Manipulating and Measuring Variations in DNN Representations

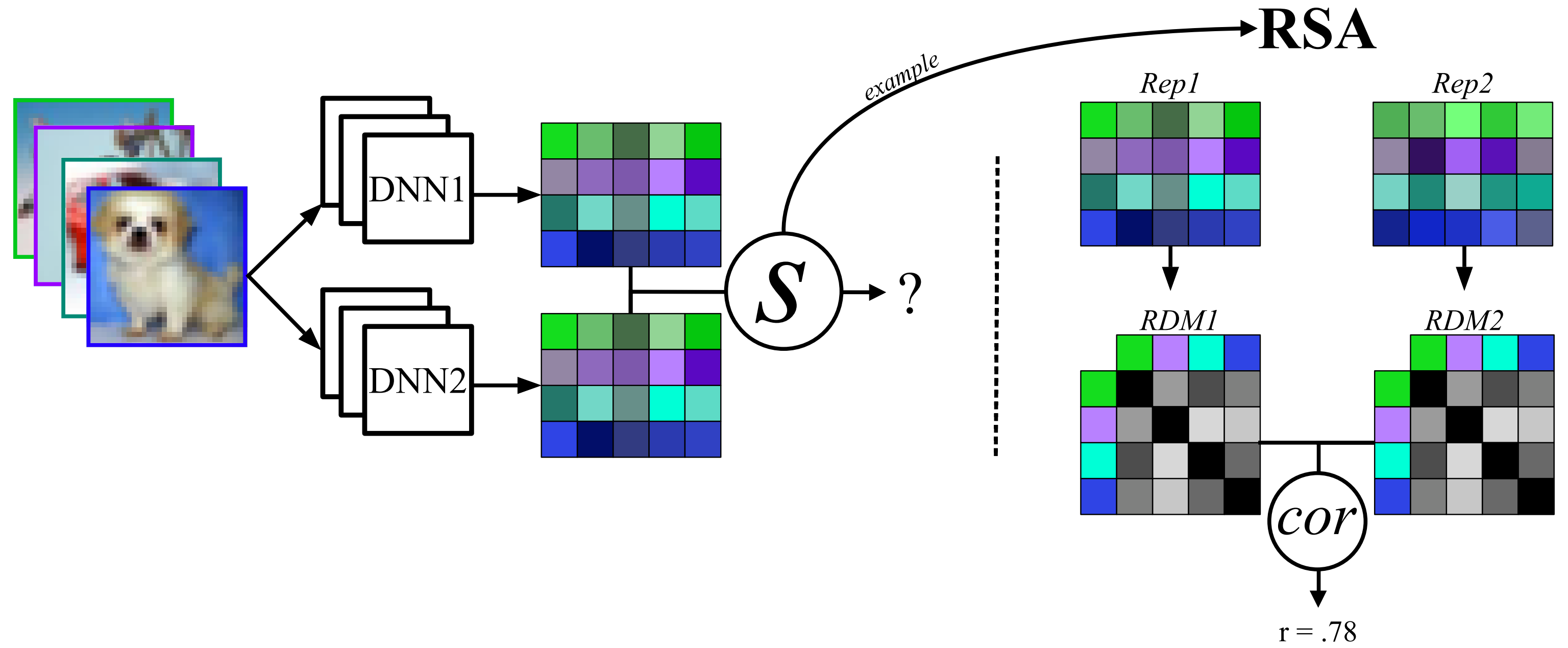
Jason Chow and Thomas Palmeri
Department of Psychology

Deep neural networks (DNNs) are useful in modeling the visual system, variations in object representations across models could be used to model individual differences in visual cognition.

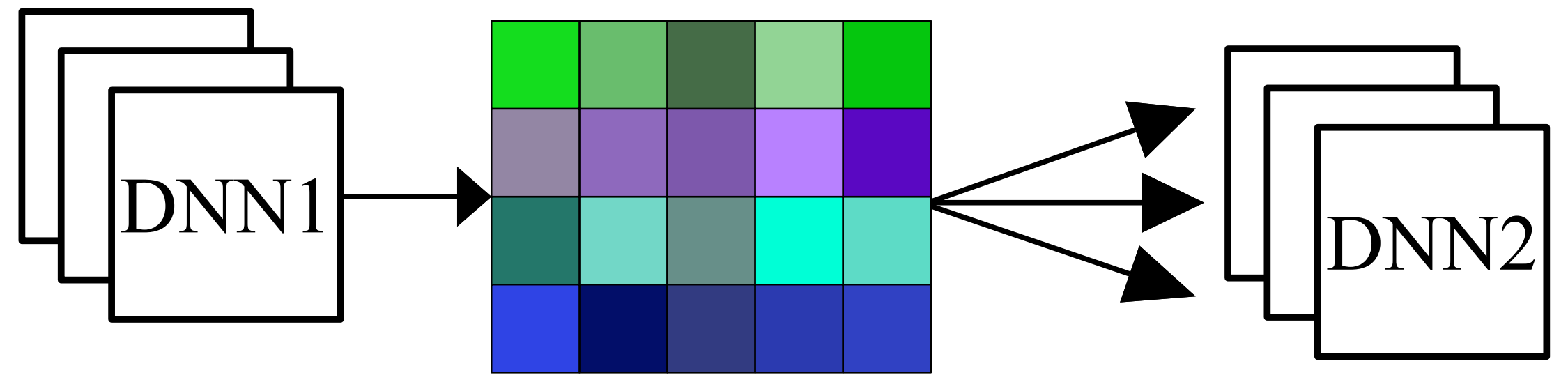
- •How should we measure representational (dis)similarity?
- •How much variation results from a range of manipulations?

