Rosetta Neurons: Mining the Common Units in a Model Zoo

Amil Dravid*, Yossi Gandelsman*, Alexei A. Efros, Assaf Shocher
Northwestern, UC Berkeley, UC Berkeley, UC Berkeley, Google

ICCV 2023

PRESENTER: JIAHANG ZHANG

2023/10/24

Outline

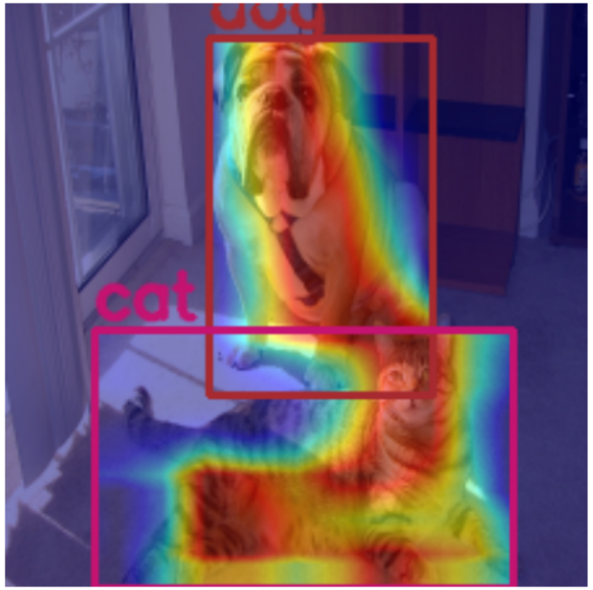
Background

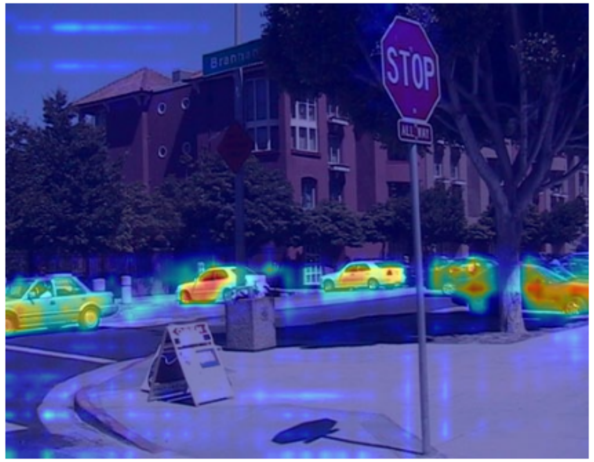
Visualizing deep representations

- CAM (Class Activation Map)
- Grad-CAM

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k A^k\right)$$

$$\alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$$



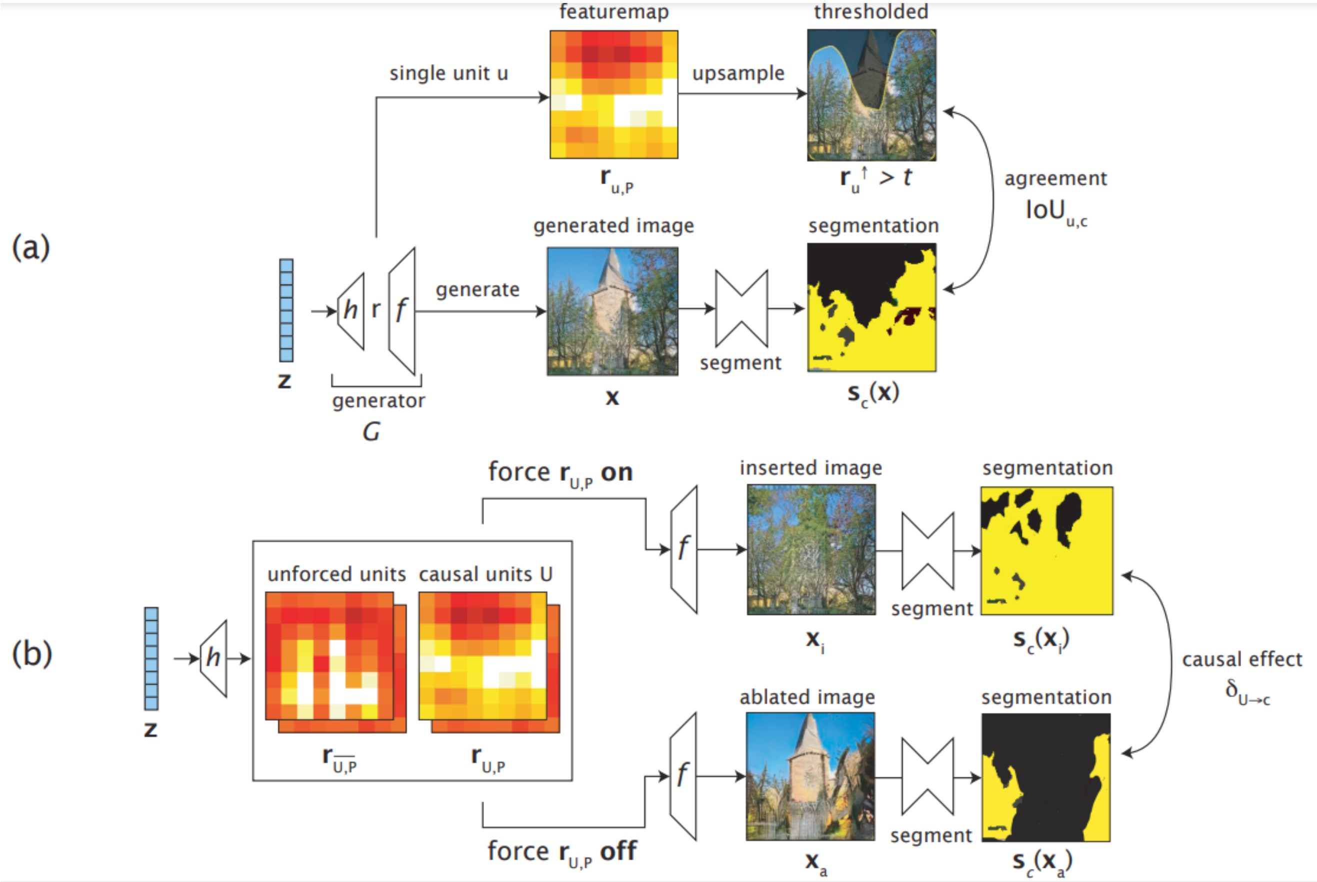


Background

Visualizing deep representations

GAN DISSECTION:
VISUALIZING AND UNDERSTANDING
GENERATIVE ADVERSARIAL NETWORKS

David Bau[1,2], Jun-Yan Zhu[1], Hendrik Strobelt[2,3], Bolei Zhou[4],
Joshua B. Tenenbaum[1], William T. Freeman[1], Antonio Torralba[1,2]
[1]Massachusetts Institute of Technology, [2]MIT-IBM Watson AI Lab,
[3]IBM Research, [4]The Chinese University of Hong Kong

