SEIS: SUBSPACE-BASED EQUIVARIANCE AND INVARIANCE SCORES FOR NEURAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how neural representations respond to geometric transformations is important for diagnosing robustness, comparing architectures, and selecting models for spatial tasks. Existing measures typically examine output predictions, aggregate individual units independently, or fit an explicit transformation between representations. We introduce SEIS (*Subspace-based Equivariance and Invariance Scores*), a label-free subspace measure that disentangles equivariance and invariance directly from paired activations of original and transformed inputs, without knowledge of the transformation. SEIS organizes activations into spatial positions and observations, applies a continuous spectral filter that softly downweights weak components, then factorizes the comparison into an observation-space alignment for equivariance and a stricter spatial alignment for invariance, with both scores chance-corrected against null expectations. Controlled experiments show that SEIS cleanly separates the two properties and is stable given sufficient observations. Applied to trained networks, SEIS reveals how representations evolve during training, how augmentation strengthens deeper layers, how decoder architecture and multi-task supervision shape transformation geometry, and how positional encodings affect translation structure in vision transformers. Across a diverse pool of ImageNet-pretrained classifiers, SEIS predicts accuracy retention under affine transformation better than validation accuracy, without labels.

1 INTRODUCTION

The remarkable success of deep learning across computer vision (He et al., 2016), physics simulation (Sanchez-Gonzalez et al., 2020), and molecular biology (Jumper et al., 2021) is largely attributed to the effective utilization of inductive biases (Bronstein et al., 2017). A representation is invariant when its response remains unchanged under a transformation, whereas an equivariant representation changes according to a corresponding action in feature space. Convolutional Neural Networks (CNNs) (LeCun et al., 1989) achieve translation equivariance through spatial weight sharing, while Group-equivariant CNNs (G-CNNs) (Cohen & Welling, 2016) extend this principle to rotations and reflections by sharing filters over a specified transformation group, and Graph Neural Networks (GNNs) (Xu et al., 2019) encode permutation symmetries as architectural constraints. However, many modern architectures, such as Vision Transformers (ViTs) (Dosovitskiy et al., 2021), do not provide explicit geometric guarantees. Consequently, determining whether and at which layers a model learns invariant or equivariant representations remains an empirical question requiring explicit measurement.

Existing measures capture only part of this picture. Output-level metrics gauge how much predictions change under a transformation (Azulay & Weiss, 2019; Hendrycks & Dietterich, 2019; Deng et al., 2022; Kvinge et al., 2022), but they are blind to the mechanism behind that stability and reveal nothing about how it develops across depth. Unit-wise statistics do examine internal activations (Goodfellow et al., 2009; Quiroga et al., 2023; Bruintjes et al., 2023), yet scoring neurons or channels in isolation overlooks structure spread across many units (Kornblith et al., 2019). Learned latent maps (Lenc & Vedaldi, 2015) address this, but at the cost of fitting a transformation-specific operator. Subspace-based similarity measures (Raghu et al., 2017; Kornblith et al., 2019) need no assumption about the transformation, but they ignore the spatial layout on which such transformations act.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we address this gap by introducing *Subspace-based Equivariance and Invariance Scores (SEIS)*. Our key idea is to arrange paired activations as matrices with spatial positions as rows and observations as columns. A spatial transformation then acts on the row space, and the two properties correspond to two different levels of agreement between the original and transformed representations. A continuous spectral filter weights reliable directions without requiring a manually selected rank, explained-variance threshold, or regularization coefficient. SEIS then measures observation-space agreement as evidence of equivariance and additionally requires spatial alignment for invariance. Finally, SEIS corrects both signals for chance using derived null expectations.

Our main contributions are as follows:

- We introduce the first subspace-based measure that separately quantifies invariance and equivariance in internal neural representations without requiring labels, an explicitly specified latent transformation action, or a fitted transformation-specific mapping.
- We develop data-adaptive spectral weighting that avoids manually selected ranks, explained-variance thresholds, and regularization strengths, together with analytical chance correction that removes incidental alignment between unrelated representations.
- We conduct an extensive empirical analysis of how transformation structure evolves across training and network depth, and how it is shaped by data augmentation, decoder architecture, task supervision, and positional encoding across classification and segmentation models.
- We demonstrate that SEIS equivariance provides a label-free indicator of performance retention under affine transformations across a diverse pool of ImageNet-pretrained classifiers.

2 Related Work

Quantifying neural-network responses to geometric transformations is crucial for determining whether learned representations exhibit meaningful structure. Modern architectures do not generally guarantee invariance or equivariance, as these properties can be disrupted by discretization and operations on finite spatial grids (Azulay & Weiss, 2019), while imposed geometric biases may be ineffective when their symmetry assumptions do not align with the data distribution (Benton et al., 2020).

Prediction Stability. A common approach to measuring invariance is quantifying the stability of the final model predictions under input transformations. Metrics such as Accuracy Drop (Azulay & Weiss, 2019) and Relative Corruption Errors (Hendrycks & Dietterich, 2019) measure performance degradation under geometric shifts. To reduce dependence on ground-truth labels, Deng et al. (2022) proposed Effective Invariance, which weighs stability by prediction confidence. Additionally, Softmax G-Empirical Equivariance Deviation (G-EED) (Kvinge et al., 2022) assesses global deviations using the Kullback–Leibler divergence between predictions and the group orbit average. While these methods provide an intuitive gauge of model invariance, they operate solely at the output level and treat the internal feature extraction as a black box. As a result, they cannot distinguish whether stability arises from robust semantic representations or shortcut learning via spurious correlations (Geirhos et al., 2020).

Element-Wise Feature Statistics. To probe internal representations, significant effort has been directed toward analyzing the statistics of individual neurons or channels. Prior work has introduced statistical measures that compare a single neuron's response stability against its global firing profile (Goodfellow et al., 2009; Quiroga et al., 2023). More recently, equivariance has been quantified by computing cosine similarity and Pearson correlation between transformed channel activations and the channel activations of transformed inputs (Kvinge et al., 2022; Bruintjes et al., 2023). These methods typically aggregate scores by averaging across neurons or channels, implicitly assuming that units operate independently. However, deep neural networks rely on distributed representations where geometric information is encoded in the joint activation space rather than any single direction (Kornblith et al., 2019).

Latent Transformation Modeling. A more rigorous line of work seeks to model how transformations act within latent representations. Lenc & Vedaldi (2015) measure equivariance by learning a

linear mapping between representations of original and transformed inputs, and identify invariant features as those that are mapped approximately to themselves. However, invariance is treated as a special case of equivariance at the level of individual feature channels, rather than as a distinct property of the representation. As a result, this approach does not provide a unified or disentangled characterization of equivariance and invariance.