Robustness of Metrics

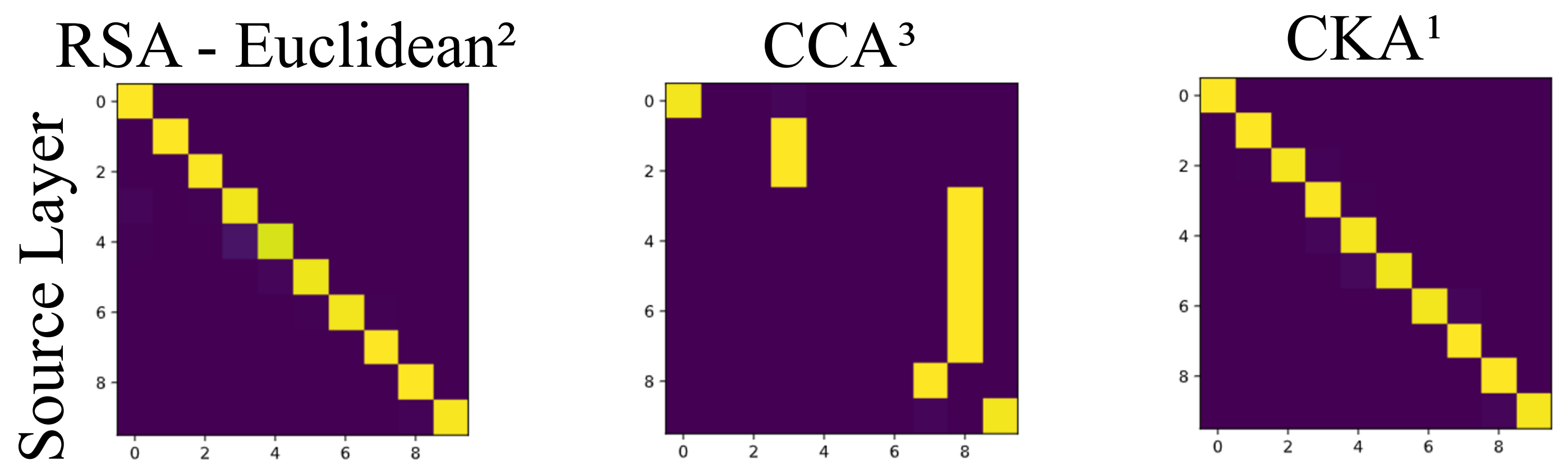
Simulations to test which measures should be used. Alternatives measures seem interchangeable but are mathematically distinct.



Correspondence Test



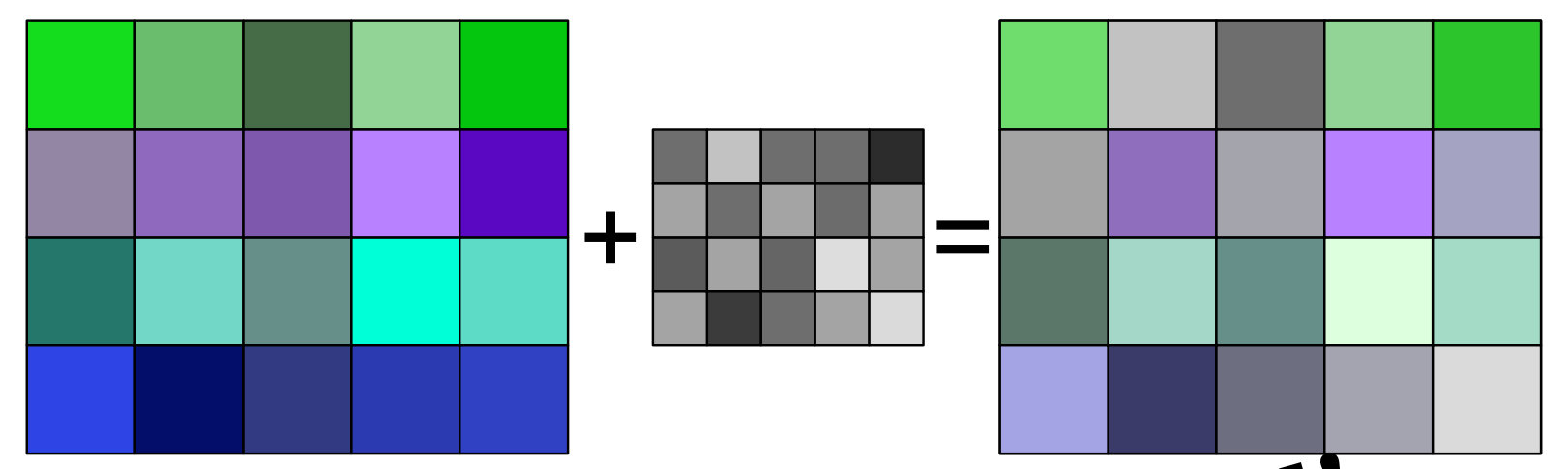
Given representations from a layer, it should be the most similar to representations from the same layer in a copy of the DNN[1].


