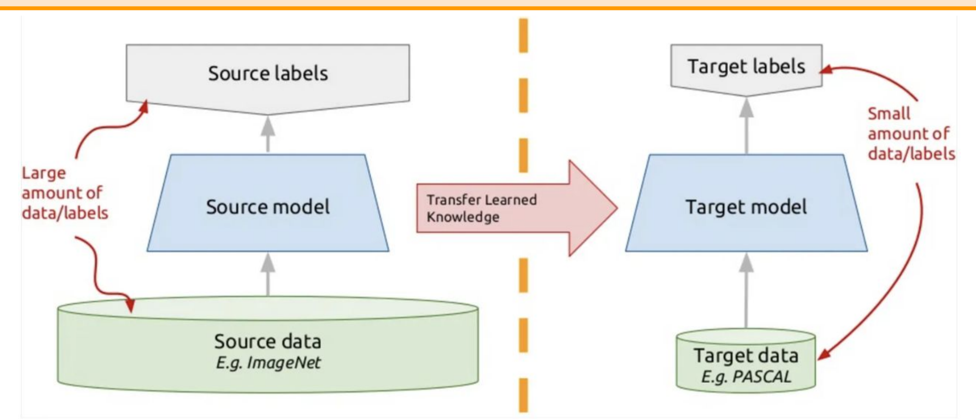
# What Variables Affect Out-of-Distribution Generalization in Pretrained Models?

Md Yousuf Harun[1*†], Kyungbok Lee[2*], Jhair Gallardo[1], Giri Krishnan[3], Christopher Kanan[2]

[1]Rochester Institute of Technology, [2]University of Rochester, [3]Georgia Tech, * Equal contribution, † Correspondence: mh1023@rit.edu

Project Website

## Motivation

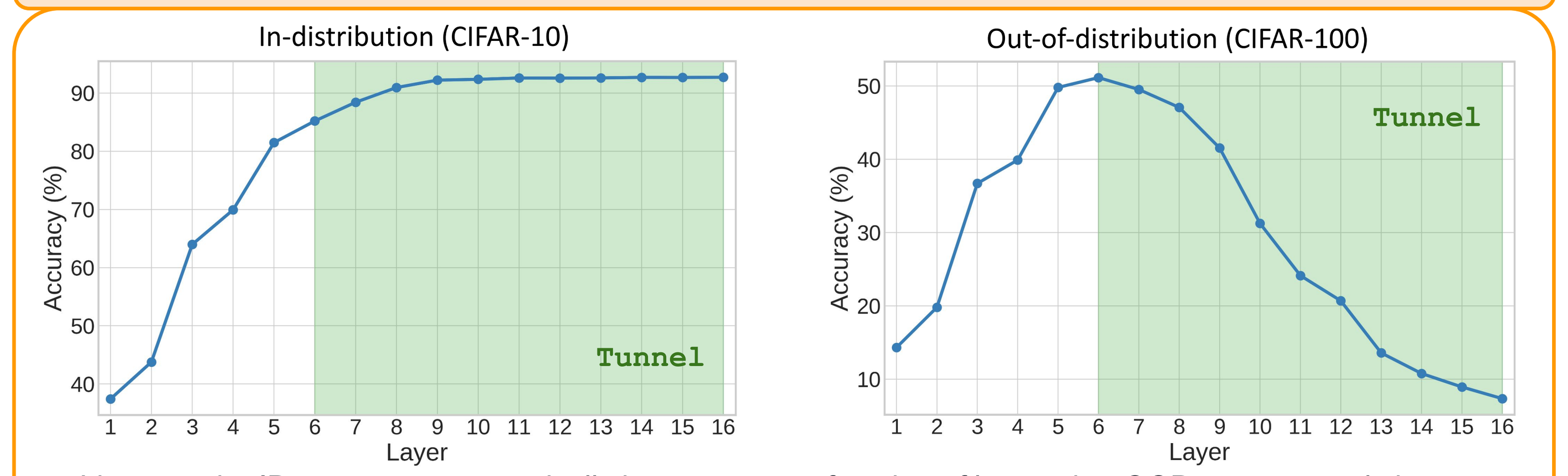



**Research Question:** How do in-distribution (ID) training data properties (resolution, # samples, and # classes), data augmentation, and architectural choices impact out-of-distribution (OOD) generalization and DNN representations?

**Research Question:** Why do methods validated on low-resolution datasets with a small number of classes (e.g., CIFAR-10) fail to generalize to high-resolution datasets with a large number of classes like ImageNet-1K?