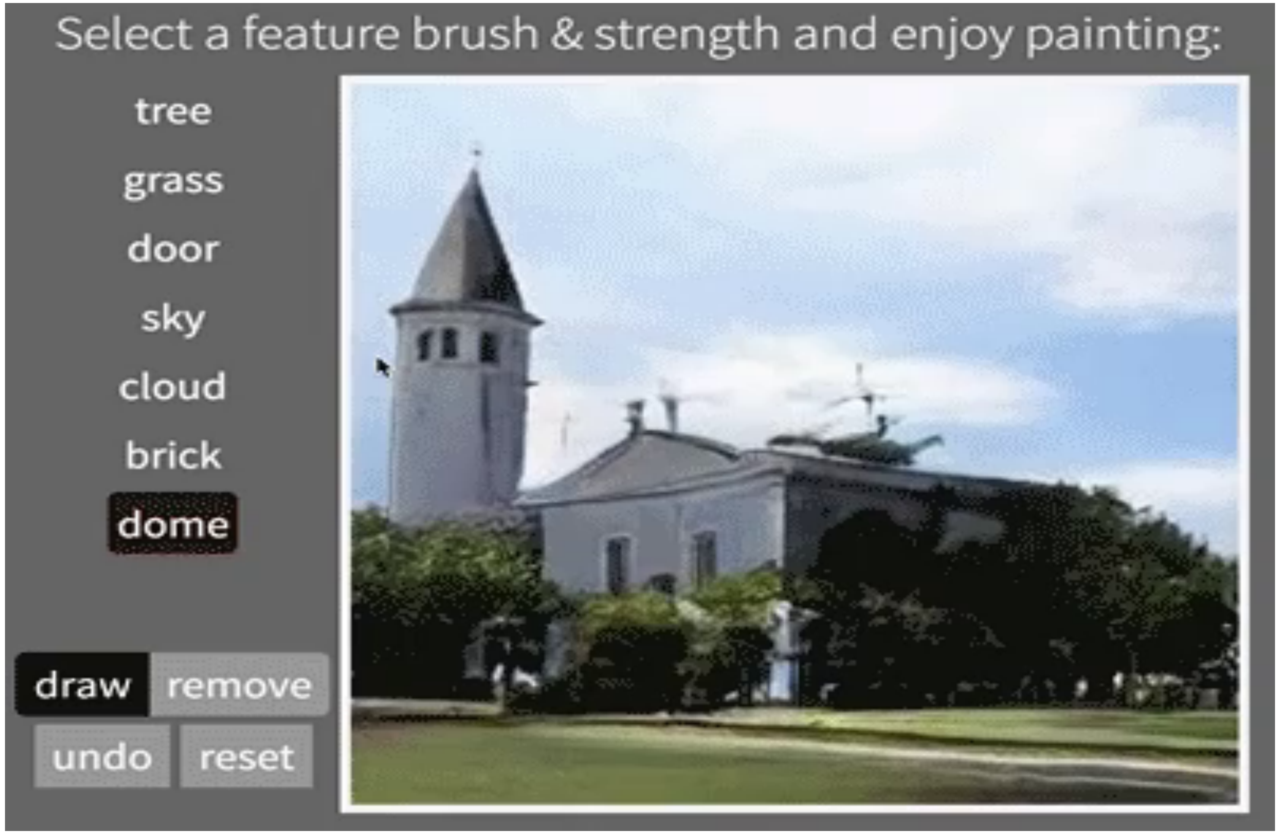


Background

Visualizing deep representations

GAN DISSECTION:
VISUALIZING AND UNDERSTANDING
GENERATIVE ADVERSARIAL NETWORKS

David Bau[1,2], Jun-Yan Zhu[1], Hendrik Strobelt[2,3], Bolei Zhou[4],
Joshua B. Tenenbaum[1], William T. Freeman[1], Antonio Torralba[1,2]
[1]Massachusetts Institute of Technology, [2]MIT-IBM Watson AI Lab,
[3]IBM Research, [4]The Chinese University of Hong Kong



Background

- Limitations
  - Need Label
  - Focus on One Model

- ***Can we make connections between different models directly through the neurons inside the model?***

Outline

Method

- **Motivation:**
  - Do different neural networks, trained for various vision tasks, share some common representations?
  - We seek to identify and match units that express similar concepts across *different models*.
  - We call them ***Rosetta Neurons Techniques***.

