DOES LEARNED FEATURE GEOMETRY PREDICT DETECTION RELIABILITY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Two ways of measuring geometry in a vision model have never been brought into one frame: the Zernike spectrum of a learned feature, and the localization error of a detector. This paper constructs the quantity that joins them. A detector already reports a displacement in units of the size of the thing that moved, the offset between a predicted box and its ground truth divided by the object's scale; we build the feature-side counterpart, the offset between an activation cluster and the same cluster recovered from an independently seeded retraining, divided by the cluster radius. Both are dimensionless and vary smoothly where IoU can flip on one pixel. The correspondence is constructed, not analogous: cross-seed clusters are matched by the harness's own assignment, cost and threshold, and one ResNet-50 trunk supplies both sides, trained as a classifier under eight seeds and then frozen inside a Faster R-CNN detector. The apparatus carries its own calibrations, and on RarePlanes a positive control separates one feature-side statistic from chance in the fixed-weight case, the cross-seed match rate, in 13 of 24 cells against 1.2 by chance; across retrainings it returns the chance value, in the shared canvas frame and after rigid registration, which lifts it fourfold for real pairs and for random positions alike. The displacement statistic the bridge rests on does not resolve the fixed-weight case, so its correlations cannot be read. Re-anchored to activation maps of real tiles, the dominant canvas pathologies are removed, measurement density rises by two orders of magnitude, and local feature density and spectral similarity show small associations with localization error, at $|\rho_S|$ near 0.1, partly through object size and unresolved at the tile level: suggestive, not yet predictive. The detection side is unaffected: unfreezing the trunk lifts the decomposable reliability score from 0.862 to 0.925 on non-overlapping intervals.

1 INTRODUCTION

The geometry of a learned feature representation has only recently become measurable, and to our knowledge no one has measured it in a form that a detection metric can be compared against. Two literatures sit on either side of that gap. The first asks what a network has learned, and projecting a feature visualization onto the Zernike basis turns the question into numbers: a spectrum, a centroid and a radius for every feature (Algire et al., 2026b). The second asks whether a detector can be trusted, and answers with average precision under IoU thresholds (Lin et al., 2014) or the scale-aware measures that decompose it (Algire et al., 2026c). Neither has asked whether the two geometries are the same.

What is new here. The gap closes on one observation, drawn in Figure 1. A detector already reports a displacement in units of the size of the object that moved: the distance from a predicted box's centre to its matched ground truth, divided by the square root of the object's area. The feature side reports no such quantity, so we construct one. Retrain the classifier under a new seed and its activation clusters reappear in different places; the distance a cluster moves, divided by its own radius, is the internal counterpart. Neither quantity carries a unit, and both move continuously as the geometry moves, where IoU reads a ratio of areas against a threshold and can flip on one pixel (Algire et al., 2026a; Yu et al., 2020). The descriptor itself was built to characterize one trained model; its variation across retrainings has never been treated as a measurable quantity, and never in units that a detection metric shares. *Predict* is used in one sense throughout: a feature-side

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.

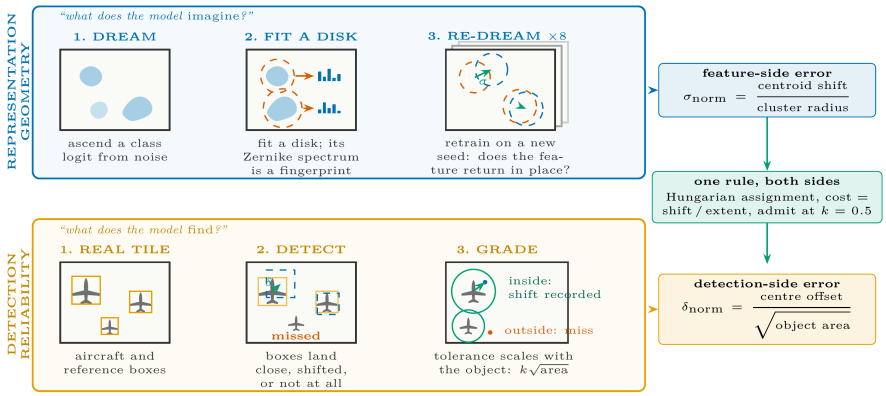


Figure 1: The two lanes and the one rule. Each half reduces a geometry to a displacement in units of the size of the thing that moved, and both use the same assignment, cost and threshold to decide what is compared. The lower lane is the detection harness introduced in (Algire et al., 2026c), unchanged; one ResNet-50 trunk supplies both (§3.2).

quantity, measured without any detector output, that is monotonically associated with the per-object localization error of a detector sharing the same trunk, reported as a rank correlation with its interval, with the tile as the unit of inference, and tested on held-out tiles in §5.5.

Stabilities, not positions. Nothing guarantees that a feature lands at the same place on two seeds' canvases, so the measure is not a position but a stability: vary only the seed, match each feature to its counterpart, and measure how far the pair moved relative to its own extent. The displacement carries no canvas unit; the matching that fixes the pair does use canvas position, and §5.3 tests whether that dependence hides anything.

Identical procedure, and one trunk. Cross-seed cluster matching reuses the detection harness's own cost, threshold and solver (§4.3). And because a correlation between internal and external geometry means little when the two are read off different weights, one ResNet-50 trunk serves both roles: trained as a classifier under eight seeds, then transferred unchanged into a Faster R-CNN R50-C4 detector, frozen in arm F so both geometries come from one set of weights and released in arm U to bound how far fine-tuning moves them.

Contributions. Varying the seed turns the feature side into a dispersion rather than a point estimate, which is what makes it a candidate predictor, and the contrast we report is whether that dispersion tracks the harness's decomposed axes more closely than it tracks AP_{50}. The paper contributes, in the order it builds them, the stability measure (§4.4); a matching protocol that makes the two sides one procedure rather than two analogous ones (§4.3); an identity bridge in which a single trunk supplies both geometries (§3.2); and a calibration suite, run on every execution, that decides whether the measure resolves feature re-localization at all. §5 is the first reading: on RarePlanes the measure returns chance on each of its nulls, and a positive control separates the one feature-side statistic that carries a reading from those that do not.

2 Related Work

Describing what a network has learned. Feature visualization synthesizes an input that maximally excites a unit or a class logit (Simonyan et al., 2013; Yosinski et al., 2015; Mahendran & Vedaldi, 2016; Olah et al., 2017), and its output is an image, so the analysis it supports is usually qualitative; network dissection makes it quantitative by aligning units to a labelled concept vocabulary (Bau et al., 2017), but only for concepts the vocabulary already contains. A basis with physical meaning avoids both limits: Zernike polynomials are the standard language for wavefront aberration in optics (Zernike, 1934; Noll, 1976), orthogonal on the unit disk with rotation-invariant magnitudes (Teague, 1980; Khotanzad & Hong, 1990), and on activation structure they yield a com-

pact spectrum for every feature (Algire et al., 2026b). We take that descriptor as given and ask what its variance predicts.

Do independently trained networks learn the same thing? Networks trained from different initializations learn largely overlapping features, arranged differently and only partially matched (Li et al., 2016), and the seed alone moves final accuracy by margins comparable to many reported architectural gains (Summers & Dinneen, 2021; Picard, 2021). Similarity indices measure how close two networks are while staying invariant to nuisance transforms (Raghu et al., 2017; Kornblith et al., 2019), but each returns one scalar per layer pair and integrates the spatial structure away. We keep that structure, and ask which feature moved how far relative to its own radius.

