SOFT-ARGMAX FOR THE PROJECTIVE PLANE VIA THE VERONESE EMBEDDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Soft-argmax is a widely used readout operator for differentiable coordinate regression, turning heatmaps into predictions by averaging candidate coordinates. While equivariant architectures account for symmetries in feature processing, the readout that turns evidence into a prediction, and the loss on that prediction, must respect the topology of the target domain as well. The parameters of undirected lines are a case in which such a symmetry breaks coordinate averaging. Each line has two sign-related representatives, so no coordinate domain admits an average that respects proximity between lines, and we prove that any readout averaging signed offsets can miss a target on which the evidence concentrates. We introduce *Veronese soft-argmax*, a parameter-free readout that averages sign-invariant quadratic features and recovers the line from their average. The same representation yields the training objective, whose value bounds the error of the recovered line. We prove that predictions approach the target as evidence concentrates and give sufficient conditions for a unique prediction, on real projective spaces of any dimension. Experiments on synthetic lines show that readout and objective each contribute, and our approach transfers to horizon estimation in real images. The construction applies wherever a heatmap ranges over candidates in a real projective space, and in particular to targets with antipodal symmetry.

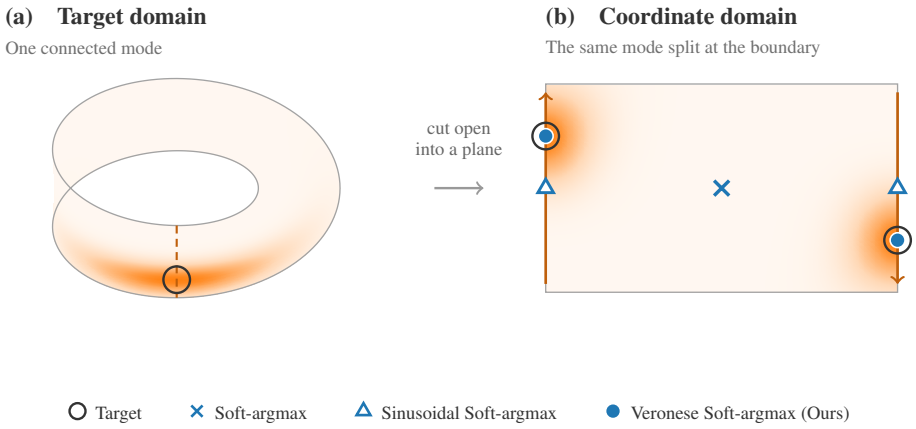


Figure 1: **Heatmap readout must respect the topology of the target domain.** For targets on the projective plane, a single heatmap peak can appear split in coordinates, as illustrated by the Möbius strip (a) and its coordinate domain (b). Soft-argmax and its sinusoidal variant can then extract a location away from the target despite well-localised evidence. Veronese soft-argmax respects the projective identification when combining that evidence. Orange shows relative density. Readout markers are schematic.

1 INTRODUCTION

A heatmap can concentrate on the correct target and still be read out as the wrong one. Soft-argmax makes prediction differentiable by averaging candidate coordinates under normalised heatmap prob-

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.

abilities (Luvizon et al., 2019; Nibali et al., 2018). For image keypoints this works because pixel coordinates are Euclidean, so averaging nearby points gives a nearby point (Ham & Oh, 2023; Sharan et al., 2021). The parameters of many geometric objects are not Euclidean. They wrap around, or give one object two distant parameter values, so averaging nearly identical objects can return a different one. Line prediction, needed for horizon estimation in camera calibration (Workman et al., 2016; Abbas & Zisserman, 2019) and needle-axis estimation in surgical guidance (Wijata et al., 2024; Steiger et al., 2026), is one instance of this problem. A readout must turn evidence into an object, so it must respect the geometry of the label space.

In the Hesse normal form of lines, a half-turn in orientation reverses the signed offset (Stillwell, 1995). Nearly identical lines across the angular boundary therefore have opposite offsets, which cancel under linear averaging even as the evidence concentrates on the target (Figure 1). Sinusoidal soft-argmax (Joung et al., 2020) handles angular wrapping but leaves this cancellation unchanged. The coupling of angle and offset gives the domain its Möbius topology and reflects the sign ambiguity of projective targets.

The Hough transform (Duda & Hart, 1972), embedded as a known operator (Maier et al., 2019), gives a heatmap over candidate lines. Existing approaches map Hough features back to image space (Lin et al., 2020), extract peaks without differentiating through selection (Zhao et al., 2022), or predict orientation without signed offset (Bachmaier et al., 2023). Transformer-based detectors instead predict lines without this prior and attribute suboptimal results of Hough-based approaches to their modelling and inference steps (Xu et al., 2021). The inference step is the readout, and the missing component is a differentiable one that combines orientation and offset while respecting their projective identification. Its prediction should approach the target whenever the evidence does.

