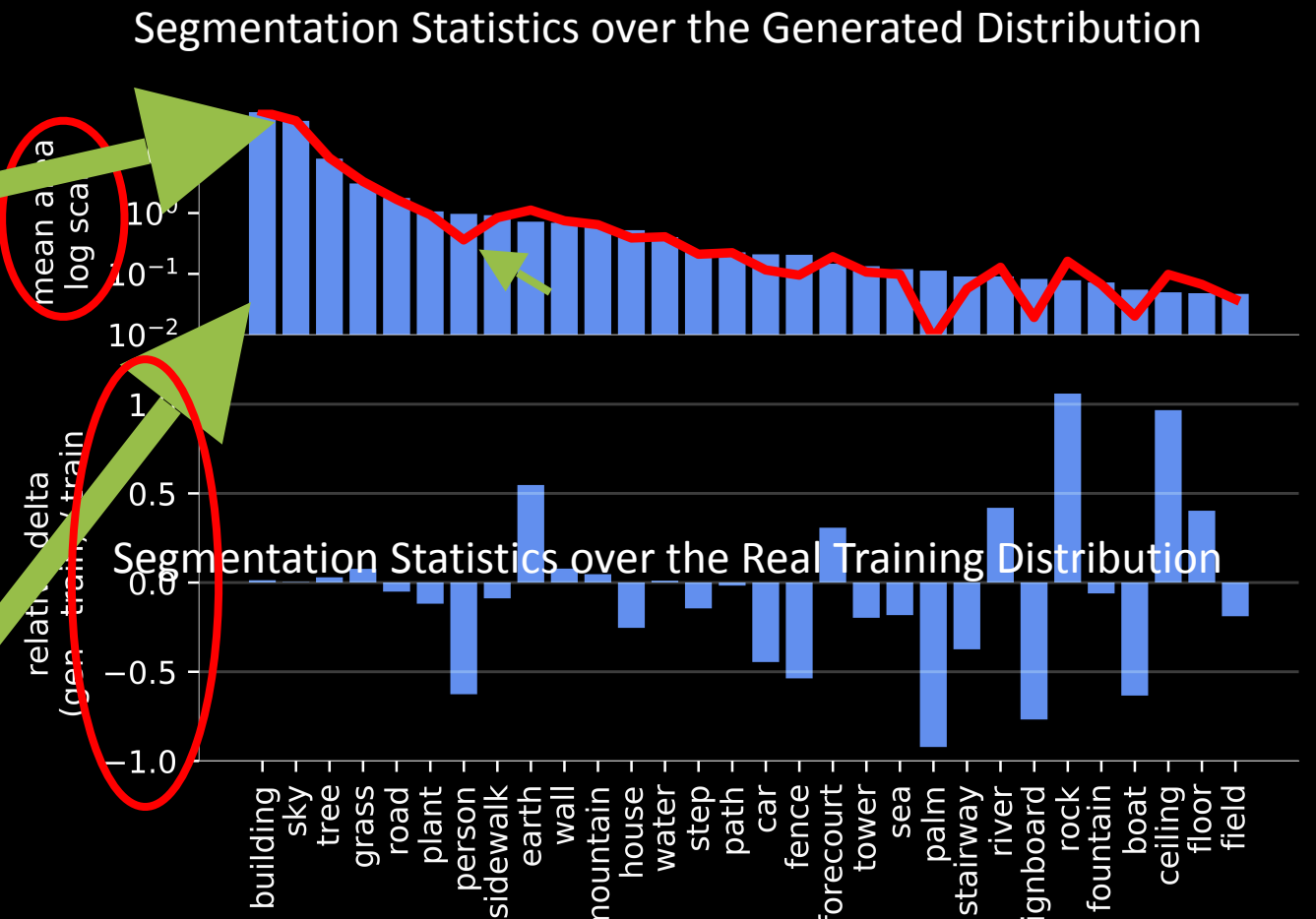
Generated image

Semantic segmentation



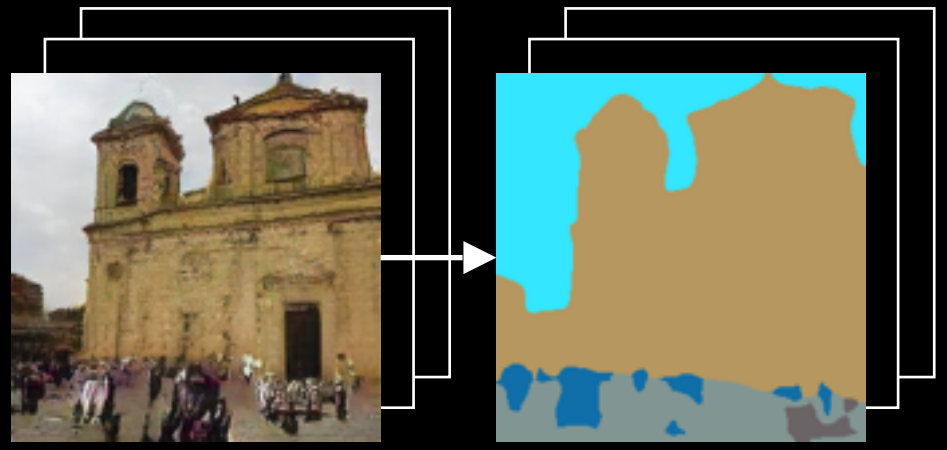
Real image

Semantic segmentation

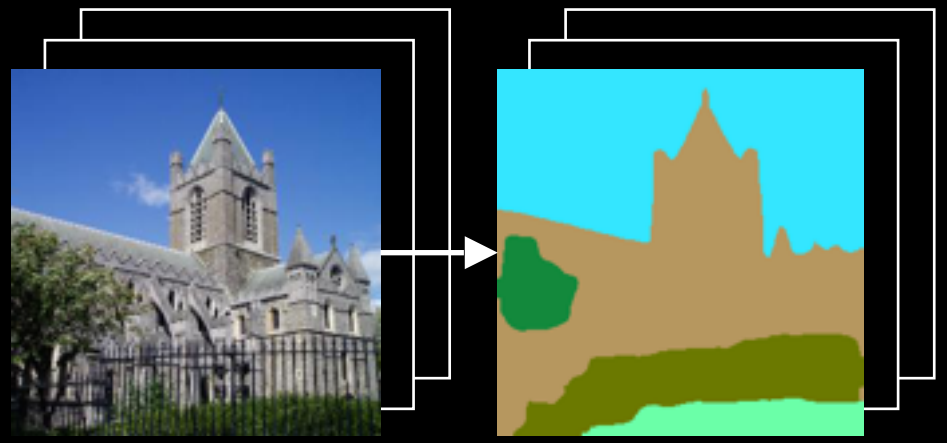


Differences in Segmentation Statistics Reveal Omissions

The distribution is missing objects

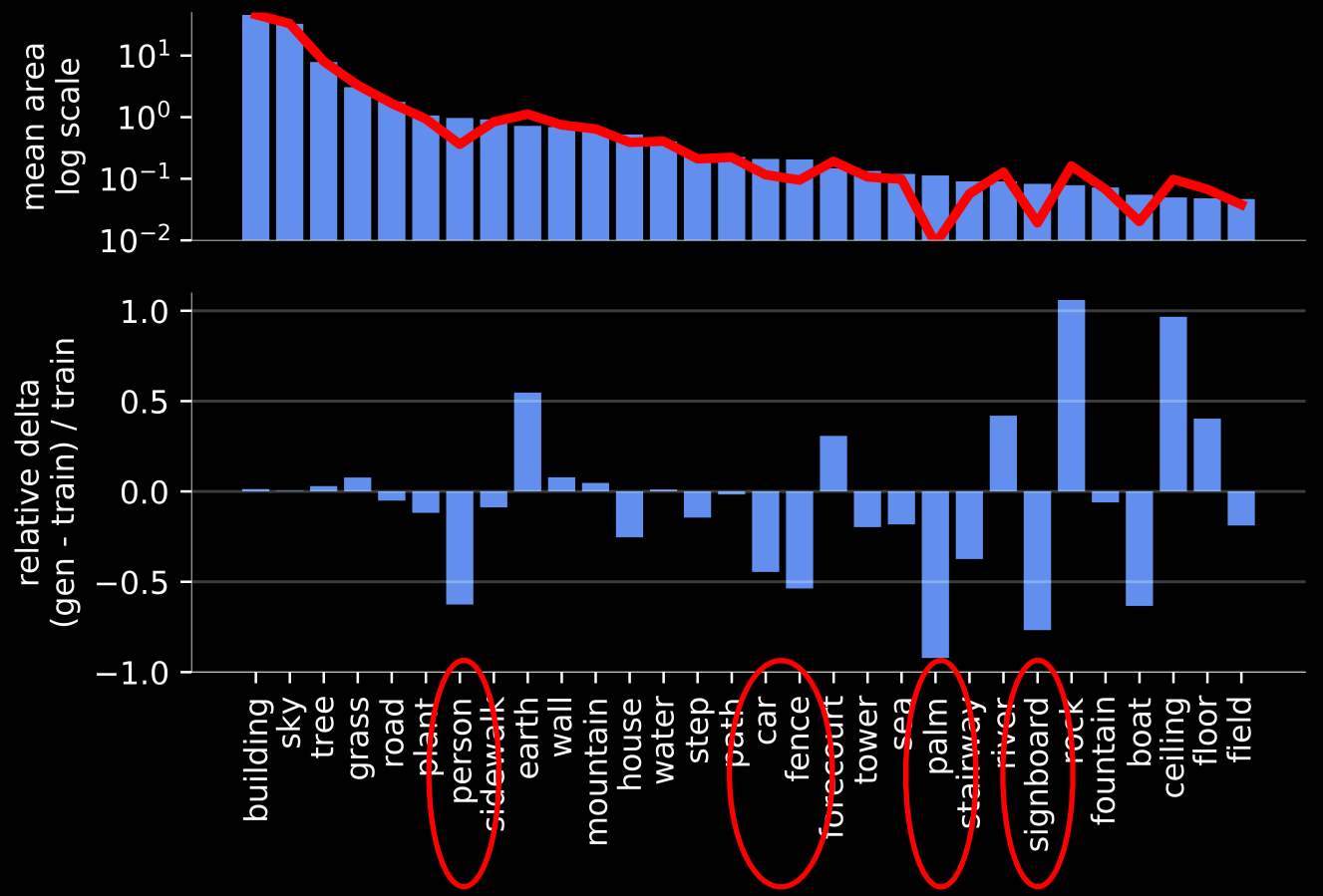


Generated image Semantic segmentation



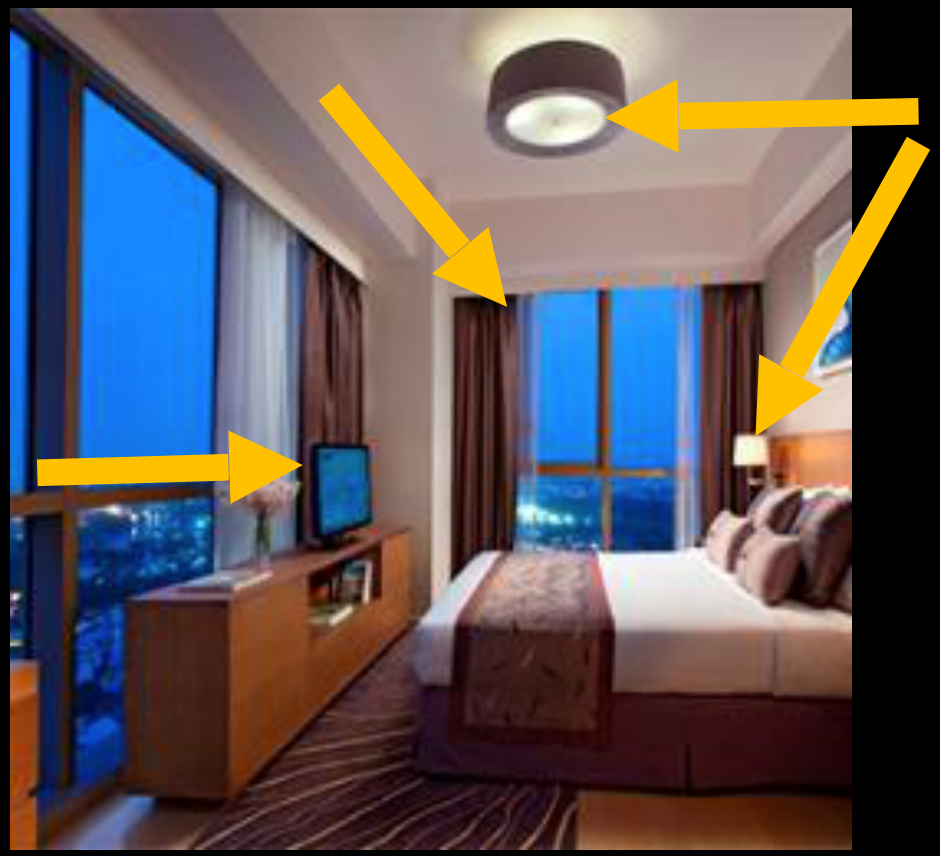
Real image Semantic segmentation

Segmentation Statistics over the Generated Distribution



Differences in Segmentation Statistics Reveal Omissions

2. Understanding omissions in individual images



real image x

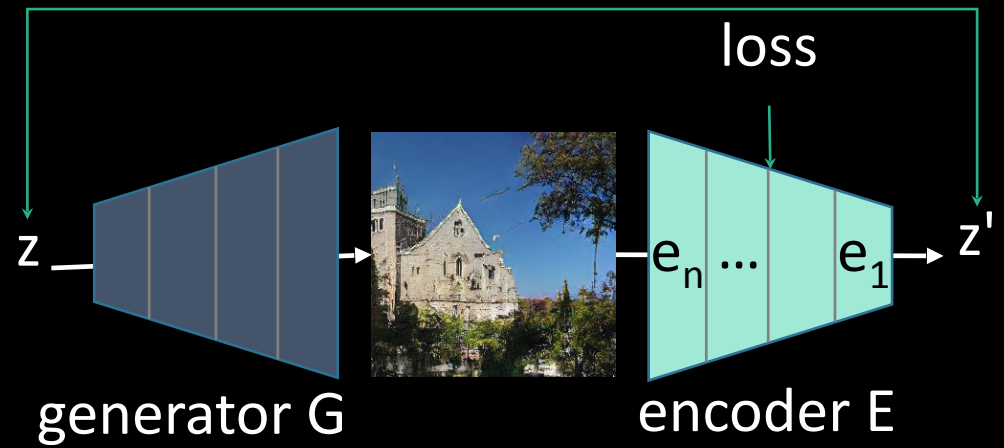
synthesized G(z)

Pairs (x, G(z*)) reveal omissions

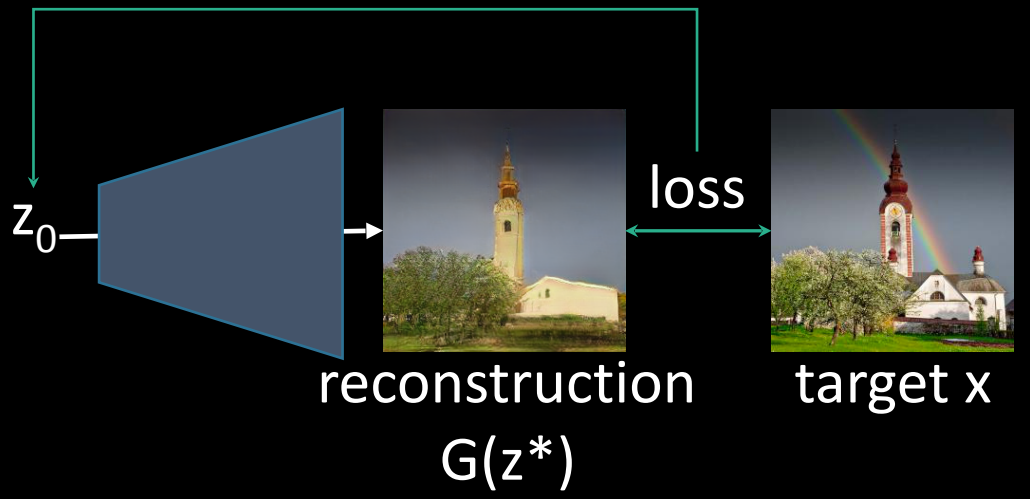
Objective: $z^* = \text{argmin}_z \mathcal{L}(x, G(z))$

Two steps to invert a large generator

1: Train encoder by layer



2: Optimize latents by layer



