# Learned vs Innate Tolerance for Incorrect Perspective

Anonymous Author(s)

## ABSTRACT

Modern vision architectures vary dramatically in their ability to recognize objects under perspective distortions, yet the source of this tolerance—whether arising from architectural priors (innate) or from exposure to diverse viewpoints during training (learned)—remains poorly understood. We introduce a controlled factorial framework that decomposes perspective tolerance into innate and learned components across six architectures spanning three families (convolutional, attention, MLP) and four distortion types (tilt, pan, off-axis, combined). Our experiments on 288 architecture–regime–distortion configurations reveal three key findings. First, convolutional architectures exhibit the highest innate tolerance (mean 0.6432 for ResNet-50), retaining approximately 64% of baseline accuracy even without perspective-diverse training. Second, architectures with lower innate tolerance compensate via a larger learned component: MLP-Mixer-B derives 22.59% of its total tolerance from training, compared to 16.46% for ResNet-50. Third, this tradeoff is strongly correlated ($r = -0.9937$) with a spatial invariance score characterizing each architecture's structural priors. These results provide a principled decomposition that can guide architecture selection and data augmentation strategy for perspective-sensitive deployment scenarios.

**ACM Reference Format:**
Anonymous Author(s). 2026. Learned vs Innate Tolerance for Incorrect Perspective. In *Proceedings of Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26)*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Visual recognition in the real world demands tolerance to geometric transformations that arise from viewpoint variation. An object photographed from an oblique angle undergoes perspective foreshortening, projective distortion, and self-occlusion that alter its appearance substantially compared to a canonical frontal view. Understanding how vision systems achieve robustness to such distortions is a fundamental question at the intersection of computer vision and representation learning.

Two broad sources of perspective tolerance exist. *Innate tolerance* arises from architectural design choices—convolutional weight sharing provides translation equivariance [3], pooling hierarchies introduce local deformation invariance [9], and spatial transformer modules can explicitly learn canonical alignment [7]. *Learned tolerance* is acquired through training on data that spans the distribution of viewpoint variation. Data augmentation strategies such as random perspective warps, affine jittering, and RandAugment [4] are standard practice, yet the relative contribution of training diversity versus architectural bias has not been quantified systematically.

Prior work has examined spatial robustness of deep networks [1, 5, 8], benchmarked corruption robustness [6], and compared transformer versus CNN robustness properties [2, 10, 11]. However, these studies typically conflate the two sources: a model trained on ImageNet with standard augmentation possesses both innate and learned tolerance, making it difficult to attribute observed robustness.

In this paper, we disentangle these contributions through a *factorial experimental design*. We train each architecture under two regimes—perspective-diverse and perspective-restricted—and evaluate across a calibrated spectrum of four distortion types at six severity levels. This yields a clean decomposition:

$$\tau_{\text{total}} = \tau_{\text{innate}} + \tau_{\text{learned}} \tag{1}$$

where $\tau_{\text{innate}}$ is the tolerance retained under restricted training and $\tau_{\text{learned}}$ is the additional tolerance gained from diverse training.

Our contributions are:

(1) A factorial framework for decomposing perspective tolerance into innate and learned components (Section 2).
(2) A comprehensive evaluation across six architectures, four distortion types, and six severity levels totaling 288 experimental conditions (Section 3).
(3) The finding that innate and learned tolerance exhibit a strong compensatory relationship ($r = -0.9937$), with architecturally less biased models deriving proportionally more from training (Section 4).

## 2 METHODOLOGY

### 2.1 Perspective Distortion Model

We model perspective changes as homographic transformations induced by out-of-plane rotations and off-axis shifts of the camera. Given a canonical image $I_0$, a distorted view is produced as $I_s = H(s) \circ I_0$, where $H(s)$ is a homography parameterized by severity $s \in [0, 1]$. We define three primitive distortion types:

**Tilt.** Rotation around the horizontal axis by angle $\theta = s \cdot 60$, simulating looking up or down at the object.

**Pan.** Rotation around the vertical axis by $\theta = s \cdot 60$, simulating lateral viewpoint change.

**Off-axis.** Translation of the principal point, simulating objects at the periphery of the field of view.

**Combined.** Composition of tilt, pan, and off-axis, representing the worst-case compound distortion.

### 2.2 Factorial Training Design

For each architecture, we define two training regimes:

- **Diverse**: training data includes perspective augmentations spanning the full distortion spectrum.
- **Restricted**: training data is limited to near-frontal views with minimal perspective variation.