**In a highly controlled study, we examine these questions through the lens of the Tunnel Effect Hypothesis, which is closely related to intermediate Neural Collapse**

## The Tunnel Effect Hypothesis

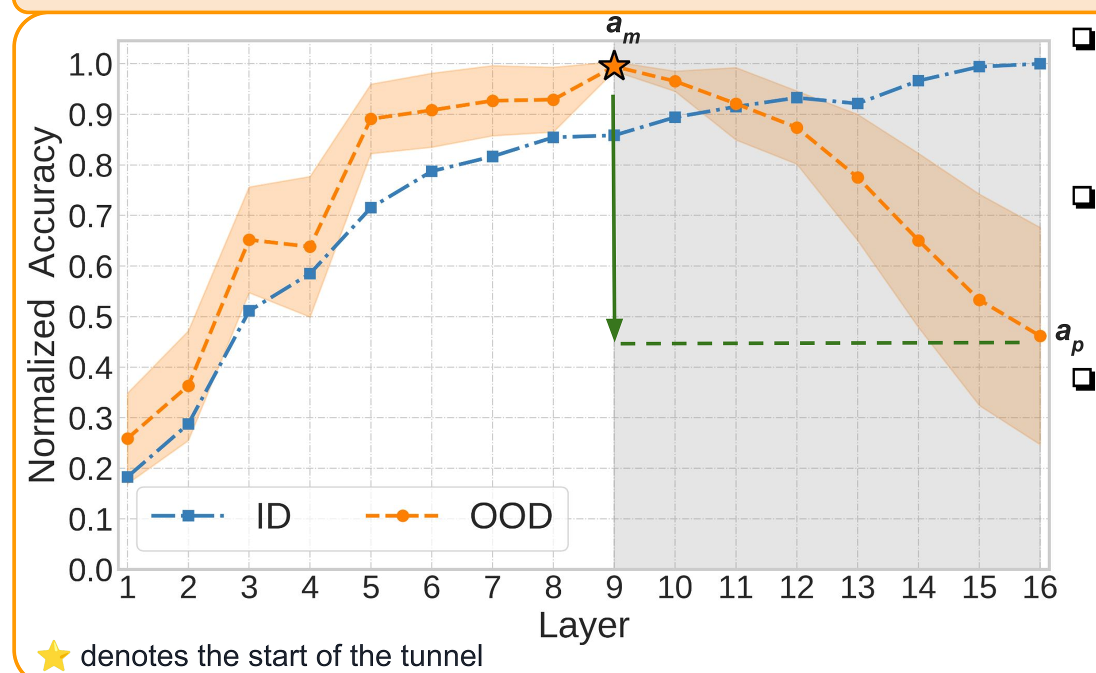



- Linear probe ID accuracy monotonically increases as a function of layers, but OOD accuracy only increases until the *tunnel* is reached and then decreases
- Earlier work used datasets with 32×32 images (CIFAR-10, etc.) for ID training data and did not measure tunnel effect *strength* (Masarczyk et al., NeurIPS 2023)
- Their findings are contrary to widely used transfer learning approaches with ImageNet-1K backbones

**Tunnel Effect Hypothesis:** An *overparameterized* N - layer DNN develops two *distinct* groups:

1. *The **extractor*** consists of the first K layers, creating linearly separable representations.
2. *The **tunnel*** comprises the remaining N − K layers, compressing representations and hindering OOD generalization.