We address this with Veronese soft-argmax (VSmax), which averages six sign-invariant quadratic features instead of domain coordinates. These are the established degree-two Veronese features (Hamm & Lee, 2008; Harandi et al., 2014). Opposite representatives contribute identical features, the target line is supervised in the same features, and a line is recovered from their average as a leading eigenvector. This follows extrinsic averaging and regression (Bhattacharya & Patrangenaru, 2003; Lin et al., 2017) and the eigenvector construction used in quaternion averaging (Markley et al., 2007). The network, probabilities and linear aggregation remain unchanged. Our contributions are:

- **A geometric obstruction to standard soft-argmax.** We prove that standard soft-argmax and any variant retaining linear signed-offset averaging can remain inaccurate even as evidence concentrates on the correct line, whatever angle it predicts (Proposition 1).
- **A readout and an objective on one projective representation.** We introduce VSmax, a parameter-free operator that replaces soft-argmax in any heatmap model over a projective target domain. A squared error in the same representation measures error in the geometry rather than in coordinates. Readout and objective each improve accuracy, and gains grow as evidence spreads across the heatmap.
- **Guarantees for both.** We prove invariance to antipodal representation, accurate recovery as evidence concentrates, and that a small loss bounds the recovered error and, below an explicit threshold, yields a unique prediction (Theorem 1). The construction and guarantees hold on real projective spaces of any dimension.

2 Problem formulation and geometric background

The failure of coordinate averaging on undirected lines is topological, so no choice of angular readout removes it. We make the space and its metric precise, formalise the two requirements a readout must meet to respect that topology, and prove that averaging signed offsets cannot meet both.

2.1 Projective geometry of undirected lines

The general equation of a line $au+bv+c=0$ is unchanged when all its coefficients are multiplied by a nonzero scalar. Normalising the coefficient vector reduces the scaling freedom to a sign ambiguity. The unit vectors x and $-x$ describe the same line, which we write as the projective point $[x] = \{x, -x\}$. These points form the real projective plane $\mathbb{RP}^2 = S^2/\{\pm 1\}$. In the Hesse normal form $\rho = u\cos\theta + v\sin\theta$, with (u, v) image coordinates relative to a chosen origin, θ the angle of the

line normal and ρ the signed distance of the line from the origin, the unit coefficient vector is

$$\ell(\theta, \rho) = \frac{(\cos\theta,\ \sin\theta,\ -\rho)^\top}{\sqrt{1+\rho^2}} \in S^2. \tag{1}$$

Since $\ell(\theta+\pi, -\rho) = -\ell(\theta, \rho)$, angular wrapping is coupled to offset reversal. Training by regression requires a unique target, so each line must be represented once. The canonical domain $C = [0, \pi) \times \mathbb{R}$ contains one representative per finite line. Its angular boundaries are glued by $(\theta, \rho) \sim (\theta+\pi, -\rho)$, which gives a Möbius strip (Figure 1a). Its surface is non-orientable, so no coordinates exist whose Euclidean differences measure distance between lines. A regression objective is a function on the label space, so for lines it must be a function of a metric on the projective points $[x]$ rather than of chart coordinates. Such a metric is the chordal distance (rationale in Appendix B.2),

$$d([x], [y]) = \sqrt{1 - (x^\top y)^2} = \sin\phi \in [0, 1], \tag{2}$$

where $x, y \in S^2$ and ϕ is the smaller angle between their unoriented axes. Antipodal representatives have $d = 0$ and orthogonal lines $d = 1$. This is the one-dimensional case of the Grassmannian chordal distance of Conway et al. (1996). For lines, we evaluate d on unit coefficient vectors at a fixed image origin and scale bounding ρ to the image boundaries (example in Appendix A).

2.2 Requirements and limitations of soft-argmax readouts

A heatmap is a finite probability distribution P over cells, each cell i carrying a unit representative $x_i \in S^2$ with probability p_i. On the Hough grid, $x_i = \ell(\theta_i, \rho_i)$ for the cell's canonical coordinates (θ_i, ρ_i). A readout $\mathcal{R}$ maps P to a line $\mathcal{R}(P) \in \mathbb{RP}^2$, possibly undefined on some P. Standard soft-argmax returns the mean of P in coordinates. This aggregation of evidence is valid only if lines that are close in $\mathbb{RP}^2$ are also close in coordinates, which the identification $(\theta, \rho) \sim (\theta+\pi, -\rho)$ prevents. We therefore state the requirements on $\mathcal{R}$ in terms of lines rather than cells. First, assigning the same probabilities to the same lines must give the same prediction, whichever representatives carry them.