Seeing What a GAN Cannot Generate [Bau, et al. ICCV 2019]

Inversion converges precisely when x = G(z)

Generated

x = G(z)

Reconstruction

x = G(z*)

When G generates x,
reconstruction is precise

Real photo

x real

Reconstruction

x ≠ G(z*)

When reconstruction is imperfect
we know G cannot generate x

Seeing What a GAN Cannot Generate [Bau, et al. ICCV 2019]

Progressive GAN trained on churches

Original photo

Reconstruction

Wedding dresses are reconstructed

Original photo

Reconstruction

Vehicles

Real

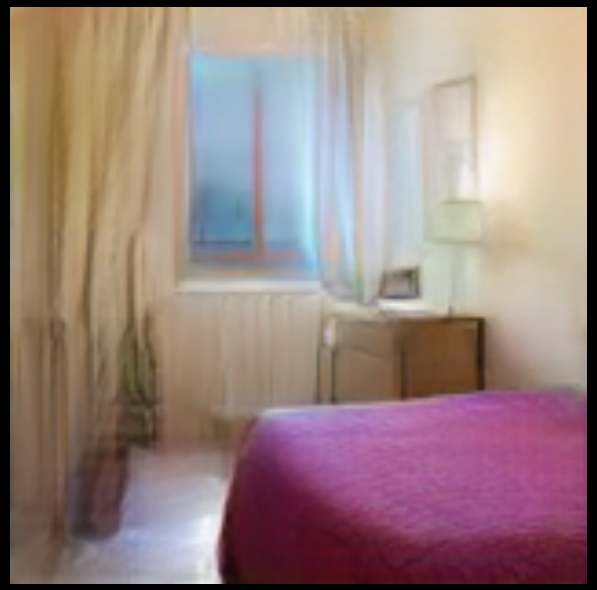
Reconstruction

Bedrooms

Real




Reconstruction




Putting it together:
Using a GAN as a Paintbrush

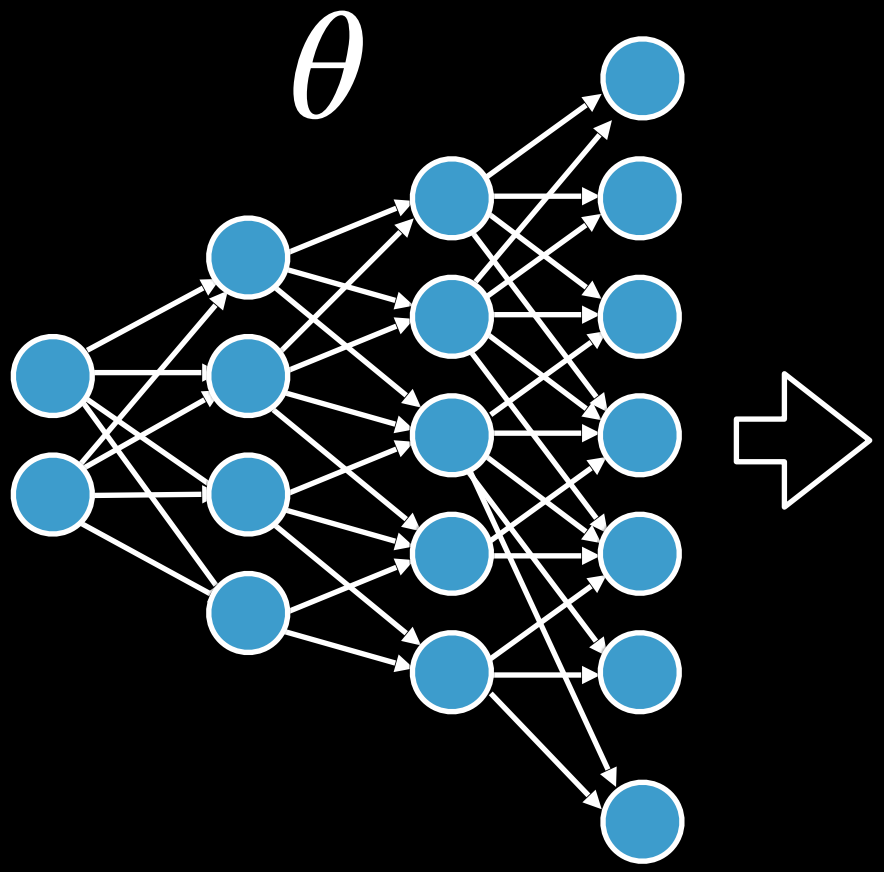
How to edit my own photo?

GAN-Synthesized Kitchen

My Kitchen Photo

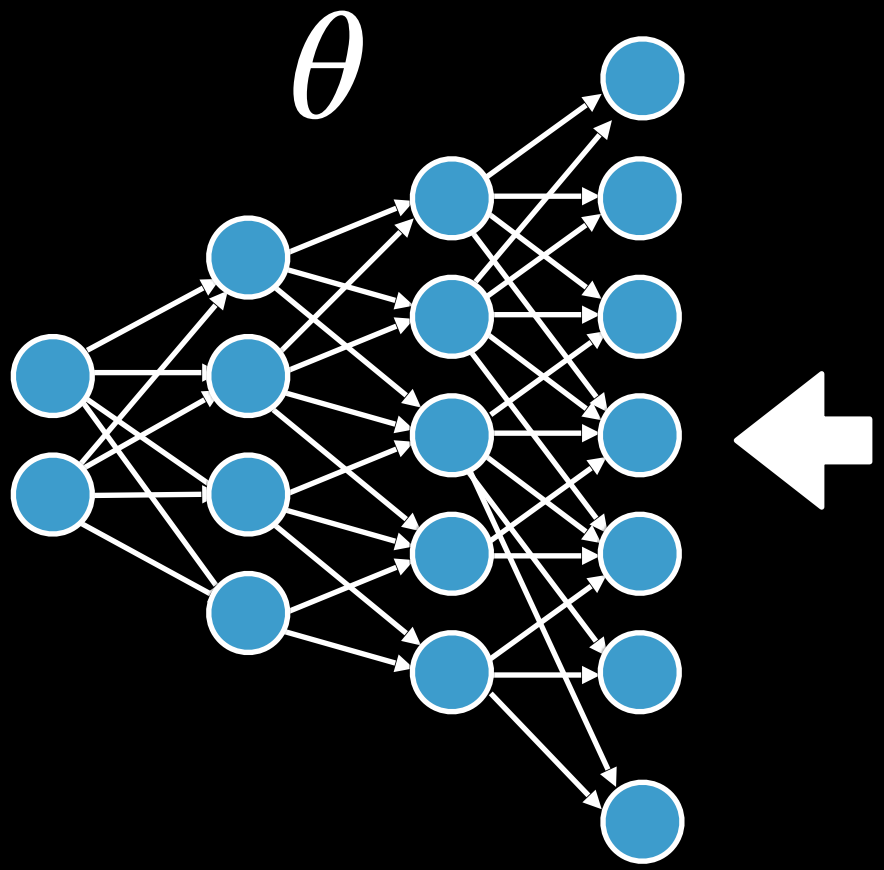
How to edit my own photo?

Random vector z?



My Kitchen Photo

How to edit my own photo?



