MAX PLANCK INSTITUTE FOR INFORMATICS

SIC Saarland Informatics Campus

ICCV OCT 19-23, 2025 HONOLULU HAWAII

VITAL: More Understandable Feature Visualization through Distribution Alignment and Relevant Information Flow

Ada Görgün

Bernt Schiele

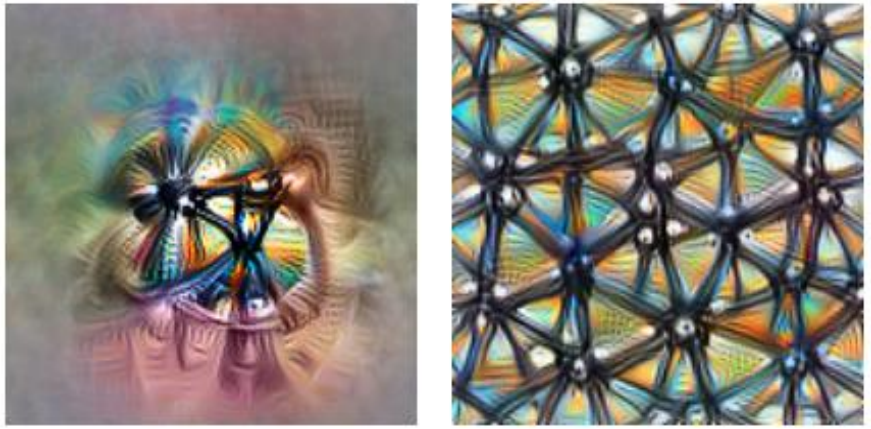
Jonas Fischer

Max Planck Institute for Informatics, Saarland Informatics Campus

Mechanistic Interpretability: Feature Visualization

Attribution Methods vs Feature Visualization

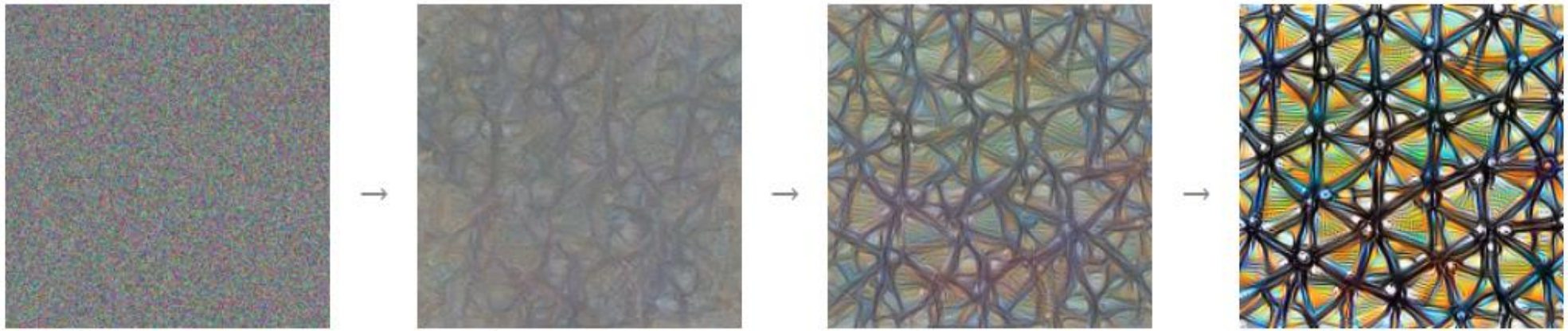
Question: What part of an example is responsible for the network activating a particular way?



Question: What a network or parts of a network are looking for by generating examples?

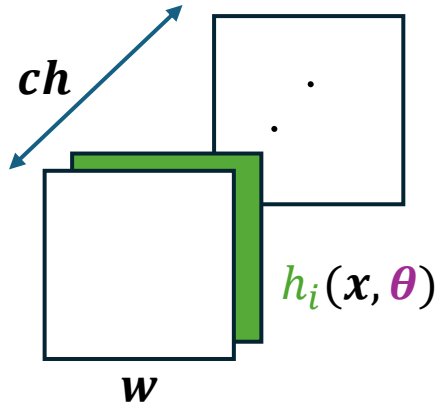
Mechanistic Interpretability: Feature Visualization

- How?
  - ➢ starting from random noise, optimize an image to activate a particular neuron



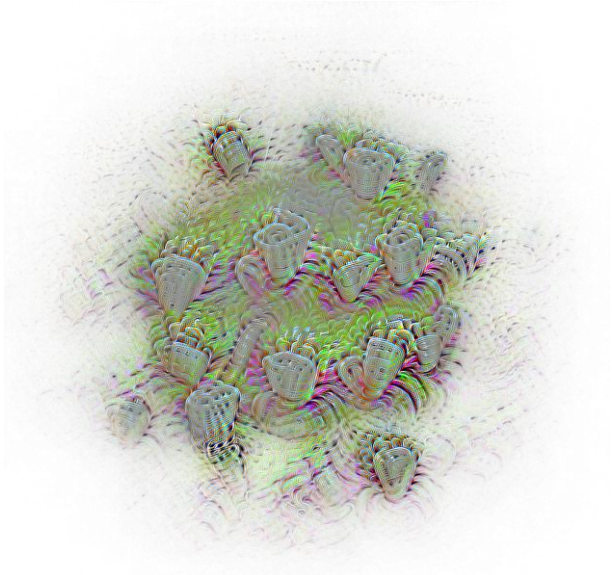
- In a way, this is the opposite of the learning process:
  - ➢ Normally, we have input data and want to learn the right weights $\boldsymbol{\theta}$
  - ➢ Here, we have the weights and want to generate data samples

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, h_i(\boldsymbol{x}, \boldsymbol{\theta})$$

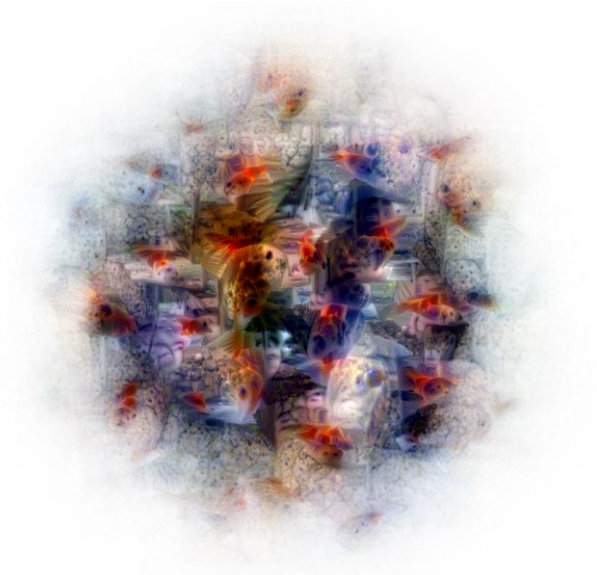


Feature Visualization: Challenges of Existing Efforts

- Unrealistically Generated Images:
  - ➢ Visualized features produce abstract, non-realistic images with repetitive patterns [1, 2].
  - ➢ Results often fail to resemble recognizable patterns or objects [1, 2].
  - ➢ Fails in larger models [1, 3].
  - ➢ Leverage statistically learned priors from auxiliary models [3].



[1] Fourier FV, Olah et. al, 2017



[2] MACO, Fel et. al, NeurIPS 2023

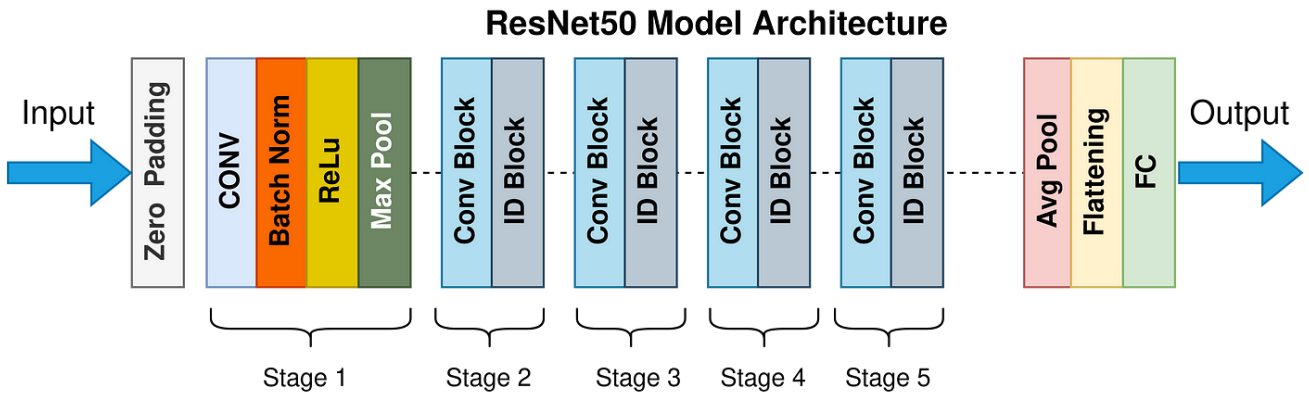
[3] GANs, Nyugen et. al, NeurIPS 2016

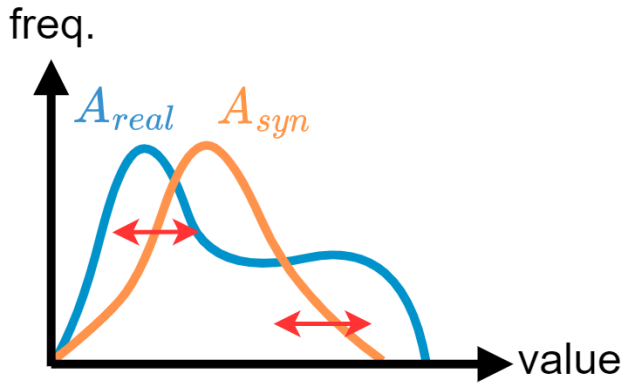
VITAL: Motivation and Contributions

- Matching the feature distribution of real images and the synthetic image to:
  - generate a **more interpretable** visualization
  - generate **non-repetitive patterns**
  - capture the intrinsic characteristics of real images
  - **not directly rely on activation maximization**
- Utilization of the method for a more intuitive grasp on:
  - how the **flow of the network** affects the visualized image
- Evaluation of the generated image in terms of:
  - how well it reflects the real images
  - **quantitative** and **qualitative** metrics

VITAL: Proposed Method

- We want to **match** the feature distribution of real (A_{real}) images and the synthetic (A_{syn}) image.

