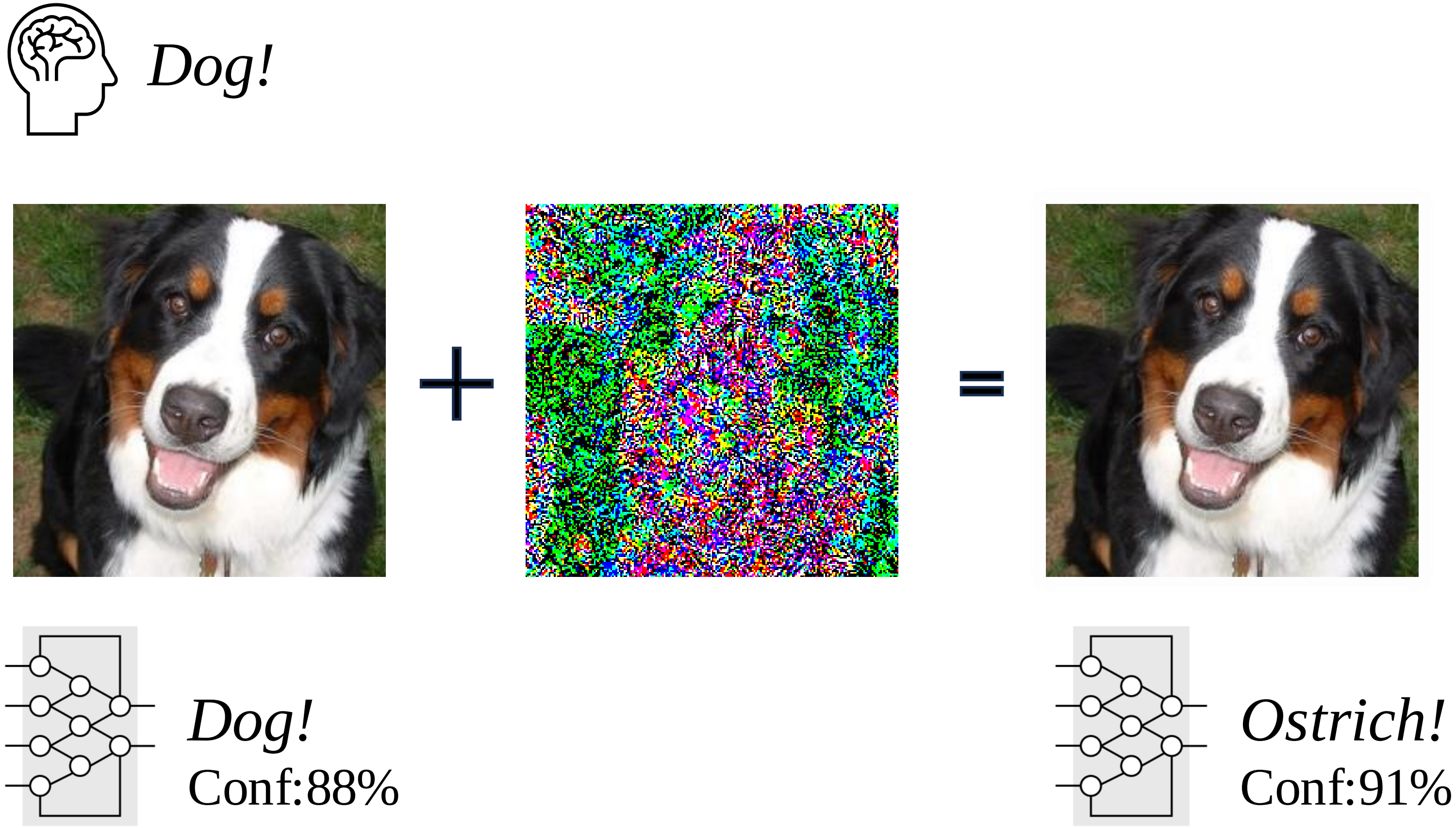
Neural-guidance by the Human Ventral Visual Stream Improves Neural Network Robustness

Zhenan Shao[1,2], Linjian Ma[3], Bo Li[3,4], Diane M. Beck[1,2]

[1]Department of Psychology, University of Illinois Urbana-Champaign
[2]Beckman Institute, University of Illinois Urbana-Champaign
[3]Department of Computer Science, University of Illinois Urbana-Champaign
[4]Department of Computer Science, University of Chicago

MLDL Workshop at Sandia National Laboratories, 2024

Vulnerable machine vision

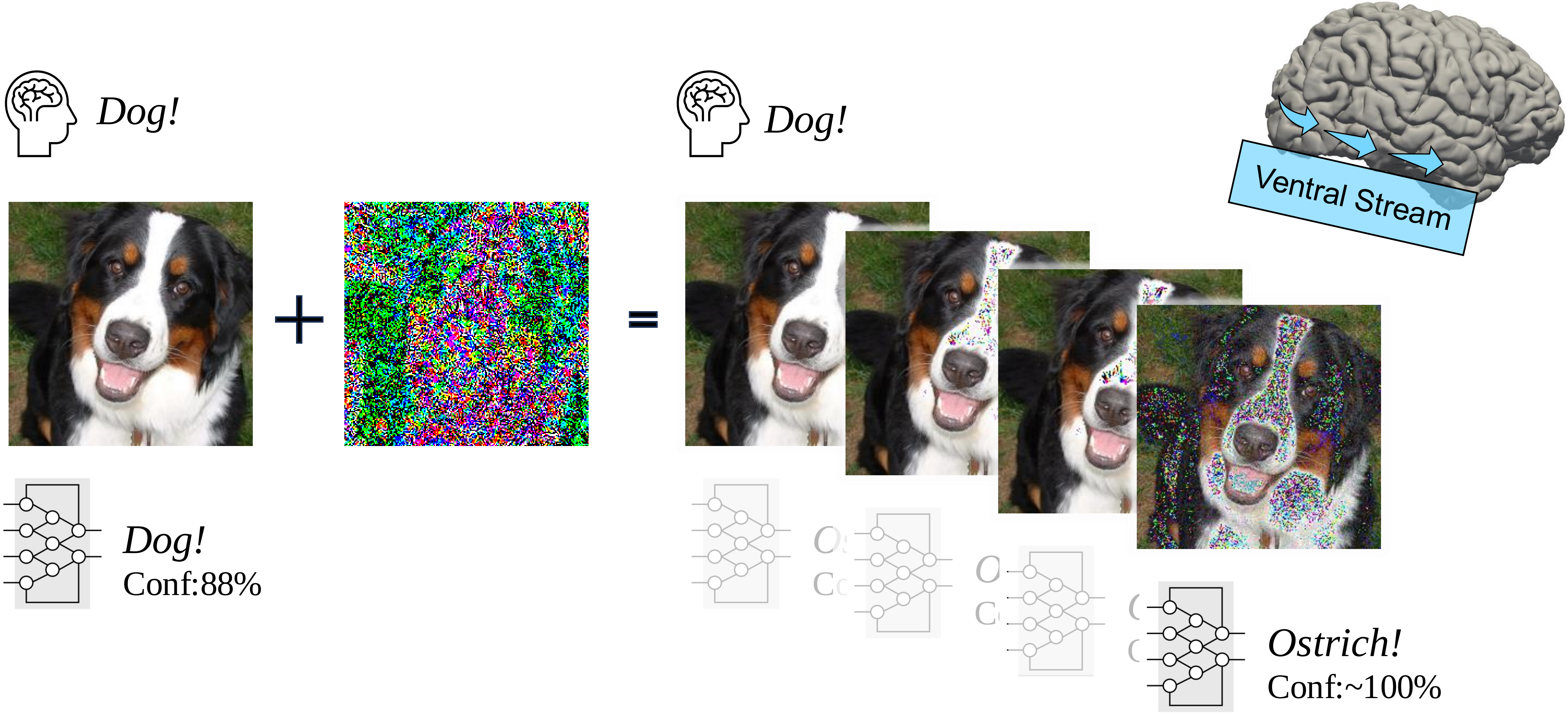


Dog!

Dog!
Conf:88%

Ostrich!
Conf:91%

Vulnerable machine vision



Dog!

Dog!

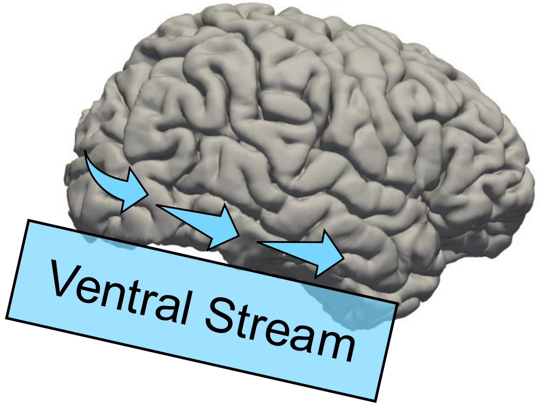
Dog!
Conf:88%

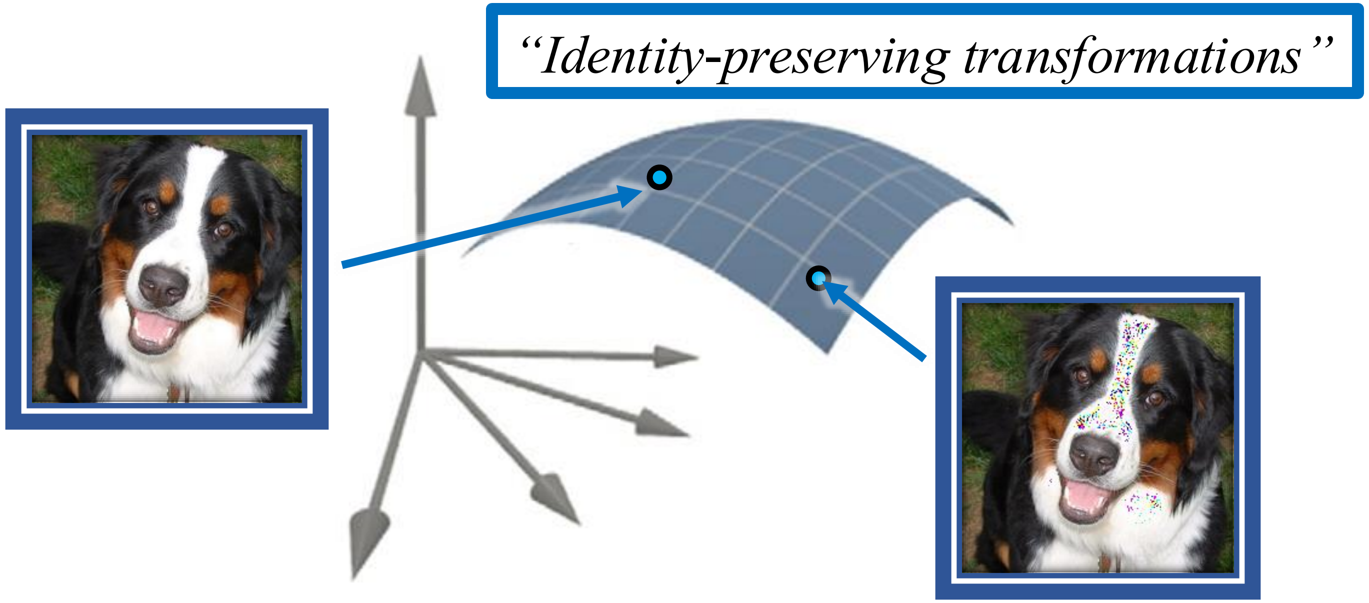
Ostrich!
Conf:~100%

Achieving invariances along visual ventral stream

- Evolving representational space achieved by disentangling object manifolds along the ventral visual stream.

  Dicarlo & Cox, Trends Cogn Sci. 2007



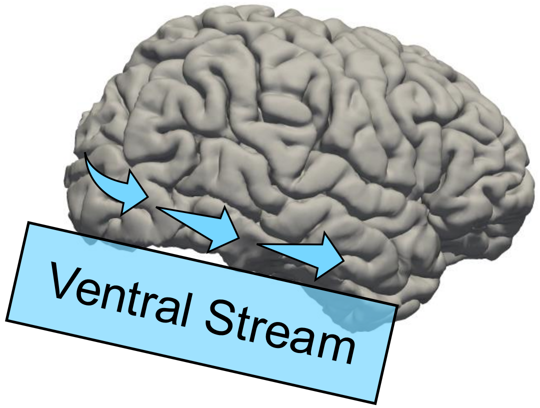


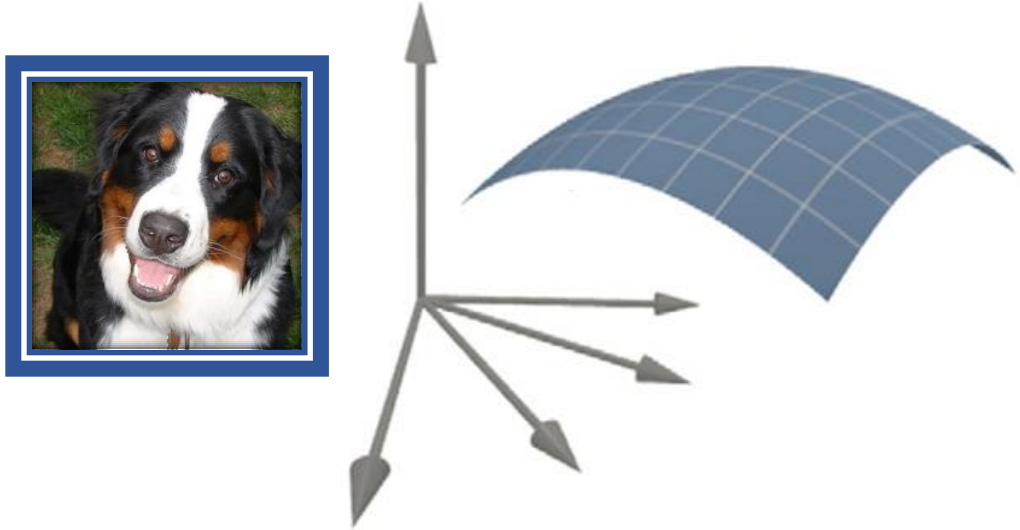
"Identity-preserving transformations"

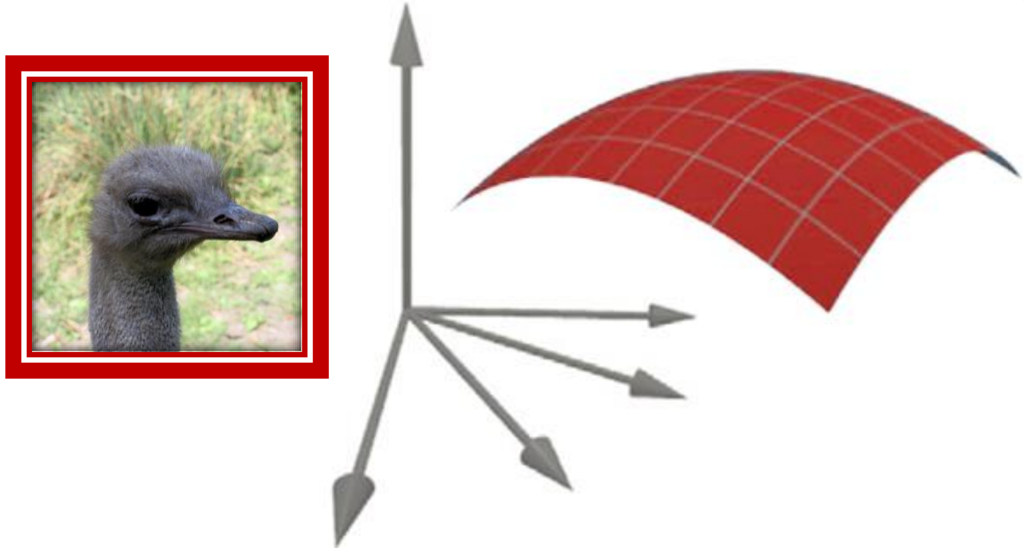
Achieving invariances along visual ventral stream

- Evolving representational space achieved by disentangling object manifolds along the ventral visual stream.

Dicarlo & Cox, Trends Cogn Sci. 2007



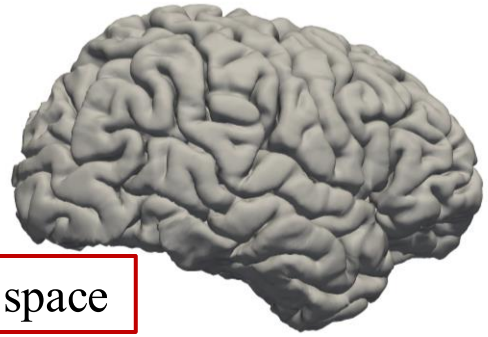




Achieving invariances along visual ventral stream

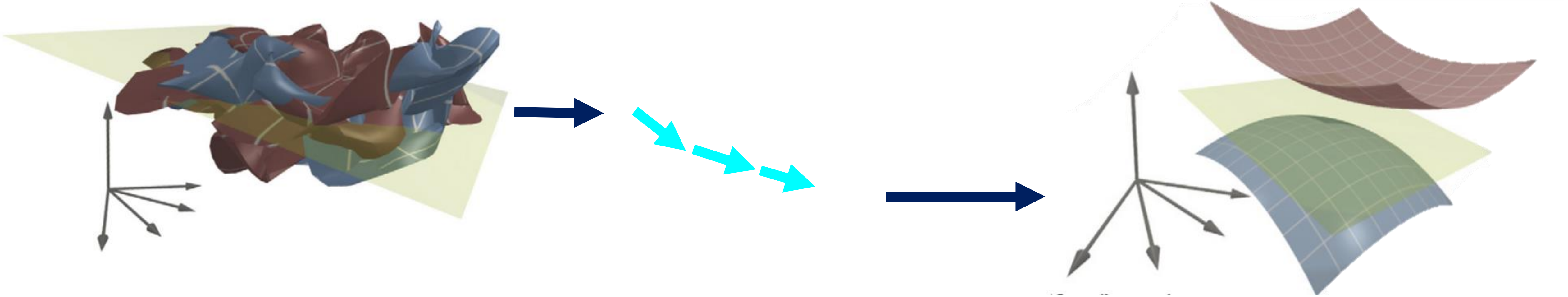
- Evolving representational space achieved by disentangling object manifolds along the ventral visual stream.

Dicarlo & Cox, Trends Cogn Sci. 2007



Pixel space

“Good” neural space

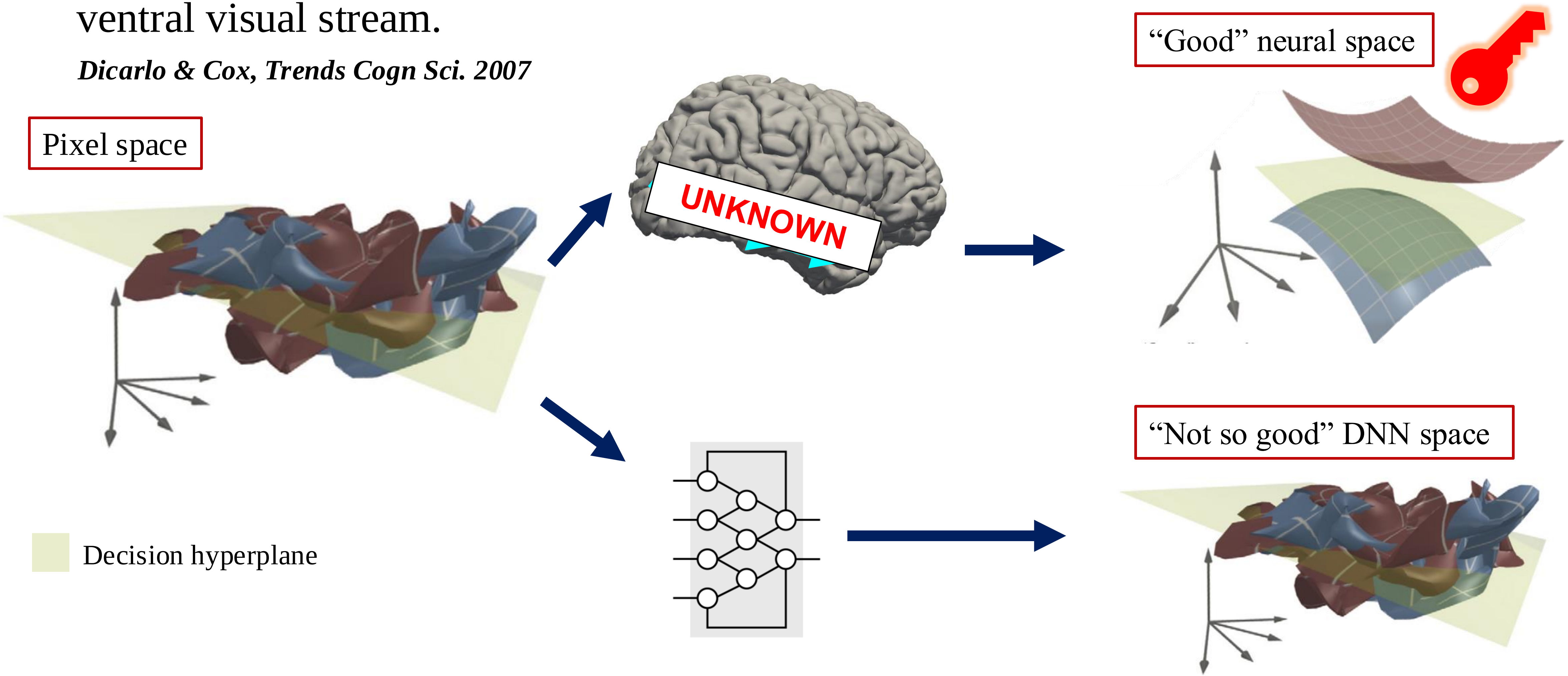


🟨 Decision hyperplane

Achieving invariances along visual ventral stream

- Evolving representational space achieved by disentangling object manifolds along the ventral visual stream.