Definition 1 (Representative invariance)**.** Let P assign masses $p_i \geq 0$, $\sum_i p_i = 1$, to finitely many cells with unit vectors x_i, and let

$$\bar{P}([x]) = \sum_{i:\,[x_i]=[x]} p_i \tag{3}$$

be the induced distribution over lines. A readout $\mathcal{R}$ is *representative-invariant* if its prediction and its domain of definedness depend on P only through $\bar{P}$.

On the Hough grid each line occupies exactly two cells i, j with $x_j = -x_i$, so $\bar{P}([x_i]) = p_i + p_j$. *Folding* moves this mass onto the canonical cell, $\tilde{p}_i = p_i + p_j$ and $\tilde{p}_j = 0$, and yields a representative-invariant readout for any rule applied to the folded heatmap. Invariance alone is necessary but not sufficient. As the expected chordal distance to a target tends to zero, the prediction should approach that target, even when the evidence is split across the angular boundary (Figure 1b).

Definition 2 (Concentration-consistency)**.** Write the expected chordal distance to a target $[x^\star]$ as

$$\bar{d}(P, x^\star) = \mathbb{E}_{X\sim P}\, d([X], [x^\star]) = \sum_i p_i\, d([x_i], [x^\star]). \tag{4}$$

The readout is *concentration-consistent* if for every target there is a function φ with $\varphi(t) \to 0$ as $t \to 0$ such that

$$d\big(\mathcal{R}(P), [x^\star]\big) \leq \varphi\big(\bar{d}(P, x^\star)\big)$$

for all P with sufficiently small $\bar{d}(P, x^\star)$, with $\mathcal{R}(P)$ defined there.

Standard and sinusoidal soft-argmax act on the canonical domain C in three steps: fold, average the coordinates of the folded heatmap to a prediction $(\hat{\theta}, \hat{\rho})$, and decode it as the line $[\ell(\hat{\theta}, \hat{\rho})]$. Writing p_i for the folded probabilities, they share the offset average $\hat{\rho} = \sum_i p_i \rho_i$ and differ in the angle:

$$\begin{aligned} \text{soft-argmax}: &\quad \hat{\theta} = \textstyle\sum_i p_i \theta_i, \\ \text{sinusoidal soft-argmax}: &\quad \hat{\theta} = \tfrac{1}{2}\operatorname{atan2}\big(\textstyle\sum_i p_i \sin 2\theta_i,\ \sum_i p_i \cos 2\theta_i\big) \bmod \pi, \end{aligned} \tag{5}$$

where the sinusoidal rule of Joung et al. (2020) is applied to the doubled angle 2θ because lines are π-periodic. Both are representative-invariant by folding, and both decode a heatmap concentrated on one cell correctly. Concentration-consistency is what fails. Near the angular boundary a line is approached from both edges with offsets of opposite sign, and the linear average cancels them. Proposition 1 turns this into a bound that holds for every predicted angle (proof in Appendix B.5).

Proposition 1 (Canonical signed-offset averaging is inconsistent)**.** *No readout on C returning $[\ell(\hat{\theta}, \hat{\rho})]$ with $\hat{\rho} = \sum_i p_i \rho_i$ and any $\hat{\theta} \in [0, \pi)$ is concentration-consistent. Fix $a \neq 0$, let $x^\star = \ell(0, a)$, and let P_ε put equal mass on (ε, a) and $(\pi - \varepsilon, -a)$ for $0 < \varepsilon < \pi/4$. Then $\bar{d}(P_\varepsilon, x^\star) \to 0$ as $\varepsilon \to 0$, but every such readout predicts $\hat{\rho} = 0$ and satisfies*

$$d\big([\ell(\hat{\theta}_\varepsilon, 0)], [x^\star]\big) \geq \frac{|a|}{\sqrt{1 + a^2}} > 0. \tag{6}$$

Both baselines are therefore representative-invariant but not concentration-consistent. Standard soft-argmax errs by 1, returning $(\pi/2, 0)$ (cross in Figure 1b), while the sinusoidal variant (triangles) gets the angle exactly right and still errs by the bound in Equation 6. A better angular readout therefore cannot restore concentration-consistency while retaining linear signed-offset averaging. Appendix D.1 reproduces this failure on synthetic heatmaps over C. For a target on the angular boundary, both baselines retain this error at every heatmap width, while VSmax converges to the target. Selecting unit representatives and averaging them also fails, since no global selection rule is continuous (Proposition B.2 in Appendix B.4). We therefore average sign-invariant features rather than signed offsets or selected unit representatives. The next section constructs such a readout.