## Measuring The Tunnel Effect Strength

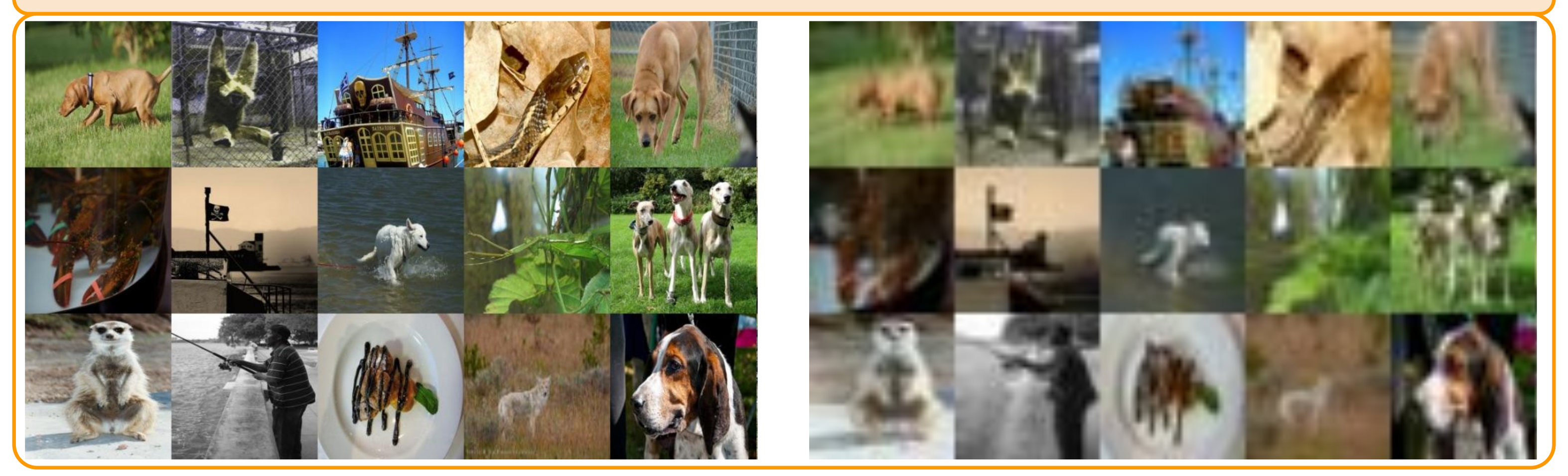



★ denotes the start of the tunnel

- **% OOD Performance Retained:** $r = (a_p / a_m) \times 100$, lower $r$ indicates stronger tunnel effect
- **Pearson Correlation:** $\rho = Corr(a_{ID}, a_{OOD})$, the higher the $\rho$, the lower the tunnel effect will be
- **ID/OOD Alignment:** $A = (a_{ID} - c_{ID}) \times (a_{OOD} - c_{OOD})$, distinguishes low-performing and high-performing models

Here, $c_{ID}$ and $c_{OOD}$ denote chance accuracy. 1st two metrics use normalized accuracy whereas 3rd one uses raw accuracy

