MIT CSAIL

Rewriting Generative Networks

David Bau

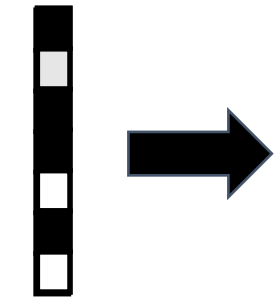
The 1870 Decatur Courthouse Tower Tree

Can we create a model
without a data set?

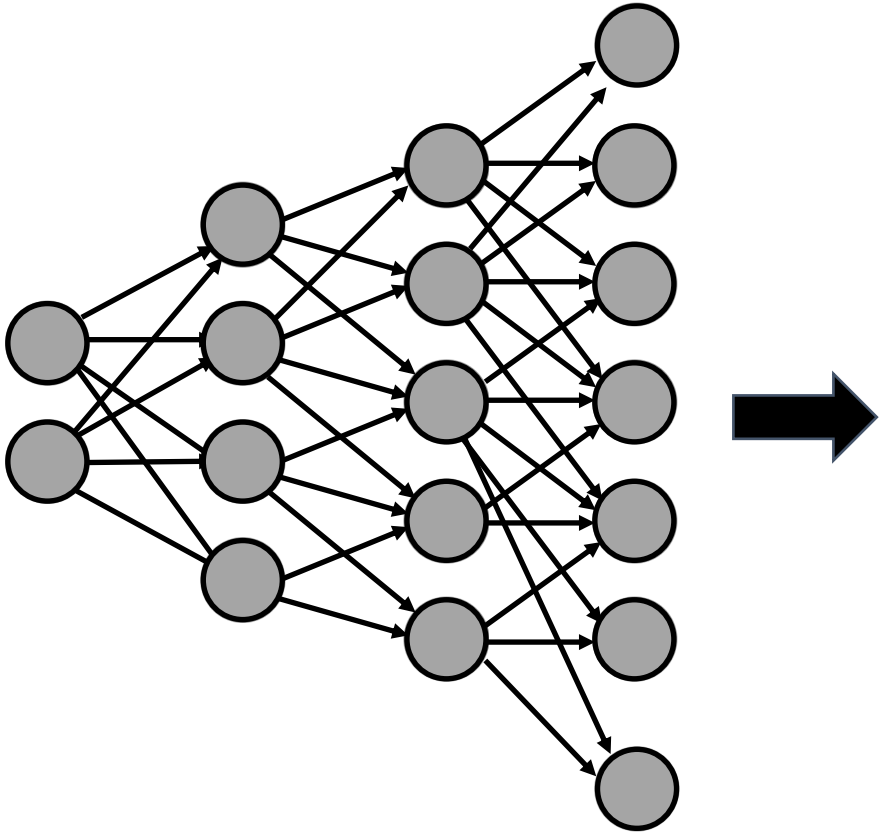
Part 1: Dissecting a GAN

Using a GAN Generator

Random vector



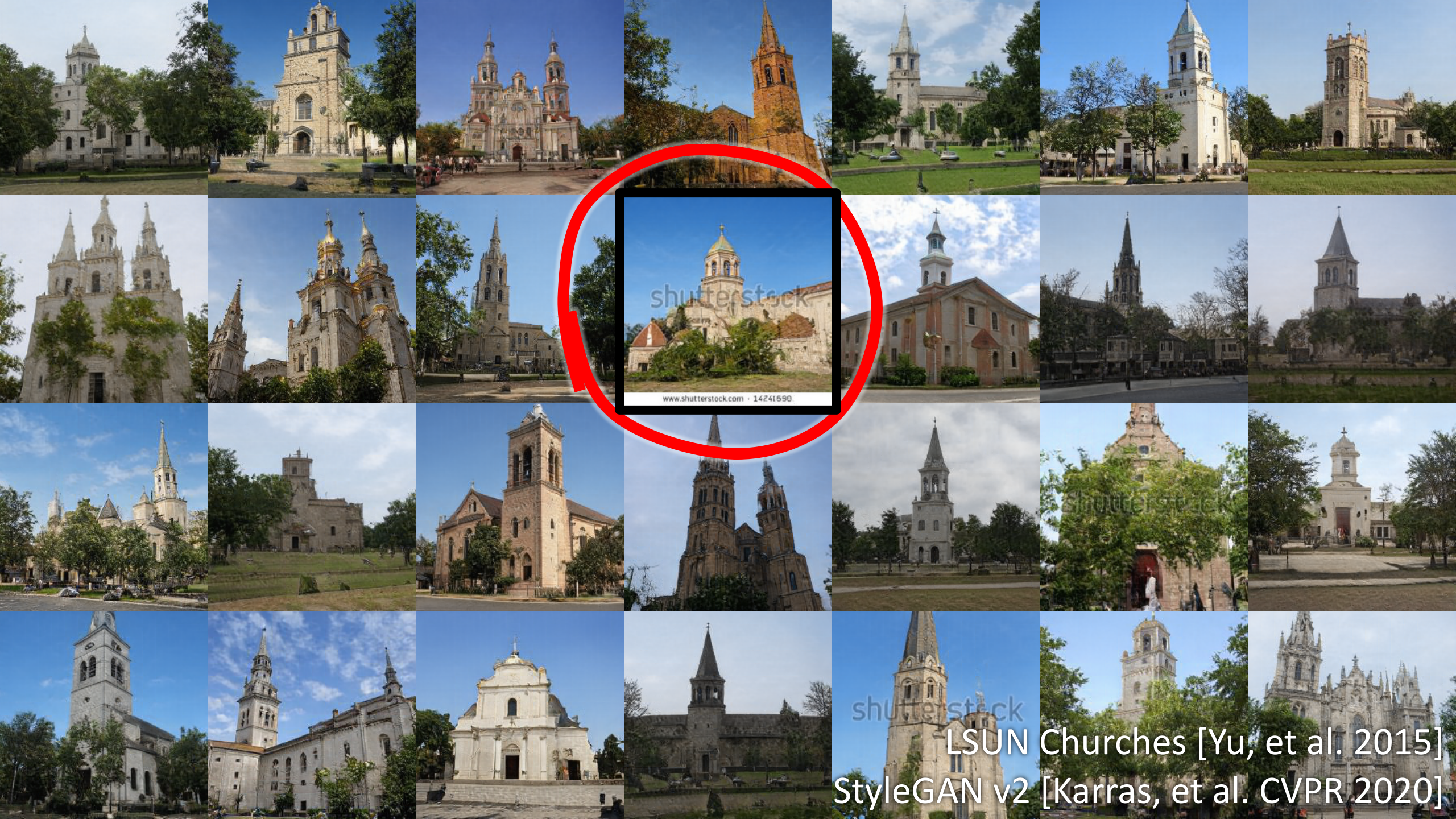
512 dimensions



Randomly generated image

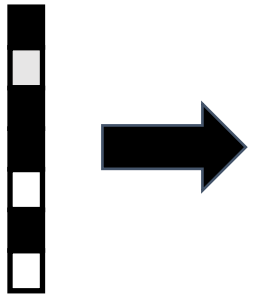
StyleGAN v2 [Karras, et al. CVPR 2020]

LSUN Churches [Yu, et al. 2015]
StyleGAN v2 [Karras, et al. CVPR 2020]

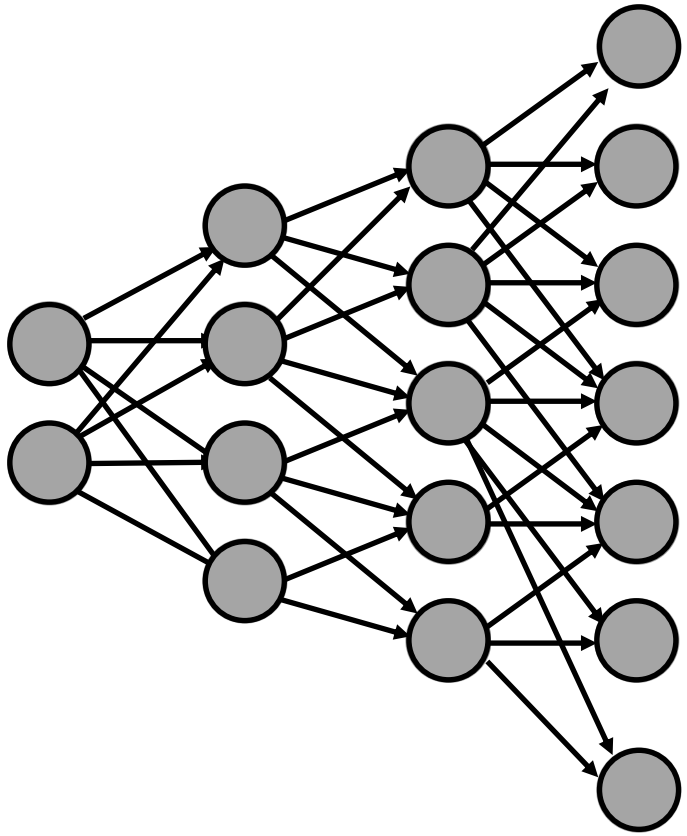

shutterstock
www.shutterstock.com · 14741690
shutterstock
shutterstock
LSUN Churches [Yu, et al. 2015]
StyleGAN v2 [Karras, et al. CVPR 2020]

Are there watermark neurons?

Random vector



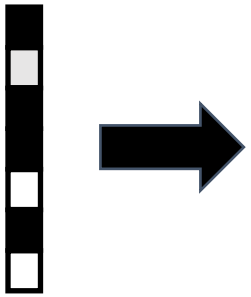
512 dimensions



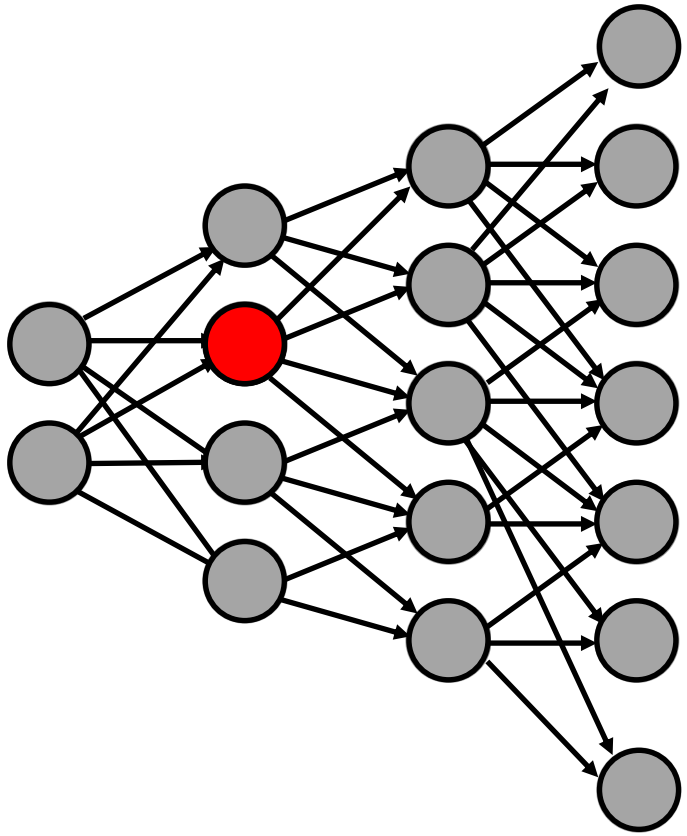
Randomly generated image

Are there watermark neurons?