3 Veronese soft-argmax

VSmax replaces the coordinate features averaged by soft-argmax with sign-invariant Veronese features. Since the operator already produces its output in this representation, the objective is stated there as well. The feature average is supervised in the embedding, and a line is recovered from it afterwards. The heatmap network as well as the linear aggregation stay the same, and the construction adds no trainable parameters.

3.1 Veronese embedding and aggregation

For a unit representative $x \in S^2$ of a line $[x]$, the established degree-two Veronese embedding (Harris, 1992; Hamm & Lee, 2008; Harandi et al., 2014) collects the six distinct quadratic products:

$$\nu_2(x) = \big(x_1^2,\ x_2^2,\ x_3^2,\ x_1x_2,\ x_1x_3,\ x_2x_3\big)^\top, \qquad \nu_2(x) = \nu_2(-x). \tag{7}$$

These features identify exactly $x \sim -x$ and vary continuously across the angular boundary, so the aggregation below needs no canonical fold and no assumption of equal antipodal mass (Theorem 1(i)). Appendix B.1 relates this unit-normalised realisation to the general projective Veronese map.

On a fixed grid of N cells $x_i = \ell(\theta_i, \rho_i)$, replace the cached coordinates $c_i = (\theta_i, \rho_i)^\top$ with these features. Each cell supplies one column of the readout's correspondence tensor:

$$\underbrace{C = \big[(\theta_i, \rho_i)^\top\big]_{i=1}^N \in \mathbb{R}^{2\times N}}_{\text{Hesse Correspondence Tensor}} \quad \longrightarrow \quad \underbrace{V = \big[\nu_2(\ell(\theta_i, \rho_i))\big]_{i=1}^N \in \mathbb{R}^{6\times N}}_{\text{Veronese Correspondence Tensor}}. \tag{8}$$

Both operators use the same normalised heatmap probabilities $p = \mathrm{vec}(P)$, with $p_i \geq 0$ and $\sum_i p_i = 1$:

$$\underbrace{(\hat{\theta}, \hat{\rho})^\top = \sum_i p_i c_i = Cp}_{\text{soft-argmax}} \quad \longrightarrow \quad \underbrace{\hat{\nu} = \sum_i p_i \nu_2(x_i) = Vp\,.}_{\text{VSmax}} \tag{9}$$

The aggregation applies to arbitrary cell vectors $x_i \in S^2$ and needs no Hough grid. Its result is a convex combination of embedded cells, hence in general not the embedding of a single line.

3.2 VERONESE SUPERVISION

Coordinate readouts are supervised on their decoded parameters $(\hat{\theta}, \hat{\rho}) = Cp$, with offsets normalised to the image (Appendix C.4). VSmax instead supervises the aggregate $\hat{\nu} = Vp$ before recovery, against the embedding of the ground-truth line. Weighting the three mixed products by $\sqrt{2}$ defines

$$W = \mathrm{diag}(1, 1, 1, \sqrt{2}, \sqrt{2}, \sqrt{2}), \qquad v_2(x) = W\nu_2(x), \qquad \hat{v} = W\hat{\nu}. \tag{10}$$

For unit representatives, the weighted feature distance satisfies (proof in Appendix B.2)

$$\|v_2(x) - v_2(y)\|_2^2 = 2d([x], [y])^2, \tag{11}$$

an isometry to the chordal distance up to the fixed factor $\sqrt{2}$. For target $[x^\star] = [\ell(\theta^\star, \rho^\star)]$ the two objectives are

$$\underbrace{\mathcal{L}_{\mathrm{hs}} = \sin^2(\hat{\theta} - \theta^\star) + (\hat{\rho} - \rho^\star)^2}_{\text{Hesse supervision}} \quad \longrightarrow \quad \underbrace{\mathcal{L}_{\mathrm{vs}} = \|W(\hat{\nu} - \nu_2(x^\star))\|_2^2 = \|\hat{v} - v_2(x^\star)\|_2^2}_{\text{Veronese supervision}}. \tag{12}$$

$\mathcal{L}_{\mathrm{hs}}$ inherits the failure stated in Proposition 1. Its π-periodic term handles the angular wrap alone and ignores its coupling to ρ. $\mathcal{L}_{\mathrm{vs}}$ acts on the aggregate itself. Its gradients pass through aggregation and probability normalisation, but not through the eigendecomposition. Training through that projection is standard in rotation estimation (Peretroukhin et al., 2020; Zhou et al., 2019; Levinson et al., 2020), but eigenvector derivatives scale with the inverse eigengap. Here the recovered line is the final output, so we place the objective before recovery. Theorem 1(iii) bounds its chordal error by $\mathcal{L}_{\mathrm{vs}}$.