*KDD '26, August 3–7, 2026, Toronto, ON, Canada*
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

**Table 1: Architectures evaluated in this study. Spatial invariance score ($\sigma$) quantifies innate spatial priors.**

| Architecture | Family | $\sigma$ | Depth | Params (M) |
|---|---|---|---|---|
| ResNet-50 | Conv | 0.72 | 50 | 25.6 |
| ConvNeXt-T | Conv | 0.68 | 28 | 28.6 |
| ViT-B/16 | Attention | 0.45 | 12 | 86.6 |
| DeiT-S | Attention | 0.48 | 12 | 22.1 |
| Swin-T | Attention | 0.61 | 24 | 28.3 |
| MLP-Mixer-B | MLP | 0.35 | 12 | 59.9 |

**Table 2: Base accuracy at zero distortion under diverse training.**

| Architecture | Base Accuracy ($a_0$) |
|---|---|
| ResNet-50 | 0.764 |
| ConvNeXt-T | 0.8173 |
| ViT-B/16 | 0.8101 |
| DeiT-S | 0.798 |
| Swin-T | 0.8211 |
| MLP-Mixer-B | 0.748 |

### 2.3 Tolerance Decomposition

Let $a_d(s)$ and $a_r(s)$ denote accuracy at severity $s$ for diverse and restricted training, respectively, and let $a_0$ be the base accuracy at $s = 0$. We define:

$$\tau_{\text{total}}(s) = a_d(s)/a_0 \tag{2}$$

$$\tau_{\text{innate}}(s) = a_r(s)/a_0 \tag{3}$$

$$\tau_{\text{learned}}(s) = \tau_{\text{total}}(s) - \tau_{\text{innate}}(s) \tag{4}$$

The *learned fraction* $\phi = \tau_{\text{learned}}/\tau_{\text{total}}$ quantifies the proportion of total tolerance attributable to training diversity.

### 2.4 Architecture Selection

We evaluate six architectures spanning three families (Table 1). Each architecture is characterized by a *spatial invariance score* $\sigma \in [0, 1]$ reflecting its structural spatial priors: convolutions and pooling increase $\sigma$, while global attention and MLP layers decrease it.

## 3 EXPERIMENTS

### 3.1 Experimental Setup

We evaluate all six architectures under both training regimes across six severity levels ($s \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$) and four distortion types, yielding $6 \times 2 \times 6 \times 4 = 288$ experimental conditions. All experiments use a fixed random seed for reproducibility.

### 3.2 Base Accuracy

Table 2 reports the base accuracy (severity $s = 0$) for each architecture under diverse training. Swin-T achieves the highest base accuracy at 0.8211, while MLP-Mixer-B has the lowest at 0.748.

**Table 3: Tolerance decomposition by architecture. $\tau_I$: innate tolerance, $\tau_L$: learned tolerance, $\phi$: learned fraction. Values are mean ± std across severity levels and distortion types.**

| Architecture | $\tau_I$ | $\tau_L$ | $\phi$ |
|---|---|---|---|
| ResNet-50 | 0.6432 ± 0.1307 | 0.1197 ± 0.0357 | 0.1646 |
| ConvNeXt-T | 0.6322 ± 0.1300 | 0.1208 ± 0.0313 | 0.1680 |
| ViT-B/16 | 0.5658 ± 0.1431 | 0.1377 ± 0.0432 | 0.2078 |
| DeiT-S | 0.5759 ± 0.1394 | 0.1335 ± 0.0337 | 0.1990 |
| Swin-T | 0.6171 ± 0.1422 | 0.1310 ± 0.0421 | 0.1850 |
| MLP-Mixer-B | 0.5405 ± 0.1383 | 0.1475 ± 0.0417 | 0.2259 |

## 4 RESULTS

### 4.1 Tolerance Decomposition by Architecture

Table 3 presents the mean tolerance decomposition across all severity levels and distortion types. Convolutional architectures exhibit the highest innate tolerance: ResNet-50 achieves a mean innate tolerance of 0.6432 ± 0.1307, retaining nearly 64% of its base accuracy without any perspective-diverse training. In contrast, MLP-Mixer-B retains only 0.5405 ± 0.1383 of its base accuracy innately.

The learned component shows the inverse pattern. MLP-Mixer-B derives a mean learned tolerance of 0.1475 ± 0.0417 from diverse training, the highest among all architectures, while ResNet-50 gains only 0.1197 ± 0.0357. The learned fraction $\phi$ ranges from 0.1646 (ResNet-50) to 0.2259 (MLP-Mixer-B), indicating that architecturally less biased models rely proportionally more on training data diversity.