Random vector



512 dimensions



Randomly generated image

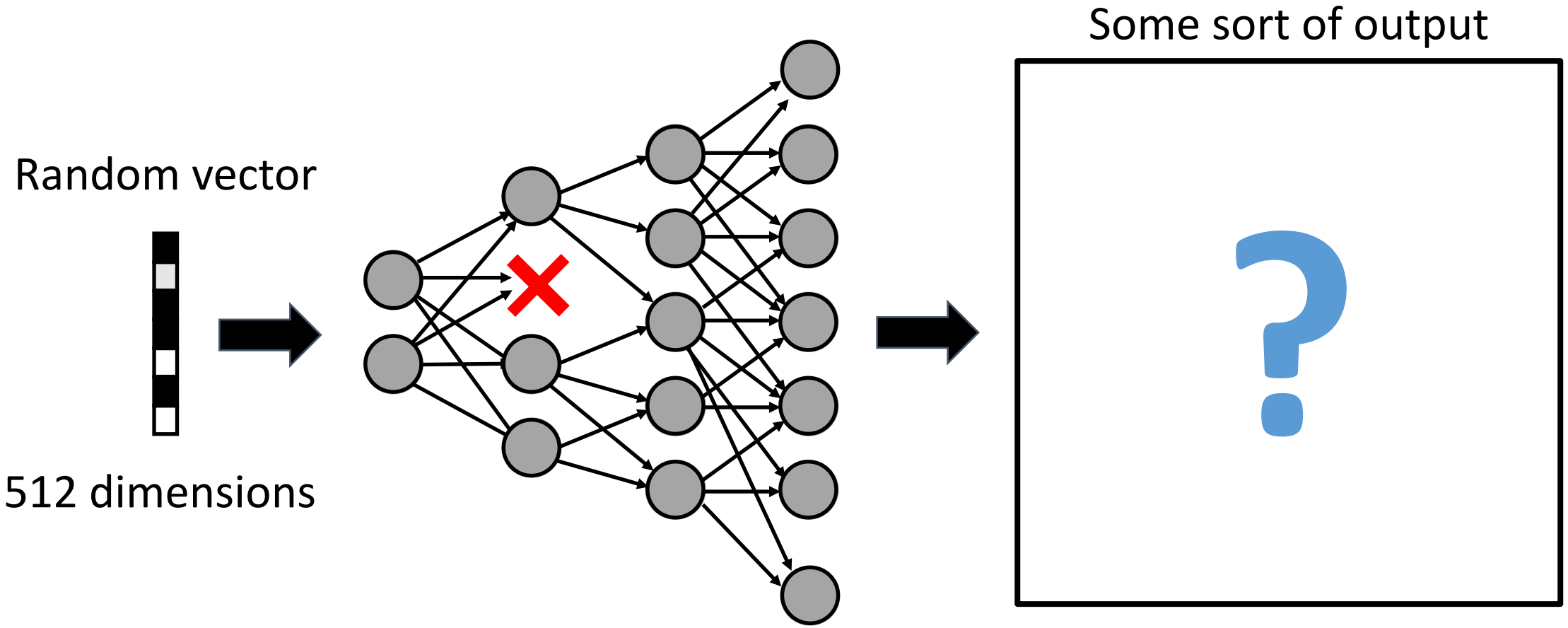
Layer 5, Neuron 304:





Layer 5, Neuron 234:

What if we turn off these neurons?



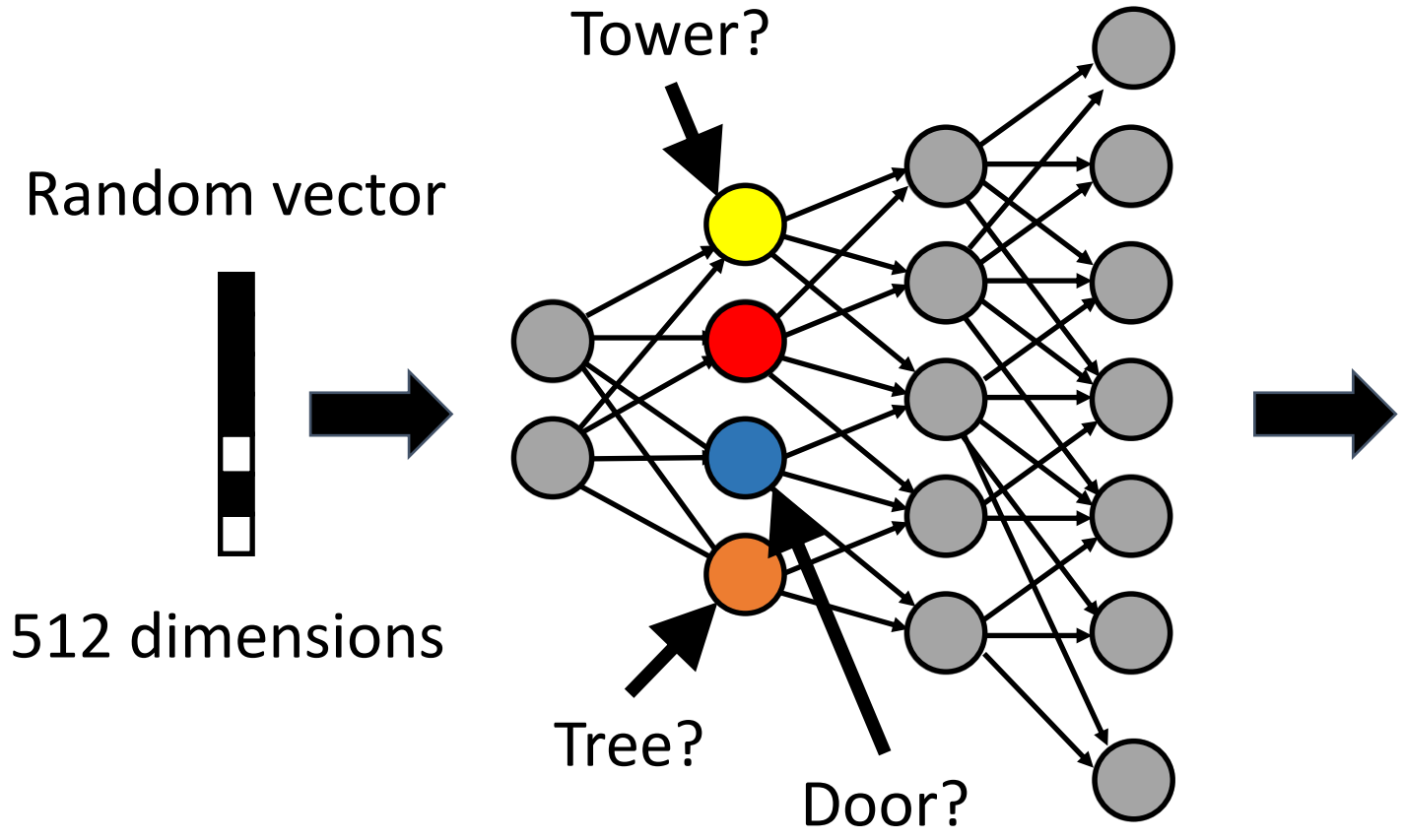
Zeroing 30 Watermark Neurons in Layer 5








Are there other objects?



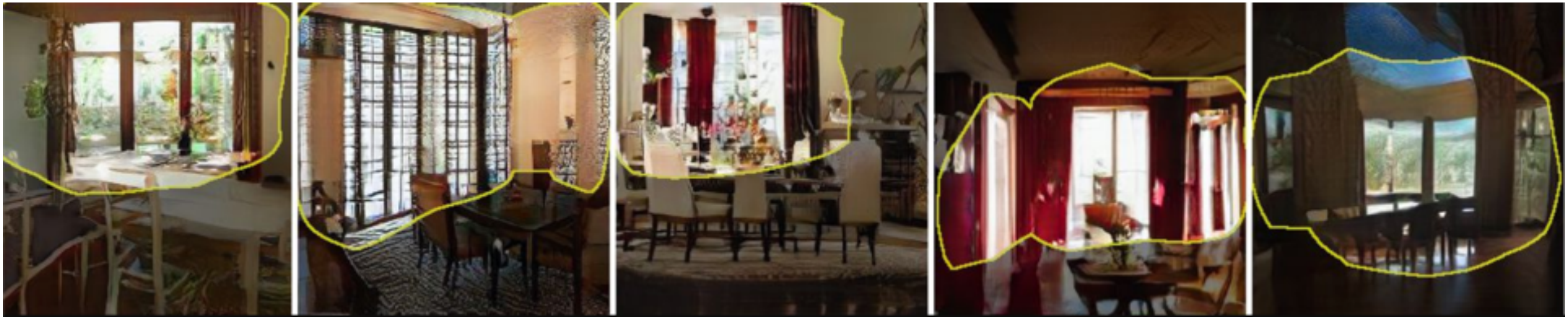
Randomly generated image

Progressive GAN [Karras, et al 2018]
GAN Dissection [Bau et al., ICLR 2019]

Layer 4, Neuron 119: tree

Layer 4, Neuron 43 : dome

Layer 4, Neuron 84: window

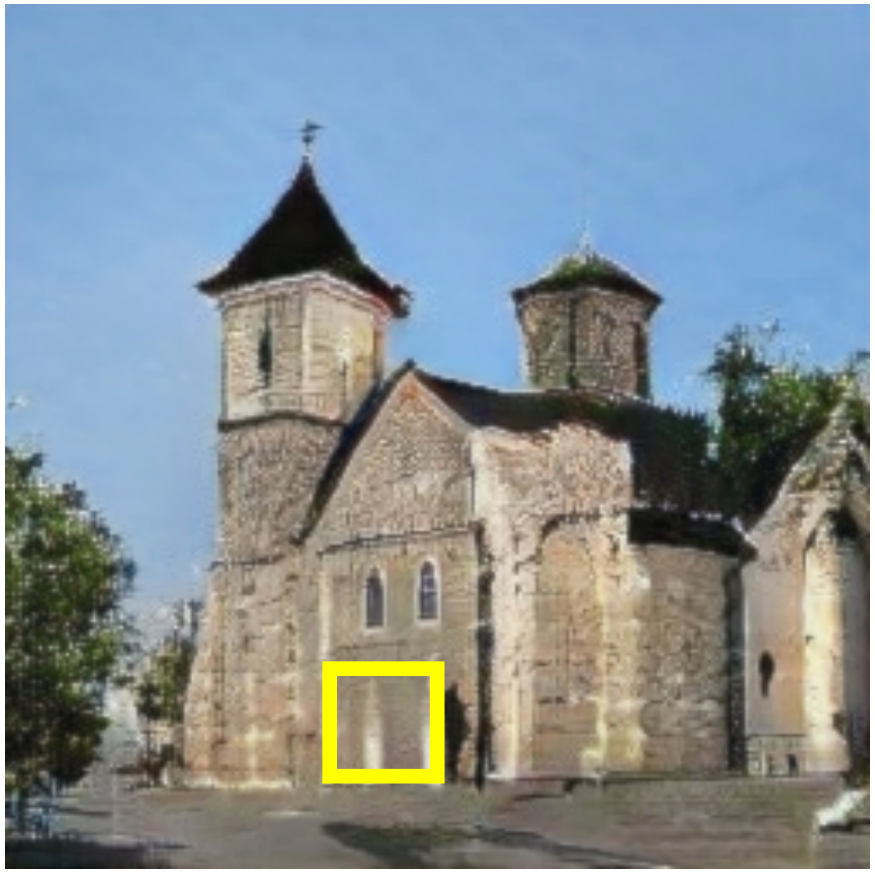


Layer 4, Neuron 315: chair

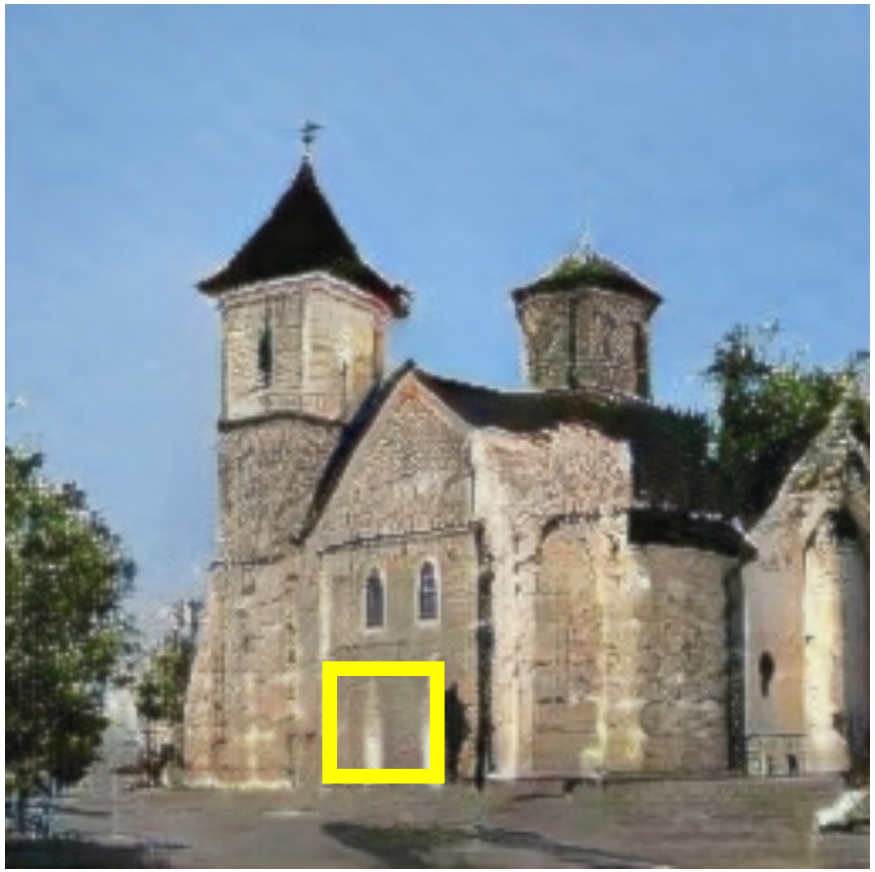
Turning off tree neurons

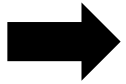
Occluded buildings are now visible

Turning on door neurons



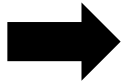
Turning on door neurons





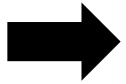
Turning on door neurons

Turning on door neurons



Turning on door neurons

Turning on door neurons



Activating neurons interactively

Demo at gandissect.csail.mit.edu

Part 2: Editing a Real Photo

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

How to edit my own photo?