What detector metrics measure. Two properties make AP a poor target for a geometric predictor: the discontinuity described in §1, where near the IoU threshold a displacement of a pixel or two changes the score by a finite amount (Algire et al., 2026a; Rezatofighi et al., 2019; Yu et al., 2020); and dilution, since AP folds ranking, classification and localization into one number, so a predictor acting on localization alone is masked by the other two, and decomposing the score is the known remedy (Oksuz et al., 2018). The harness we adopt decomposes it for small objects: admissibility follows centre distance normalized by object size rather than IoU, recall is averaged over size octaves, and a position score comes from the median normalized shift, the three combining into a Decomposable Reliability Score (Algire et al., 2026c). We use it unchanged as the target side and reuse its matching rule on the feature side.

3 Experimental Setup

The pipeline runs in twenty-one stages, from data preparation through classifier training, gradient ascent, cluster geometry, detector training, evaluation and the bridge, to the calibration stages of §C. Every value is pinned in one configuration module and stamped into a provenance sidecar beside each artifact. Because the measured quantity is a spread across seeds, the eight seeds are pinned rather than sampled and each run is deterministic on its own, so a variance is reproducible; the reproducibility statement gives the settings and four correctness invariants that are enforced in code rather than trusted.

3.1 Data and the shared trunk

RarePlanes (Shermeyer et al., 2021) supplies the imagery in its real, tiled PS-RGB form at 512×512 px with the publisher's tile-space COCO annotations. Its civil-size scheme gives three foreground categories, the Small, Medium and Large Civil Transport/Utility roles, shortened to small, medium and large throughout. §A.1 covers the annotation and split handling, including a train/test location overlap that scopes detector results to official-split performance rather than to unseen airports.

One ResNet-50 (He et al., 2016) is trained as a multi-label tile classifier over those three categories, at native tile resolution, with layer4 dilated to keep an effective stride of 16. Hyperparameters follow the published Zernike feature-descriptor configuration (Algire et al., 2026b) and are listed in §A.4, so the cluster geometry analyzed here is the published one rather than a re-tuned variant. Training repeats for seeds $\{0, \ldots, 7\}$, and nothing else varies, with one documented deviation: that study's tile configuration defaults to focal loss, whereas we use its class-weighted binary cross-entropy branch.

3.2 Two detector arms

Detection uses Faster R-CNN (Ren et al., 2015) in the R50-**C4** profile under Detectron2 (Wu et al., 2019), not the more common FPN profile, because C4's backbone is exactly conv1 through res4 while the classifier's dilation is in res5. The trunk is therefore architecturally identical in both roles, with no dilation conflict and no surgery beyond a name mapping.

Arm F freezes the trunk through res4, so only the RPN and the res5 RoI head train. Cluster geometry and detection geometry then come from identical weights, which makes this the identity bridge: a correlation observed here cannot be attributed to the two sides describing different features. **Arm**

U starts from the same initialization but leaves the trunk free, and measures how far detection fine-tuning displaces the geometry. Both train from the seed-0 trunk and are evaluated at the harness's stated operating point (Algire et al., 2026c); arm **F** is additionally trained from each of the other seven trunks for the test of §6. Settings in §A.4.

3.3 Gradient-ascent canvases

Per-class canvases are optimized at 224×224 for 800 image-update steps under a cosine schedule with annealed total-variation and ℓ_2 regularization and periodic Gaussian smoothing (§A.4); without them, gradient ascent leaves artifacts the cluster detector reads as spurious clusters. Three classes by eight seeds gives 24 canvases.

4 The Geometric Bridge

4.1 Feature side: clusters, centroids, radii

Images are indexed (y, x) from the top left, $\|\cdot\|$ is the Euclidean norm, and $\mathrm{med}(\cdot)$ the median. A class canvas is reduced to a saliency map, thresholded at five percentiles, and split into connected components; §A.1 gives the parameters. A surviving component with member pixels $\{(y_i, x_i)\}$ and saliency weights w_i has centroid $(c_y, c_x) = (\sum_i w_i y_i, \sum_i w_i x_i)/\sum_i w_i$ and radius

$$r = \max\Big(P_{85}\big(\|(y_i, x_i) - (c_y, c_x)\|\big),\ r_{\mathrm{floor}}\Big), \qquad r_{\mathrm{floor}} = 15 \text{ px}. \tag{1}$$

Centroid and radius are the geometric half of the published descriptor, computed from the saliency map before any moment is taken; the spectrum of §4.2 is its other half, computed on the patch they define. The matching of §4.3 and the measure of §4.4 use only the geometric half. The floor is a measurement artifact and is treated as one: every row at the floor is flagged and every analysis involving r is reported with those rows and without, since a censored radius in the denominator of (3) would compress the quantity this paper measures.

4.2 Feature side: the local Zernike spectrum

Each cluster contributes a square patch of half-width $p = \lceil r \rceil$, centred on its centroid. We convert the patch to grayscale, normalize it to $[0, 1]$, and map it onto the unit disk, where the pixel-sum estimator of (6) projects it onto the Zernike basis of (5). Only the rotation-invariant magnitudes $|\widehat{Z}_n^m|$ for $n \le 8$ are retained, 45 values per cluster, since a patch a few tens of pixels across aliases before it resolves higher orders. The disk normalization discards absolute scale, so the radius of (1) survives as separate metadata, the denominator (3) needs.

Two quantities derived from the spectrum enter the bridge as predictors. Spectral dispersion D_{spec} is the mean across seeds of the coefficient of variation of the radial energy $E_n^{(q)}$ at the declared exponent $q = 2$; spectral identity S_{ab} is the cosine similarity between the magnitude spectra of two matched clusters, reported with and without the $(0, 0)$ component. §A.3 defines both in the canonical notation of the corpus standard, and §A.2 gives the index-set convention behind the count of 45, together with four recorded deviations.

4.3 Matching features across seeds

Fix a class c and an ordered seed pair (t, u). Let A be the clusters found on seed t's canvas and B those on seed u's. Assign A to B by Hungarian assignment (Kuhn, 1955) with cost

$$\mathrm{cost}(i, j) = \frac{\|(c_y^i, c_x^i) - (c_y^j, c_x^j)\|}{\bar{r}_{ij}}, \qquad \bar{r}_{ij} = \tfrac{1}{2}(r_i + r_j), \tag{2}$$

admitting a pair only when $\mathrm{cost}(i, j) \le k_{\mathrm{match}}$.

The normalized centre-distance cost, the admissibility threshold and the Hungarian solver are not chosen to resemble the detection harness; they are the harness's own rule, with $k_{\mathrm{match}} \equiv k = 0.5$, applied to a different substrate. That identity is the methodological core of the paper.

	Feature side			Arm F (frozen)				Arm U (unfrozen)			
Class	M_c	med σ_{norm}	D_{spec}	med δ_{norm}	P_{score}	$\bar{R}$	AP_{50}	med δ_{norm}	P_{score}	$\bar{R}$	AP_{50}
Small	0.142	0.304	0.875	0.083	0.835	0.741	0.822	0.038	0.924	0.887	0.930
Medium	0.117	0.356	0.885	0.051	0.897	0.811	0.830	0.026	0.948	0.870	0.890
Large	0.134	0.312	1.011	0.061	0.877	0.647	0.722	0.029	0.943	0.856	0.912

Table 1: Both sides of the bridge by civil-size class (`bridge/cluster_class_stats.csv`, `detection_class_stats_{F,U}.csv`). Decomposable Reliability Score: F 0.862 (bootstrap 0.841–0.883), U 0.925 (0.908–0.939). Censored-radius fraction 0.68, 0.77, 0.77; large-class $\bar{R}$ rests on truncated support (§B.1).