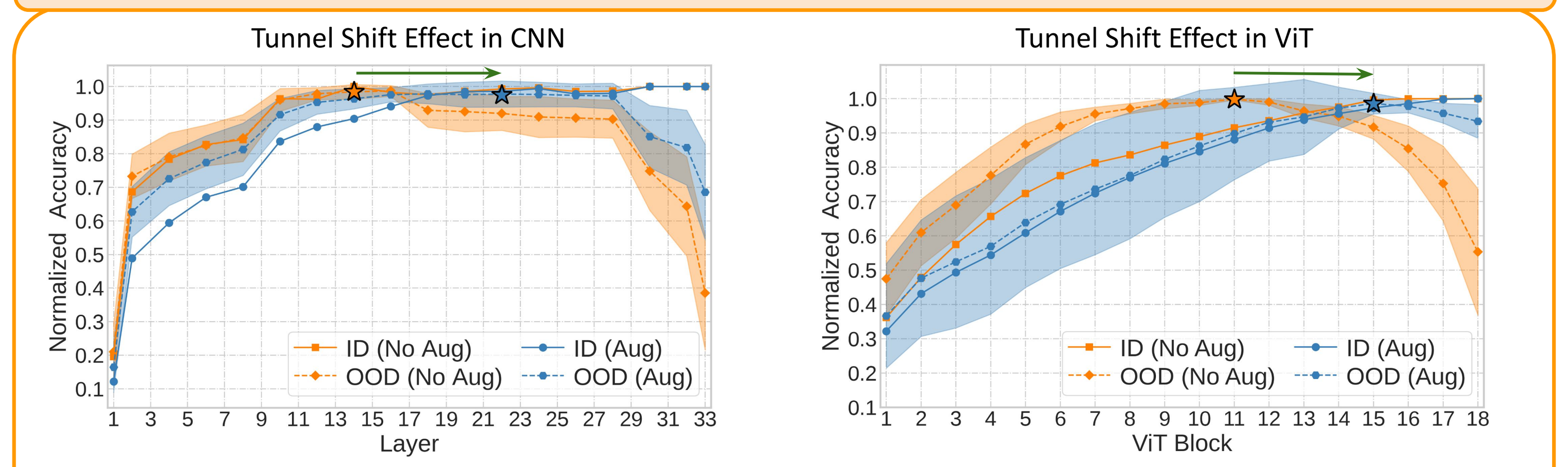
## ImageNet at 224×224 VS 32×32

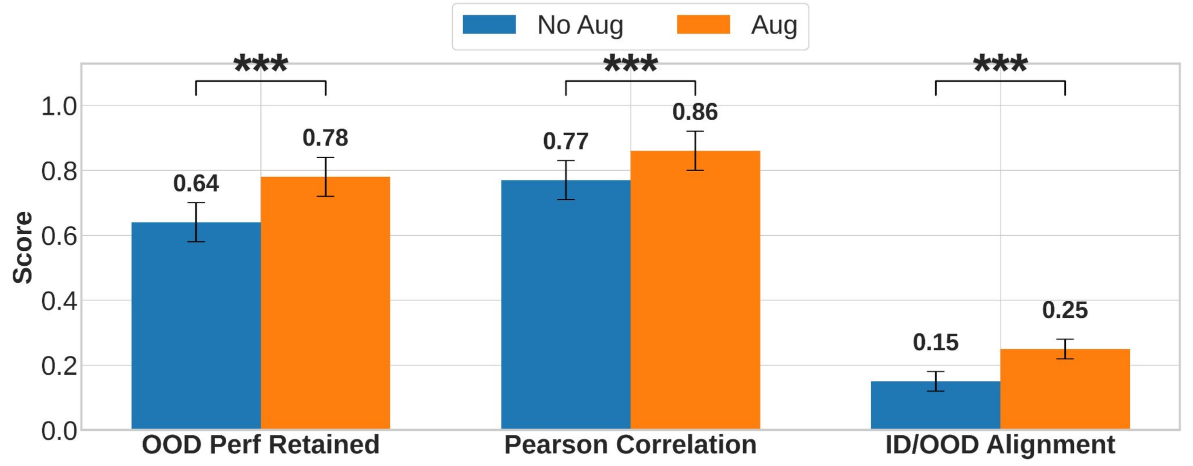


## Experimental Setup

- Used ImageNet-100 and its variants as ID datasets and 9 widely-used OOD datasets e.g., NINCO, ImageNet-R, CIFAR-10, CIFAR-100, Oxford 102 Flowers, CUB-200, Aircrafts, Oxford Pets, and STL-10
- Studied total **8 variables:** 1) image resolution (32/64/128/224), 2) ID class count, 3) augmentations (random crop & flip), 4) overparameterization, 5) depth, 6) spatial reduction, 7) stem, and 8) CNN vs. ViT
- Trained and assessed **64 ID backbones** and **8,604 linear probes,** resulting in 512 values per metric
- Performed *paired* tests to study impact of each variable in isolation for every combination of other variables
- Paris are constructed to control for the impact of other variables
- Conducted hypothesis testing for statistical significance. *P*-values are denoted by stars according to $*P<0.05$, $**P<0.01$, $***P<0.001$, $****P<0.0001$
- Jointly analyzed and ranked variables using "SHAP Slope", our SHAP-based analysis