Method

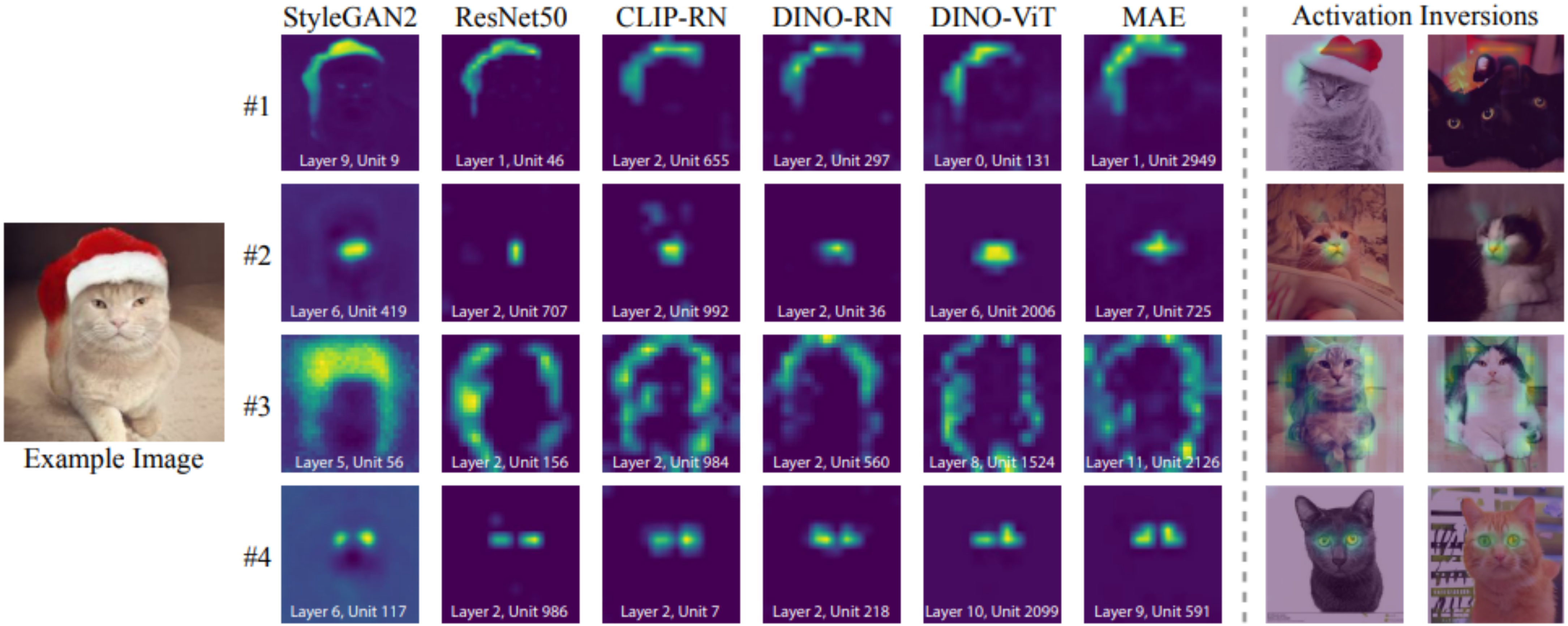


Figure 1: Mining for "Rosetta Neurons." Our findings demonstrate the existence of matching neurons across different models that express a shared concept (such as object contours, object parts, and colors). These concepts emerge without any supervision or manual annotations. We visualize the concepts with heatmaps and a novel inversion technique (two right columns).

Method - 1 Mining Common Units

- **Settings:**
  - **Two models,** $F^{(1)}, F^{(2)}$
    - **A generative model and a discriminative model**
  - **Connecting the two models:**
    - $F^{(1)} \rightarrow I \rightarrow F^{(2)}$
  - $F^{(1)j}_{i,x}$
    - **The x location of j-th intermediate activation map when applied $F^{(1)}$ to the i-th input image**

Method - 1 Mining Common Units

- **Filtering "best buddies" pairs:**
  - **select the pairs that are nearest neighbors**

$$KNN(F^{(a)j}, F^{(b)act}; K) = \underset{q_1 \ldots q_K \subseteq F^{(b)act}}{\operatorname{argmin}} \sum_{k=1}^{K} d(F^{(a)j}, q_k)$$

  - **the distance is defined as the *Pearson correlation:***

$$d(F^{(1)j}, F^{(2)k}) = \frac{\sum_{i,x} \left(F_{i,x}^{(1)j} - \overline{F^{(1)j}}\right)\left(F_{i,x}^{(2)k} - \overline{F^{(2)k}}\right)}{\sqrt{var(F^{(1)j}) \cdot var(F^{(2)k})}} \quad (2)$$

Method - 1 Mining Common Units

- **Filtering "best buddies" pairs:**
  - **select the pairs that are mutual nearest neighbors**
  - **the distance is defined as the *Pearson correlation***
  - **we have found the similar activation maps across different models!**

$$
\begin{aligned}
BB(F^{(1)}, F^{(2)}; K) = \{(j, k) | \\
F^{(1)k} \in KNN(F^{(2)j}, F^{(1)act}; K) \\
\wedge \;\; F^{(2)j} \in KNN(F^{(1)k}, F^{(2)act}; K)\}
\end{aligned}
$$

Method - 1 Mining Common Units

- **Cluster them!**
  - **Merging units between different models to obtain Rosetta units:**

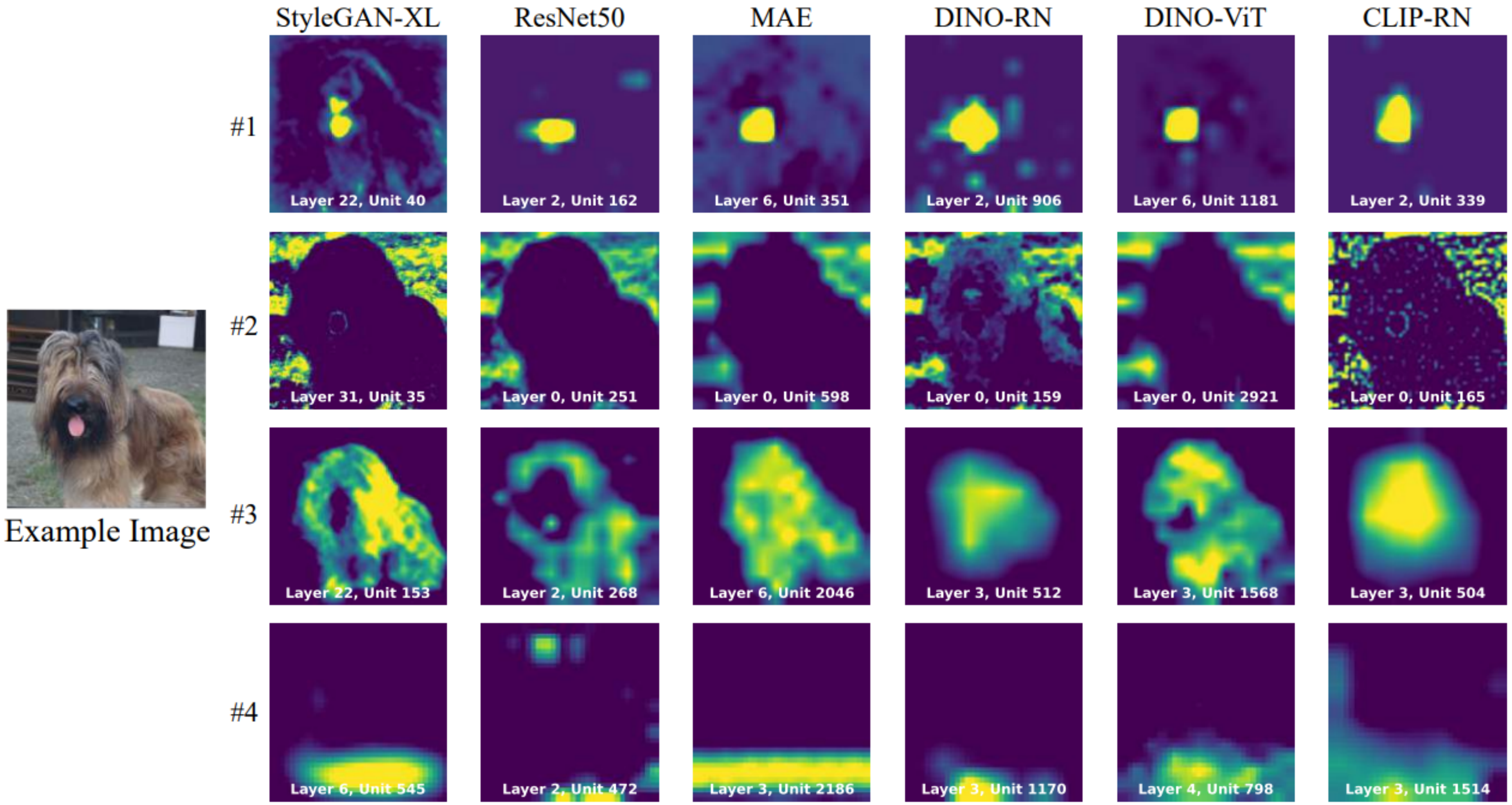
$$R(G, D_1 ... D_m) = \{(j, k_1, ..., k_m) | \forall i : (j, k_i) \in BB(G, D_i)\}$$