RSA - Euclidean[2]
CCA[3]
CKA[1]

Most Similar Layer

RSA and CKA work well, CCA confuses "pooling" layers.

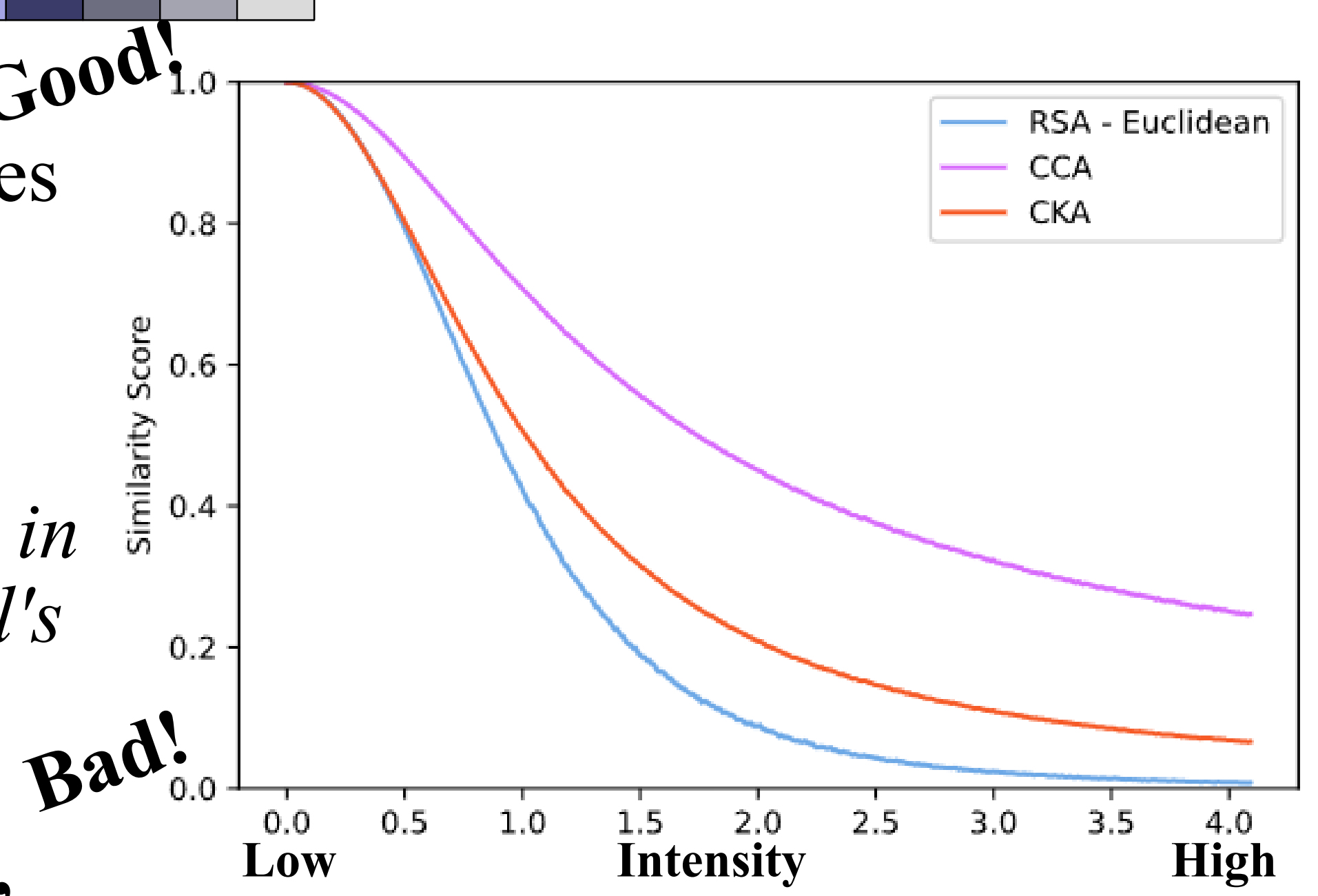
Add Noise



Adding noise should lower similarity but there is still signal. **Goal: 0 < s < 1**