Dicarlo & Cox, Trends Cogn Sci. 2007



Pixel space

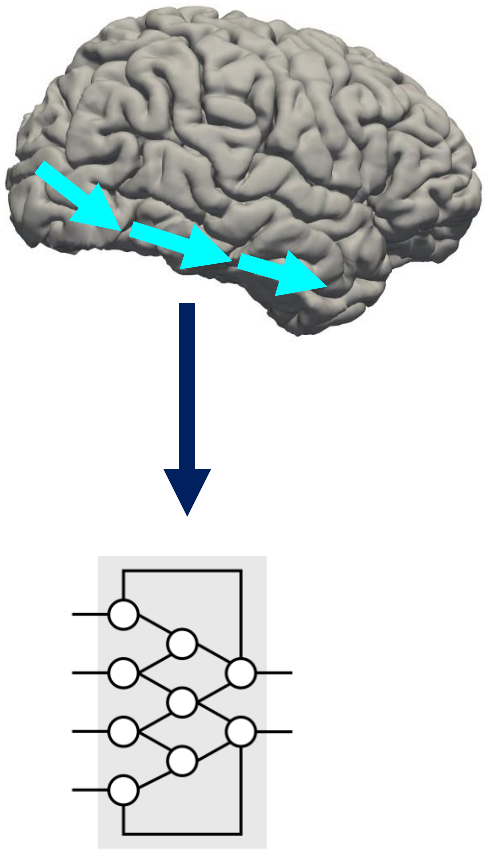
"Good" neural space

"Not so good" DNN space

🟩 Decision hyperplane

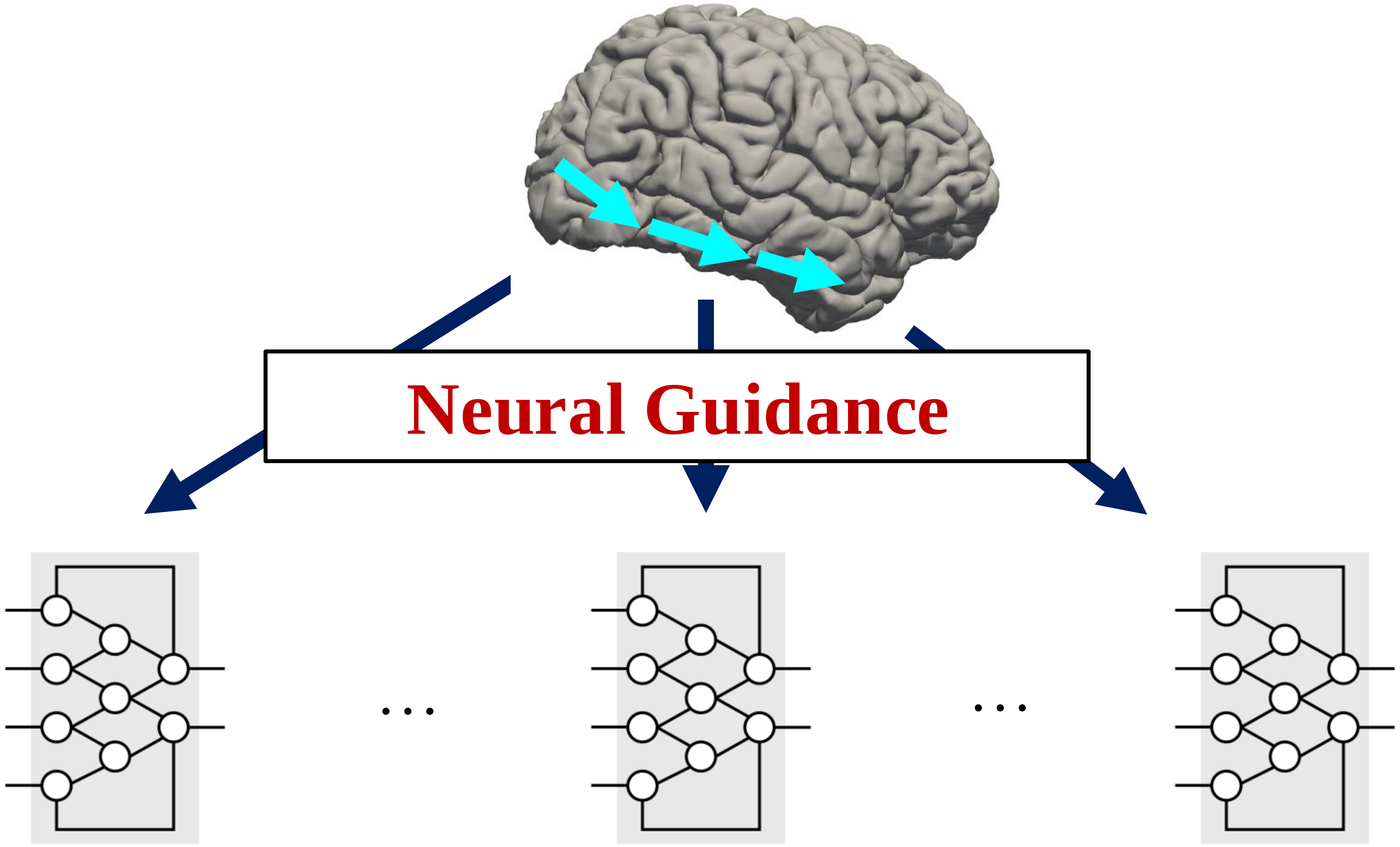
Our questions

1. Does training guided by human ventral cortex activity improve DNN robustness?



Our questions

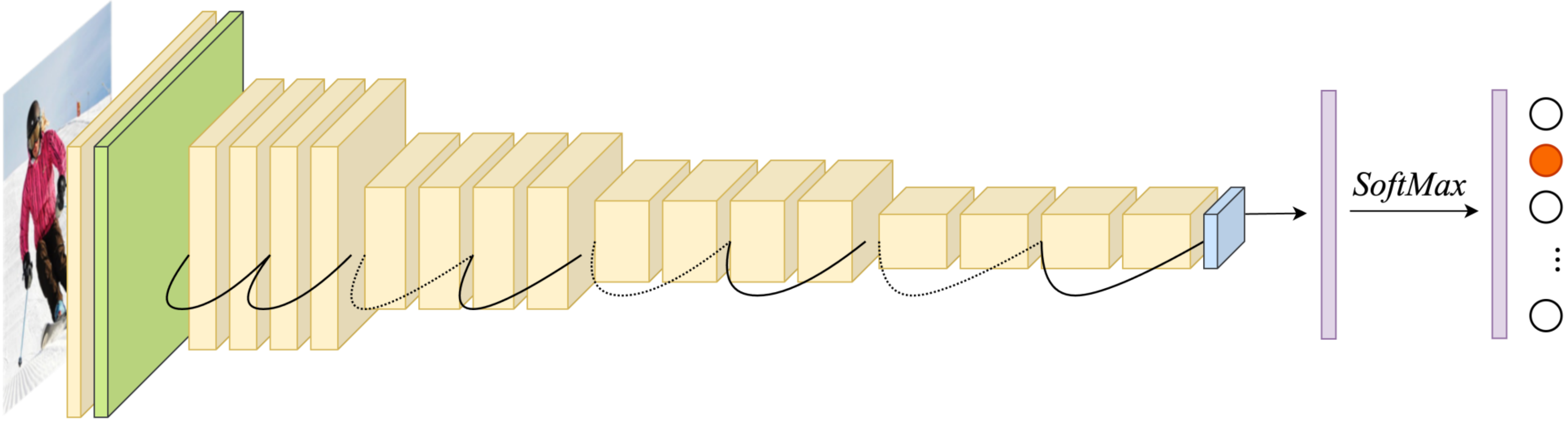
1. Does training guided by human ventral cortex activity improve DNN robustness?
2. Does such improvement increase as we ascend the ventral visual cortex?



Neural Guidance

Training with **Neural Guidance**

$$Loss_{total} = L_{task}$$



🟩 Max Pooling

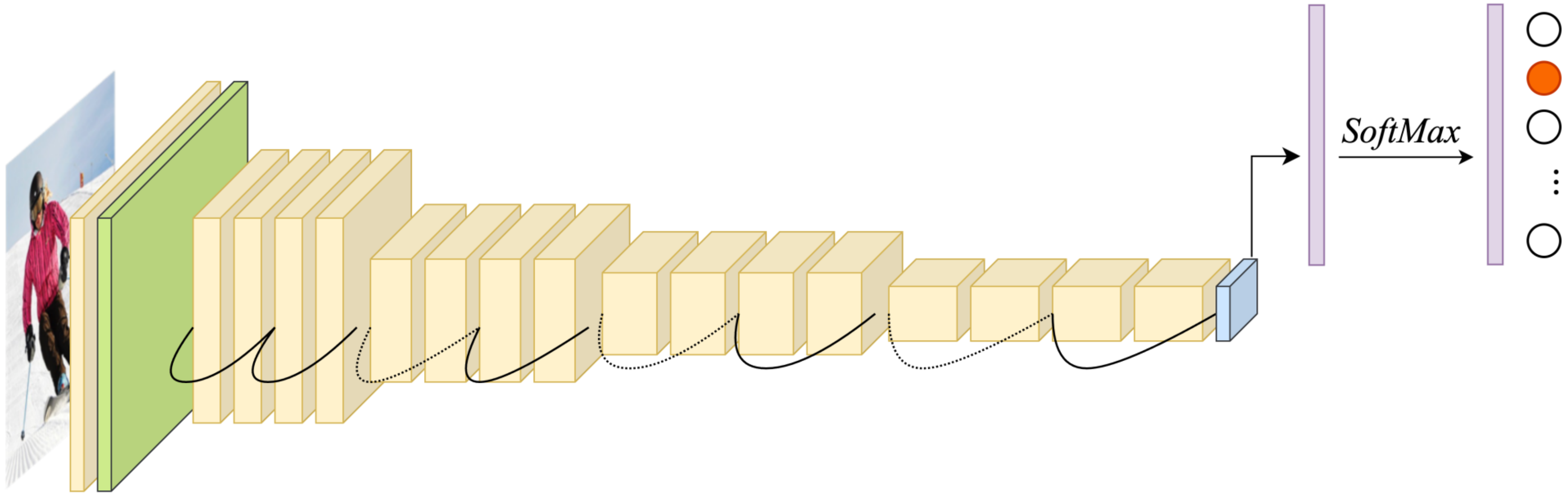
🟦 Avg Pooling

🟨 Conv

🟪 FC

Training with **Neural Guidance**

$$Loss_{total} = L_{task}$$

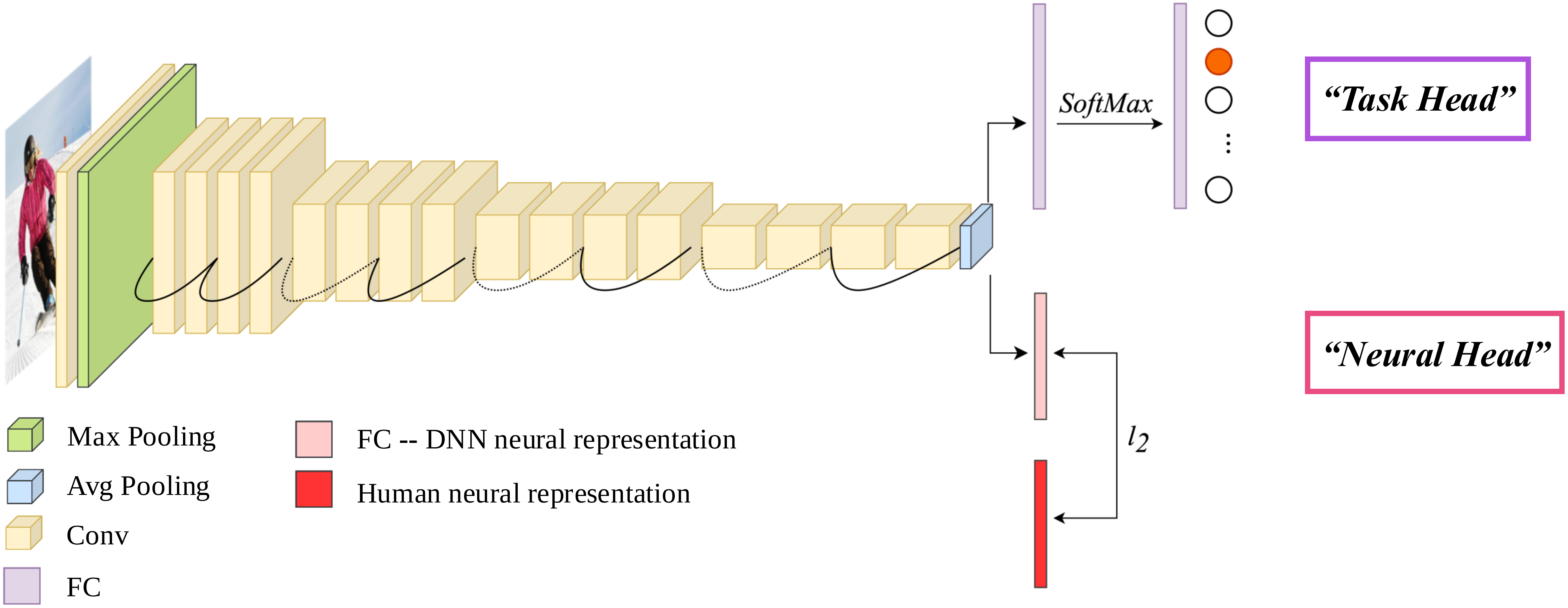


"Task Head"

- 🟩 Max Pooling
- 🟦 Avg Pooling
- 🟨 Conv
- 🟪 FC

Training with **Neural Guidance**

$$Loss_{total} = L_{task} + \left|\left|R_{DNN} - R_{neural}\right|\right|_2$$



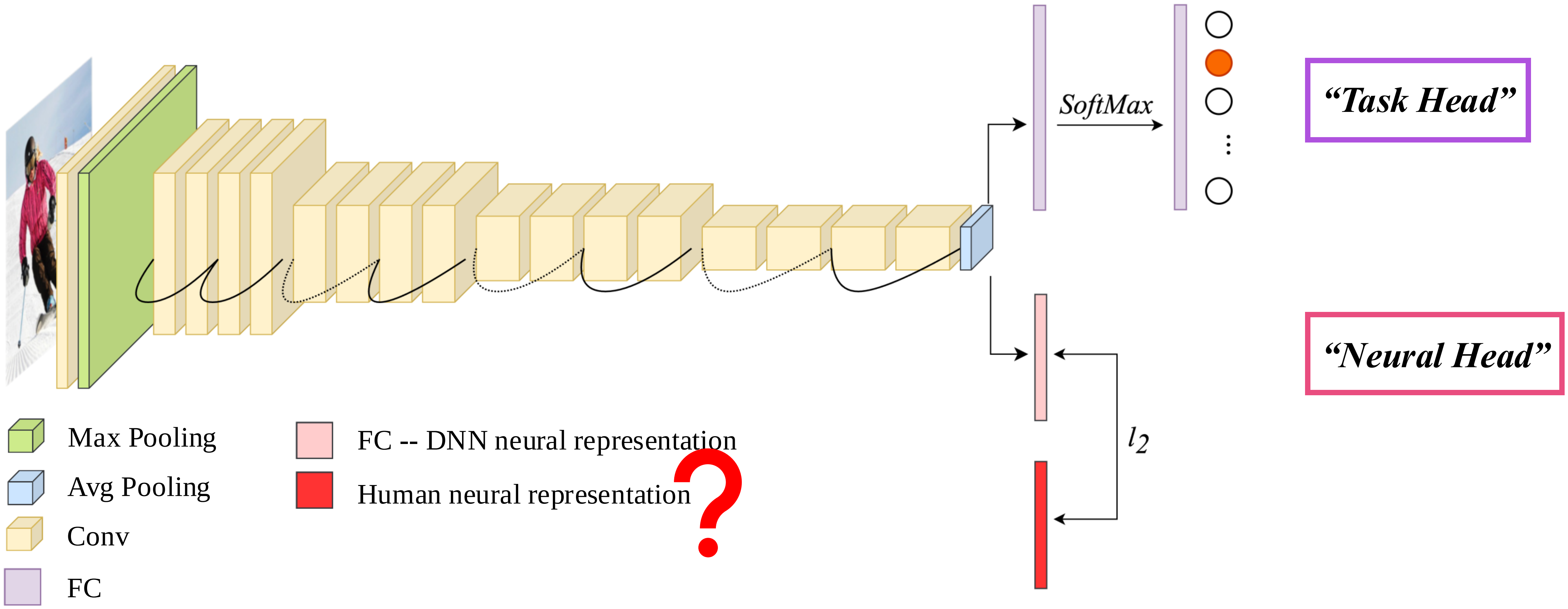
“Task Head”

“Neural Head”

- 🟩 Max Pooling
- 🟦 Avg Pooling
- 🟨 Conv
- 🟪 FC
- 🟥 FC -- DNN neural representation
- 🟥 Human neural representation

Training with **Neural Guidance**

$$Loss_{total} = \underline{\alpha} L_{task} + \underline{(1-\alpha)} ||R_{DNN} - R_{neural}||_2$$



SoftMax

l_2

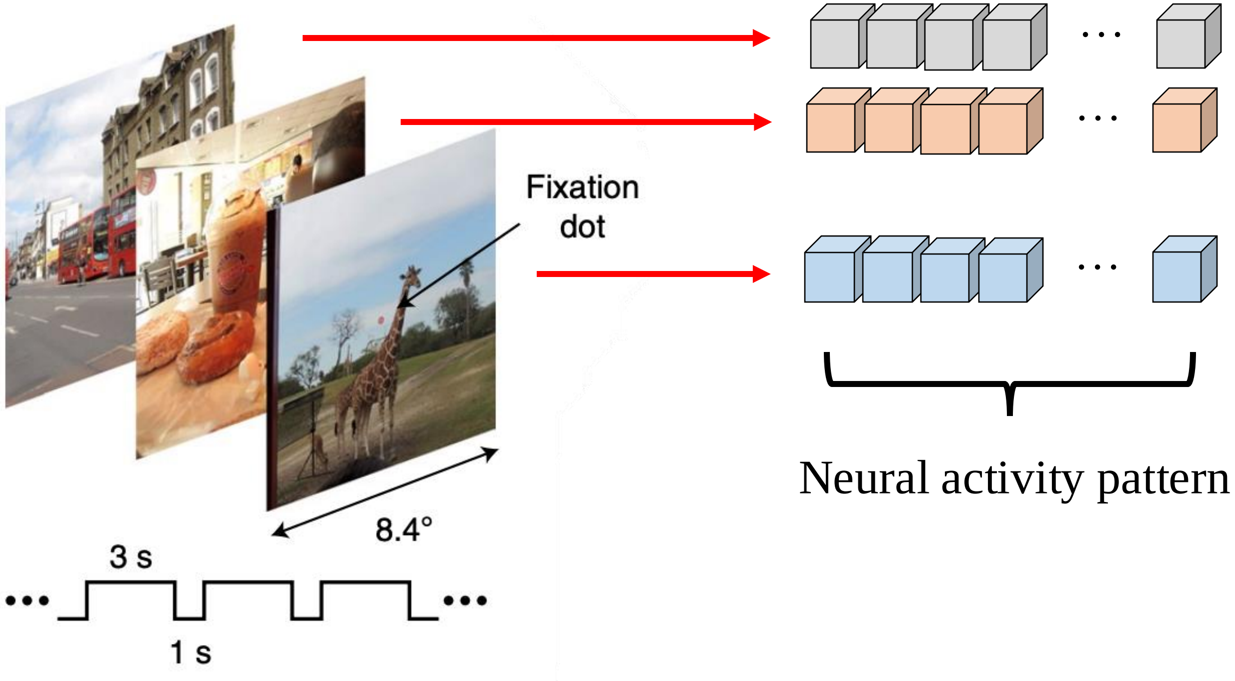
“Task Head”

“Neural Head”

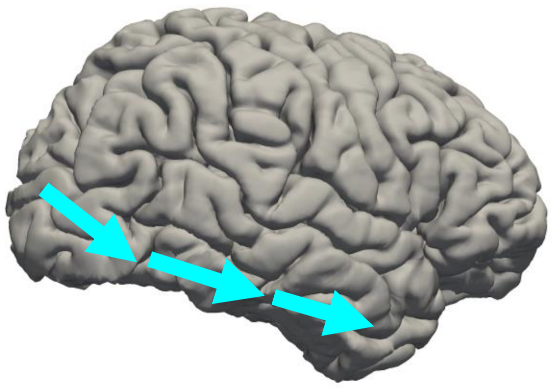
- 🟩 Max Pooling
- 🟦 Avg Pooling
- 🟨 Conv
- 🟪 FC
- 🟥 FC -- DNN neural representation
- 🟥 Human neural representation ?