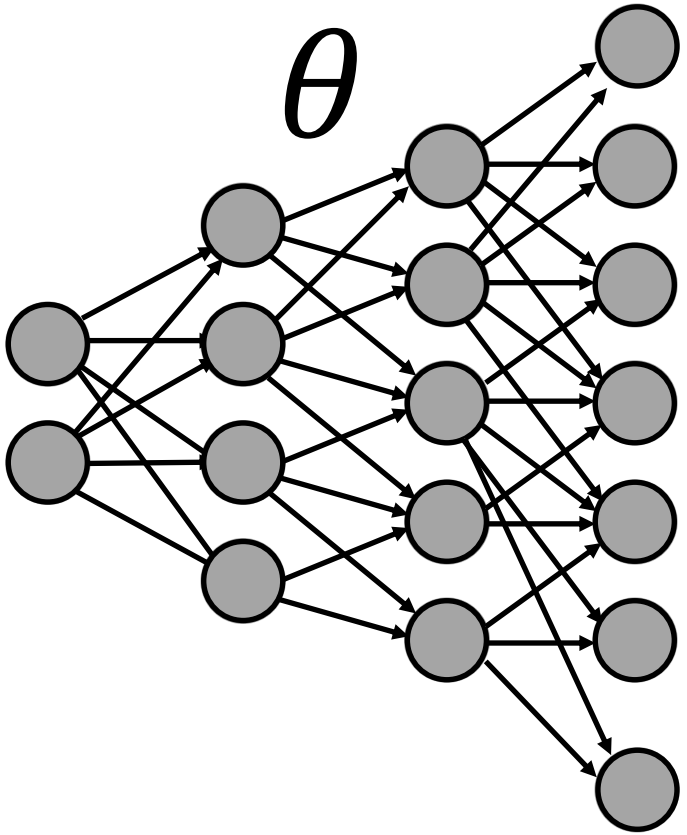
GAN-Synthesized Kitchen

My Kitchen Photo

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

How to edit my own photo?

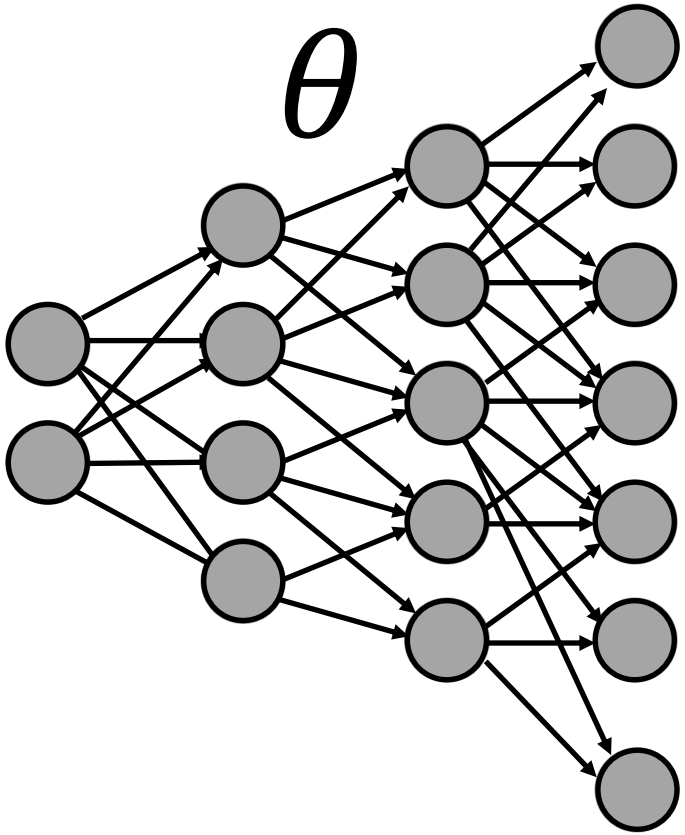
Random vector z?



My Kitchen Photo

How to edit my own photo?

Optimized vector $\hat{z}$

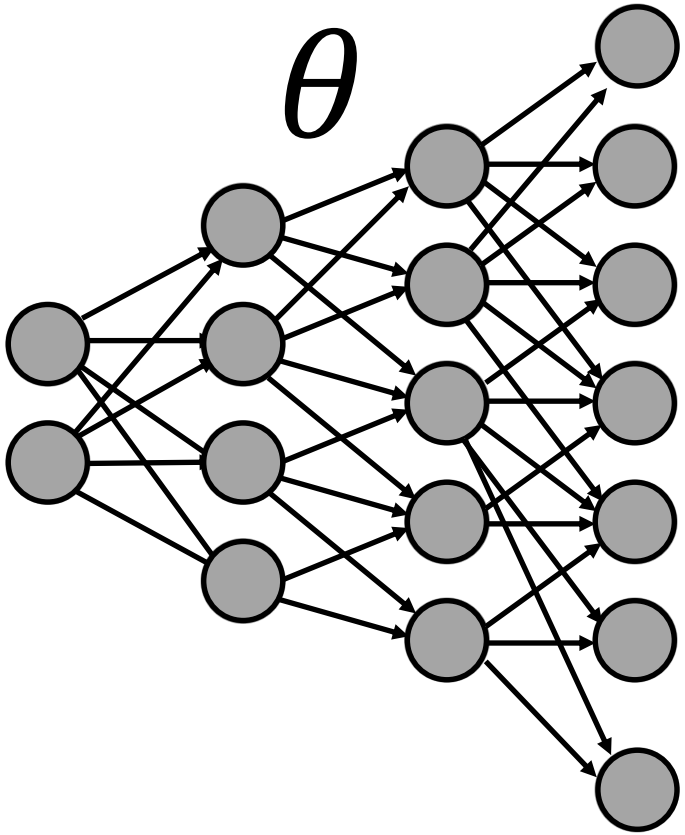


My image I

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

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

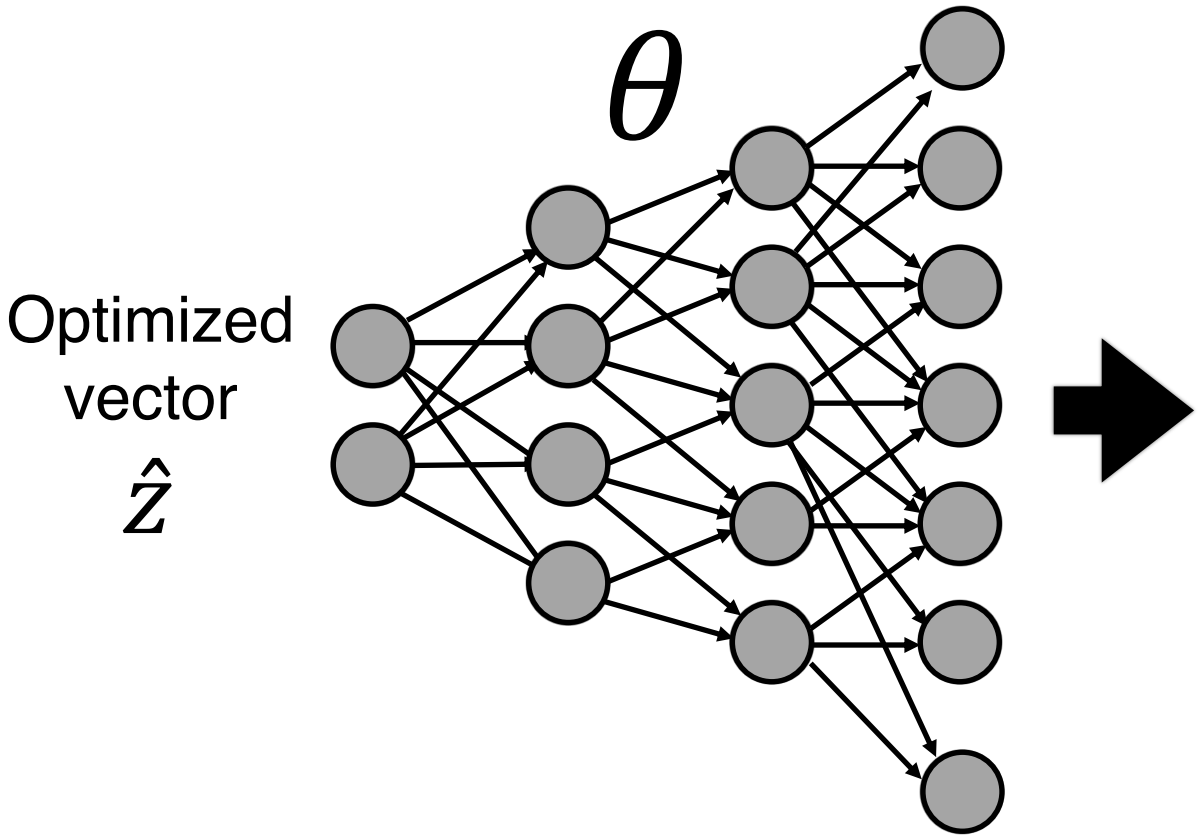
How to edit my own photo?



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

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

How to edit my own photo?

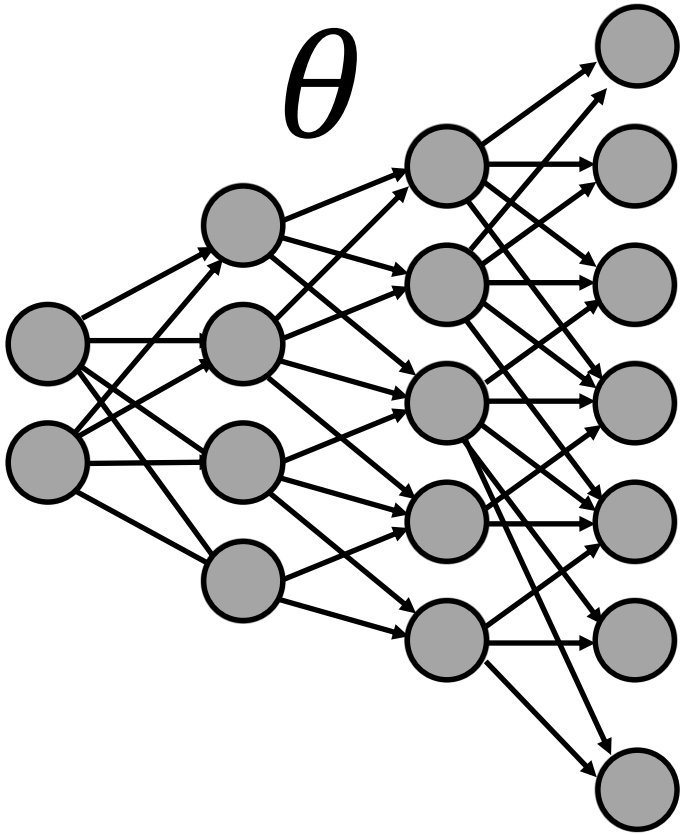


$$\hat{z} = \underset{z}{\operatorname{argmin}} L_{rec}(I, G(z, \theta))$$

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

How to edit my own photo?