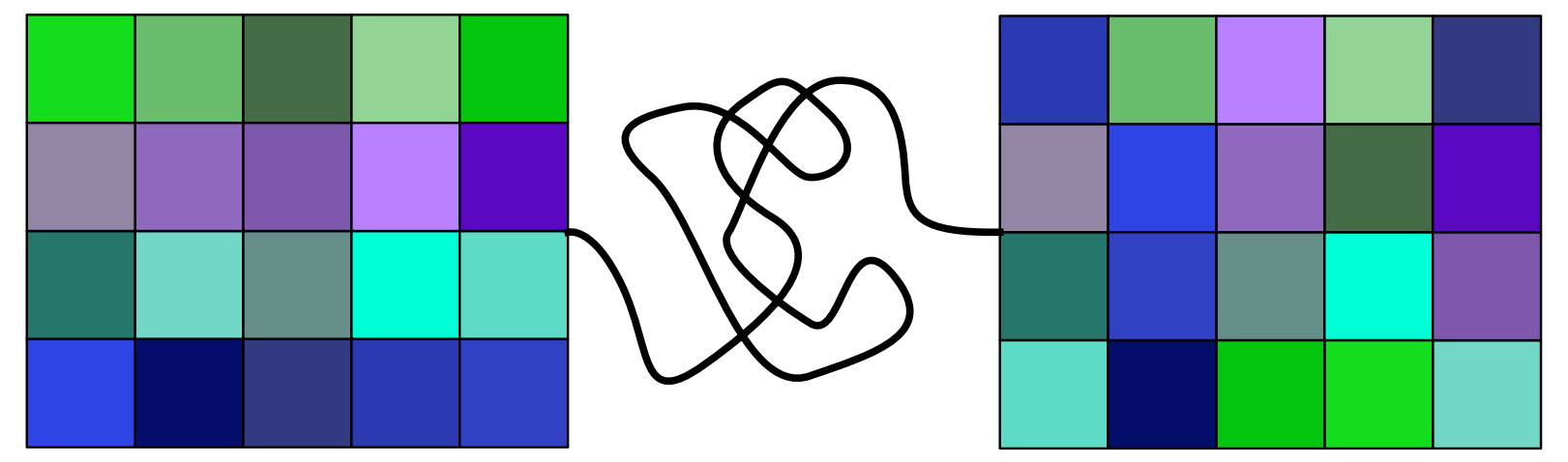
Representations were compared to noisy copies using each measure.

Normally-distributed noise N(0, SD) *with SD in multiples of the original's SD were added.*



For low levels of noise, we expected high similarity scores. For high levels of noise, we expected low similarity scores.