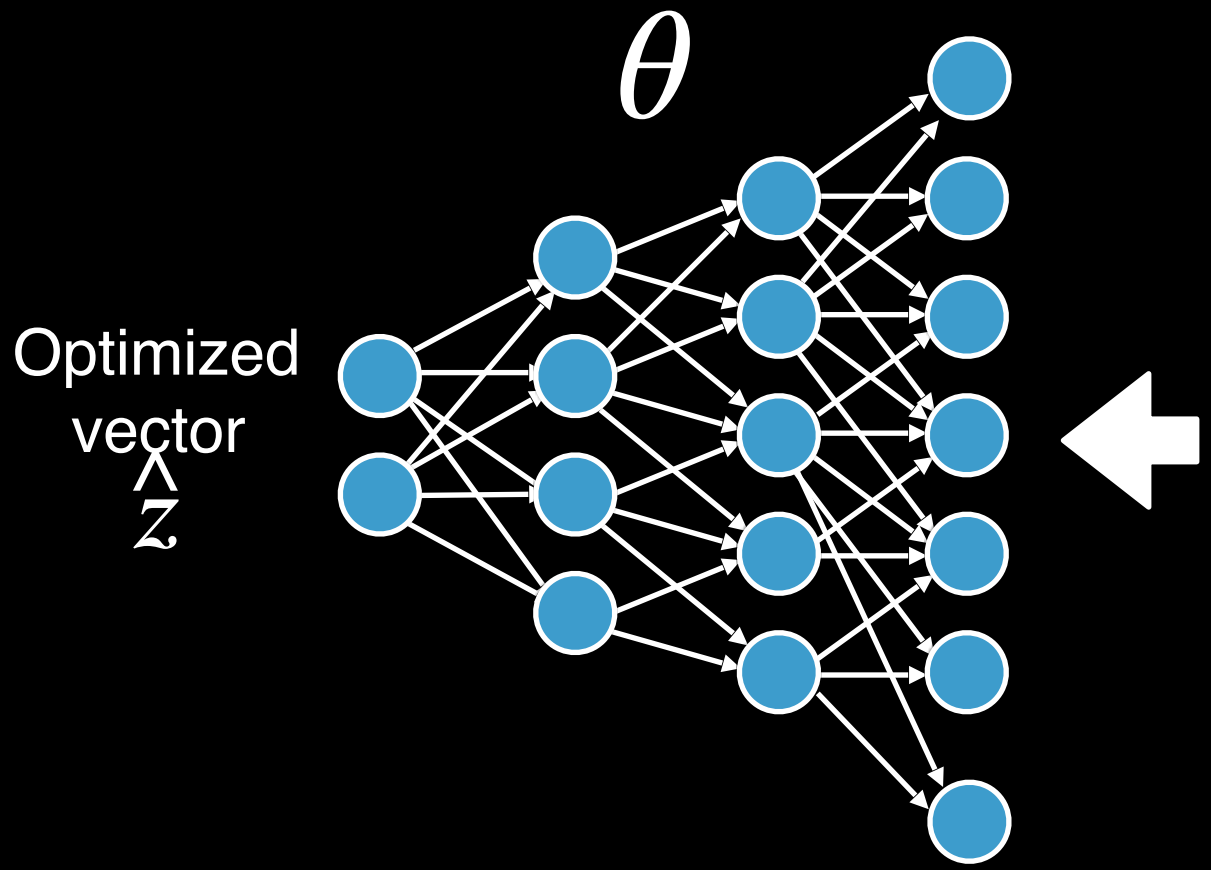
My image

$$\hat{z} = \underset{z}{\operatorname{argmin}} L_{rec}(I, G(z, \theta))$$

[Zhu et al., 2016]

[Dosovitskiy and Brox., 2016]

How to edit my own photo?

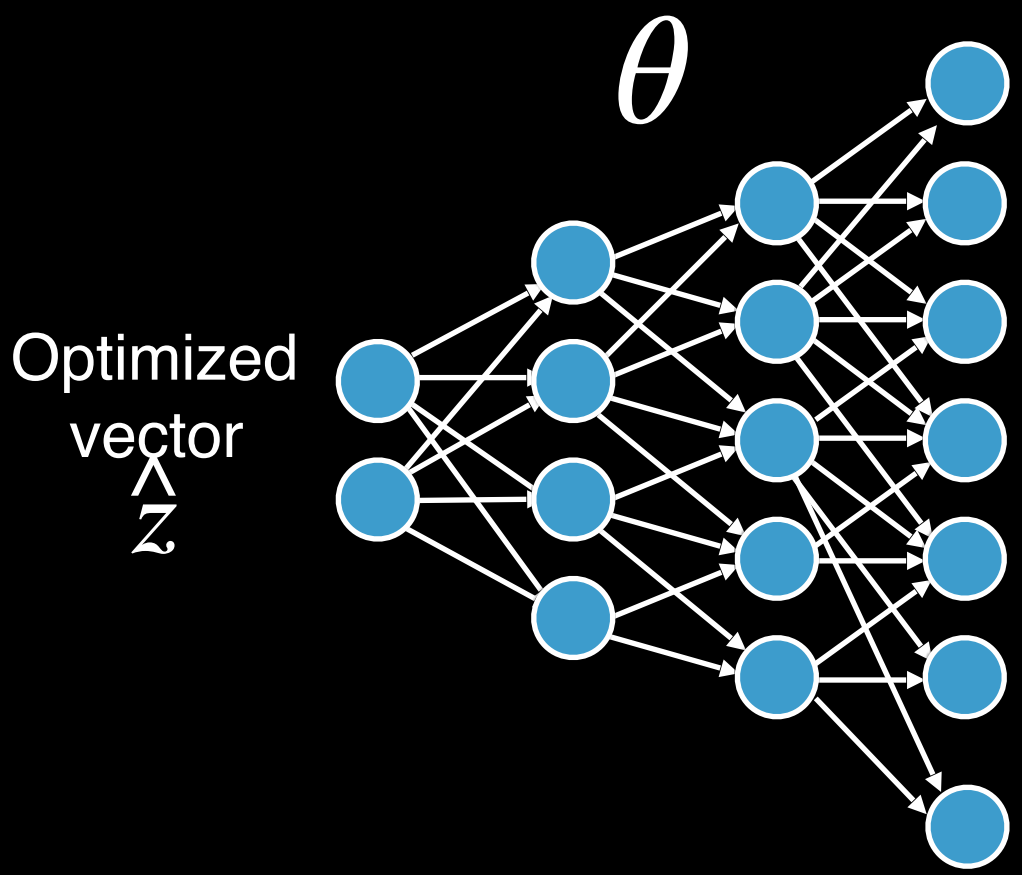


My image

$$\hat{z} = \underset{z}{\operatorname{argmin}} L_{rec}(I, G(z, \theta))$$

[Zhu et al., 2016]
[Dosovitskiy and Brox., 2016]

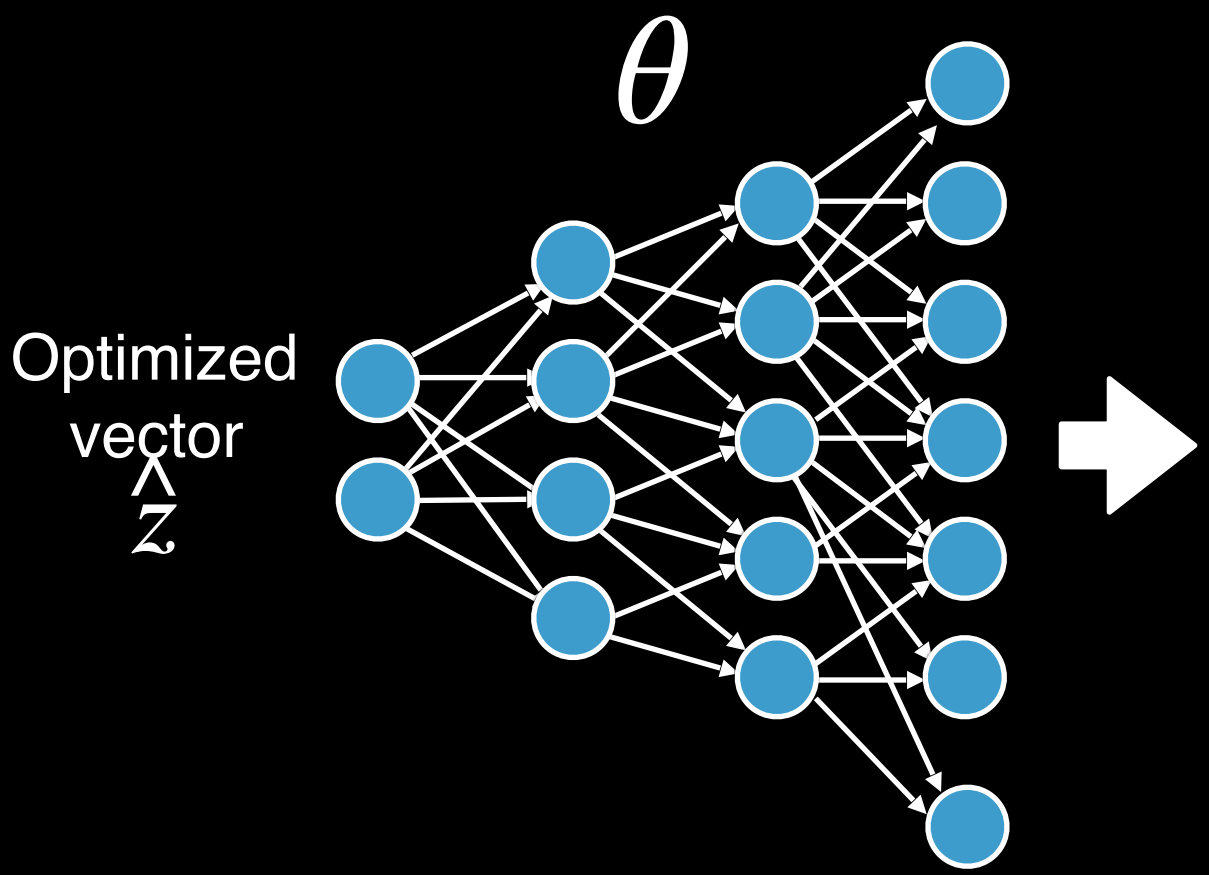
How to edit my own photo?



$$\hat{z} = \underset{z}{\operatorname{argmin}} L_{rec}(I, G(z, \theta))$$

[Zhu et al., 2016]
[Dosovitskiy and Brox., 2016]

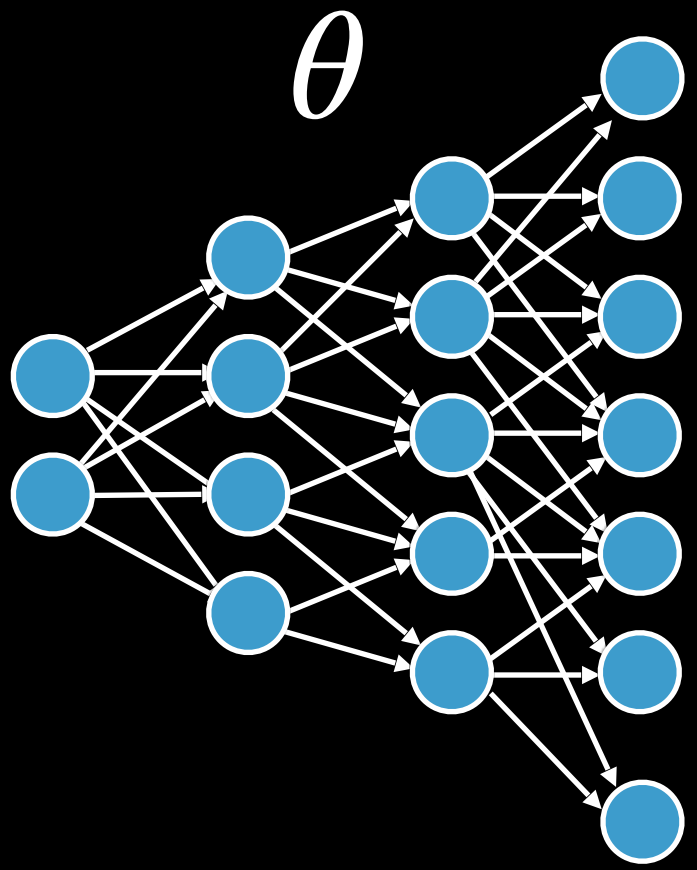
How to edit my own photo?



$$\hat{z} = \underset{z}{\operatorname{argmin}} L_{rec}(I, G(z, \theta))$$

[Zhu et al., 2016]
[Dosovitskiy and Brox., 2016]

How to edit my own photo?