Optimized vector $\hat{z}$



Reconstructed image $G(\hat{z}, \theta)$

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

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

Find the differences…

Original image

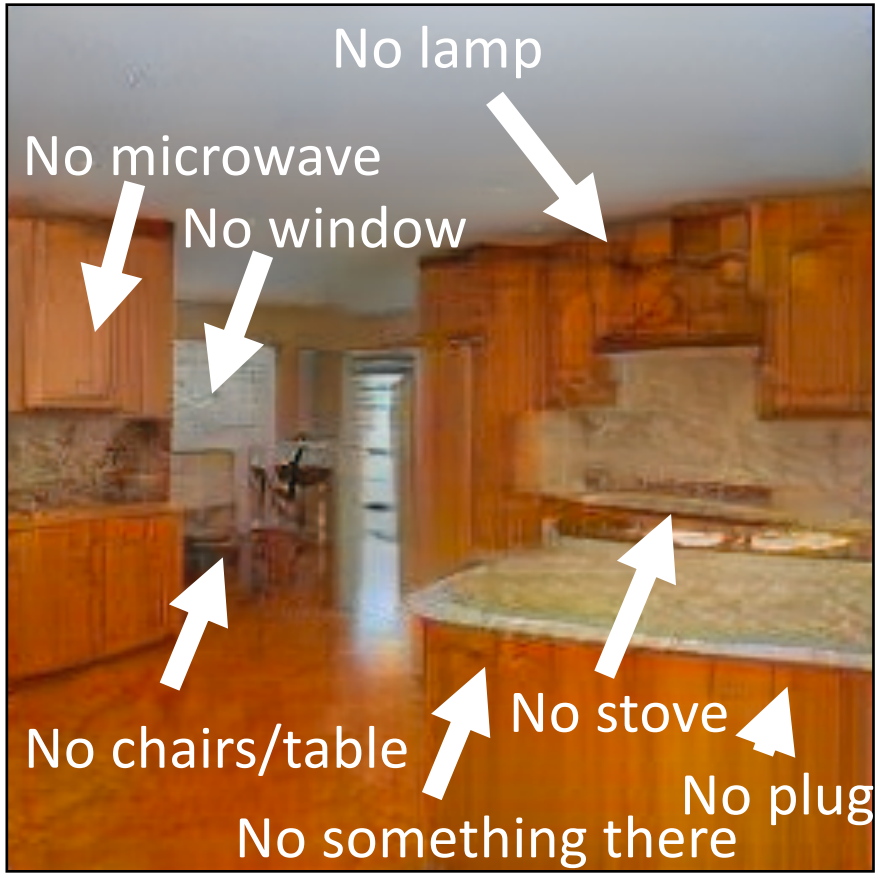
Find the differences…

Original image

GAN reconstructed image

Find the differences…

Original image

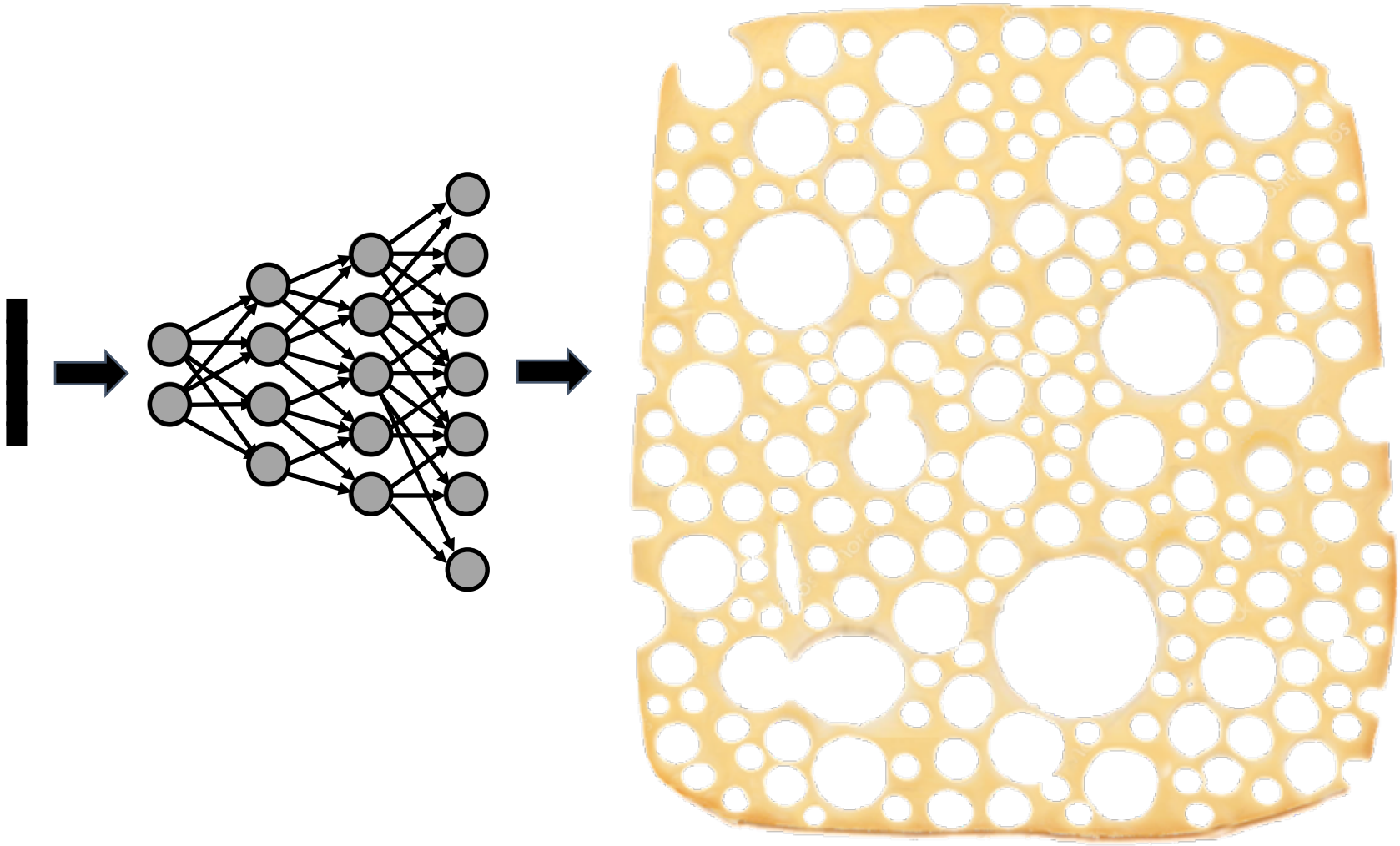


GAN reconstructed image

The cheese hypothesis

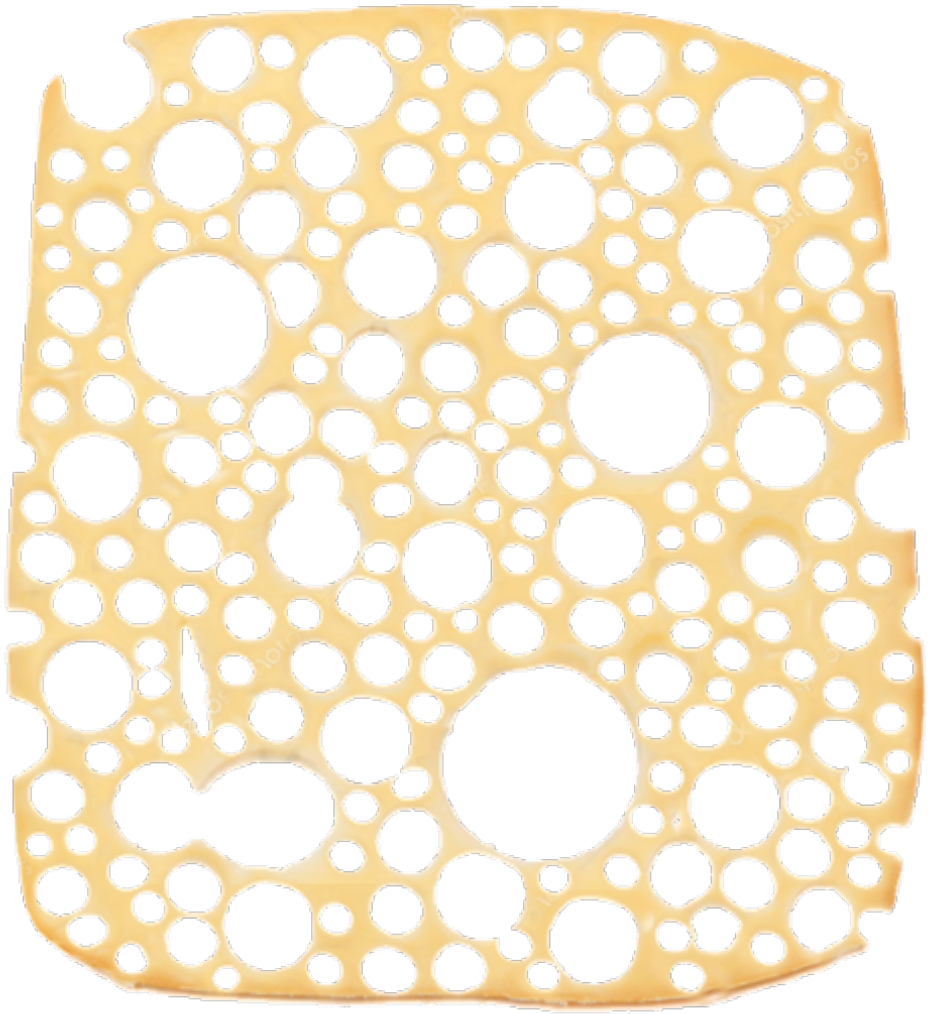
The cheese hypothesis

The cheese hypothesis



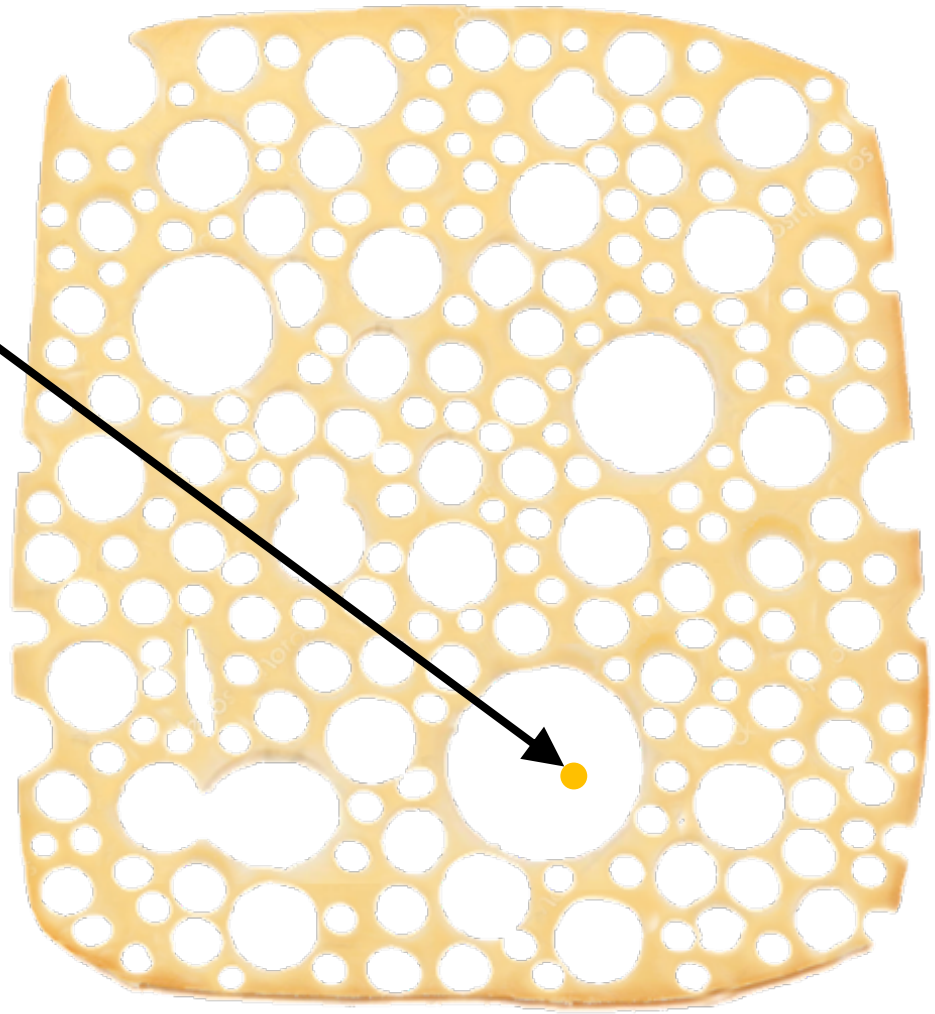
The cheese hypothesis

