LOCATION: WHAT VIEW MANIFOLDS CAN AND CANNOT TELL US ABOUT INVARIANT RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep vision networks could recognise certain objects from every viewpoint well but fail on others, and we hypothesise two possible ways in which recognition invariance is achieved through the viewpoint manifold of the object: gradual collapse in later layers, or reorganisation. We test both possibilities against the networks' classification heads and found three factors influencing readout consistency, which are how far the view set sits from the decision boundary, how widely they spread, and how much of the spread the readout measures on the viewpoint manifold. We measured the factors using 15 ShapeNet objects rendered at forty angles in a full 360° azimuth rotation with ViT-B/16 and ResNet-50. In the two networks tested, correct recognition is rarely achieved, with two and five in fifteen objects, respectively. Relative collapse of the manifold across depth is not a requirement for correct recognition either. Compactness of viewpoints is positively related to the consistency of the recognition and shows no detectable association with its correctness. Correctness is related to where the viewpoint set lies relative to the boundary. Summaries computed only from distances between views cannot certify recognition.

1 INTRODUCTION

The same object seen from a new angle can produce an image that shares few pixels with any previous view of it, yet a competent visual system still recognises it. Deep neural networks trained on large-scale datasets have these invariant properties for some objects but also fail for others, particularly objects in unusual poses and viewpoints. (Alcorn et al., 2019; Barbu et al., 2019; Abbas & Deny, 2023; Madan et al., 2022). Why a network succeeds on one object and fails on another is a question about how its representation meets its readout.

Representational geometry offers two possible answers. In the *collapse* hypothesis, a network discards nuisance variation and compresses different views of an object towards a single point first termed 'neural collapse' at the end of training by (Papyan et al., 2020) and in accounts of invariance as discarded information (Achille & Soatto, 2018). In the *reorganisation* picture, sometimes also called untangling, the network retains pose information while untangling object manifolds so that identity becomes linearly readable (DiCarlo & Cox, 2007; Chung & Abbott, 2021; Cohen et al., 2020). Directional neural collapse supports this theory: variation shrinks along class-separating directions while total within-class variation stays large (Luthra et al., 2025; 2026). Both pictures are often tested with label-free geometric measurements, or with readouts fitted afterwards to class-level data. However, previous studies rarely evaluate the network's predictions for different views of the same object.

In this study, we test the latter with three measurements. For a fixed affine head, the margin between two classes is a linear function of the representation, and the influence of consistency of an object's recognition result across all its views could be decomposed into three factors (Section 3): the viewpoint set's average distance from the boundary, how much the view points spread, and the fraction of the spread that the readout could see, each of them pointing to a way to higher invariance (Figure 1a): shrink the viewpoints set (a1, collapse), turn its variation away from the readout direction (a2, directional collapse), or move the whole set away from the decision boundary (a3, location). Using 15 ShapeNet objects rendered at forty angles in a full 360° azimuth rotation, we tested them

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.