Neural representation

- Brain activities were recorded with 7T fMRI while each human subject viewing 10,000 natural images. ***(Natural Scene Dataset, Allen et al., Nat. Neurosci. 2022)***



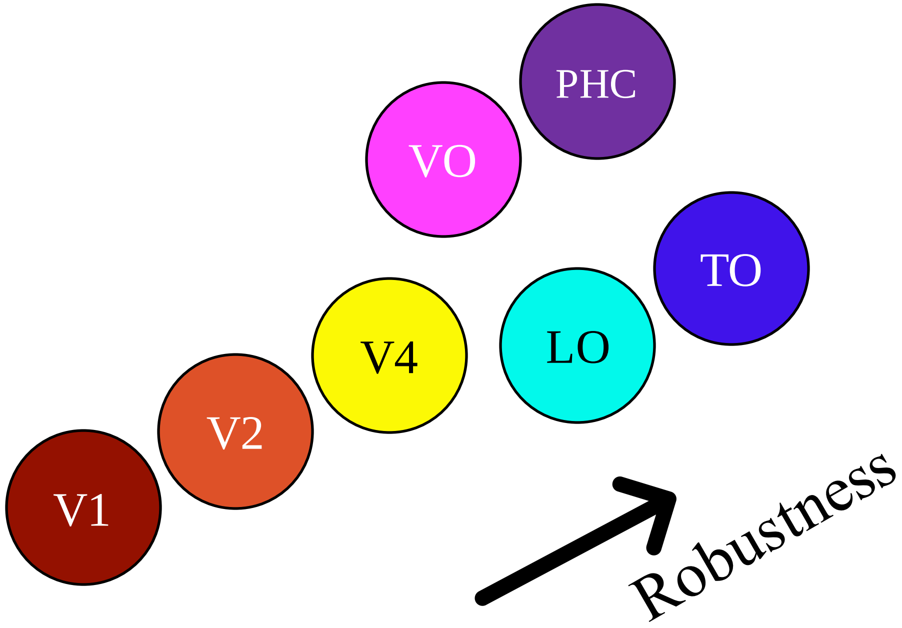
Neural activity pattern



Neural representation

- Brain activities were recorded with 7T fMRI while each human subject viewing ~30,000 natural images.
- 7 bilateral Regions of Interest (ROIs) were used

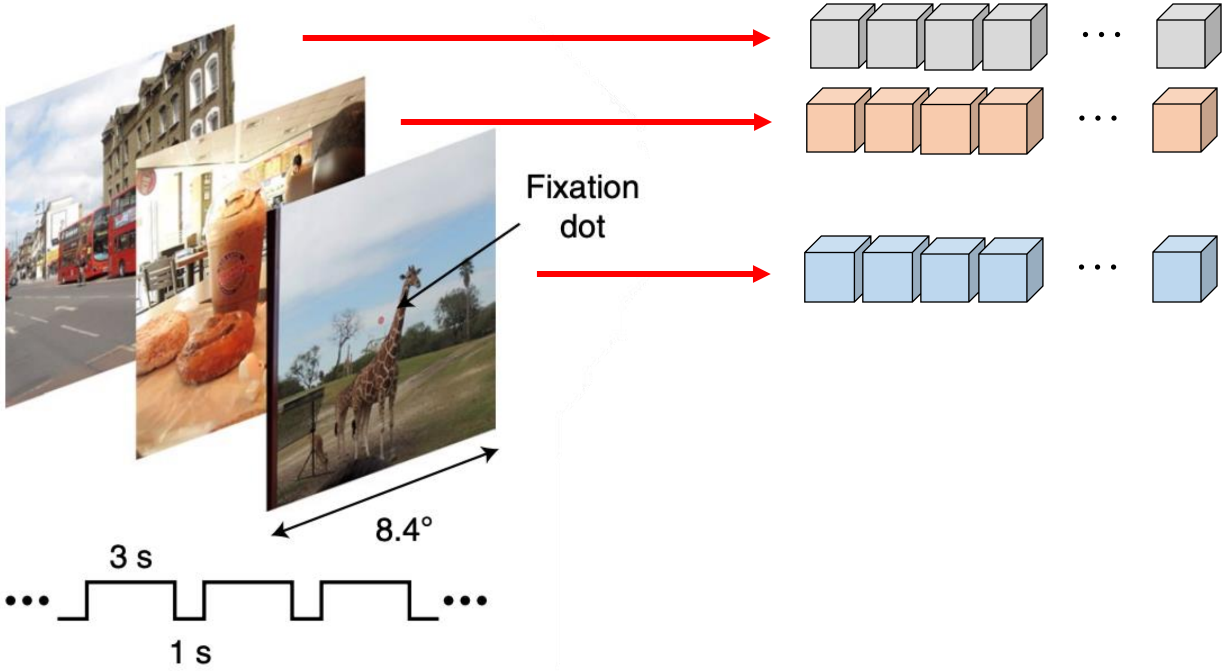
  (Wang et al., Cereb. Cortex, 2015)



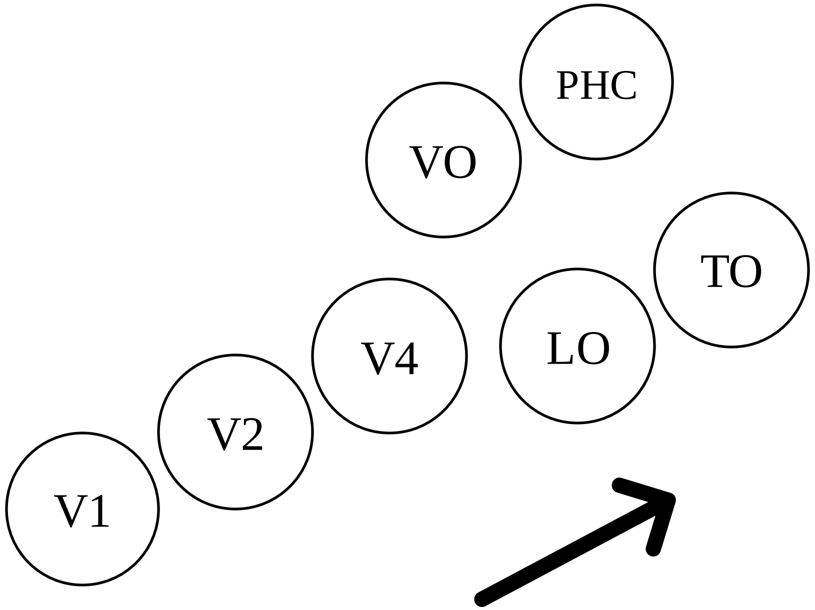
Ventral Visual Stream Hierarchy

Neural representation

- Brain activities were recorded with 7T fMRI while each human subject viewing ~30,000 natural images.
- 7 bilateral Regions of Interest (ROIs) were used

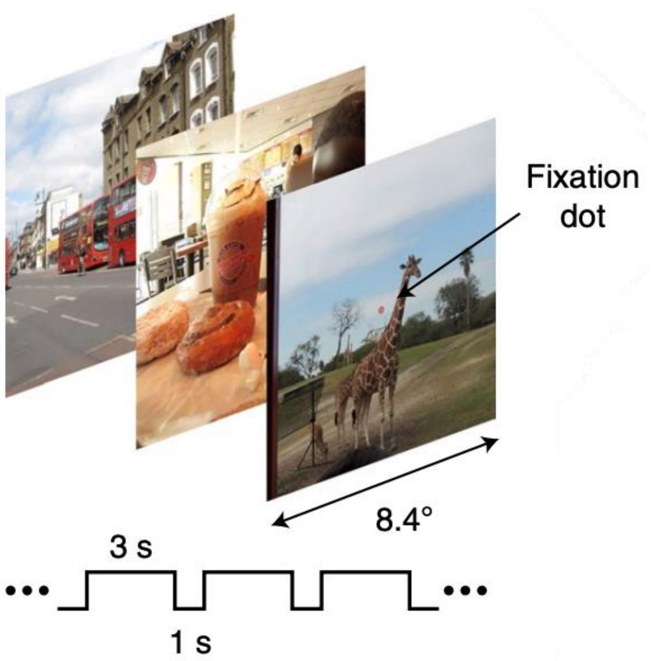


(NSD, Allen et al., Nat. Neurosci. 2022)

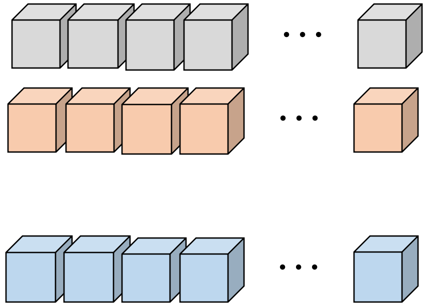


Neural representation

- Brain activities were recorded with 7T fMRI while each human subject viewing ~30,000 natural images.
- 7 bilateral Regions of Interest (ROIs) were used

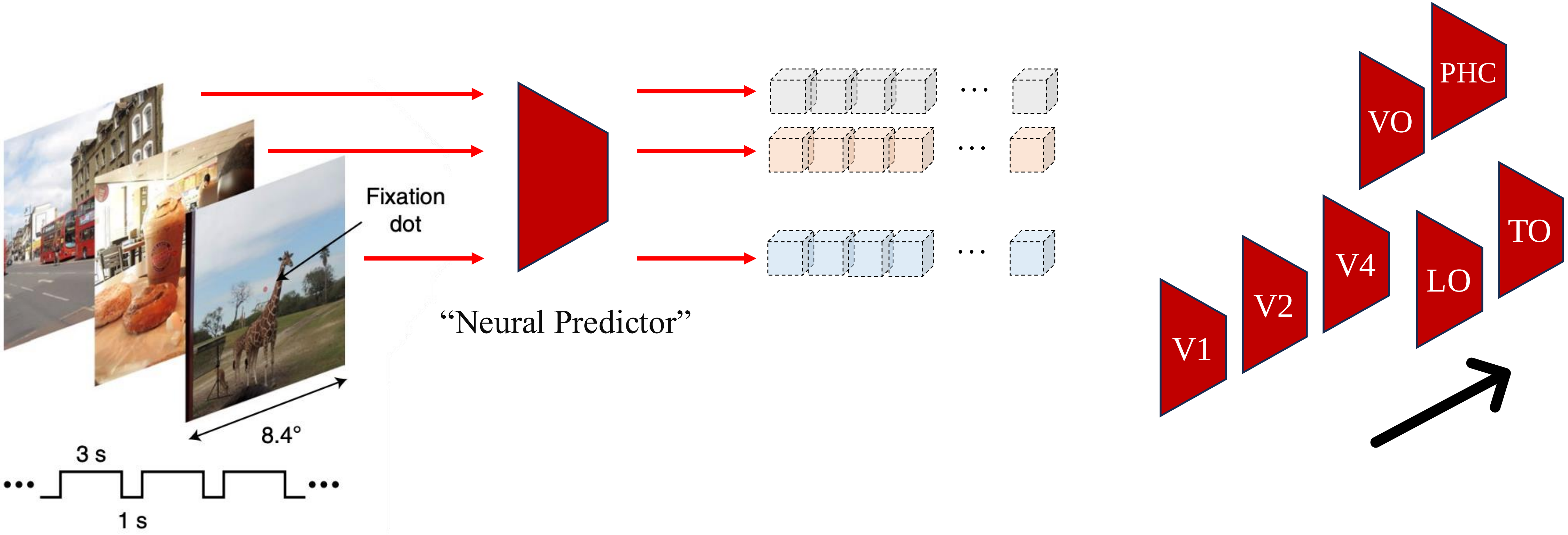


(NSD, Allen et al., Nat. Neurosci. 2022)



Neural representation

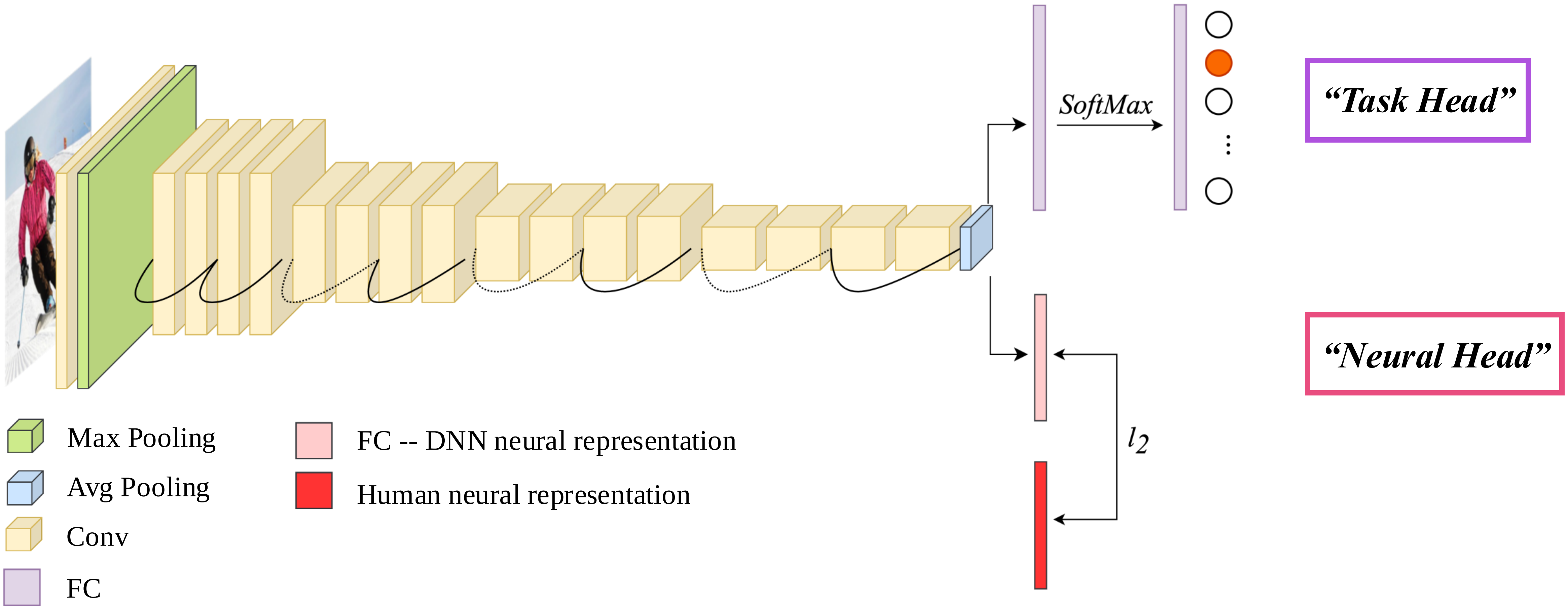
- Brain activities were recorded with 7T fMRI while each human subject viewing ~30,000 natural images.
- 7 bilateral Regions of Interest (ROIs) were used



(NSD, Allen et al., Nat. Neurosci. 2022)

Training with **Neural Guidance**

$$Loss_{total} = \alpha L_{task} + (1-\alpha)\left|\left|R_{DNN} - R_{neural}\right|\right|_2$$



“Task Head”

“Neural Head”

🟩 Max Pooling

🟦 Avg Pooling

🟨 Conv

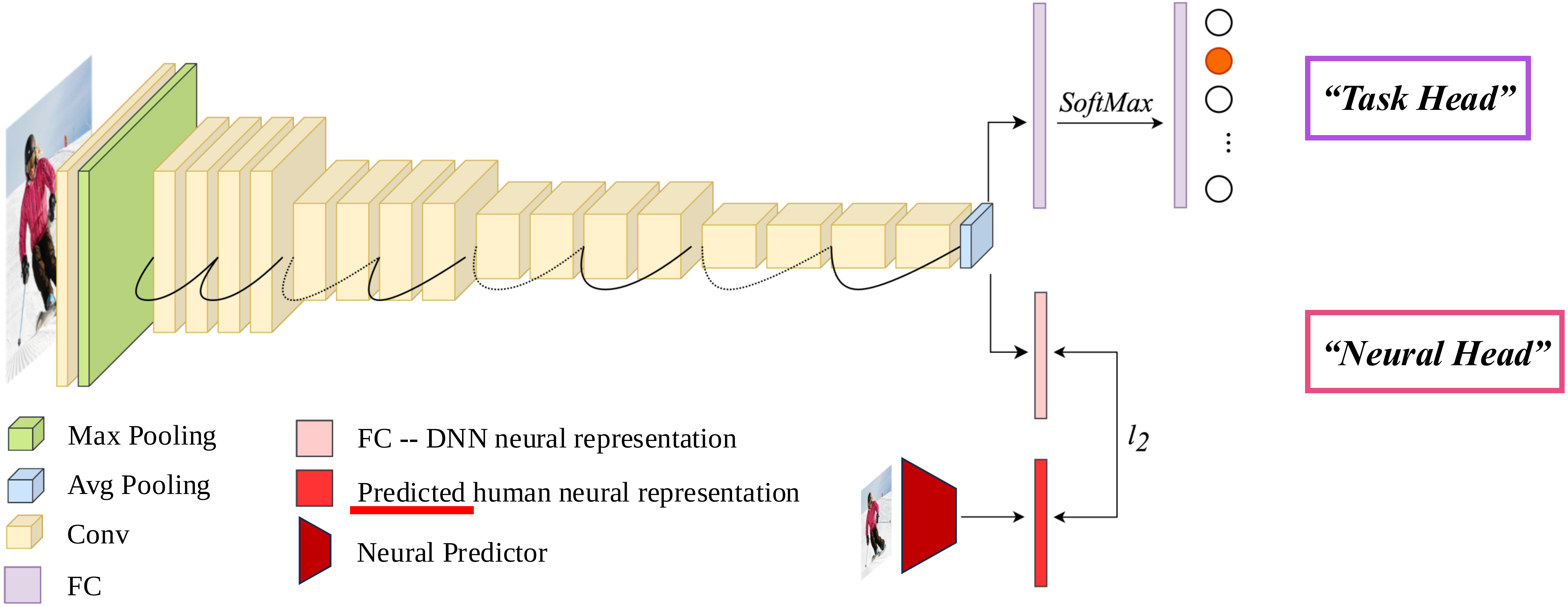
🟪 FC

🟥 FC -- DNN neural representation

🟥 Human neural representation

Training with **Neural Guidance**

$$Loss_{total} = \alpha L_{task} + (1-\alpha)||R_{DNN} - R_{neural}||_2$$



SoftMax

l_2

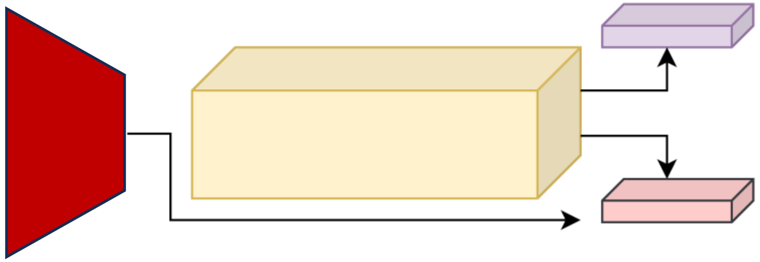
“Task Head”

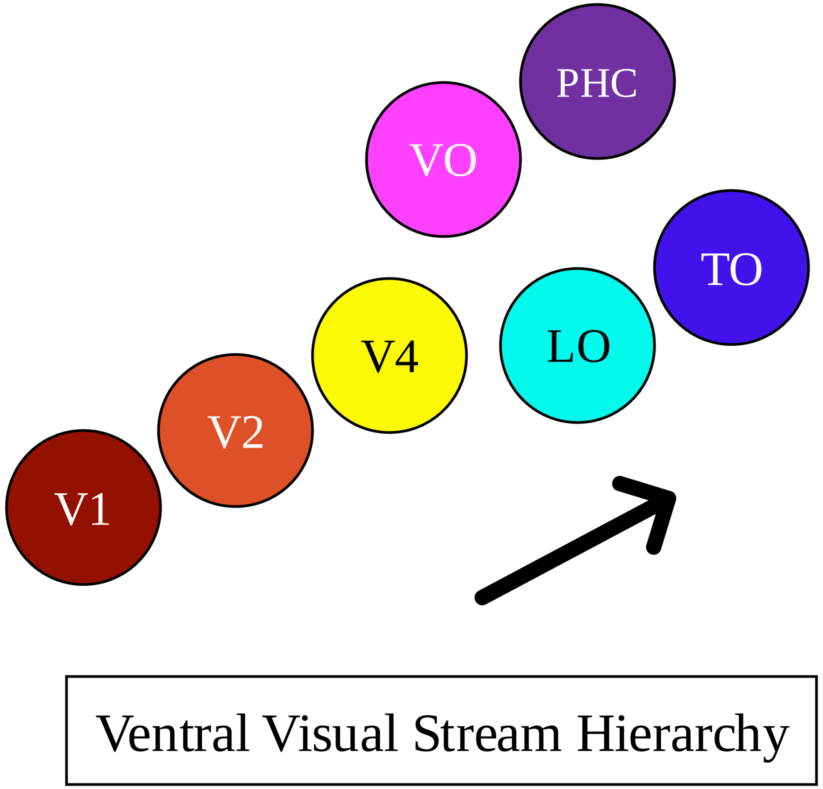
“Neural Head”

- Max Pooling
- Avg Pooling
- Conv
- FC
- FC -- DNN neural representation
- Predicted human neural representation
- Neural Predictor