Original image



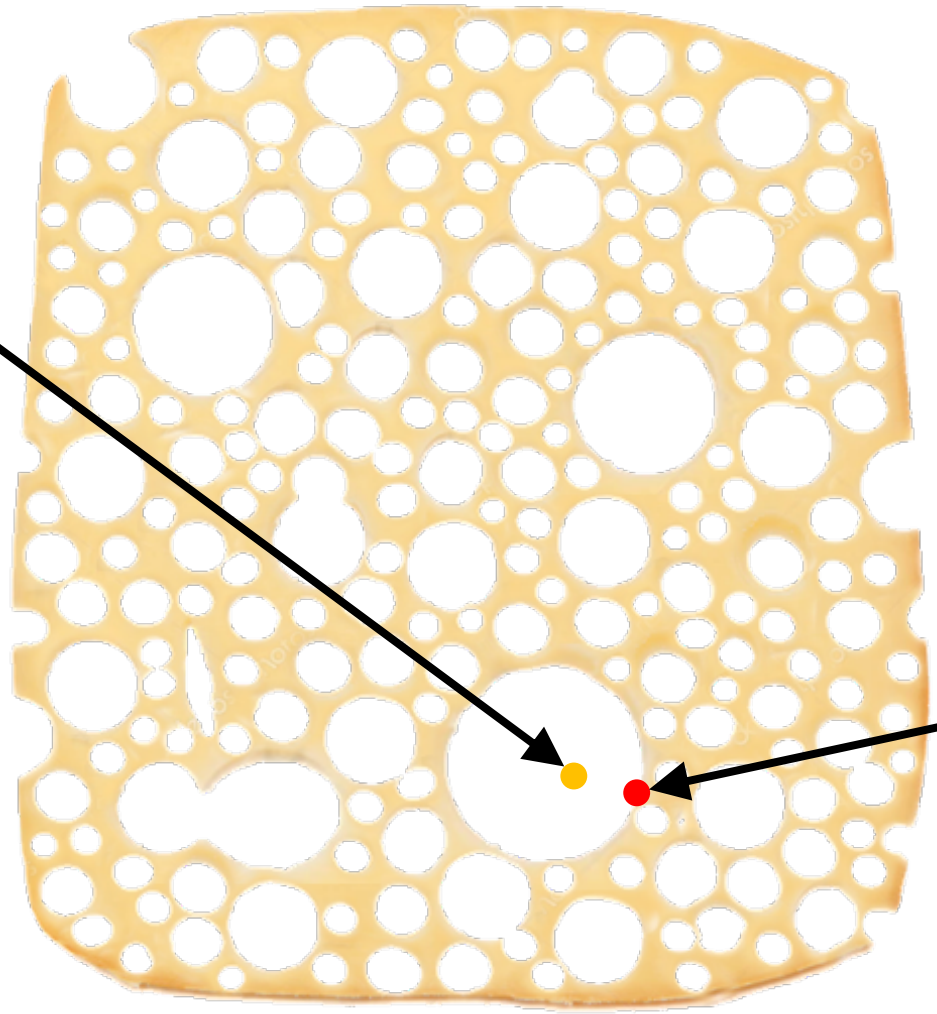
The cheese hypothesis

Original image



The cheese hypothesis

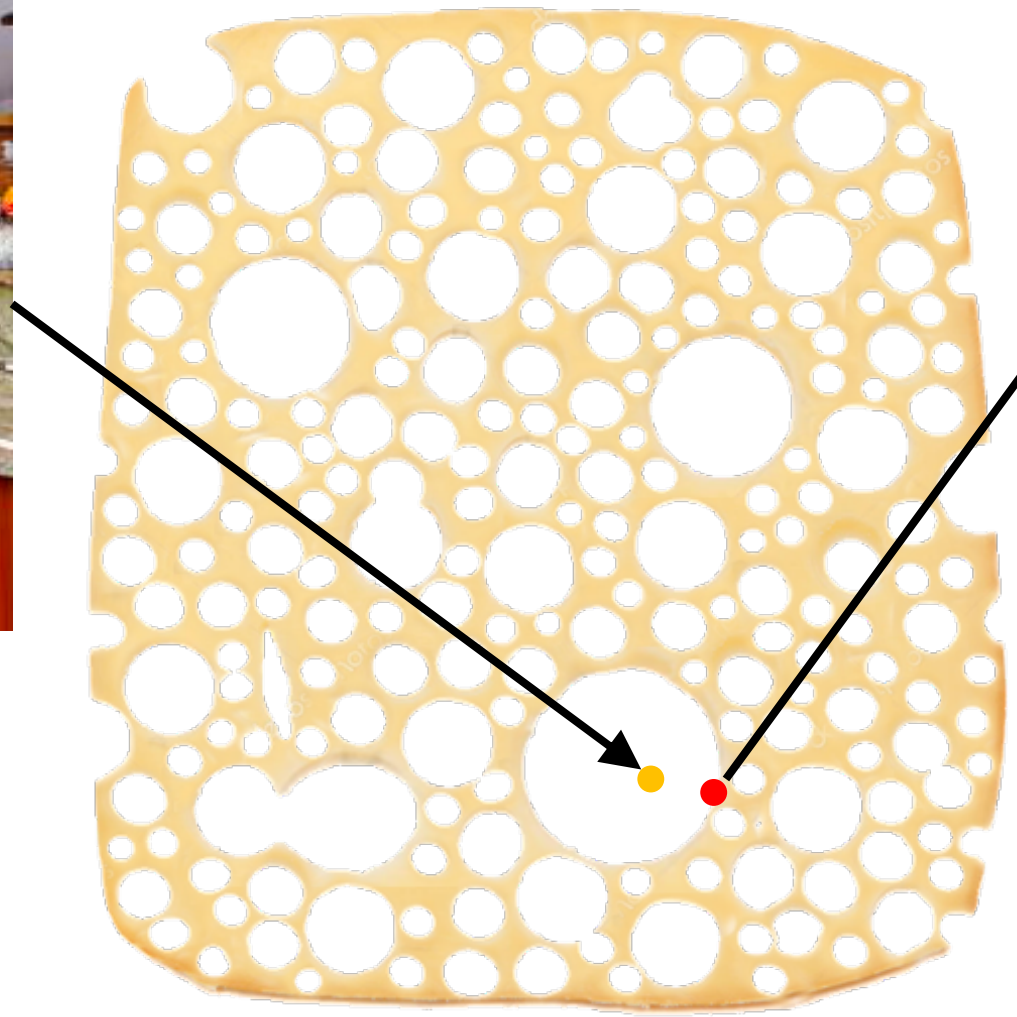
Original image



Optimized z

The cheese hypothesis

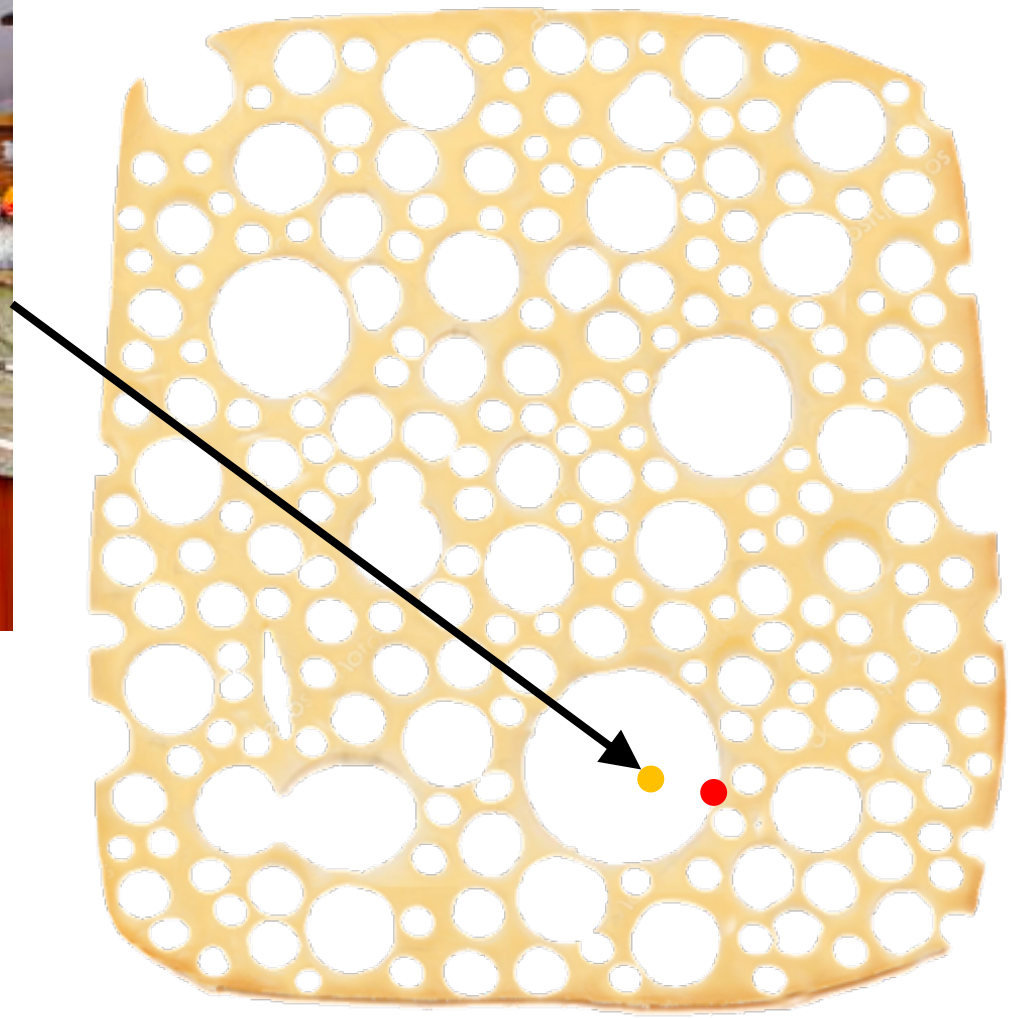
Original image



Optimized z

The cheese hypothesis

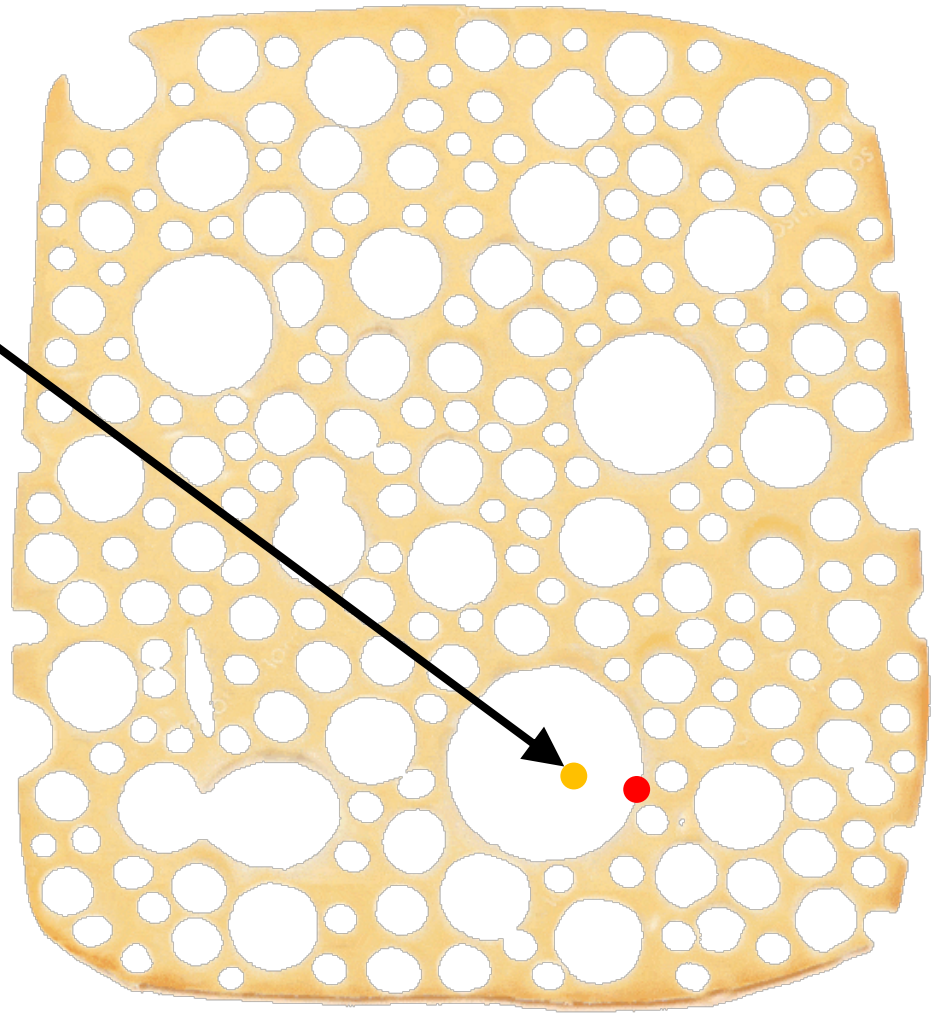
Original image



The cheese hypothesis

Original image

Adapted cheese



Reconstructing my own photo

Original image

Optimized $\hat{z}$

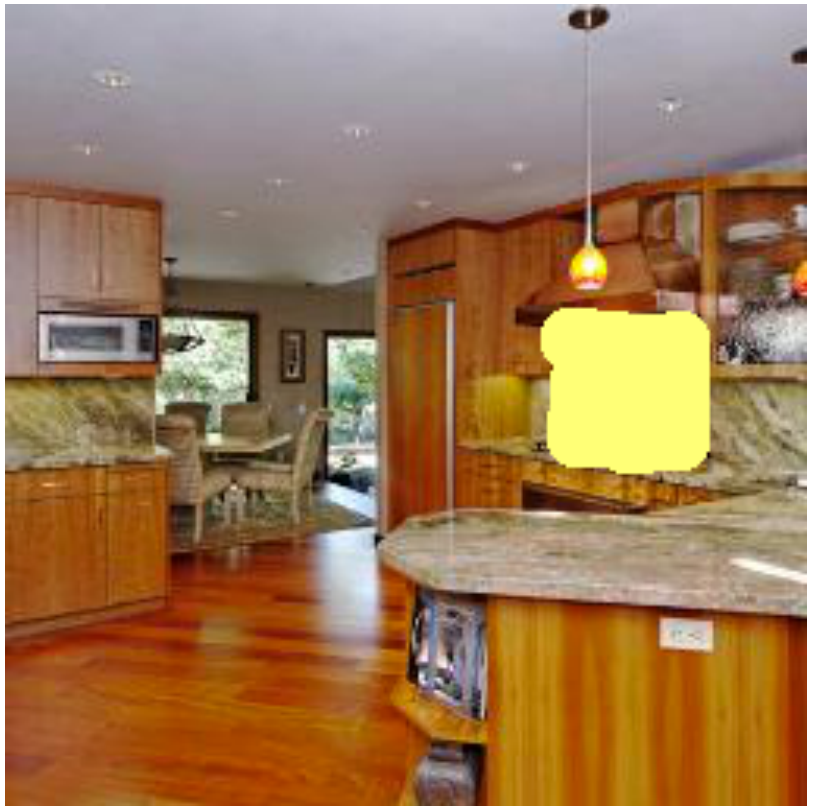
Optimized $\hat{z}$ and $\widehat{\theta}$

Will editing work?