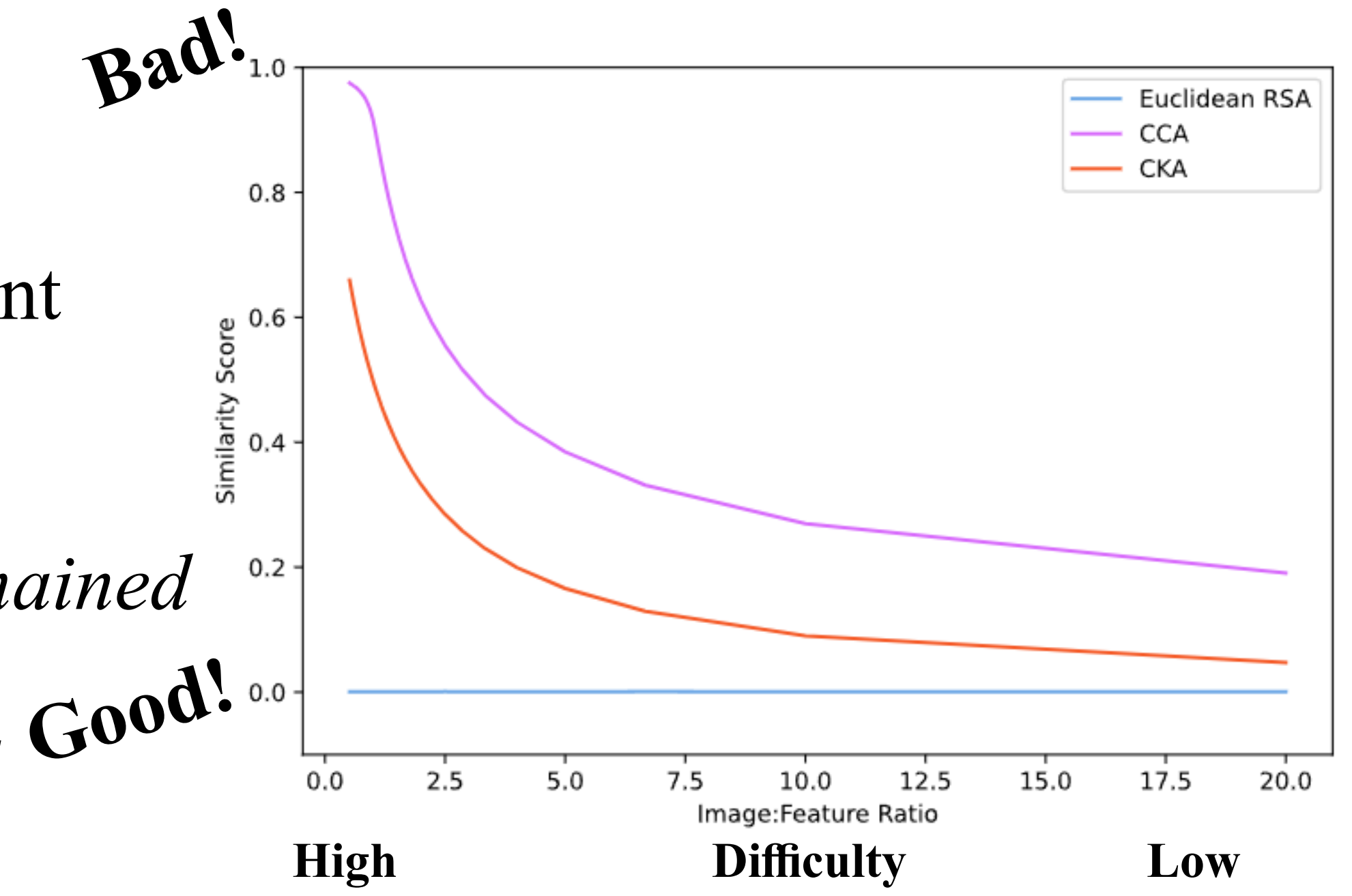
Randomize Activation



Randomizing activation should result in perfect dissimilarity. **Goal: s = 0**

Randomized representations were compared using each measure across different image:feature ratios.

Number of images remained the same (1000) while the number of features varied (50-950).

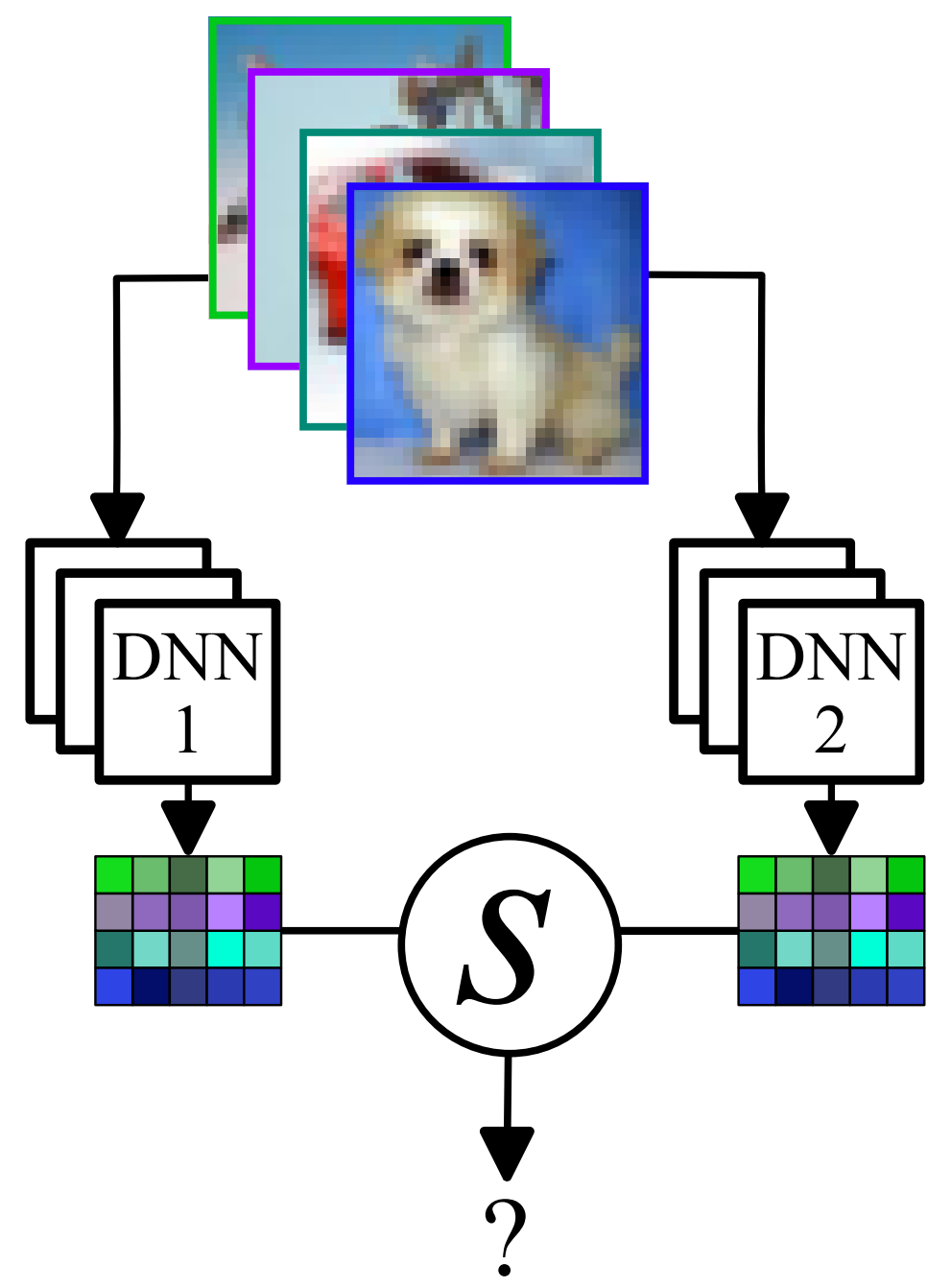


We always expect perfect dissimilarity.