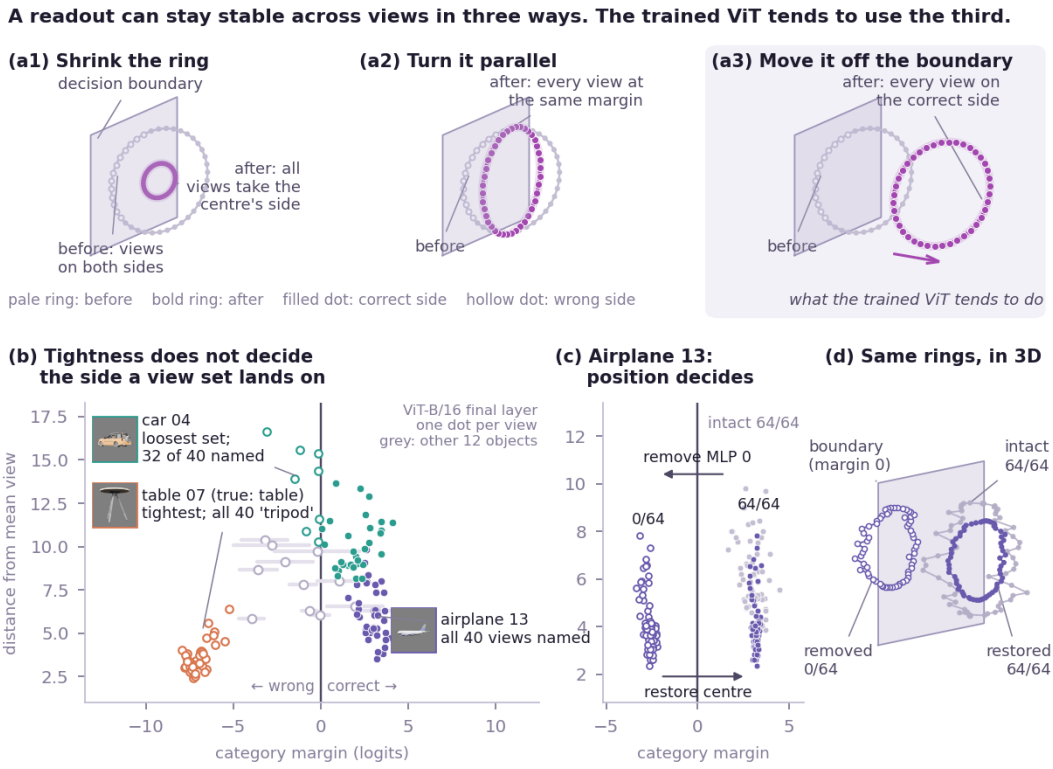

Figure 1: **Three ways a linear readout can keep an object's views on one side of its margin, and the last one is what our pretrained ViT tends to use.** (a) Light purple sheet denotes the margin of a fixed linear readout; the ring represents one object's all viewpoint set in the manifold, forming a closed loo the the rotation finishes in 360 degrees. Filled dots are viewpoint recognised on the correct side; hollow dots are views on the wrong side; Faint grey ring is the starting condition. The orchid ring is the after condition. A network can have every view on one side by shrinking the view set (a1, collapse), by turning it parallel to the margin so that all views are within one side of the margin (a2, directional collapse), or by moving the whole set away from the boundary (a3). (b) ViT-B/16, 15 ShapeNet objects at 40 rotations, final layer. Horizontal: each view's category margin (positive when the category is named correctly). Vertical: its distance from the object's mean view set centre. purple: Airplane 13, orange: table 07 and green: car 04; The grey bars show each other object's mean $\pm$ s.d. margin (filled: recognised at $\geq 90\%$ of views). The tightest set (table 07) is wrongly classified as "tripod" at all 40 views; 32 in 40 of the loosest set(car 04) are named correctly. (c) Airplane 13 at 64 rotations instead of 40, same axes. lilac-grey: Original recognition (64/64 correct). hollow purple: Removing the update of the first MLP block we get 0/64 correct. Filled in purple: moving the damaged set back to the intact centre without changing its shape: 64/64 correct (a move computed exactly, because the head is affine). (d) The three states of (c) in 3D. the light purple surface denotes the margin $= 0$.

with ViT-B/16 and ResNet-50. We first established which objects each network actually recognises, and then measured the three terms with the native heads.

We summarise the results in Table 1. Relative collapse across depth is neither necessary nor sufficient for recognition, and final compactness tracks the consistency of the recognition result without obvious associations with its correctness (Section 5). Viewpoint variation is more aligned with each object's decision direction than a random orientation (Section 6). Correctness is related to the location of the set. For one recognised airplane 13 under ViT, recognition error was caused mainly by displacement of the view set rather than by changing the distribution of viewpoint set(Section 7).

This study provides the following empirical contributions. (1) Relative collapse is not necessary (ResNet-50) and not sufficient (both networks) for viewpoint-invariant recognition, and compactness

Table 1: **Three ways of achieving higher invariance.** Recommended preprocessing; Spearman ρ over the 15 objects.

Route (Fig. 1a)	Test	ViT-B/16	ResNet-50
a1: shrink the set (collapse §5)	Does contraction correlate with recognition?	Not sufficient: 6 of 8 contracting objects are not recognised (2 recognised contract most). $\rho(A_{\text{final}}, C) = +0.01$	Not necessary: cars 03 and 04 are recognised while A grows. Not sufficient: table 08 contracts most, $\rho(A_{\text{final}}, C) = -0.24$
a2: turn it away (directional collapse §6)	Does the readout see the spread less than for a random orientation?	No: Sees more, for 15 of 15 objects (median $2.8\times$)	No: Sees more, for 15 of 15 objects (median $4.0\times$)
a3: move it (location §7)	Does correctness relate to centre location?	Yes: $\rho(m_{\text{loc}}, C) = 0.95$; re-centring restores 64/64	Yes: $\rho(m_{\text{loc}}, C) = 0.97$

tracks consistency rather than correctness. (2) A test of directional collapse: changes caused by altering viewpoints occur more than randomly along directions that matter to the classifier head (3) We hypothesise that a summary computed from distances between views alone could not ensure whether a network recognises the object.