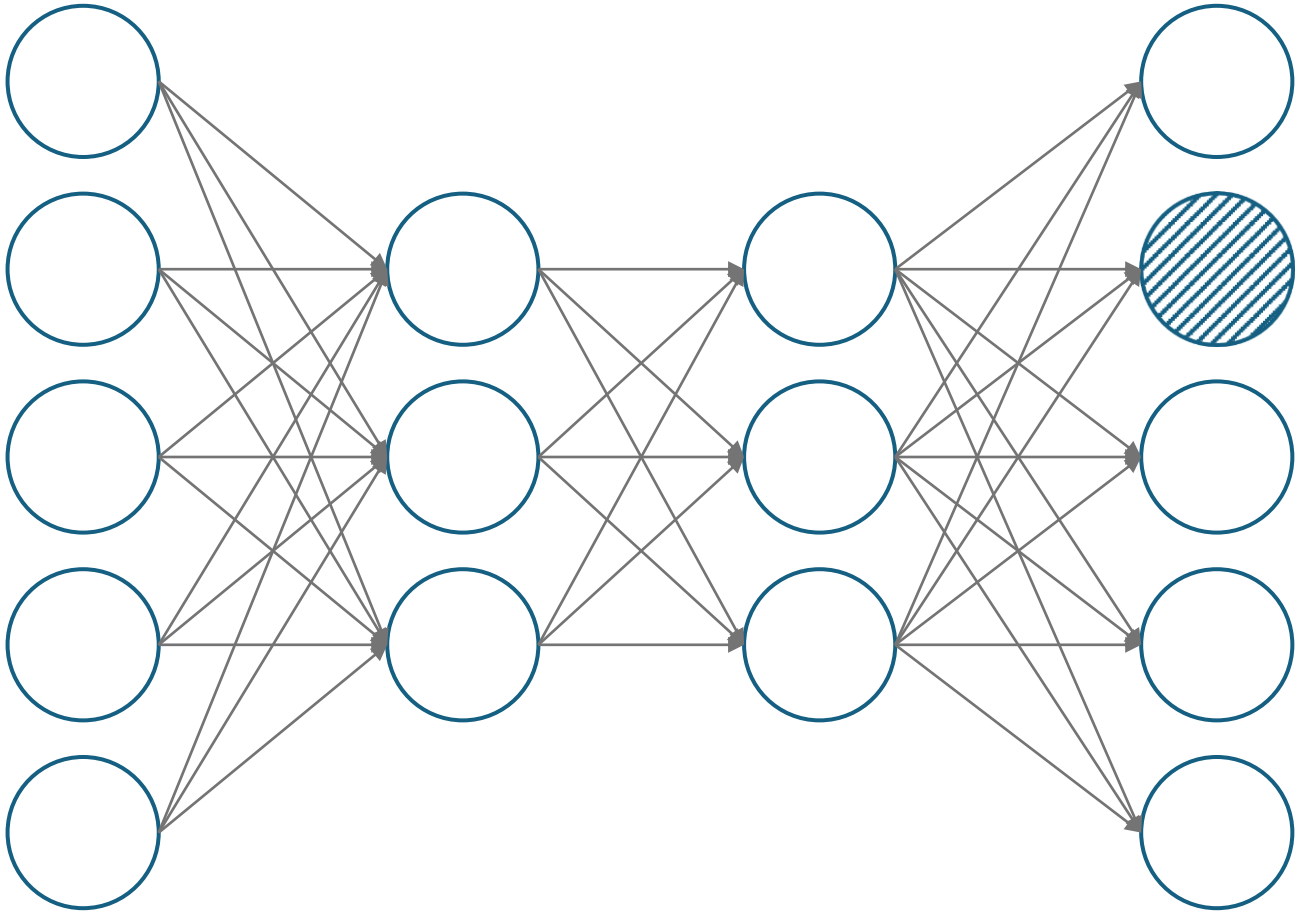




Matching is done by the sort matching algorithm (Zhang et al. CVPR 2022)

- We want to achieve visualizations for
  - class neurons
  - intermediate neurons

VITAL: Class Neuron Visualization



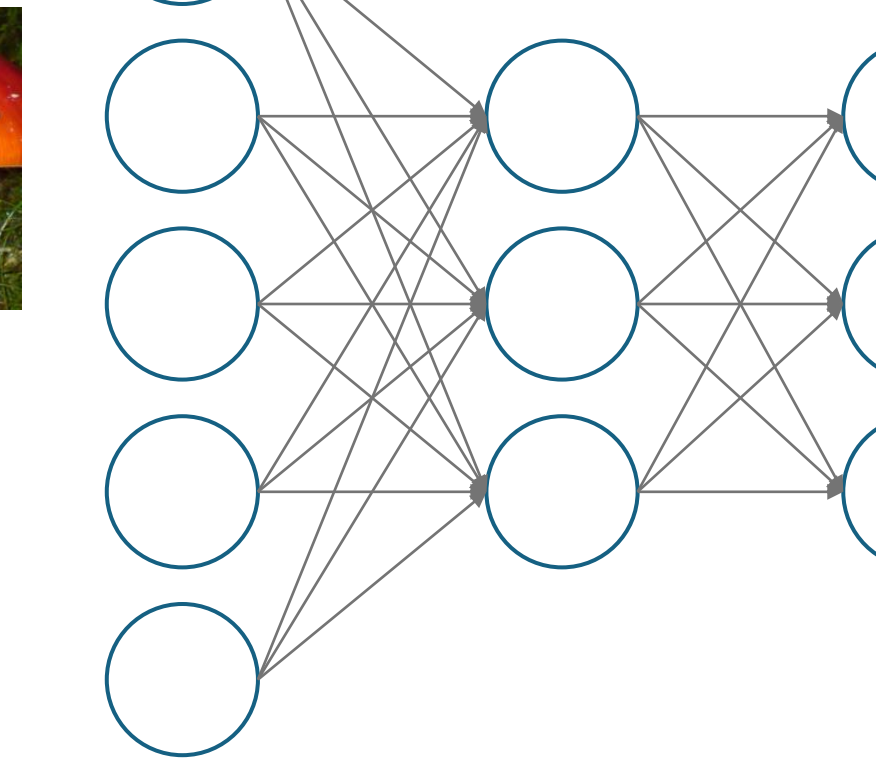
1. select target neuron

agaric

Pretrained Model Architecture

VITAL: Class Neuron Visualization

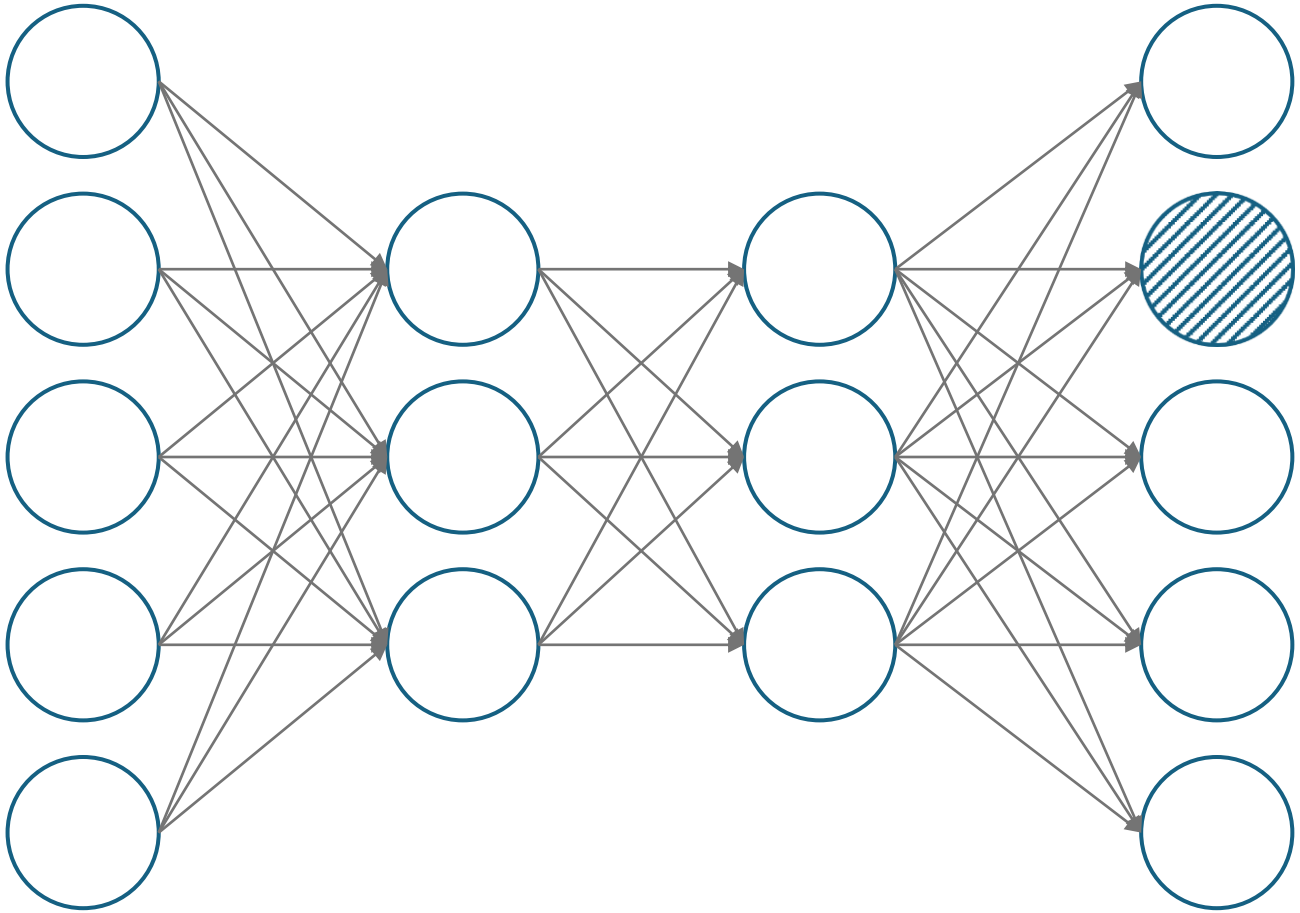
2. select random N images of the target class neuron



Pretrained Model Architecture

VITAL: Class Neuron Visualization

Real Image Inputs

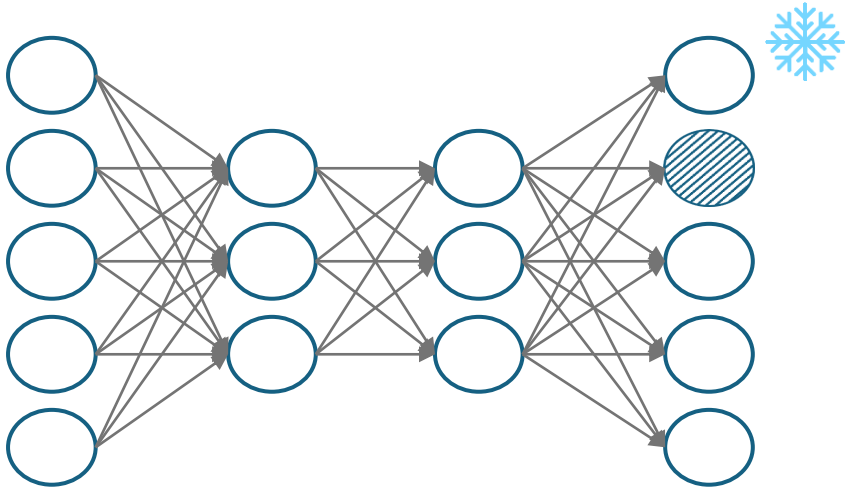


Pretrained Model Architecture

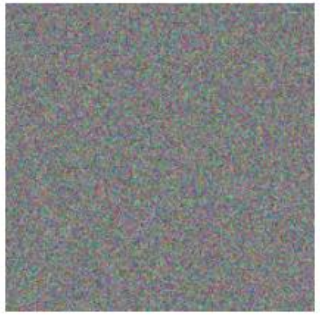
3. feed-forward the real image inputs to the pretrained model

