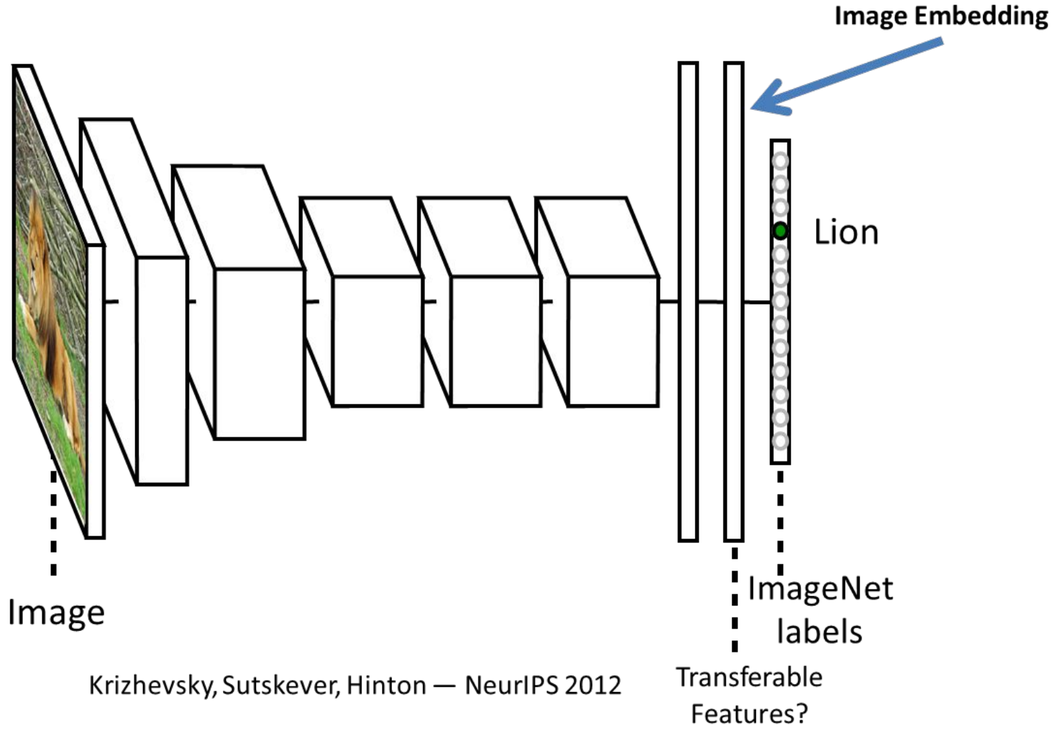
What Variables Affect Out-of-Distribution Generalization in Pretrained Models?

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

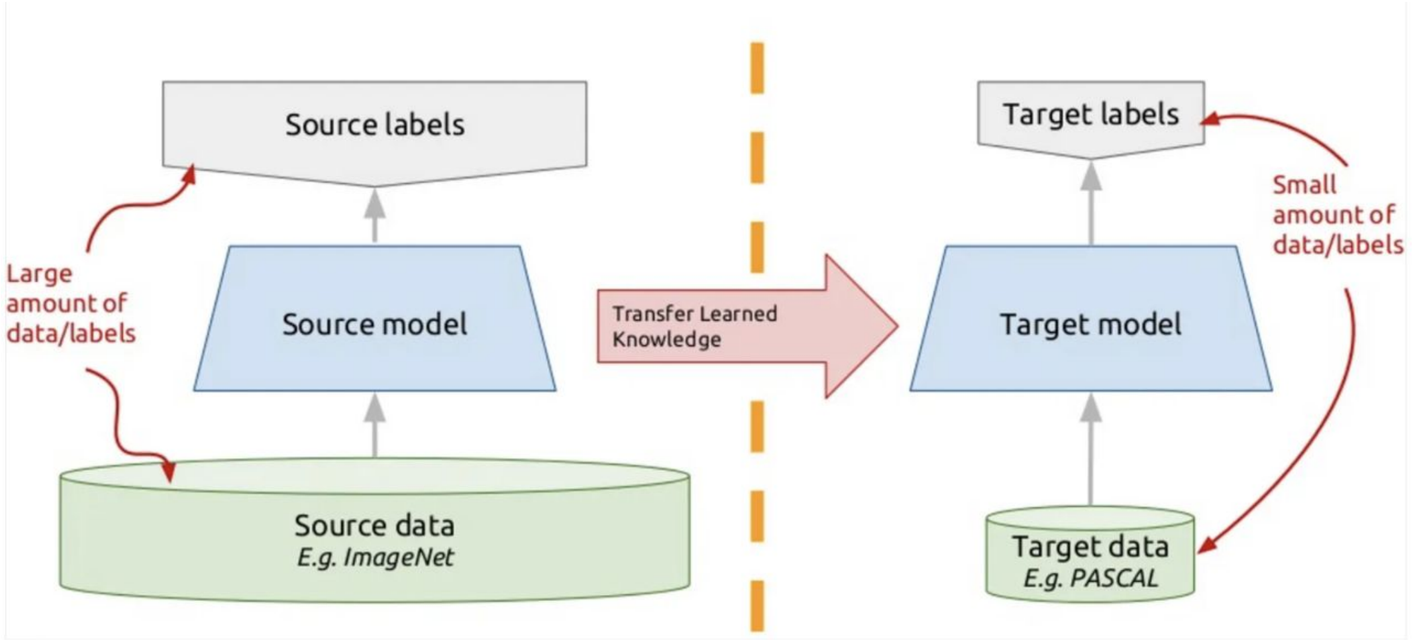
[1]*Rochester Institute of Technology,* [2]*University of Rochester,* [3]*Georgia Tech*

Project Website

Motivation: How do we build good representations?



Krizhevsky, Sutskever, Hinton — NeurIPS 2012

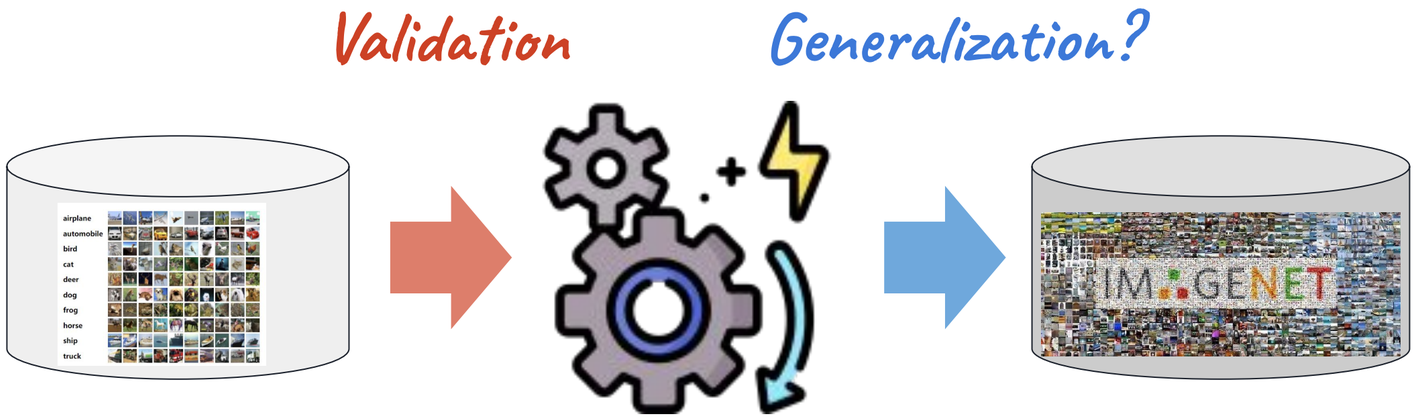


Research Question:

How do variables e.g., training data, object categories, image resolution, architecture etc. impact representations and out-of-distribution (OOD) generalization?

Motivation

Findings on toy datasets often fail to generalize!



Toy Datasets
e.g., CIFAR
(10 classes, 32x32 images)

Methods & Findings

Real-world Datasets
e.g., ImageNet
(1K classes, 224x224 images)

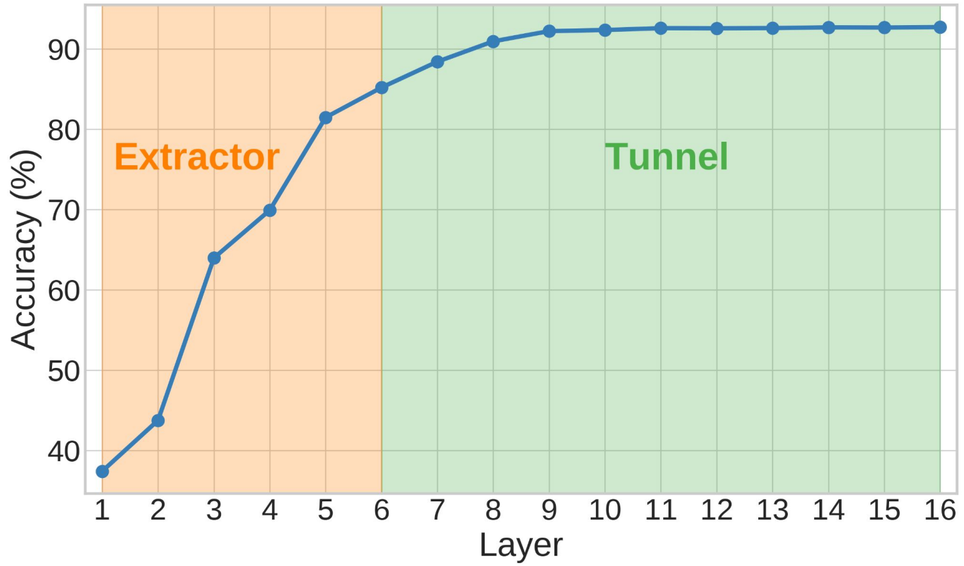
This happens in areas:

- Continual Learning
- Active Learning
- Open Set Recognition
- OOD Detection
- Uncertainty Quantification
- Dataset Distillation
- And so on..