3.3 GEOMETRIC RECOVERY

Recovering a line from the embedding means choosing the one closest to the cells in the chordal metric, solved in closed form once the aggregate is written as a matrix. With $Q(x) = xx^\top$ and half-vectorisation vech retaining its distinct entries in the order of Equation 7,

$$M = \mathrm{vech}^{-1}(\hat{\nu}) = \mathbb{E}_{X \sim P}\left[XX^\top\right] = \sum_i p_i Q(x_i). \tag{13}$$

M is the second-moment matrix of the heatmap, a probability-weighted mixture of the cells' line projectors. It is symmetric, positive semidefinite, of unit trace, and of rank one exactly when all probability lies on one line. By linearity, the expected squared chordal distance of a unit y to the cells is the *recovery objective*

$$\sum_i p_i \, d([y], [x_i])^2 = 1 - y^\top M y = 1 - \sum_i p_i (y^\top x_i)^2. \tag{14}$$

This is minimised by $\hat{x} \in \arg\max_{\|y\|_2 = 1} y^\top M y$, a leading eigenvector of M by the Rayleigh–Ritz theorem, which also spans its best rank-one approximation (Eckart & Young, 1936). VSmax returns the projective point $[\hat{x}]$, the extrinsic mean of the cells (Lin et al., 2017; Markley et al., 2007), recovered as in rotation prediction (Peretroukhin et al., 2020). The direction is unique when $\lambda_1(M) > \lambda_2(M)$, for which Theorem 1 gives sufficient conditions.

4 THEORETICAL ANALYSIS

Section 2.2 asked two things of a projective readout. Invariance to the choice of representative (Definition 1) and accuracy on evidence that concentrates on the target (Definition 2). Canonical coordinate readouts can meet the first but, by Proposition 1, cannot meet the second. We now prove that VSmax meets both, and provide a third guarantee bounding the recovered error by the training loss $\mathcal{L}_{\mathrm{vs}}$ of Equation 12.

Theorem 1 (Representative-invariant and concentration-consistent recovery)**.** *Let P be a heatmap with cells $x_i \in S^2$ (Section 2.2), $x^\star \in S^2$ a representative of the target and $\bar{d} = \bar{d}(P, x^\star)$ (Equation 4). Let $\hat{x}$ be any unit leading eigenvector of the second-moment matrix M of Equation 13, whose eigenvalues are $\lambda_1(M) \geq \lambda_2(M) \geq \lambda_3(M)$.*

(i) ***Representative invariance.*** *M and $\hat{v}$ depend only on the class distribution $\bar{P}$, as does the set of optimal projective directions. When $\lambda_1(M) > \lambda_2(M)$, the recovered projective point is unique and VSmax is representative-invariant independently of the eigensolver's selection convention.*

(ii) ***Concentration-consistent recovery.*** *If $\bar{d} < 1/2$, the recovered projective point is unique and*

$$\lambda_1(M) - \lambda_2(M) \geq 1 - 2\bar{d}, \qquad d([\hat{x}], [x^\star]) \leq \varphi(\bar{d}), \quad \varphi(t) = \frac{t}{1-t}. \tag{15}$$

(iii) ***Supervision and recovered error.*** *Without any assumption on $\bar{d}$,*

$$d([\hat{x}], [x^\star])^2 \leq 2\mathcal{L}_{\mathrm{vs}}, \qquad \lambda_1(M) - \lambda_2(M) \geq 1 - \sqrt{2\mathcal{L}_{\mathrm{vs}}}, \tag{16}$$

so $\mathcal{L}_{\mathrm{vs}} < 1/2$ also suffices for uniqueness.

Appendix B.6 proves the theorem on $\mathbb{RP}^{m-1}$ for any $m \geq 2$. In words, part (i) says that moving probability between x and $-x$ leaves the prediction untouched. The invariance is built into the features themselves, since $\nu_2(x) = \nu_2(-x)$ makes antipodal representatives one and the same point of the embedding, so no step of the readout can depend on the choice. Coordinate readouts reach it only by folding the heatmap first (Proposition B.1).