2 Related Work

Collapse and reorganisation in invariance. Neural collapse describes the terminal phase of training, in which within-class variability vanishes and class means form a simplex (Papyan et al., 2020). Collapse like statistics are also shown to decrease across layers (He & Su, 2023; Rangamani et al., 2023), although collapse on training data usually does not have to carry over to test data (Hui et al., 2022). The reorganisation/untangling view instead states that the ventral stream and visual deep networks reformat object manifolds until identity is linearly separable while pose information is retained (DiCarlo & Cox, 2007; Hong et al., 2016; Chung & Abbott, 2021). Manifold capacity theory on the other hand measures separability quantitatively using manifold radius, dimension and correlations (Chung et al., 2018; Cohen et al., 2020; Wakhloo et al., 2023). Collapse implies separability to an extreme, but separability does not require collapse: a large manifold could remain separable if its spread along the directions which the readout uses is small relative to the margin. Directional neural collapse points out exactly this phenomenon(Luthra et al., 2025; 2026), and Avidan et al. (2026) found that rotating each class's residuals around its centre improves prototype classification. We test these against the native heads of the pretrained networks with single objects.

Measuring invariance and representations. Invariance is usually measured without labels, either as the stability of unit or feature responses under transformations (Goodfellow et al., 2009; Kvinge et al., 2022) or as prediction consistency, such as flip probability under perturbation (Hendrycks & Dietterich, 2019) or shift consistency (Zhang, 2019). Such measures record whether an output changes, not whether it is correct. Representations are also summarised by intrinsic dimension (Ansuini et al., 2019), topology (Naitzat et al., 2020), and similarity indices (Kornblith et al., 2019), which can change while the function stays fixed (Davari et al., 2023; Ding et al., 2021). Probing shows what is decodable, not what the network uses (Alain & Bengio, 2016; Hewitt & Liang, 2019; Elazar et al., 2021), and interventions on representations can mislead about mechanism (Makelov et al., 2024).

Class-level and object-level readouts. Analyses of directional collapse and prototype coding read decisions off class means (Luthra et al., 2025; Avidan et al., 2026). Prototype-based analyses fail to detect this because they define each class centre to be correct from the onset. We instead evaluated the representations using the network's actual classifier so we were able to detect it.

3 SETUP

Stimuli. Fifteen exemplar ShapeNet (Chang et al., 2015) objects were selected, three from each of five categories (chair, car, table, lamp, airplane) labelled as object 00–14. Each object was rendered upright at 40 azimuths (spaced 9° apart), at 224×224 pixels on a uniform grey background, with fixed low camera elevation, distance, and lighting. For additional interventions in Section 7, Airplane 13 is also rendered at 64 azimuths (spaced $5.625°$ apart).

Models and view representation. We applied pretrained ViT-B/16 (Dosovitskiy et al., 2021) and ResNet-50 (He et al., 2016) with torchvision ImageNet-1k weights (`IMAGENET1K_V1` and `IMAGENET1K_V2`, respectively). Both models retained their native linear classification head $z(h) = Wh + b$. We passed each rendered image through the models and, at every block, took the class token (ViT) or the spatially averaged feature map (ResNet) as the representation of the view. The input h to the classification head, namely, the class token after the final LayerNorm (ViT; $p = 768$) or the output of global average pooling (ResNet; $p = 2048$), was used as the final representation. Extracted representations were indexed by their *relative depth* $d \in [0, 1]$, from block 0 (ViT) or the stem (ResNet) at $d = 0$ to the final representation at $d = 1$. We defined the *middle window* as ViT blocks 5–10 ($d \in [0.42, 0.83]$) and the ResNet blocks at the same relative depth (`layer3.0`–`layer4.0`).

Object recognition. We classified every view of each object with both models and recorded the top-1 ImageNet label. From these labels, for each object, we computed its *stability*, S, or the fraction of views that receive the object's most common label, and *category correctness*, C, or the fraction of views whose label matches the object's correct category. The mapping from labels to our five object categories was predefined, with ambiguous labels counted as incorrect (e.g. for airplanes, `airliner` and `warplane` are correct, whereas `wing` is not). An object is consider *recognised* by a model when $C \geq 0.9$.

Normalised view variation. Representation manifolds may change scale substantially across layers, so the raw spread of an object's views is not directly comparable across different layers. For object i, let h_{iv} denote the representation of view v, and let $\bar{h}_i$ be the mean representation across its 40 views. We measure its view-dependent spread as $V_i = \frac{1}{40}\sum_v \|h_{iv} - \bar{h}_i\|^2$. At the same layer, we also define a reference scale to quantify the separation of the 15 objects, and $B = \frac{1}{15}\sum_i \|\bar{h}_i - \bar{h}_{\text{all}}\|^2$,where $\bar{h}_{\text{all}}$ is their grand mean. We then define the relative view variation to be $A_i = \frac{V_i}{B}$. Thus, A_i measures how widely the views of one object spread out relative to the scale. A decrease in A_i with depth therefore indicates relative contraction, even when the raw view variance value V_i increases.

To quantify late-layer contractions, we define $\Delta A_i = A_{i,\text{mid}} - A_{i,\text{final}}$, where $A_{i,\text{mid}}$ is the mean over the middle layer window. $\Delta A_i > 0$ indicates relative contraction from the middle of the network to the final representation. We use $A_{i,\text{final}}$ separately as a measure of final view-set compactness, where smaller values indicate a more compact final representation.