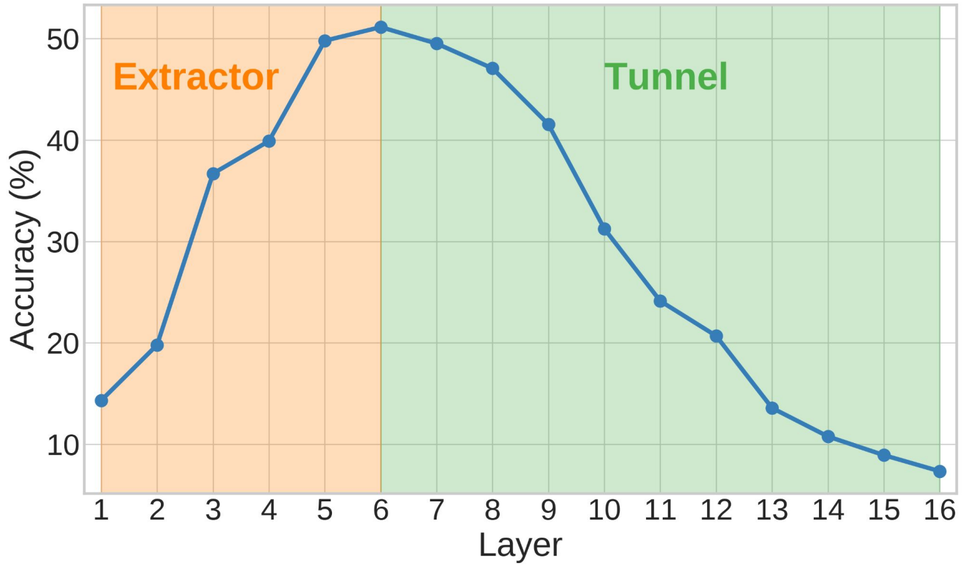
Research Question:

Why do methods & findings validated on toy datasets fail to generalize to real-world datasets?

Prior Work & Background



ID Dataset: CIFAR-10



OOD Dataset: CIFAR-100

Findings:

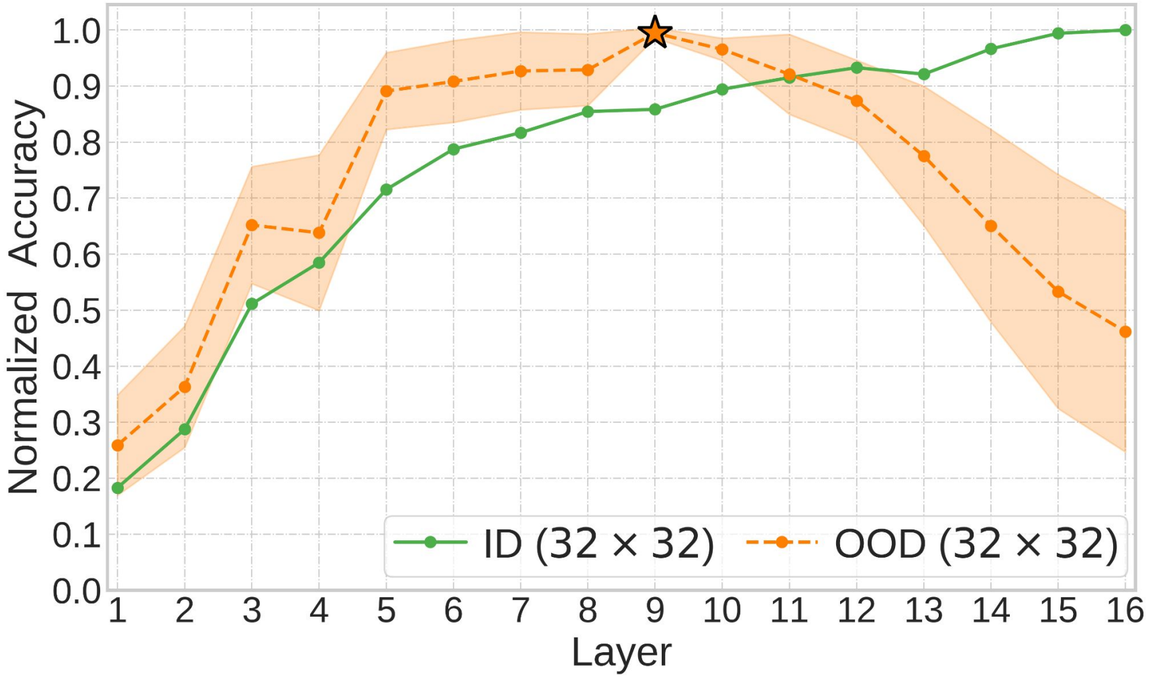
- Linear probe ID accuracy monotonically increases as a function of layers
- Linear probe OOD accuracy goes up and then down
- The tunnel is where the OOD accuracy starts to go down
- Earlier work mainly studied low-resolution datasets and did not measure the strength of the tunnel effect (Masarczyk et al., NeurIPS 2023)

This suggests that when tunnel exists, current practice i.e., using embeddings from penultimate layer will be sub-optimal for downstream tasks!

The Tunnel Effect Hypothesis

An *overparameterized* N - layer DNN forms 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.



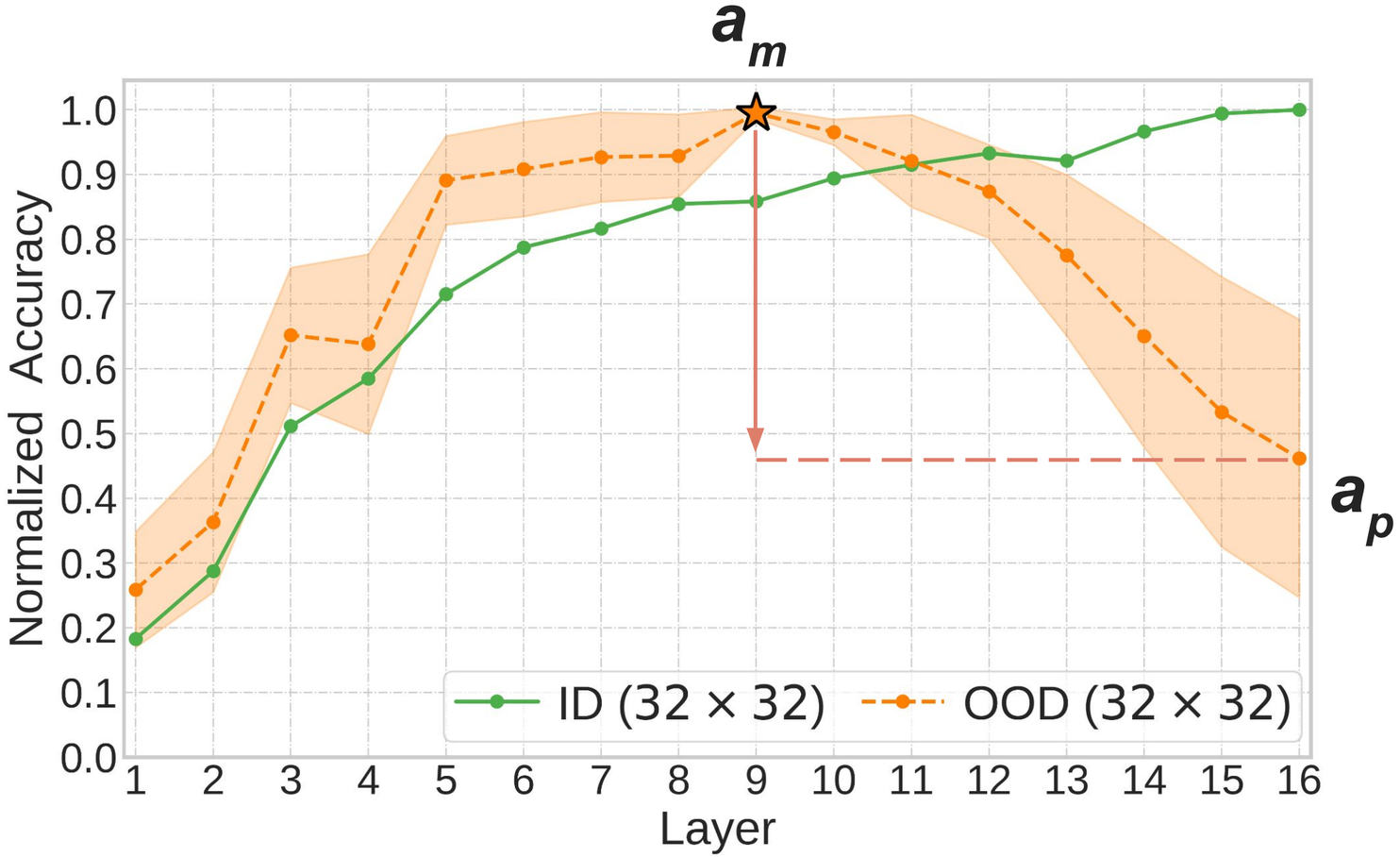
- VGGm-17 was trained on ImageNet-100 (32x32 images)
- Linear probes were trained on ID and OOD datasets for each layer
- Y-axis shows normalized accuracy (divided by max)
- **Extractor** consists of first 8 layers
- **Tunnel** spans from layer 9 to 16