  - **Cluster the similar Rossetta units:**

two Rosetta Neurons in R to belong to the same cluster if their corresponding units in G are in $BB(G, G; K)$.

Method - 1 Mining Common Units

Cluster Results (Rosetta Neuron Dictionary)



Method - 1 Mining Common Units

Cluster Results (Rosetta Neuron Dictionary)

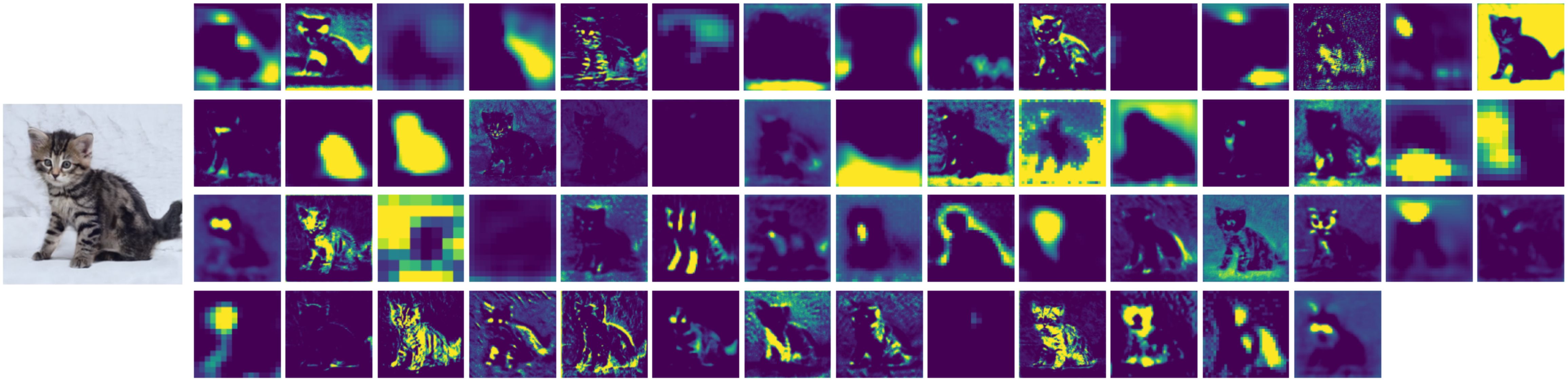
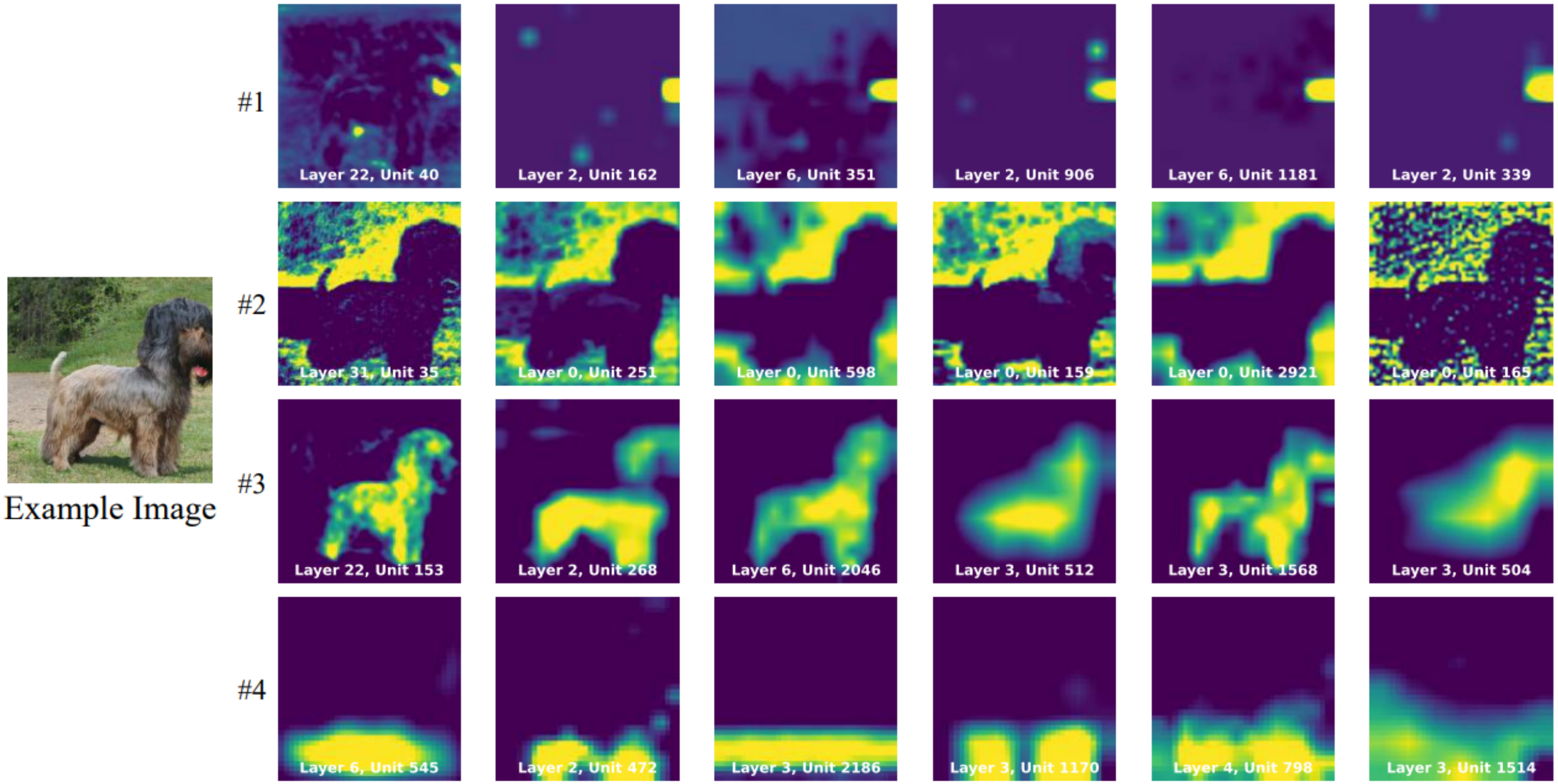