Margins and readout For an object tagged with ImageNet labels $\mathcal{C}$, the category margin of a single viewpoint v is denoted as $m_v^{\text{cat}} = \max_{c\in\mathcal{C}} z_c(h_v) - \max_{j\notin\mathcal{C}} z_j(h_v)$, which is positive when the view is classified into the correct category. We define the centre margin m_{loc} by evaluating the same measure with the mean for representation $\bar{h}$, which is essentially where the centre of the whole view set sits relative to the category decision region.

To test how strongly viewpoint variation is influencing the readout along the readout direction, we let a and r be the top choice class pick and runner-up choice at $\bar{h}$, and $u = w_a - w_r$ the readout direction. We then define $\kappa = \frac{\text{sd}_v[u^\top(h_v - \bar{h})]}{\sqrt{V}}$, as the variation along this decision direction with respect to the spread of the view set. Smaller κ means that the viewpoint variation imposes less influence on this pairwise decision margin. Section 6 compares κ with a random-orientated baseline.

Statistics. All analyses are performed at the level of exemplar objects ($n = 15$). The models, representation extraction, recognition criterion, and bootstrap procedure were determined a priori. Behavioural correlations and alignment tests were performed post hoc. We report Spearman cor-

relations and means with 95% percentile bootstrap intervals over objects, from 10,000 and 50,000 draws, respectively.

4 Recognition Under Viewpoint Variation

Recognition correctness. Both models showed substantial variation in recognition across azimuth rotations, with a mean category correctness of $C = 0.39$ (95% CI [0.20, 0.59]) for the ViT-B/16 and $C = 0.46$ (95% CI [0.26, 0.66]) for the ResNet-50. Object-level correctness was strongly correlated across the two networks ($\rho = 0.80$, 95% CI [0.47, 0.96]). Of the 15 objects, the ViT recognised 2 (airplane 12 and airplane 13, both $C = 1.00$) and the ResNet recognised 5 (chair 00, $C = 0.95$; car 03, $C = 0.93$; car 04, $C = 0.90$; airplane 12 and airplane 13, both $C = 1.00$), generally indicating a high variation in recognition under different viewpoints.

Stability without correctness. Stability varied considerably across objects in both models and was largely dissociated from category correctness (Figure 2c,e). Both models classified chair 02 and lamp 11 unstably and incorrectly, and airplane 12 and airplane 13 stably and correctly, whereas the ResNet classified car 04 correctly but unstably, under several labels ($C = 0.90$, $S = 0.60$). Notably, both models labelled table 07 as *tripod* at all 40 views, making it perfectly stable but never correct ($S = 1.00$, $C = 0$). We verified that this was not the default response of the models, but an outlier. No view of any other object was labelled *tripod*, and classifying the mean representation across all 15 objects yielded the label *table lamp* rather than *tripod*, with table 07's mean view representation lies farther from this overall mean in both models. For this object, the ViT assigned high confidence to the incorrect label (mean top-1 probability 0.70), whereas the ResNet did not (0.13).

5 Spread: Relative Collapse Across Depth

Absolute and relative view variation. In both models, absolute view variance V increased with fraction depth d for every object. From $d = 0$ to $d = 1$, V grew by a factor of 3.7×10^3 to 4.9×10^4 across objects in the ViT and 4.7×10^2 to 2.7×10^4 in the ResNet, and between-object variance B showed similar increment. Since both V and B grew with depth, collapse of an object's view representations is best measured in relative terms via its relative view variation A. For example, in the ViT, V for airplane 13 rose from 0.32 at block 5 to 38.8 at the final representation, while B rose from 1.18 to 207.8, hence A decreased from 0.27 to 0.19, indicating a relative contraction of view representations (Figure 2a).

ViT recognise with contraction. In the ViT, objects' view representation generally showed no net contraction ($\overline{\Delta A} = -0.010$, 95% CI [$-0.055$, 0.033]), with 8 objects contracting and 7 expanding. Yet, the two objects recognised by the ViT, namely airplane 12 and airplane 13, showed strong signs of contraction ($\Delta A = +0.134$ and $+0.169$ against a median of -0.016 for the other 13 objects). For both objects, A peaked at block 9 (0.43) and roughly halved by the final representation (0.21 and 0.19, respectively; Figure 2b).