☆ denotes the start of the tunnel

Measuring Tunnel Effect Strength

Percentage OOD Performance Retained

- The OOD performance drops in the tunnel as a function of layer index
- The lower the OOD accuracy in the last layer, the stronger the tunnel effect
- We introduce a metric to capture this



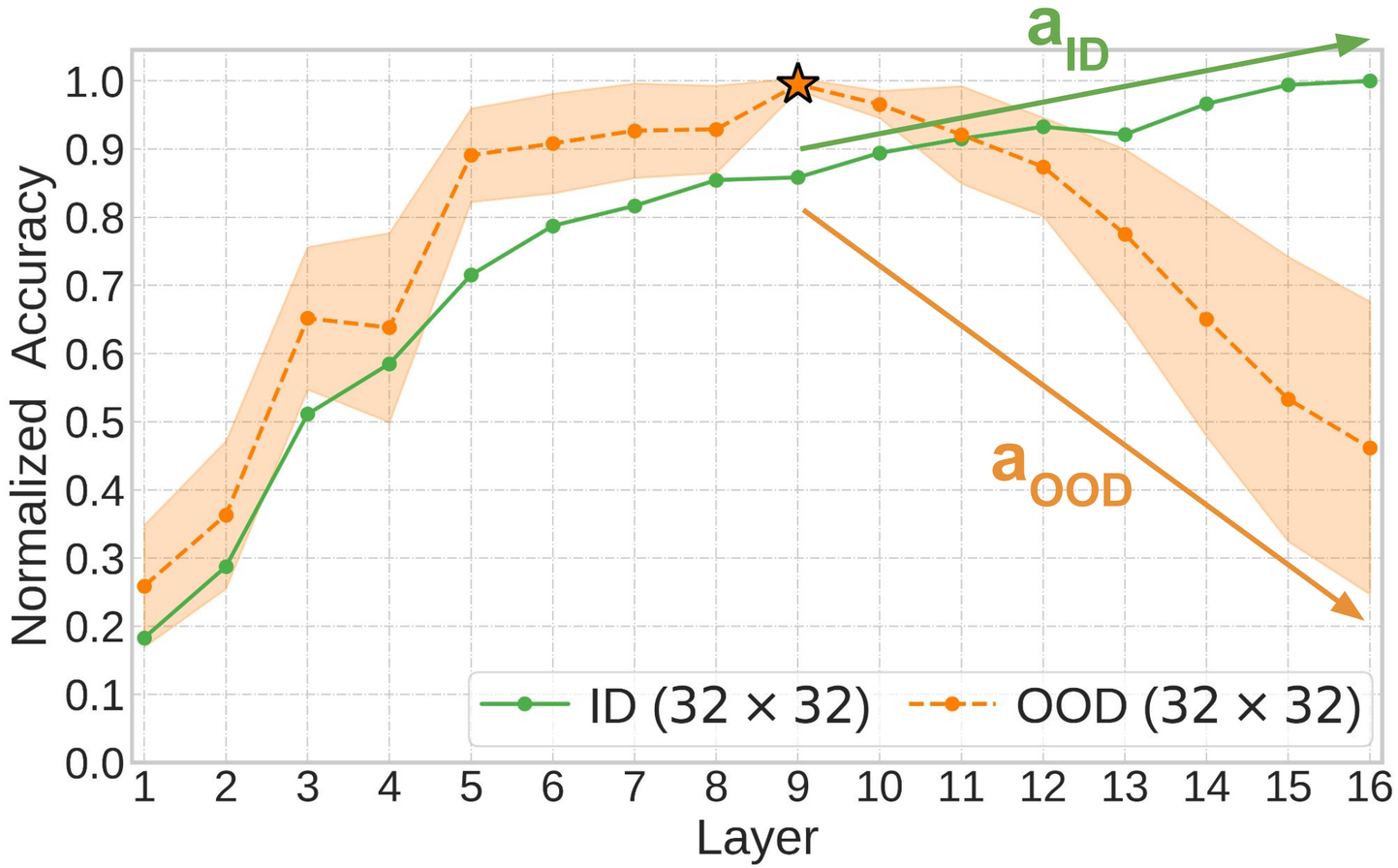
% OOD Perf Retained,
$r = (a_p / a_m) \times 100$

- Higher r indicates higher OOD generalization, hence a weaker tunnel and vice-versa
- When $a_p = a_m$, there is no tunnel

☆ denotes the start of the tunnel

Pearson Correlation between ID & OOD

- **No tunnel effect:** Both ID and OOD curves will uniformly go upward
- **Tunnel effect:** While ID is going up, OOD goes down and diverges from ID
- The higher the correlation between ID & OOD, the lower the tunnel effect will be



Pearson Correlation,
$\rho = Corr\ (a_{ID}\ ,\ a_{OOD})$

- Higher ρ indicates less tunnel effect and vice-versa

☆ denotes the start of the tunnel

ID / OOD Alignment

We also introduce another metric based on raw accuracy to distinguish between low-performing and high-performing models.

ID/OOD Alignment,

$$A = (a_{ID} - c_{ID}) \times (a_{OOD} - c_{OOD})$$

Weak Model: Low ID accuracy and/or OOD accuracy

Strong Model: High ID & OOD accuracy

- Higher ***A*** indicates greater alignment between ID and OOD performance
- ***c*** is random guess

Design of Our Study

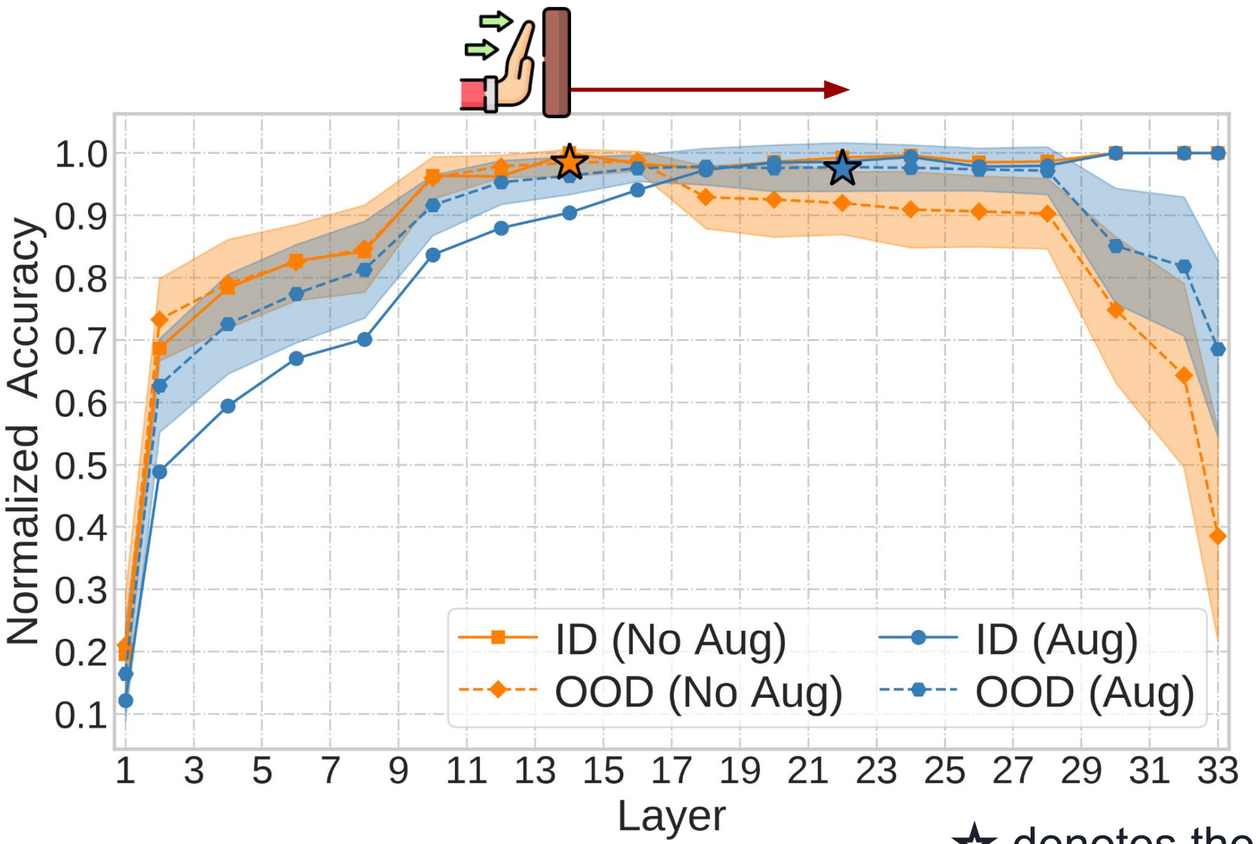
- **3 Dataset variables**:
  - Image Resolution (32x32, 64x64, 128x128, 224x224)
  - ID Class Count
  - Augmentations
- **5 DNN variables**:
  - Overparameterization level
  - Depth
  - Spatial reduction ratio
  - Stem
  - DNN type (CNN vs. ViT)