Part (ii) is the property coordinate readouts cannot obtain. By Equation 14, each cell contributes $p_i(y^\top x_i)^2 \geq 0$ to $y^\top M y$. The two cells of one line contribute the same amount, since $(y^\top x_i)^2 = (y^\top(-x_i))^2$. Evidence therefore adds up and cannot cancel as the offsets do in Proposition 1. VSmax averages projectors, whose distance is the line distance, $\|Q(x) - Q(y)\|_{\mathrm{op}} = d([x], [y])$, so $\|M - Q(x^\star)\|_{\mathrm{op}} \leq \bar{d}$. Averaging cannot move the aggregate farther from the target than the evidence is. Standard and sinusoidal soft-argmax average chart coordinates, in which nearby lines can lie far apart, so near the angular boundary a small $\bar{d}$ gives no control over their prediction (Figure 7 in Appendix D.1). The first inequality of (ii) keeps the eigengap positive, so recovery stays well-conditioned. The second holds with the same modulus $\varphi(t) = t/(1-t)$ for every target, which is stronger than Definition 2 requires.

$\mathcal{L}_{\mathrm{vs}}$ compares an aggregate of embedded cells with the embedding of the target line, not one line with another. Part (iii) is what connects it to the predicted line, while assuming nothing about how concentrated the heatmap is. Whatever the evidence, a small $\mathcal{L}_{\mathrm{vs}}$ bounds the chordal error of the recovered line, and $\mathcal{L}_{\mathrm{vs}} < 1/2$ ensures that line is unique. Both bounds hold on every prediction of the trained pipeline (Section 5.2).

5 Experiments

5.1 Synthetic setup

Setup and data. The known-operator pipeline in Figure 2 applies a fixed Hough transform, a 37.6K-parameter refinement CNN with cylindrical convolutions (Kim et al., 2020) on the double cover of Hough space (Appendices C.2 and C.3), and a readout. We compare standard soft-argmax, sinusoidal soft-argmax and VSmax on this architecture. VSmax is invariant to the duplication, whereas the coordinate baselines first fold antipodal probabilities onto the canonical domain (Proposition B.1). Training samples synthetic 256×256 single-line images over the 127×127 Hough grid, with zero-mean Gaussian pixel noise $\sigma \sim \mathcal{U}[0, 0.5]$. Early stopping uses loss on a fixed $1{,}000$-image clean validation set, and all synthetic configurations share optimisation settings (Appendix C.4). We test the same $14{,}383$ lines at $\sigma \in \{0.5, 0.6, 0.7, 0.8\}$, where $\sigma = 0.5$ is the in-distribution (ID) reference at the training boundary and $\sigma \in \{0.6, 0.7, 0.8\}$ are out-of-distribution (OOD) levels beyond it. The bound analysis uses the complete $\sigma \in [0, 0.8]$ sweep to span the observed evidence dispersion.

Supervision and metrics. Coordinate readouts are trained with $\mathcal{L}_{\mathrm{hs}}$ and VSmax with $\mathcal{L}_{\mathrm{vs}}$ of Equation 12. Method comparisons report mean Euclidean and Angular distance score (EA-score) (Zhao et al., 2022), an angular-and-positional agreement score in $[0, 1]$ with higher values better (Appendix C.5). The bound analysis reports the chordal distance d of Equation 2, the quantity Theorem 1 controls, with the loss $\mathcal{L}_{\mathrm{vs}}$, the eigengap and the evidence dispersion.

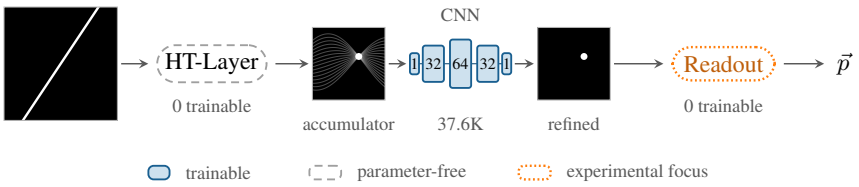


Figure 2: Known-operator line regression: a fixed Hough transform, learned accumulator refinement and a differentiable readout predicting $\vec{p}$. Filled modules are trainable, dashed or dotted operators are parameter-free. The highlighted readout is varied across the main comparisons.

5.2 Recovery guarantees in practice

Both bounds of Theorem 1 hold on the trained model. We evaluate them on the 129,447 predictions of the VSmax pipeline across $\sigma \in [0, 0.8]$. Each stored embedding average gives M, the loss, the eigengap and the decoded error. Part (iii) bounds the decoded error by $\sqrt{2\mathcal{L}_{\text{vs}}}$ for any evidence, and Figure 3a places that bound on the diagonal. The per-bin maximum stays below it over the whole range, without a single violation. For part (ii) we measure dispersion by the RMS dispersion $\bar{d}_2 = \sqrt{1 - x^{\star\top} M x^{\star}}$, the recovery objective of Section 3.3 evaluated at the target and read directly from the aggregate. It dominates $\bar{d}$, so the hypothesis and bound of part (ii) hold at $\bar{d}_2$ (Appendix D.2). Across noise levels (Figure 3b), mean $\bar{d}_2$ rises from 0.027 to 0.4955, the decoded error stays well below it throughout, and the eigengap falls only from 0.998 to 0.622, so recovery never approaches the degenerate regime.