RSA using Euclidean distances to form RDMs performs well. CKA is good but is sensitive to image:feature ratio.

Manipulating Representational Variations

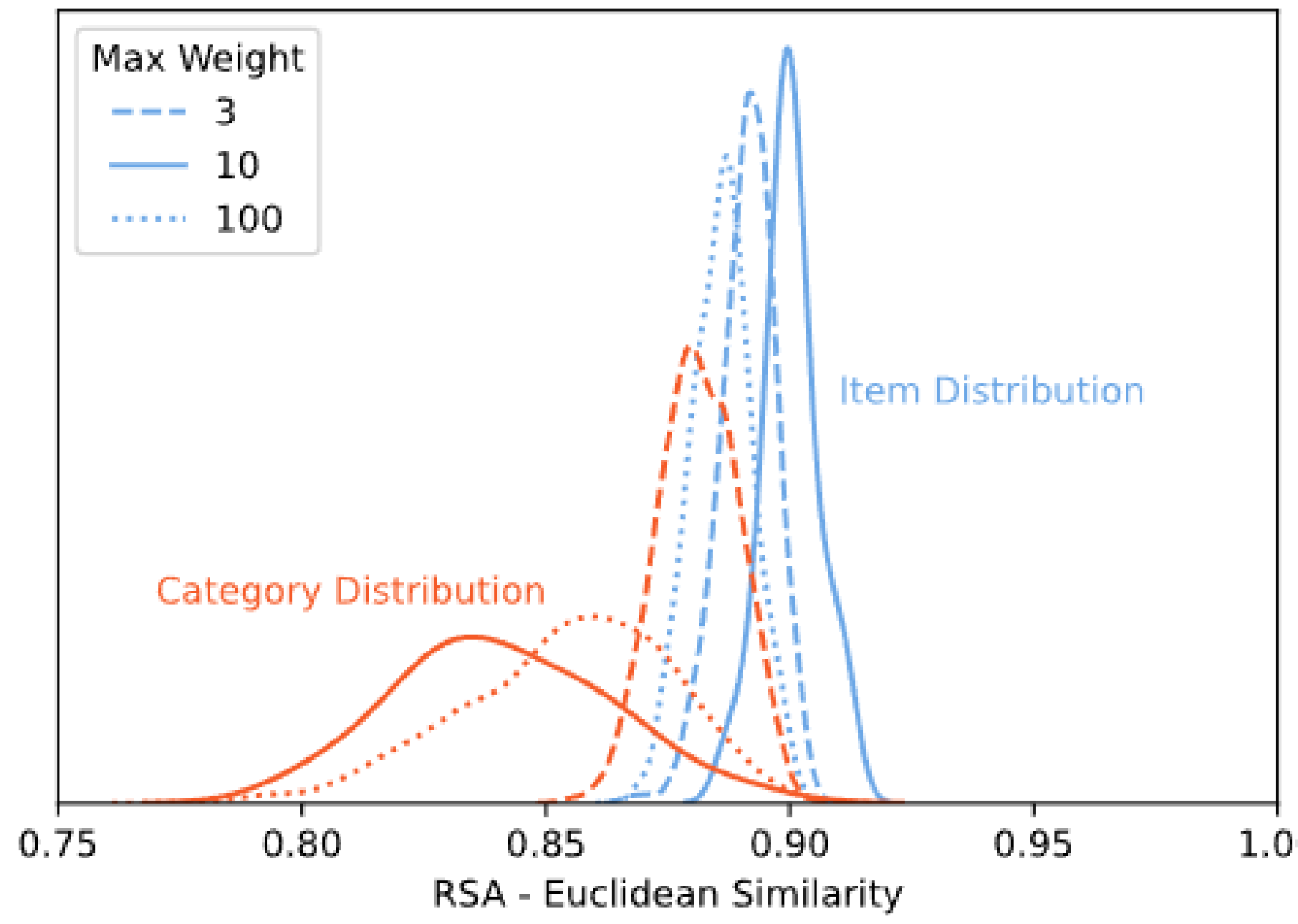
Training Manipulations



Variation in representations as a result of training manipulations was measured.

DNNs differed in three ways:

1. Initial weights/data randomization[4]
2. Distribution of images in the dataset
3. Distribution of categories in the dataset