## Research Question: How does image augmentation impact the tunnel strength?



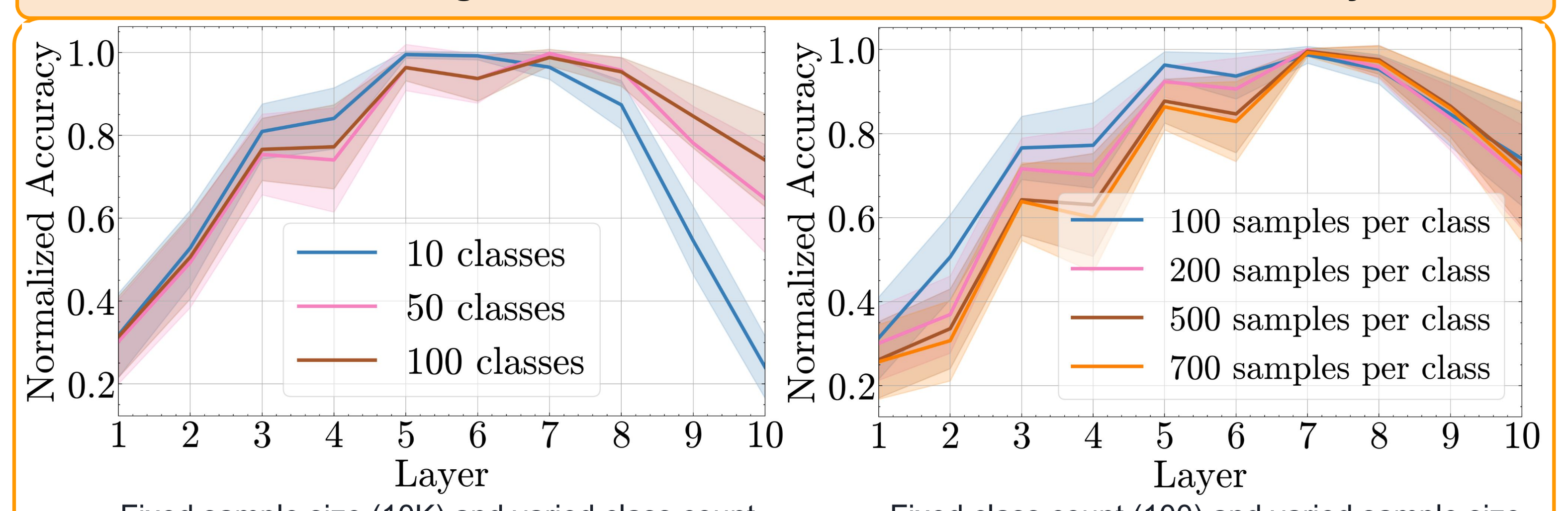




**Findings:** 256 *paired* experiments without augmentations and another 256 with augmentations

- Augmentations shifted & weakened the tunnel across various architectures
- Augmentations improved OOD generalization

**Takeaway:** Augmentations increase training data diversity and decrease the tunnel strength

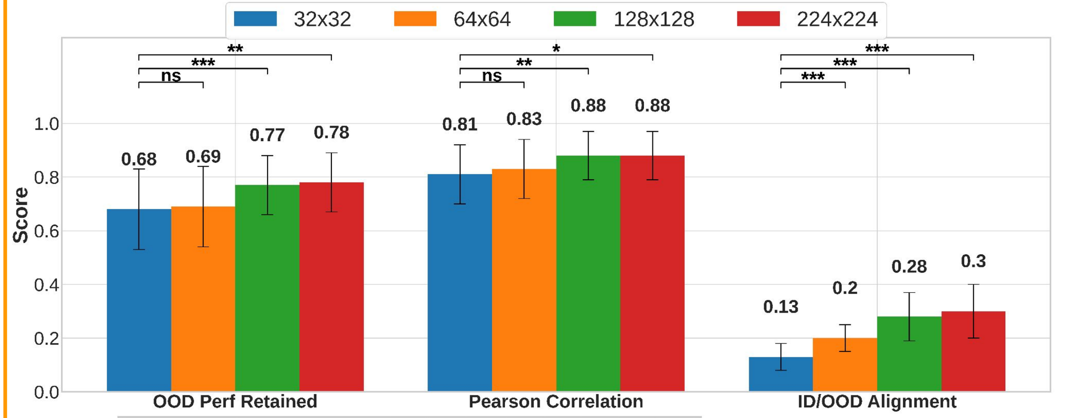
## Wider Coverage of Semantics Matters More Than Data Quantity




Fixed sample size (10K) and varied class count

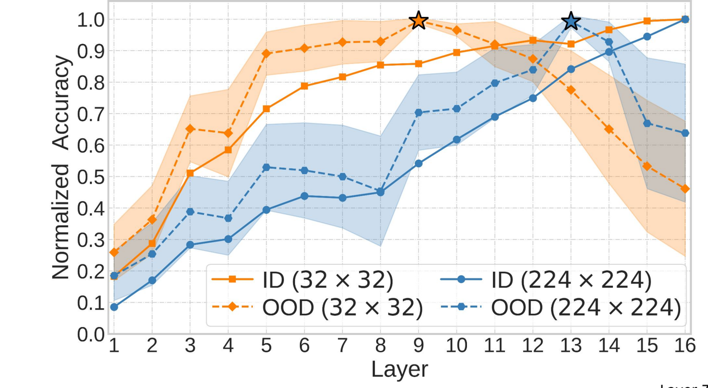
Fixed class count (100) and varied sample size

## Research Question: How does image resolution impact the tunnel strength?




**Findings:** Conducted *paired* tests between models trained with 32×32 images and those trained with 64×64, 128×128, or 224×224 resolution images (48 paired experiments per group)

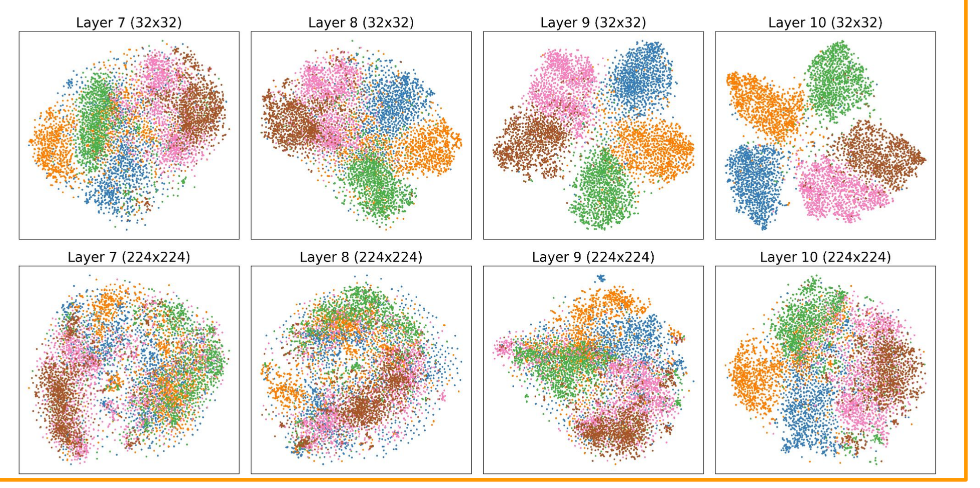
- Increasing image resolution improves OOD performance across all criteria