5.3 Readout and supervision controls

Does the readout alone matter? VSmax leads at every tested noise level and the margin widens with corruption (Figure 4a). The heatmaps are those of one frozen VSmax checkpoint and only the inference readout changes, so the difference is attributable to aggregation alone.

Does the objective alone matter? With the VSmax pipeline fixed, $\mathcal{L}_{\text{vs}}$ leads at every tested level, reaching mean EA-score 0.636 at $\sigma = 0.8$ against 0.333 for $\mathcal{L}_{\text{hs}}$ differentiated through recovery (Figure 4b). The training objective matters even when inference uses the same readout.

How do complete pipelines compare? Releasing both controls, separately trained pipelines with each readout under its native objective ($\mathcal{L}_{\text{hs}}$ for the coordinate readouts, $\mathcal{L}_{\text{vs}}$ for VSmax) reach

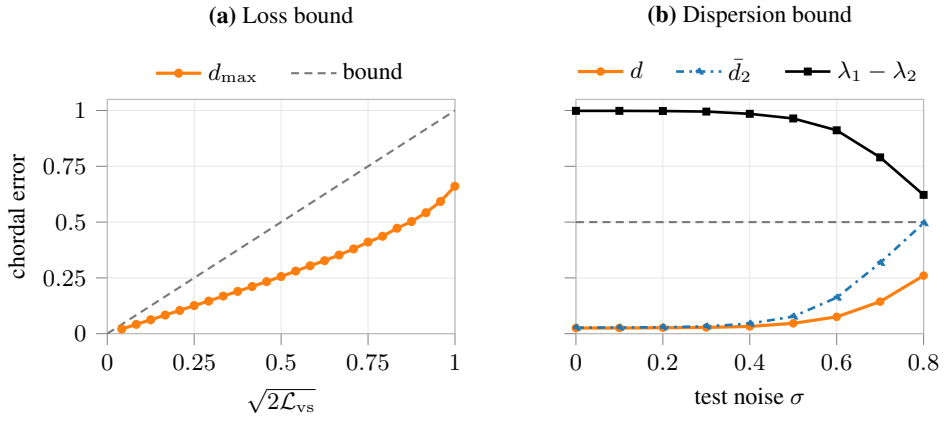


Figure 3: Recovery bounds over 129,447 predictions. (a) Maximum chordal error $d_{\max}$ over the predictions in each bin of $\sqrt{2\mathcal{L}_{\text{vs}}}$, with the part (iii) bound dashed. (b) Chordal error d, RMS dispersion $\bar{d}_2$ and eigengap $\lambda_1 - \lambda_2$, each averaged over the predictions at a given noise level σ. The line at $1/2$ marks the per-example certificate $\bar{d}_2 < 1/2$ of part (ii).

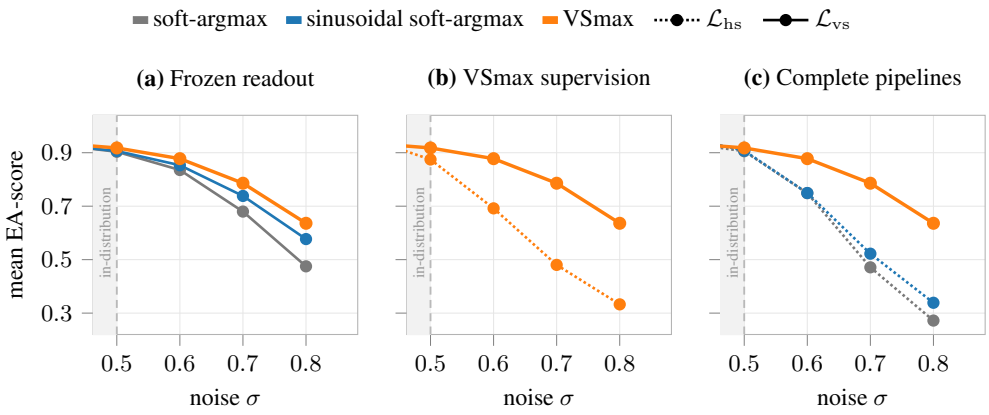


Figure 4: Mean EA-score over 14,383 paired test lines. Colour identifies the readout, dotted and solid curves $\mathcal{L}_{\text{hs}}$ and $\mathcal{L}_{\text{vs}}$. (a) Frozen $\mathcal{L}_{\text{vs}}$-trained checkpoint with only the inference readout changed. (b) VSmax retrained under each objective. (c) Separately trained pipelines, each readout with its native objective. The solid orange curve is the same checkpoint throughout. Dashed lines mark the training maximum $\sigma = 0.5$, with the training side shaded. Standard deviations are listed in Table 3.