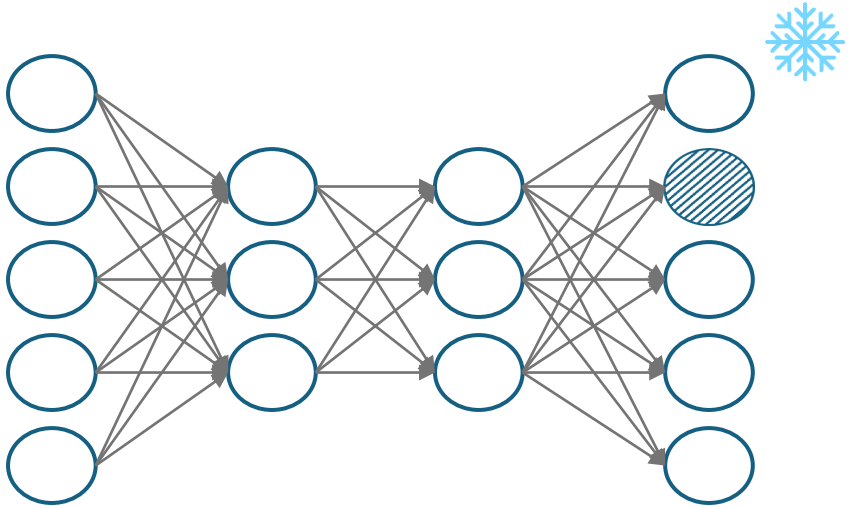
VITAL: Class Neuron Visualization

Real Image Inputs



Synthetic Input

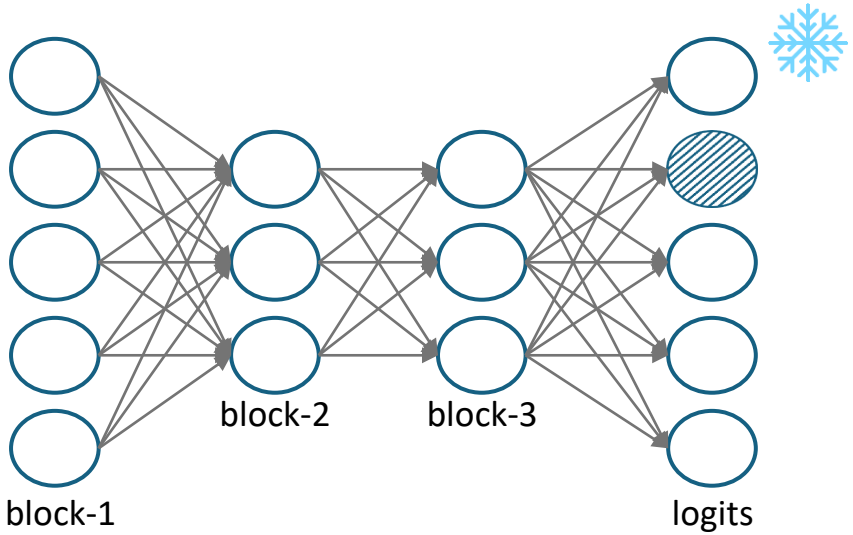




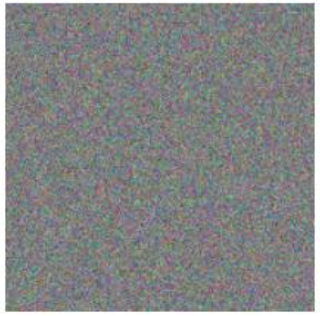
4. generate a noisy input and feed-forward to the same frozen model

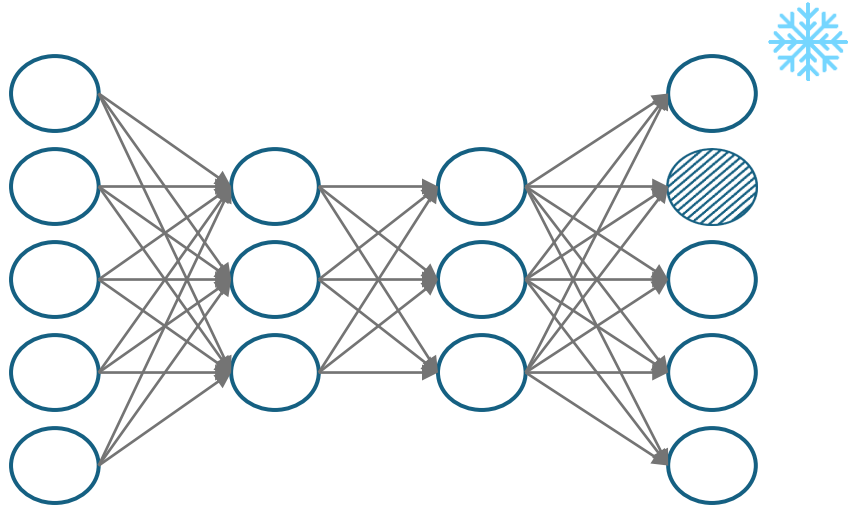
VITAL: Class Neuron Visualization

Real Image Inputs



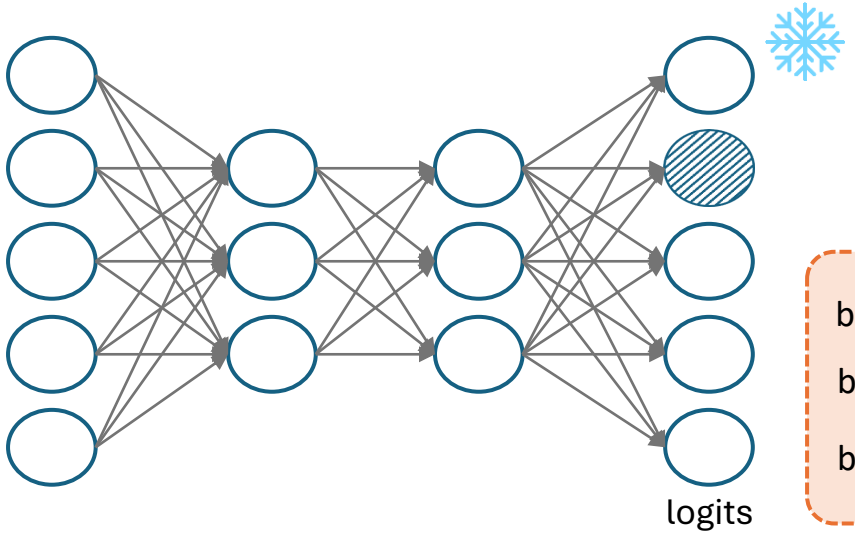
Synthetic Input





VITAL: Class Neuron Visualization

Real Image Inputs



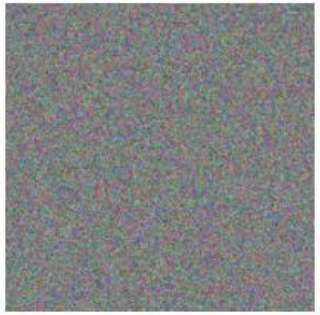
block-1

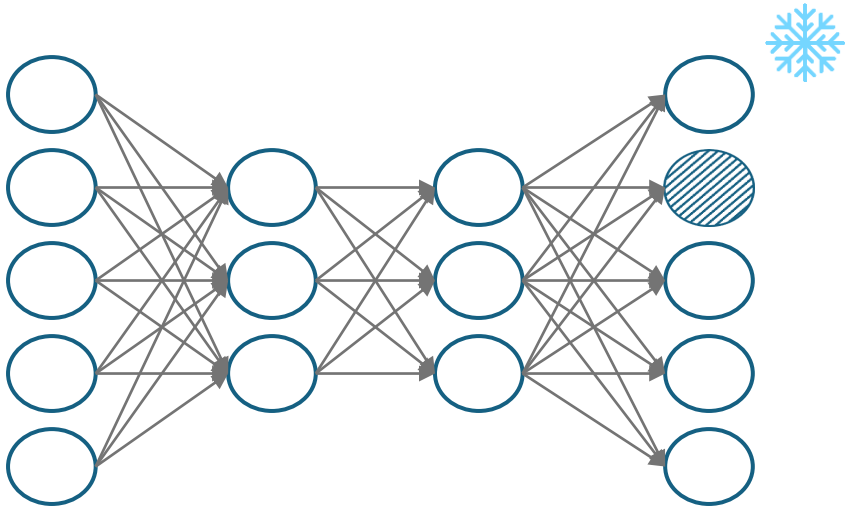
block-2

block-3

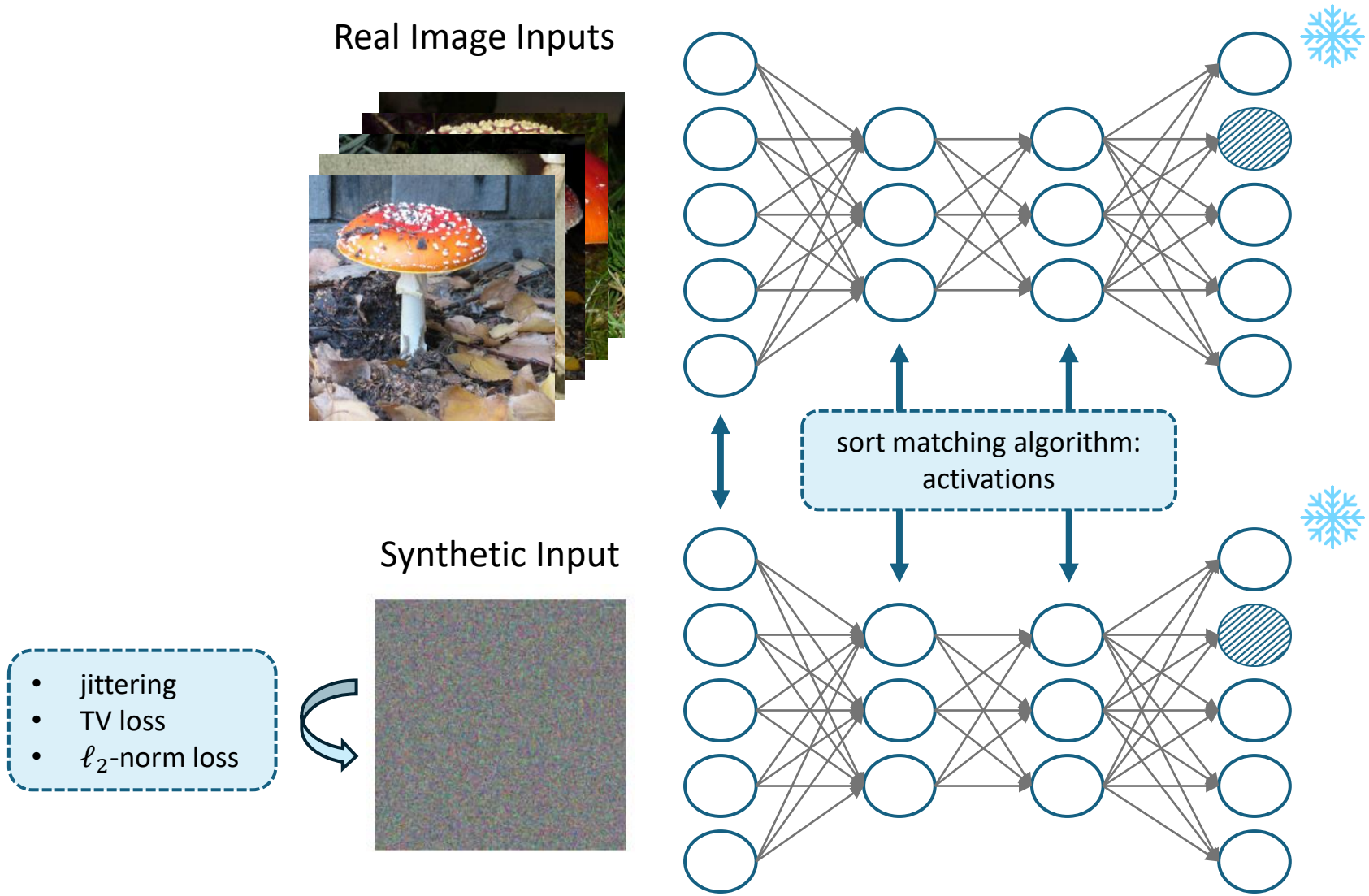
These blocks will be used for aligning the distribution of latent features in each block using the **sort-matching algorithm**.