- Trained and assessed **64 ID backbones** and over **10K linear probes**
- Performed paired tests to study impact of each variable in isolation for every combination of other variables
- Jointly analyzed and ranked variables using “SHAP Slope”, our proposed SHAP-based analysis

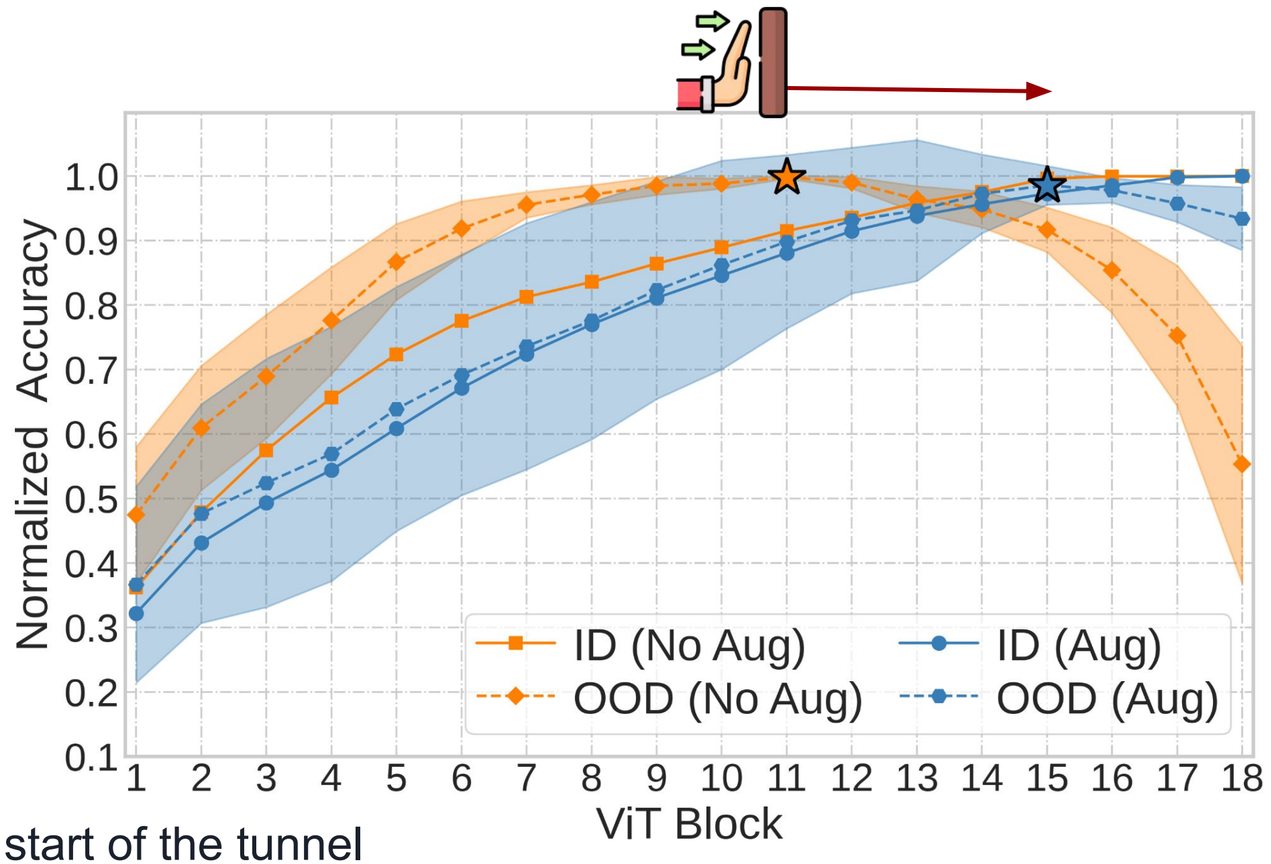
Research Question: How does image augmentation impact the tunnel strength?

Tunnel Effect is Influenced by Augmentation

Tunnel shift effect in CNN



Tunnel shift effect in ViT

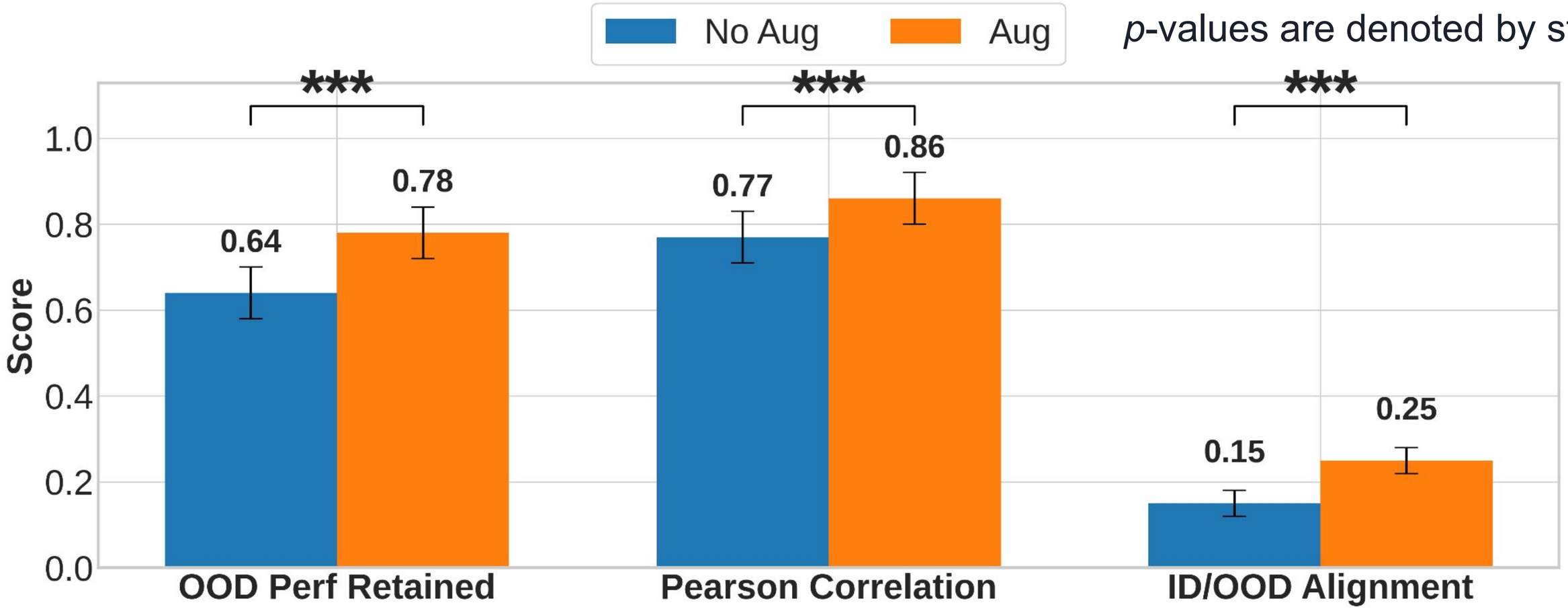


☆ denotes the start of the tunnel

We conducted 256 *paired* experiments (256 without aug and 256 with aug), 512 in total.
We used random resized crop and random horizontal flip augmentations.

Takeaway: Augmentations increase data diversity and thereby decrease the tunnel strength