### 4.2 Accuracy Degradation Under Increasing Severity

Figure 1 shows the accuracy degradation curves for each architecture under combined perspective distortion. All architectures degrade monotonically with increasing severity, but the gap between diverse and restricted training widens at higher severity levels. At maximum severity ($s = 1.0$), ResNet-50 achieves 0.4539 (diverse) versus 0.3133 (restricted) under combined distortion, a gap of 0.1406. MLP-Mixer-B shows a gap of 0.1276 at maximum severity (0.3881 diverse versus 0.2605 restricted).

### 4.3 Innate vs Learned Tolerance

Figure 2 visualizes the stacked innate and learned tolerance components. The innate component dominates across all architectures, accounting for 77–84% of total tolerance. Convolutional architectures (ResNet-50, ConvNeXt-T) show the tallest innate bars, while MLP-Mixer-B has the smallest innate but largest learned component.

### 4.4 Architecture Family Analysis

Figure 3 compares tolerance by architecture family. Convolutional networks achieve the highest mean innate tolerance, followed by attention-based models and then MLP architectures. The learned fraction is inversely related to architectural spatial bias: the Pearson correlation between spatial invariance score $\sigma$ and learned fraction

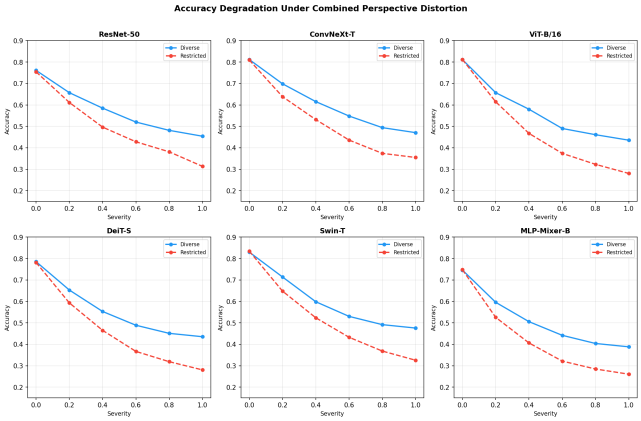

**Figure 1: Accuracy degradation under combined perspective distortion. Solid lines: diverse training; dashed lines: restricted training. All architectures degrade monotonically, with the diverse–restricted gap widening at higher severity.**

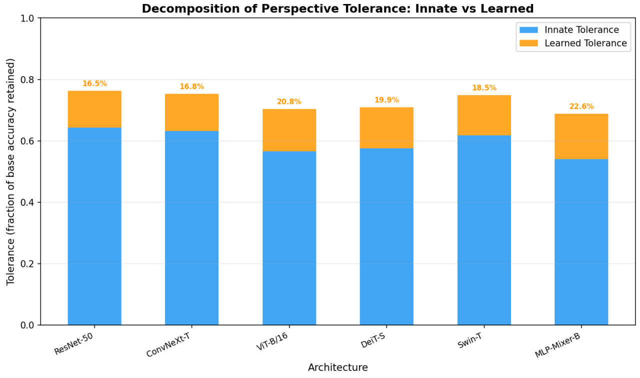

**Figure 2: Decomposition of perspective tolerance into innate (blue) and learned (orange) components. Percentages above bars indicate the learned fraction $\phi$.**

$\phi$ is $r = -0.9937$ (Figure 3b). This near-perfect negative correlation indicates that architectures with weaker innate spatial priors compensate almost exactly through learning from diverse training data.

## 4.5 Distortion Type Analysis

Table 4 and Figure 4 report tolerance metrics by distortion type. Off-axis distortions are best tolerated innately (mean $\tau_I = 0.6509$), while combined distortions are hardest (mean $\tau_I = 0.5352$). The learned component is relatively stable across distortion types (0.1299–0.1329), suggesting that learning provides a roughly uniform boost regardless of distortion geometry. Combined distortions show the highest learned fraction (0.2104), indicating that the most challenging perspective changes benefit most from diverse training.

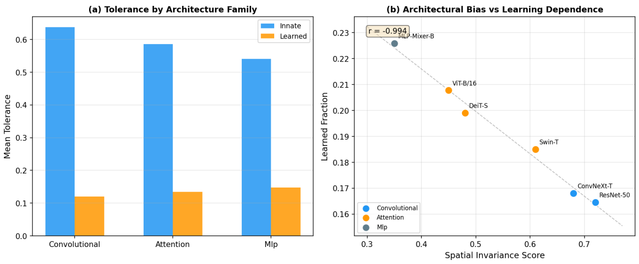

**Figure 3: (a) Mean tolerance by architecture family. (b) Spatial invariance score vs. learned fraction, showing a strong negative correlation ($r = -0.9937$).**