Synthetic Input





VITAL: Class Neuron Visualization



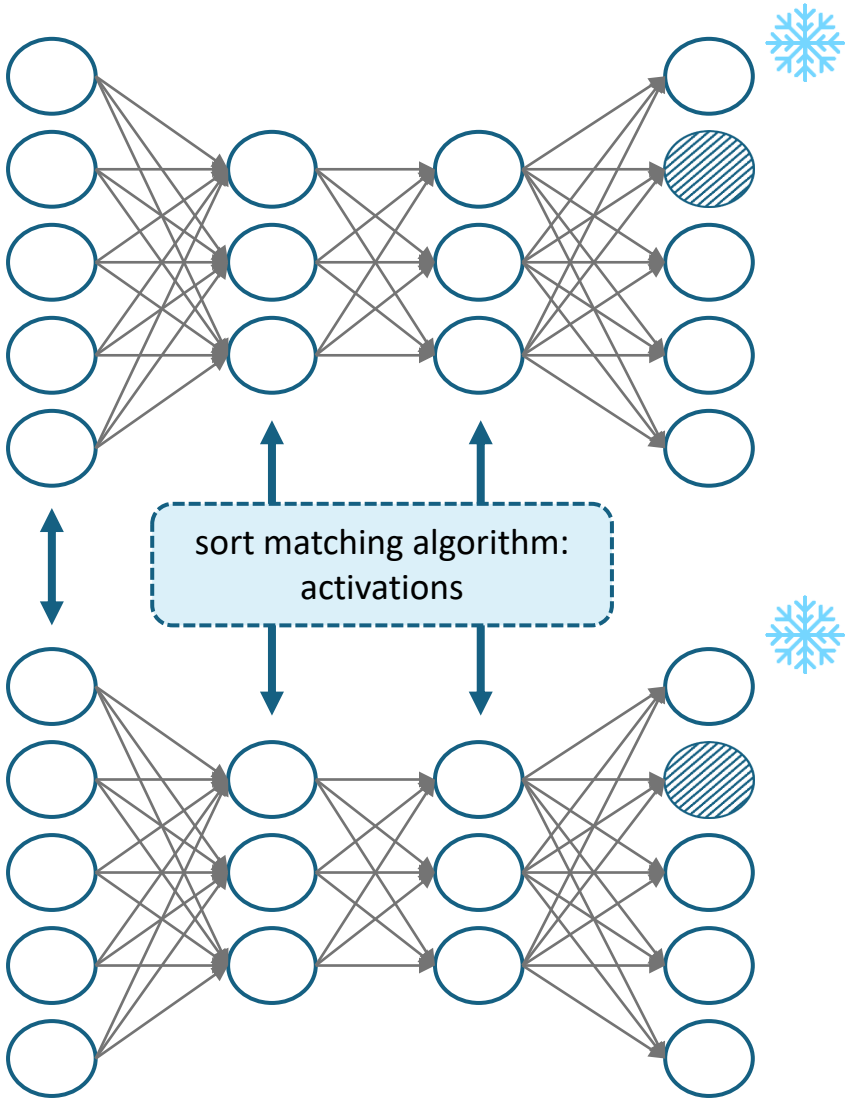
5. optimize the synthetic image with the sort matching loss, TV loss and ℓ_2-norm loss

VITAL: Class Neuron Visualization

Real Image Inputs

Synthetic Input

- jittering
- TV loss
- ℓ_2-norm loss



5. optimize the synthetic image with the sort matching loss, TV loss and ℓ_2-norm loss

VITAL: Class Neuron Visualization

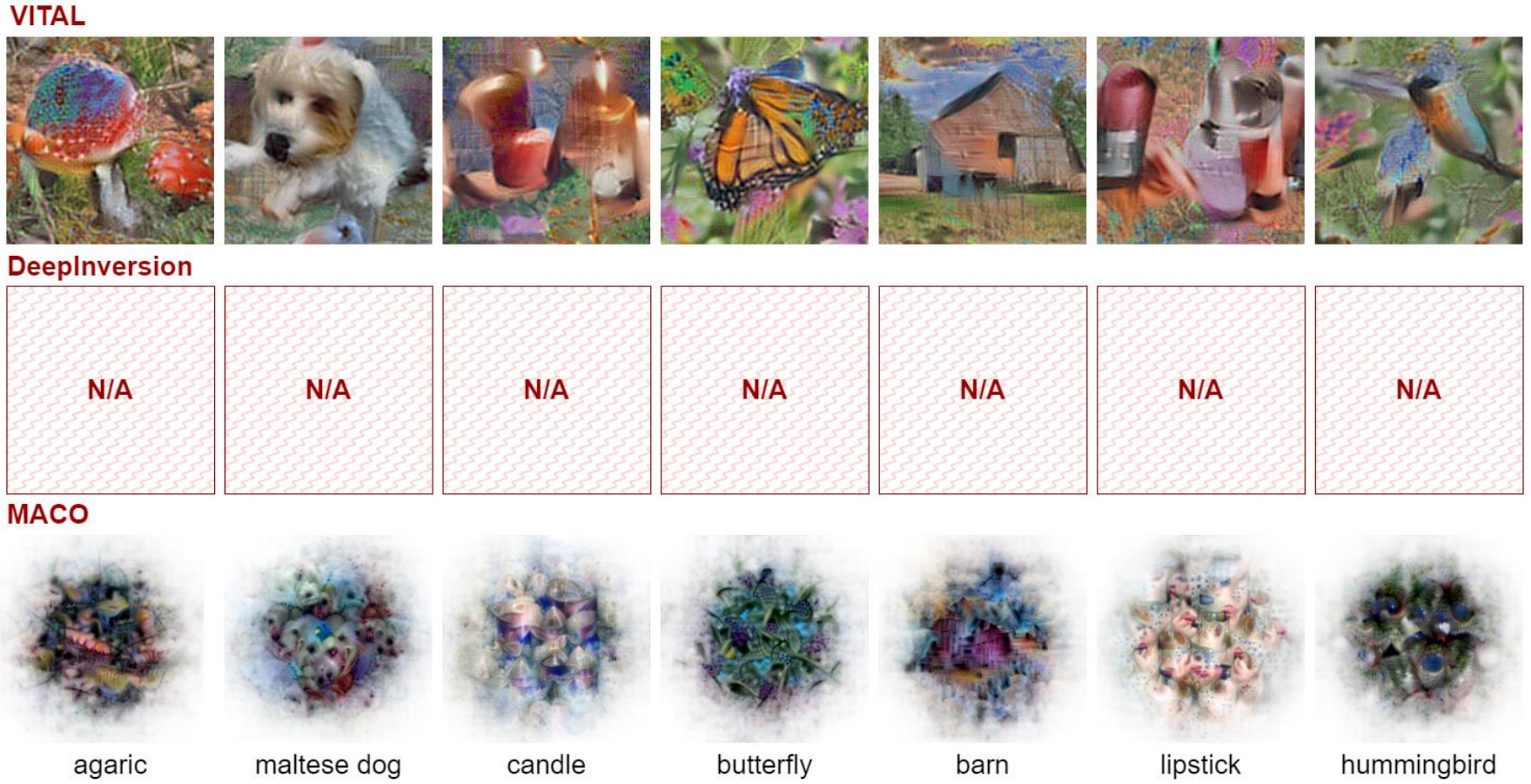
- Model tested: ResNet50

VITAL: Class Neuron Visualization

- Model tested: ConvNext-base

VITAL: Class Neuron Visualization

- Model tested: ViT-L-32



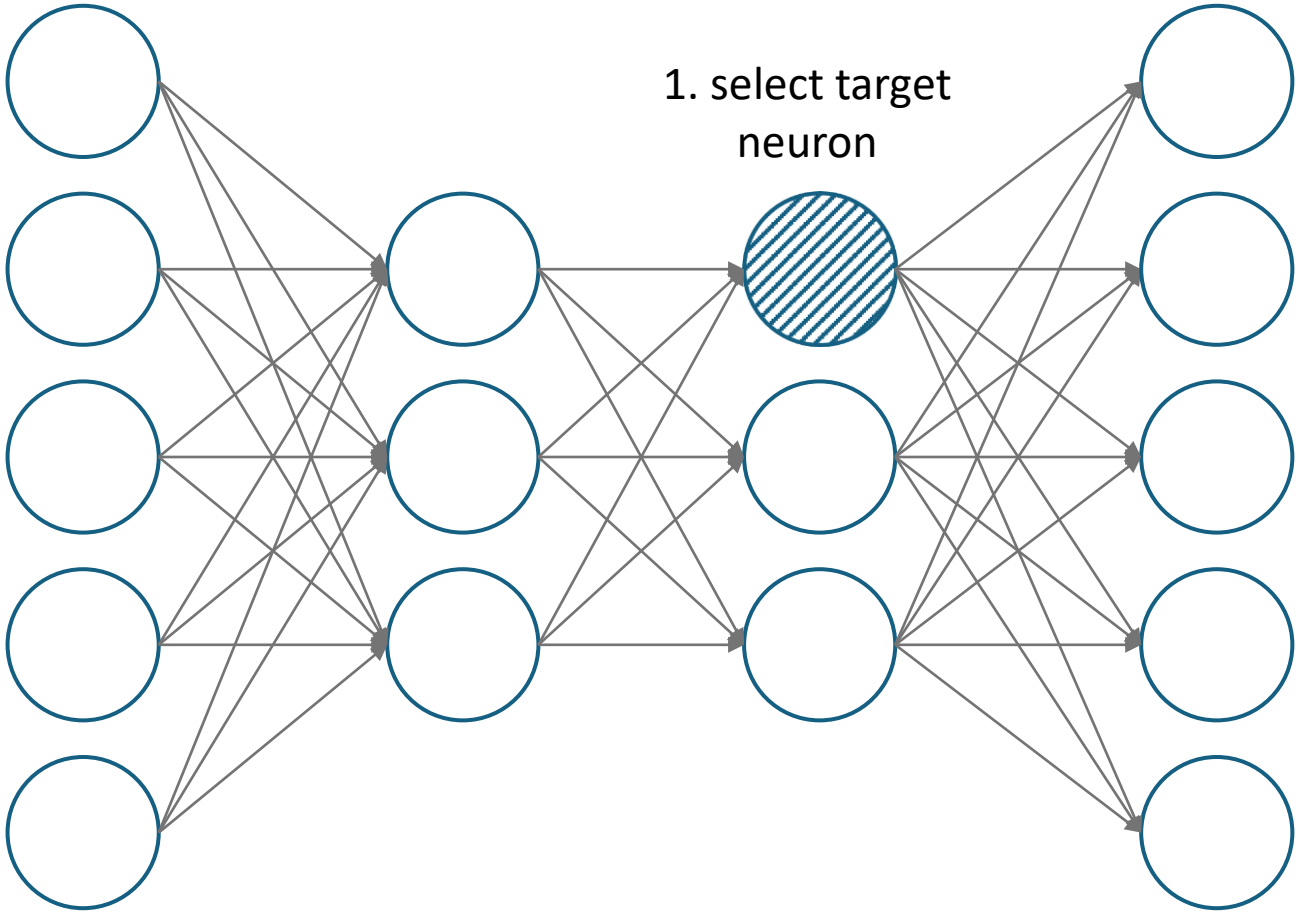
VITAL: Class Neuron Visualization

- Quantitative Results
  - ResNet50, ConvNeXt, DenseNet121, ViT-L-16, ViT-L-32
  - Compared Methods: Fourier, MACO, DeepInversion
- Metrics:

1. **Top1 Accuracy** of the generated images
2. **FID score** to quantify visualization realism
3. **CLIP Zero-Shot Prediction** to assess general perceptual quality

	Method	Setup	Acc. Top1 (↑)	FID (↓)	Zero-Shot Prediction Top1 (↑)	Zero-Shot Prediction Top5 (↑)
ResNet50	*ImageNet*		-	-	69.11	92.23
	MACO	r: 224	29.43	360.74	12.87	29.73
	Fourier	r: 224	21.30	422.44	6.73	18.27
	DeepInv	bs: 64	**100.00**	**35.76**	29.90	55.20
	VITAL		99.90	58.79	**66.62**	**92.56**
ConvNeXt	*ImageNet*		-	-	65.66	89.80
	MACO	r: 224	66.07	62.55	7.20	19.77
	Fourier	r: 224	60.07	59.60	2.77	8.30
	VITAL		**99.97**	**3.92**	**63.53**	**90.30**
DenseNet121	*ImageNet*		-	-	70.64	93.16
	MACO	r: 224	9.20	1.80	9.33	23.20
	Fourier	r: 224	15.53	1.63	4.87	12.17
	DeepInv	bs: 64	**100.00**	**0.20**	10.00	25.47
	VITAL		99.93	0.27	**58.70**	**86.93**
ViT-L-16	*ImageNet*		-	-	64.78	89.31
	MACO	r: 224	44.33	946.96	3.93	10.57
	Fourier	r: 224	25.30	990.51	1.67	5.13
	VITAL		**99.80**	**126.29**	**68.17**	**92.80**
ViT-L-32	*ImageNet*		-	-	65.83	90.03
	MACO	r: 224	24.87	2318.90	17.53	37.23
	Fourier	r: 224	17.03	1983.09	10.30	28.10
	VITAL		**89.60**	**147.33**	**55.97**	**85.47**

VITAL: Intermediate Neuron Visualization



Pretrained Model Architecture

VITAL: Intermediate Neuron Visualization

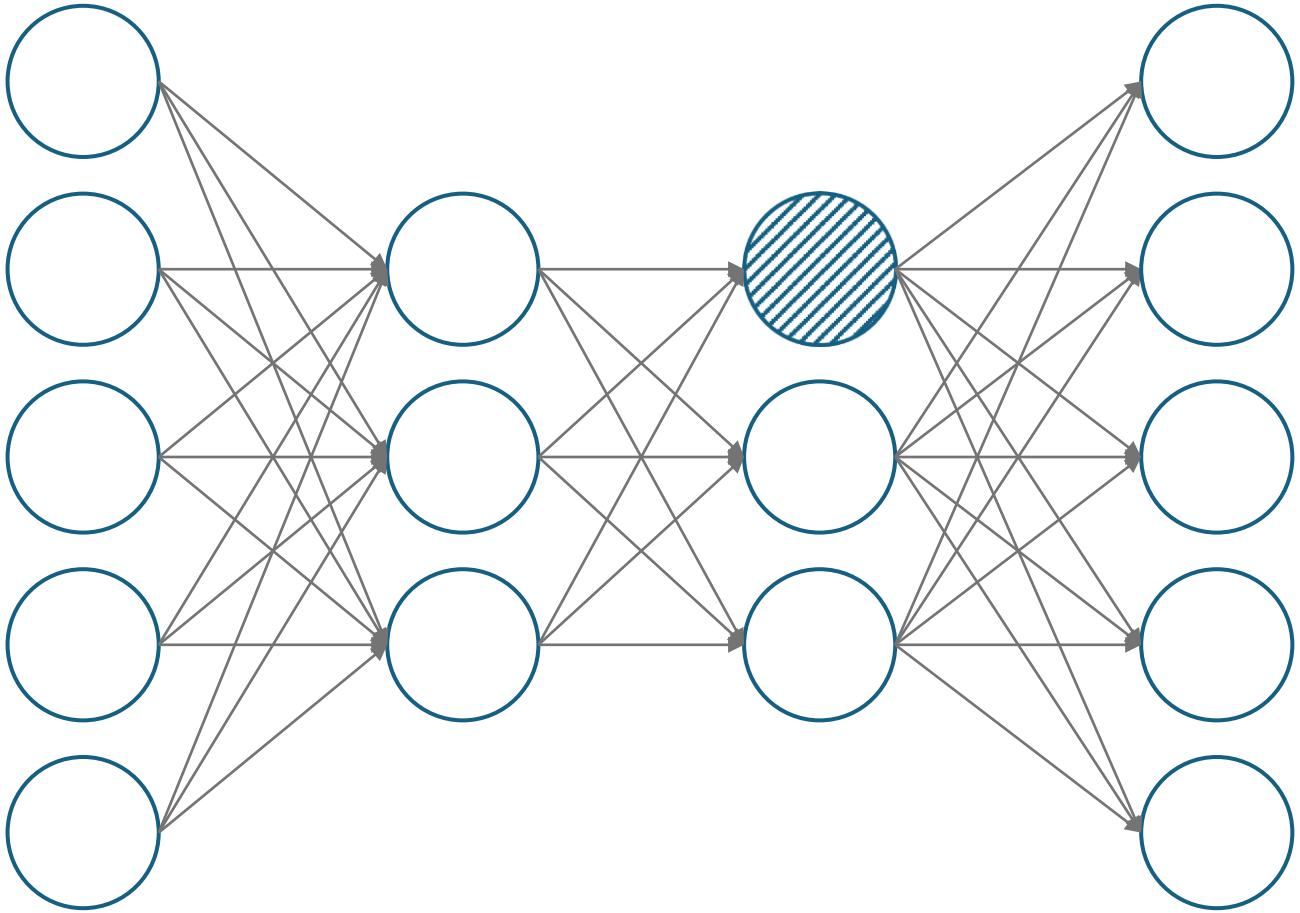
2. find top-k images that activate the target neuron the most

top activating image crops

Pretrained Model Architecture

VITAL: Intermediate Neuron Visualization

Real Image Inputs

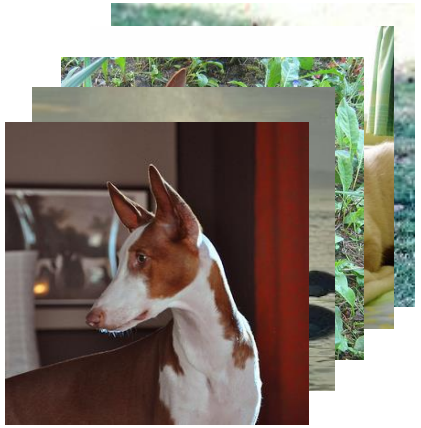


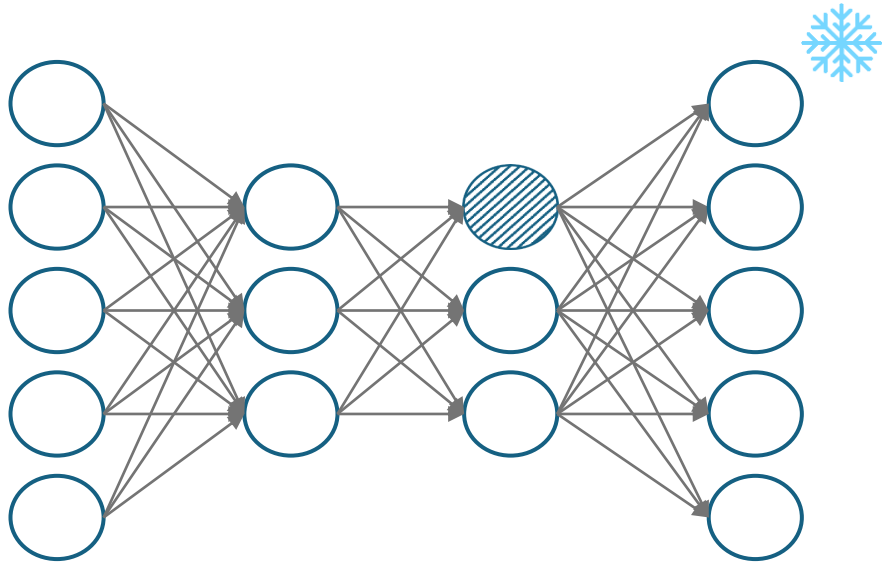
Pretrained Model Architecture

3. feed-forward the real image inputs to the pretrained model

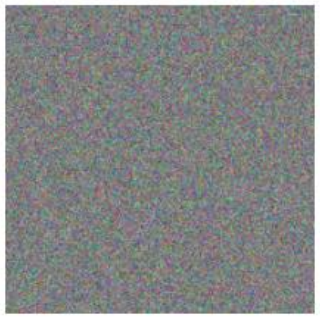
VITAL: Intermediate Neuron Visualization