Reconstructed image

$$\hat{z} = \underset{z}{\operatorname{argmin}} L_{rec}(I, G(z, \theta))$$

[Zhu et al., 2016]
[Dosovitskiy and Brox., 2016]

Find the differences…

Original image

Find the differences…

Original image

GAN reconstructed image

Find the differences…

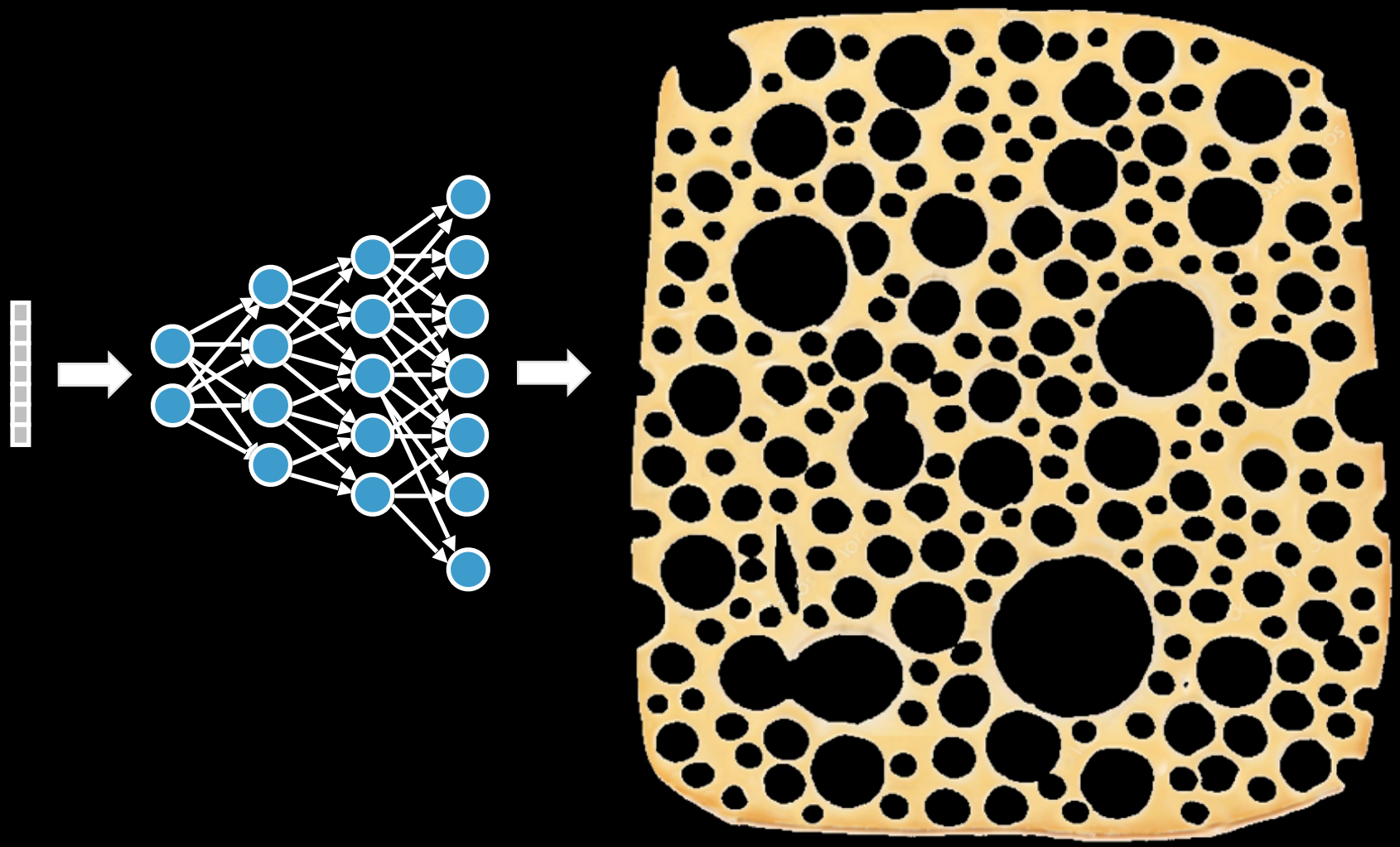
Original image

GAN reconstructed image

The cheese hypothesis

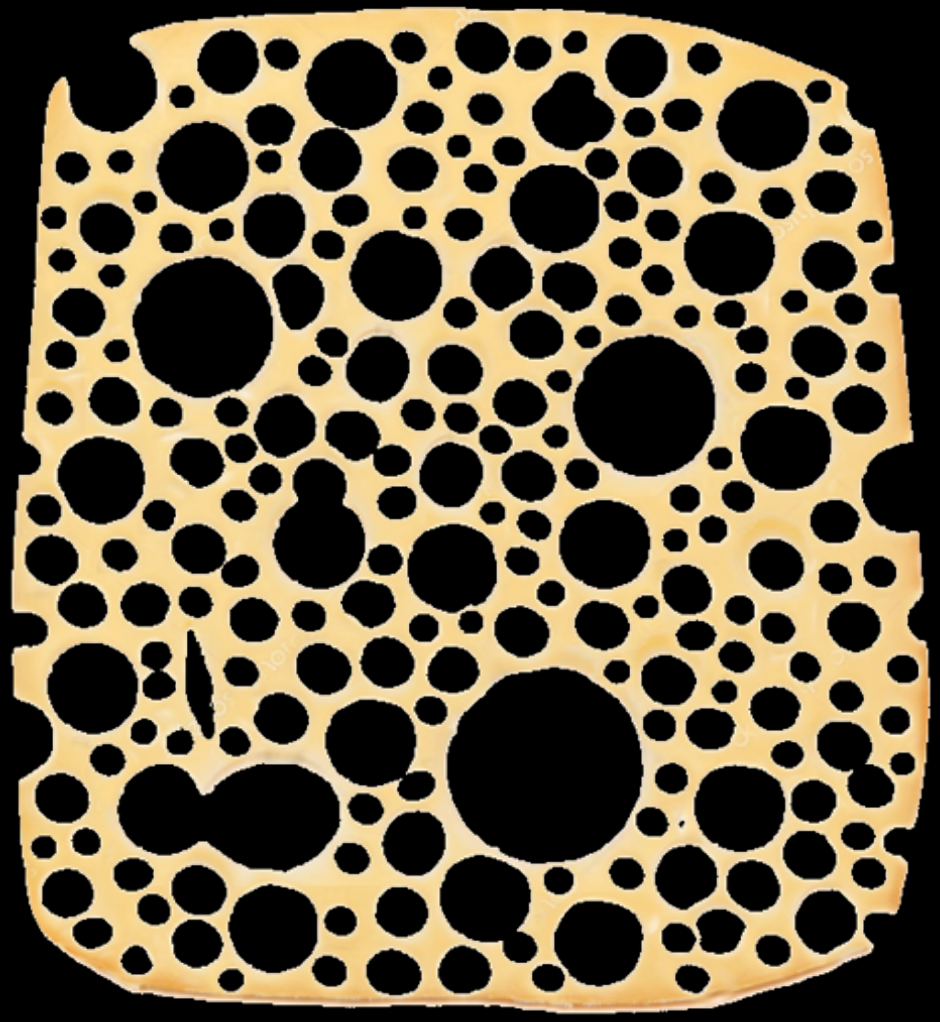
The cheese hypothesis

The cheese hypothesis



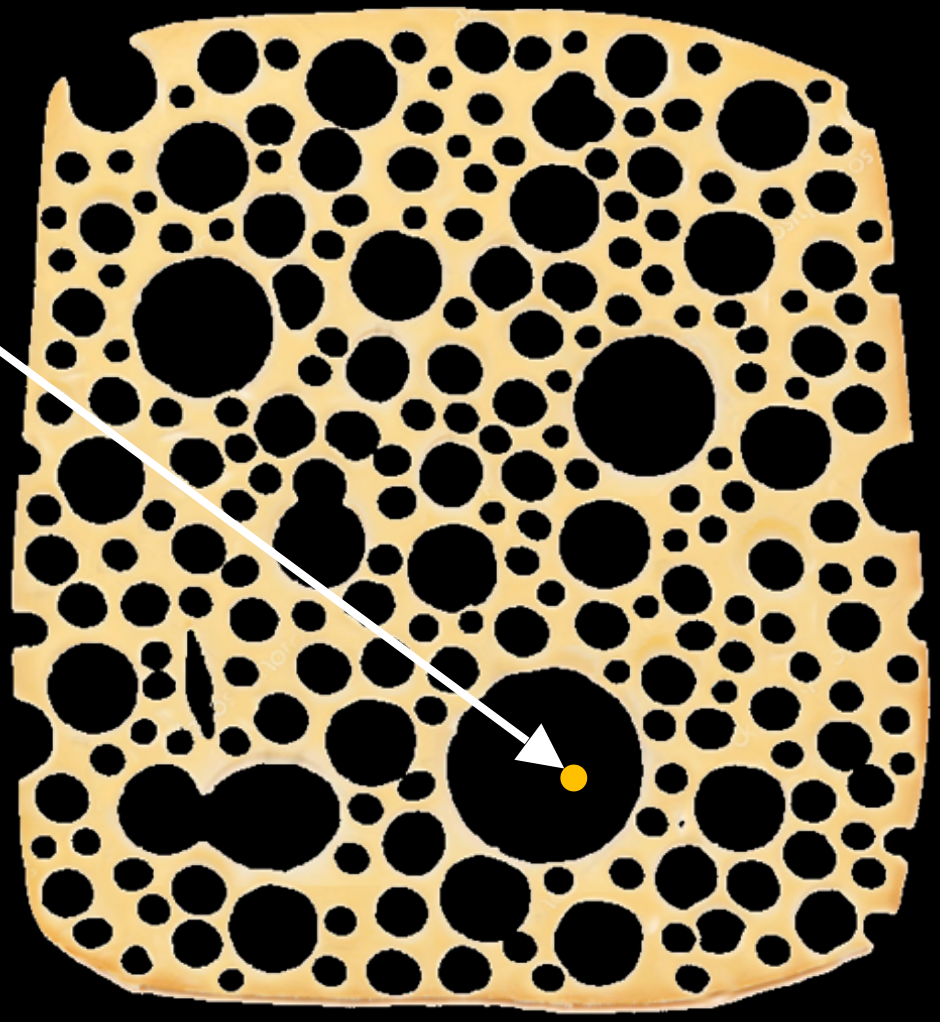
The cheese hypothesis