ResNet recognise without contraction. In the ResNet, objects' view representations showed large expansion ($\overline{\Delta A} = -0.137$, 95% CI [$-0.227$, -0.077]). Unlike in the ViT, its recognised objects were not accompanied with contraction. Car 03 and car 04 reached their largest A at the final representation ($\Delta A = -0.128$ and -0.183, respectively), and chair 00 ended within 6% of its peak, yet all three were recognised. Car 04, for instance, was classified correctly at 36 of 40 views, although its A_{final} ranked only tenth of 15 objects. The two recognised airplanes did contract, but only after the middle window. In both, A peaked at the first block of stage 4 and fell by 63% and 55%, respectively, by the final representation, while raw view variance V decreased over the last block (from 15.4 to 7.1 and from 14.6 to 8.6). Because this peak lies at the end of the middle window, A_{mid} and A_{final} were similar, and ΔA remained close to zero ($+0.012$ and -0.013).

Collapse and compactness are not sufficient for recognition. Collapse of view representations is not sufficient for recognition. In the ViT, six of the eight objects with $\Delta A > 0$ were not recognised (chair 00, chair 01, table 06, table 07, lamp 09 and lamp 10), although their collapse ($+0.012$ to $+0.045$) was much smaller than that of the recognised airplanes ($+0.134$ and $+0.169$). In the

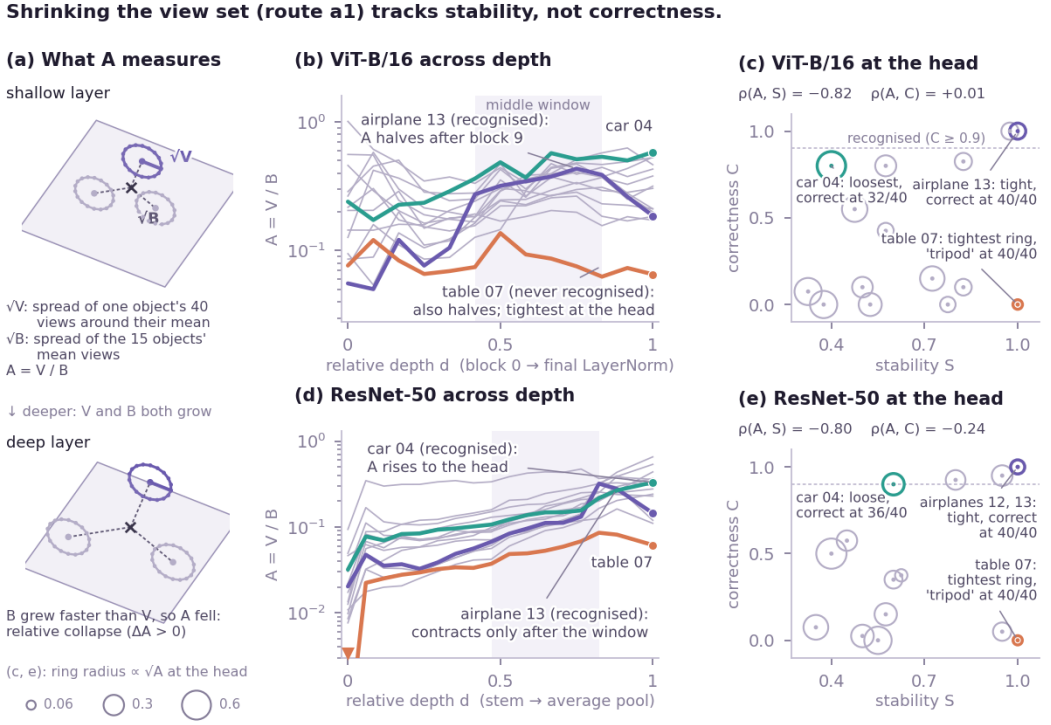

Figure 2: **Collapsing of view representations (route a1) tracks stability, not correctness.** (a) Geometric interpretation of relative view variation A; a decrease in A indicates a collapsing view representations relative to object spacing. (b, d) A (log scale) for all 15 objects as a function of relative depth in the ViT (b) and the ResNet (d). Shading marks: middle window. In the ViT, A falls sharply from its peak for both airplane 13 and table 07, yet only airplane 13 is recognised. In the ResNet, car 04 and airplane 13 are both recognised, yet A for car 04 rises continuously, whereas A for airplane 13 collapses after the middle window. (c, e) Stability S against category correctness C at the final representation. Ring radius is proportional to $\sqrt{A_{\text{final}}}$, where smaller rings indicate more compact view representations. In both models, S is strongly correlated with A_{final} ($\rho = -0.82$ in the ViT and -0.80 in the ResNet), whereas C is not. Table 07 is the most compact and stable object in both models, yet it is never classified correctly.

ResNet, table 08 showed the largest collapse ($\Delta A = +0.050$), yet only 6 of its 40 views were classified correctly. Likewise, compactness is not sufficient for recognition. In both models, table 07 had the most compact final view representations ($A_{\text{final}} = 0.066$ and 0.062), yet it was never classified correctly. In the ViT, this compactness arose through collapse of representations, with A halving from a peak at block 6 ($\Delta A = +0.023$); in the ResNet, such compactness was present from shallow depth ($A = 3 \times 10^{-4}$ at $d = 0$), likely due to the object's inherent symmetry. In both cases, the readout head receives a tight view set on the wrong side of the decision boundary.