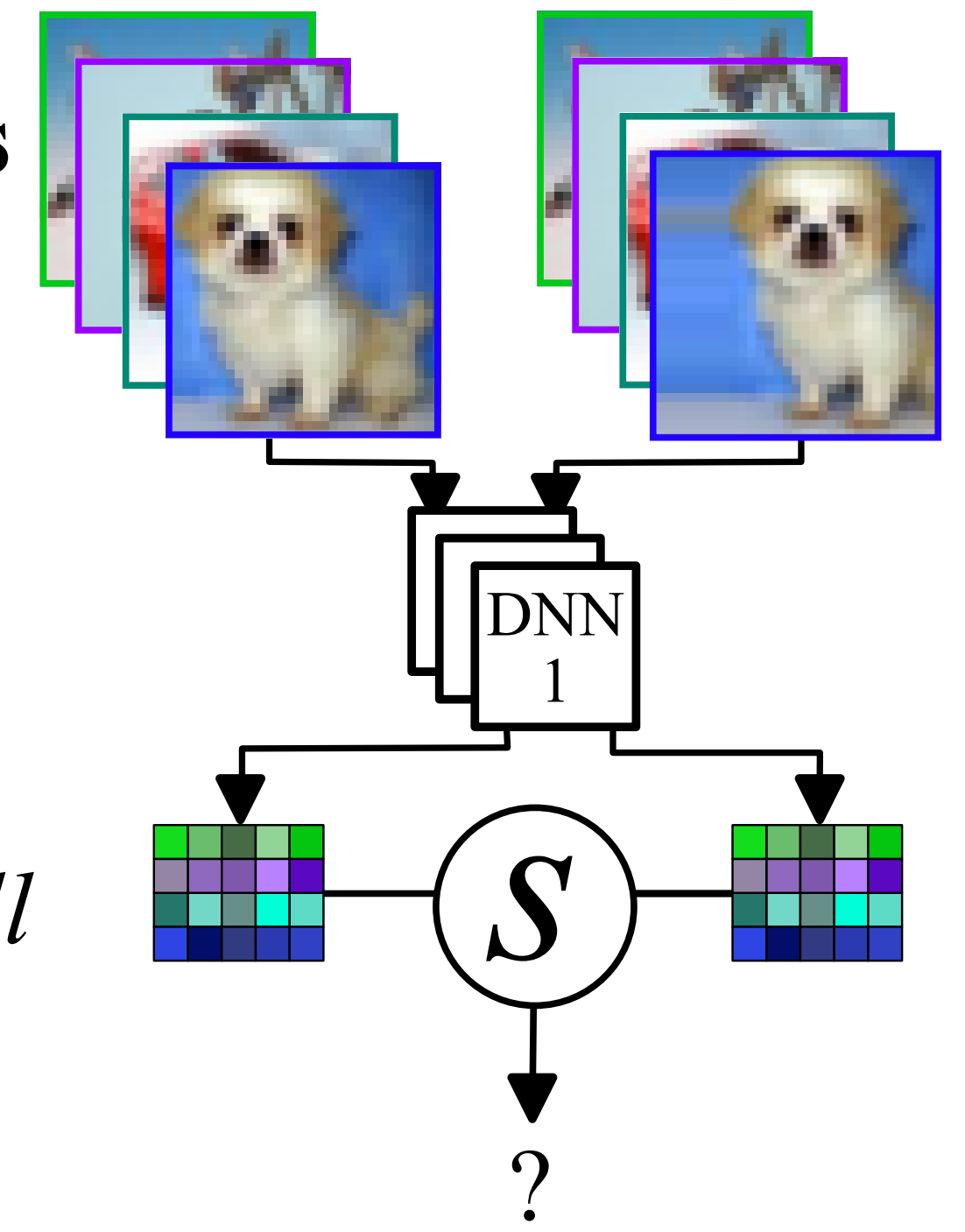
Dataset manipulations could vary in intensity.

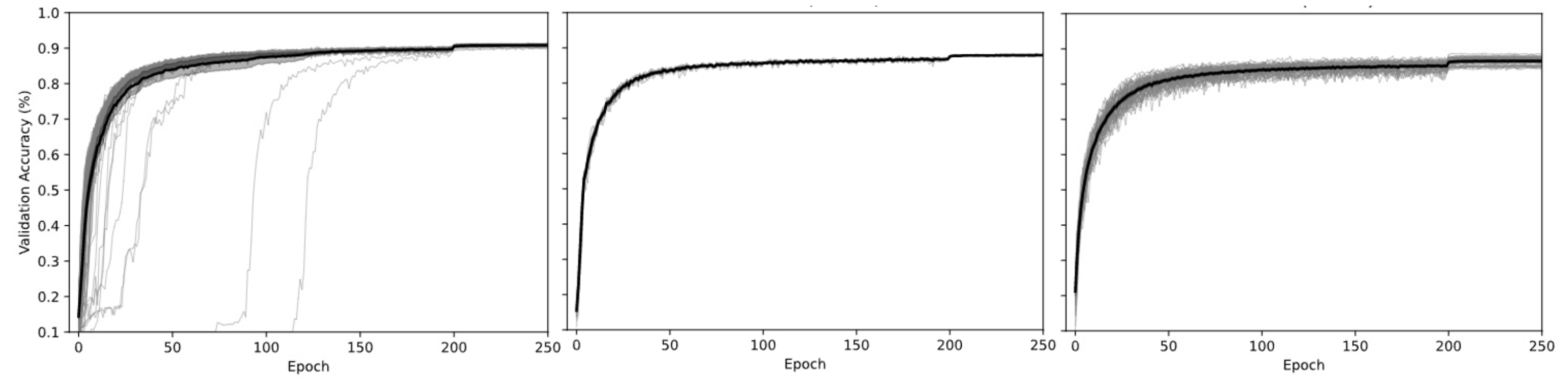
Even modest deviations from the original dataset produces a range of variations in representations.