Representation Similarity. Subspace-based methods have been widely used to compare neural representations across models and layers by analyzing the eigenvectors or singular values of covariance or Gram matrices. CCA (Hotelling, 1992) and its variants, including SVCCA (Raghu et al., 2017) and CKA (Kornblith et al., 2019), capture similarity up to invertible linear transformations, while Procrustes alignment (Ding et al., 2021) enforces orthogonal correspondence. These methods have not been used to directly assess equivariance or invariance within a representation under input transformations. They treat spatial locations as independent samples when constructing similarity measures, thereby discarding the underlying grid structure. However, their reliance on intrinsic subspace structure rather than explicit transformation priors suggests that they could, in principle, be adapted to analyze geometric stability even when the form of the transformation is unknown.

3 METHOD

3.1 MATRICIZATION: SPATIAL POSITIONS OVER OBSERVATIONS

Our first step exposes the spatial axis on which a transformation acts. Given paired activations, we form $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{p \times n}$, where p is the number of spatial positions and n is the number of observations. For CNN activations in $\mathbb{R}^{B \times C \times H \times W}$, B, C, H and W denote the batch size, number of channels, height and width, respectively; we therefore set $p = HW$ and $n = BC$. For transformer activations in $\mathbb{R}^{B \times L \times D}$, L and D denote the number of spatial tokens and embedding dimensions, respectively; we set $p = L$ and $n = BD$.

This orientation expresses a spatial transformation as $\mathbf{Y} \approx \mathbf{T}_g \mathbf{X}$, where $\mathbf{T}_g \in \mathbb{R}^{p \times p}$. Invariance corresponds to $\mathbf{T}_g = \mathbf{I}$, whereas an invertible spatial action preserves the subspace spanned by the spatial activation profiles across observations. The matricization therefore retains spatial correspondence for measuring invariance and shared observation-space structure for measuring equivariance, without removing information about how the activations are spatially reorganized.

3.2 CONTINUOUS SPECTRAL REPRESENTATION

We center each spatial position across observations and normalize each matrix to unit Frobenius norm, yielding $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$. For either $\widetilde{\mathbf{M}} \in \{\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}\}$, let

$$\widetilde{\mathbf{M}} = \mathbf{U} \operatorname{diag}(\mathbf{s}) \mathbf{V}^\top, \qquad \mathbf{U} \in \mathbb{R}^{p \times q}, \quad \mathbf{V} \in \mathbb{R}^{n \times q}, \tag{1}$$

where $\mathbf{s} = (s_1, \ldots, s_q)$ contains the singular values in descending order. Since centering removes the constant observation direction, the effective observation dimension is $d = n - 1$, and $q = \min(p, d)$.

Rather than choosing the effective rank by hard-thresholding singular values (Gavish & Donoho, 2014) or retaining a fixed fraction of explained variance as in SVCCA (Raghu et al., 2017), we assign each spectral component j the continuous trust weight

$$w_j = \frac{s_j^2}{s_j^2 + \rho}, \qquad \rho = \sqrt{s_1^2 \operatorname{median}_j(s_j^2)}. \tag{2}$$

This weight has the form of a classical Tikhonov spectral filter (Engl et al., 1996). The data-derived ridge ρ is the geometric mean of the largest spectral power and a median-spectrum proxy for the typical noise level.

The filtered representation is therefore

$$\mathbf{Z} = \mathbf{U} \operatorname{diag}(\mathbf{w}) \mathbf{V}^\top. \tag{3}$$

Applying the construction independently produces $\mathbf{Z}_X$ and $\mathbf{Z}_Y$, smoothly suppressing weak directions without a user-selected rank or regularization parameter.

3.3 FACTORIZED SPATIAL AND OBSERVATION ALIGNMENT

The filtered representations $\mathbf{Z}_X = \mathbf{U}_X \operatorname{diag}(\mathbf{w}_X)\mathbf{V}_X^\top$ and $\mathbf{Z}_Y = \mathbf{U}_Y \operatorname{diag}(\mathbf{w}_Y)\mathbf{V}_Y^\top$ separate each spectral component into a spatial pattern and an observation pattern. Their weighted observation-space agreement is collected in

$$\mathbf{C} = \operatorname{diag}(\mathbf{w}_X)\mathbf{V}_X^\top \mathbf{V}_Y \operatorname{diag}(\mathbf{w}_Y), \tag{4}$$

where each entry compares one component from each representation and downweights weak components on either side. The observation-alignment signal sums the squared similarities between all trusted observation-pattern pairs

$$\mathcal{S}_{\text{obs}} = \|\mathbf{C}\|_F^2 = \left\langle \mathbf{Z}_X^\top \mathbf{Z}_X, \mathbf{Z}_Y^\top \mathbf{Z}_Y \right\rangle_F . \tag{5}$$

This relaxed signal captures observation-space structure that can be preserved under an unknown spatial action and therefore forms the basis of the *equivariance* score.

Invariance additionally requires the associated spatial patterns to agree. Their pairwise alignment is $\mathbf{U}_X^\top \mathbf{U}_Y$, yielding

$$\mathcal{S}_{\text{sp}} = \left\langle \mathbf{C}, \mathbf{U}_X^\top \mathbf{U}_Y \right\rangle_F = \left\langle \mathbf{Z}_X, \mathbf{Z}_Y \right\rangle_F . \tag{6}$$

Thus, the same weighted component correspondences support two levels of comparison: $\mathcal{S}_{\text{obs}}$ preserves evidence under spatial reorganization, while $\mathcal{S}_{\text{sp}}$ requires both observation and spatial agreement.

3.4 CHANCE-CORRECTED SCORES

The raw values of $\mathcal{S}_{\text{sp}}$ and $\mathcal{S}_{\text{obs}}$ cannot be compared directly. Representations with more trusted components tend to receive larger values, and unrelated representations can still overlap by chance. The expected chance overlap is therefore derived by randomly orienting the patterns of $\mathbf{X}$ and $\mathbf{Y}$ while keeping their trust weights fixed.

Invariance. Let $\tau = \left(\sum_i w_{X,i}^2\right)\left(\sum_j w_{Y,j}^2\right) = \|\mathbf{Z}_X\|_F^2 \|\mathbf{Z}_Y\|_F^2$, the Cauchy–Schwarz inequality gives $\mathcal{S}_{\text{sp}}^2/\tau \in [0,1]$. Under independent random spatial and observation orientations, its expected value is

$$\mathbb{E}\left[\frac{\mathcal{S}_{\text{sp}}^2}{\tau}\right] = \frac{1}{pd} \tag{7}$$

as derived in Appendix A. The chance-corrected invariance score is therefore

$$\mathcal{S}_{\text{inv}}(\mathbf{X}, \mathbf{Y}) = \left[\frac{\mathcal{S}_{\text{sp}}^2/\tau - 1/(pd)}{1 - 1/(pd)}\right]_0^1, \tag{8}$$

where $[x]_0^1 = \min\{1, \max\{0, x\}\}$. The numerator removes the expected chance overlap, while the denominator preserves one as the maximum score.