Original image



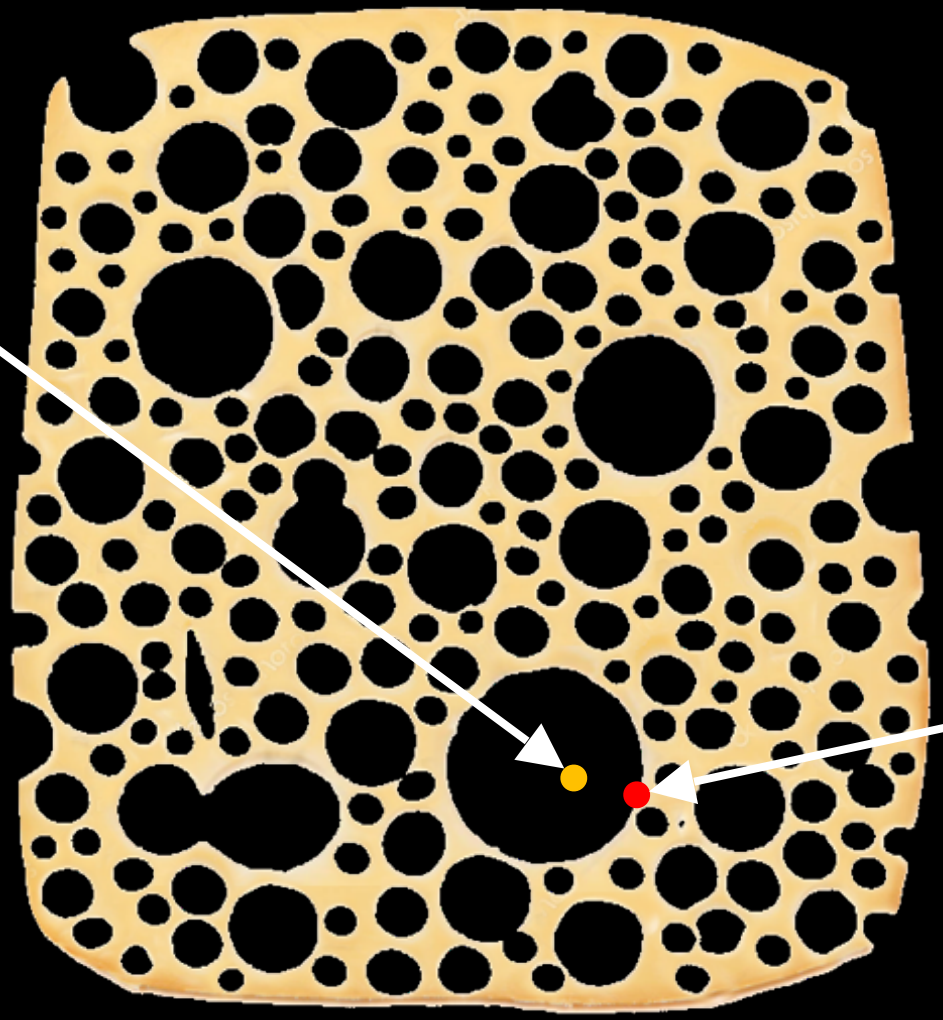
The cheese hypothesis

Original image



The cheese hypothesis

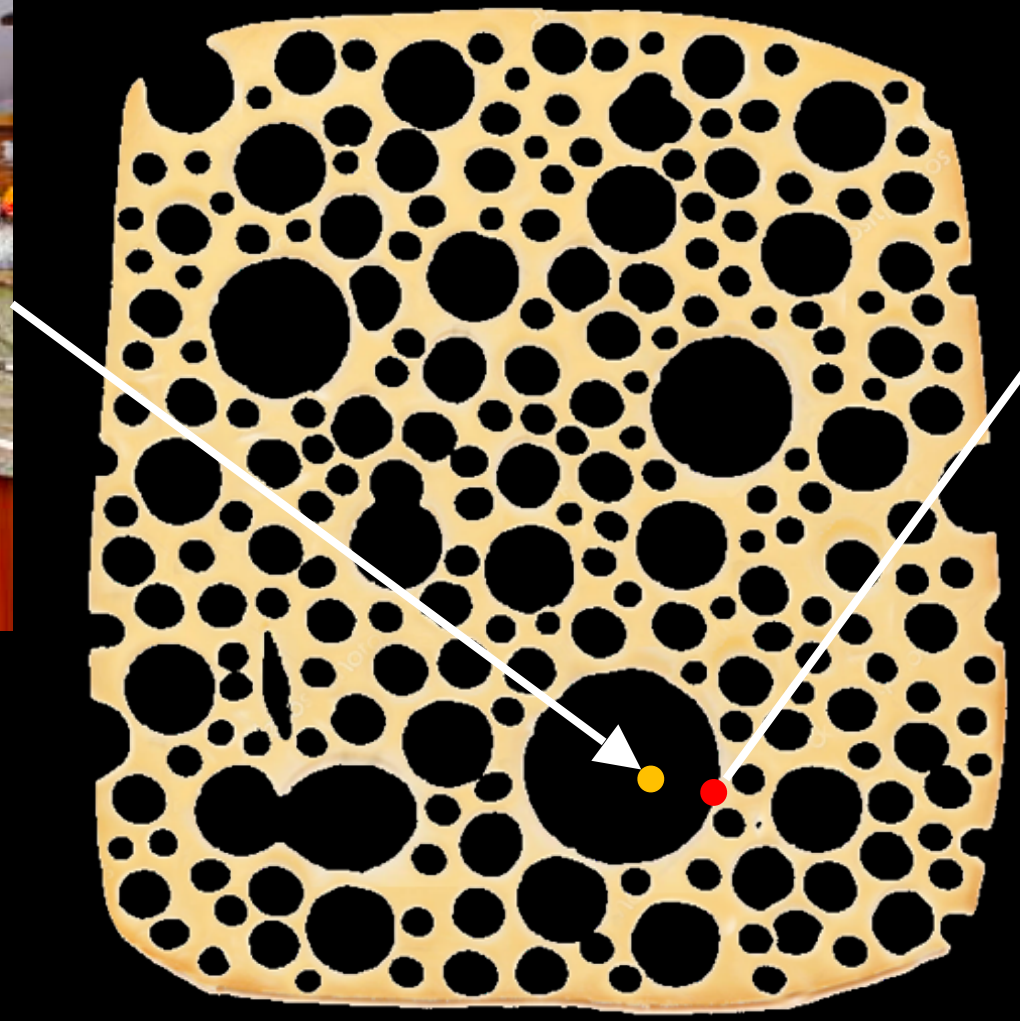
Original image



Optimized z

The cheese hypothesis

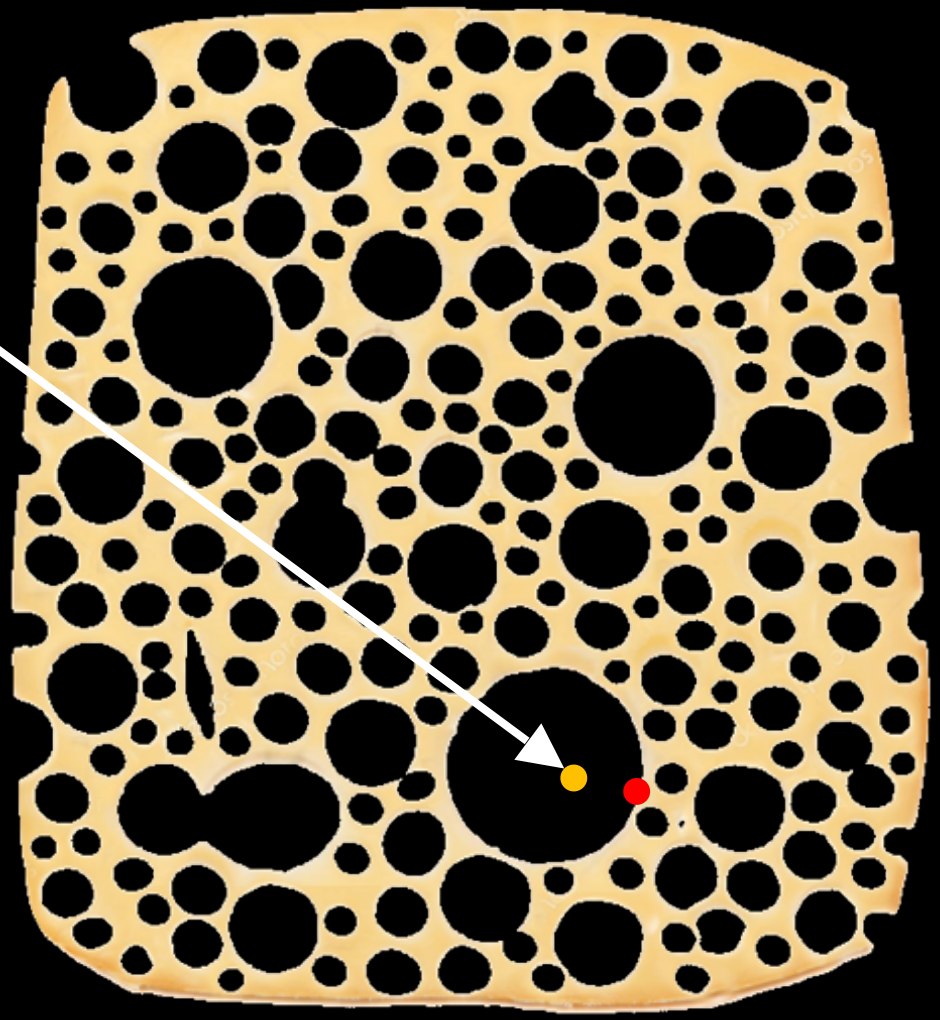
Original image



Optimized z

The cheese hypothesis

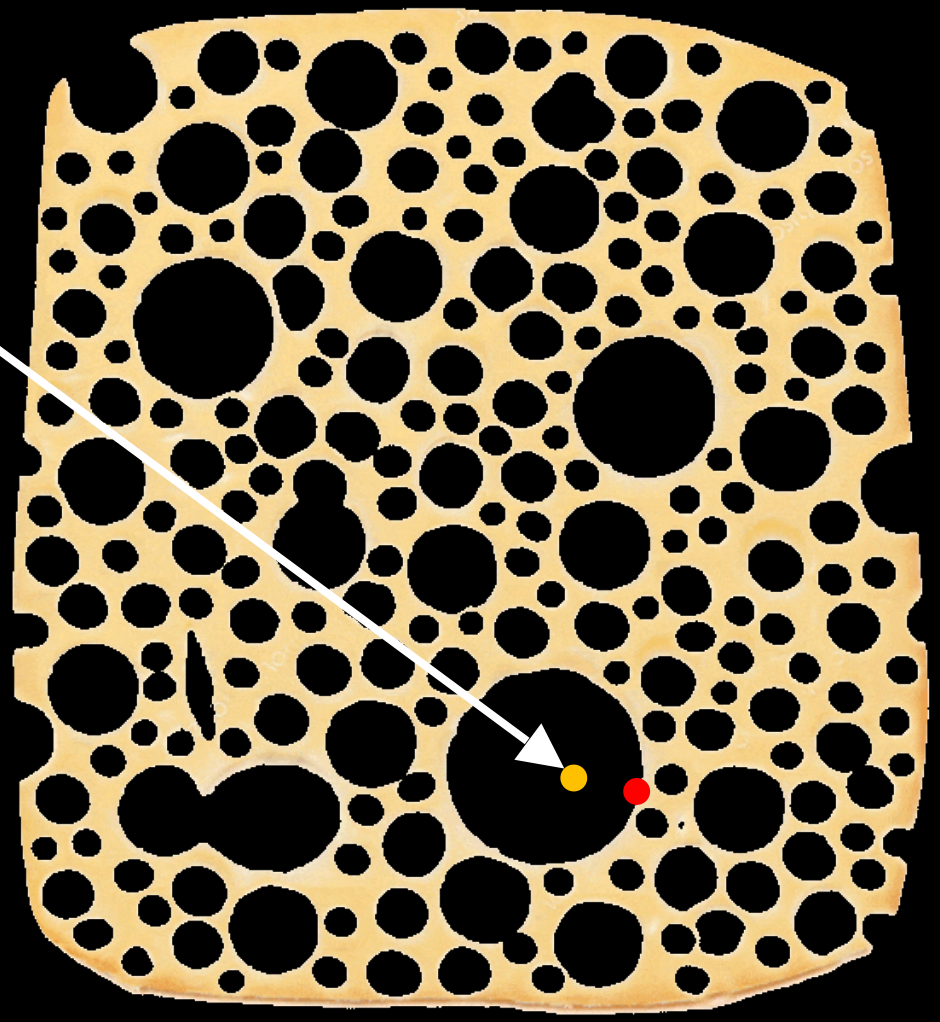
Original image



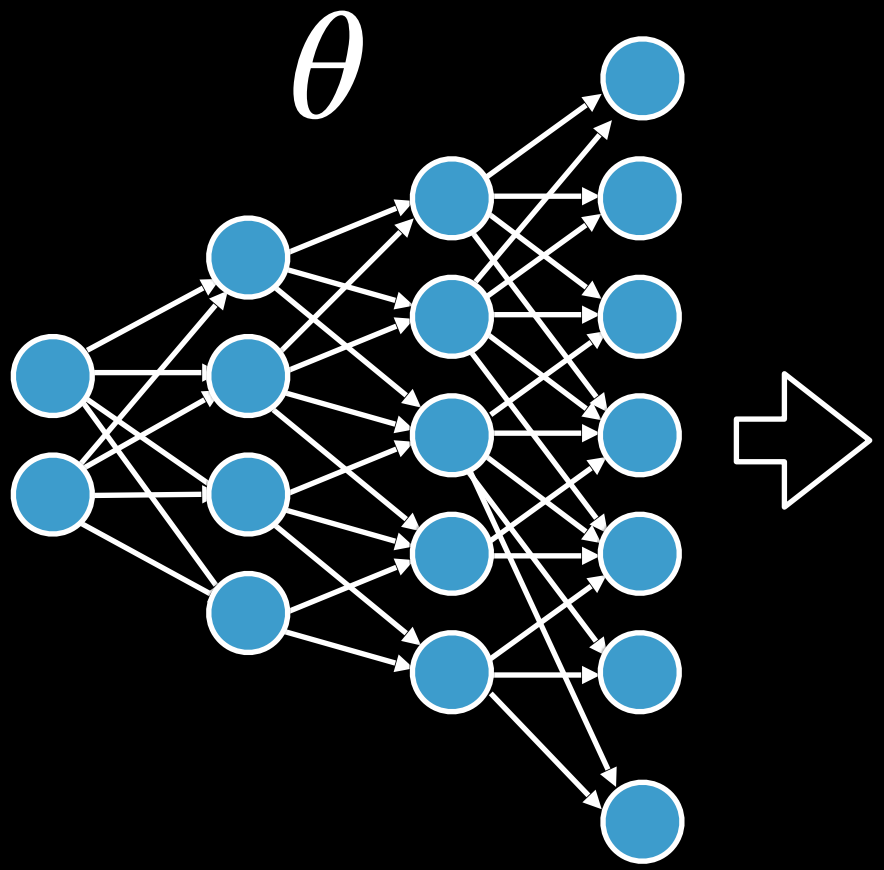
The cheese hypothesis

Adapted cheese

Original image



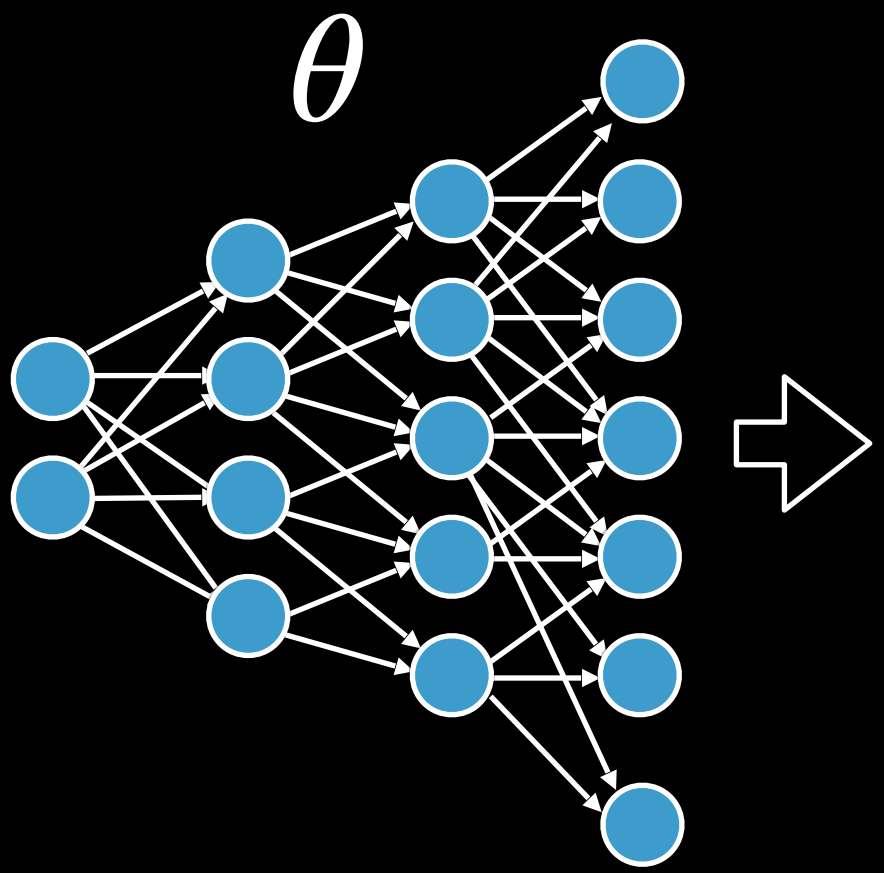
Reconstructing my own photo