Matches over all $\binom{8}{2} = 28$ seed pairs are then stitched into tracks: edges merge in increasing order of cost, only between components whose seed sets are disjoint, and a track must span at least 3 of the 8 seeds before it enters the stability statistics. The match rate M_c, the fraction of eligible pairs actually matched, is itself a stability statistic, since a feature that never reappears is maximally unstable and dropping it without comment would bias (3) toward the features that happened to survive.

4.4 The two locality errors

For a matched cluster pair (a, b), the cross-seed centroid shift is

$$\sigma_{\text{norm}}(a, b) = \frac{\|(c_y^a, c_x^a) - (c_y^b, c_x^b)\|}{\bar{r}_{ab}}. \tag{3}$$

For a matched detection and ground-truth pair, the harness's normalized center shift is

$$\delta_{\text{norm}} = \frac{\|(c_x^{\text{det}} - c_x^{gt},\ c_y^{\text{det}} - c_y^{gt})\|}{s_{gt}}, \quad s_{gt} = \sqrt{w_{gt} h_{gt}}. \tag{4}$$

The two expressions have the same form. At the median of this run they read $\sigma_{\text{norm}} = 0.313$ on a cluster of mean radius 21.3 px and $\delta_{\text{norm}} = 0.062$ on a box of median $\sqrt{A_{\text{ref}}} = 49.4$ px. The form is shared; the measurand is not. Equation (3) is a dispersion across retrainings and (4) an error against an annotation, an asymmetry forced by training one detector from one trunk, and §6 names the experiment that removes it. Equation (3) is dimensionless, since the displacement is taken in units of the feature's own radius, but it is not frame-invariant: the displacement is a distance in the shared canvas frame, and (2) fixes the pair on canvas position. §5.3 tests that dependence by registering the canvases before matching.

The remaining harness axes are used unchanged. Two appear throughout the results: octave-balanced recall $\bar{R}$ and the position score $P_{\text{score}} = \max(0, 1 - \text{med}(\delta_{\text{norm}})/k)$, whose harmonic mean with class accuracy is the Decomposable Reliability Score (§A.4).

5 Results

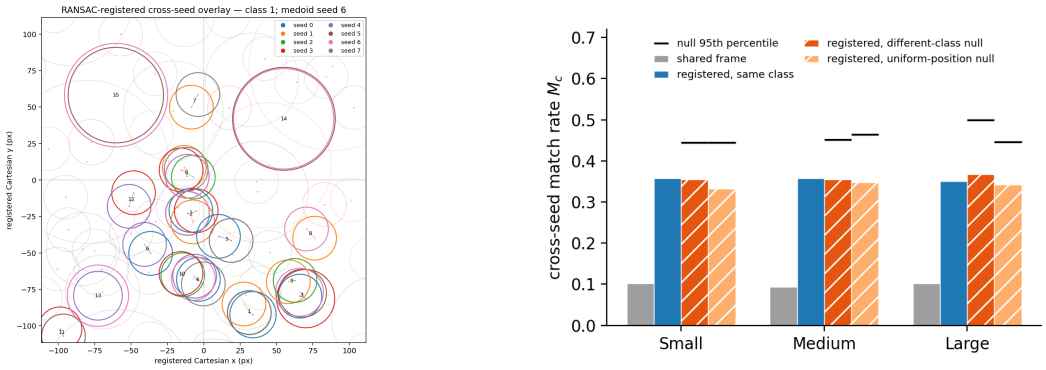
The reported run completed every stage on the official split: eight classifier seeds, 24 class canvases, 416 clusters, two detector arms, 2710 test tiles. §5.2 asks whether the feature-side columns carry a measurement at all, and every number after it is read under that answer.

5.1 Stability and reliability, side by side

Table 1 carries both sides. Its feature-side columns keep two questions apart: how often a feature reappears, M_c, and how far it moved when it did, σ_{norm}.

The detection side behaves as a detector should: releasing the trunk improves every axis in every class, with octave-balanced recall rising by 0.06 to 0.21 and the reliability score moving from 0.862 to 0.925 on non-overlapping intervals.

The feature side is thinner than the design anticipated (Figure 2): 416 clusters yield 162 admissible cross-seed matches and 30 tracks spanning at least 3 of 8 seeds, at match rates between 0.117 and



(a) (b)

Figure 2: The feature side in one frame. **(a)** All clusters of the Medium class from all eight seeds, after each canvas is registered to a medoid seed by the rigid transform of §C.2; bold circles are the sixteen registered components. **(b)** Cross-seed match rate in the shared frame and after rigid registration, under the pipeline's null search (§C.2), for the same-class pairs, for pairs from *different* classes and for uniformly random positions; bars are means over 28 seed pairs, ticks each null's 95th percentile. Registration lifts real pairs and pairs that share nothing alike (`registration_null/*.csv`).

		Rotation null, p			vs. random pairs, Δ			Pos. control	
Statistic	Favours	S	M	L	S	M	L	cells/24	Reading
M_c	higher	0.030	0.461	0.130	—	—	—	**13**	read
med σ_{norm}	lower	0.114	0.603	0.133	—	—	—	1	none
mean S_{ab}	higher	0.140	0.002	0.021	$+0.014$	$+0.014$	$+0.002$	3	none
mean S^{+}_{ab}	higher	0.950	0.085	0.769	-0.017	$+0.016$	-0.015	2	none

Table 2: The calibration ledger. *Rotation null*: one-sided p over 1,000 rotations of one seed's field about the canvas centre, on the inscribed disk (§B.3); three of twelve cells fall below 0.05 where 0.6 are expected. *Random pairs*: observed mean minus the median of 5,000 random same-class pairings. *Positive control*: of 24 model-by-class cells on fixed-weight restarts, where the weights fix the correspondence, how many beat the same null, against 1.2 by chance; the cells share eight models and three classes, so the count is read against that chance value rather than as a binomial tail (§B.3). Upstream of the spectral rows, the median intraclass correlation of the spectrum is 0.32, 0.45 and 0.26 (§C.1). Files: `bridge/bridge_null_rotation.csv`, `spectral_identity_null.csv`, `positive_control_summary.csv`.

0.142, a median σ_{norm} of 0.313, and a mean S_{ab} of 0.943 that looks like strong correspondence and is not.

5.2 The statistical indicators

Equation (3) measures stability only if the two clusters it compares really are one feature recovered twice, and its spread reflects the seed only if the seed is what moved them. Three nulls test those premises on every run and a positive control decides which statistics are entitled to a reading; Table 2 collects them.

Three nulls. The classifier's augmentation makes it approximately equivariant to the eight symmetries of a square, so any symmetry applied to one seed's canvas is an equally valid rendering of that training; rotating one field about the canvas centre by a continuous angle destroys the configuration the same way and, unlike the eight symmetries, draws without limit. Of the three cells below 0.05 in Table 2, two belong to S_{ab}, which does not separate from chance in its positive control; the third, match rate in the small class, is a draw from a broad distribution, since the 0.142 this run reports sits two standard deviations above the repeat mean of Table 4. The chance-similarity null

asks whether S_{ab} exceeds the cosine similarity of randomly paired same-class clusters, and it does so by at most 0.014; with the brightness-dominated $\widehat{Z}_0^0$ component dropped, one class clears the random-pair median by 0.016 and two fall below, so the high absolute values come from min-max normalization of similar-sized patches rather than from correspondence. The seed-variance null sits upstream of both: re-running canvas optimization ten times per seed shows that 55 to 74 percent of the variation in a typical spectral coefficient is repeat-to-repeat noise at fixed weights (§C.1); a comparison on a quantity that is half its own noise cannot separate much.