Impact of Image Augmentation

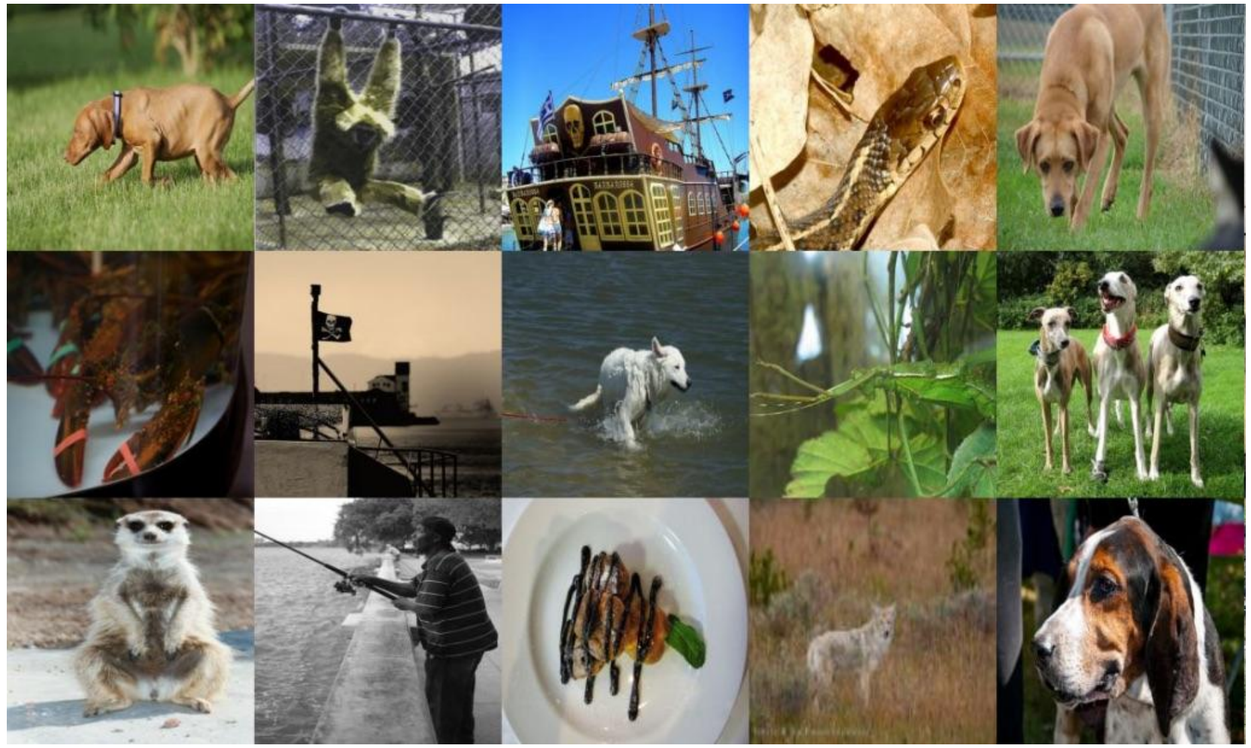


- **Augmentations greatly improved OOD generalization across all metrics.**
- *p*-values are based on Wilcoxon signed-rank tests.
- $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$
- All paired groups achieved *medium* effect size (Cliff's delta).

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

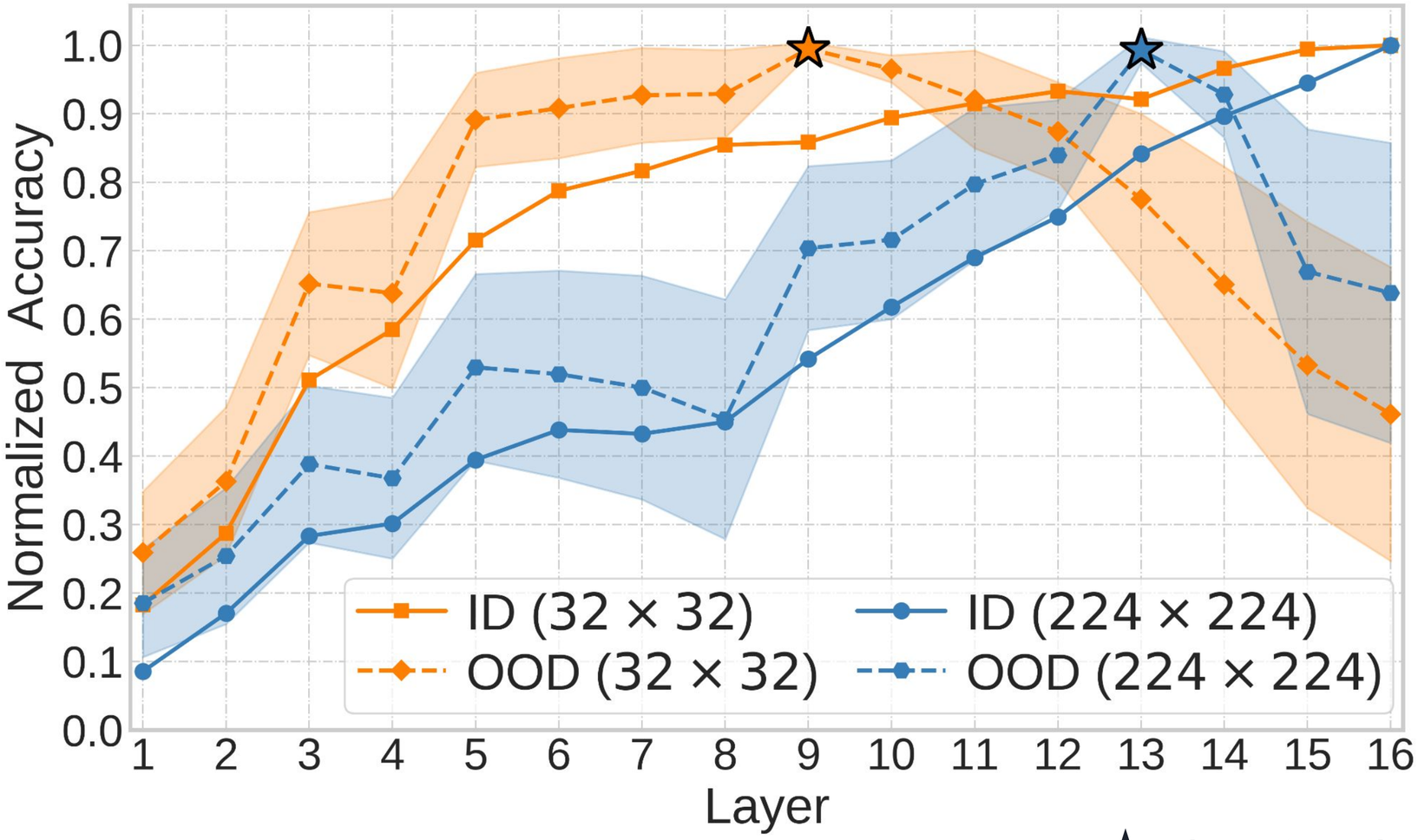
ImageNet at 224×224 VS 32×32

224×224 Resolution



32×32 Resolution

Impact of Image Resolution

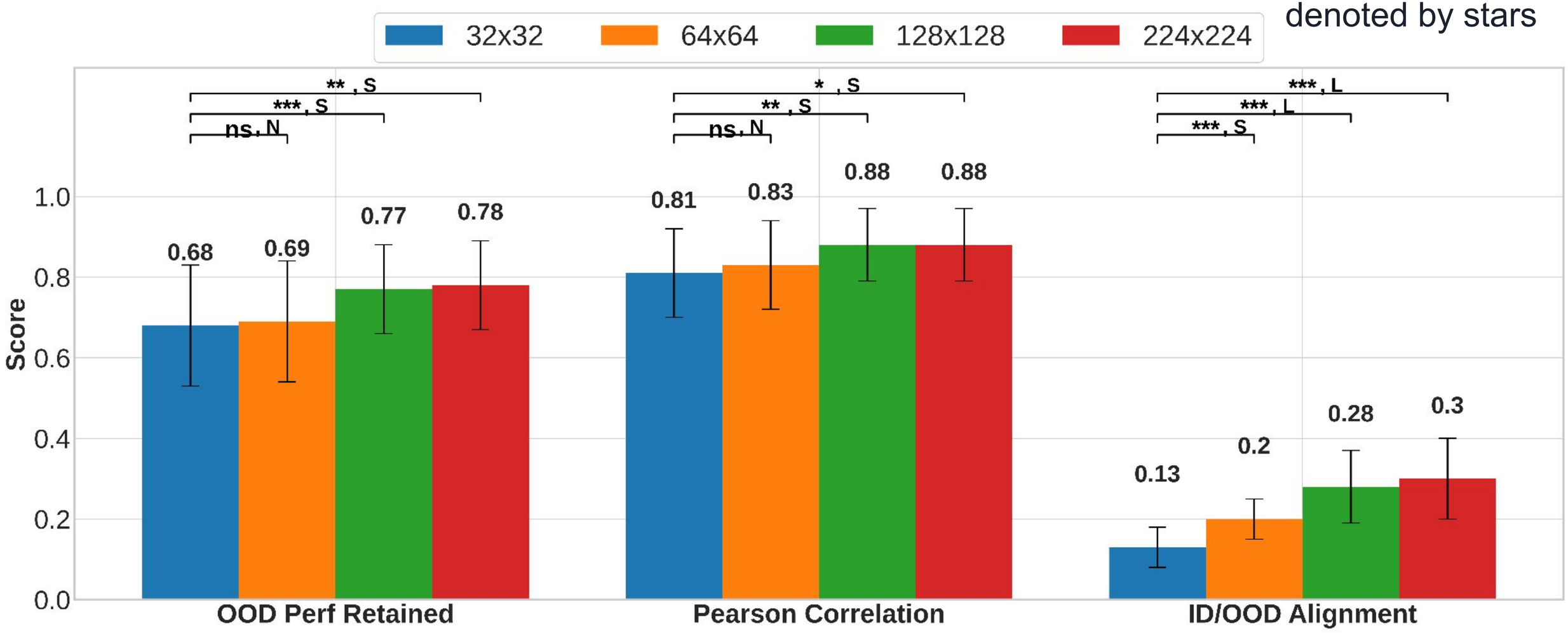


☆ denotes the start of the tunnel

Takeaway: Models trained on *low-resolutions* images develop *longer* tunnel than models trained on high-resolution images.