Figure 13: All the concepts for LSUN-cats. Shown for one StyleGAN2 generated image.

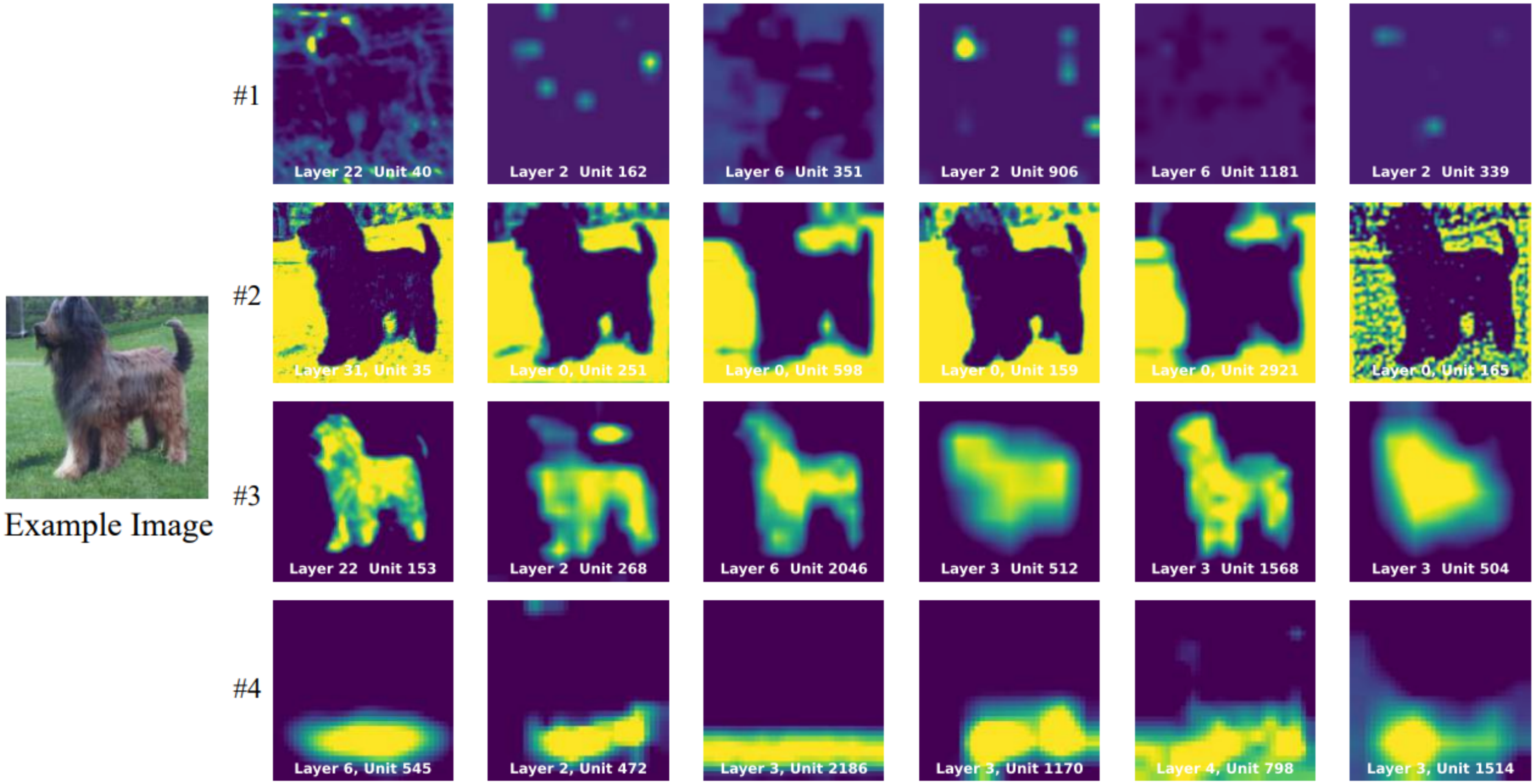
Method - 1 Mining Common Units

Cluster Results (Rosetta Neuron Dictionary)