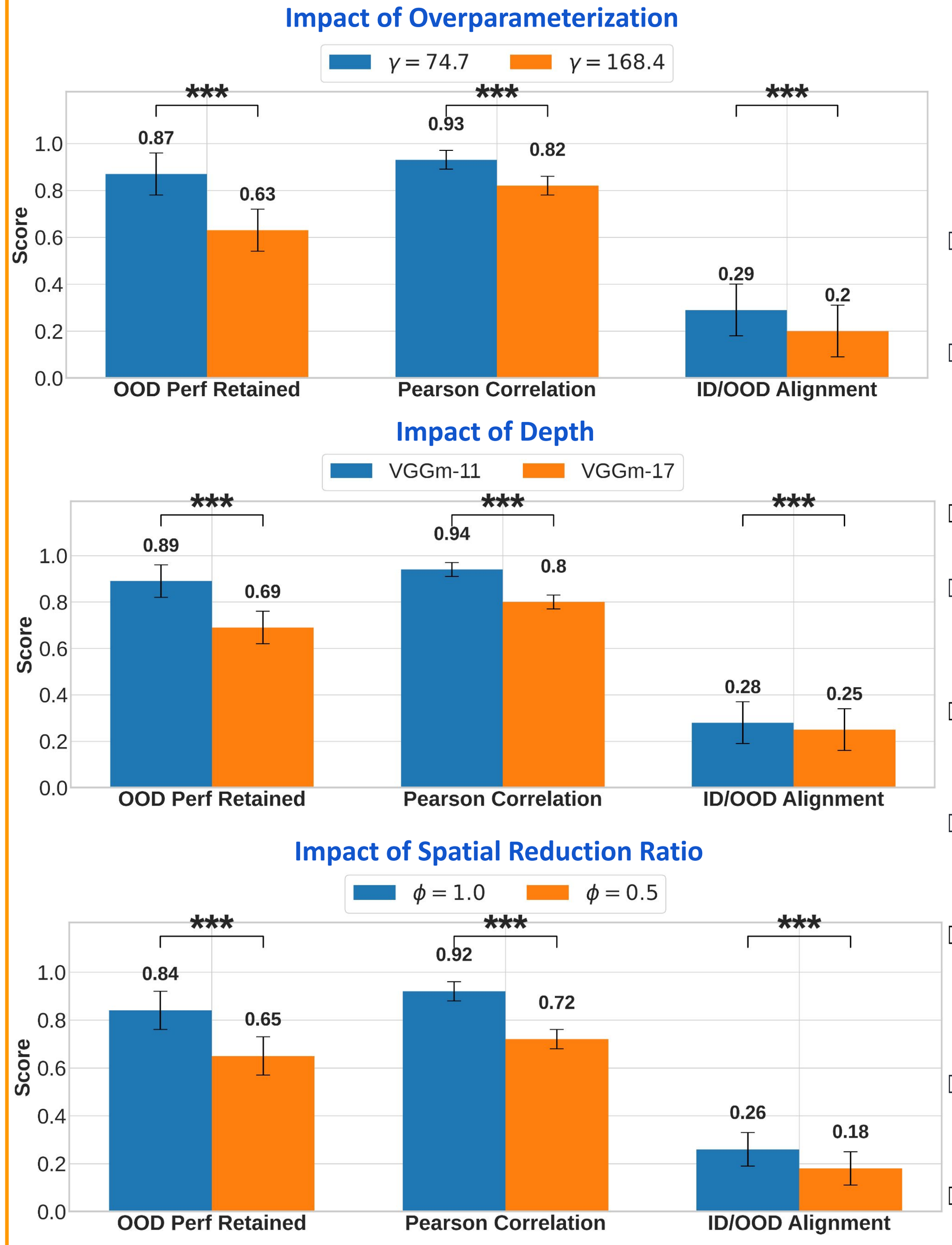

- Models trained on low-resolution images develop *longer* tunnel than the models trained on high-resolution images
- **Takeaway:** High image resolution reduces the tunnel strength and improves OOD generalization