Impact of Image Resolution

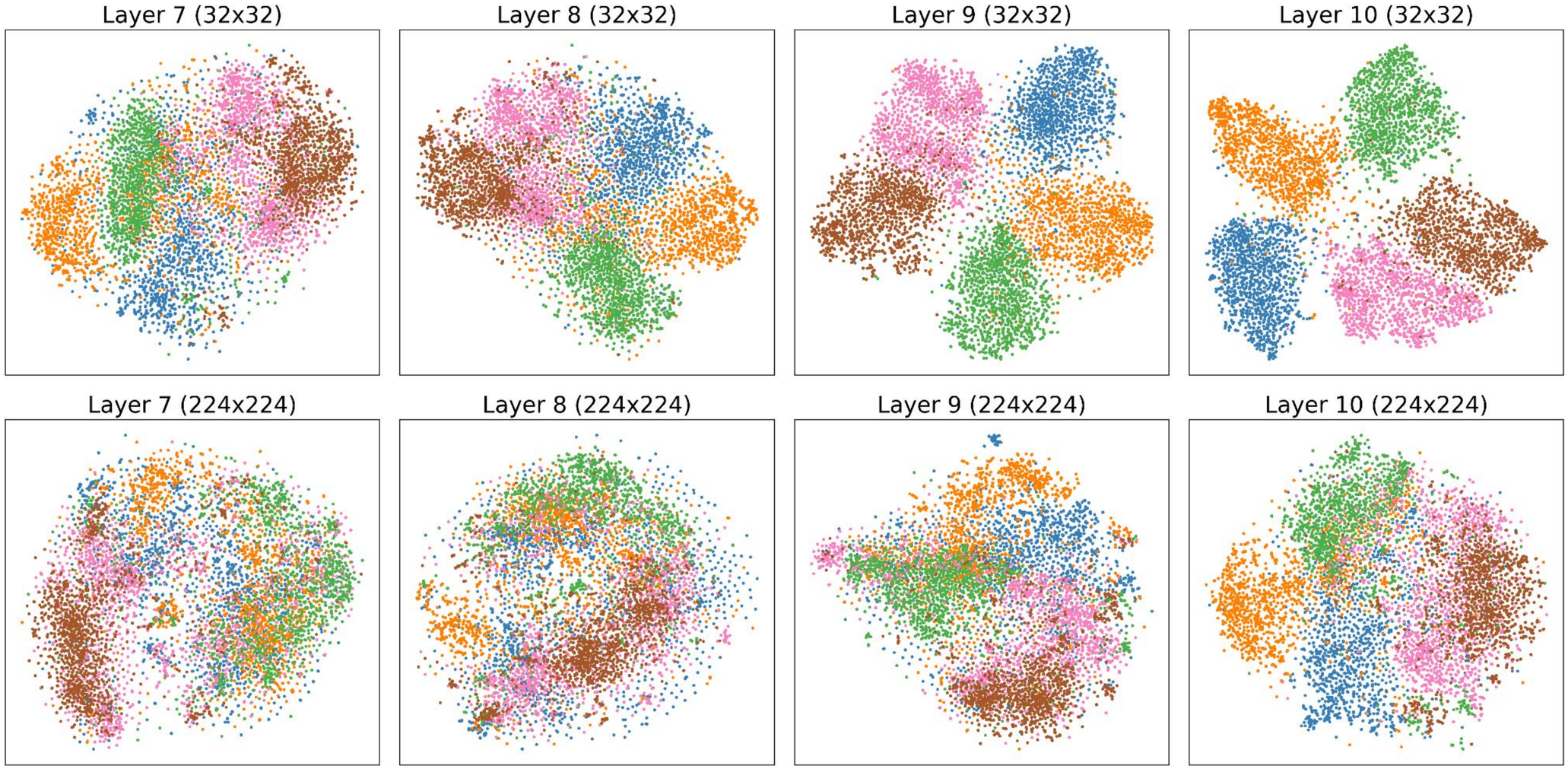
p-values are denoted by stars



- **Increasing image resolution improves OOD generalization across all metrics.**
- *p*-values are based on Wilcoxon signed-rank tests. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$. ns means *not significant*. 48 paired tests per resolution, 192 total
- For mean effect size (Cliff's delta), N, S, M, & L denote negligible, small, medium & large respectively.

Representation Compression

Findings: Models trained on low-resolution inputs exhibit much greater representation compression than models trained on high-resolution inputs.

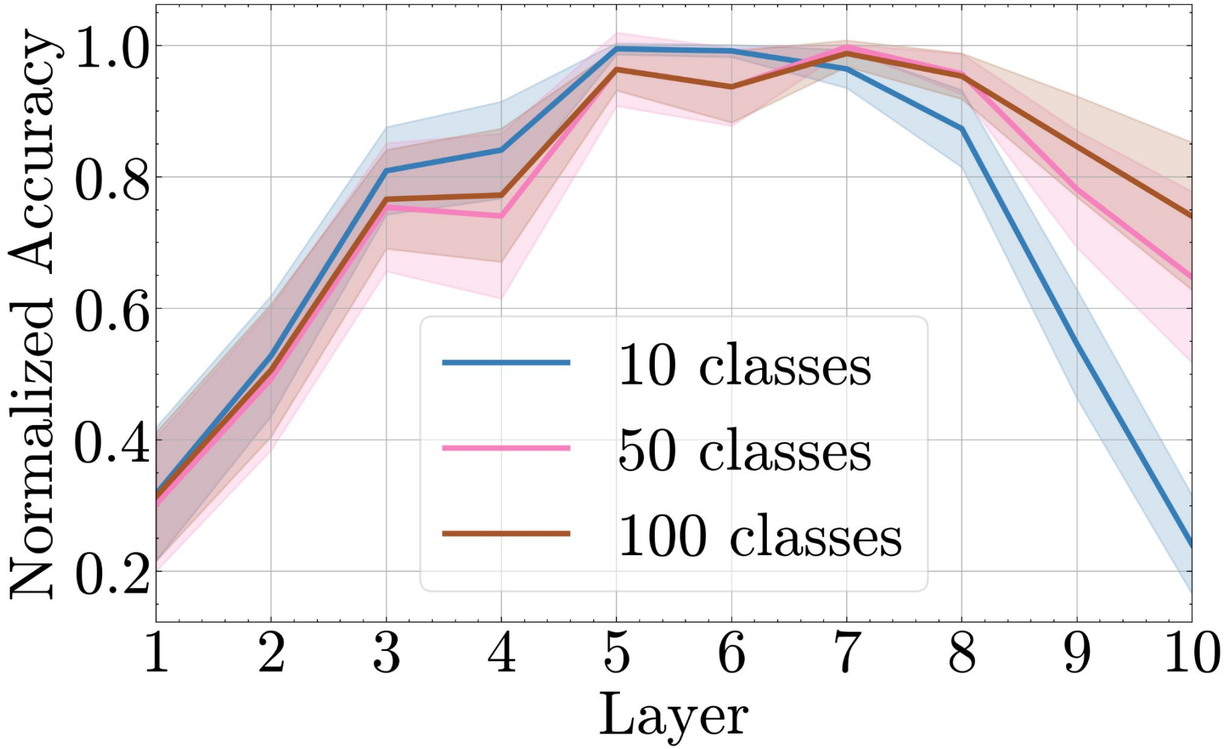


Research Question: **Which has a greater impact on tunnel strength—data quantity or semantic variability?**

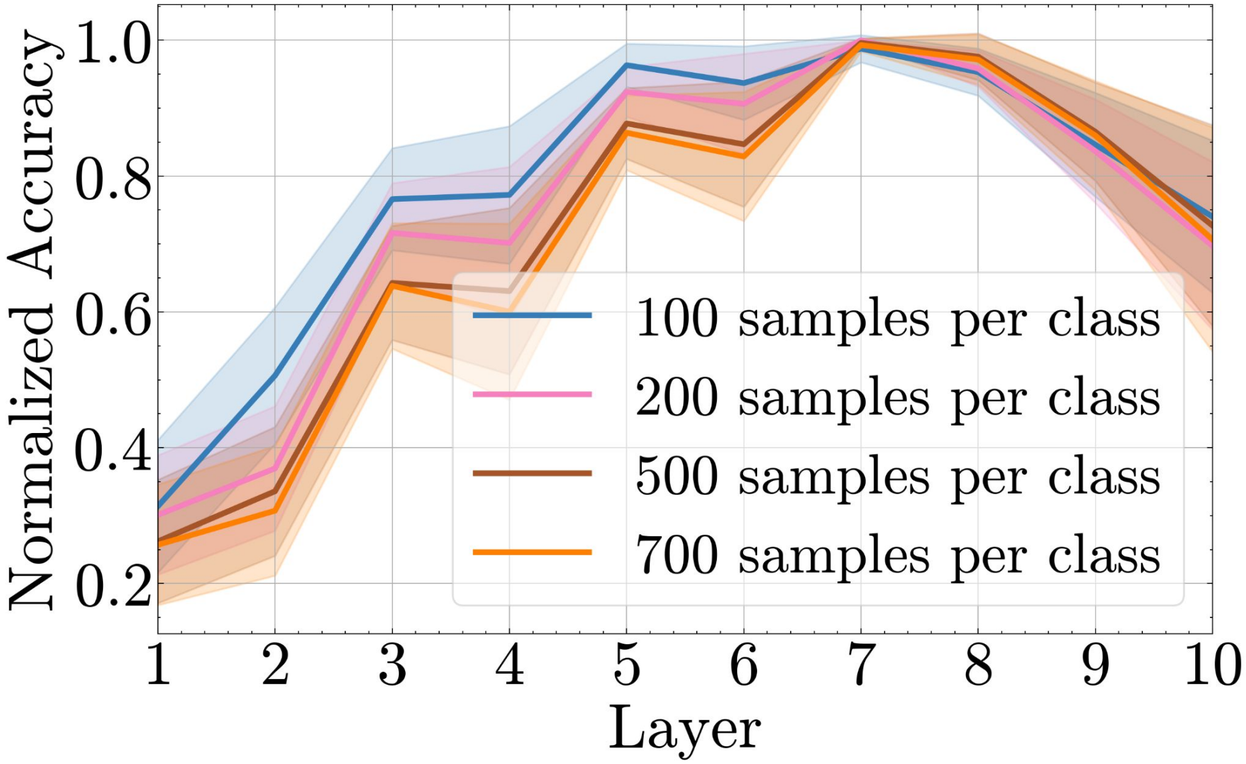
Data Quantity VS Semantic Variability

- **Fixed Sample size (10K) and varied class counts**: increasing class counts decreases the tunnel strength
- **Fixed class counts (100) and varied sample size**: minimal impact