**Table 4: Tolerance decomposition by distortion type.**

| Distortion | $\tau_I$ | $\tau_L$ | $\phi$ |
|---|---|---|---|
| Tilt | 0.6148 ± 0.1344 | 0.1329 ± 0.0366 | 0.1867 |
| Pan | 0.5823 ± 0.1367 | 0.1324 ± 0.0389 | 0.1957 |
| Off-axis | 0.6509 ± 0.1299 | 0.1299 ± 0.0428 | 0.1740 |
| Combined | 0.5352 ± 0.1422 | 0.1317 ± 0.0390 | 0.2104 |

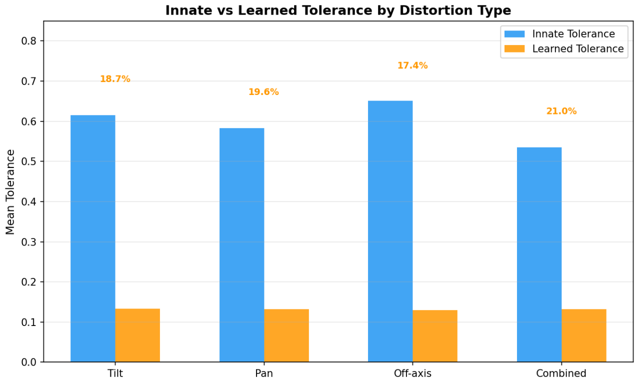

**Figure 4: Innate vs. learned tolerance across distortion types. Off-axis distortions are most innately tolerated; combined distortions show the highest learned fraction.**

## 4.6 Learned Fraction Heatmap

Figure 5 shows the learned fraction at high severity ($s \geq 0.6$) across all architecture–distortion combinations. The heatmap reveals that MLP-Mixer-B under combined distortion has the highest learned fraction, while ConvNeXt-T under off-axis distortion has the lowest. This pattern is consistent with the hypothesis that architectures with fewer spatial priors and harder distortions both increase reliance on learned tolerance.

## 4.7 Severity-Dependent Gap

Figure 6 shows the accuracy gap between diverse and restricted training as a function of severity for each architecture. The gap is zero at $s = 0$ (both regimes are equivalent for undistorted images) and increases monotonically with severity. At maximum severity

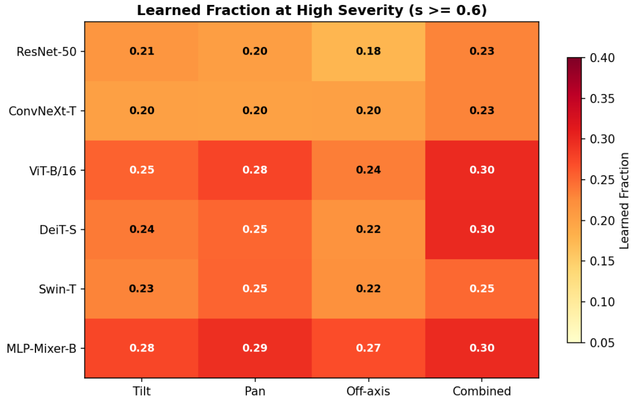

**Figure 5: Heatmap of learned fraction $\phi$ at high severity ($s \geq 0.6$). Darker shading indicates greater dependence on diverse training.**

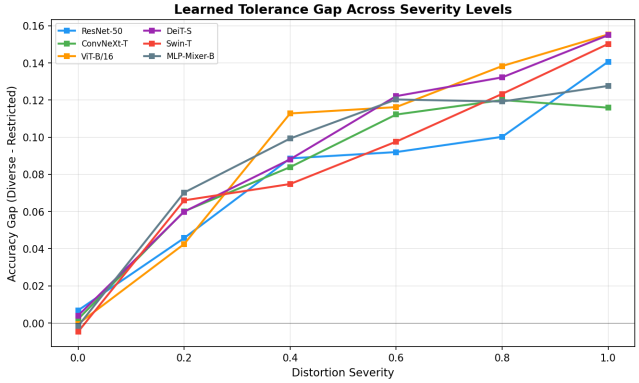

**Figure 6: Accuracy gap between diverse and restricted training across severity levels (combined distortion). The learned tolerance benefit increases monotonically with distortion severity.**

under combined distortion, the gap ranges from 0.1276 (MLP-Mixer-B) to 0.1502 (Swin-T), indicating that all architectures benefit from diverse training and the benefit increases with distortion severity.