Reconstructed image

$$\hat{z} = \underset{z}{\operatorname{argmin}} L_{rec}(I, G(z,\ \theta))$$

Reconstructing my own photo

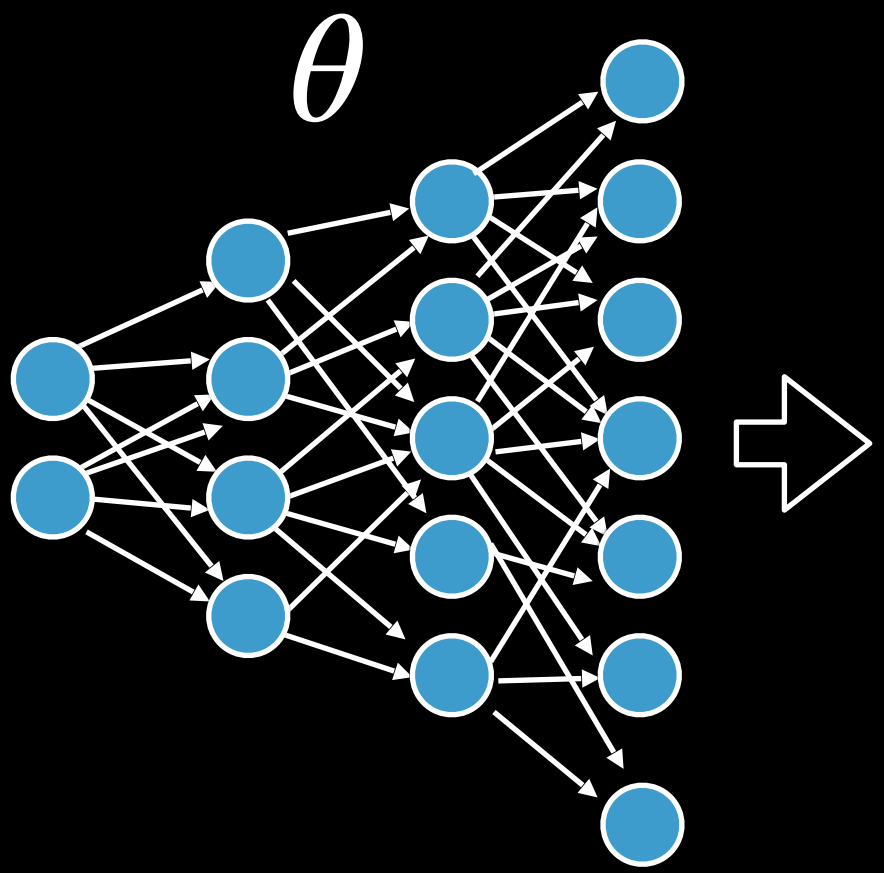


Reconstructed image

$$\hat{z}, \hat{\theta} = \underset{z,\theta}{\operatorname{argmin}} L_{rec}(I, G(z, \theta)) + R(\theta)$$

← Regularizer

Reconstructing my own photo

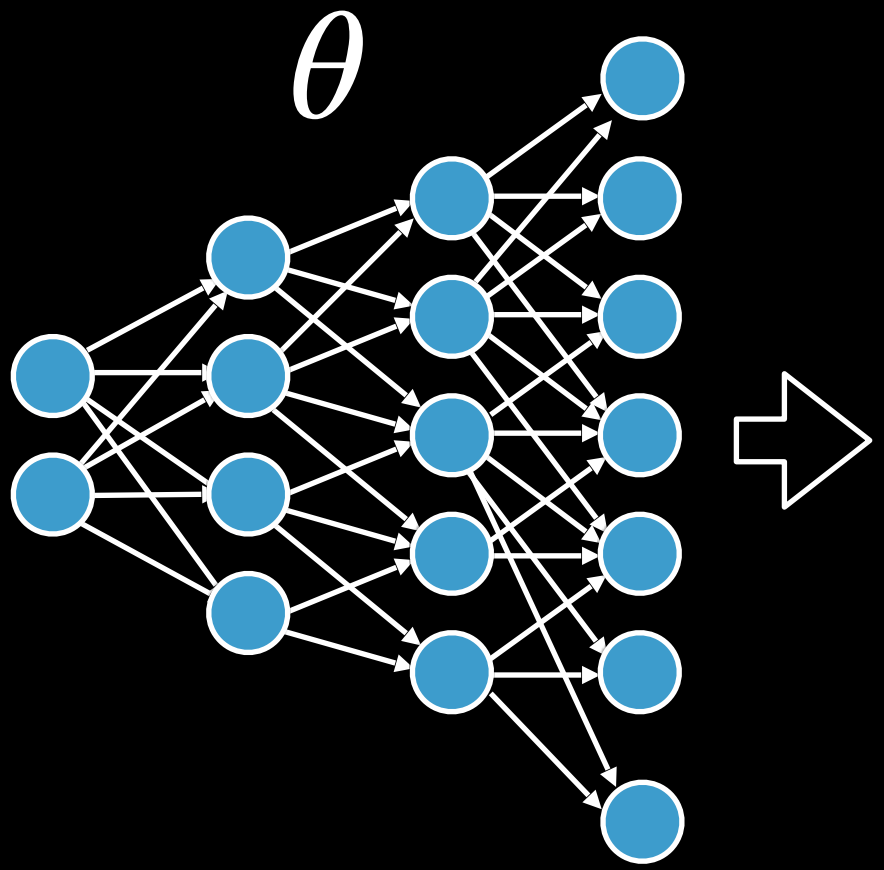


Reconstructed image

$$\hat{z}, \hat{\theta} = \underset{z,\theta}{\operatorname{argmin}} L_{rec}(I, G(z, \theta)) + R(\theta)$$