Method - 1 Mining Common Units

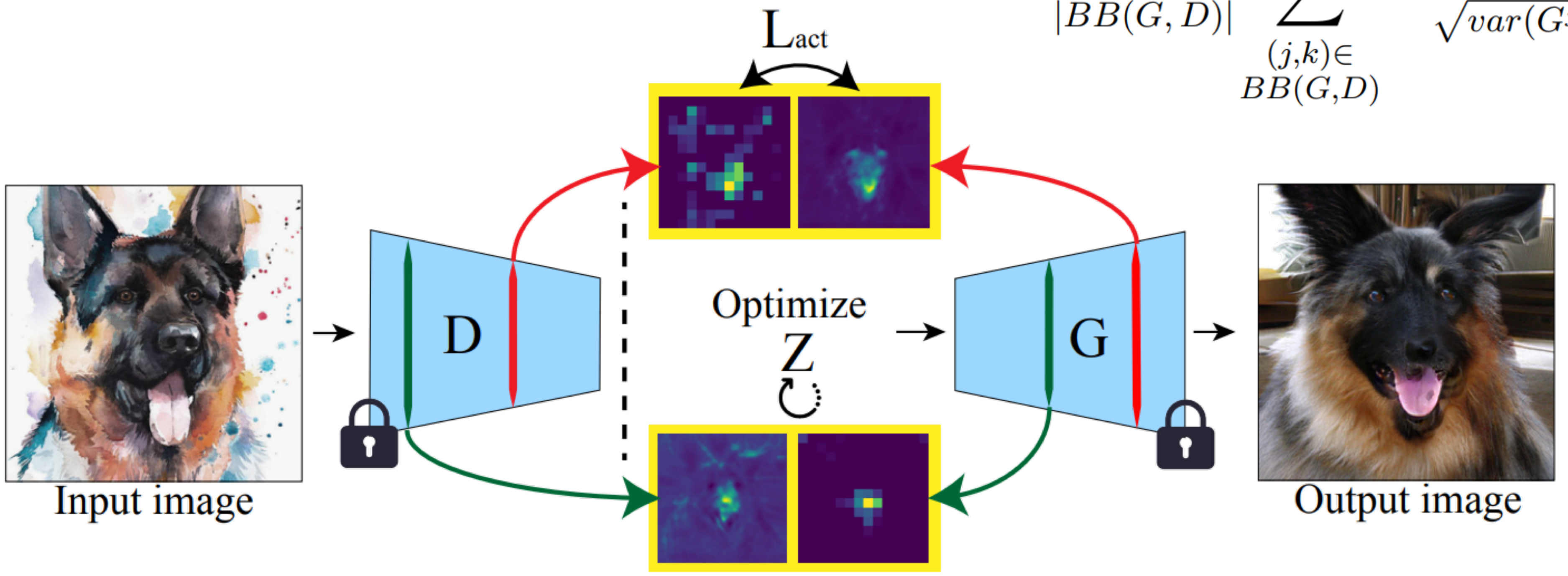
Cluster Results (Rosetta Neuron Dictionary)



Method - 2 Visualize the Unit (Application)

- Rosetta Neurons-Guided Inversion

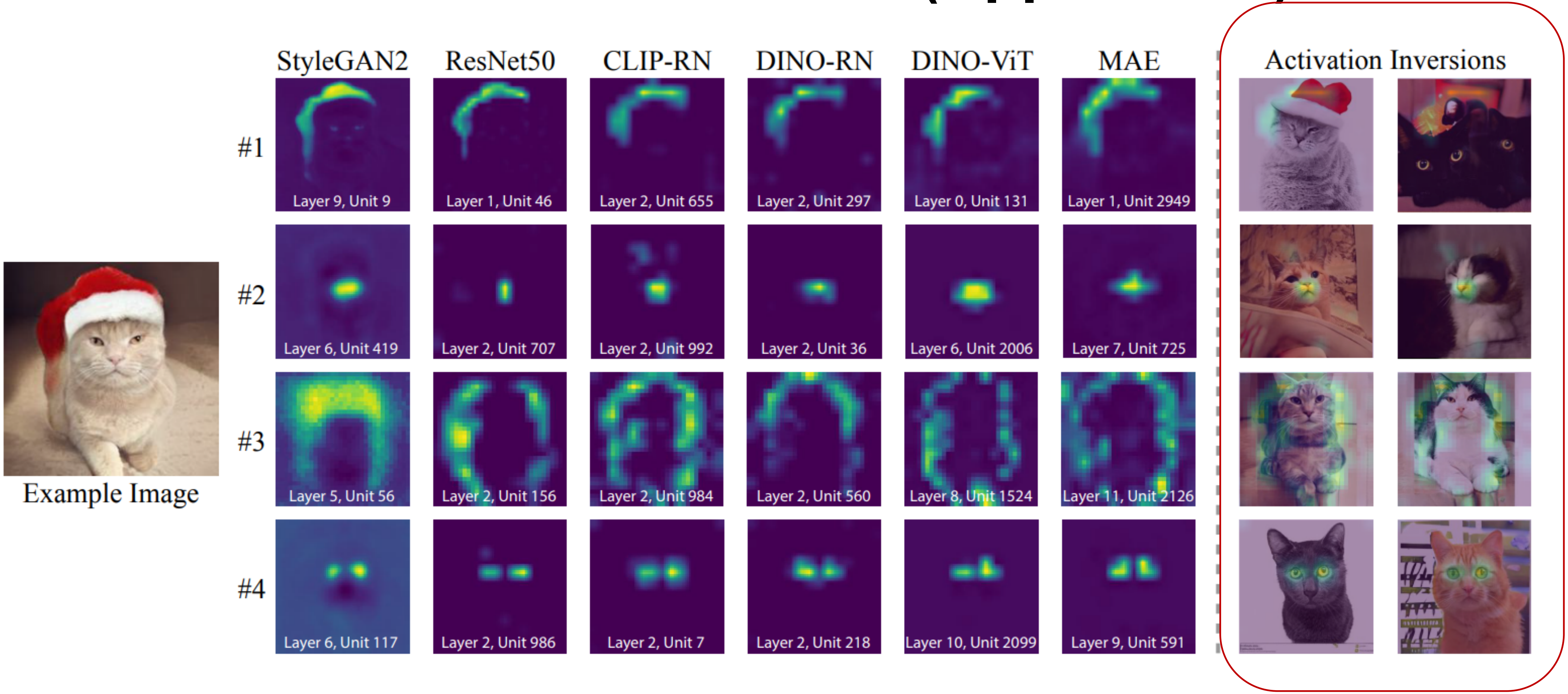
$$L_{act}(z, I_v) = \frac{1}{|BB(G,D)|} \sum_{\substack{(j,k)\in \\ BB(G,D)}} \frac{\sum_x \left(G_x^j - \overline{G^j}\right)\left(D_x^k - \overline{D^k}\right)}{\sqrt{var(G^j)\cdot var(D^k)}}$$



Method - 2 Visualize the Unit (Application)

- Rosetta Neurons-Guided Inversion - Results
  - The corresponding visual concepts will be consistent
    - emerge or disappear
    - property alignment
  - Consistency in images generated over the regions that these neurons focus on