Summary of models

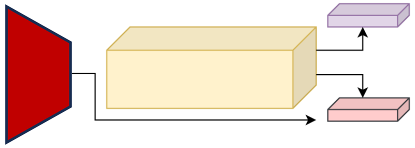
- 7 DNNs trained with Neural Guidance



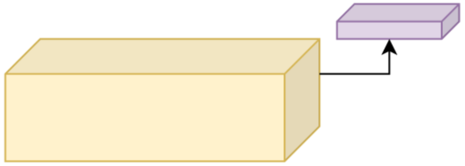


Summary of models

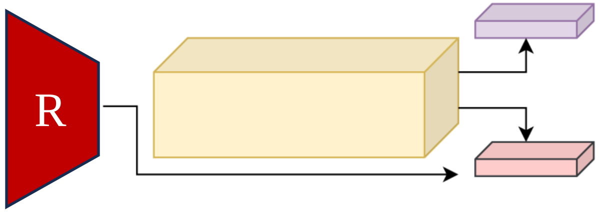
- 7 DNNs trained with Neural Guidance



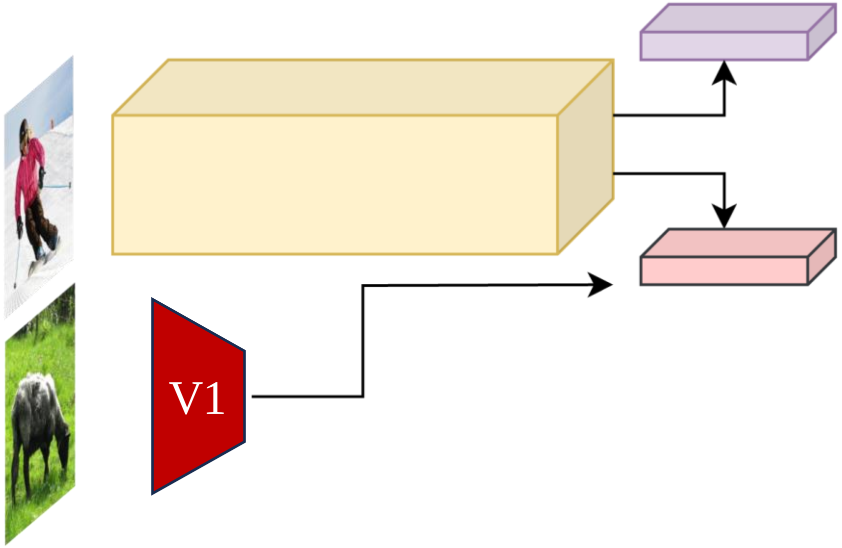
- 4 baseline models for comparison



“None”



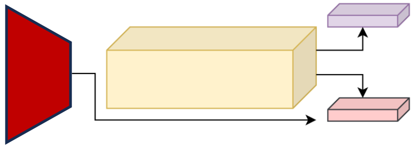
“Random”



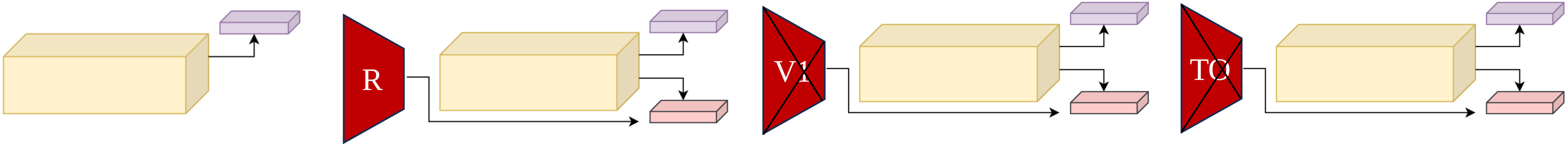
“V1-shuffle”

Summary of models

- 7 DNNs trained with Neural Guidance



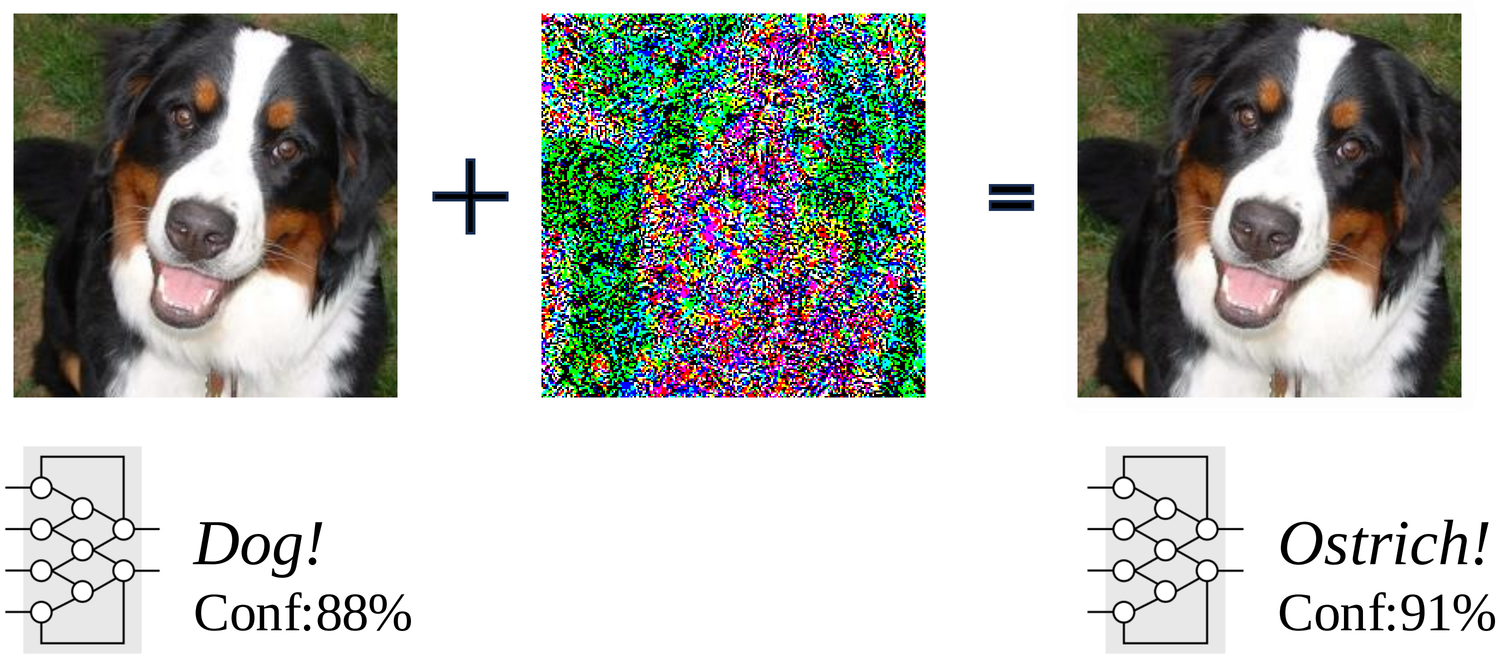
- 4 baseline models for comparison



"None" *"Random"* *"V1-shuffle"* *"TO-shuffle"*

Evaluating DNN robustness

- $\boldsymbol{l_p}$**-based adversarial attack:** $\max_{||\tau||_p<\epsilon} l(f_\theta(x+\tau),y)$

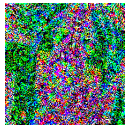


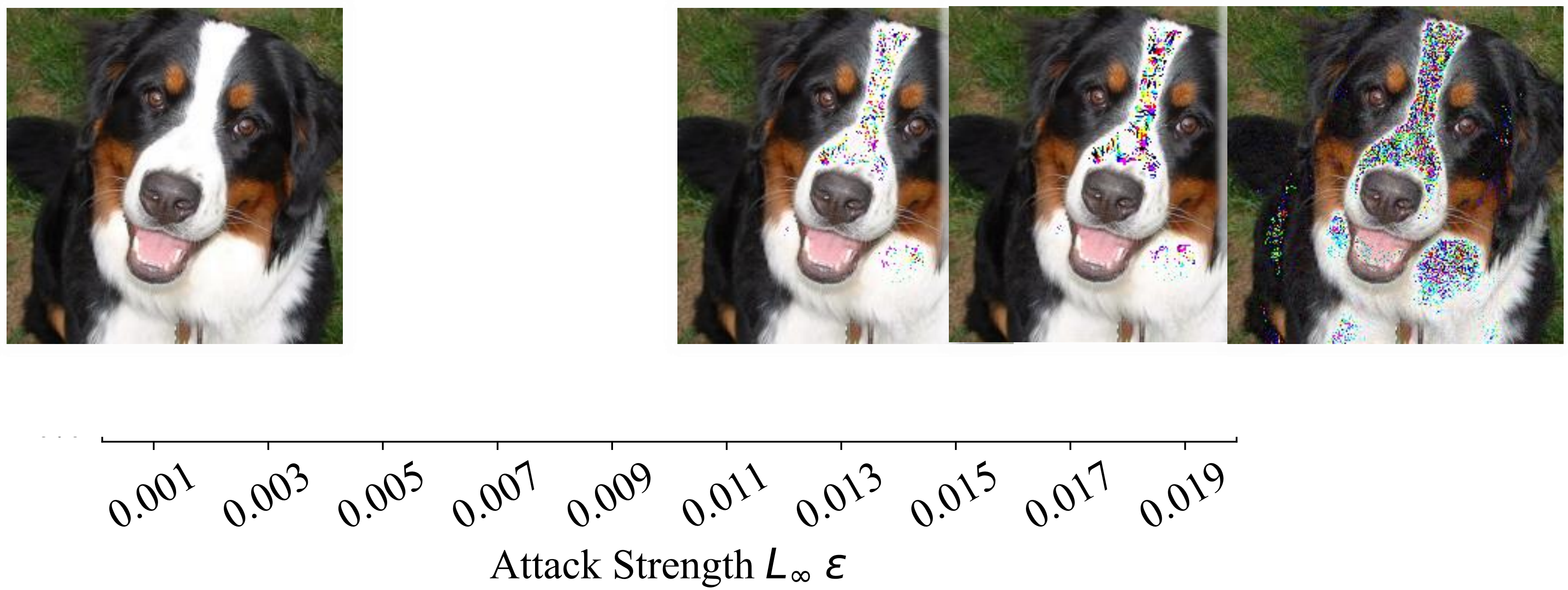
Dog!
Conf:88%

Ostrich!
Conf:91%

Evaluating DNN robustness

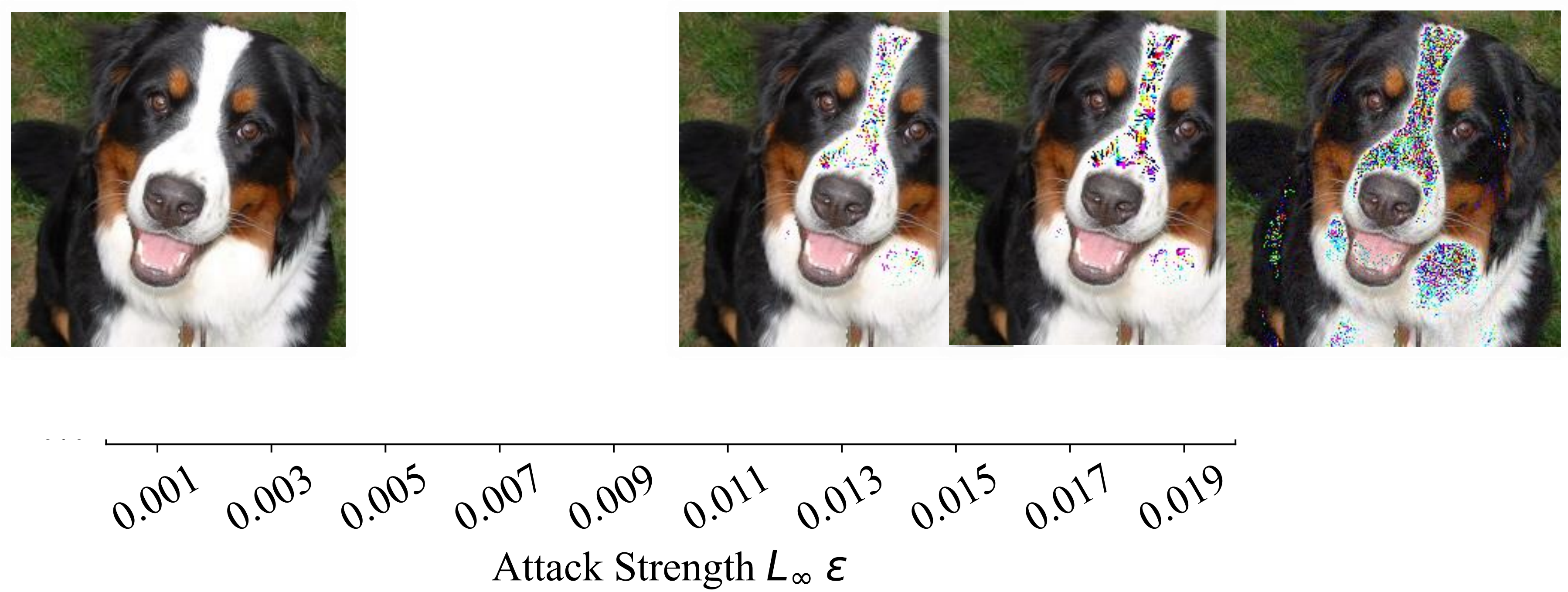
- ***l_p*-based adversarial attack:** $\max_{||\tau||_p <\epsilon} l(f_\theta(x+\tau), y)$



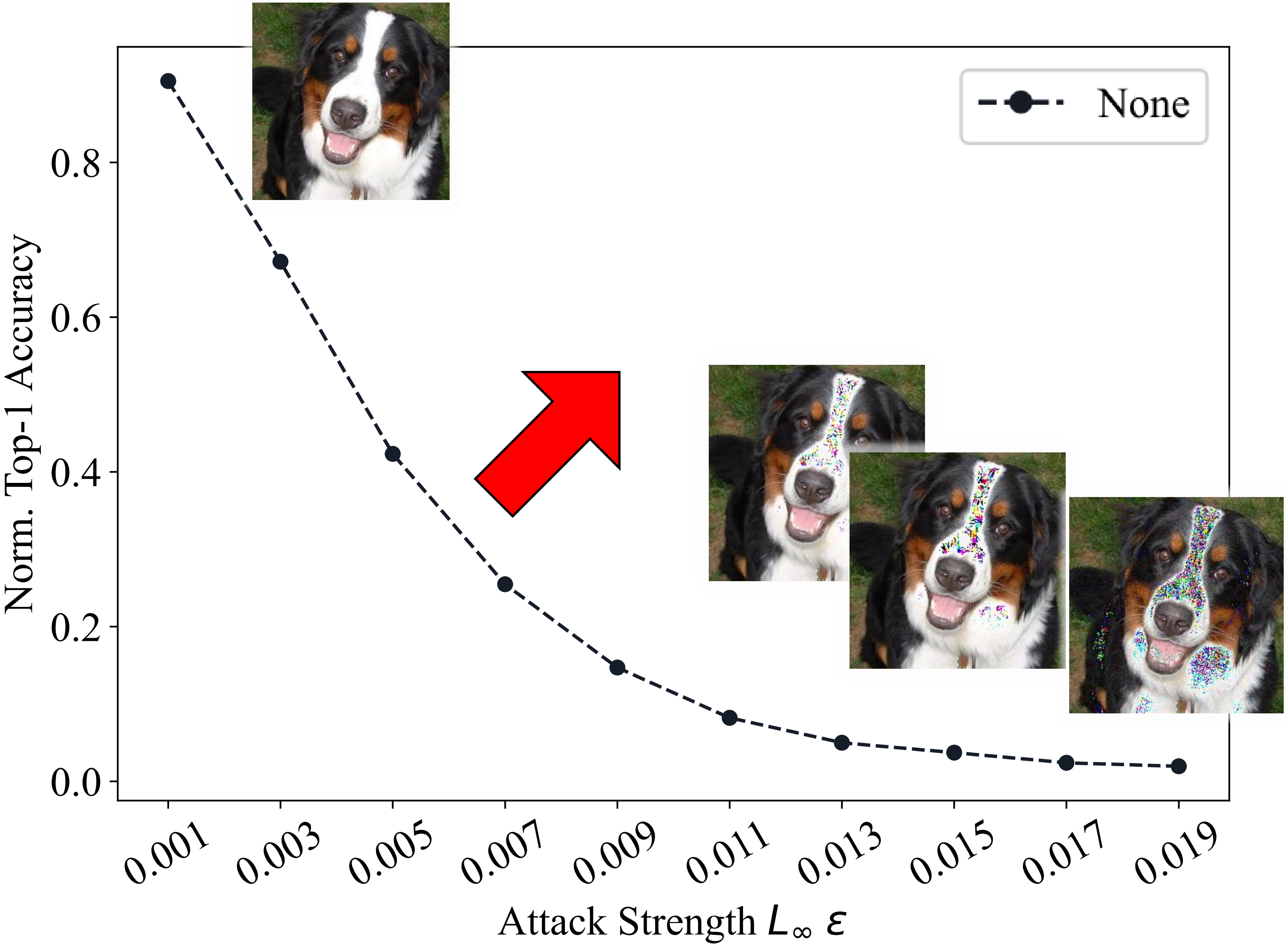


Evaluating DNN robustness

Task: Image Classification
Dataset: ImageNet (Deng et al., 2009)
Attack: l_∞-based PGD attack



Evaluating DNN robustness



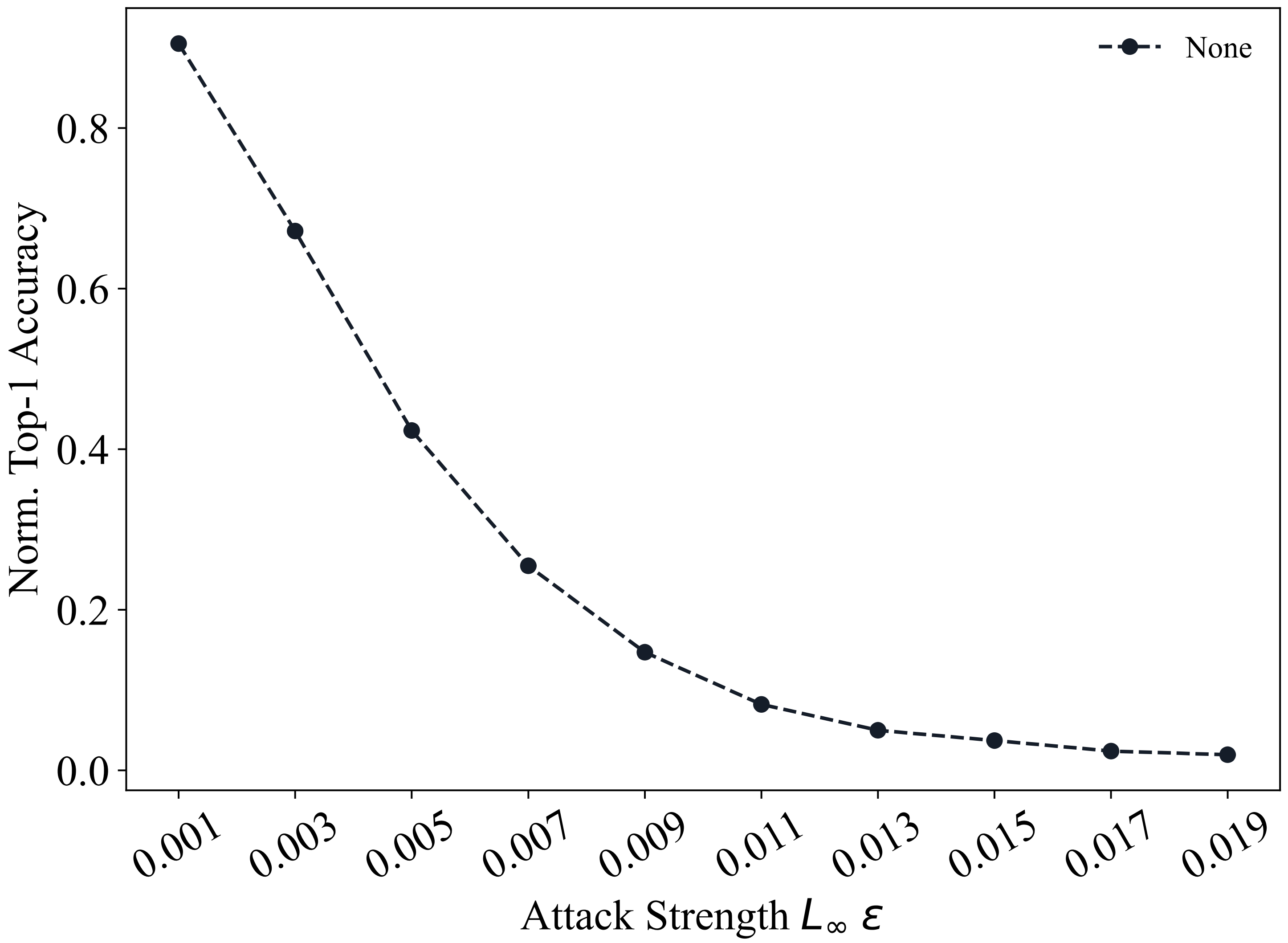
Task: Image Classification

Dataset: ImageNet (Deng et al., 2009)