Equivariance. Let $\kappa = \|\mathbf{Z}_X^\top \mathbf{Z}_X\|_F \|\mathbf{Z}_Y^\top \mathbf{Z}_Y\|_F$. The Cauchy–Schwarz inequality gives $\mathcal{S}_{\text{obs}}/\kappa \in [0,1]$. Under a uniformly random relative observation orientation, its expected value is

$$\mathbb{E}\left[\frac{\mathcal{S}_{\text{obs}}}{\kappa}\right] = \frac{\tau}{d\kappa} \tag{9}$$

Unlike invariance, which requires agreement in both the p-dimensional spatial space and the d-dimensional observation space, observation alignment depends only on the latter. Its chance fluctuations can therefore be larger, particularly when d is small. We account for this using $\delta_d = \sqrt{\frac{2}{(d-1)(d+2)}}$, for which $\delta_d \kappa$ bounds one null standard deviation of $\mathcal{S}_{\text{obs}}$; see Appendix B.

The chance-corrected equivariance score is

$$\mathcal{S}_{\text{equiv}}(\mathbf{X}, \mathbf{Y}) = \max\left\{\mathcal{S}_{\text{inv}}(\mathbf{X}, \mathbf{Y}), \left[\frac{\mathcal{S}_{\text{obs}} - \tau/d}{\kappa - \tau/d + \delta_d \kappa}\right]_0^1\right\}, \tag{10}$$

because invariance is the identity-action special case of equivariance. An ablation demonstrating the effect of this correction is provided in Appendix C.

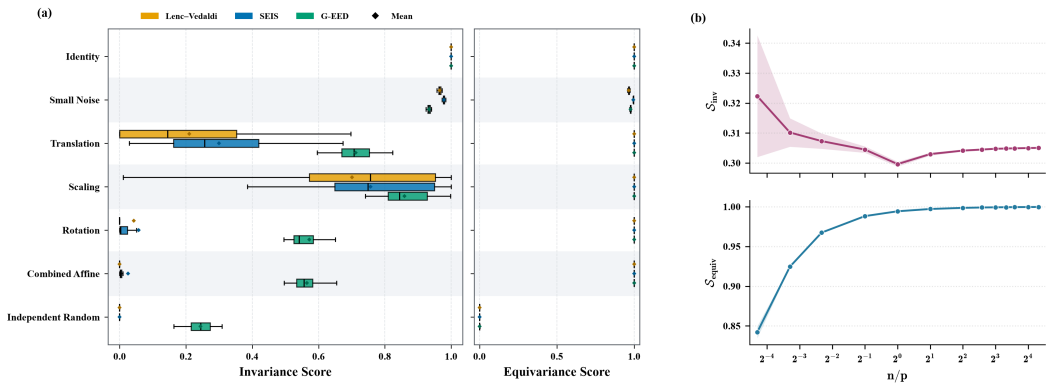

Figure 1: Controlled verification and observation sensitivity of SEIS. (a) Comparison with Lenc–Vedaldi and G-EED over 50 synthetic trials. (b) Sensitivity of SEIS to the observation-to-position ratio (n/p) under translation. SEIS distinguishes invariance from equivariance, rejects independent activations, and becomes stable as n/p increases.

4 Experiments

4.1 Controlled Validation and Observation Sensitivity

Controlled transformations. We validate SEIS on controlled synthetic activations and compare it with Lenc–Vedaldi (Lenc & Vedaldi, 2015) and G-EED (Kvinge et al., 2022). Each trial generates low-rank activations, then applies identity, additive noise ($\sigma = 0.1$), translation sampled within $\pm 15\%$ of the spatial size, scaling in $[0.8, 1.2]$, rotation in $[-180°, 180°]$, a combined affine transformation, or replacement with independent random activations. We conduct 50 independent trials. Lenc–Vedaldi uses a 70/30 train/test split to estimate its linear map, whereas G-EED is supplied with the known transformed target. For comparability, all outputs are mapped to $[0, 1]$, where higher values indicate stronger invariance or equivariance.

As shown in Fig. 1a, SEIS exhibits the expected behavior across all controls. Identity and small noise produce near-unit invariance and equivariance; geometric transformations preserve near-perfect equivariance while reducing invariance according to transformation severity; and independent activations score approximately zero on both dimensions. Lenc–Vedaldi follows the same overall trend but shows greater variability in its invariance estimates. G-EED correctly detects equivariance but assigns comparatively high invariance scores to strong transformations and even to independent activations (0.24 on average). Overall, SEIS cleanly distinguishes invariance from equivariance without fitting a transformation-specific map or requiring the transformation action as input. Appendix D further confirms this behavior on an untrained D_4-equivariant G-CNN.

Sensitivity to the number of observations. We examine the sensitivity of SEIS to the observation-to-position ratio n/p, where $p = HW$ is the number of spatial positions and $n = BC$ is the number of observations. We fix a 16×16 spatial grid ($p = 256$) and generate activations with additive Gaussian noise. A fixed translation is applied to generate the matched transformed activations. We vary n/p from 0.05 to 20, keeping the spatial basis and transformation fixed. Within each trial, we first generate the largest observation set and obtain smaller sets by retaining subsets of the same observations, ensuring that only n changes. Fig. 1b reports the mean and standard deviation across trials.

Equivariance is underestimated under severe undersampling, decreasing from 0.99 at $n/p = 0.50$ to 0.84 at $n/p = 0.05$. Invariance is less sensitive, remaining close to 0.30. Both scores are effectively stable once n is a few times larger than p: from $n/p = 2$ to 20, invariance changes by less than 0.01, while equivariance remains above 0.99. Fixed scaling, rotation, and combined affine transformations exhibit the same convergence pattern.

These results suggest using $n \geq 2p$ as a practical minimum. Because $n = BC$, the preferred strategy is to increase the number of evaluated images or aggregate additional batches. If additional

observations are unavailable, p may be reduced through spatial pooling, but this changes the spatial subspace being measured and can suppress fine-scale structure; consequently, the same pooling policy and measured resolution should be used across all compared models.