The positive control. Nulls at chance say nothing on their own, since a matcher that recovered nothing would return chance too; what supports reading them is whether the apparatus separates a correspondence present by construction, and stage 09 supplies one: ten canvas restarts at fixed weights, where the weights fix the features and only the rendering differs. Match rate beats the rotation null in 13 of 24 model-by-class cells where 1.2 are expected; the other three statistics return chance. The feature side therefore splits. D_{spec}, S_{ab} and S_{ab}^{+} rest on coefficients that are at least half instrument noise and do not separate the fixed-weight case, so nothing rests on them. Equation (3) is reproducible across repeats (Table 4) and still does not separate it, so what it resolves at 416 clusters is undetermined, and the ordering of three classes by it carries nothing either way, since the whole between-class spread it ranks on, 0.304 to 0.356, is about one repeat standard deviation. The match rate alone is read: it separates the fixed-weight case, and across seeds it returns the chance value.

5.3 Correspondence under rigid registration

That reading rests on a premise the rule of (2) takes for granted, that two seeds' canvases share a frame; the classifier is approximately rotation-equivariant, so a feature that recurs at a rotated position is one the shared-frame rule cannot see. Stage 16 removes the premise. For each of the 28 seed pairs it searches, by RANSAC over 750 hypotheses, for the rigid transform that brings the most clusters into coincidence under a position gate of about 10 px calibrated from fixed-weight restarts and a gate on the log radius ratio (§C.2). Under it the match rate rises fourfold, no seed pair returns empty, and components span up to six of the eight seeds; Figure 2a shows the result, and it looks like correspondence. Figure 2b shows what the same search returns for pairs of cluster sets drawn from *different* classes, which share no feature, and for sets whose positions are uniformly random: the same value. The same-class pairs sit at the median of both nulls, every bootstrap interval on the difference covers zero, and none approaches the 95th percentile the pre-specified criterion required. Two internal diagnostics agree (§C.2): a reflection, which no shared feature layout should want, is chosen in about half the pairs at no change in match rate, and composing the pairwise transforms through a medoid seed returns the rate to 0.12 to 0.15. That is evidence about how features move between retrainings rather than about the instrument: at this density no population of canvas features recurs across retrainings in the shared frame or in the best rigid frame searched, the one feature-side statement this run supports.

5.4 The bridge correlations

The hypothesis as first posed is a table of Spearman correlations over the three object categories, five feature-side predictors by five detection-side targets per arm (§B.2). Three points admit only $\rho_S \in \{0, \pm 0.5, \pm 1\}$ and a two-sided permutation p no lower than 0.33; the class medians it ranks differ by about one repeat standard deviation, and ten re-runs move the headline cell across its range (§C.1). The table displays effect sizes, and nothing rests on it.

5.5 The measure on real activation maps

Whether the reading of §5.2 belongs to the canvas or to (3) itself is settled by changing the substrate and leaving the measure alone: it replaces class canvases with activation maps of real test tiles, anchors clusters to annotated objects, and applies (2) unchanged.

Two of the three nulls are then answered by construction, since real tiles share a frame and a response map is a deterministic function of weights and tile; the chance-similarity null is run by the final stage: matched pairs exceed random same-class pairings in spectral identity by 0.066 at tile-level sign-flip $p \leq 0.001$ in every class, so all three nulls are answered. The measurement also becomes far denser

and the censoring nearly vanishes: 32,811 clusters and 34,732 admissible pairs against 162, and a censored fraction under 0.02 against roughly 0.7 (§C.3).

At that density the test of §1 can be run on hundreds of objects per size octave rather than three class points: three predictors in three octaves per arm against δ_{norm}, each with a tile-clustered bootstrap interval, a per-object permutation p under Holm correction (descriptive), a partial correlation given object size and a tile-block permutation p (Table 5). At the object level the local cluster count carries the association in both arms, $\rho_S = -0.17$ and -0.16 at octaves 1 and 2 in arm F and -0.15, -0.13 and -0.12 across the three octaves in arm U, five of the six intervals excluding zero. Mean S, the spectral statistic, is -0.12 and -0.13 at octave 2 with both intervals excluding zero, the one cell that repeats across arms. Two checks bound them. Cluster count rises with size within an octave ($\rho_S = 0.33$ to 0.36); holding size fixed leaves -0.09 to -0.11 at octaves 1 and 2 and nothing at 3, while mean S carries little size and keeps -0.09 to -0.12. At the tile level, where whole tiles are permuted with their objects intact, nominal $p < 0.05$ appears only in arm U, for $\mathrm{med}\,\sigma$ and mean S at octaves 1 and 2, and no cell survives Holm. A split-half test over 500 random halves of the tiles preserves the sign of the strongest cell in 83% and 88% of splits by arm, at a median held-out ρ_S of 0.10 and 0.07 (§C.3). On real activations, feature density and spectral similarity are associated with localization error at $|\rho_S|$ near 0.1, partly through object size, consistent across arms at the object level and unresolved at the tile level: suggestive rather than predictive.

6 Discussion

Two readings of chance. An experiment of this shape returns chance either because the measurement works and the relationship is absent, or because the measurement does not resolve what it claims to; the positive control of §5.2 assigns each statistic to one case. Match rate is the first case: it uses the descriptor's centroid and radius and never its spectrum, it separates the fixed-weight case from chance, and across seeds it returns chance in every frame searched. Median σ_{norm} and the two spectral statistics are the second, so Table 3, which rests on σ_{norm}, is evidence for neither. On real activation maps (§5.5) the measure reaches the first case, and feature density and spectral similarity carry associations that are suggestive rather than predictive.

Three conditions, all addressable. The chance reading has a specific shape: the radius floor, which pins the denominator of (3) to a constant for two thirds of clusters; the sparsity, at which a wrong pairing is about as likely to be admissible as a right one; and the gradient-ascent substrate, which is not a repeatable rendering of fixed weights. Under a threshold the data itself sets, no canvas feature recurs past three of the eight seeds, and registration raises that to six only as far as random positions do (§C.2). None is intrinsic, as §5.5 shows by removing all three at once.

The crossed-trunk test. Two further limits are ceilings: the class-level p-value floor of 0.33, retired by moving the unit of analysis to the object, and the seed-0 trunk behind both arms of Table 1, which leaves their detection side without across-seed variance. Training one detector per trunk removes that, and this run carries its cheaper half, arm F from all eight trunks: an object's predicted-centre spread across trunks is a dispersion in the units of (3), paired per object on the substrate of §5.5. The run supplies its positive-control floor: trunk-0 restarts put the same-trunk box spread at a median of 0.028 against 0.042 across trunks, within twice the detector's training noise, so the dispersion is not yet a reading. Against it, $\mathrm{med}\,\sigma$ gives $\rho_S = +0.05$ pooled and $+0.11$ at octave 2 (Figure 3c); §B.1 states the remaining bounds.

7 Conclusion

The title's question has a bounded answer. On gradient-ascent canvases the one calibrated feature-side statistic, the cross-seed match rate, returns chance in the shared frame and after rigid registration, so no population of canvas features recurs across retrainings at this density and the displacement statistic the bridge was built on cannot be read. On real activation maps the dominant canvas pathologies are removed and feature density and spectral similarity show small associations with localization error, at $|\rho_S|$ near 0.1, partly through object size and unresolved at the tile level: suggestive, and not yet the prediction the title asks for. The contribution is the apparatus separating the two readings; the crossed-trunk design can decide whether the second is more than a small effect.