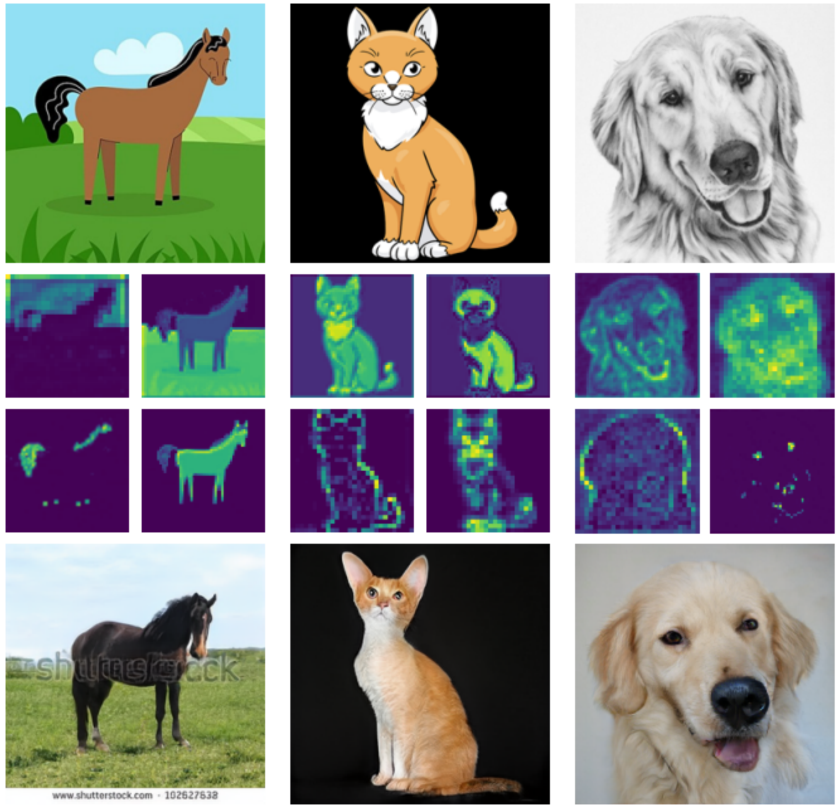
Method - 2 Visualize the Unit (Application)



Method - 2 Visualize the Unit (Application)

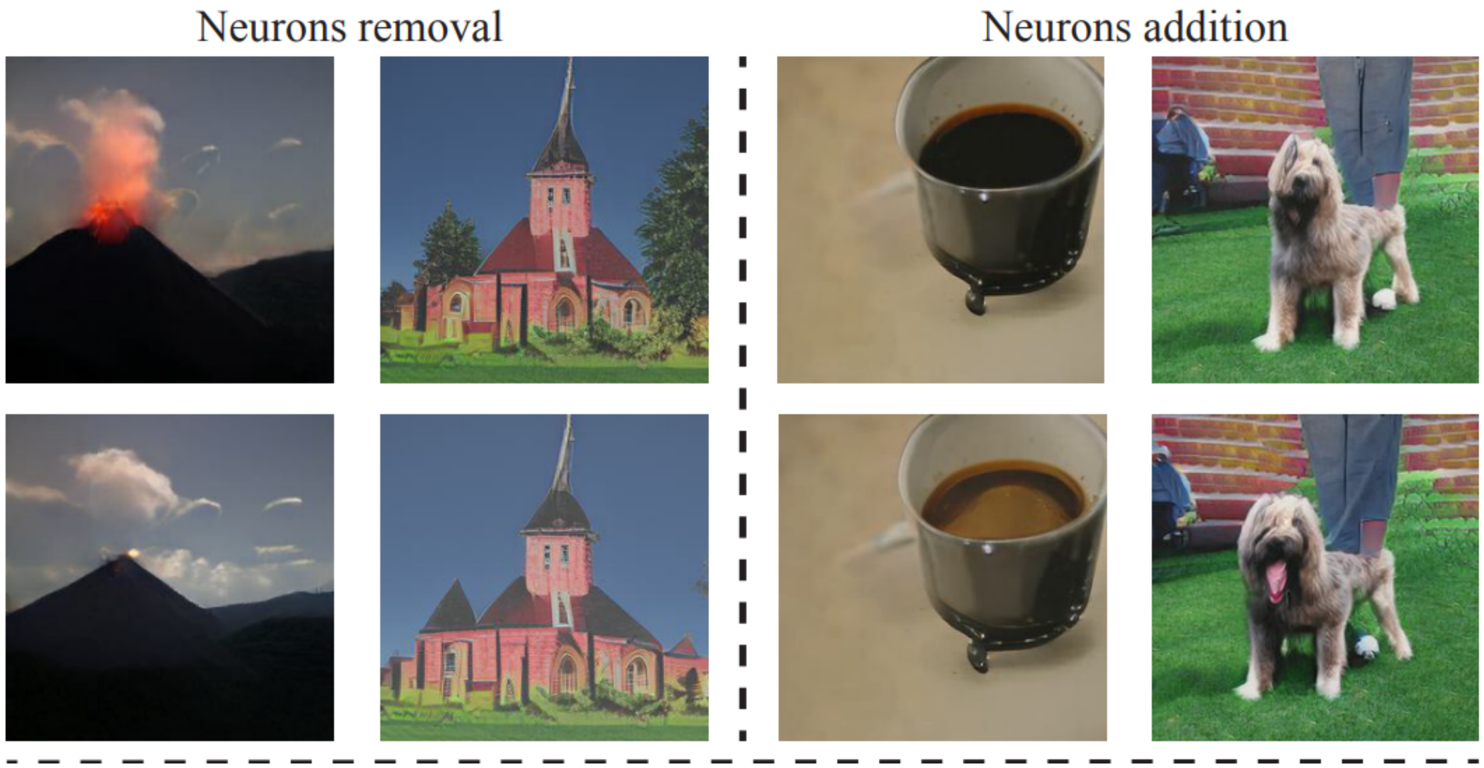
- Rosetta Neurons-Guided Inversion - Results
  - We can perform invertion to generate out-of-distribution images