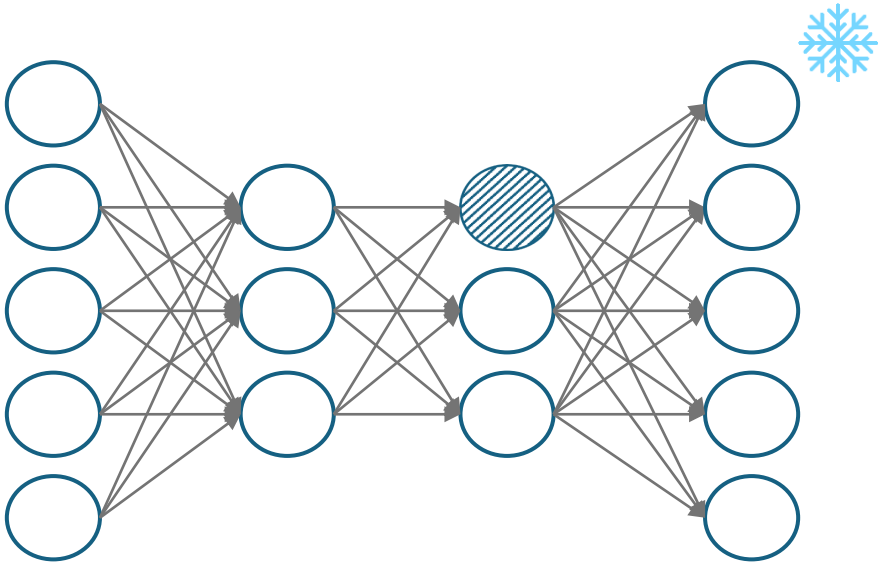
Real Image Inputs





Synthetic Input

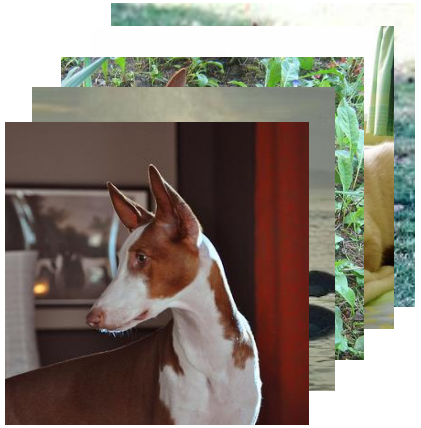


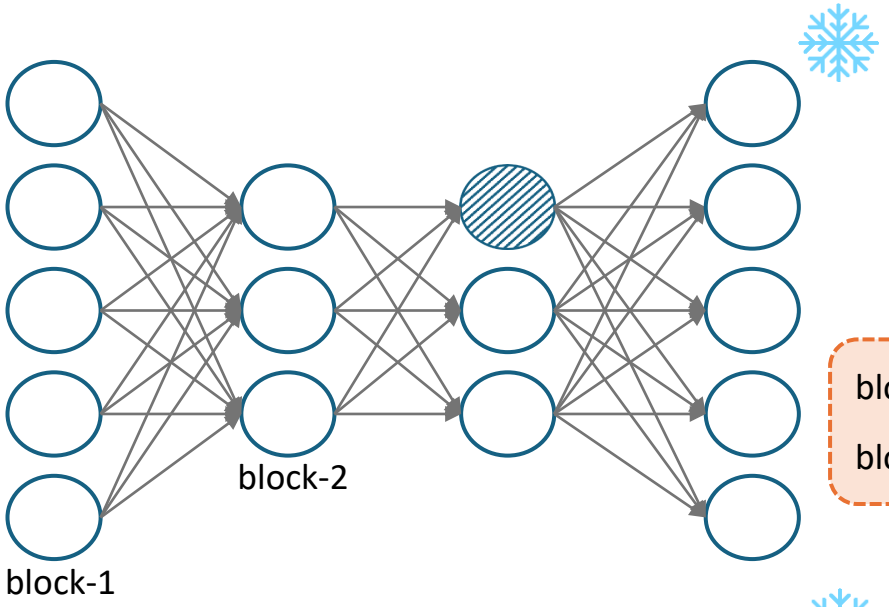


4. generate a noisy input and feed-forward to the same frozen model

VITAL: Intermediate Neuron Visualization

Real Image Inputs



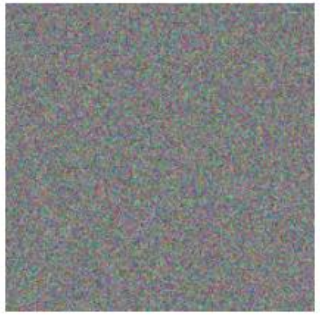


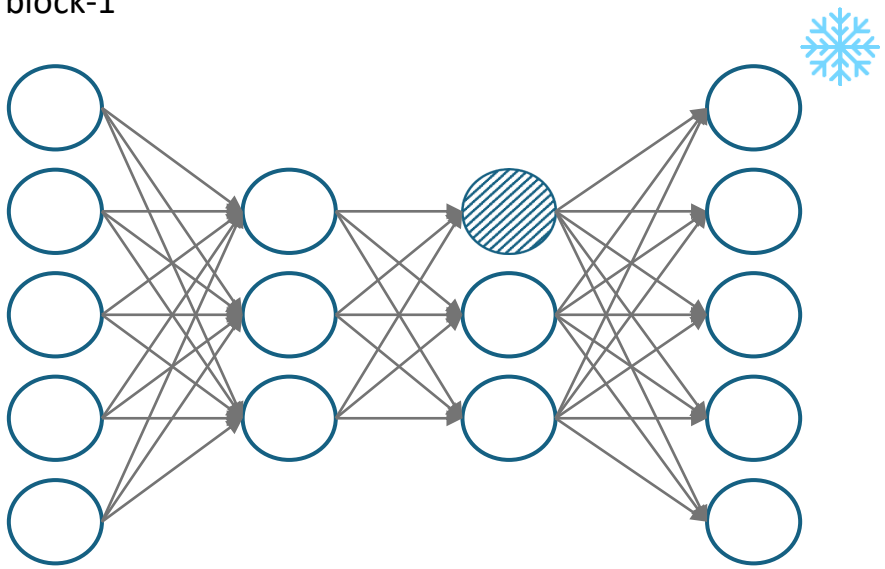
block-1

block-2

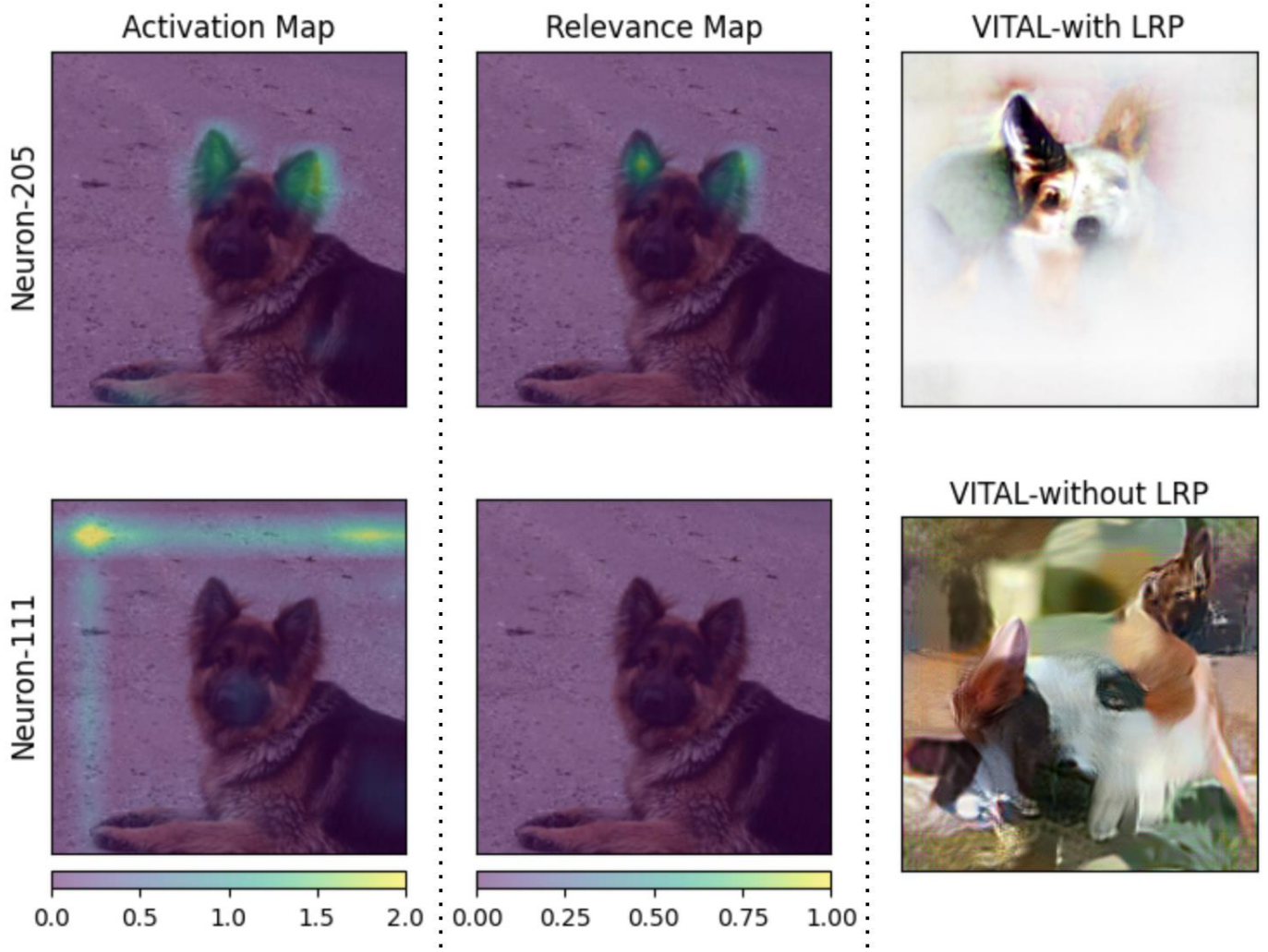
These blocks will be used for aligning the distribution of latent features in each through the **sort matching algorithm**.

Synthetic Input





VITAL: Intermediate Neuron Visualization

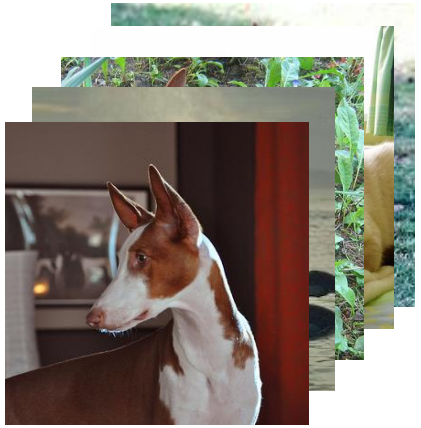


For intermediate neuron visualization, we need explicit supervision from the target neuron to find the regions for matching the distribution.

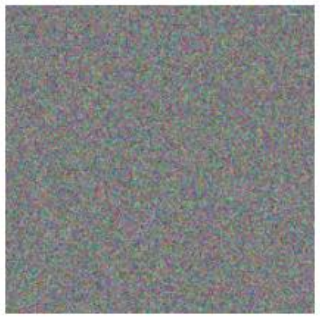
This supervision can be obtained through an attribution method (e.g., LRP).

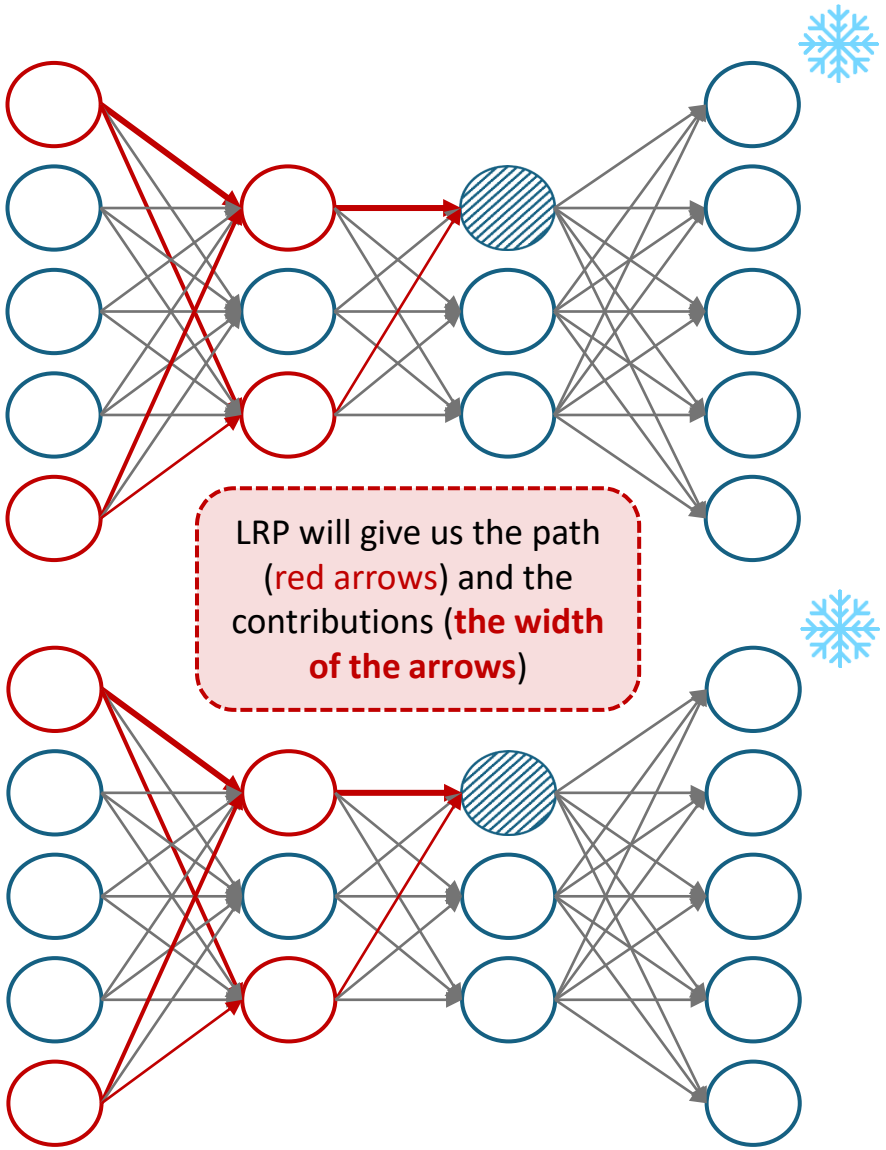
VITAL: Intermediate Neuron Visualization