Compactness tracks stability, not correctness. Across objects, the compactness of the final view representations was strongly correlated with stability ($\rho(A_{\text{final}}, S) = -0.82$, 95% CI $[-0.95, -0.48]$ in the ViT and -0.80, $[-0.95, -0.49]$ in the ResNet), but not reliably with correctness ($\rho(A_{\text{final}}, C) = +0.01$, $[-0.62, +0.58]$ in the ViT and -0.24, $[-0.80, +0.36]$ in the ResNet). Relative collapse showed a similar link with stability in both models, with $\rho(\Delta A, S) = +0.82$ and $+0.70$, but a moderate correlation with correctness in the ViT ($\rho(\Delta A, C) = +0.52$, $[-0.08, +0.86]$), driven mostly by the two recognised airplanes, and a weak correlation with correctness in the ResNet ($+0.26$, $[-0.32, +0.76]$).

6 Orientation: Viewpoint Variation Is Not Turning Away from the Readout

Alignment against a random-orientation baseline. Test a2 predicts that viewpoint variation should be less visible to the classifier head than a randomly oriented one does. Therefore, we compared κ with the random-orientation baseline $\kappa_{\text{null}} = \frac{\|w_a - w_r\|}{\sqrt{p}}$. and to our surprise, we observed alignment way exceeding the random baseline for all 15 objects in both networks (Figure 3a), with median $\kappa/\kappa_{\text{null}} = 2.80$ in the ViT and 3.99 in ResNet-50. We see the networks do not suppress viewpoint variation by orienting it away from the readout but instead express it more strongly along the relevant decision direction than the random baseline. Although only about 1% of the total view variance is visible along this direction, this small value is expected in a very high-dimensional representation and still substantially larger than the random baseline.

Rotating a recognised view set. To verify this result directly, we randomly rotated the view set around its centre while keeping its mean and all pairwise distances untouched to only alter its orientation relative to the classifier head. For airplane 13 in the ViT, we applied 32 such rotations to the 64 final-layer representations and all rotations remained correctly classified, while the minimum category margin increased from 1.67 in the native representation to 2.36–2.94 after rotation (Figure 3b).

The original viewpoint set was also more strongly aligned with relevant decision directions than almost every rotated version. Its readout gain κ compared to the runner-up category*wing* ones was 2.3 times the rotation mean and exceeded 31 of 32 rotations against *airship* which was the closest rival at the worst view, and it was 5.0 times the rotation mean and exceeded all 32. The same pattern held true across all repeats that the native κ exceeded the rotation mean for every object in both networks. Thus, recognition does not need to rely on rotating viewpoint variation away from the classifier readout. The trained representations show stronger viewpoint variation to the relevant decision directions than random.

7 Location: The View Set Relative to the Decision Boundary

Centre margin and correctness. Correctness of recognition is strongly associated with where the view set is with respect to the margin. The centre category margin m_{loc} correlates with correctness at $\rho = 0.95$ in the ViT and $\rho = 0.97$ in ResNet-50. The recognised objects have the largest positive centroid margins, whereas compactness does not distinguish recognised from unrecognised objects. For example, table 07 has the most compact final view set in both networks but a strongly negative centre margin, and is never correctly classified. A positive centre margin is not sufficient for every view to be recognised correctly, while a sufficiently scattered set of viewpoints can still go cross the boundary.

Ablations work mainly through location. Airplane 13 provides a direct test of whether recognition depends on changes in the viewpoint set geometry or on where the set is located. In the ViT representation, views from the two opposite side become extremely close, and selective attention and MLP ablation destroy both the geometry and the recognition. We separated these effects by moving the intact set to the ablated centroid that changes only location, and moving the ablated set back to the original centre restores location while preserving the ablated geometry. Restoring recovers all 64 views in every ablation condition, even when the opposing-view distance remains substantially altered. On the other hand, moving the intact geometry to the ablated centre reproduces the recognition loss. The strongest case is removal of the block-0 MLP, which moves the centroid across the *airliner–wing* boundary and yields 0/64 correct views; moving the ablated set back to its original centroid restores 64/64. These counterfactuals indicate that the recognition failures arise primarily from changes in location rather than from altered geometry inside the set.