Attack: l_∞-based PGD attack

None

Evaluating DNN robustness - Results

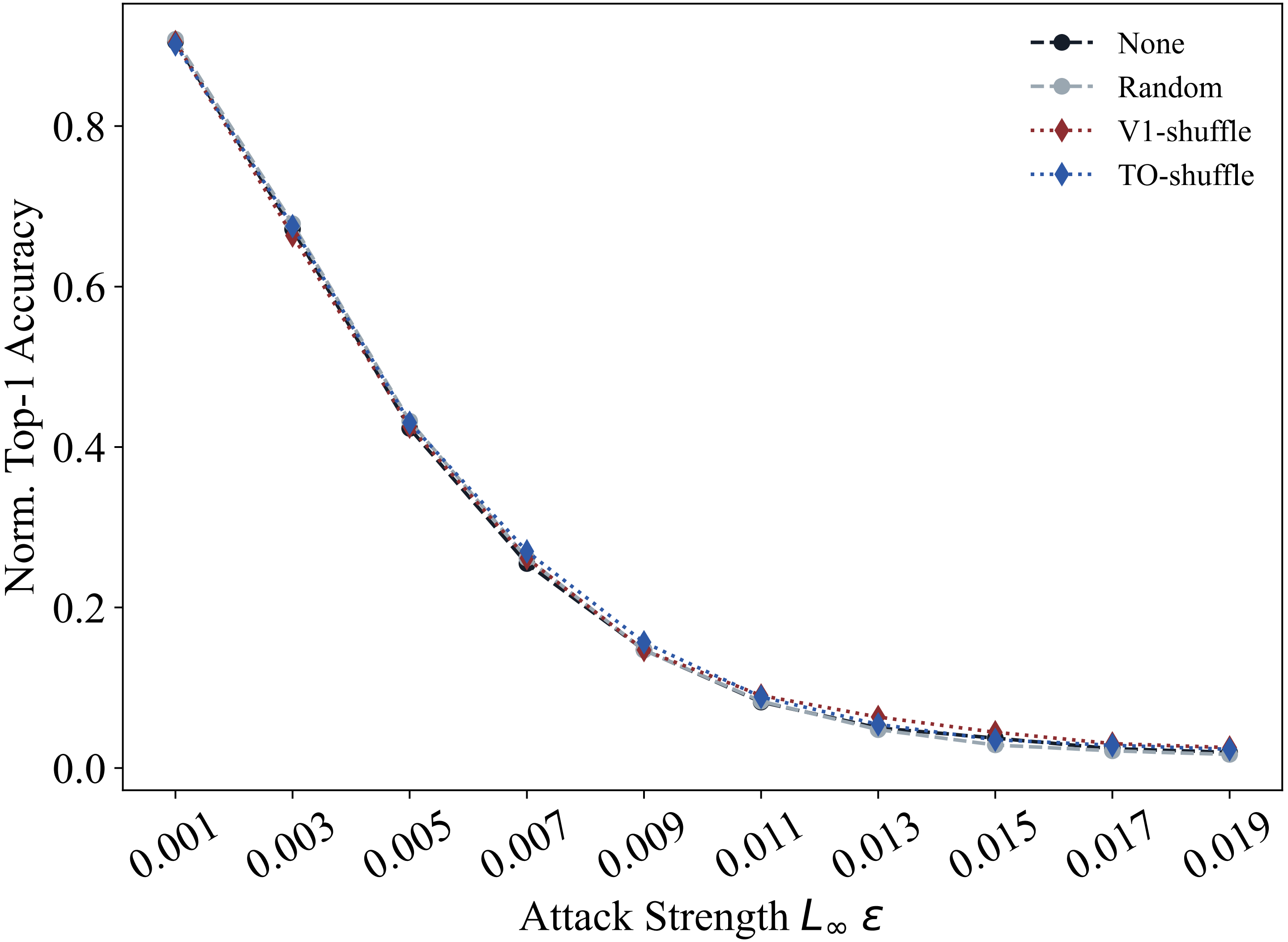


Task: Image Classification
Dataset: ImageNet (Deng et al., 2009)
Attack: l_∞-based PGD attack

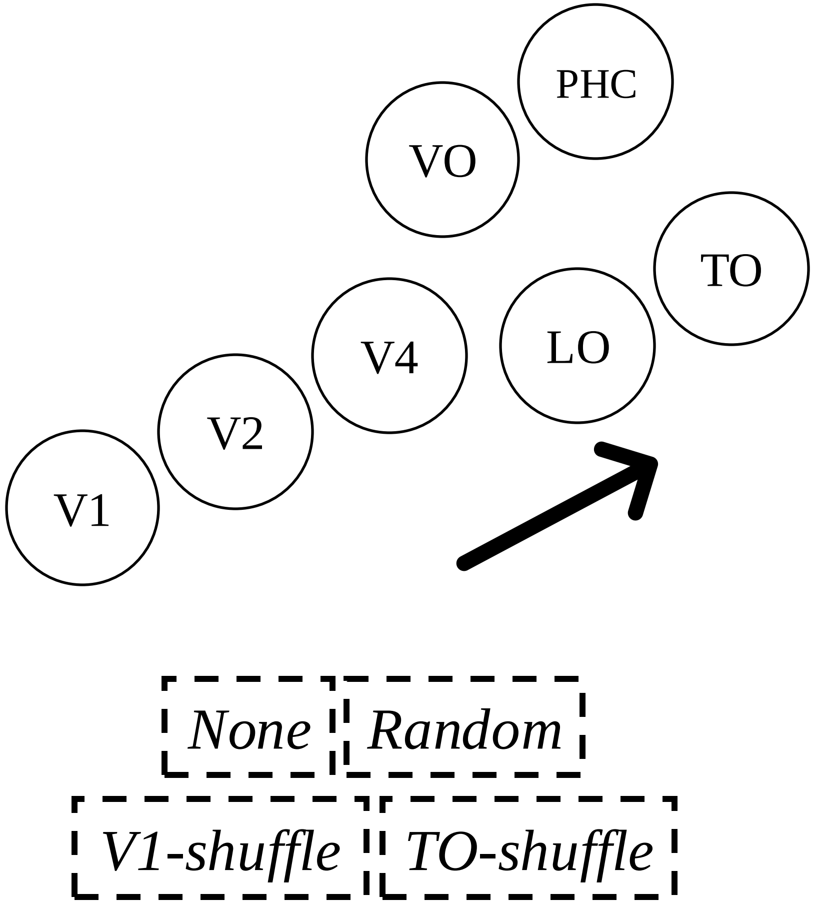
None *Random*

V1-shuffle *TO-shuffle*

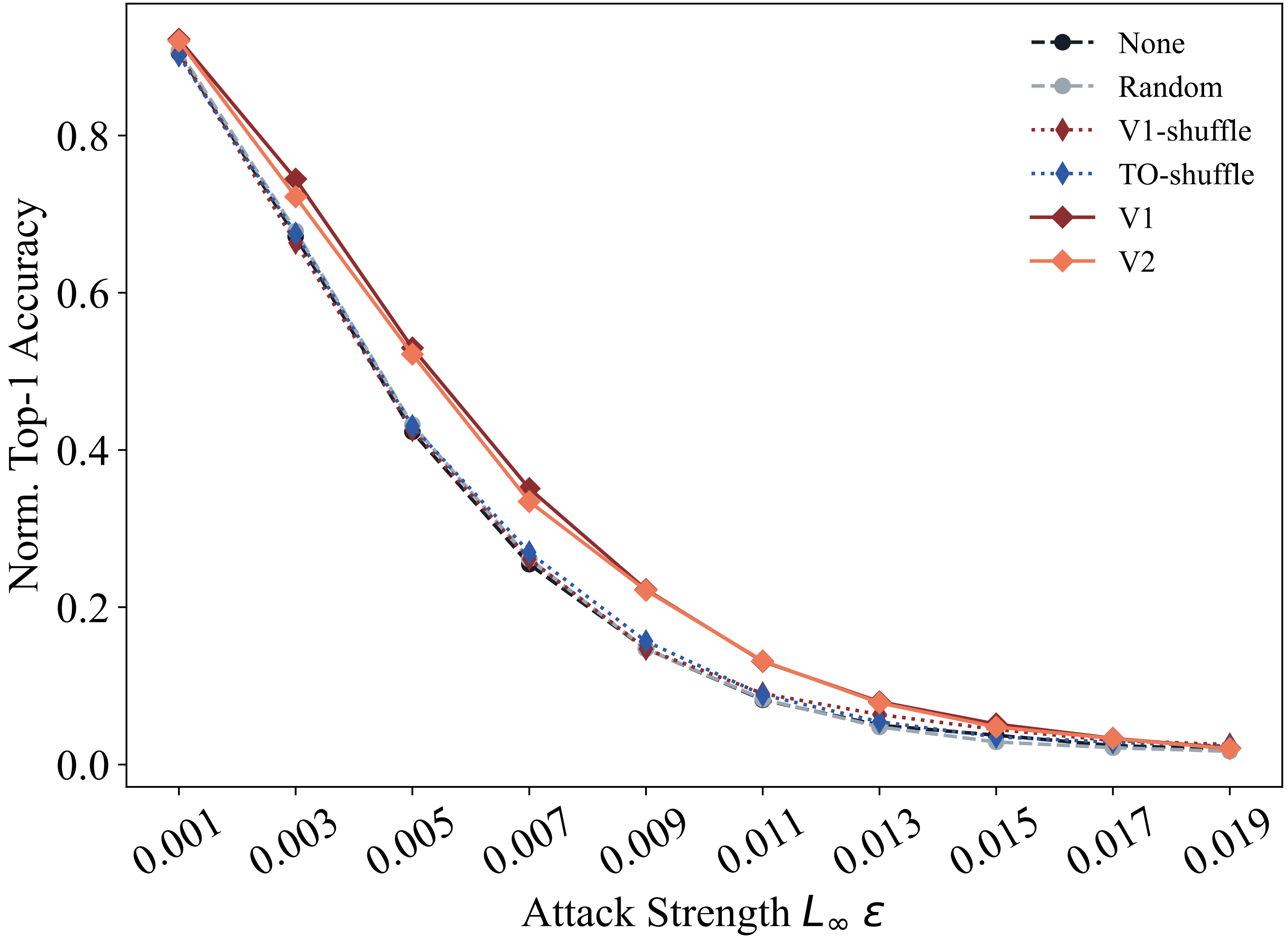
Evaluating DNN robustness - Results



Task: Image Classification
Dataset: ImageNet (Deng et al., 2009)
Attack: l_∞-based PGD attack



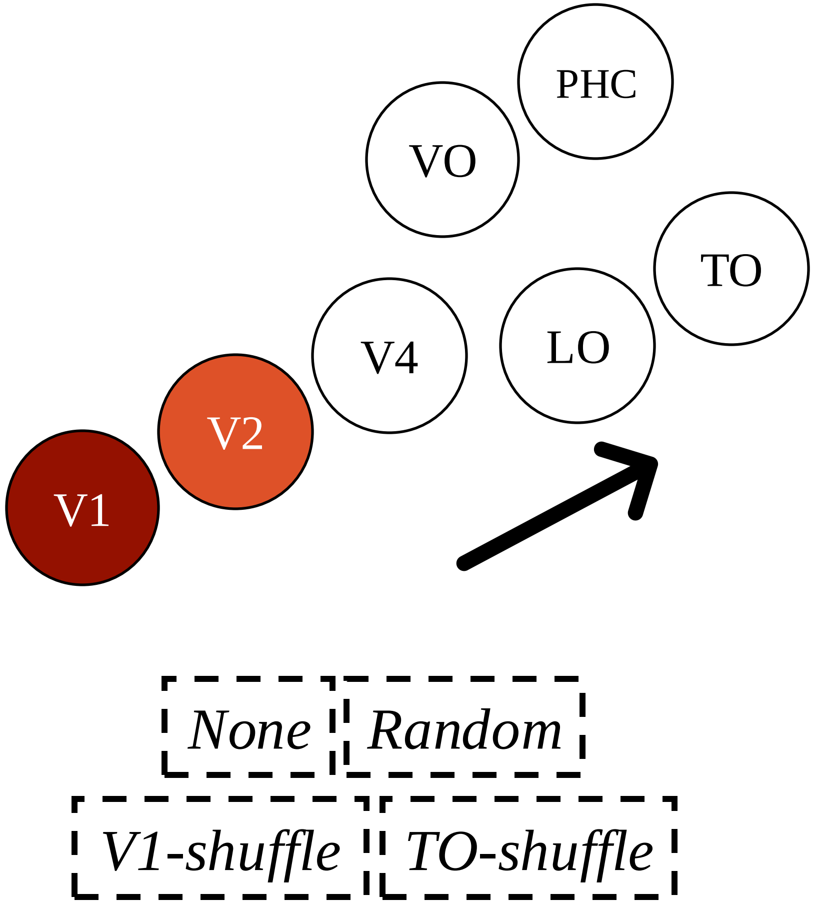
Evaluating DNN robustness - Results



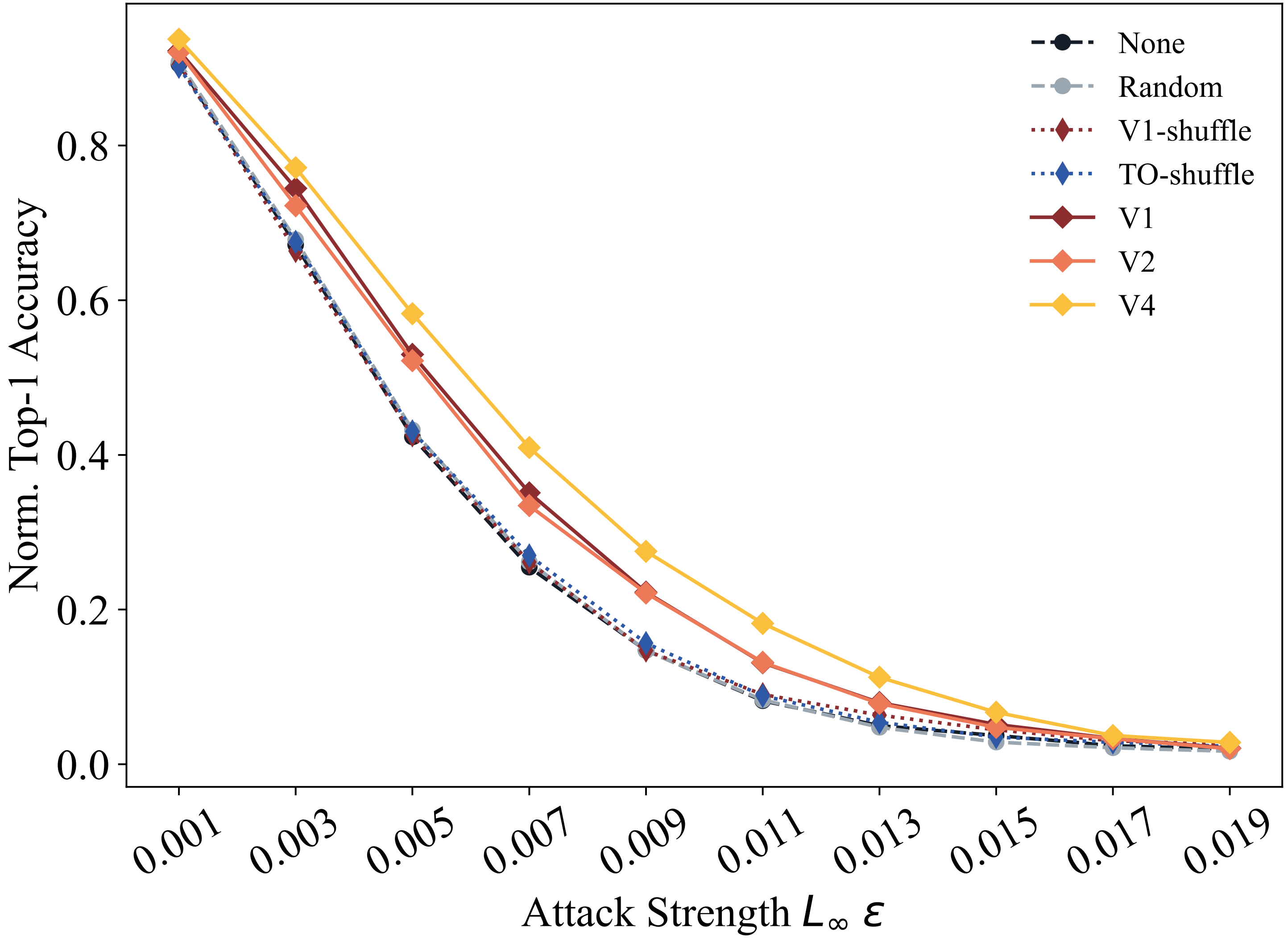
Task: Image Classification

Dataset: ImageNet (Deng et al., 2009)

Attack: l_∞-based PGD attack



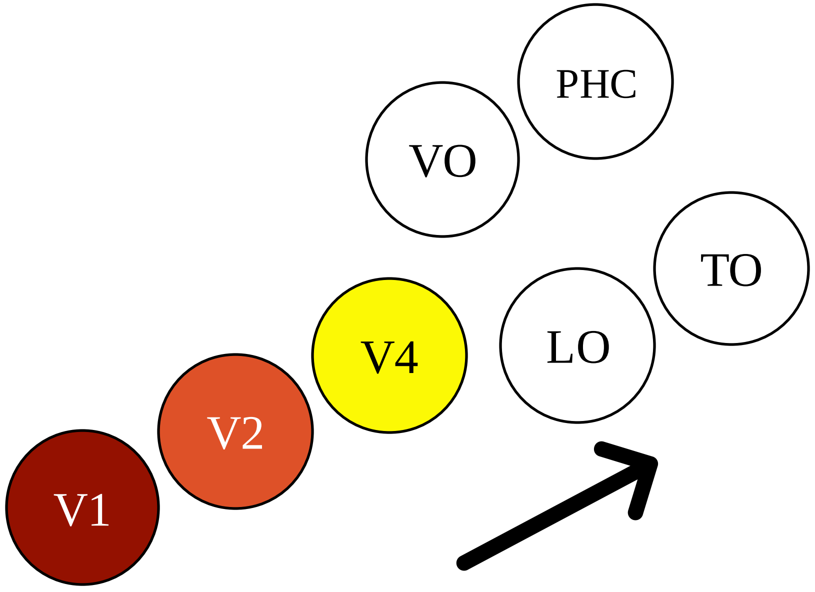
Evaluating DNN robustness - Results



Task: Image Classification

Dataset: ImageNet (Deng et al., 2009)