Varied Class Counts



Varied Sample Size

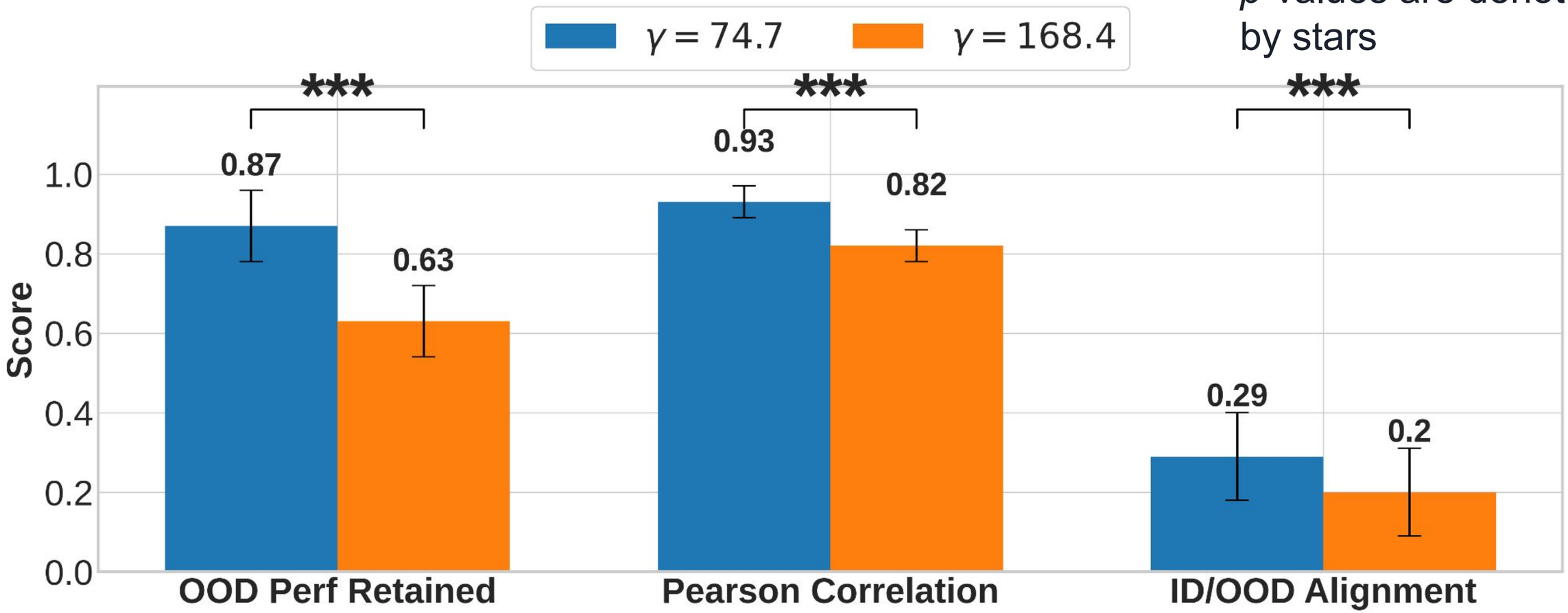


Takeaways: This observation challenges the common wisdom that *more data helps*. Wider coverage of semantics matter more than data quantity.

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

Impact of Overparameterization

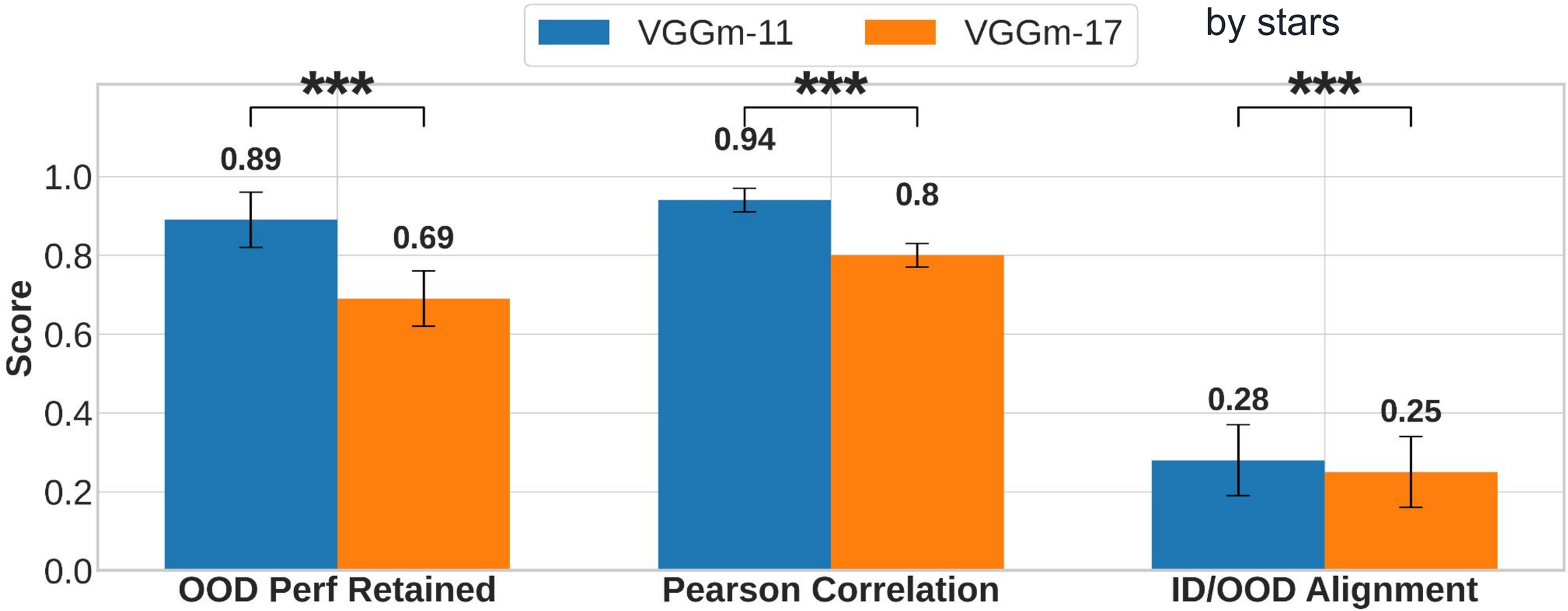
p-values are denoted by stars



- **Increasing overparameterization (γ) impairs OOD generalization across all metrics.**
- p-values are based on Wilcoxon signed-rank tests. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$
- In terms of mean effect size (Cliff's delta), OOD perf retained, pearson correlation, and ID/OOD alignment achieve large, large, and medium effect size, respectively.

Impact of Depth

p-values are denoted by stars



- **Increasing DNN depth hurts OOD generalization across all metrics.**
- The p-values are based on Wilcoxon signed-rank tests. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$
- In terms of mean effect size (Cliff's delta), OOD perf retained, pearson correlation, and ID/OOD alignment achieve large, large, and small effect size, respectively.