**Representation Compression:**

- Models trained on low-resolution images exhibit much greater *representation compression* than models trained on high-resolution images
- This is likely *why* many methods that work well for low-resolution datasets e.g., CIFAR fail to perform effectively for high-resolution datasets e.g., ImageNet



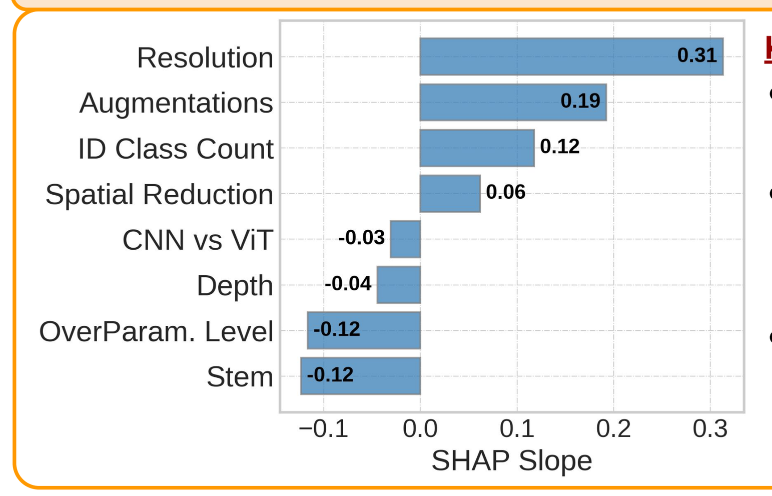

## Research Question: How do DNN architecture variables influence the tunnel effect?




**Findings:** Conducted 416 experiments using 8 DNN architectures drawn from 3 families: VGG, ResNet, and ViT. Each architecture uses the same # parameters across image resolutions.

- Overparameterization $\gamma = P / N$ where P and N denote # DNN parameters and # training samples
- Impact of overparameterization ($\gamma$): increasing $\gamma$ impairs OOD performance
- Depth refers to number of layers for CNN and blocks for ViT
- Impact of depth: increasing depth hurts OOD generalization
- Stem refers to the kernel size (k × k) in the first layer of CNN. It is the patch size for ViT
- Impact of stem: large stem size impairs OOD generalization
- Spatial reduction ratio is defined as the ratio of the output spatial dimension to the input spatial dimension
- Impact of spatial reduction ratio ($\phi$): decreasing $\phi$ impairs OOD transfer
- CNN vs ViT: negligible impact

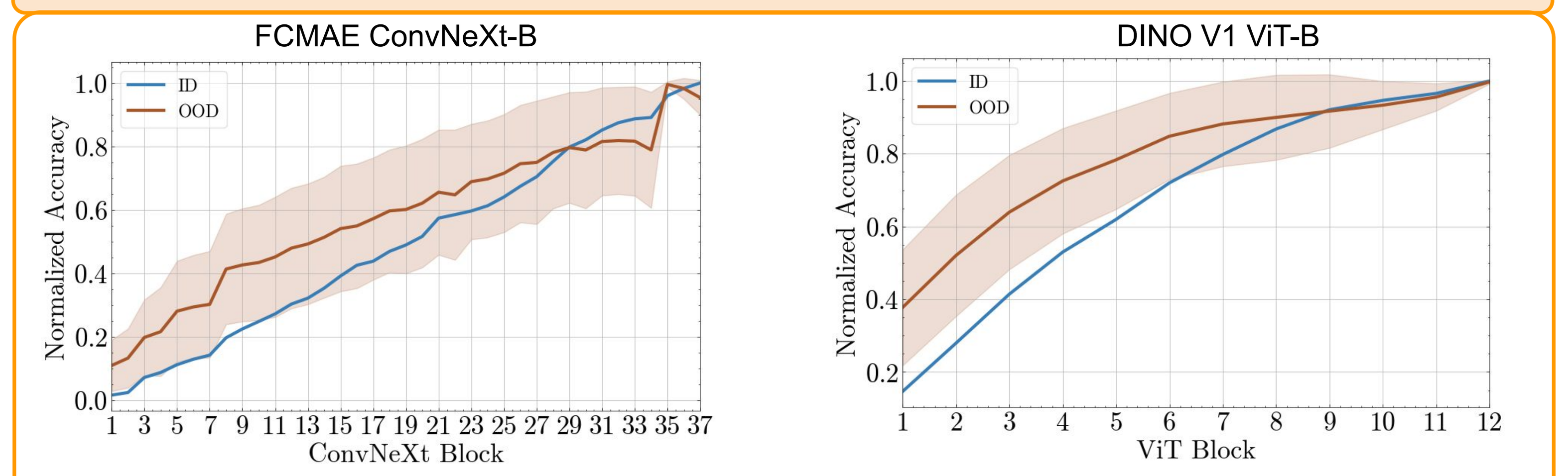
## SHAP Analysis: What Variable Matters Most?