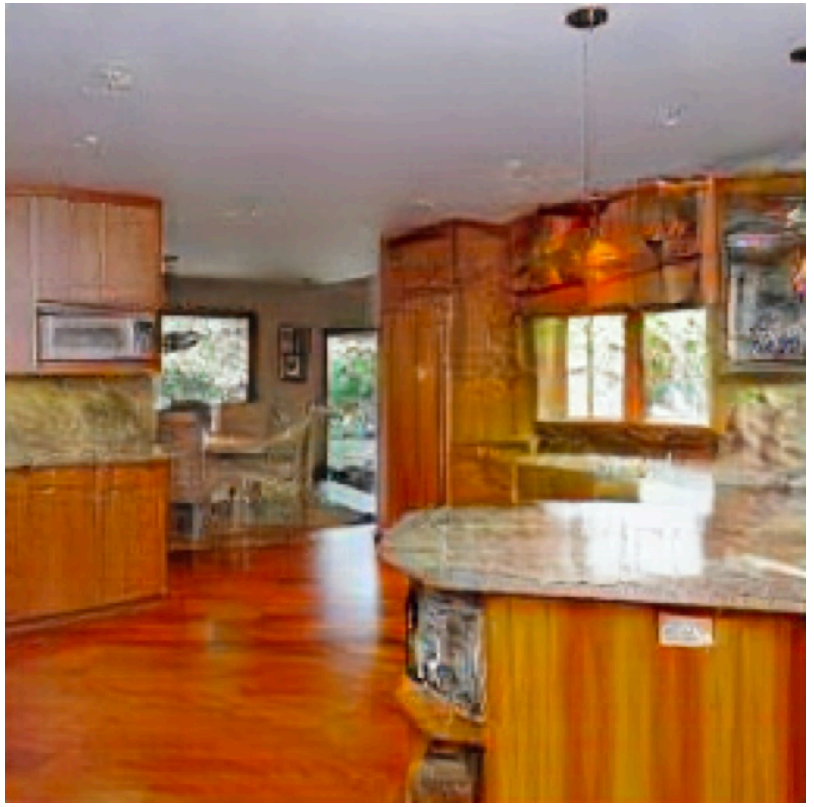
Optimized $\hat{\boldsymbol{z}}$ and $\widehat{\boldsymbol{\theta}}$

Will editing work?

Optimized $\hat{\boldsymbol{z}}$ and $\widehat{\boldsymbol{\theta}}$



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

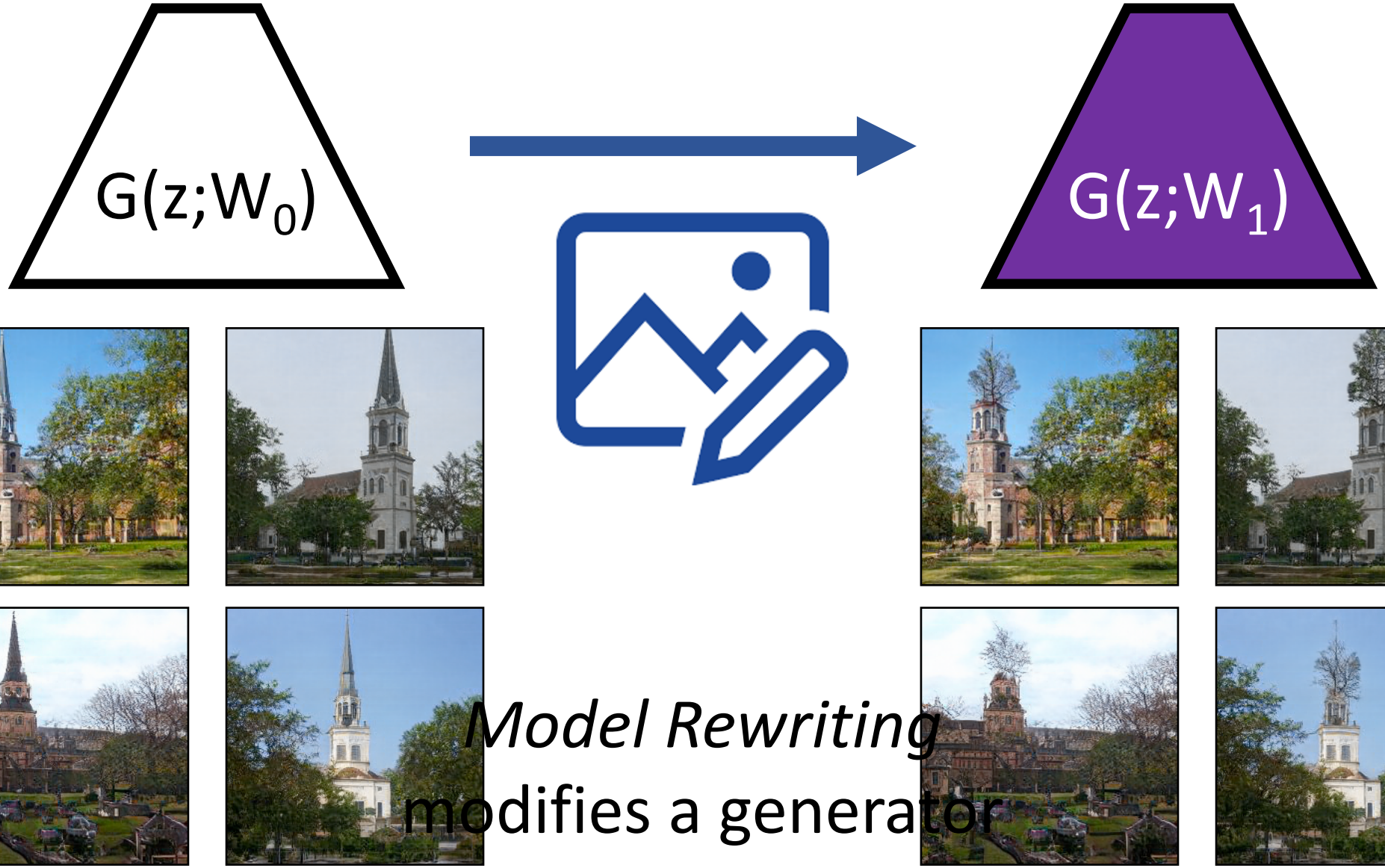
Part 3: Editing a Model

Rewriting a Deep Generative Model [Bau et al., ECCV 2020]

The 1870 Decatur Courthouse Tower Tree

Can we create a model
without a data set?

An image editor can
create a single image


G(z;W0)
G(z;W1)
Model Rewriting
modifies a generator

Rewriting a Deep Generative Model

Search | Toggle Original | Execute Change

Add to Context | Show Context Matches

Rule Selector

Copy | Paste

Goal Selector

Rewriting a Deep Generative Model

Search | Toggle Original | Execute Change

Add to Context | Show Context Matches

Copy | Paste

[StyleGANv2: Karras 2020, LSUN: Yu 2015]

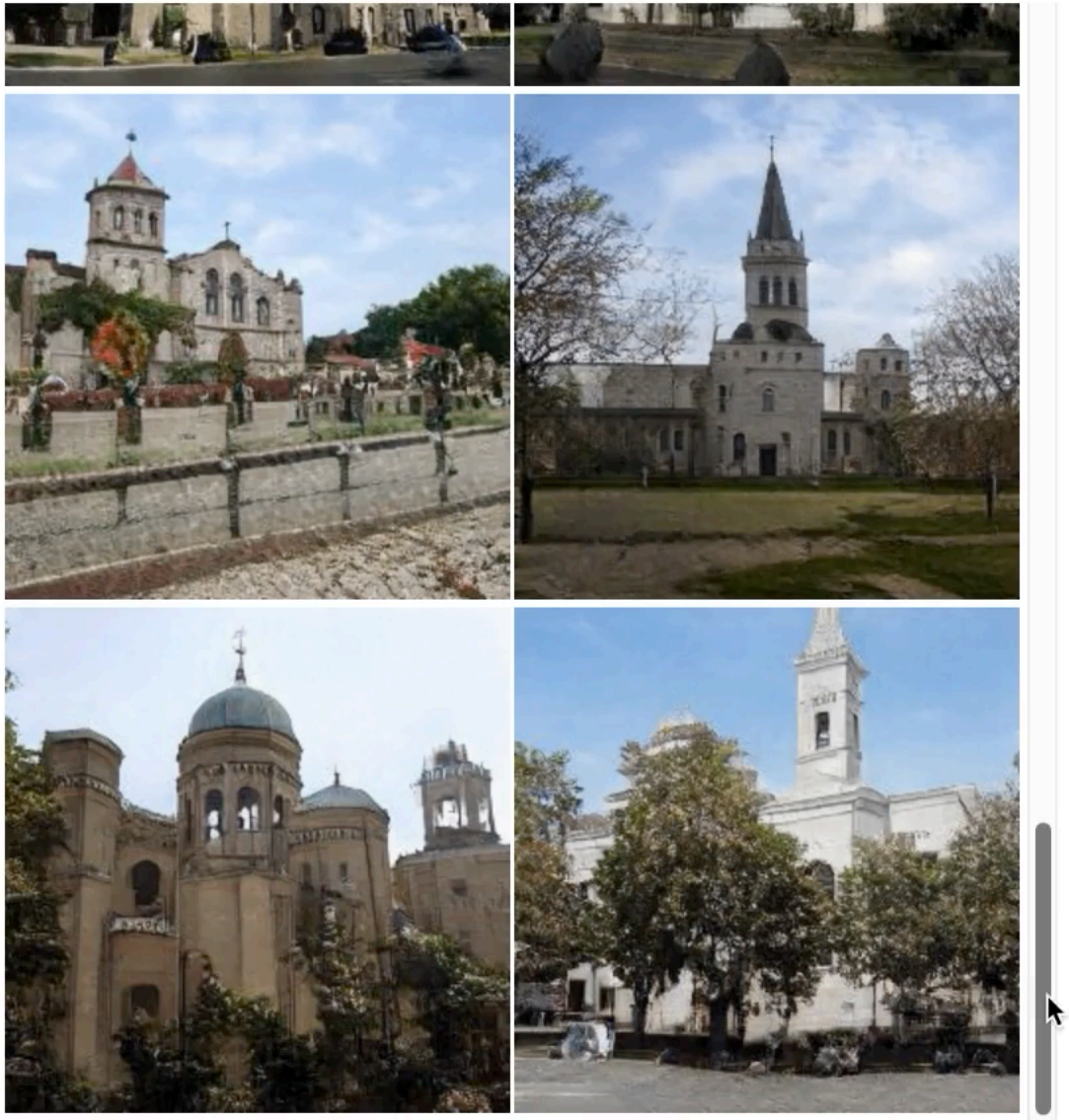
Rewriting a Deep Generative Model

Search | Toggle Original | Execute Change

Add to Context | Show Context Matches

Rule Selector

Copy | Paste



Rewriting a Deep Generative Model

Search | Toggle Original | Execute Change

Add to Context | Show Context Matches

Copy | Paste

Goal Selector

Rewriting a Deep Generative Model

Search | Toggle Original | Execute Change

Add to Context | Show Context Matches

Copy | Paste

Rewriting a Deep Generative Model

Search | Toggle Original | Execute Change

Add to Context | Show Context Matches

Copy | Paste

Changing a model of physics

Progressive GAN models **light reflections**

Has Window

No Window

W_0

Has Reflection

No Reflection

Changing a model of physics

Progressive GAN models **light reflections**

Has Window

Rewritten Model

No Window

NO Reflection!