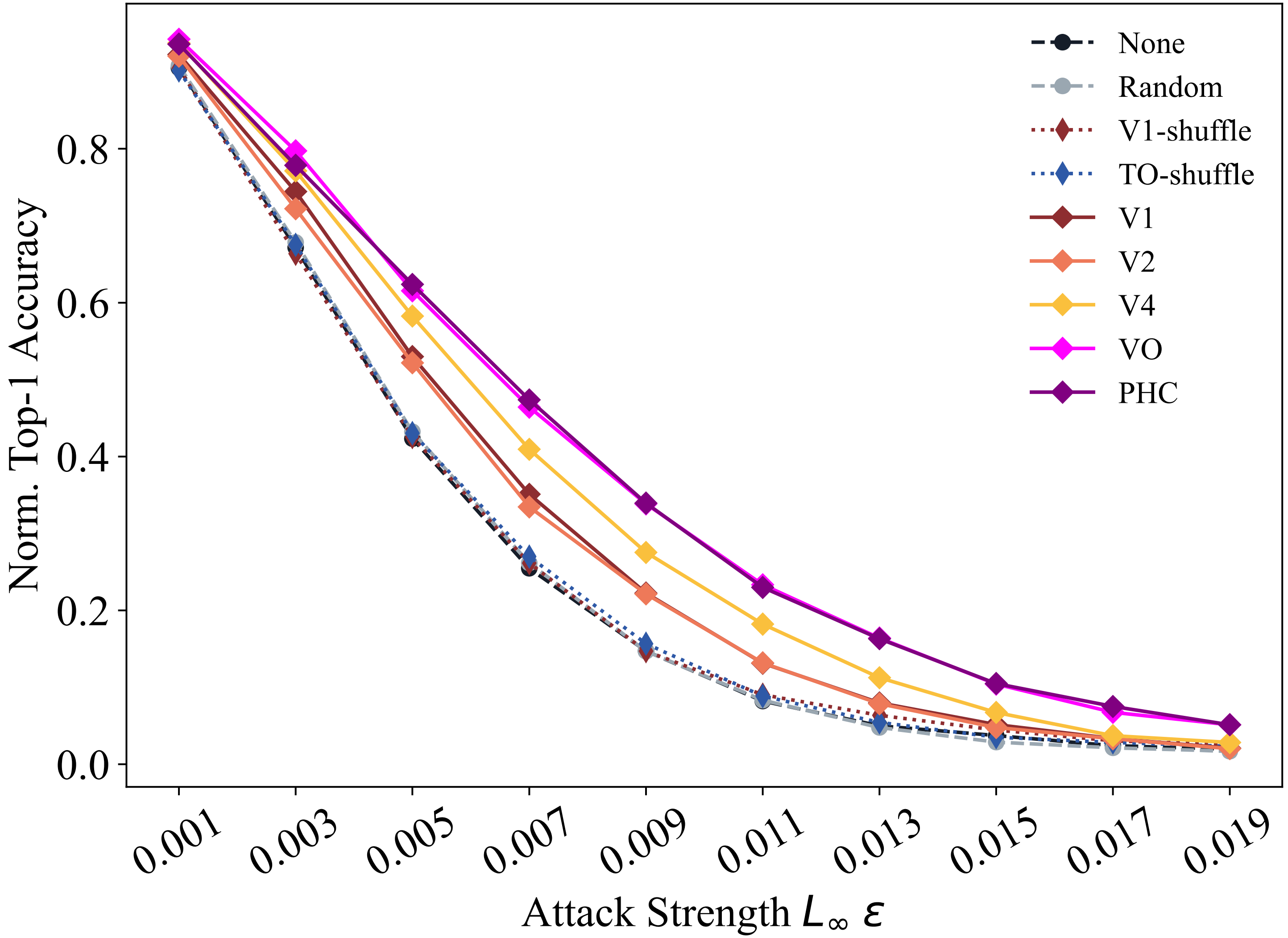
Attack: l_∞-based PGD attack



None *Random*

V1-shuffle *TO-shuffle*

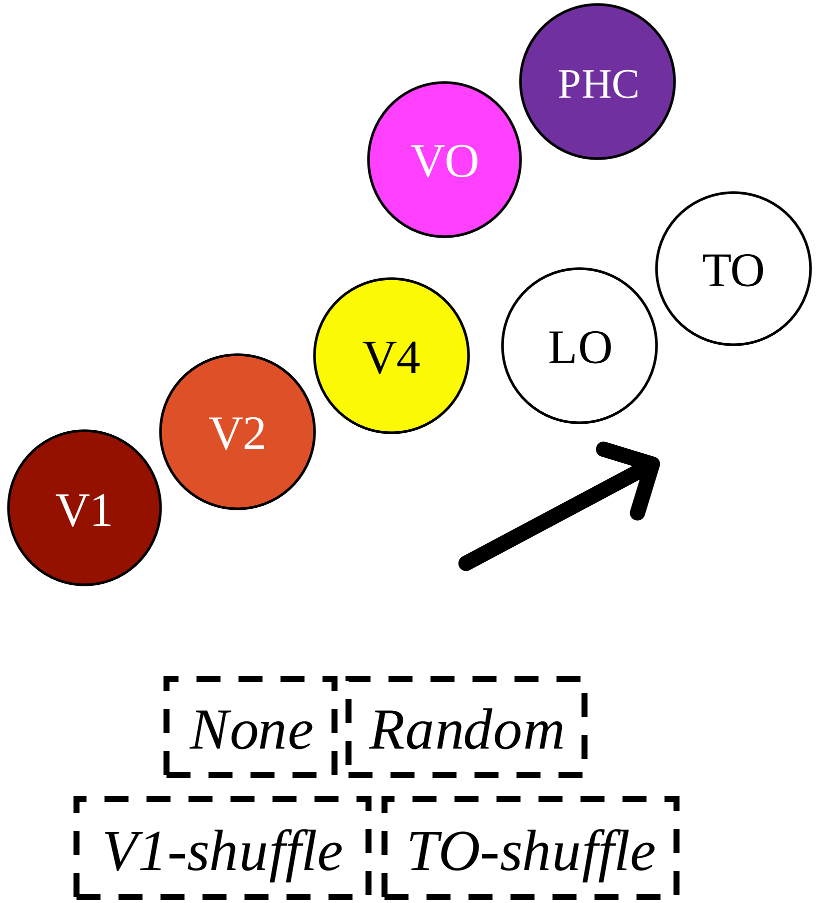
Evaluating DNN robustness - Results



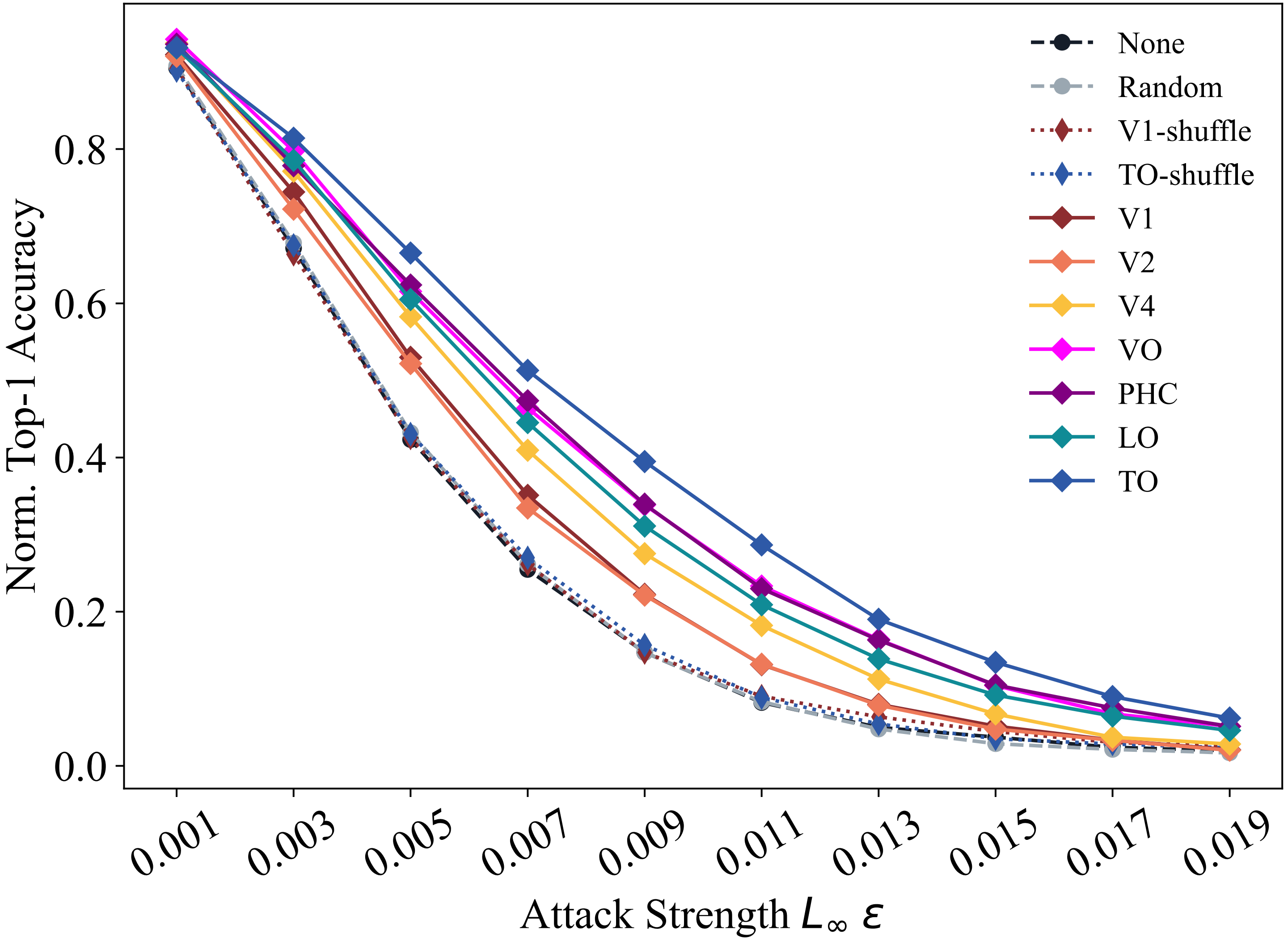
Task: Image Classification

Dataset: ImageNet (Deng et al., 2009)

Attack: l_∞-based PGD attack



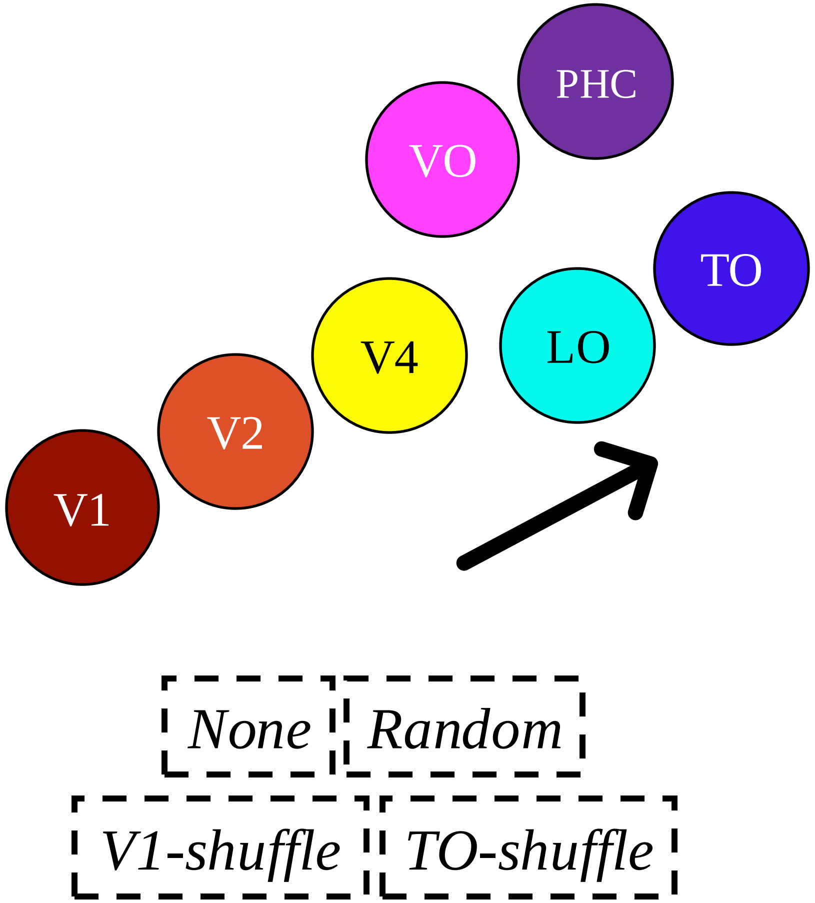
Evaluating DNN robustness - Results



Task: Image Classification

Dataset: ImageNet (Deng et al., 2009)

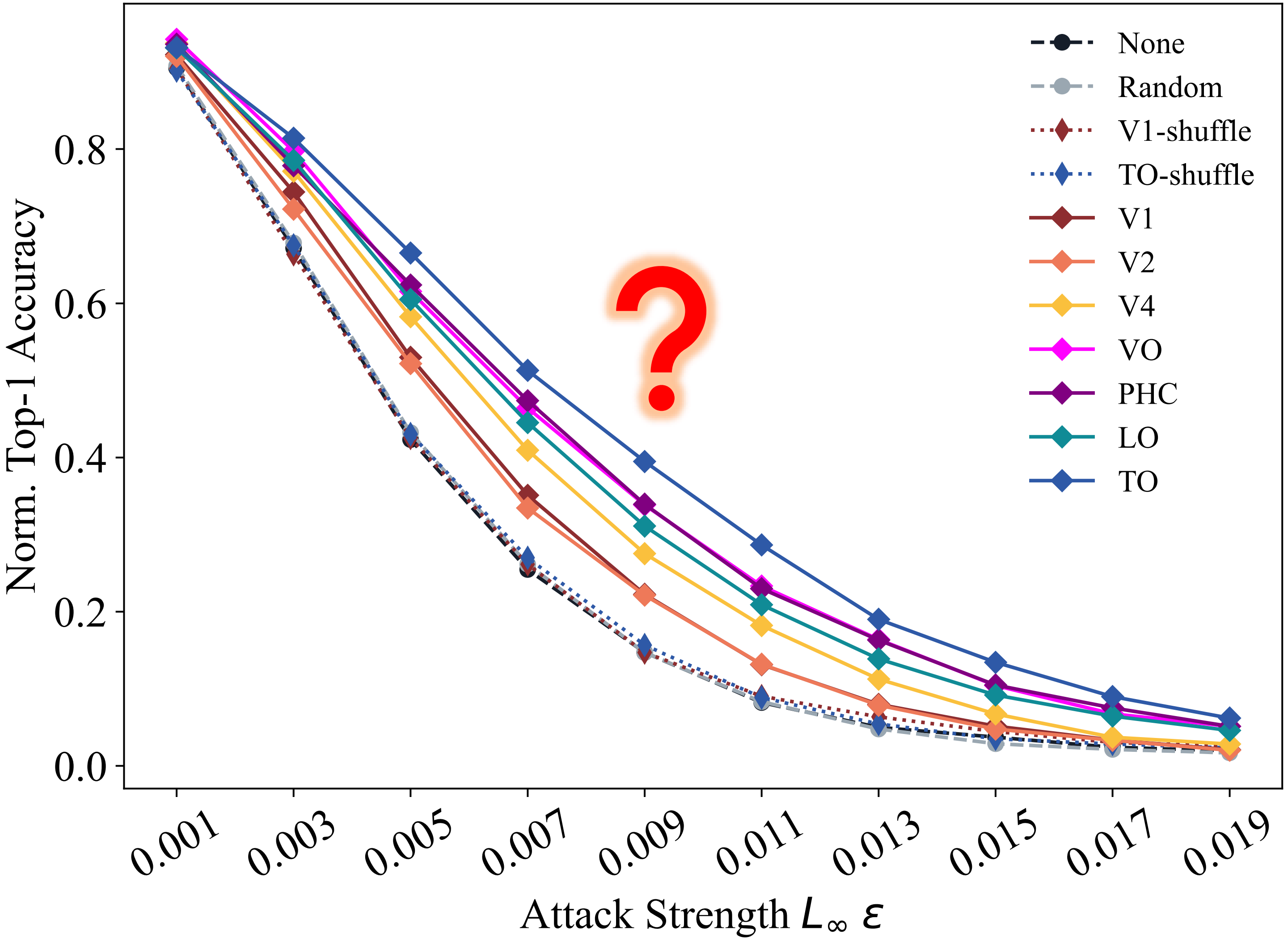
Attack: l_∞-based PGD attack



None *Random*

V1-shuffle *TO-shuffle*

Evaluating DNN robustness - Results



Task: Image Classification

Dataset: ImageNet (Deng et al., 2009)

Attack: l_∞-based PGD attack

☑ **Neural guidance improves robustness** (max: 22% accuracy increase)

☑ **There exists a hierarchy of improvement's magnitude**

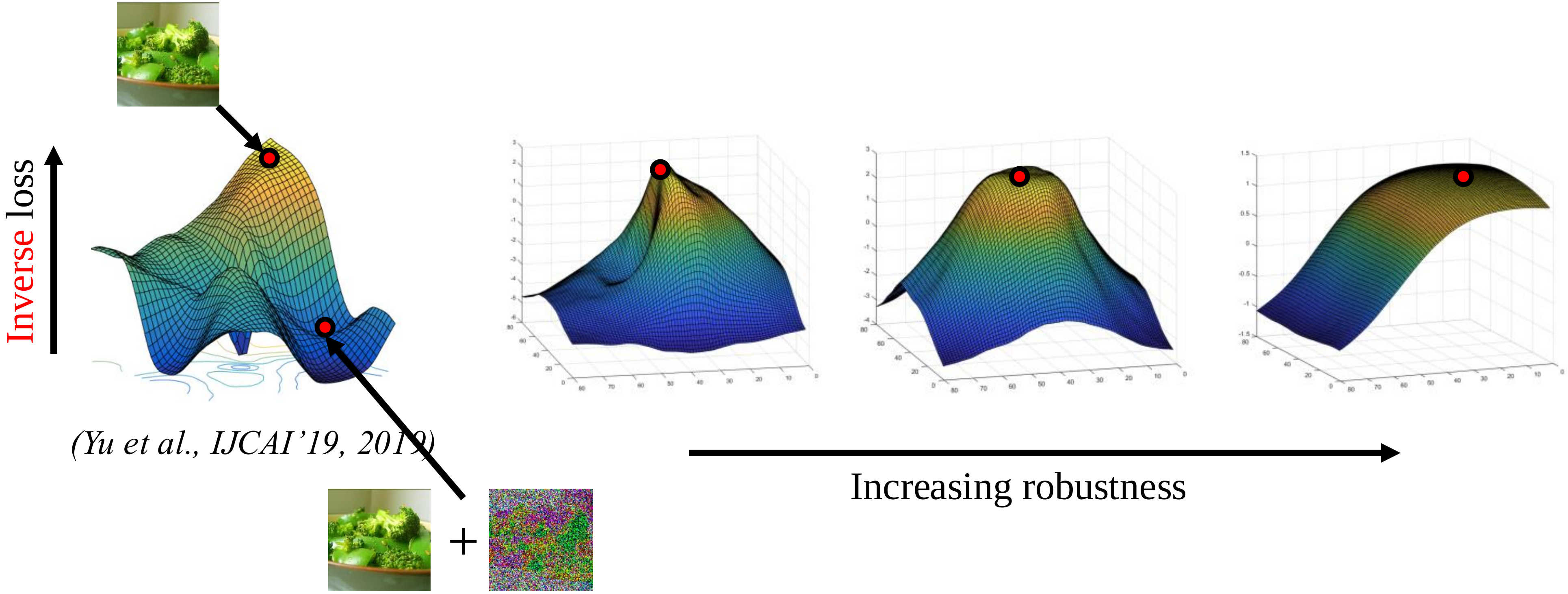
☑ ***Replicated across datasets, tasks, attacks…***

- CIFAR-100
- MSCOCO
- Image Captioning

- L_∞ FGSM
- Auto-Attack (APGD-CE, APGD-T, FAB square)
- L_2 FGM
- L_2 Deepfool

Intriguing properties of neurally-guided DNNs

- Robust DNNs have smoother output surfaces



Inverse loss

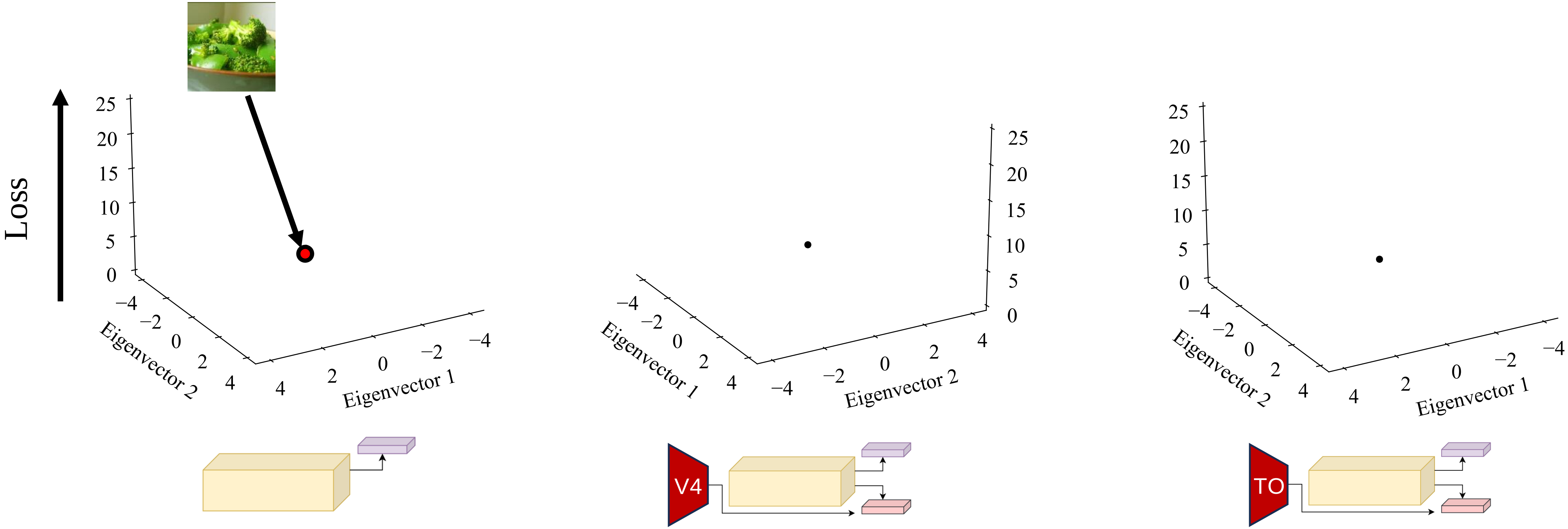
(Yu et al., IJCAI'19, 2019)

+

Increasing robustness

Intriguing properties of neurally-guided DNNs

- Robust DNNs have smoother output surfaces



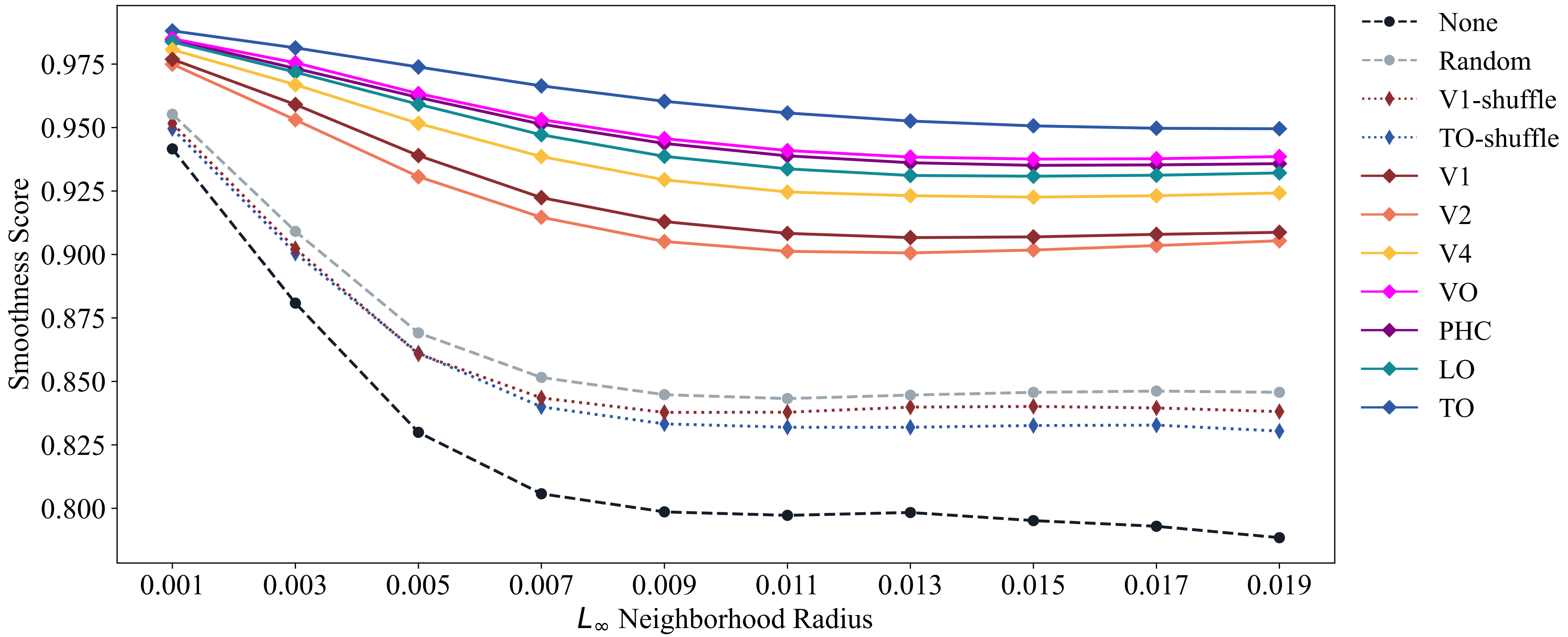
None

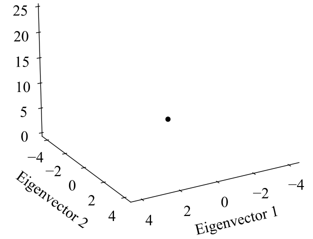
V4-guided

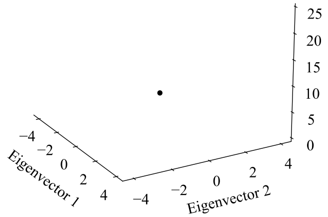
TO-guided

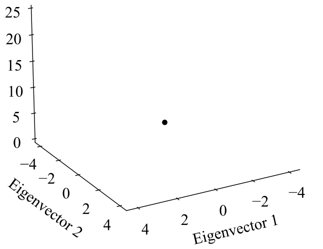
Intriguing properties of neurally-guided DNNs