**Key Findings:**

- SHAP results using ID/OOD alignment as the target enable us to measure each variable's impact
- Resolution was found most dominant followed by augmentations and ID class count in terms of reducing tunnel strength
- Overparameterization level, stem, and depth increase the tunnel strength but their impact is much less than dataset variables

## Do Self-Supervised Models Develop Tunnel?




**Findings:** Most large-scale pretrained models including self-supervised ones did not exhibit any tunnel effect

## Additional Findings & Summary

**Additional Findings:**

- SHAP analysis using % OOD performance retained and Pearson correlation as targets revealed that
  - Dataset variables greatly improved OOD generalization where ID class count had the greatest impact followed by augmentations, spatial reduction, and resolution
  - DNN architecture variables e.g., overparameterization and depth impaired OOD generalization
  - Dataset variables had a stronger influence on the tunnel effect than DNN architecture variables
- Among 64 ID backbones, 4 did not exhibit any tunnel effect, suggesting that the tunnel effect is not a universal phenomenon and its strength depends on various factors
- Training on datasets with the properties similar to CIFAR results in very *different representations* in the "tunnel" compared to training on ImageNet, as well as much stronger tunnels
- Tunnel plays *task-specific* role and impacts forgetting in continual learning

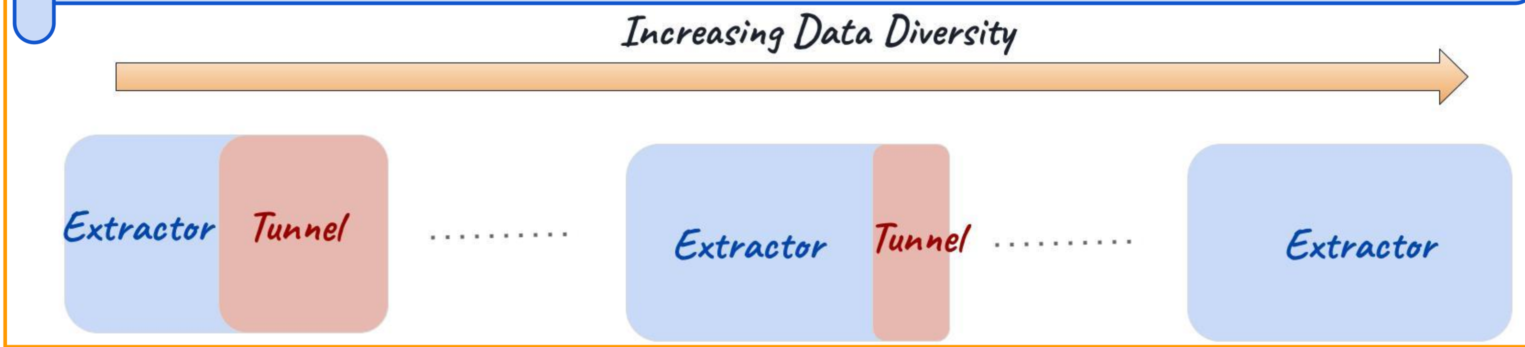
**Summary:**

- ★ Increasing ID class count (*between-class diversity*), using augmentations (*within-class diversity*), and using higher image resolution (*hierarchical features*) greatly reduce the tunnel strength and improve OOD transfer
- ★ DNN variables e.g., over-parameterization and depth, increase the tunnel effect, but their impact is much smaller than that of the aforementioned dataset variables
- ★ Concretely, we observe that increasing dataset diversity plays a major role in mitigating the tunnel effect
- ★ This leads us to *revise* the tunnel effect hypothesis as follows:

**Revised Tunnel Effect Hypothesis:** An *overparameterized* N - layer DNN develops two *distinct* groups:

1. *The **extractor*** consists of the first K layers, creating linearly separable representations.
2. *The **tunnel*** comprises the remaining N − K layers, compressing representations and hindering OOD generalization.

K is proportional to the diversity of training inputs, where if diversity is sufficiently high, N = K (no tunnel)




## Acknowledgments

We thank US National Science Foundation for supporting our research