AI USE STATEMENT

We used generative AI tools in four ways. First, to assist with the code optimizations including post-hoc analysis stages after training including the positive control and the provenance sidecars. All such code was reviewed by the authors and is covered by an automated test suite, and numerical claims about the augmentation and resampling paths were verified against reference implementations rather than accepted on assertion. Second, to draft and edit text, tables and figures of this paper from the emitted artifacts, including the scripts under `figures/` that render every figure and table from the run's own CSV files, and to check the manuscript against the author kit. Third, to draft three analyses over the pipeline's emitted tables: the rigid registration search that draws the two nulls of §5.3; the tile-clustered bootstrap intervals, Holm correction and split-half held-out test of §5.5; and the size-partial correlations and tile-level permutation test of the same section. The authors reviewed each draft, re-implemented it as the pipeline's final stage and re-ran the experiment; every value reported for those analyses comes from that stage's emitted tables. The tile-block permutation and the random-pair spectral null added in revision were likewise specified in AI-assisted review discussion and implemented by the authors as pipeline code. The scripts that render Figures 2b and 3 from emitted tables were AI-written as well. Fourth, to argue against the authors' own readings of the results to strengthen weak arguments where further clarification or amendments are required.

We did not use generative AI to generate synthetic data, to train or select models, or to derive the equations of §4 beyond checking their stated forms. Every scientific decision, including the hypothesis, the choice of descriptor, the two-arm design and the decision to compare stabilities rather than positions, was made by the authors, and every interpretation in this paper was reviewed by the authors against the raw artifacts. The conclusion reported is the authors' own.

The decisions above are recorded with their rationale in a decision log. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses RarePlanes, a public overhead-imagery dataset, under its published terms, with the label space restricted to its three civil transport and utility size classes. No human subjects are involved and no personally identifying information is processed.

The study is scoped to civil transport and utility aircraft in public overhead imagery, and its contribution is a measurement protocol rather than a detector improvement: the detectors we train are standard baselines replicated from published work for reproducibility. We note one direction in which the analysis is protective. The paper's central argument is that headline accuracy conceals localization behaviour, and that reliability should be reported on decomposed, scale-aware axes. Overstated small-object detection performance is a risk wherever these methods inform civil monitoring, from airport operations to infrastructure and traffic analysis, and a metric that hides where a box actually landed is the mechanism by which that risk is realized.

REPRODUCIBILITY STATEMENT

The measurand is a spread across a fixed seed set $\{0, \ldots, 7\}$, so the reported quantity is reproducible even though it is a variance: the seeds are pinned rather than sampled, and each run is deterministic given its seed and pinned library versions. Every experimental parameter is pinned in a single configuration module, listed in §3, and hashed into a provenance sidecar written beside each artifact together with the stage, runtime settings and the seeds or arms involved, so a run differing only in an accelerator flag is distinguishable after the fact. Deterministic algorithm selection is on in the three stages that produce the ensemble, classifier training, canvas optimization and cluster geometry; the detector stages run with it off, because the global pin forces RoI-align onto a compiled fallback that aborts the unfrozen arm and slows the frozen arm about twentyfold, and the cross-seed spread is a classifier-side quantity in any case. The classifier's batched on-device augmentation was matched per operation against the reference per-sample chain, and GPU stages run one whole run per device, so every run stays bit-identical to running it alone.

What is *not* claimed: determinism holds within a configuration, but bit-identical replay across different backend policies does not, because kernel autotuning changes the order in which partial products

accumulate. Four paths would run to completion while invalidating the result, and each is enforced in code rather than trusted: geographic coordinates are never read as tile pixels and every box is checked against the image bounds; model selection uses the holdout of §A.1 and never the shipped test split; the evaluated image universe is recorded in full, including images with neither ground truth nor detections, because they belong in false-positive denominators and bootstrap populations; and the disjoint-seed merge rule of §4.3 is enforced, since inconsistent transitive links could otherwise put two clusters from one seed into one track. The results archive of the reported run is provided as supplementary material: every table a reported value is read from, together with the run's logs, detector outputs and provenance sidecars, and every reported value names the file it comes from. Model weights and the per-cluster dumps from which no reported value is read are omitted for size.

REFERENCES

Christopher Algire, Steven Israel, and Kannappan Palaniappan. When One Pixel Tips the Scale: The Geometric Instability of IoU-Thresholded Metrics for Small Object Detection. In *Proceedings of the 54th Annual Applied Imagery Pattern Recognition Workshop (AIPR)*, Lecture Notes in Computer Science, Washington, DC, USA, 2026a. Springer. To appear.

Christopher Algire, Steven Israel, and Kannappan Palaniappan. Quantifying learned feature representations in vision models using Zernike moments. In *Pattern Recognition and Prediction XXXVII*, volume 14032, pp. 140320R. International Society for Optics and Photonics, SPIE, 2026b. doi: 10.1117/12.3115563. URL https://doi.org/10.1117/12.3115563.

Christopher Algire, Steven Israel, and Kannappan Palaniappan. Beyond mAP_{50}: A Decomposed and Scale-Aware Test Harness for Small Object Detection, 2026c. Manuscript; venue not yet confirmed.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6541–6549, 2017.

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.

Alireza Khotanzad and Yaw Hua Hong. Invariant image recognition by Zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):489–497, 1990.

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

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *International Conference on Learning Representations (ICLR)*, 2016.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016.

Robert J. Noll. Zernike polynomials and atmospheric turbulence. *Journal of the Optical Society of America*, 66(3):207–211, 1976.

Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. Localization recall precision (LRP): A new performance metric for object detection. In *European Conference on Computer Vision (ECCV)*, pp. 504–519, 2018.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

David Picard. Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision. *arXiv preprint arXiv:2109.08203*, 2021.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6076–6085, 2017.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 91–99, 2015.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, 2019.

Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. RarePlanes: Synthetic data takes flight. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 207–216, 2021.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Cecilia Summers and Michael J. Dinneen. Nondeterminism and instability in neural network optimization. In *International Conference on Machine Learning (ICML)*, pp. 9913–9922, 2021.

Michael Reed Teague. Image analysis via the general theory of moments. *Journal of the Optical Society of America*, 70(8):920–930, 1980.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. In *ICML Deep Learning Workshop*, 2015.

Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1257–1265, 2020.

F. Zernike. Beugungstheorie des Schneidenverfahrens und seiner verbesserten Form, der Phasenkontrastmethode. *Physica*, 1(7–12):689–704, 1934.

A Method details

A.1 Data handling and cluster detection

The dataset's GeoJSON geometry is WGS84, and reinterpreting it as pixel coordinates would corrupt every box, so the pipeline consumes the publisher's tile-space COCO annotations instead. The label space holds the three civil size classes and nothing else: annotations outside it are left unlabelled, and a tile with no in-scope annotation is kept as background rather than discarded, so background statistics are not quietly biased toward empty scenes.

Model selection uses a fixed 10% holdout of whole physical locations, drawn under a split seed independent of the eight model seeds and shared by all eight models; the shipped test split is first touched at detection evaluation. One property of the published split bounds the claim: 27 location prefixes occur in both the train and test file lists, across 92 train and 47 test locations, and we keep the official split for comparability with published numbers, which is what scopes the detector claim (§3.1).

The cluster detector of §4.1 builds its saliency map from colour saturation, Sobel gradient magnitude and local variance, Gaussian smoothed. Its five percentiles are $\{65, 70, 75, 80, 85\}$, overlapping

components are suppressed by non-maximum suppression at factor 0.6, and components outside $[0.003, 0.25]$ of canvas area are discarded.