Real Image Inputs



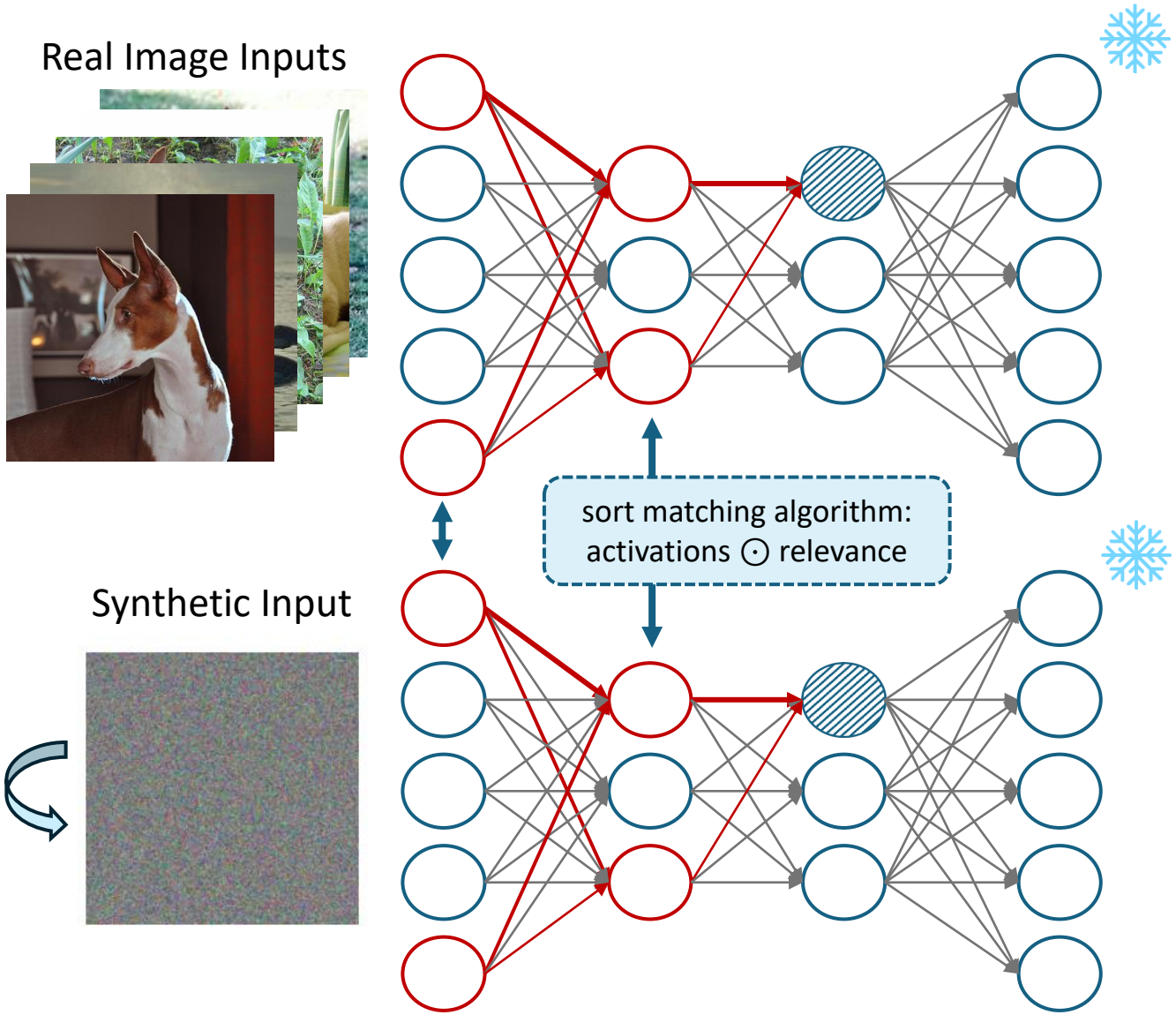
Synthetic Input





LRP will give us the path (red arrows) and the contributions (**the width of the arrows**)

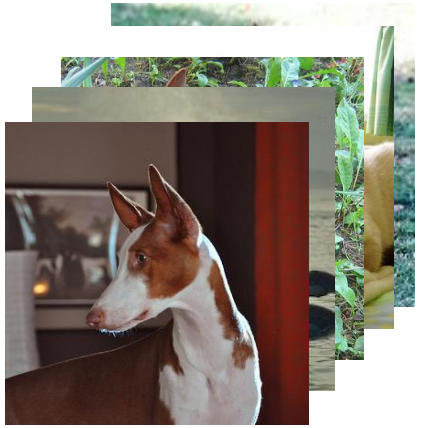
VITAL: Intermediate Neuron Visualization



5. optimize the synthetic image with the sort matching loss, TV loss and ℓ_2-norm loss

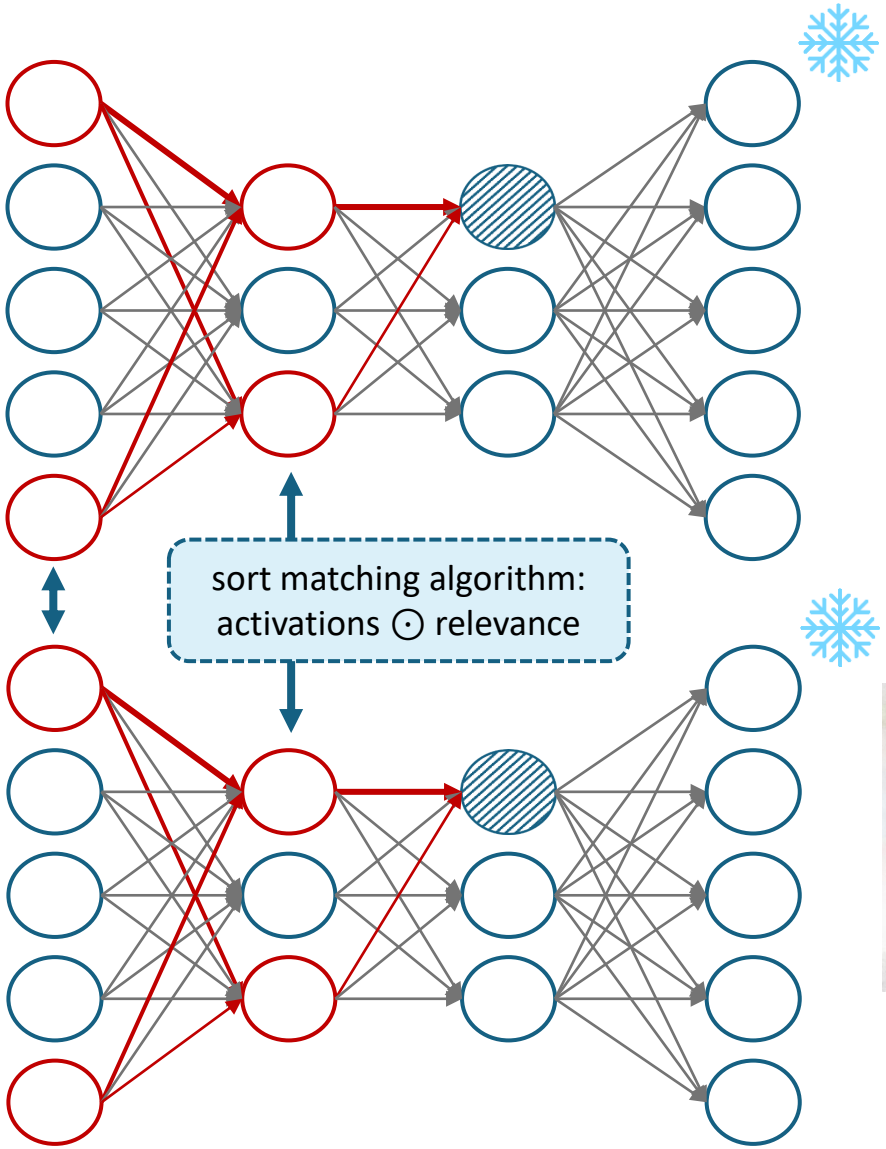
VITAL: Intermediate Neuron Visualization

Real Image Inputs

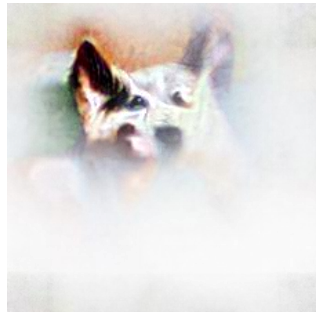


Synthetic Input

- jittering
- TV loss
- ℓ_2-norm loss



5. optimize the synthetic image with the sort matching loss, TV loss and ℓ_2-norm loss