8 What Distance-Based Measures Can and Cannot Tell us

What can distance measures tell Sections 5–7 measured the three factors that decide what a fixed head does with an object's views: spread, orientation and location. A distance-only summary is any

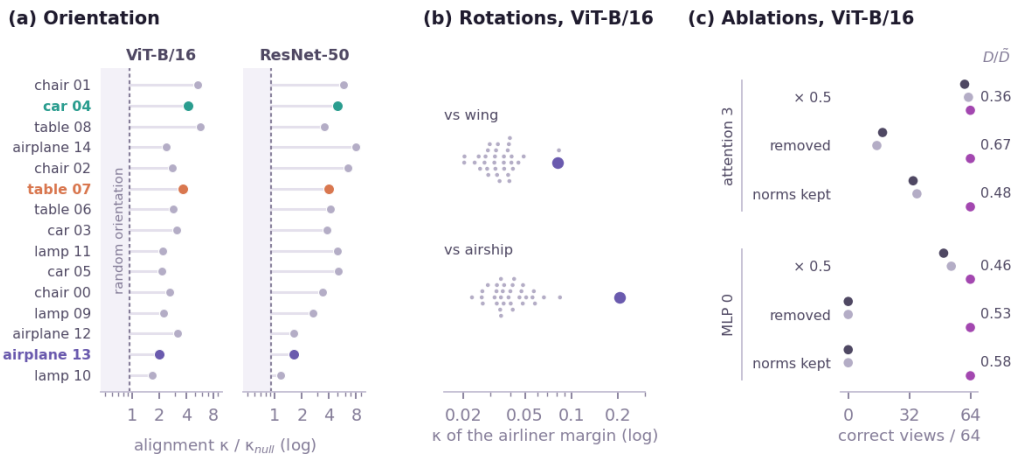

Figure 3: **Testing the orientation and rotation influence.** (a) Alignment $\kappa/\kappa_{\text{null}}$ of each object's view set with its readout direction u (top class against runner-up class at the mean view), at the final layer of ViT-B/16 and ResNet-50; 15 objects, 40 viewpoints each. Dashed lines: random orientation baseline of the same view set (0.93 and 0.90). Shaded region denotes where turning view variation away from the readout (route a2) would place an object. (b) ViT-B/16, airplane 13 with 64 views: κ of the original view set (violet) and of 32 random rotations of the centred set (grey) compared against the runner-up class*wing* and *airship*, the closest rival at the worst viewpoint in log scale. The original value surpasses almost all rotations(31 out of 32). (c) ViT-B/16, airplane 13 with 64 views, after the block-3 attention update or the block-0 MLP update is halved ($\times 0.5$), removed, or removed with post-residual norms restored (norms kept). Dark: correct views after ablation. Grey: intact set moved to the ablated centre (location only). Orchid: the ablated set moved back to the intact centre (geometry only). Right: normalised distance $D/\tilde{D}$ between the two side views (intact 0.27).

quantity computed from the pairwise distances between the views alone, such as view variance, intrinsic dimension or curvature estimates. These distance summaries have been widely associated to accuracy across networks (Ansuini et al., 2019). Pairwise distances determine a view set up to an isometry because they only fix the spread and shape but not where it sits or how it is oriented relative to the head. Therefore, A distance-only summary sees route a1 of Figure 1a, not a2 or a3. Relative variation A belongs to this class when all view sets move together; measures that depend on where sets sit relative to one another, such as manifold capacity, do not. Proposition 1 states the consequence.

Proposition 1. *Let the classifier head be affine. (a) If the weight vector of some rival class lies outside the convex hull of the compatible classes' weight vectors, every view set, including one recognised at every view, has a translate on which no view is named correctly. (b) Let a and r be the two top classes at the centroid. If the distance between their weight vectors, times the largest distance of any view from the centroid, exceeds the margin of a over r at the centroid, some rotation about the centroid gives at least one view a label other than a. In both cases, every distance-only summary takes the same value on the moved set as on the original.*

Translations preserve geometry but not correct recognition One exposed limitation of distance-only measurement is that translating a viewpoint set leaves every pairwise distance unchanged, but can move the entire set across the classifier's decision regions, which could be proven against with the location test. Moving airplane 13's whole set to the centre produced after removing the block-3 attention update preserves its geometry, yet reduces recognition from 64/64 to 15/64 views. We see that even complete knowledge of the pairwise distance geometry alone cannot determine whether a view set is recognised by a fixed classifier.

Orientation is invisible to distance geometry. Distance geometry also does not specify how a viewpoint set is oriented relative to the classifier head. Rotating a centred set around its centre preserves every pairwise distance but changes its projection onto the readout directions. For airplane 13, the 32 random rotations leave the intrinsic geometry unchanged but greatly change the readout gain

and decision margins (Section 6). More generally, Part (b) shows that, when the view set is sufficiently large relative to the centre margin, a change in the orientation alone can move a view across a decision boundary and shift recognition result.

Together with the translation result, this identifies two pieces of information that distance-only summaries ignore related to the task, which are 1) where the view set sits, and 2) how its variation is oriented against the readout. These geometric summaries should therefore be interpreted alongside such readout-relative quantities when the aim is to explain recognition rather than to describe representation geometry alone.