A.2 Deviations from the original descriptor

Equations (5) to (10) are the canonical forms of the corpus notation standard and follow its rules as written: the estimator wears a hat where the continuous moment does not, the conjugated angular factor is spelled $e^{-im\theta}$, the pooling exponent is declared once and never mixed, and *energy* names the $q = 2$ quantity only. This paper is the standard's local chain throughout, so its spectra are not comparable with whole-map values. Three definitions nonetheless deviate. First, (8) squares the moment magnitudes whereas the original summed them under the same name; the squared form is the Parseval-consistent energy that (9) assumes. Second, dominant order is $n^* = \arg\max_{n>0} E_n^{(2)}$, excluding $n = 0$, whose DC term reports patch brightness rather than structure. Third, the bridge chain pools over every admissible pair, $\mathcal{A}(8)$, whereas the canonical descriptor keeps $m \geq 0$ only, $\mathcal{A}^+(8)$. For a real-valued map $\widehat{Z}_n^{-m}$ is the complex conjugate of $\widehat{Z}_n^m$, so the two sets carry the same content at different lengths, 45 components against 25, but not uniformly across orders: odd orders admit no $m = 0$ term, so their all-m sum is exactly twice the $\mathcal{A}^+$ sum, while even orders keep that term once. A coefficient of variation is invariant to a constant factor, so (9) is unaffected at the four odd orders and shifted only at the five even ones, and (10) becomes a cosine weighted by $2 - \delta_{m0}$. Both conventions run in one pipeline, deliberately: every bridge quantity in §5 uses all-m, while §C.1 and §C.3 use $\mathcal{A}^+(8)$ and report 25 modes, so the two families cannot be compared coefficient by coefficient. All three make $E_n^{(2)}$, n^*, D_{spec} and S numerically incomparable with archived values from the original study, and none affects (3), which does not use the spectrum.

A.3 Spectral variability statistics

The patch of §4.2 is mapped to the unit disk, where the basis function of radial order n and azimuthal repetition m is a radial polynomial times an angular harmonic,

$$V_n^m(\rho, \theta) = R_n^{|m|}(\rho)\, e^{im\theta}, \tag{5}$$

with $R_n^{|m|}$ the usual radial polynomial, equal to 1 at the disk rim, and an index pair admissible when $|m| \leq n$ and $n - |m|$ is even. The moment is the projection of the patch onto that basis. It has one continuous definition and one pixel-sum estimator, and because this paper measures rather than derives, it uses the estimator:

$$\widehat{Z}_n^m = \frac{n+1}{\pi} \sum_{(x,y)\in D_\Delta} f(x, y)\, R_n^{|m|}(\rho)\, e^{-im\theta}, \tag{6}$$

where f is the min–max conditioned patch of §4.2, a step the corpus standard states as part of the measurement rather than as incidental preprocessing, the sum runs over the pixel centres D_Δ inside the inscribed disk, and the prefactor $(n + 1)/\pi$ is the reciprocal of the basis function's squared size. The magnitude spectrum of a cluster collects the $|\widehat{Z}_n^m|$ over the admissible index pairs,

$$\mathbf{z} = \big(\big|\widehat{Z}_n^m\big| \big)_{(n,m)\,\in\,\mathcal{A}(8)}, \tag{7}$$

which is 45 components under the convention of §A.2. Radial pooling raises each magnitude to a declared exponent before summing over repetitions, and this paper declares $q = 2$, so the pooled quantity is an energy in the Parseval-consistent sense:

$$E_n^{(q)} = \sum_{m\,:\,(n,m)\in\mathcal{A}(8)} \big|\widehat{Z}_n^m\big|^{\,q}, \qquad q = 2. \tag{8}$$

For a track T of the same feature recovered under several seeds, spectral dispersion is the mean coefficient of variation over orders,

$$D_{\text{spec}} = \operatorname*{mean}_n \frac{\operatorname{std}_{t\in T} E_{n,t}^{(2)}}{\operatorname{mean}_{t\in T} E_{n,t}^{(2)}}, \tag{9}$$

where $E^{(2)}_{n,t}$ is the energy at order n on seed t. Spectral identity between two matched clusters a and b is the cosine similarity of their magnitude spectra, reported both over all components and with the $(0,0)$ component removed,

$$S_{ab} = \frac{\mathbf{z}_a \cdot \mathbf{z}_b}{\|\mathbf{z}_a\| \, \|\mathbf{z}_b\|}, \qquad S^{+}_{ab} = \text{the same with the } (0,0) \text{ component removed.} \tag{10}$$

Basis orthogonality is what makes (10) worth computing: the magnitude spectrum is a coordinate-free signature, so drift in S_{ab} is drift in the representation rather than in how the canvas rendered. S^{+}_{ab} is a necessity rather than a refinement, because min-max normalization makes the $(0,0)$ magnitude large and similar in every cluster, and S_{ab} then saturates near 1 whatever the structured orders do.

A.4 Training and canvas hyperparameters

Classifier. 200 epochs, batch size 32, differential learning rates of 10^{-3} at the head and 10^{-4} at the backbone, warmup into a cosine schedule, weight decay 10^{-4}, dropout 0.3, early stages conv1, bn1, layer1 and layer2 frozen, validation holdout fraction 0.10 at split seed 0. **Detector.** The recipe fixes how much data the detector sees rather than an iteration count: 120,000 image presentations at a per-image learning rate of 5×10^{-4}, so iterations and base rate follow from the batch; the reported run used a global batch of 16, hence 7500 iterations at base rate 0.008. Score threshold 0.30 at evaluation, detections written down to 0.05 so the precision-recall tail survives on disk, every reliability statistic computed at the operating point, RGB input with ImageNet normalization. **Canvases.** Adam at lr $= 0.02$, cosine annealed over 800 image-update steps, total-variation weight annealed 10^{-2} to 10^{-3}, ℓ_2 at 10^{-4}, periodic Gaussian smoothing, best-scoring image kept; the checkerboard and wavelet artifacts this suppresses are read as spurious small clusters and inflate high-order Zernike energy. **Harness.** $k = 0.5$, Hungarian assignment, overlap required, scale range 8 to 128 px square-root area, minimum 10 ground-truth boxes per octave, 200 bootstrap resamples at seed 0.

B Additional results

B.1 Bounds on the numbers

Three bounds apply to every number in §5 beyond those of §6. All 28 seed pairs come from the same eight trained models, so bootstrap intervals over pairs and the between-class p-values are pseudoreplication; every emitted row is stamped `inference_status = exploratory_nonindependent_pairs`, and nothing is claimed from them. The radius floor is upstream of every feature-side number, which is why every r-dependent analysis appears with the censored rows and without. And $\bar{R}$ for the large class rests on truncated support, since only 48% of its reference boxes have square-root area inside the harness's 8–128 px range, so that column should not be read across classes. Two limits belong to the instrument: gradient-ascent canvases are synthetic renderings rather than the tiles the detector sees, so cluster geometry describes what the classifier prefers rather than where objects are, and the classifier uses class-weighted binary cross-entropy where the original configuration defaults to focal loss, so weights, canvases and clusters all differ from that study's.

B.2 The bridge correlation tables

Table 3 gives the hypothesis of §5.4 in both arms, five of the ten detection targets the stage computes (the rest, including AP@[.5:.95], are in `bridge/bridge_correlations_{F,U}.csv`), read under §5.2.