4.2 Representation Dynamics During Training and Data Augmentation

Experimental setup. We study representation dynamics and augmentation effects using a ResNet-18 (He et al., 2016) trained on CIFAR-100 classification (Krizhevsky & Hinton, 2009) and a ResNet-18 encoder followed by a fully convolutional decoder (ResNet-FCN) on PASCAL VOC2012 segmentation (Everingham & Winn, 2011). At every 10 epochs, $\mathcal{S}_{\text{equiv}}$ and $\mathcal{S}_{\text{inv}}$ are measured on the validation set using fixed pairs of original and affine-transformed images. Transformations are sampled with rotation in $[-15°, 15°]$, translation up to 5% per axis, and scaling in $[0.9, 1.1]$. Training dynamics are summarized by the mean absolute change from the first epoch across classification layers and, separately, the segmentation encoder and decoder. To evaluate augmentation, we compare layer-wise scores at the best validation checkpoint between models trained with and without random affine augmentation using the same transformation ranges. All reported results are averaged over five random seeds for each model.

Representation dynamics during training. As shown in Fig. 2a, classification representations change rapidly early in training and stabilize after roughly 40 epochs, by which point the mean absolute changes $|\Delta S|$ in equivariance and invariance reach 0.32 and 0.36, respectively. This sensitivity is consistent with evidence that standard CNNs do not inherently preserve geometric transformations (Azulay & Weiss, 2019; Zhang, 2019). Segmentation trained from scratch exhibits stronger depth-dependent dynamics (Fig. 2b): by epoch 100, equivariance and invariance change by $0.21/0.28$ in the encoder and $0.46/0.57$ in the task-specific decoder. This observation is consistent with evidence that deep representations converge progressively from lower to higher layers during training (Raghu et al., 2017). Appendix E further shows that initializing the encoder with ImageNet-pretrained weights substantially reduces these changes relative to training from scratch, consistent with evidence that pretrained networks stay closer to their lower-layer features during subsequent optimization (Raghu et al., 2019).

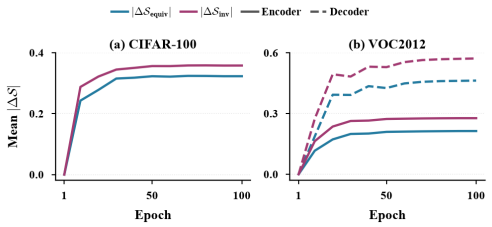

Figure 2: Mean absolute changes in $\mathcal{S}_{\text{equiv}}$ and $\mathcal{S}_{\text{inv}}$ throughout training.

Effect of data augmentation. Without augmentation, classification equivariance generally decreases with depth, while invariance initially falls and then recovers in the final stages (Fig. 3a). Affine augmentation largely prevents this degradation, keeping most equivariance scores above 0.75 and raising final-layer invariance from 0.45 to 0.74, with the largest gains occurring in deeper blocks. Segmentation exhibits a pronounced encoder–decoder pattern (Fig. 3b): both scores decline through the encoder but recover as spatial resolution is restored by the decoder. Augmentation has little influence on the earliest encoder layers but consistently strengthens deeper representations, improving both scores by approximately 0.09 across the decoder. These results agree with prior work showing that data augmentation can promote both invariance and equivariance in learned representations (Benton et al., 2020; Bruintjes et al., 2023).

4.3 Effects of Supervision and Decoder Architecture

Experimental setup. We examine how task weighting and decoder architecture shape the transformation geometry of representations learned on PASCAL VOC2012. We compare the final decoder representations of ResNet-FCN and ResNet-UNet under segmentation-only supervision and joint classification–segmentation supervision. For multi-task learning (MTL), the objective is $\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{cls}}$, where $\lambda \in \{1, 0.75, 0.5, 0.25, 0.1\}$ controls the contribution of image-level classification supervision. SEIS is evaluated at the best validation checkpoint using validation images. We apply affine transformations with rotation in $[-15°, 15°]$, translation of up to 5%, and scaling in

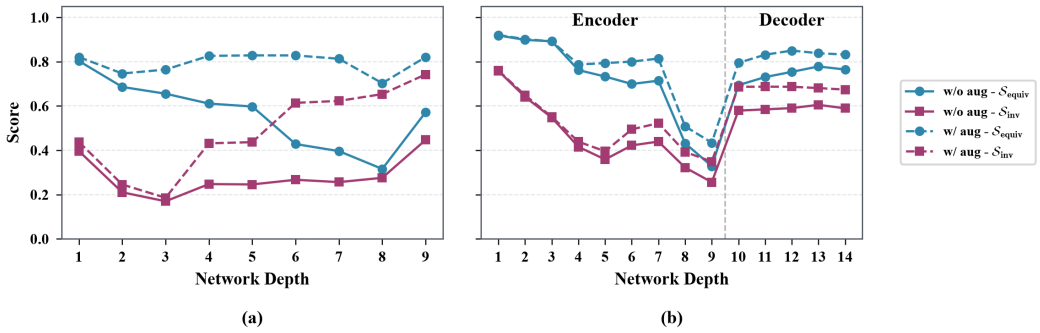

Figure 3: Effect of data augmentation on layer-wise $\mathcal{S}_{\text{equiv}}$ and $\mathcal{S}_{\text{inv}}$ for (a) CIFAR-100 and (b) Pascal VOC 2012, w/ and w/o augmentation. Augmentation generally improves both scores, particularly in deeper layers.

$[0.9, 1.1]$. The same images and transformation draws are used for every configuration. All reported results are averaged over five random seeds for each model.

Task weighting and decoder architecture. Fig. 4 reveals both an architectural difference and an architecture-dependent response to task weighting. Across supervision settings, UNet produces more invariant and equivariant final-decoder representations than FCN. Reducing the classification weight below $\lambda = 1$ leads to similar mean invariance/equivariance scores for FCN ($0.60/0.77$ versus $0.60/0.78$), but higher scores for UNet ($0.63/0.80$ versus $0.61/0.79$). Because $\lambda < 1$ increases the relative contribution of segmentation supervision, these opposing trends suggest that task weighting interacts with the pathway through which spatial information reaches the decoder. UNet skip connections transmit spatially localized, multi-scale encoder features directly to the decoder (Ronneberger et al., 2015), allowing stronger relative segmentation supervision to shape these features throughout reconstruction. In contrast, the skip-free FCN must recover spatial predictions entirely from its semantic bottleneck. More broadly, the results reinforce that multi-task effects cannot be attributed to loss weighting alone but depend on the architecture through which the objectives share representations (Kendall et al., 2018; Standley et al., 2020).

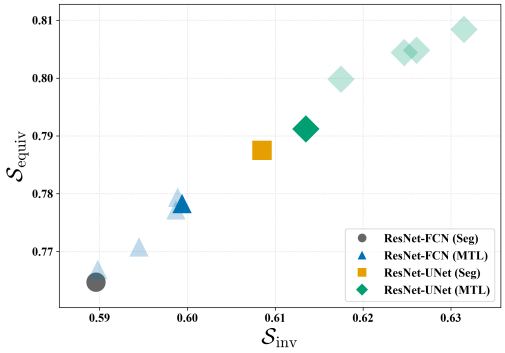

Figure 4: SEIS scores for ResNet-FCN and ResNet-UNet models trained with segmentation (Seg) or multi-task learning (MTL); transparent markers indicate $\lambda < 1$.