- Robust DNNs have smoother output surfaces





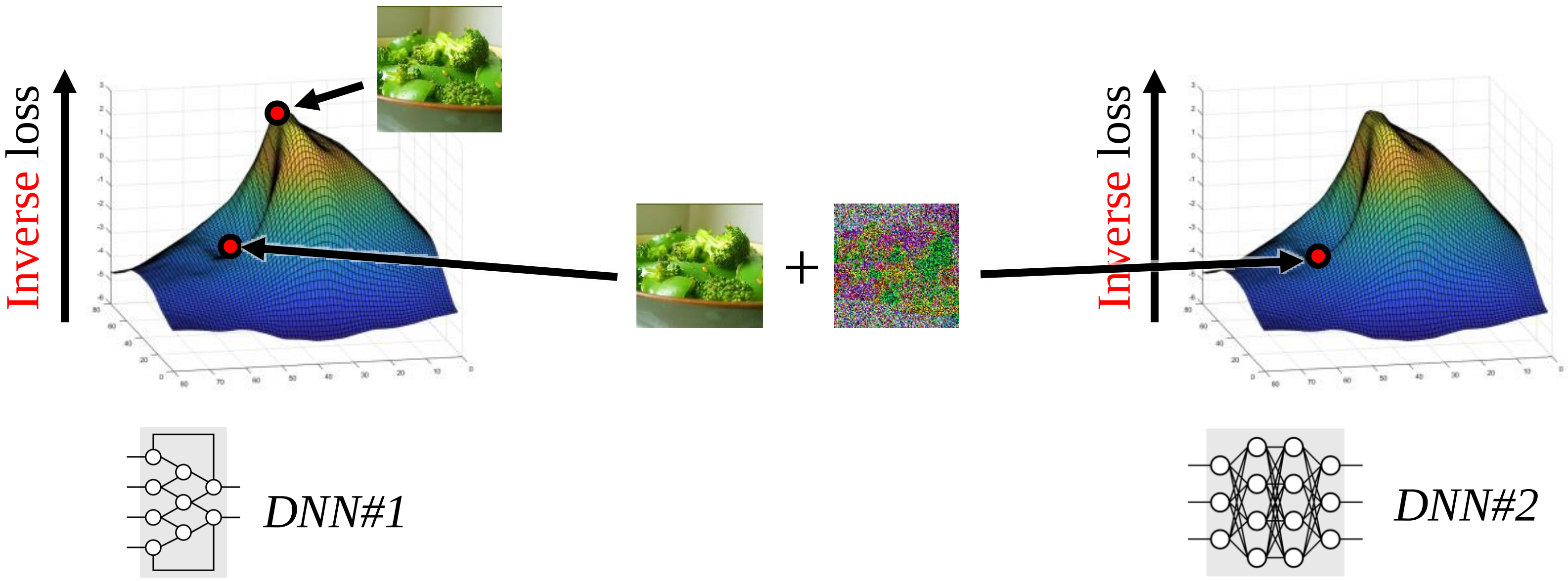




Intriguing properties of neurally-guided DNNs

- Robust DNNs have smoother output surfaces -- neurally-guided DNNs are indeed smoother!

- Conventional DNNs usually develop highly homogenous output surfaces



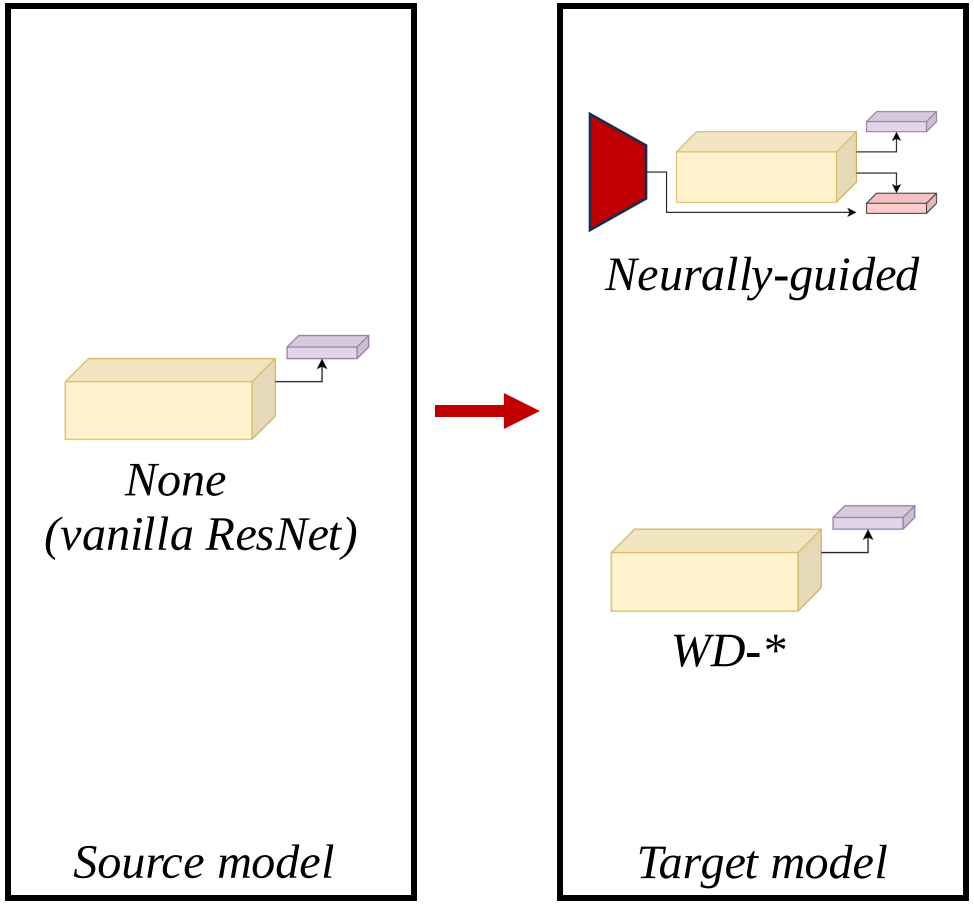
“Transfer attack”

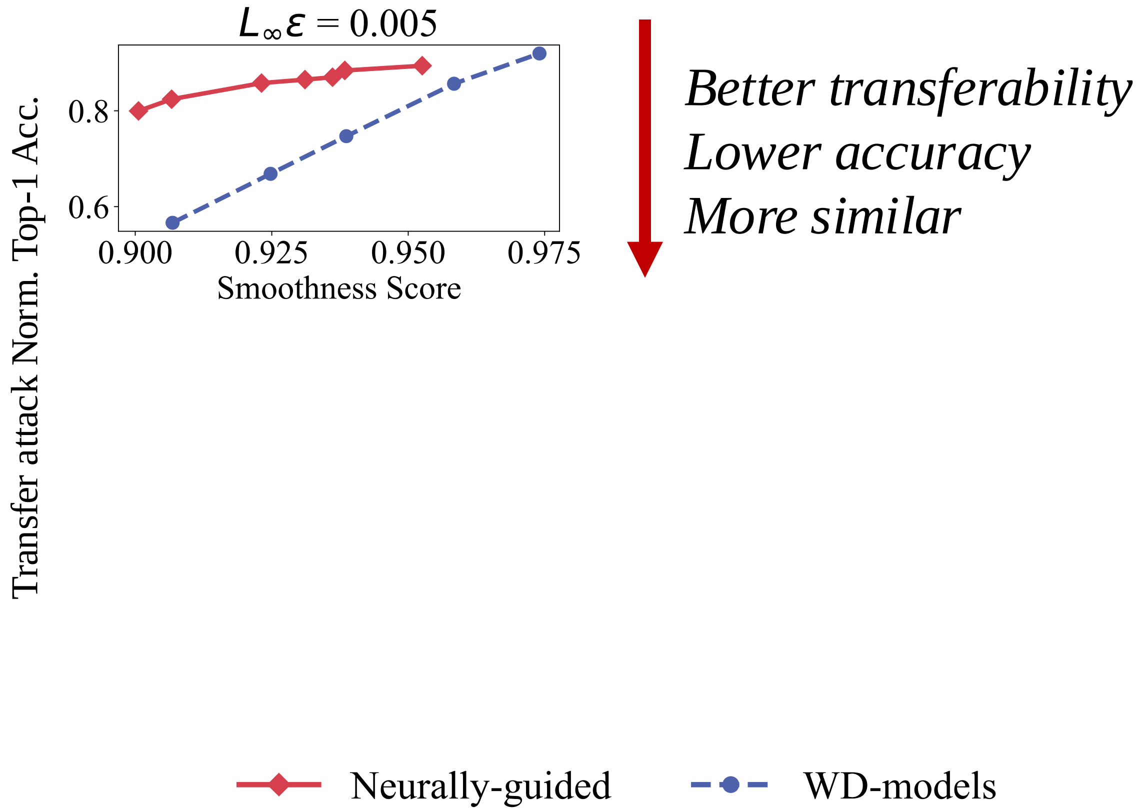
Adversarial examples are transferable across:

- Architectures *(Liu et al., 2017)*
- ML techniques *(Papernot et al., 2016)*
- Training datasets *(Lu et al., 2020)*
- Tasks *(Richards et al., 2021)*

Intriguing properties of neurally-guided DNNs

- Conventional DNNs usually develop highly homogenous output surfaces

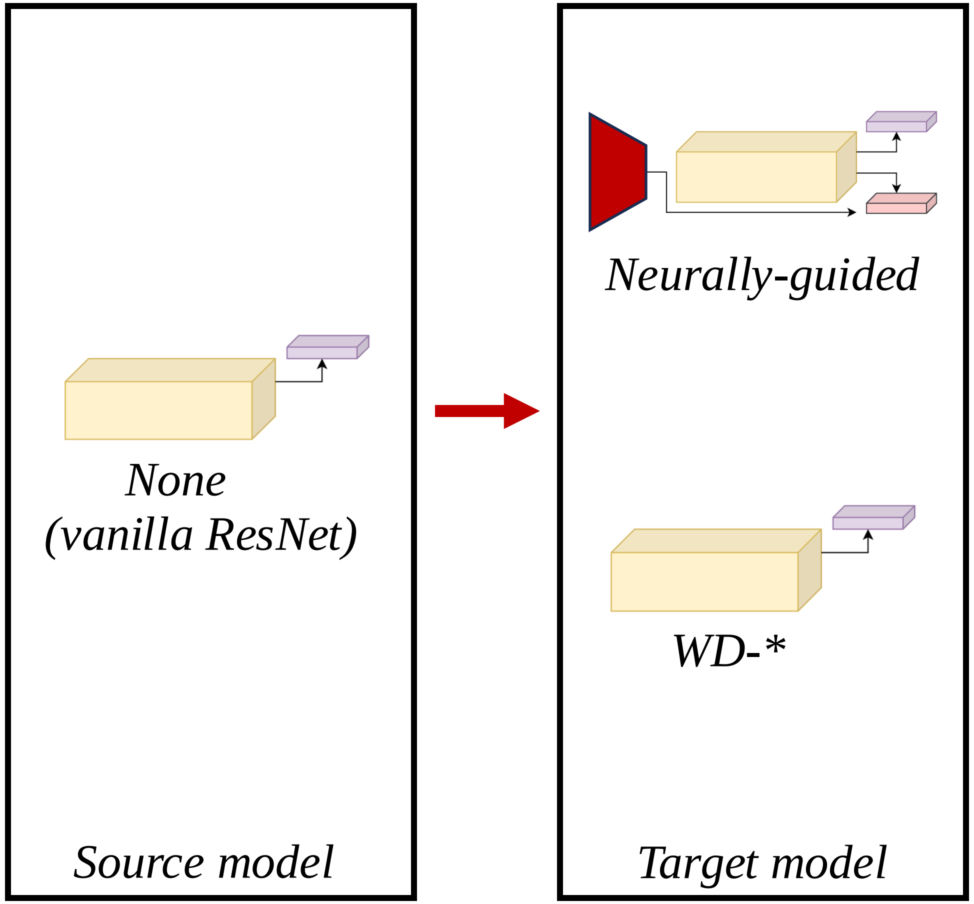




Better transferability
Lower accuracy
More similar

Intriguing properties of neurally-guided DNNs

- Conventional DNNs usually develop highly homogenous output surfaces



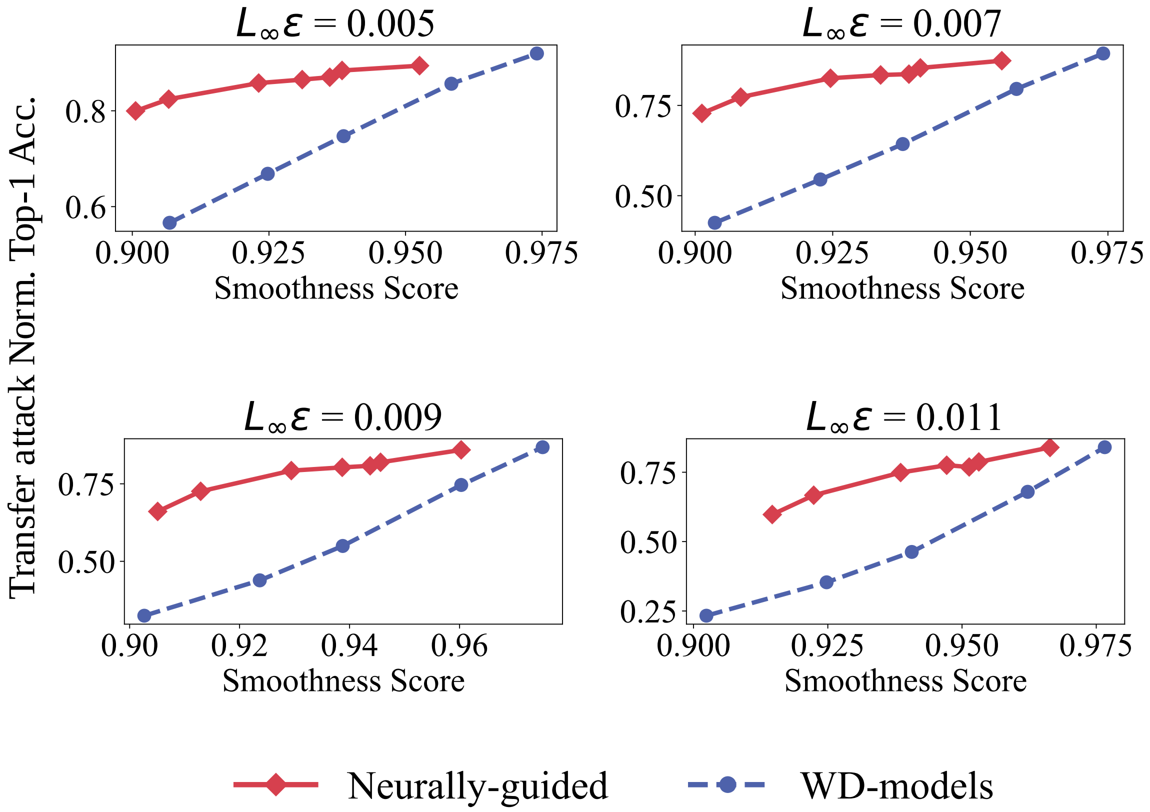
None (vanilla ResNet)

Source model

Neurally-guided

*WD-**

Target model

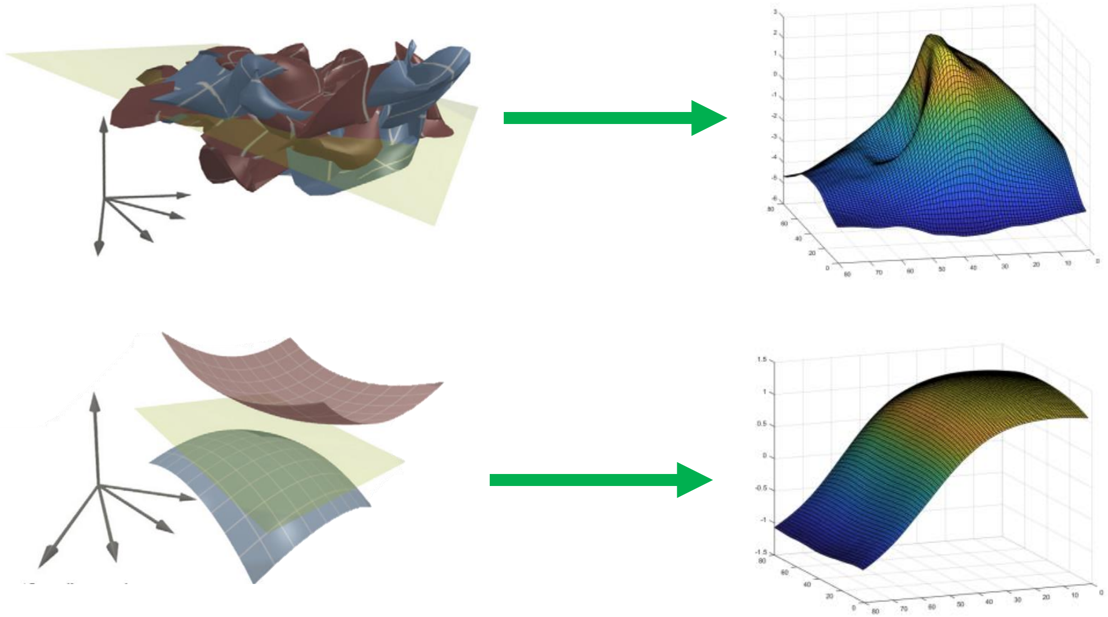


Intriguing properties of neurally-guided DNNs

- Robust DNNs have smoother output surfaces -- neurally-guided DNNs are indeed smoother!

- Conventional DNNs usually develop highly homogenous output surfaces -- but neurally-guided DNNs have distinct surfaces!