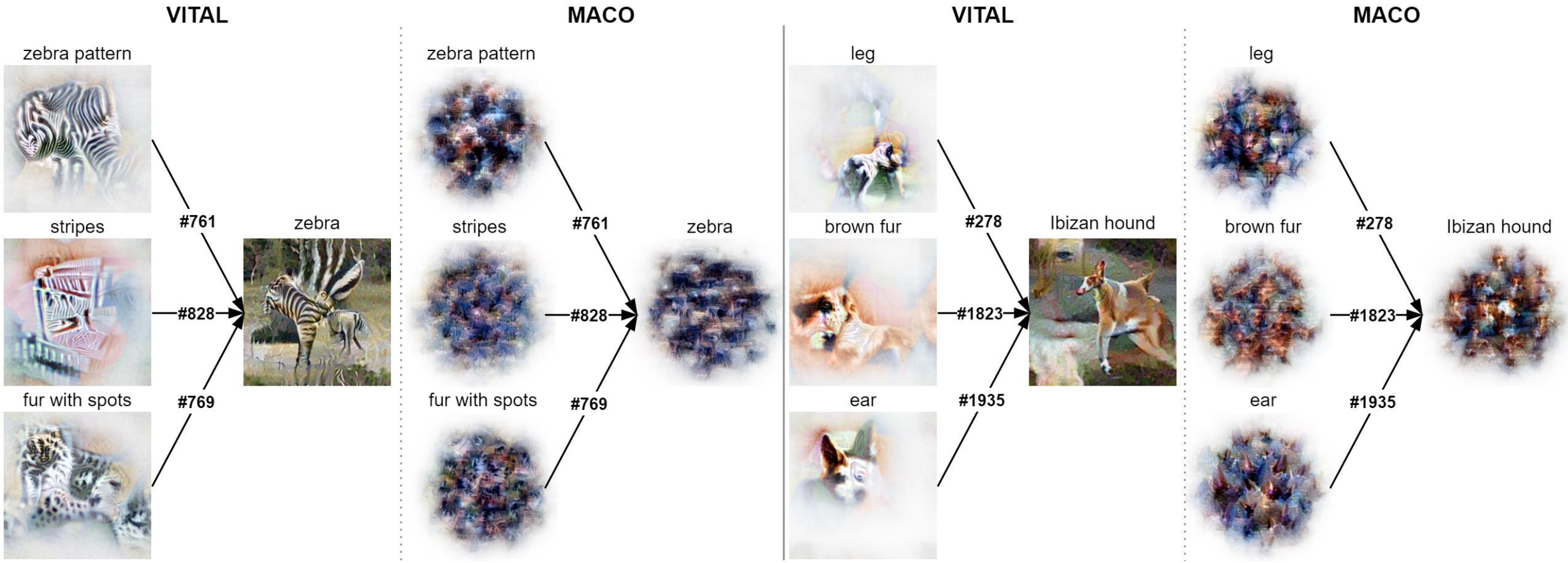


transparency map

top-k image patches

VITAL: Intermediate Neuron Visualization

- Model tested: ResNet50



VITAL: Human User Study

- The user study aims to evaluate **interpretability / understandability** of the generated feature visualizations (FVs) from a **human perspective**.
  - ➢ complements the quantitative metrics (e.g. classification accuracy, FID, CLIP zero-shot)
- The study is organized into **3 tasks**, with increasing difficulty / less prior information.

Task	Description / Input	What participants do	Scoring / Measure
(1) *Class-level match*	A class name (single word) + a FV	Rate how well the FV matches that class	1 (worst) to 5 (best) scale
(2) *Inner neuron match*	Reference images that highly activate a hidden neuron + the FV for that neuron	Rate how well the FV corresponds to those images	1 (worst) to 5 (best) scale
(3) *Labeling / open description*	A FV only (no class name)	Write a word / short description for what the FV shows	Compare description to ground truth via embedding similarity (Universal Sentence Encoder)

VITAL: Human User Study

- The user study aims to evaluate **interpretability / understandability** of the generated feature visualizations (FVs) from a **human perspective**.
  - ➢ complements the quantitative metrics (e.g. classification accuracy, FID, CLIP zero-shot)
- The study is organized into **3 tasks**, with increasing difficulty / less prior information.

Task	Description / Input	What participants do	Scoring / Measure
(1) *Class-level match*	A class name (single word) + a FV	Rate how well the FV matches that class	1 (worst) to 5 (best) scale
(2) *Inner neuron match*	Reference images that highly activate a hidden neuron + the FV for that neuron	Rate how well the FV corresponds to those images	1 (worst) to 5 (best) scale
(3) *Labeling / open description*	A FV only (no class name)	Write a word / short description for what the FV shows	Compare description to ground truth via embedding similarity (Universal Sentence Encoder)

VITAL: Human User Study

- The user study aims to evaluate **interpretability / understandability** of the generated feature visualizations (FVs) from a **human perspective**.
  - ➢ complements the quantitative metrics (e.g. classification accuracy, FID, CLIP zero-shot)
- The study is organized into **3 tasks**, with increasing difficulty / less prior information.

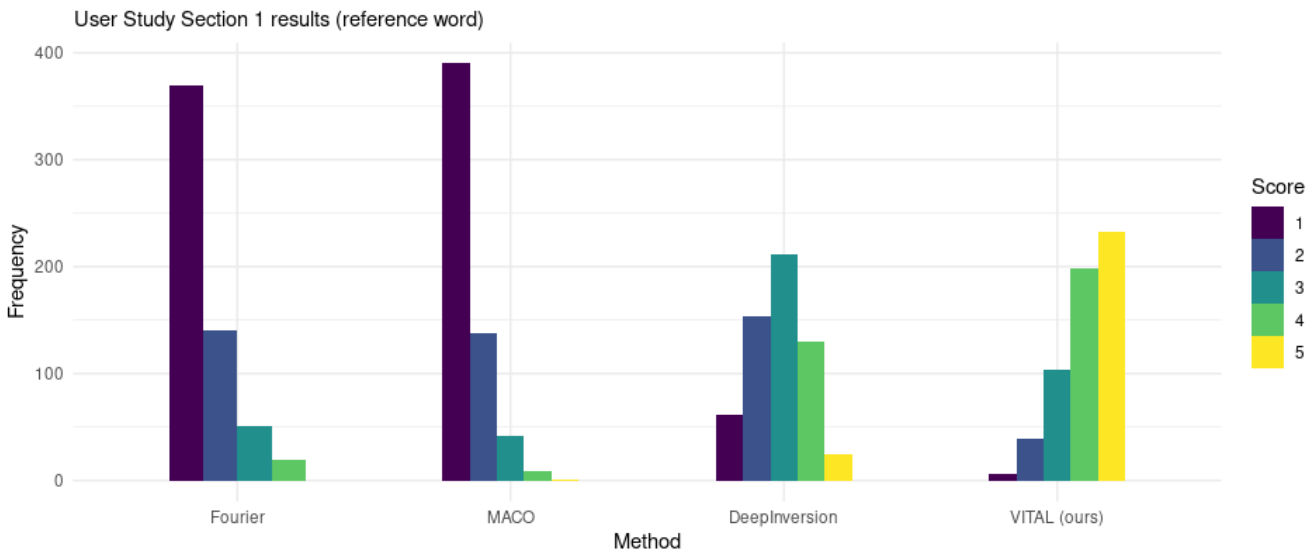
Task	Description / Input	What participants do	Scoring / Measure
(1) *Class-level match*	A class name (single word) + a FV	Rate how well the FV matches that class	1 (worst) to 5 (best) scale
(2) *Inner neuron match*	Reference images that highly activate a hidden neuron + the FV for that neuron	Rate how well the FV corresponds to those images	1 (worst) to 5 (best) scale
(3) *Labeling / open description*	A FV only (no class name)	Write a word / short description for what the FV shows	Compare description to ground truth via embedding similarity (Universal Sentence Encoder)

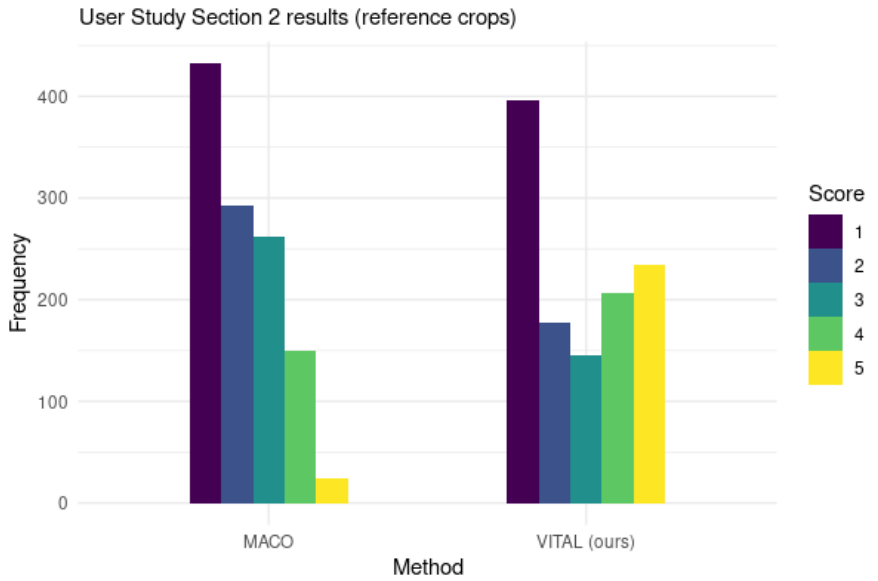
VITAL: Human User Study

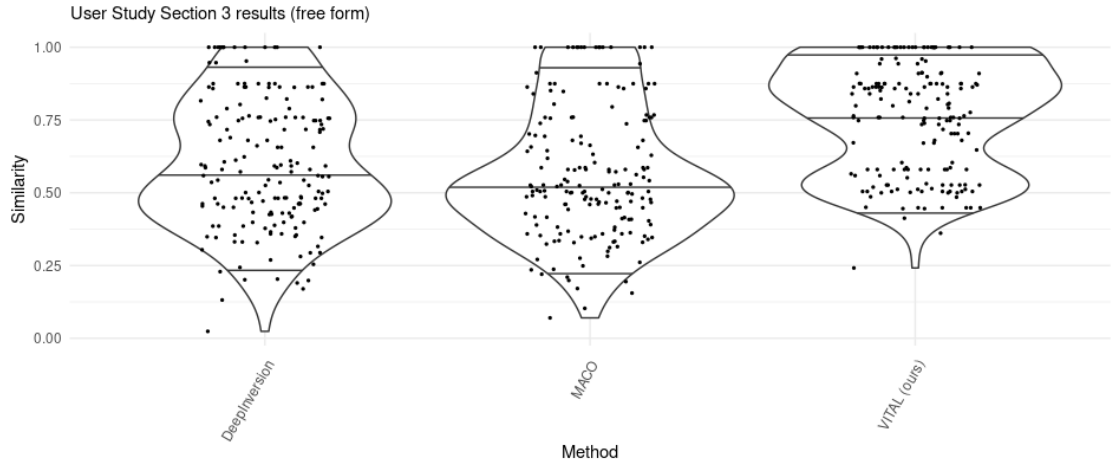
- The user study aims to evaluate **interpretability / understandability** of the generated feature visualizations (FVs) from a **human perspective**.
  - ➢ complements the quantitative metrics (e.g. classification accuracy, FID, CLIP zero-shot)
- The study is organized into **3 tasks**, with increasing difficulty / less prior information.

Task	Description / Input	What participants do	Scoring / Measure
(1) *Class-level match*	A class name (single word) + a FV	Rate how well the FV matches that class	1 (worst) to 5 (best) scale
(2) *Inner neuron match*	Reference images that highly activate a hidden neuron + the FV for that neuron	Rate how well the FV corresponds to those images	1 (worst) to 5 (best) scale
(3) *Labeling / open description*	A FV only (no class name)	Write a word / short description for what the FV shows	Compare description to ground truth via embedding similarity (Universal Sentence Encoder)

VITAL: Human User Study







VITAL: Limitations and Future Work

- While our visualizations, like those of other methods, are not photo-realistic, they sometimes lack spatial coherence, particularly for non-rigid objects.
- Exhaustive benchmarks and well-defined evaluation metrics for inner neurons are actively discussed but still lacking.
- Future Works:
  - ➢ understanding networks in the medical domain
  - ➢ downstream applications in global representation learning
  - ➢ studying what neurons encode after knowledge transfer
  - ➢ how pruning affects the representations

Thanks for listening!