B.3 The nulls and the positive control, in full

The frame-artifact null has two forms. The first re-ran the matching of §4.3 under each of the eight symmetries of a square, requiring the identity to dominate the other seven; on match rate it places second, sixth and fifth of eight by class (`bridge/bridge_null_dihedral.csv`), and the repeats of §C.1 move that rank across most of its range. The second rotates one seed's field about the canvas centre by a continuous angle, which holds each cluster's radius and distance from

	Arm F (frozen)					Arm U (unfrozen)				
Predictor	med δ_{norm}	P_{score}	$\bar{R}$	Rec.	AP_{50}	med δ_{norm}	P_{score}	$\bar{R}$	Rec.	AP_{50}
med σ_{norm}	-1.0	$+1.0$	$+0.5$	$+0.5$	$+0.5$	-1.0	$+1.0$	-0.5	$+0.5$	-1.0
D_{spec}	-0.5	$+0.5$	-0.5	-0.5	-0.5	-0.5	$+0.5$	-1.0	$+1.0$	-0.5
mean S_{ab}	$+1.0$	-1.0	-0.5	-0.5	-0.5	$+1.0$	-1.0	$+0.5$	-0.5	$+1.0$
M_c	$+1.0$	-1.0	-0.5	-0.5	-0.5	$+1.0$	-1.0	$+0.5$	-0.5	$+1.0$
med $\tilde{r}$	—	—	—	—	—	—	—	—	—	—
spacing	-0.5	$+0.5$	-0.5	-0.5	-0.5	-0.5	$+0.5$	-1.0	$+1.0$	-0.5

Table 3: Bridge correlations (ρ_S) in both arms (`bridge_correlations_{F,U}.csv`). Each rests on three points, so ρ_S takes only $0, \pm 0.5, \pm 1.0$; med $\tilde{r}$ is undefined because the median radius equals the 15 px floor in every class. The last row is a sensitivity companion to med σ_{norm}. Ten of the twenty-five defined cells change between arms, eight by a sign flip.

the centre fixed and draws without limit. Rotation is closed only on the inscribed disk, and 25% of clusters sit outside it, so a rotated field is sparser where the fixed field lives, which flatters the identity; Table 2 therefore restricts both sides to the disk, and the correction runs in that direction, the medium class moving from $p = 0.11$ over the full canvas to 0.46 on the disk. The two forms carry opposite artifacts, the symmetries preserving the pixel lattice and the disk leaking no clusters across its boundary, and on match rate they agree in two classes of three.

The positive control, stage 18, runs the rotation form on stage 09's ten gradient-ascent restarts at fixed weights, where the features are the same and only the rendering differs, so a statistic that measures correspondence has to separate from chance. Each model's repeats give 24 model-by-class cells per statistic, read as a count of cells beating the null at $\alpha = 0.05$ against the 1.2 chance would give (`bridge/positive_control_summary.csv`): match rate 13, median σ_{norm} 1, mean S_{ab} 3 and mean S_{ab}^{+} 2. A binomial tail treating the cells as independent would put the first at 1.8×10^{-11}, but the cells share eight models and three classes, so the count is the whole reading. Median σ_{norm} and both forms of S_{ab} do not separate, so no cross-seed reading of them is available in either direction; the guarantee holds at the level of the weights rather than the canvas, since a restart renders the same features differently, which is the condition the cross-seed comparison also faces.

C Calibrations added after the bridge

Thirteen stages run after the figures, and their job is to bound how far the chance reading reaches: to this run, to the gradient-ascent canvas, or to the descriptor itself.

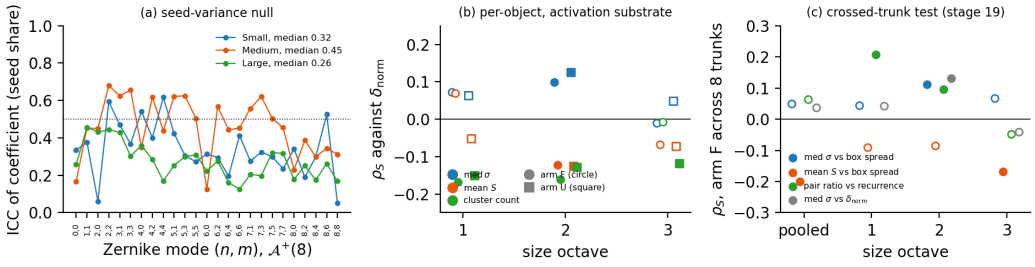


Figure 3: **(a)** Intraclass correlation of each canonical Zernike mode across eight seeds and ten canvas restarts; dotted line at one half. **(b)** Per-object ρ_S against δ_{norm} by size octave on the activation substrate, both arms. **(c)** The four limbs of the crossed-trunk test, arm F from all eight trunks. Filled markers: $p_{\text{perm}} < 0.05$. From `figures/make_appendix_panels.py`.

C.1 How much of the feature side is noise, and whose noise it is

Stage 09 re-runs cluster detection, matching, tracking and the square-symmetry null ten times against the same eight checkpoints, varying only the stage's own seed; Table 4 gives the spread (`repeats/geometry_repeats_summary.csv`).

Over ten repeats, mean $\pm$ SD	Small	Medium	Large
Clusters	114.3 ± 4.8	138.6 ± 9.2	157.6 ± 9.2
Admissible pairs	39.3 ± 5.5	49.4 ± 8.9	62.4 ± 11.1
Tracks spanning ≥ 3 of 8 seeds	7.0 ± 1.9	10.2 ± 2.3	12.9 ± 2.8
Match rate M_c	0.113 ± 0.014	0.120 ± 0.019	0.133 ± 0.022
$\operatorname{med} \sigma_{\mathrm{norm}}$	0.348 ± 0.026	0.342 ± 0.017	0.345 ± 0.041
Identity rank of 8 (range)	3.0 ± 1.8 (1–6)	4.0 ± 1.9 (1–6)	2.5 ± 2.0 (1–7)

Table 4: Repeat-to-repeat spread of the feature side at fixed weights, from `repeats/geometry_repeats_summary.csv`.

The identity rank spans most of its range, so the ranks this run reports are draws rather than an ordering, and the small-class match rate of 0.142 that reaches $p = 0.03$ under rotation is two repeat standard deviations above its own mean. The repeat standard deviation of $\operatorname{med} \sigma_{\mathrm{norm}}$ is about the size of the between-class spread the bridge table ranks on, 0.304 to 0.356, so that ranking is unstable under re-running the stage that produces it, which is stronger than the three-point caveat of §B.1 and independent of it.

Stage 15 asks how much of a Zernike coefficient's variation across the eight canvases belongs to the seed and how much to the canvas optimizer. Eight seeds by ten repeats give eighty observations per coefficient, and a one-way decomposition separates a seed component σ^2_{seed} from an instrument component σ^2_{inst}; their intraclass correlation is the fraction of variance the measure is entitled to interpret. Over the 25 canonical modes of $\mathcal{A}^+(8)$ (Figure 3a) it has median 0.324, 0.453 and 0.263 by class, so 55 to 74 percent of a typical coefficient's variation is the instrument. This is the seed-variance null of §5.2, and it sits upstream of the other two.

The same stage puts the gradient-ascent spectrum 1.54, 1.52 and 1.59 away in relative ℓ_2 from the shared-aperture spectrum of real activations at `layer4`, more than the spectra themselves: a different measurement rather than a rescaled one.

C.2 Cross-seed correspondence, with and without a shared frame

Stage 16 first replaces the harness's $k = 0.5$ rule with a threshold the data sets, the pooled cross-seed standard deviation of cluster radius at 5.6, 8.4 and 6.1 px by class: 15, 12 and 10 of the 28 seed pairs return no admissible correspondence, the survivors contribute 16, 25 and 24 matches, and stitching those into components gives 51 features across the three classes, 48 spanning exactly two seeds and 3 spanning three, so no feature reaches a seed fraction above 0.375.