Table 1: Correlation between SEIS scores and mIoU ($N = 60$).

	r	ρ
$\mathcal{S}_{\text{inv}}$	0.92	0.88
$\mathcal{S}_{\text{equiv}}$	0.93	0.87

Correlation with segmentation performance. Table 1 shows a strong correlation between the final-decoder geometry and the segmentation performance. Invariance is strongly positively correlated with mIoU according to both Pearson's correlation coefficient (Pearson, 1895) ($r = 0.92$) and Spearman's rank correlation coefficient (Kendall, 1948) ($\rho = 0.88$). Equivariance shows a similar strong association ($r = 0.93$, $\rho = 0.87$). This interpretation is consistent with prior work showing that transformation-consistent invariance and equivariance provide effective inductive biases for pixel-level representation learning (Wang et al., 2020; Cho et al., 2021).

4.4 Effect of Positional Encoding in ViTs

Experimental setup. We study how positional encoding affects the translation structure learned by ViTs. We train four otherwise matched DeiT-Tiny-scale models (Touvron et al., 2021) on Imagenette

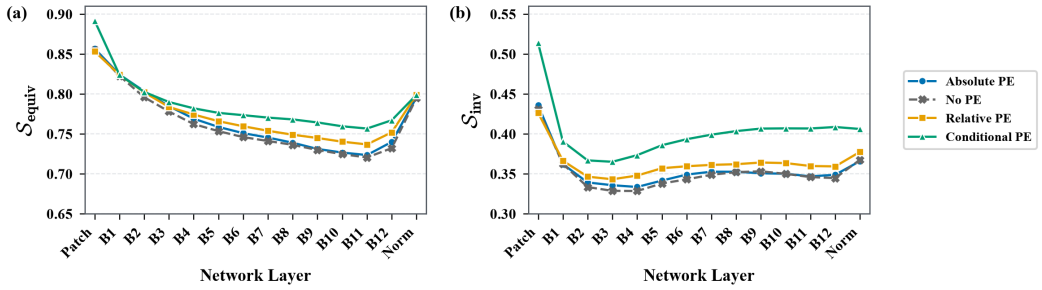

Figure 5: Layer-wise $\mathcal{S}_{\text{equiv}}$ (a) and $\mathcal{S}_{\text{inv}}$ (b) for DeiT-Tiny models using absolute, no, relative, or conditional positional encoding (PE) on Imagenette. Scores are averaged over horizontal and vertical image translations of 4–16 pixels. Conditional PE preserves the strongest overall translation structure and develops a pronounced invariance advantage.

(Howard, 2019) using learned absolute positional embeddings, no positional encoding, relative positional encoding (RPE) (Wu et al., 2021), and conditional positional encoding (CPE) (Chu et al., 2023). All models use 224×224 inputs, 16×16 patches, and 12 transformer blocks. Training includes translations of up to 10% of the image dimensions. We train each variant from scratch with five random seeds and select the checkpoint with the highest validation accuracy for each seed. We then compute layer-wise SEIS on validation images under horizontal and vertical translations of 4, 8, 12, and 16 pixels. Reported values are means and standard deviations across seeds after averaging over the transformations.

Results. Fig. 5 shows that equivariance generally decreases with depth before partially recovering at the final normalization layer, indicating that successive transformer blocks progressively modify translation structure. CPE preserves both properties most effectively: at the final transformer block, it achieves equivariance/invariance scores of $0.77/0.41$, compared with $0.75/0.36$ for RPE, $0.74/0.35$ for absolute encoding, and $0.73/0.34$ without positional encoding. CPE's invariance advantage becomes increasingly pronounced in deeper blocks, while RPE provides the next strongest preservation. These patterns are consistent with conditional encodings using local image context instead of fixed coordinates and with relative encodings representing spatial relationships through relative offsets (Chu et al., 2023; Wu et al., 2021). Both designs therefore introduce less dependence on fixed image coordinates than absolute positional embeddings, helping preserve translation structure across layers.

4.5 Label-Free Prediction of Affine Robustness

Experimental setup. We evaluate whether SEIS provides information about model robustness beyond that conveyed by validation accuracy (A_{val}). The experiment includes 50 frozen ImageNet-pretrained classifiers spanning CNN, ViT, and mixer-style architectures. For each model, SEIS is computed from its final spatial representation using paired clean and transformed ImageNet (Russakovsky et al., 2015) validation images. The fixed affine transformation combines a $15°$ rotation, translation by $(+5\%, -5\%)$, and scaling by 0.9. We then evaluate clean and transformed accuracy on ImageNetV2 (Recht et al., 2019). We compare A_{val} and $\mathcal{S}_{\text{equiv}}$ as predictors of transformed accuracy A_{trans} and performance change $A_{\text{trans}} - A_{\text{clean}}$. The latter isolates transformation robustness from the model's baseline accuracy, with values closer to zero indicating better performance preservation.

Results. As shown in Fig. 6, A_{val} strongly predicts absolute performance after transformation, correlating with A_{trans} at $r = 0.90$ and $\rho = 0.92$. $\mathcal{S}_{\text{equiv}}$ is less predictive of absolute transformed accuracy ($r = 0.74$, $\rho = 0.80$), but more directly reflects whether performance is preserved: its correlation with $A_{\text{trans}} - A_{\text{clean}}$ reaches $r = 0.77$ and $\rho = 0.82$, compared with $r = 0.62$ and $\rho = 0.77$ for A_{val}. Thus, A_{val} primarily indicates *how accurate* a model will be, whereas $\mathcal{S}_{\text{equiv}}$ indicates *how well that accuracy is retained* under the affine transformation. $\mathcal{S}_{\text{inv}}$ is a weaker indicator in this experiment. Importantly, unlike A_{val}, SEIS requires only paired representations from original and transformed inputs and no ground-truth class labels. It is therefore suitable for assessing