← Regularizer

Reconstructing my own photo

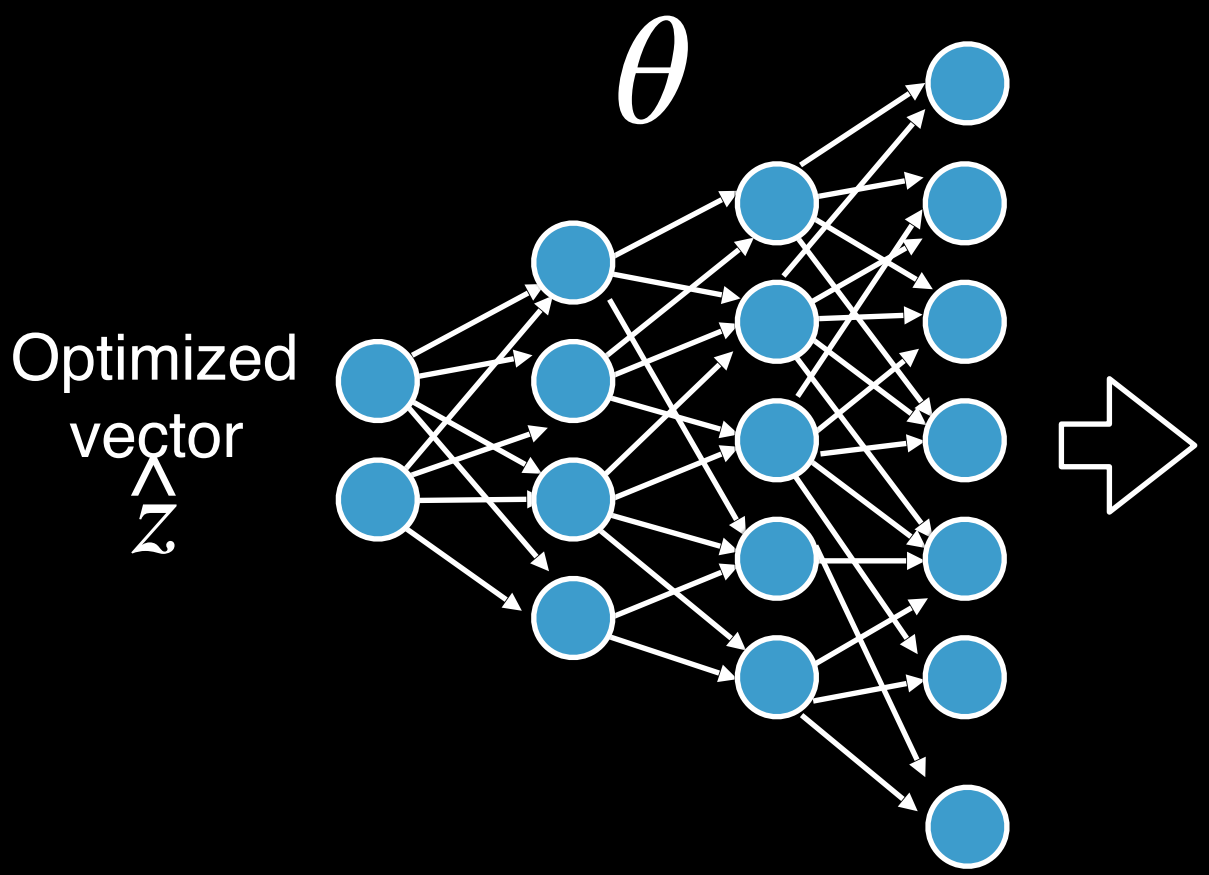


Reconstructed image

$$\hat{z}, \hat{\theta} = \underset{z,\theta}{\operatorname{argmin}} L_{rec}(I, G(z, \theta)) + R(\theta)$$

← Regularizer

Reconstructing my own photo

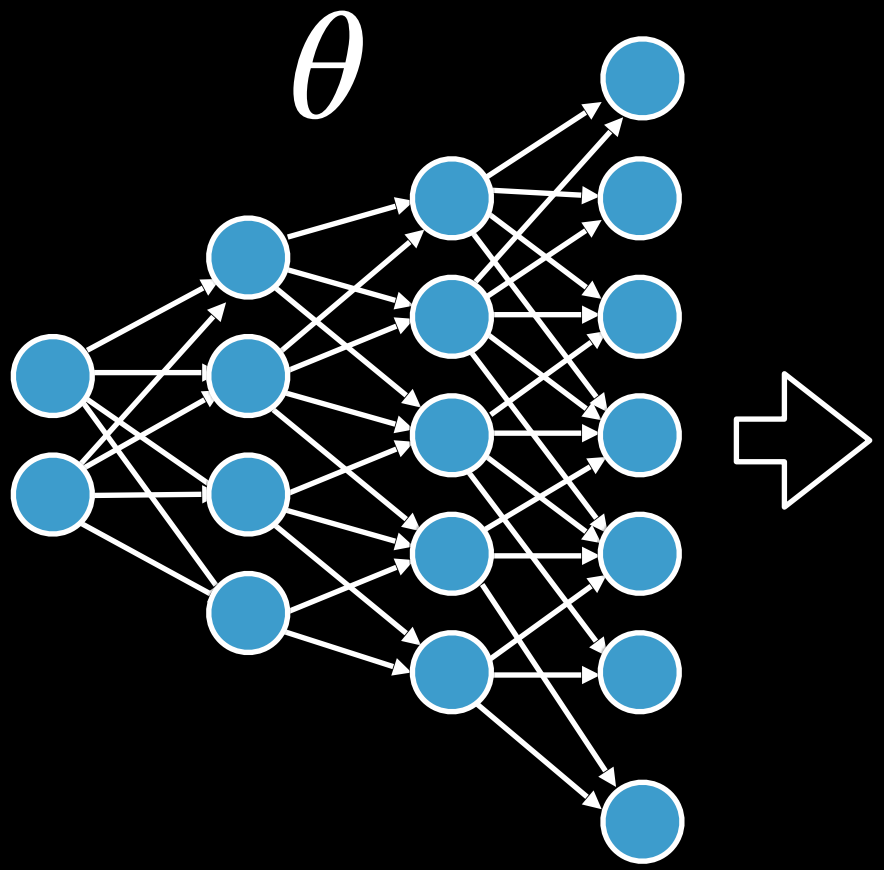


Reconstructed image

$$\hat{z}, \hat{\theta} = \underset{z,\theta}{\operatorname{argmin}} L_{rec}(I, G(z, \theta)) + R(\theta)$$

← Regularizer

Reconstructing my own photo

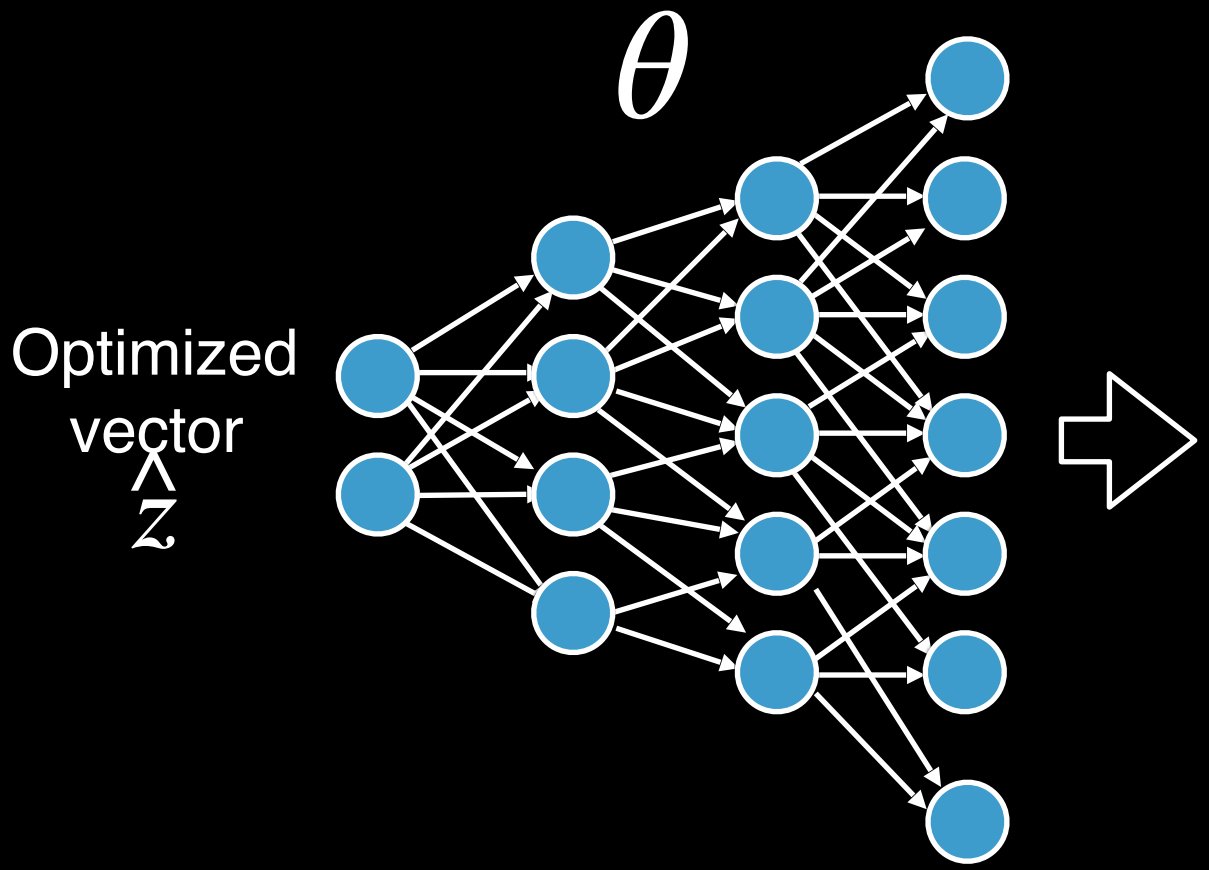


Reconstructed image

$$\hat{z}, \hat{\theta} = \underset{z,\theta}{\operatorname{argmin}} L_{rec}(I, G(z, \theta)) + R(\theta)$$

← Regularizer

Reconstructing my own photo

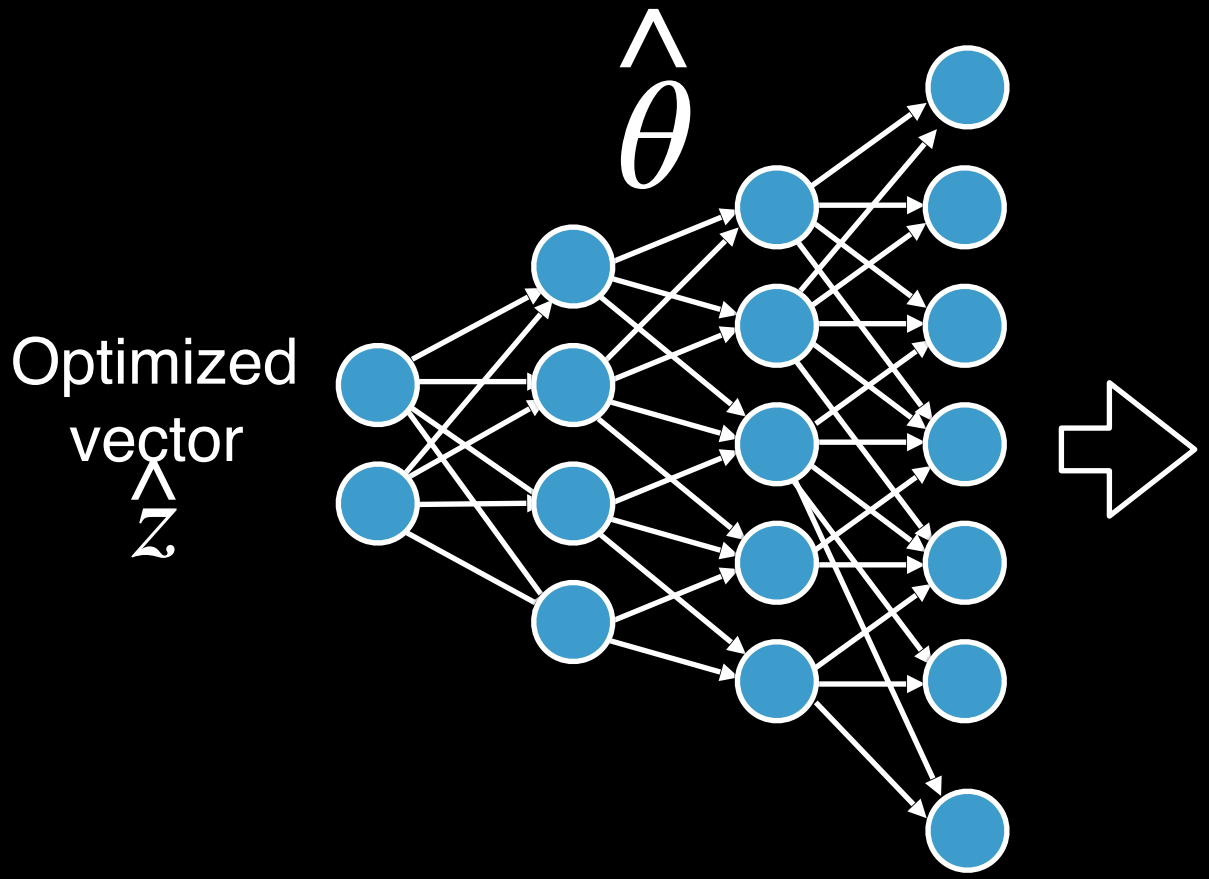


Reconstructed image

$$\hat{z}, \hat{\theta} = \underset{z,\theta}{\operatorname{argmin}} L_{rec}(I, G(z, \theta)) + R(\theta)$$

← Regularizer

Reconstructing my own photo



Reconstructed image

$$\hat{z}, \hat{\theta} = \underset{z,\theta}{\operatorname{argmin}} L_{rec}(I, G(z, \theta)) + R(\theta)$$

← Regularizer

Reconstructing my own photo

Original image

Optimized

Optimized and

Inspired by Deep Image Prior [Ulyanov et al., 2018] and [Shocher et al., 2017]

Will editing work?