HAS Reflection!

Changing a model of physics

Progressive GAN models **light reflections**

Has Window

No Window

W_0

Has Reflection

No Reflection

Changing a model of physics

Progressive GAN models **light reflections**

Has Window

Has Reflection

Rewritten Model

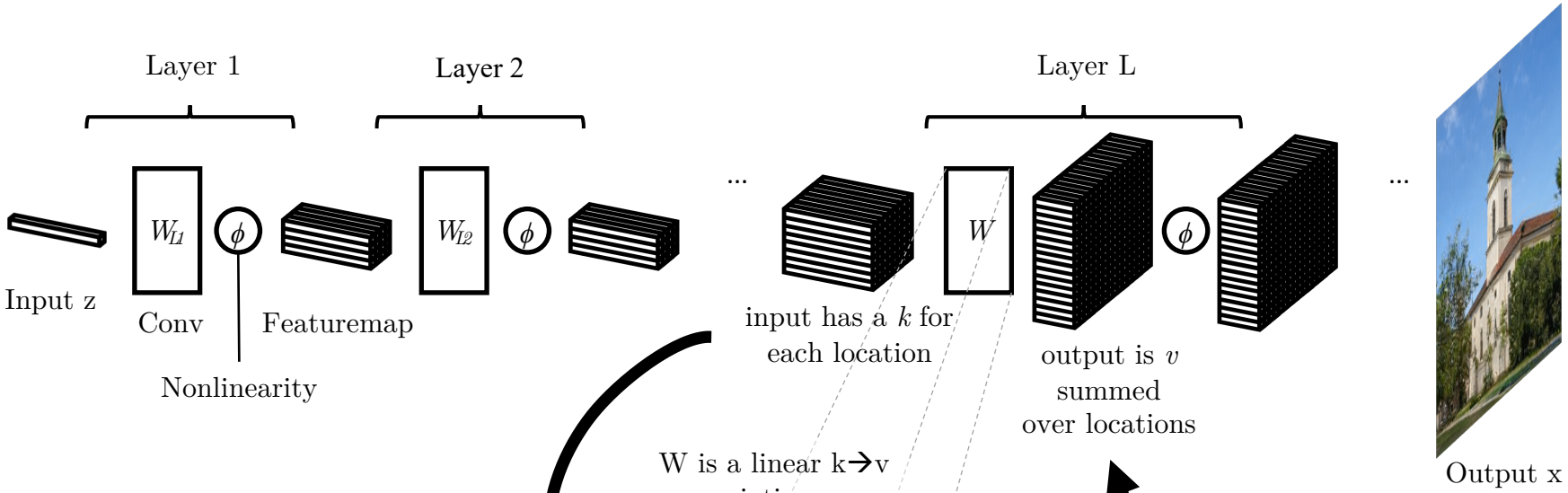
W_0

Changing ONE Rank-1 Rule Reverses Reflections

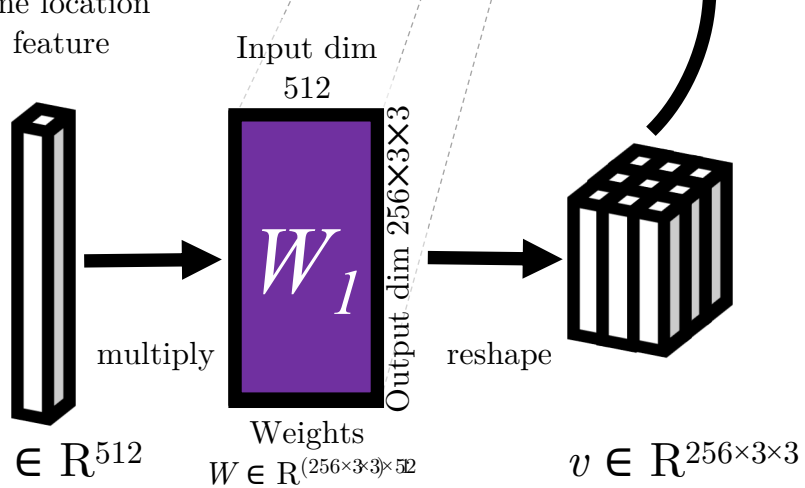
No Window

No Reflection

Hypothesis: Weights act as associative memory



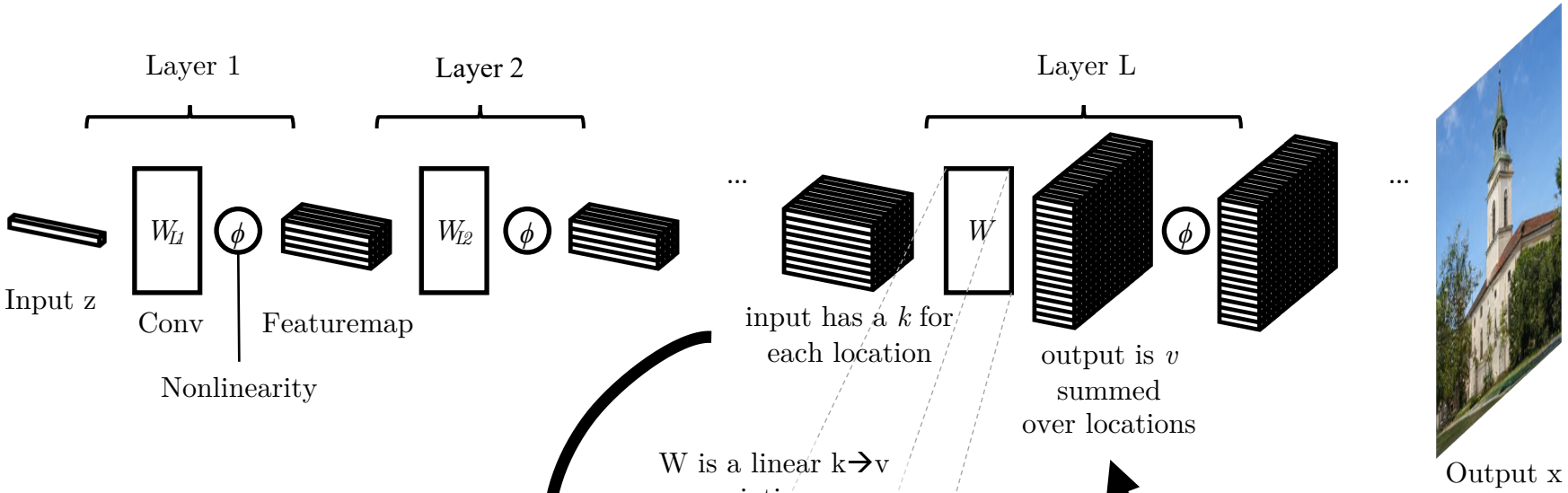
"**Context**" key



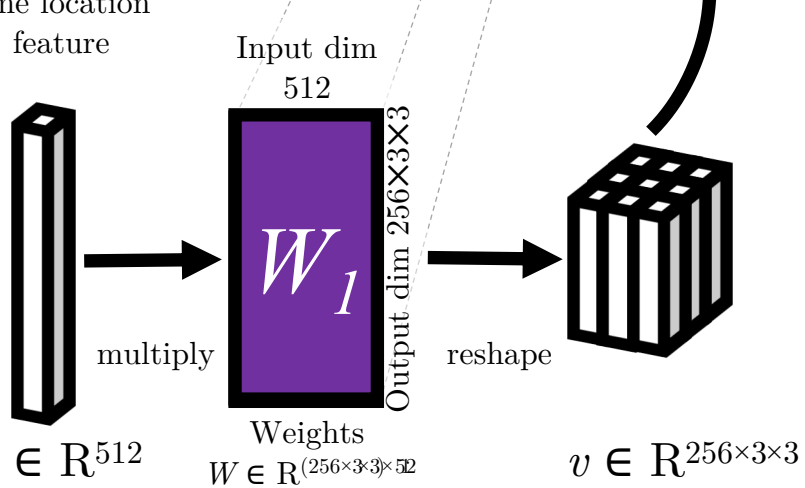
$k_* \rightarrow v_*$

"**Copy-Paste**" goal

Hypothesis: Weights act as associative memory



"**Context**" key



$k_* \rightarrow v_*$

"**Copy-Paste**" goal

1. What the hypothesis implies

Assume: the job of a layer is to recall k→v with minimal error.

$$W_0 \triangleq \arg\min_W \sum_i ||v_i - W k_i||^2$$

Then: weights satisfy Least Squares.

$$W_0 K K^T = V K^T$$

[Kohonen 1972]

2. What an ideal model edit would do

We wish to set $k_* \rightarrow v_*$ while still minimizing error in old k→v

$$W_1 = \arg\min_W ||V - WK||^2$$

$$\text{subject to } v_* = W_1 k_*.$$

This is Constrained Least Squares, and has this solution:

$$W_1 K K^T = V K^T + \Lambda {k_*}^T$$

3. Implications for making an ideal edit

Subtracting LS assumption from CLS solution cancels terms.

$$W_1 K K^T = V K^T + \Lambda k_*^T$$
$$W_0 K K^T = V K^T$$
$$W_1 K K^T = W_0 K K^T + \Lambda k_*^T$$
$$W_1 = W_0 + \Lambda (C^{-1} k_*)^T$$

The solution is a *rank-one update* invariant to v_*

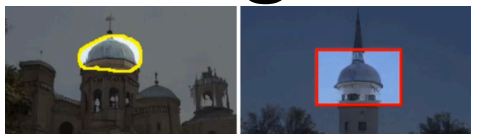
Method: constrained optimization

$$d \triangleq C^{-1} k$$

1. The update direction *d* is the **rule**.

$$k_* \rightarrow v_*$$

2. The copy-paste example is the **goal**.



3. Change *W* in direction *d* only:

$$\Lambda_1 = \underset{\Lambda \in \mathbb{R}^M}{\arg\min} \, ||v_* - f(k_*; W_0 + \Lambda \, d^T)||$$

$$W_1 = W_0 + \Lambda_1 d^T$$

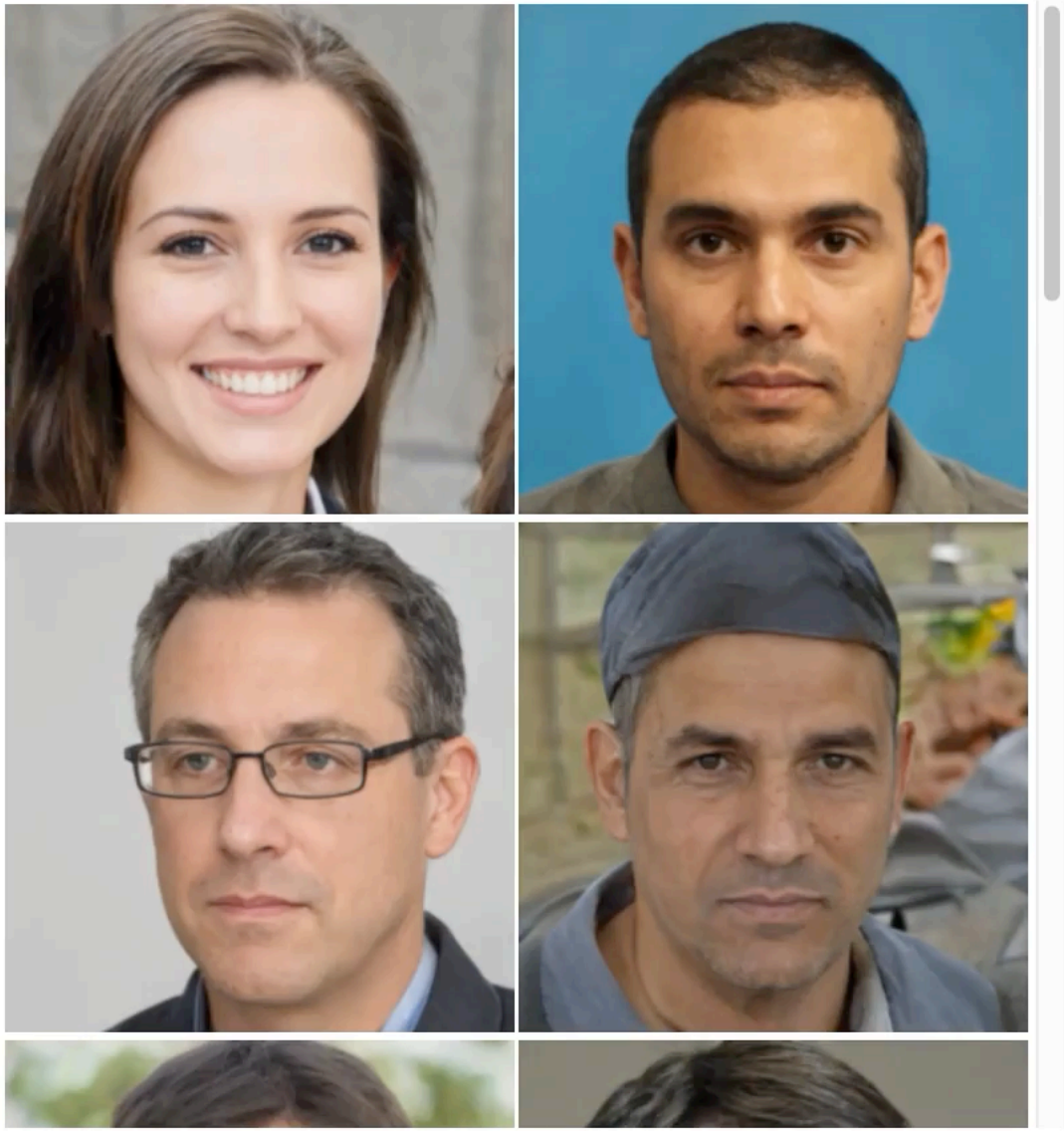
Avoids changing other rules!