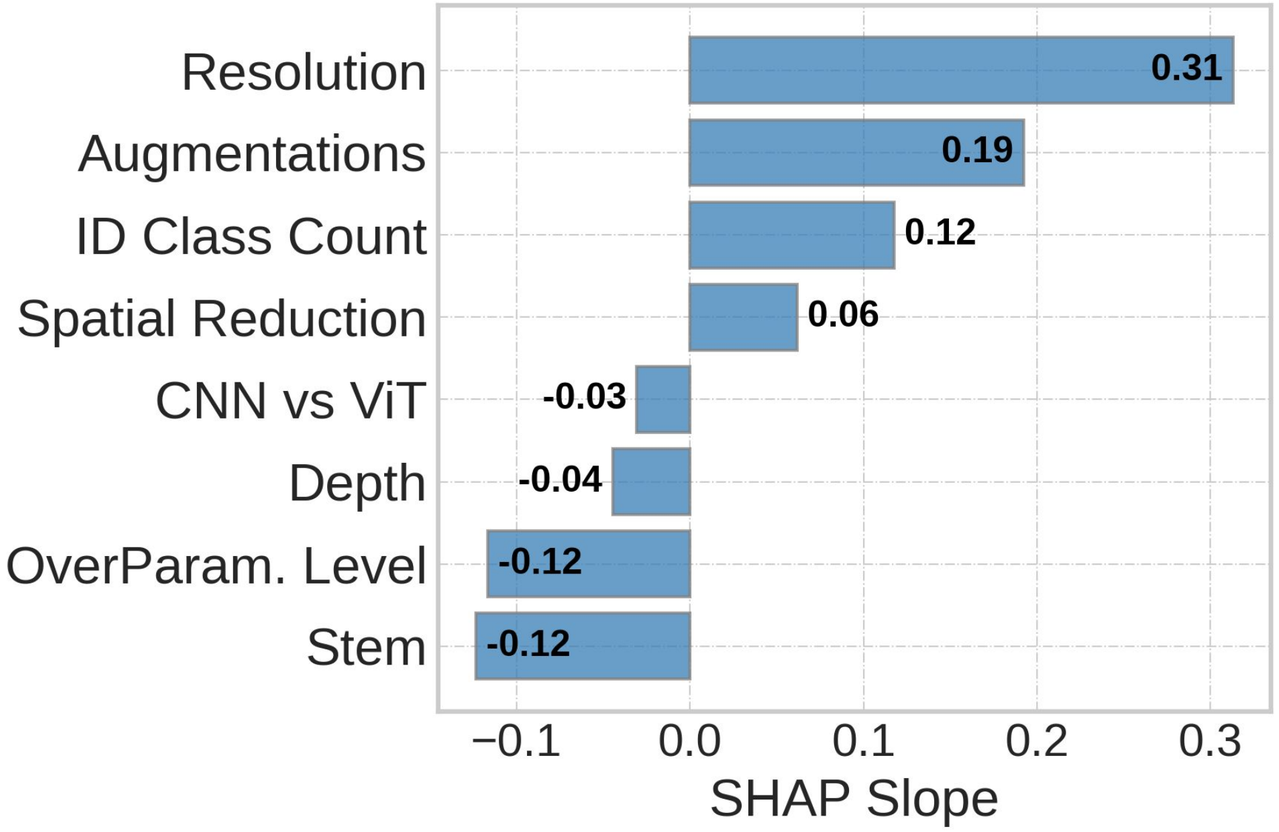
What We Observed

- ❏ Most variables except ViT vs CNN had an impact on tunnel strength
- ❏ Large stem size impairs OOD transfer across all metrics
- ❏ Decreasing spatial reduction ratio impairs OOD transfer across all metrics
- ❏ The tunnel effect is not universal and its strength varies. Among 64 ID backbones, 4 did not exhibit any tunnel effect.

Now we explore what variable matters most using SHAP analysis

SHAP Analysis

- We propose SHAP Slope, a novel SHAP-based analysis to disentangle the contribution of eight variables to three targets (our 3 metrics).
- Our SHAP Slope indicates both magnitude and direction of impact.



ID/OOD alignment as target

Key Findings:

- ❏ Resolution was found most dominant followed by augmentations and ID class counts in terms of reducing the tunnel strength
- ❏ DNN variables e.g., overparameterization, stem, and depth increase the tunnel strength but their impact is less than dataset variables.

Summary:

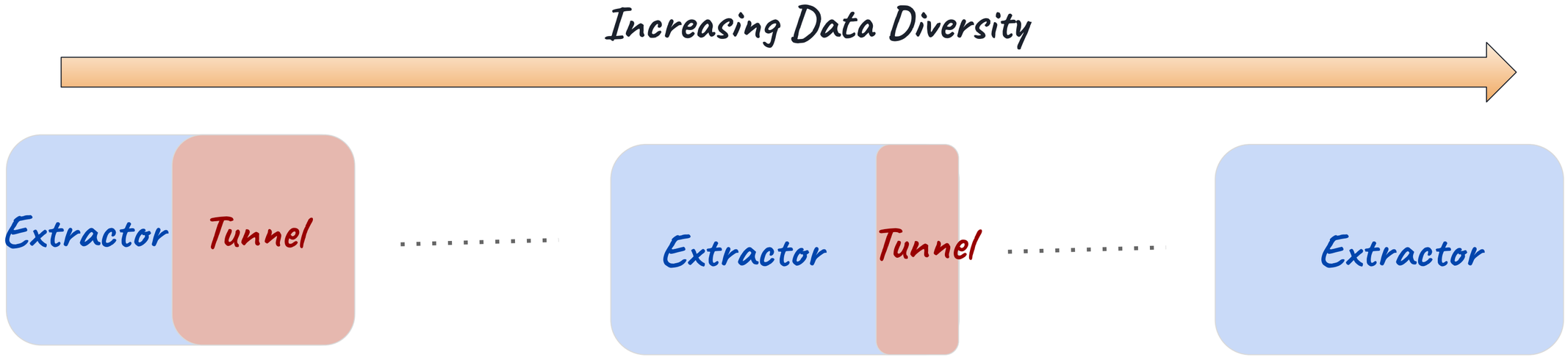
- It is evident that dataset variables e.g., image resolution, ID class counts, and augmentations show dominance in altering the tunnel effect
- Increasing ID class counts (*between-class diversity*), using more augmentations (*within-class diversity*), and using higher image resolution (*hierarchical features*) reduce the tunnel effect and improve OOD transfer.
- DNN variables e.g., over-parameterization, depth etc. increase the tunnel effect but their impact is less compared to the 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.

Revised Tunnel Effect Hypothesis

An *overparameterized* N - layer DNN forms 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).



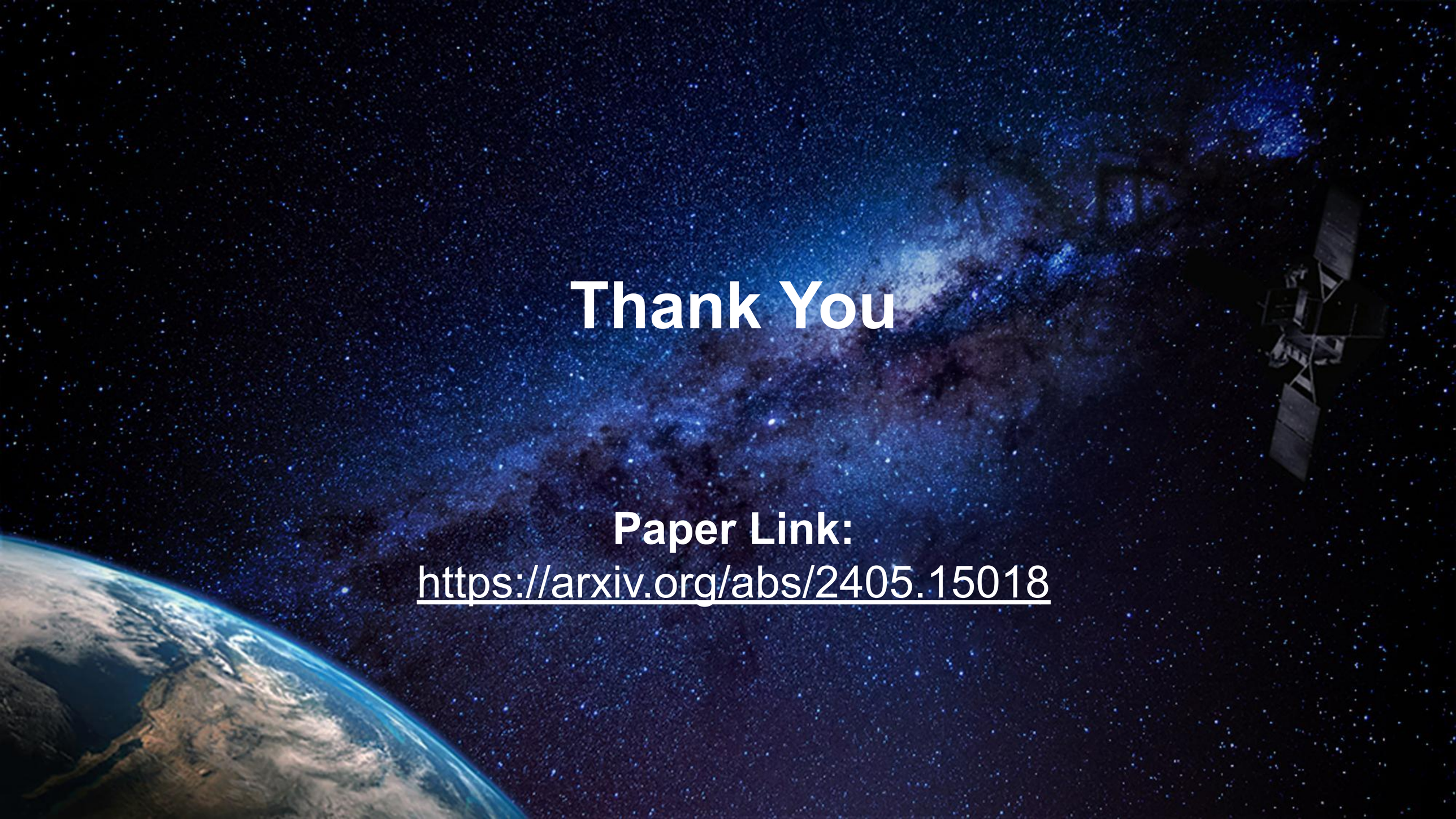
Conclusion & Acknowledgements

- We disentangled the causes of the tunnel effect and showed how its strength varies
- Our findings and insights can inform future research to build better models

Future directions:

- Theoretical framework
- Other modalities beyond vision e.g., language, multimodal etc.
- Self-supervised learning paradigm
- Regularization and architectural approaches to control the tunnel effect

Acknowledgments: We thank NSF for supporting our research


Thank You
Paper Link:
https://arxiv.org/abs/2405.15018