9 Limitations and Conclusion

Limitations. Fifteen objects are good for existence cases, but the correlation intervals span about ± 0.6 and rule out only strong associations; several effects run mostly through the two planes while scaling could give us more. The two networks differ in architecture and training recipe. The stimuli are synthetic renders on grey from one low elevation, and cars and end-on airplanes are small in the frame, without the influence of illumination or background colours.

Conclusion. We ask whether deep vision networks achieve invariant recognition by collapsing an object's view manifold, or by reorganising it. In the two networks studied here, relative collapse is neither necessary nor sufficient for recognition in both networks, and compactness tracks how consistently a network answers rather than whether it is right. Neither network turns view variation away from its readout. The correctness of recognition is more aligned with where the view set lies relative to the decision boundary. For one recognised airplane, ablations that break recognition act by moving the view set instead of by changing its shape.

AI use statement

We used generative AI tools (Claude, ChatGPT, Kimi, Gemini) for: feedback on experimental design and protocols; writing and running analysis code; checking results against saved outputs; drafting and polishing text; and finding references. All experiments, results and references were checked by the authors. We have reviewed all AI-assisted work and take full responsibility for the content of this paper, including text, claims and outcomes produced with the aid of generative AI.

Reproducibility statement

Section 3 specifies the stimuli, checkpoints, preprocessing, layers, measures and statistics. The supplementary material contains the object list and rendering settings, the frozen category mapping, the written protocol, the scripts that produce every figure and number, and the saved features and logits from which all analyses are computed without further forward passes.

References

Amro Abbas and Stéphane Deny. Progress and limitations of deep networks to recognize objects in unusual poses. In *AAAI Conference on Artificial Intelligence*, volume 37, 2023. doi: 10.1609/aaai.v37i1.25087.

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. URL https://jmlr.org/papers/v19/17-646.html.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Michael A. Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Yehonatan Avidan, Daniel D. Lee, and Haim Sompolinsky. Sparse prototype code underlies classification and prediction across modalities. *arXiv preprint arXiv:2608.15632*, 2026. URL https://arxiv.org/abs/2608.15632.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/97af07a14cacba681feacf3012730892-Abstract.html.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. URL https://arxiv.org/abs/1512.03012.

SueYeon Chung and L. F. Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, 70:137–144, 2021. doi: 10.1016/j.conb.2021.10.010.

SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018. doi: 10.1103/PhysRevX.8.031003.

Uri Cohen, SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, 11(1):746, 2020. doi: 10.1038/s41467-020-14578-5.

Mohammadreza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.16156.

James J. DiCarlo and David D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, 2007. doi: 10.1016/j.tics.2007.06.010.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/0c0bf917c7942b5a08df71f9da626f97-Abstract.html.

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. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.11929.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Ian Goodfellow, Honglak Lee, Quoc V. Le, Andrew Saxe, and Andrew Y. Ng. Measuring invariances in deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 22, 2009.

Hangfeng He and Weijie J. Su. A law of data separation in deep learning. *Proceedings of the National Academy of Sciences*, 120(35):e2221704120, 2023. doi: 10.1073/pnas.2221704120.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1903.12261.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2733–2743, 2019. URL https://aclanthology.org/D19-1275/.

Ha Hong, Daniel L. K. Yamins, Najib J. Majaj, and James J. DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, 19(4):613–622, 2016. doi: 10.1038/nn.4247.

Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022. URL https://arxiv.org/abs/2202.08384.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Henry Kvinge et al. In what ways are deep neural networks invariant and how should we measure this? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Achleshwar Luthra, Tianbao Yang, and Tomer Galanti. Self-supervised contrastive learning is approximately supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=mf4V1SK0np. arXiv:2506.04411.

Achleshwar Luthra, Yash Salunkhe, and Tomer Galanti. Directional neural collapse explains few-shot transfer in self-supervised learning. *arXiv preprint arXiv:2603.03530*, 2026. URL https://arxiv.org/abs/2603.03530.

Spandan Madan, Timothy Henry, Jamell Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Frédo Durand, Hanspeter Pfister, and Xavier Boix. When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence*, 4, 2022. doi: 10.1038/s42256-021-00437-5.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1–40, 2020. URL https://jmlr.org/papers/v21/20-345.html.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A. Poggio. Feature learning in deep classifiers through intermediate neural collapse. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28729–28745, 2023. URL https://proceedings.mlr.press/v202/rangamani23a.html.

Albert J. Wakhloo, Tamara J. Sussman, and SueYeon Chung. Linear classification of neural manifolds with correlated variability. *Physical Review Letters*, 131(2):027301, 2023. doi: 10.1103/PhysRevLett.131.027301.

Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, 2019. URL https://proceedings.mlr.press/v97/zhang19a.html.