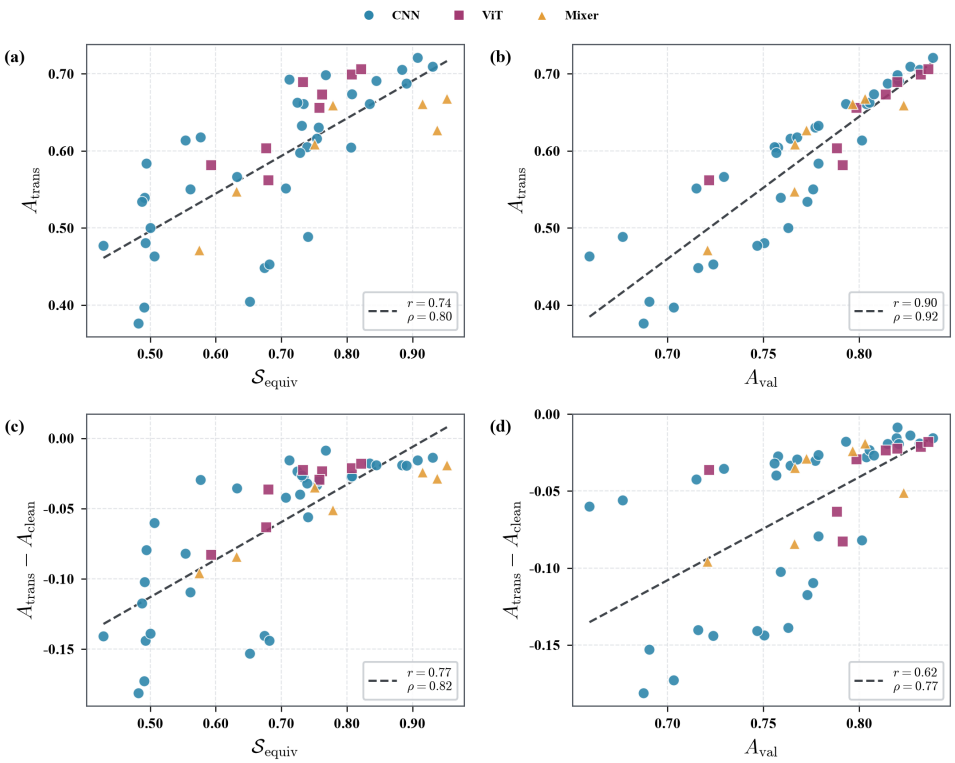

Figure 6: Comparison of validation accuracy (A_{val}) and $\mathcal{S}_{\text{equiv}}$ as model indicators. (a–b) show how $\mathcal{S}_{\text{equiv}}$ and ImageNet A_{val} relate to accuracy on affine-transformed ImageNetV2. (c–d) show how the same indicators relate to the change in ImageNetV2 accuracy caused by the affine transformation. Dashed lines show linear fits; r and ρ denote Pearson and Spearman correlations, respectively.

transformation robustness on newly collected data for which labels are unavailable or expensive to obtain.

5 Conclusion

We introduced SEIS, a label-free subspace framework for quantifying invariance and equivariance within neural representations. By separating observation-space preservation from spatial correspondence, SEIS provides complementary scores for transformation-induced reorganization and representation stability. Its continuous spectral weighting avoids hard rank selection, while chance correction reduces overlap caused by representation dimensionality.

Controlled experiments confirm that SEIS distinguishes invariant, equivariant, and unrelated activation pairs. Across classification and segmentation, SEIS reveals systematic differences in how transformation structure develops during training and across network depth. Affine augmentation primarily strengthens deeper representations, segmentation produces distinct encoder–decoder dynamics, and decoder architecture changes how representations respond to multi-task supervision. In ViTs, conditional and relative positional encodings preserve translation structure more effectively than absolute encoding or no positional encoding. Across 50 ImageNet-pretrained classifiers, SEIS equivariance is also strongly associated with performance retention under affine transformation, complementing validation accuracy while requiring no ground-truth labels.

Limitations and future work. SEIS assumes that transformations act approximately linearly on the measured representation and may not capture nonlinear latent transformation laws. It also assesses compatibility with an unknown linear spatial action rather than recovering the action itself. Future work could extend SEIS to nonlinear transformation models and explicitly estimate the latent action, enabling direct comparison with the known input-space transformation.

AI USE STATEMENT

We have not used generative AI tools for any part of this work, including research ideation, writing, coding, or data analysis. All text, code, and results in this paper were produced by the authors.

REFERENCES

Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G Wilson. Learning invariances in neural networks from training data. *Advances in Neural Information Processing Systems*, 33:17605–17616, 2020.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Robert-Jan Bruintjes, Tomasz Motyka, and Jan van Gemert. What affects learned equivariance in deep image recognition models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4839–4847, 2023.

Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16794–16804, June 2021.

Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. *International Conference on Learning Representations*, 2023.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Weijian Deng, Stephen Gould, and Liang Zheng. On the strong correlation between model invariance and generalization. *Advances in Neural Information Processing Systems*, 35:28052–28067, 2022.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in Neural Information Processing Systems*, 34:1556–1568, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, 8 (5):2–5, 2011.

Matan Gavish and David L Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. *Advances in Neural Information Processing Systems*, 22, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations*, 2019.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in Statistics: Methodology and Distribution*, pp. 162–190. Springer, 1992.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet. *GitHub*, 2019. URL `https://github.com/fastai/imagenette`.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

Maurice George Kendall. Rank correlation methods. 1948.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMlR, 2019.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Henry Kvinge, Tegan Emerson, Grayson Jorgenson, Scott Vasquez, Tim Doster, and Jesse Lew. In what ways are deep neural networks invariant and how should we measure this? In *Advances in Neural Information Processing Systems*, volume 35, pp. 32816–32829, 2022.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 991–999, 2015.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

Facundo Manuel Quiroga, Jordina Torrents-Barrena, Laura Cristina Lanzarini, and Domenec Puig-Valls. Invariance measures for neural networks. *Applied Soft Computing*, 132:109817, 2023.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in Neural Information Processing Systems*, 30, 2017.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems*, 32, 2019.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241. Springer, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, pp. 8459–8468. PMLR, 2020.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pp. 9120–9132. PMLR, 2020.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12272–12281. IEEE, 2020.

Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10013–10021. IEEE, 2021.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *International Conference on Learning Representations*, 2019.

Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*, pp. 7324–7334. PMLR, 2019.

A NULL MOMENTS FOR THE INVARIANCE STATISTIC

Let $\mathbf{u}_{X,i}$ and $\mathbf{v}_{X,i}$ denote the spatial and observation singular vectors associated with component i, and define the corresponding vectors for $\mathbf{Y}$ analogously. Expanding Eq. 6 gives

$$\mathcal{S}_{\mathrm{sp}} = \sum_{i,j} w_{X,i} w_{Y,j} \left\langle \mathbf{u}_{X,i}, \mathbf{u}_{Y,j} \right\rangle \left\langle \mathbf{v}_{X,i}, \mathbf{v}_{Y,j} \right\rangle . \tag{11}$$

Condition on the trust weights and independently randomize the relative orientations of the spatial and observation singular vectors. For the spatial overlaps

$$A_{ij} = \left\langle \mathbf{u}_{X,i}, \mathbf{u}_{Y,j} \right\rangle , \tag{12}$$

Under the null, the relative orientation of the two spatial bases is uniformly random. Consequently, the component overlaps satisfy

$$\mathbb{E}[A_{ij} A_{k\ell}] = \frac{1}{p} \, \mathbb{1}\{i = k, \; j = \ell\}. \tag{13}$$