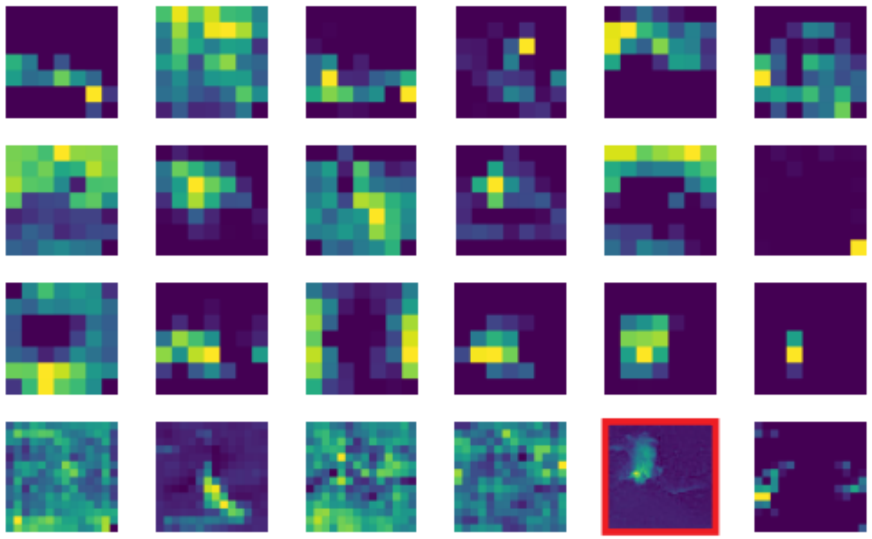
Figure 6: Cross-class image-to-image translation. Rosetta Neurons guided inversion of input images (top row) into a StyleGAN2 trained on LSUN cats [35], allows us to preserve the pose of the animal while changing it from dog to cat (bottom row). See



Method - 2 Visualize the Unit (Application)

- **Rosetta Neurons-Guided Inversion - Results**





Method - 2 Visualize the Unit (Application)

Rosetta Neurons-Guided Inversion - Results

Inverting in-distribution images

$$\arg\min_{z}(L_{rec}(G(z), I_v) + \alpha L_{reg}(z) - \beta L_{act}(z, I_v))$$

	PSNR ↑	SSIM ↑	LPIPS ↓
Perceptual loss	13.99	0.340	0.48
+DINO matches	15.06	0.360	0.45
+CLIP matches	15.20	0.362	**0.44**
+All matches	**15.42**	**0.365**	0.46

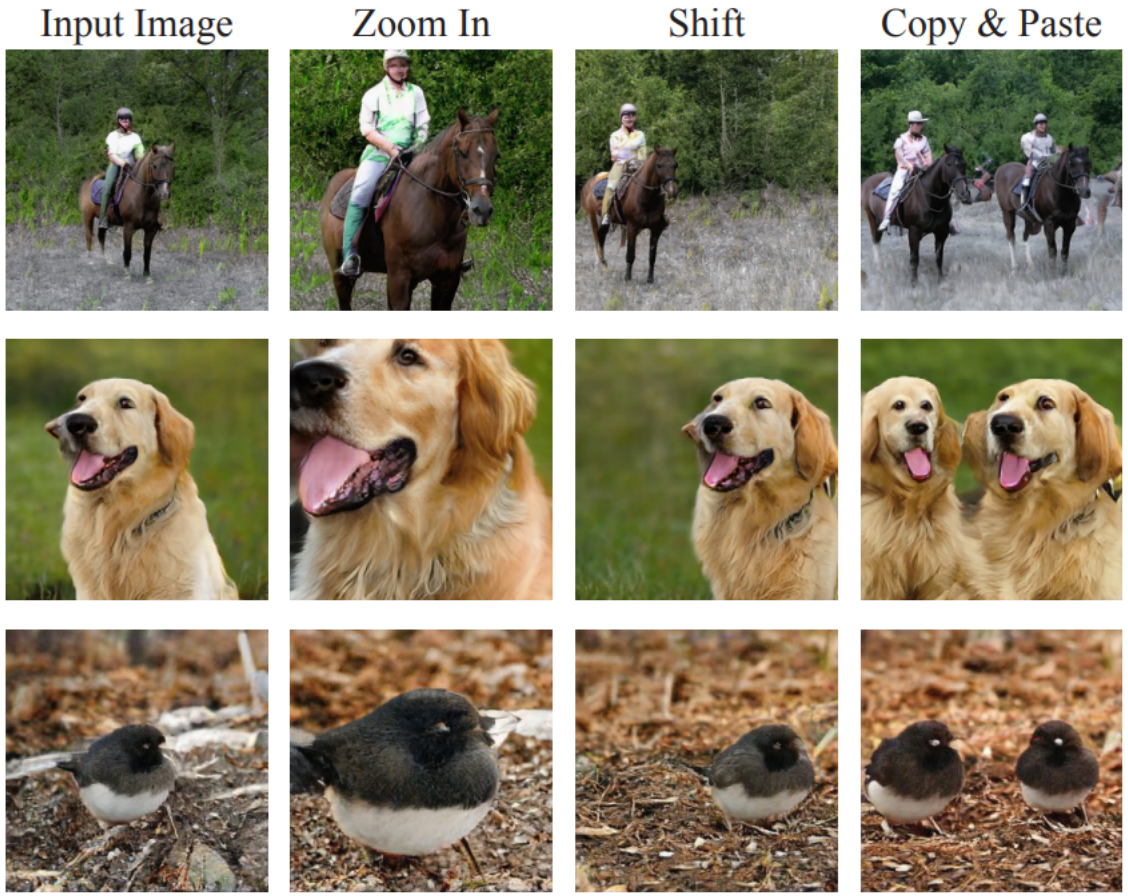
Table 1: Inversion quality on ImageNet. We compare the inversion quality for StyleGAN-XL when Rosetta Neurons guidance is added, for 3 sets of matches - StyleGAN-XL & DINO-RN, StyleGAN-XL & CLIP-RN and all the models from figure 3.

Method - 2 Visualize the Unit (Application)

- Rosetta Neurons Guided Editing
  - Change the activation map of R units
    - Zoom-in, Shift, Copy & Paste,
  - Re-optimize the latent z to match the edited activation maps

Method - 2 Visualize the Unit (Application)

- Rosetta Neurons Guided Editing - Results



Outline

Discussion

- A new method for mining and visualizing common representations that emerge in different visual models.
- Promising in some advanced generative tasks.
- Limitations:
  - Fails in GAN-GAN matching.
  - Fails in Diffusion models.
  - May suffer from spurious correlations.

Thanks!