Image-level manipulations to provide a baseline of representational variability. **Variations of a magnitude less than this baseline would not be of sufficient interest to inform individual differences.**

Image augmentation techniques (image translation, reflection, color shift) were used to train the DNNs. Representational variations caused by them should be small and would be from an "uninteresting" source.

Image-level Manipulations




Randomization
Image Dist.
Category Dist.

Training manipulations produce small differences in the classification performance across instances but **large differences in representations were found throughout the DNNs**.

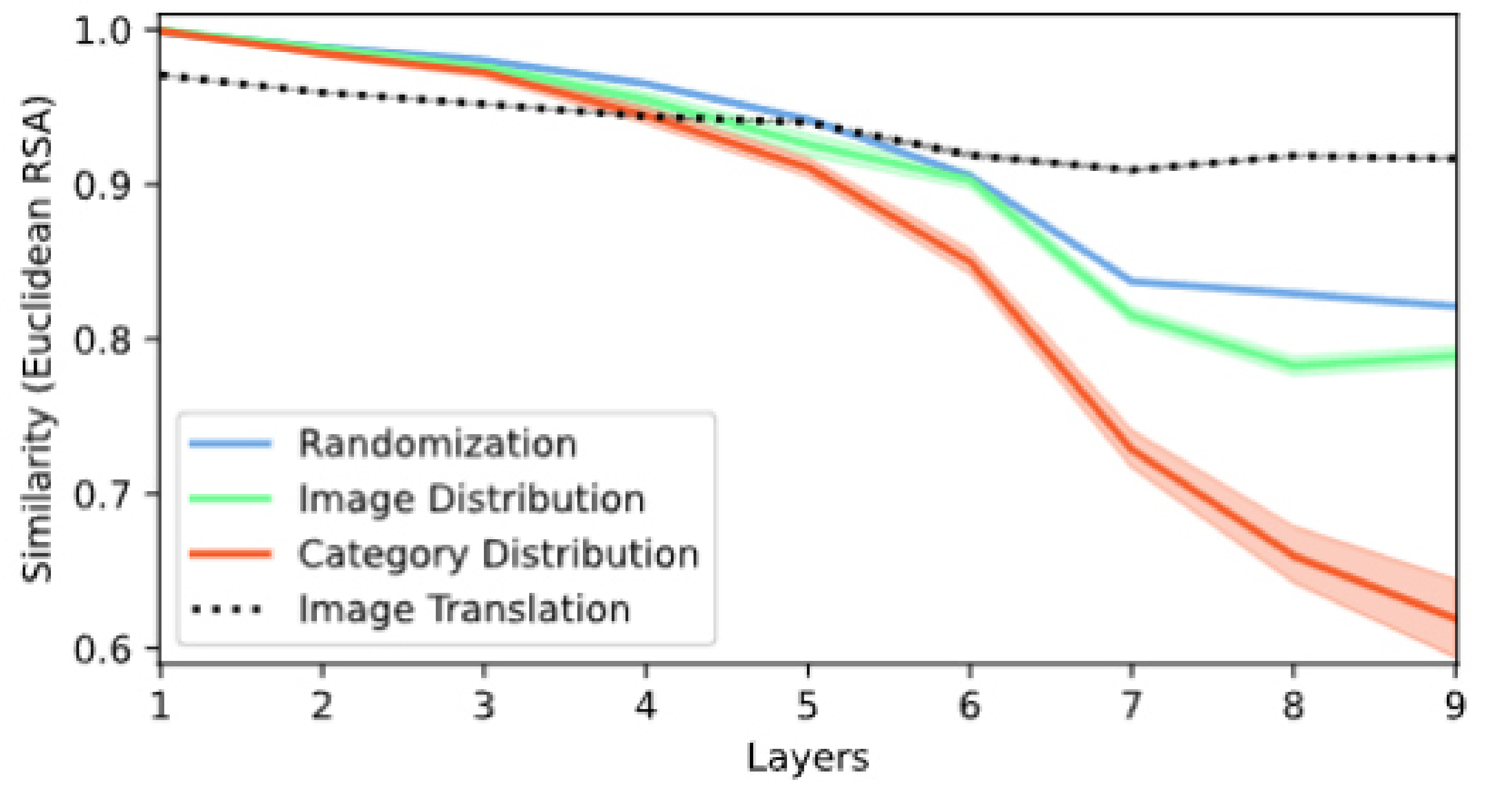


Image translation was used establish a baseline for variations in representations. *Other augmentation techniques caused similar magnitudes of variability.*

Dataset manipulations exceeded baselines earlier than the randomization manipulation.

Image distribution manipulation only caused variations at a magnitude similar to randomization manipulations.

Modifying image distribution only causes slightly more representational variability than modifying initial weights/ data randomization. Modifying category distribution causes large representational variations and may be useful to model individual differences in high-level visual cognition.

References

1. Kornblith, S., et al. (2019). Similarity of neural network representations revisited. ICML.
2. Kriegeskorte, N., et al. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2., 4.
3. Raghu M., et al. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. NIPS, 30.
4. Kornblith, S., et al. (2019). Similarity of neural network representations revisited. ICML.