Similarly, the observation vectors lie in the d-dimensional centered observation space. For

$$B_{ij} = \left\langle \mathbf{v}_{X,i}, \mathbf{v}_{Y,j} \right\rangle , \tag{14}$$

the corresponding second moment is

$$\mathbb{E}[B_{ij} B_{k\ell}] = \frac{1}{d} \, \mathbb{1}\{i = k, \; j = \ell\}. \tag{15}$$

The spatial and observation orientations are independent under the null. Consequently, all cross-component terms vanish when Eq. 11 is squared, yielding

$$\mathbb{E}[\mathcal{S}_{\mathrm{sp}}^2] = \sum_{i,j} w_{X,i}^2 w_{Y,j}^2 \mathbb{E}[A_{ij}^2] \mathbb{E}[B_{ij}^2] \tag{16}$$

$$= \frac{1}{pd} \left(\sum_i w_{X,i}^2 \right) \left(\sum_j w_{Y,j}^2 \right) = \frac{\tau}{pd}. \tag{17}$$

Dividing by τ gives the null expectation in Eq. 7.

B Null Moments for the Equivariance Alignment Statistic

In Section 3.4, we model the chance observation alignment by assuming that the relative observation-space orientation between the representations $\mathbf{X}$ and $\mathbf{Y}$ is uniformly random.

Let $\mathbf{K}_X = \mathbf{Z}_X^\top \mathbf{Z}_X$ and $\mathbf{K}_Y = \mathbf{Z}_Y^\top \mathbf{Z}_Y$ denote the observation kernels. Because the representations are centered, the observation kernels act on a $d = n - 1$ dimensional space. To model chance alignment, we keep their eigenvalues fixed and randomize only their relative orientation. Let $\mathbf{Q} \in \mathcal{O}(d)$ be a uniformly random orthogonal matrix. Then $\mathbf{Q}\mathbf{K}_Y\mathbf{Q}^\top$ has the same spectrum as $\mathbf{K}_Y$ but a random orientation relative to $\mathbf{K}_X$. Accordingly, the observation alignment under the null is

$$\mathcal{S}_{\text{random}} = \left\langle \mathbf{K}_X, \mathbf{Q}\mathbf{K}_Y\mathbf{Q}^\top \right\rangle_F = \operatorname{tr}\left(\mathbf{K}_X\mathbf{Q}\mathbf{K}_Y\mathbf{Q}^\top\right). \tag{18}$$

Averaging over uniformly random orientations gives $\mathbb{E}_{\mathbf{Q}}\left[\mathbf{Q}\mathbf{K}_Y\mathbf{Q}^\top\right] = \frac{\operatorname{tr}(\mathbf{K}_Y)}{d}\mathbf{I}_d$. Using this identity and the linearity of the trace gives

$$\mathbb{E}[\mathcal{S}_{\text{random}}] = \frac{\operatorname{tr}(\mathbf{K}_X)\operatorname{tr}(\mathbf{K}_Y)}{d} = \frac{\tau}{d}. \tag{19}$$

Here, $\operatorname{tr}(\mathbf{K}_X) = \operatorname{tr}(\mathbf{Z}_X^\top \mathbf{Z}_X) = \|\mathbf{Z}_X\|_F^2$, and similarly for $\mathbf{Y}$, so $\operatorname{tr}(\mathbf{K}_X)\operatorname{tr}(\mathbf{K}_Y) = \tau$. This is the chance expectation subtracted in Eq. 10.

To compute the standard deviation, we require the variance of $\mathcal{S}_{\text{random}}$. Using the second-order rotational average gives

$$\operatorname{Var}[\mathcal{S}_{\text{random}}] = \frac{2}{(d-1)(d+2)}\left(\|\mathbf{K}_\mathbf{X}\|_F^2 - \frac{(\operatorname{tr}(\mathbf{K}_X))^2}{d}\right)\left(\|\mathbf{K}_Y\|_F^2 - \frac{(\operatorname{tr}(\mathbf{K}_Y))^2}{d}\right). \tag{20}$$

Since $\mathbf{K}_X$ and $\mathbf{K}_Y$ are positive semi-definite, $(\operatorname{tr}(\mathbf{K}_X))^2/d \geq 0$ and $(\operatorname{tr}(\mathbf{K}_Y))^2/d \geq 0$. Dropping the nonnegative trace corrections gives the conservative bound

$$\operatorname{Var}[\mathcal{S}_{\text{random}}] \leq \frac{2}{(d-1)(d+2)}\|\mathbf{K}_X\|_F^2\|\mathbf{K}_Y\|_F^2. \tag{21}$$

Recall that $\kappa = \|\mathbf{K}_X\|_F\|\mathbf{K}_Y\|_F$, which means $\kappa^2 = \|\mathbf{K}_X\|_F^2\|\mathbf{K}_Y\|_F^2$. Substituting κ, we obtain the upper bound for the variance:

$$\operatorname{Var}[\mathcal{S}_{\text{random}}] \leq \frac{2}{(d-1)(d+2)}\kappa^2. \tag{22}$$

Consequently,

$$\sqrt{\operatorname{Var}[\mathcal{S}_{\text{random}}]} \leq \delta_d\kappa. \tag{23}$$

This bounded standard deviation $\delta_d\kappa$ is added to the denominator of Eq. 10 to stabilize the chance-corrected equivariance score against large chance fluctuations in low dimensions.

C Ablation of Chance Correction

We evaluate whether the chance-correction terms in SEIS are necessary for distinguishing meaningful representation alignment from overlap that arises by chance in finite-dimensional spaces. The ablation considers the null case in which two activation representations are statistically independent and should therefore receive invariance and equivariance scores close to zero.

We independently sample two activation matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{p\times n}$ with $X_{ij}, Y_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, where $p = 64$ and the target observation ratios are $n/p \in \{0.05, 0.1, 0.2, 0.5, 1, 2, 4, 6, 8, 10, 12, 16, 20\}$. Each setting is evaluated over 50 independent trials. We compare the complete SEIS formulation with an ablated variant that uses the same spectral factors and soft spectral weights but does not subtract the expected null overlap.

The independent condition is shown in Fig. 7. Chance correction has only a small numerical effect on the $\mathcal{S}_{\text{inv}}$ because simultaneous random alignment in both the spatial and observation domains is

already unlikely. The effect on $\mathcal{S}_{\text{equiv}}$ is substantially larger. Without chance correction, independent representations receive an $\mathcal{S}_{\text{equiv}}$ of 0.96 at $n/p = 0.05$, despite sharing no underlying structure. The uncorrected score remains 0.31 at $n/p = 2$, and at $n/p = 20$, it remains 0.05. After chance correction, the corrected score remains close to its intended null value across both undersampled and well-sampled regimes.