Rewriting a Deep Generative Model

Search | Toggle Original | Execute Change

Add to Context | Show Context Matches

Copy | Paste

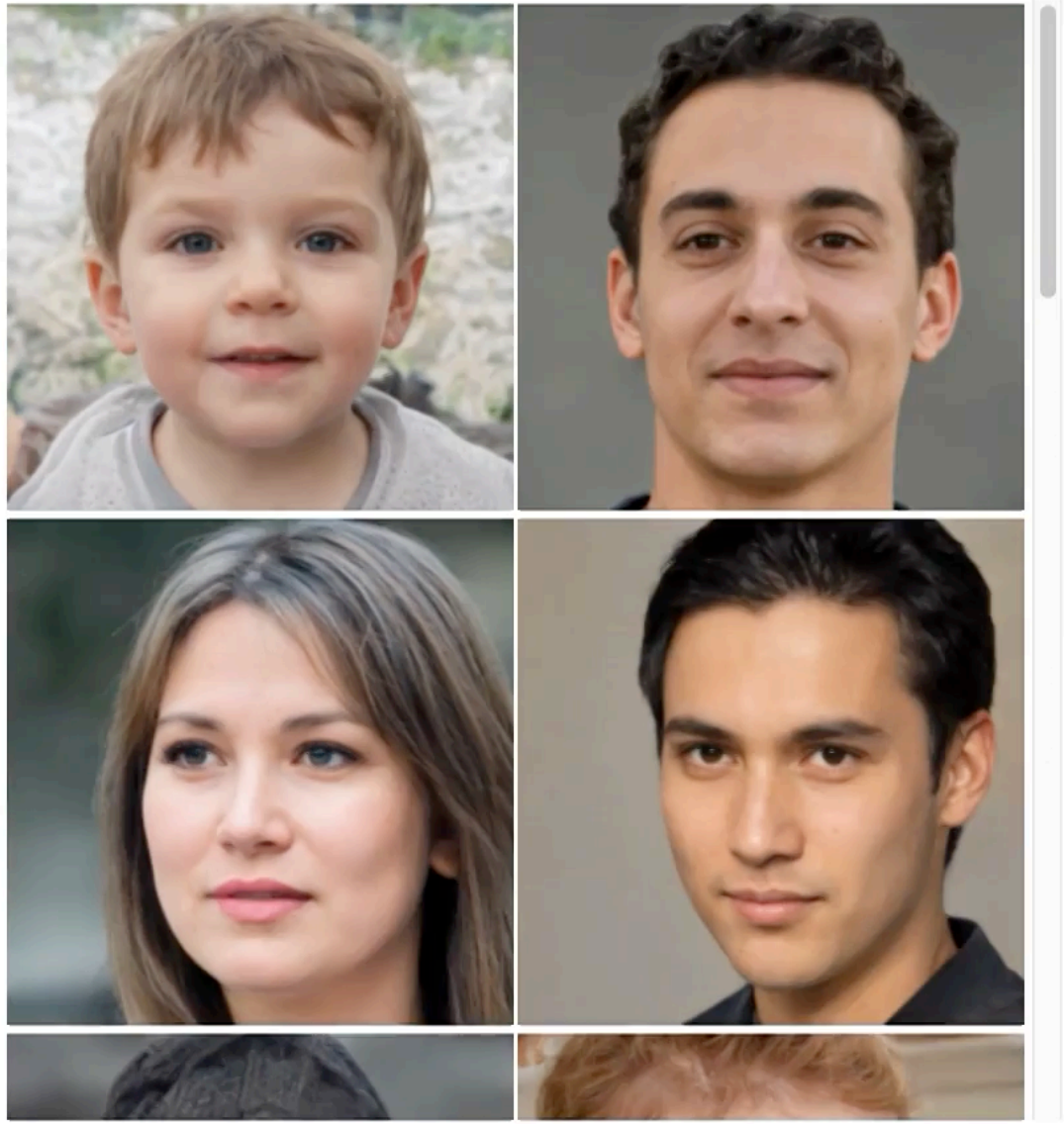


Rewriting a Deep Generative Model

Search | Toggle Original | Execute Change

Add to Context | Show Context Matches

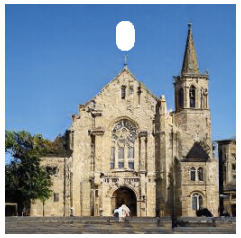
Copy | Paste



Doors in the sky? A difficult case.

Rule context: patch of sky over rooftop



Rule goal: put a door there instead



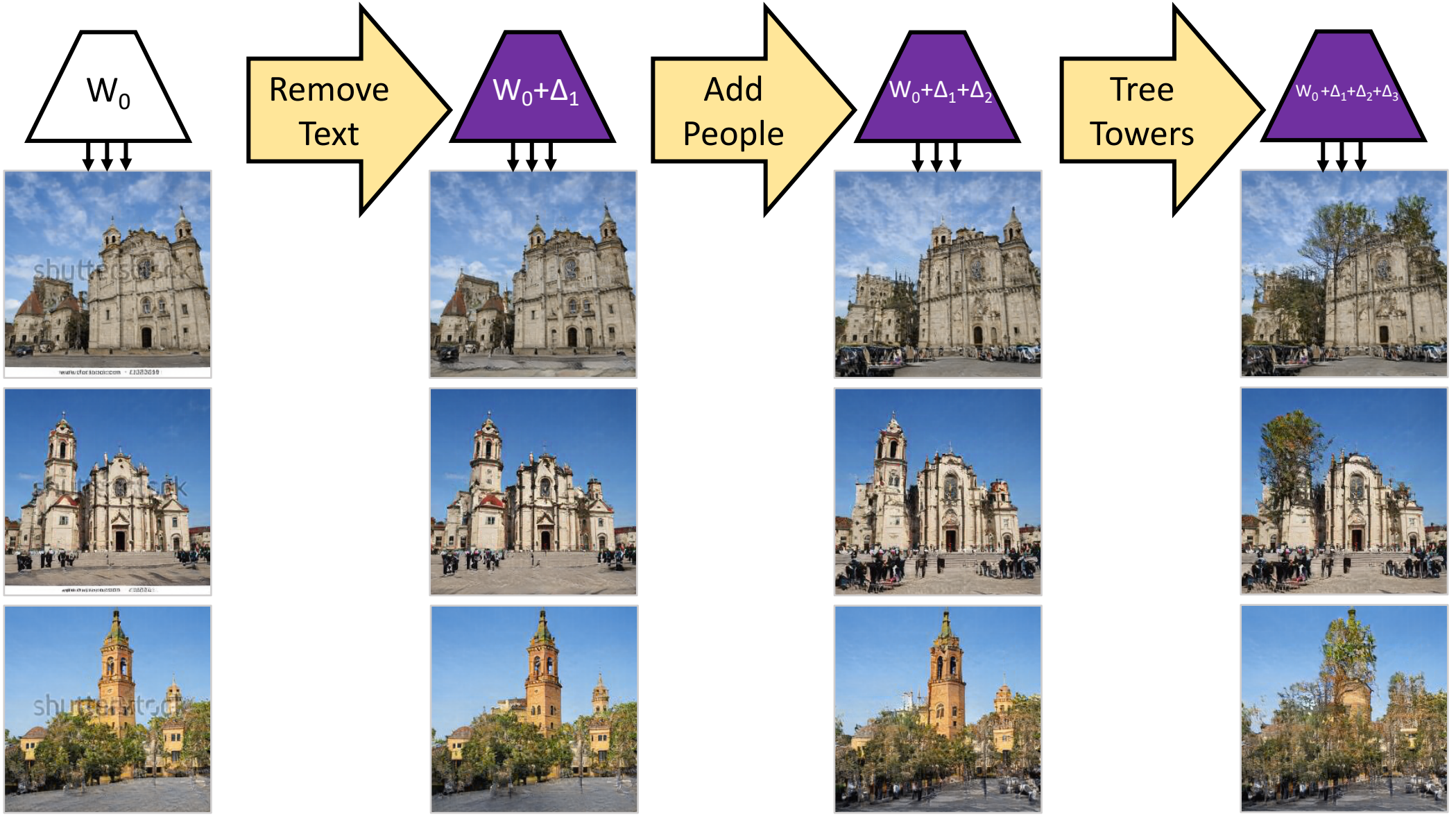


Rewriting a Deep Generative Model

Search | Toggle Original | Execute Change

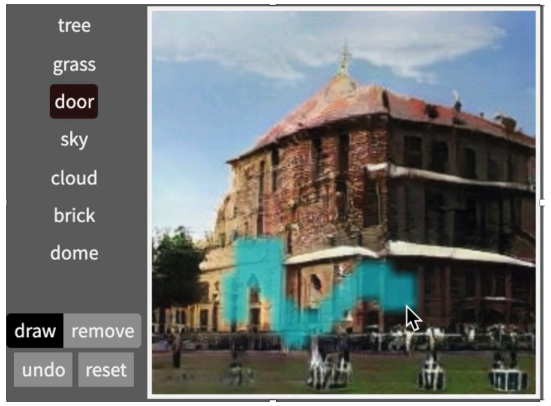
Add to Context | Show Context Matches

Copy | Paste


W_0
Remove Text
$W_0+\Delta_1$
Add People
$W_0+\Delta_1+\Delta_2$
Tree Towers
$W_0+\Delta_1+\Delta_2+\Delta_3$
Original model
Rewritten Models

Dissection

Understanding Structure



GAN Paint

Exploiting Structure

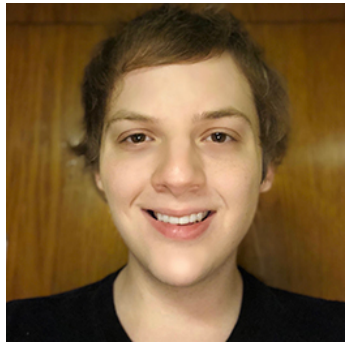
Photo Manipulation

Adapting to one Image

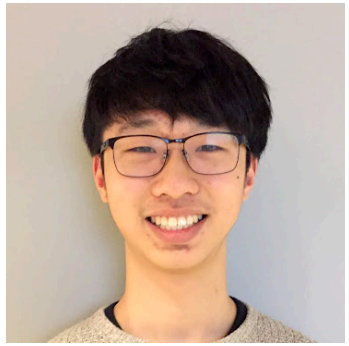
Model Rewriting

Editing Generalizable Structure

Hendrik



William

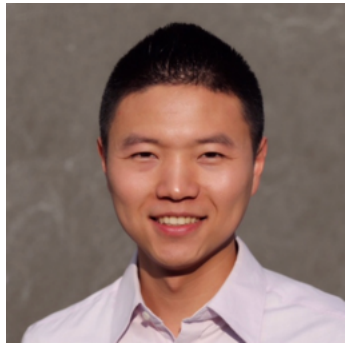


Steven

Jonas

Tongzhou

Agata



Bolei

Jun-Yan

Antonio

Thank you!

http://gandissect.csail.mit.edu
http://rewriting.csail.mit.edu