Optimized and

Will editing work?

Optimized and

Activate Window Neurons

Modified image

Semantic Photo Manipulation [Bau, et al. SIGGRAPH 2019]

Non-local editing effects

Original image and edit area

Edited result with adapted network

Semantic Photo Manipulation [Bau, et al. SIGGRAPH 2019]

Manipulating a real photo

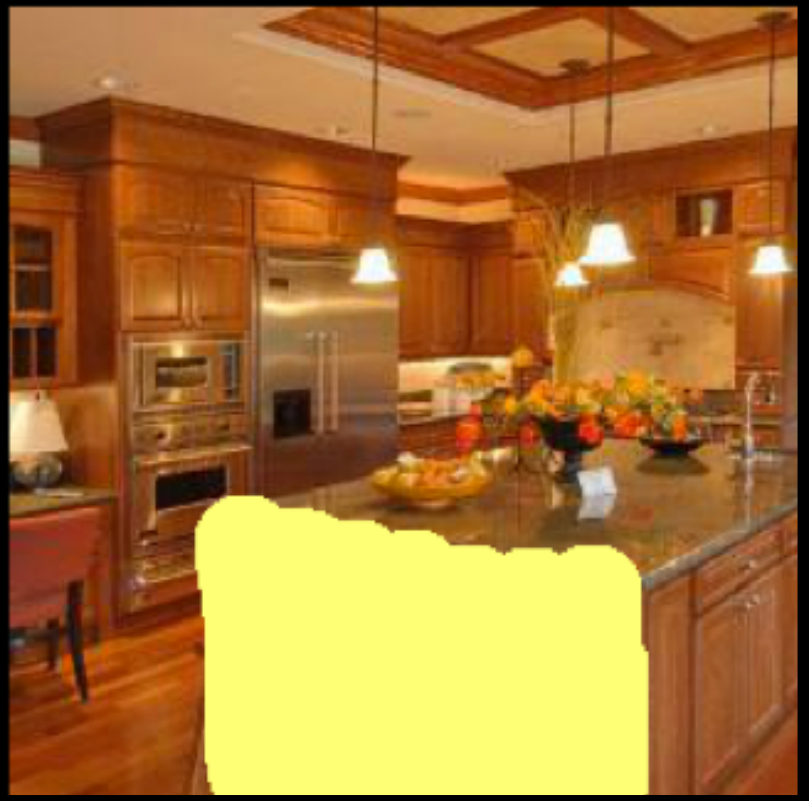
Input image

Add windows

Output result

Manipulating a real photo

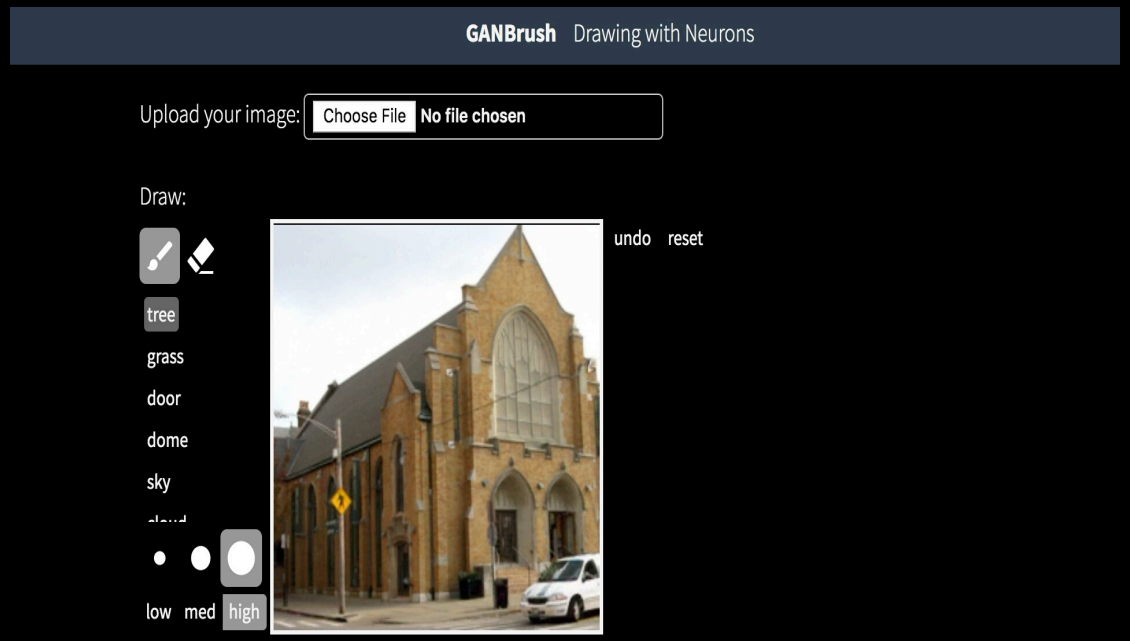
Input image



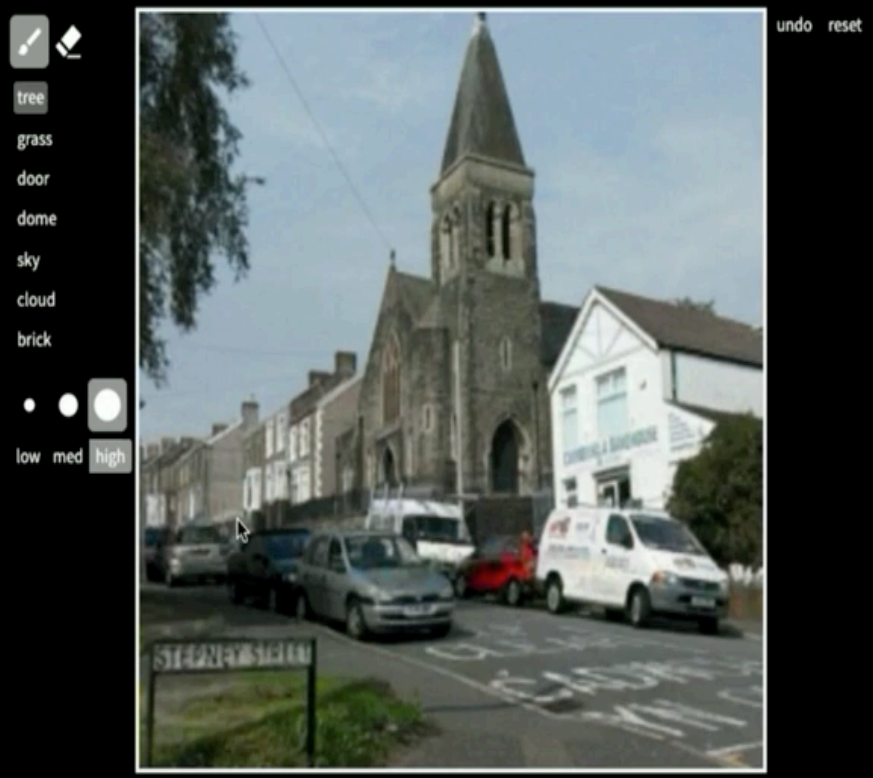
Remove chairs

Output result

Realtime editing



Cincinnati, OH

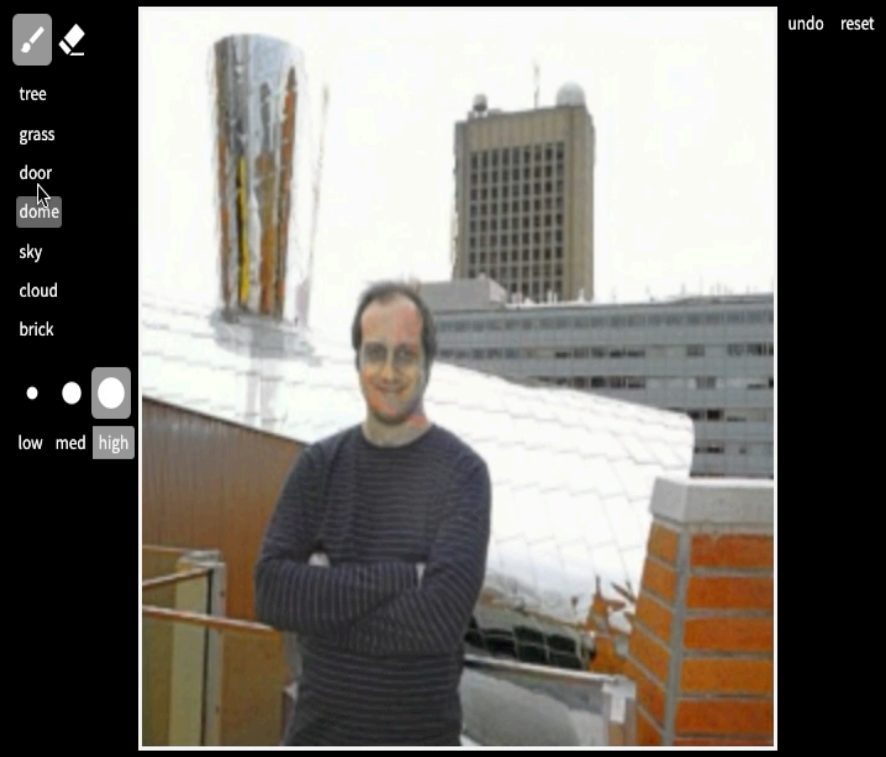


Swansea, Wales
and Los Angeles, CA

Online demo: **ganpaint.io**

Sensitivity to context

MIT Stata Center



Fantastical Scene

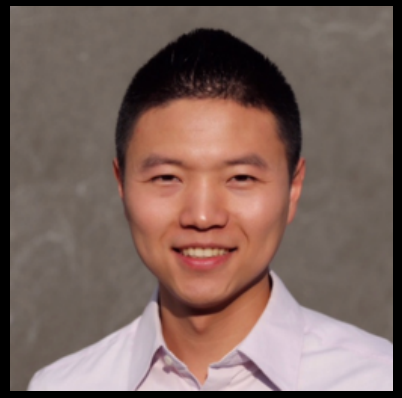
A GAN is more opinionated than photoshop.

Thank You!

Hendrik Strobelt

William Peebles

Jonas Wulff



Bolei Zhou

Jun-Yan Zhu

Antonio Torralba

Demo: **ganpaint.io**