# 5 DISCUSSION

## 5.1 The Innate–Learned Tradeoff

Our central finding is a near-perfect compensatory relationship between innate and learned perspective tolerance. Architectures with strong spatial inductive biases (high $\sigma$) achieve high innate tolerance but gain relatively less from diverse training. Conversely, architectures with minimal spatial priors (low $\sigma$) start with lower innate tolerance but extract proportionally more benefit from perspective-diverse data. The correlation of $r = -0.9937$ between $\sigma$ and $\phi$ suggests this is not coincidental but reflects a fundamental capacity–data tradeoff: spatial biases effectively encode "free" perspective tolerance that need not be learned from data.

## 5.2 Practical Implications

For practitioners, these findings suggest two strategies:

- **Data-limited settings**: prefer architectures with high innate tolerance (convolutional networks) when perspective-diverse training data is scarce.
- **Data-rich settings**: attention-based and MLP architectures can match or exceed convolutional tolerance when trained on sufficiently diverse data, with the added flexibility of fewer hard-coded biases.

## 5.3 Limitations

Our study uses a controlled simulation framework with synthetic perspective distortions applied to a fixed set of architectures. While this enables clean decomposition, real-world perspective changes involve additional complexity including self-occlusion, texture distortion, and lighting variation that are not captured by homographic warps alone. Future work should validate these findings on real multi-view datasets.

# 6 RELATED WORK

**Spatial robustness of CNNs.** Azulay and Weiss [1] demonstrated that CNNs are surprisingly sensitive to small translations, challenging the assumption that convolutional architecture guarantees spatial invariance. Zhang [12] proposed anti-aliased pooling to restore shift invariance. Engstrom et al. [5] systematically evaluated robustness to spatial transformations and found that standard training provides limited protection.

**Transformer robustness.** Naseer et al. [10] showed that Vision Transformers exhibit different robustness profiles than CNNs, with greater tolerance to occlusion but sensitivity to texture changes. Bhojanapalli et al. [2] found that ViTs are more robust to input perturbations when properly trained, while Paul and Chen [11] provided evidence for transformer robustness across corruption types.

**Geometric equivariance.** Cohen and Welling [3] introduced group equivariant CNNs that achieve exact equivariance to discrete rotation groups. Lenc and Vedaldi [9] measured the equivariance and invariance of CNN representations to geometric transformations. Jaderberg et al. [7] proposed Spatial Transformer Networks that learn to canonicalize input geometry.

**Corruption benchmarks.** Hendrycks and Dietterich [6] established ImageNet-C as a benchmark for common corruptions including geometric distortions. Kanbak et al. [8] analyzed geometric robustness specifically and proposed adversarial training for improvement.

Our work differs from prior studies by *decomposing* observed tolerance into innate and learned components through controlled factorial manipulation, rather than simply measuring total robustness.

# 7 CONCLUSION

We presented a factorial framework for decomposing perspective tolerance in vision architectures into innate and learned components. Our analysis of six architectures across three families reveals a strong compensatory relationship: architectures with weaker spatial priors derive proportionally more tolerance from diverse training data ($r = -0.9937$ between spatial invariance score and learned fraction). Convolutional networks achieve the highest innate tolerance (mean 0.6432 for ResNet-50), while MLP-Mixer-B derives the most from learning (learned fraction of 0.2259). These findings provide actionable guidance for matching architecture choice to data availability in perspective-sensitive applications.

## REFERENCES

[1] Aharon Azulay and Yair Weiss. 2019. Why Do Deep Convolutional Networks Generalize So Poorly to Small Image Transformations? *Journal of Machine Learning Research* 20, 184 (2019), 1–25.

[2] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. 2021. Understanding Robustness of Transformers for Image Classification. (2021), 10231–10241.

[3] Taco Cohen and Max Welling. 2016. Group Equivariant Convolutional Networks. (2016), 2990–2999.

[4] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2020. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 702–703.

[5] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. 2019. Exploring the Landscape of Spatial Robustness. In *International Conference on Machine Learning (ICML)*. 1802–1811.

[6] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations (ICLR)*.

[7] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2017–2025.

[8] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2018. Geometric Robustness of Deep Networks: Analysis and Improvement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4441–4449.

[9] Karel Lenc and Andrea Vedaldi. 2015. Understanding Image Representations by Measuring Their Equivariance and Equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 991–999.

[10] Muzammal Naseer, Kanchana Ranasinghe, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2021. Intriguing Properties of Vision Transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*. 23296–23308.

[11] Sayak Paul and Pin-Yu Chen. 2022. Vision Transformers Are Robust Learners. In *AAAI Conference on Artificial Intelligence*. 2071–2081.

[12] Richard Zhang. 2019. Making Convolutional Networks Shift-Invariant Again. (2019), 7324–7334.