These results demonstrate that the chance correction is necessary for interpreting $\mathcal{S}_{\text{equiv}}$. Without it, finite-dimensional observation-subspace overlap can cause statistically independent representations to appear strongly equivariant, particularly when n is small.

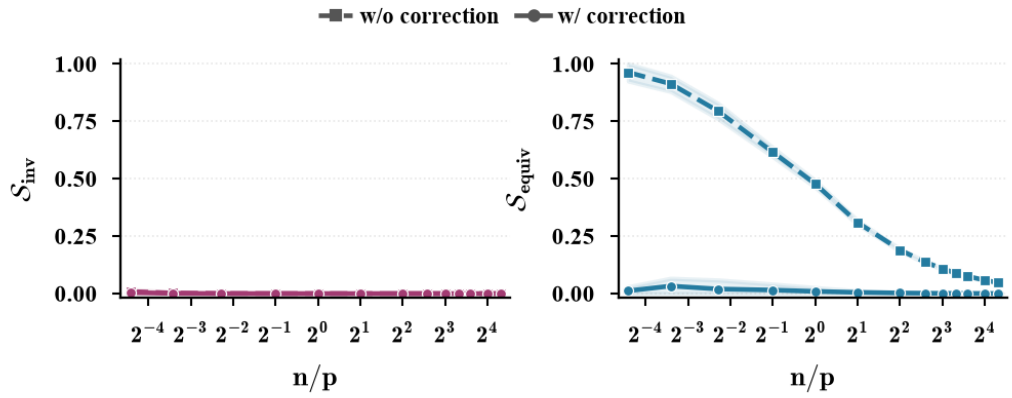

Figure 7: Chance-correction ablation for independent Gaussian activation matrices. Dashed lines show scores without chance correction, and solid lines show the complete chance-corrected SEIS scores. Points denote means over 50 independent trials, and shaded regions indicate the standard deviation.

D Verification of SEIS on an Equivariant G-CNN

We evaluate whether SEIS recognizes equivariance that is guaranteed by architectural design. We construct an untrained $D4$-equivariant lifting convolution (Cohen & Welling, 2016), in which each randomly initialized 5×5 filter is shared across 4 rotations and their reflected counterparts. Because the model is not trained, its transformation behavior arises solely from its architecture rather than learned augmentation or supervision.

We apply the model to 10 disjoint batches of 32 MNIST test images and evaluate rotations of $90°$, $180°$, and $270°$, a fixed spatial translation, and horizontal reflection. For G-CNN activations in $\mathbb{R}^{B\times C\times G\times H\times W}$, the group and image coordinates jointly define the positional domain, giving $p = GHW$ and $n = BC$.

Fig. 8 shows consistently high equivariance scores of approximately 0.99 for every transformation. Invariance remained low, ranging from 0.02 to 0.11, because the representations moved according to the D4 action rather than remaining spatially fixed. These results demonstrate that SEIS identifies the rotation-, translation-, and reflection-equivariant structure built into the G-CNN architecture.

E Effect of Encoder Pretraining

We compare the SEIS dynamics of ResNet–FCN models trained on Pascal VOC 2012 with an ImageNet-pretrained encoder and from scratch. For every layer, we report the absolute score change from epoch 1 and average it over the encoder or decoder layers.

As shown in Fig. 9, pretraining substantially stabilizes the encoder. At epoch 100, its equivariance and invariance changes are both around 0.01, compared with 0.21 and 0.28 when trained from

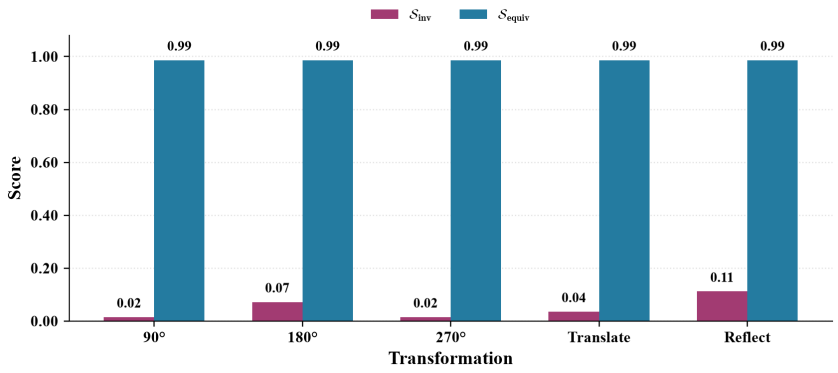

Figure 8: Mean $\mathcal{S}_{\text{inv}}$ and $\mathcal{S}_{\text{equiv}}$ for an untrained $D4$-equivariant G-CNN, averaged over 10 MNIST batches. SEIS assigns consistently high equivariance while maintaining low invariance, correctly identifying that the representations undergo structured spatial reorganization.

scratch. The decoder changes more than the encoder under both initializations and continues evolving later in training, both reaching 0.06 for the pretrained model and 0.46 and 0.57 from scratch. This suggests that pretrained features remain comparatively stable while the decoder undergoes most of the task-specific adaptation.

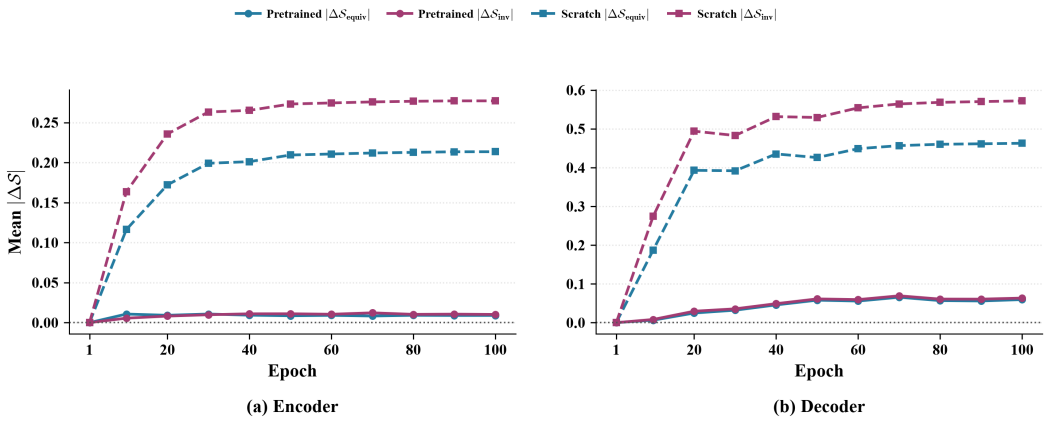

Figure 9: Mean absolute changes in equivariance and invariance relative to epoch 1. Solid lines denote the pretrained model and dashed lines denote training from scratch.