- Representational space

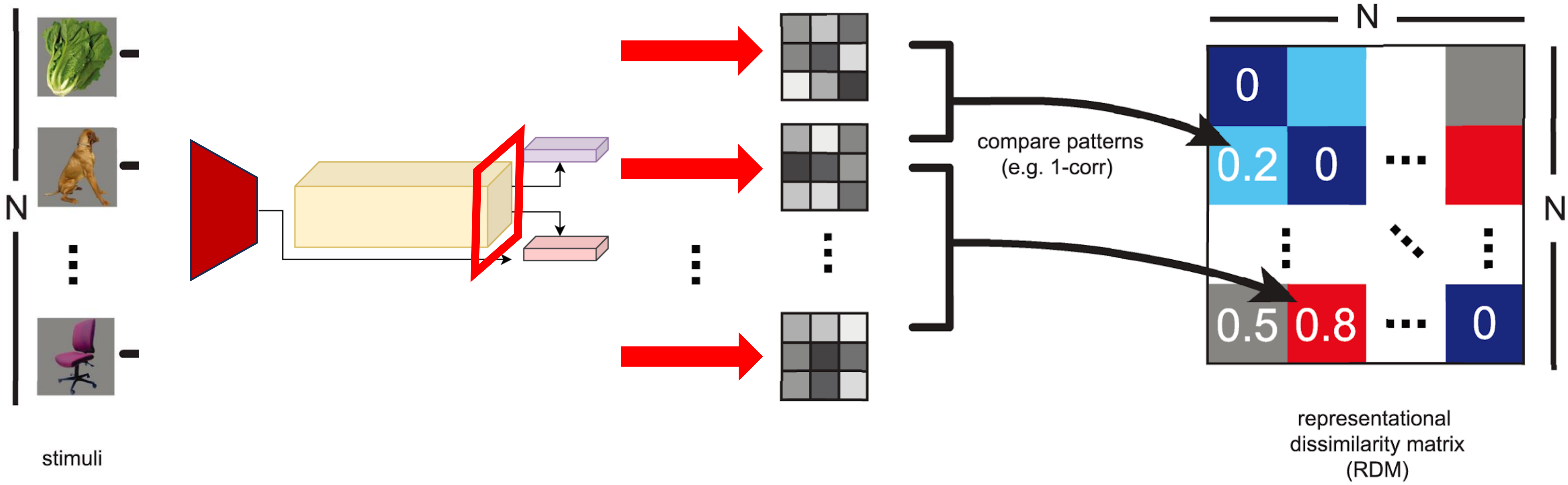


Representational space

Output surface

Neural guidance leads to distinct representational space

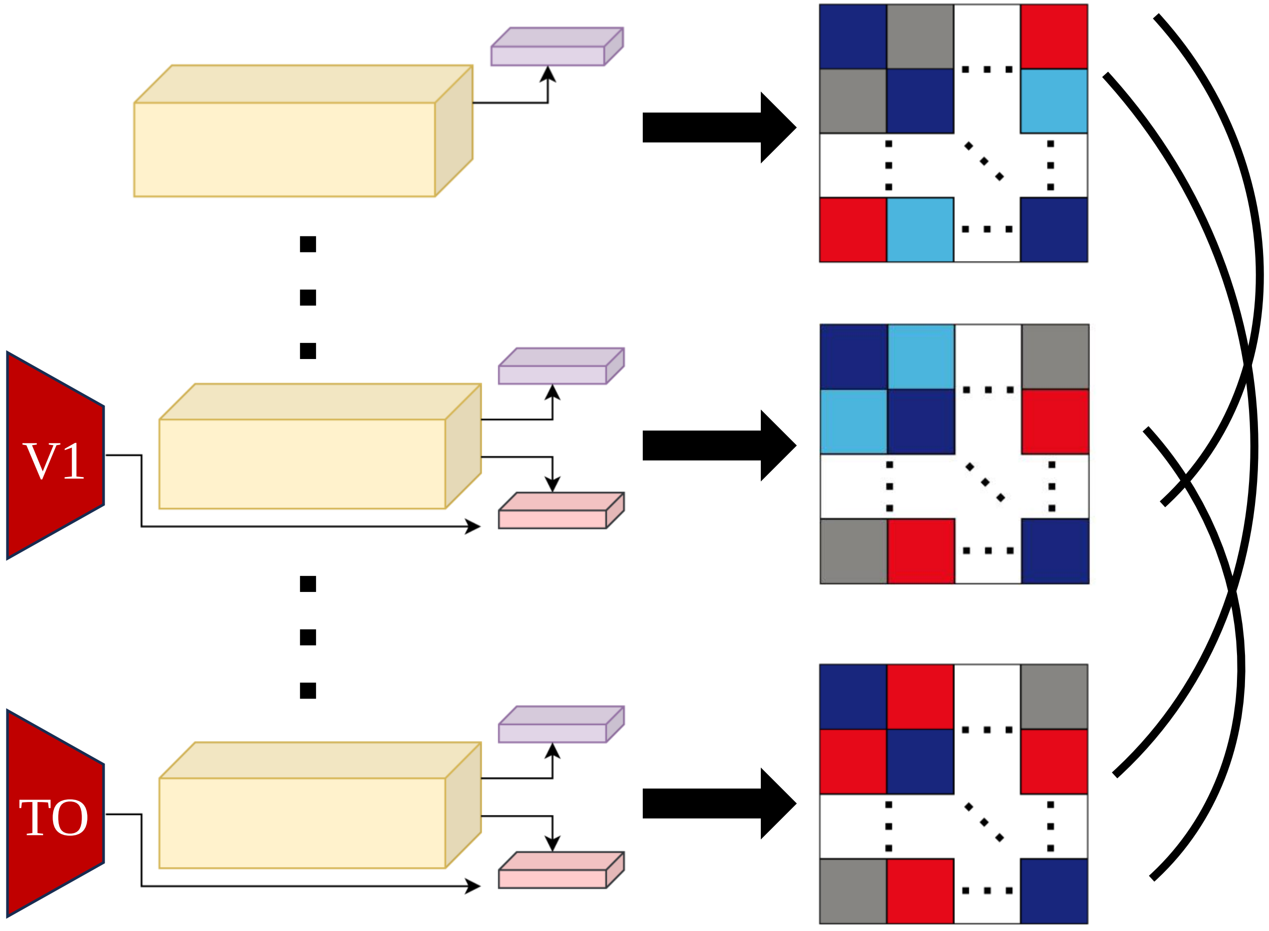
- Representational spaces of neurally guided DNNs (Representational Similarity Analysis) ***(Kriegeskorte et al., Front. Syst. Neurosci., 2008)***

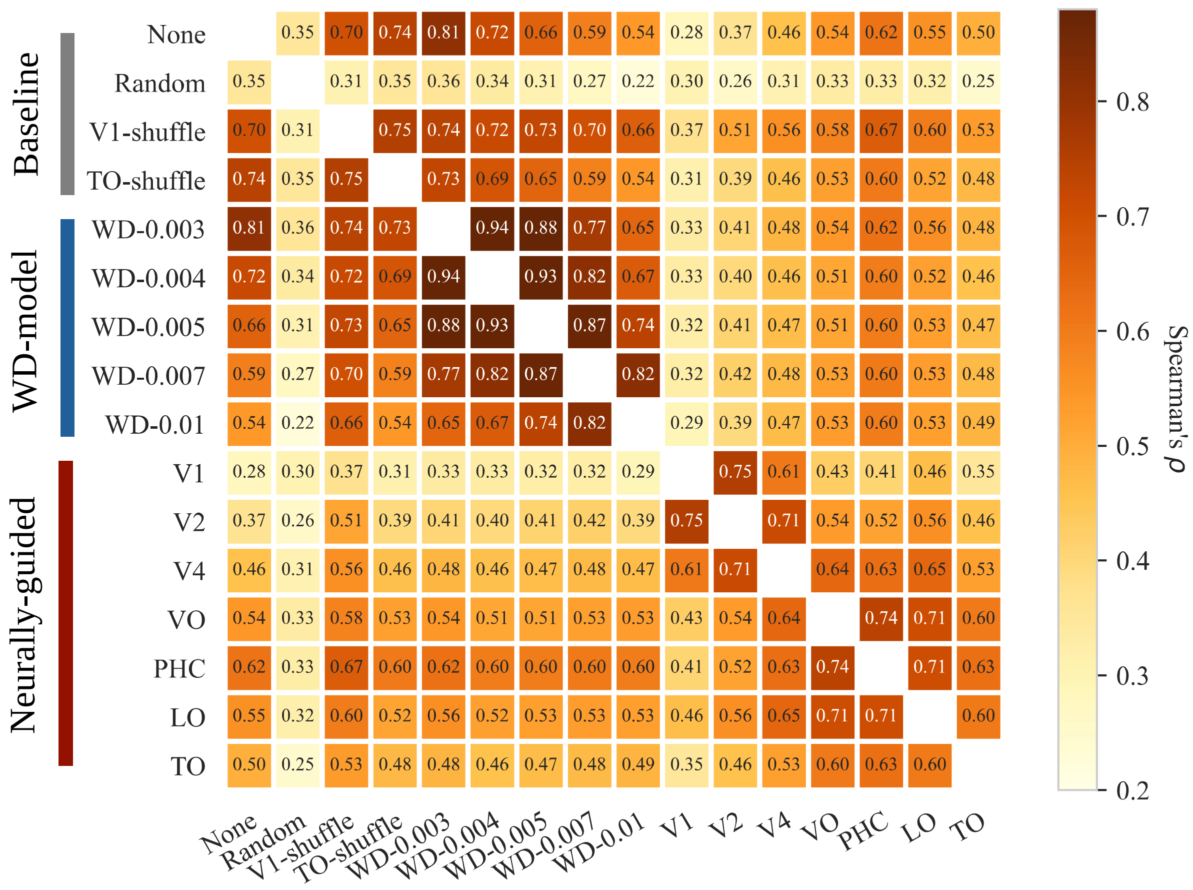


(Nilli et al., Plos. Comp. Bio., 2014)

Neural guidance leads to distinct representational space

- Representational spaces of neurally guided DNNs (Representational Similarity Analysis) ***(Kriegeskorte et al., Front. Syst. Neurosci., 2008)***

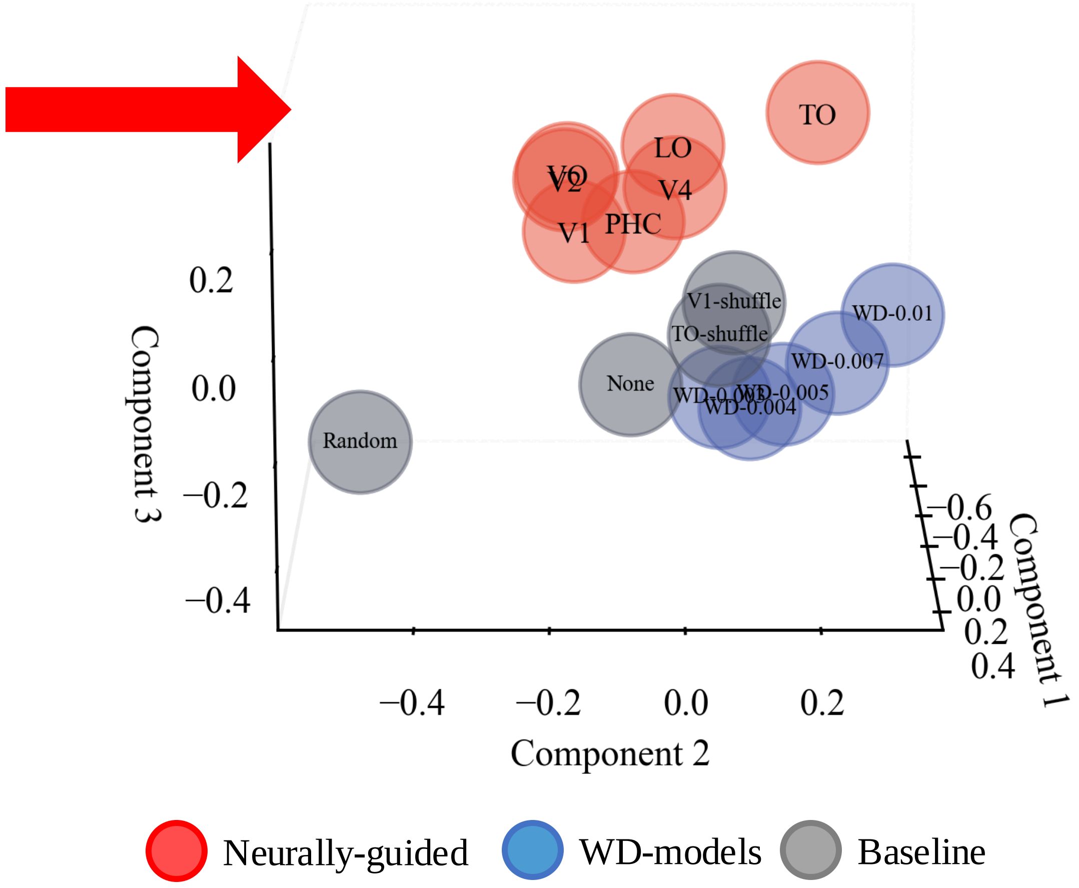


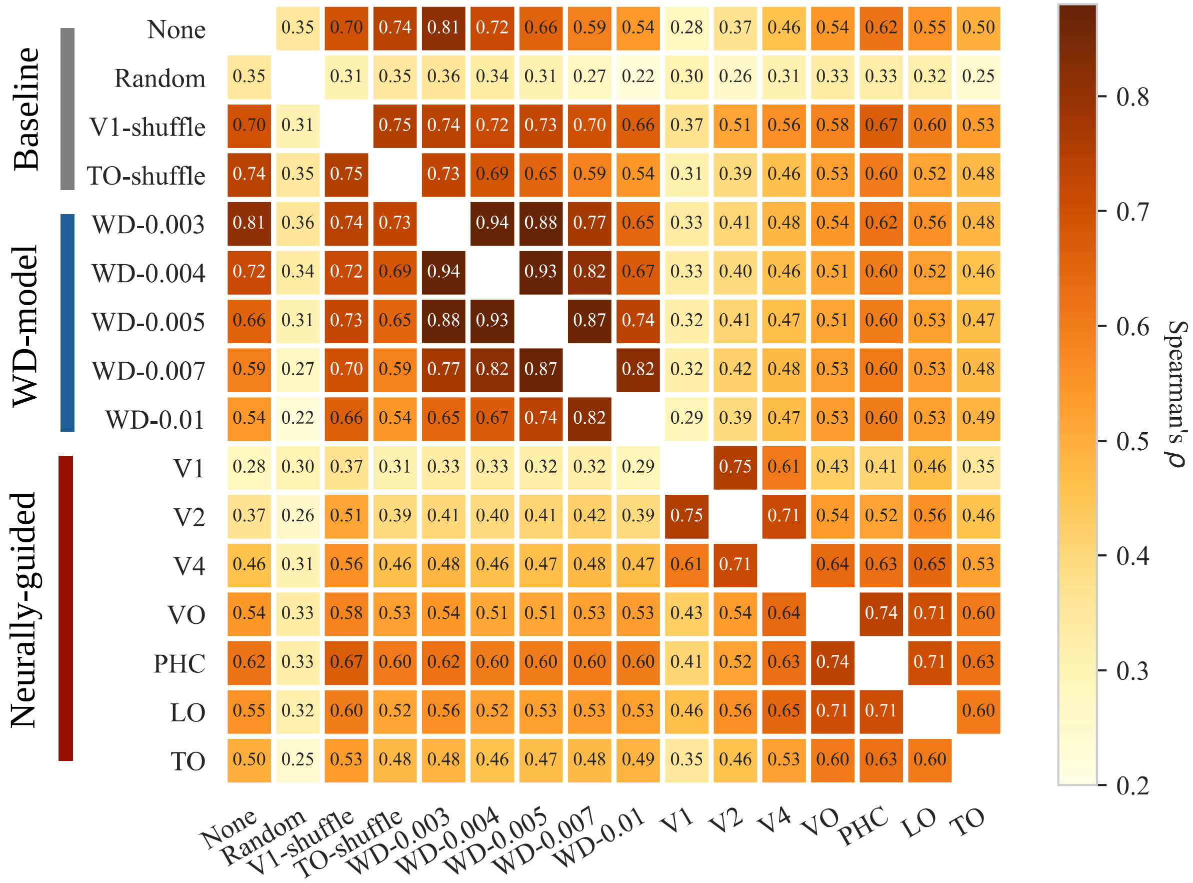


		None	Random	V1-shuffle	TO-shuffle	WD-0.003	WD-0.004	WD-0.005	WD-0.007	WD-0.01	V1	V2	V4	VO	PHC	LO	TO
Baseline	None		0.35	0.70	0.74	0.81	0.72	0.66	0.59	0.54	0.28	0.37	0.46	0.54	0.62	0.55	0.50
	Random	0.35		0.31	0.35	0.36	0.34	0.31	0.27	0.22	0.30	0.26	0.31	0.33	0.33	0.32	0.25
	V1-shuffle	0.70	0.31		0.75	0.74	0.72	0.73	0.70	0.66	0.37	0.51	0.56	0.58	0.67	0.60	0.53
	TO-shuffle	0.74	0.35	0.75		0.73	0.69	0.65	0.59	0.54	0.31	0.39	0.46	0.53	0.60	0.52	0.48
WD-model	WD-0.003	0.81	0.36	0.74	0.73		0.94	0.88	0.77	0.65	0.33	0.41	0.48	0.54	0.62	0.56	0.48
	WD-0.004	0.72	0.34	0.72	0.69	0.94		0.93	0.82	0.67	0.33	0.40	0.46	0.51	0.60	0.52	0.46
	WD-0.005	0.66	0.31	0.73	0.65	0.88	0.93		0.87	0.74	0.32	0.41	0.47	0.51	0.60	0.53	0.47
	WD-0.007	0.59	0.27	0.70	0.59	0.77	0.82	0.87		0.82	0.32	0.42	0.48	0.53	0.60	0.53	0.48
	WD-0.01	0.54	0.22	0.66	0.54	0.65	0.67	0.74	0.82		0.29	0.39	0.47	0.53	0.60	0.53	0.49
Neurally-guided	V1	0.28	0.30	0.37	0.31	0.33	0.33	0.32	0.32	0.29		0.75	0.61	0.43	0.41	0.46	0.35
	V2	0.37	0.26	0.51	0.39	0.41	0.40	0.41	0.42	0.39	0.75		0.71	0.54	0.52	0.56	0.46
	V4	0.46	0.31	0.56	0.46	0.48	0.46	0.47	0.48	0.47	0.61	0.71		0.64	0.63	0.65	0.53
	VO	0.54	0.33	0.58	0.53	0.54	0.51	0.51	0.53	0.53	0.43	0.54	0.64		0.74	0.71	0.60
	PHC	0.62	0.33	0.67	0.60	0.62	0.60	0.60	0.60	0.60	0.41	0.52	0.63	0.74		0.71	0.63
	LO	0.55	0.32	0.60	0.52	0.56	0.52	0.53	0.53	0.53	0.46	0.56	0.65	0.71	0.71		0.60
	TO	0.50	0.25	0.53	0.48	0.48	0.46	0.47	0.48	0.49	0.35	0.46	0.53	0.60	0.63	0.60	

Intriguing properties of neurally-guided DNNs

- Representational spaces of neurally guided DNNs are distinct from conventionally trained ones





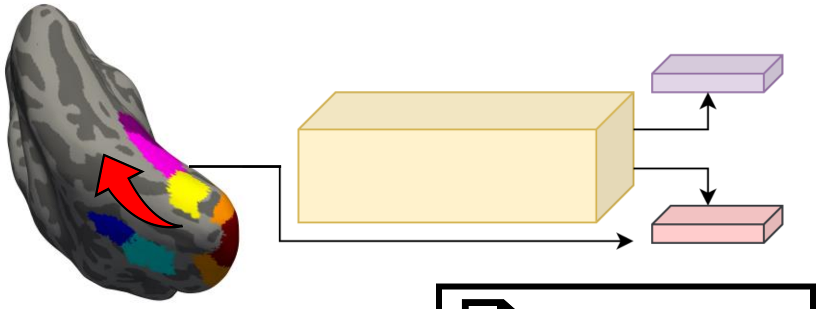
	None	Random	V1-shuffle	TO-shuffle	WD-0.003	WD-0.004	WD-0.005	WD-0.007	WD-0.01	V1	V2	V4	VO	PHC	LO	TO
None		0.35	0.70	0.74	0.81	0.72	0.66	0.59	0.54	0.28	0.37	0.46	0.54	0.62	0.55	0.50
Random	0.35		0.31	0.35	0.36	0.34	0.31	0.27	0.22	0.30	0.26	0.31	0.33	0.33	0.32	0.25
V1-shuffle	0.70	0.31		0.75	0.74	0.72	0.73	0.70	0.66	0.37	0.51	0.56	0.58	0.67	0.60	0.53
TO-shuffle	0.74	0.35	0.75		0.73	0.69	0.65	0.59	0.54	0.31	0.39	0.46	0.53	0.60	0.52	0.48
WD-0.003	0.81	0.36	0.74	0.73		0.94	0.88	0.77	0.65	0.33	0.41	0.48	0.54	0.62	0.56	0.48
WD-0.004	0.72	0.34	0.72	0.69	0.94		0.93	0.82	0.67	0.33	0.40	0.46	0.51	0.60	0.52	0.46
WD-0.005	0.66	0.31	0.73	0.65	0.88	0.93		0.87	0.74	0.32	0.41	0.47	0.51	0.60	0.53	0.47
WD-0.007	0.59	0.27	0.70	0.59	0.77	0.82	0.87		0.82	0.32	0.42	0.48	0.53	0.60	0.53	0.48
WD-0.01	0.54	0.22	0.66	0.54	0.65	0.67	0.74	0.82		0.29	0.39	0.47	0.53	0.60	0.53	0.49
V1	0.28	0.30	0.37	0.31	0.33	0.33	0.32	0.32	0.29		0.75	0.61	0.43	0.41	0.46	0.35
V2	0.37	0.26	0.51	0.39	0.41	0.40	0.41	0.42	0.39	0.75		0.71	0.54	0.52	0.56	0.46
V4	0.46	0.31	0.56	0.46	0.48	0.46	0.47	0.48	0.47	0.61	0.71		0.64	0.63	0.65	0.53
VO	0.54	0.33	0.58	0.53	0.54	0.51	0.51	0.53	0.53	0.43	0.54	0.64		0.74	0.71	0.60
PHC	0.62	0.33	0.67	0.60	0.62	0.60	0.60	0.60	0.60	0.41	0.52	0.63	0.74		0.71	0.63
LO	0.55	0.32	0.60	0.52	0.56	0.52	0.53	0.53	0.53	0.46	0.56	0.65	0.71	0.71		0.60
TO	0.50	0.25	0.53	0.48	0.48	0.46	0.47	0.48	0.49	0.35	0.46	0.53	0.60	0.63	0.60	

Intriguing properties of neurally-guided DNNs

- Robust DNNs have smoother output surfaces -- neurally-guided DNNs are indeed smoother!
- Conventional DNNs usually develop highly homogenous output surfaces -- but neurally-guided DNNs have distinct surfaces!
- Representational space -- neurally-guided DNNs developed distinct and better representational geometry!

Conclusion & Discussion

- We found:
  - Hierarchical improvements in DNN adversarial robustness with neural guidance
  - Neurally-guided DNNs developed distinct and hierarchically smoother output surfaces directly contributing to robustness
  - Neurally-guided DNNs developed distinct representational spaces
  - Neurally-guided DNNs are progressively more shape biased.



Paper

- Implications:
  - Robustness emerges from the evolving representational space along the ventral visual stream
  - Potential for understanding human representational space and advancing DNN architectural developments

Acknowledgement

Thank you! Questions?

Zhenan Shao

zhenans2@illinois.edu

Linjian Ma

Bo Li

Diane M. Beck

Paper

ACCESS

NSF

This work used NCSA Delta GPU through allocation SOC230011 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.