Both rules assume the canvases share a frame, and the stage then drops that assumption. For each seed pair it seeks the rigid transform, a rotation and a translation, that brings the most clusters of one canvas into coincidence with the other, by RANSAC over 750 two-point hypotheses, and assigns the coincident clusters by the Hungarian solver under two gates: a position gate equal to the 68th percentile of residual shift when the same registration is run on stage 09's fixed-weight restarts, 10.0, 9.8 and 10.1 px by class, which is the precision the instrument itself reaches when the weights fix the correspondence; and a gate on the log radius ratio equal to its pooled standard deviation, 0.25, 0.32 and 0.25. A pair needs three inliers to count, and the match rate is the matched count over the smaller cluster set. Under registration it is 0.396 ± 0.066, 0.398 ± 0.069 and 0.403 ± 0.075 over the 28 pairs, against 0.102, 0.093 and 0.102 in the shared frame under the same gates; no pair returns empty, and the components reach four, six and six of the eight seeds (Figure 2a).

The pre-analysis plan for this test, written before the run reported, requires a null under the same search; the null was added after a first run had reported, a recorded deviation, and runs as the pipeline's final stage. That stage re-runs a rigid two-point RANSAC with the same gates, hypothesis count, inlier minimum and assignment rule on the emitted cluster tables, returning 0.358, 0.357 and

0.351 on the same-class pairs, below the calibrated search above; every comparison stays within its search. Two nulls are drawn under it: one canvas replaced by a canvas of a *different* class from the same seed set, 56 pairs per class; and one canvas's cluster centres redrawn uniformly over the canvas, radii kept (`registration_null/*.csv`). Against the first the observed class means fall at one-sided $p = 0.41$, 0.46 and 0.64, below 95th percentiles of 0.444, 0.451 and 0.500; against the second at $p = 0.36$, 0.25 and 0.50, below 0.444, 0.465 and 0.445. Every difference from the null median is within 0.025 with a bootstrap interval covering zero, and the pre-specified criterion is met in no class: registration finds a third of any two cluster sets of this density coincident, whatever they contain.

The two pipeline diagnostics of §5.2 agree: with a reflection permitted the match rate is 0.411, 0.404 and 0.408, and composed through a medoid seed it is 0.12, 0.15 and 0.14 (`correspondence/reflection_sensitivity_summary.csv`, `medoid_frame_pairwise_summary.csv`). Under a threshold the data sets, in a shared frame and in the best frame a rigid search can find, the canvas carries no population of features that recur across retrainings, which is the premise (3) needs and the reason the nulls of §5.2 come out as they do.

C.3 Object-anchored activation geometry, in full

Stage 10 takes the classifier's response map at the chosen tap for each seed on a fixed sample of test tiles, reduces it with the rules of §A.1, keeps only clusters whose centre falls inside an annotated box, and matches exactly as in §4.3, with a 16 px radius floor on the tile grid that almost never binds. The sample is 191, 248 and 183 tiles per class, giving 32,811 clusters and 34,732 admissible pairs at censored fractions below 0.02 (`activation/activation_class_stats.csv`). Recomputing a fixed subsample reproduces the emitted rows byte for byte (`activation/repeat_check.json`), so the response map is a function of the weights and the tile alone.

Every per-object test is in `activation/object_correlations_{F,U}.csv`, nine per arm at octaves 1 to 3 (octave 0 holds one object); Table 5 and Figure 3b give them with three companions the pipeline's final stage computes from `activation/object_pairs_{F,U}.csv`: a 95% bootstrap interval resampling tiles rather than objects, since objects on one tile share a response map; a Holm-corrected p over the eighteen per-object permutation tests, both descriptive since they treat objects as independent; inference rests on the intervals and the tile-block test; and a split-half test in which, for each of 500 random halves of the 500 tiles, the cell with the largest $|\rho_S|$ is chosen on one half and evaluated on the other in that sign, with a one-sided permutation p. Two further companions: a partial rank correlation given object size, since cluster count rises with size within an octave at $\rho_S = 0.33$ to 0.36 at octaves 1 and 2 and 0.71 at octave 3 (`publication/confound_summary.csv`); and a tile-block permutation that shuffles δ_{norm} between tiles with each tile's objects kept intact, so the permuted unit is the tile (`publication/tile_block_summary.csv`). The same stage draws the substrate's random-pair spectral null, pairing same-class clusters at random: matched pairs exceed it in S by 0.056 to 0.074 and in S^+ by 0.123 to 0.152 per class, at a sign-flip p surviving Holm at 0.001 (`publication/activation_spectral_null_summary.csv`). The split-half selection picks cluster count in 68% and 62% of splits by arm, preserves the sign in 83% and 88% at a median held-out ρ_S of 0.10 and 0.07, and reaches $p < 0.05$ on the held-out half in 45% and 17% (`publication/heldout_splits.csv`).

Class-conditional apertures. Stages 11 to 13 test the remaining explanation, that the spectrum carries little class signal for want of a class-conditional aperture. Grad-CAM supplies one deterministically and stably, at a median S^+ of 0.992 against the unperturbed aperture over 96,465 perturbations of centre and radius; the same harness also re-measures each cluster at its moment-consistent centre, where the $(1, \pm 1)$ moments vanish, about a pixel from the saliency centroid the bridge matches on, and reports that mode beside the shipped one. The aperture does not make the spectrum class-discriminative: at `layer4` every between-class median S^+, 0.799 to 0.812, falls inside the within-class range of 0.777 to 0.822, which bounds any aperture-based repair.

	Arm F					Arm U				
Predictor, by octave	n	ρ_S [95% CI, by tile]	p_{Holm}	$\rho \mid$ size	p_{tile}	n	ρ_S [95% CI, by tile]	p_{Holm}	$\rho \mid$ size	p_{tile}
$\mathrm{med}\,\sigma$, octave 1	303	$+0.07$ [-0.05, +0.21]	1.00	$+0.07$	0.494	312	$+0.06$ [-0.06, +0.19]	1.00	$+0.06$	**0.040**
$\mathrm{med}\,\sigma$, octave 2	449	$+0.10$ [-0.02, +0.21]	0.36	$+0.07$	0.415	464	$+0.13$ **[+0.01, +0.24]**	0.10	$+0.12$	**0.011**
$\mathrm{med}\,\sigma$, octave 3	327	-0.01 [-0.13, +0.11]	1.00	-0.01	0.610	329	$+0.05$ [-0.06, +0.16]	1.00	$+0.06$	0.103
mean S, octave 1	303	$+0.07$ [-0.04, +0.18]	1.00	$+0.07$	0.579	312	-0.05 [-0.16, +0.06]	1.00	-0.05	**0.025**
mean S, octave 2	449	-0.12 **[-0.22, -0.01]**	0.12	-0.09	0.265	464	-0.13 **[-0.23, -0.00]**	0.13	-0.12	**0.014**
mean S, octave 3	327	-0.07 [-0.19, +0.05]	1.00	-0.07	0.217	329	-0.07 [-0.20, +0.05]	1.00	-0.07	0.110
cluster count, octave 1	397	-0.17 **[-0.27, -0.07]**	0.02	-0.09	0.342	413	-0.15 **[-0.24, -0.05]**	0.10	-0.10	0.079
cluster count, octave 2	483	-0.16 **[-0.25, -0.07]**	0.02	-0.10	0.600	499	-0.13 **[-0.22, -0.03]**	0.10	-0.11	0.940
cluster count, octave 3	327	-0.01 [-0.12, +0.10]	1.00	$+0.03$	0.319	329	-0.12 **[-0.23, -0.00]**	0.34	-0.00	0.232

Table 5: Every per-object test against δ_{norm}: tile-clustered bootstrap interval (bold where it excludes zero), Holm correction over the eighteen per-object permutation tests (descriptive), partial rank correlation given object size, and the tile-block permutation p (bold below 0.05; no cell survives its own Holm correction). From `publication/heldout_summary.csv`, `confound_summary.csv` and `tile_block_summary.csv`.