mean EA-score 0.272 and 0.339 for standard and sinusoidal soft-argmax at $\sigma = 0.8$, against 0.636 for VSmax, which leads at every tested level (Figure 4c). The gap exceeds either single control, and the baselines score lower than on the Veronese-trained heatmap, so readout and objective contribute jointly rather than redundantly.

Where do the errors occur? The cost of coordinate readouts is not confined to the angular boundary. Figure 5 maps per-line EA-score at $\sigma = 0.6$, the first OOD level, for the coordinate baselines retrained with the same Veronese targets, metric, architecture and optimisation as VSmax, supervising embeddings of decoded lines rather than the pre-recovery average (mean EA-score 0.674, 0.700 and 0.878). Baseline low-score regions are orientation-dependent and reach across the chart interior. VSmax holds a broader high-scoring region, and its remaining errors sit at extreme offsets where little of the line is visible.

Appendix D.4 ablates the known-operator pipeline for a CNN and multi-layer perceptron (MLP) regressor, which generalises worse to held-out regions despite roughly $300\times$ more parameters. Appendix D.3 compares across the full noise sweep against a classical Hough transform with argmax

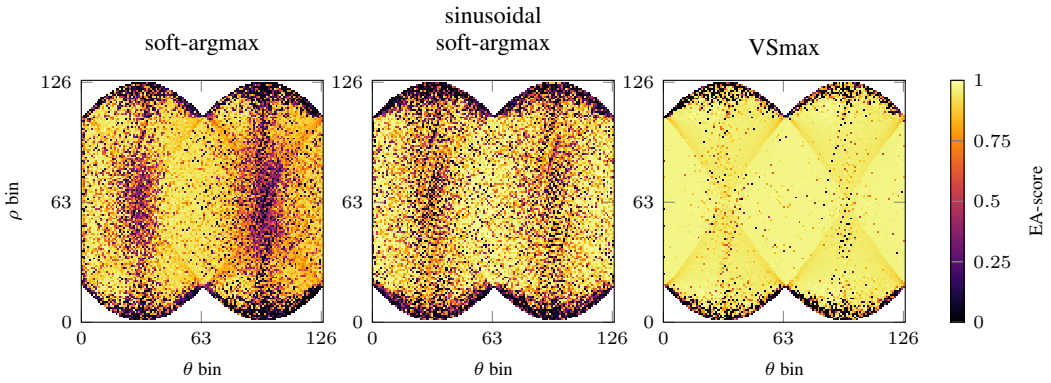


Figure 5: Per-line EA-scores at $\sigma = 0.6$ for separately trained double-cover pipelines with common Veronese targets: standard soft-argmax, sinusoidal soft-argmax and VSmax (left to right). Scores are shown at ground-truth Hough bins. White bins contain no evaluated line.

Table 1: HLW-visible normalised $\mathrm{AUC}_{0.25}$ (↑) and approximate total parameters (M). Results concern the filtered subset, not full HLW.

	HLW-CNN Workman et al. (2016)	LETR Xu et al. (2021)	LINEA Janampa & Pattichis (2025)	Ours
$\mathrm{AUC}_{0.25}$ (ID)	0.7963	0.7993	0.6949	**0.8133**
$\mathrm{AUC}_{0.25}$ (OOD)	0.7859	0.7574	0.7496	**0.8063**
$\mathrm{AUC}_{0.25}$ (combined)	0.7927	0.7847	0.7139	**0.8109**
Parameters (M)	10.5	41.5	3.9	6.3

readout and the heatmap-supervised baseline of Zhao et al. (2022), neither differentiable through the readout.

5.4 Transfer to real images

Setup. A U-Net (Ronneberger et al., 2015) feature extractor produces the evidence for the fixed Hough transform, followed by the refinement and VSmax readout above. We evaluate on HLW-visible, a visibility-filtered subset of Horizon Lines in the Wild (Workman et al., 2016) with 934 seen-scene (ID) and 499 held-out-scene (OOD) images, and report the dataset's $\mathrm{AUC}_{0.25}$, the normalised area under the horizon-error distribution up to 0.25 (Equation 36). Swept baselines use their best combined-test configuration, fixed across splits (Appendix D.5).

Result. Our model attains the highest $\mathrm{AUC}_{0.25}$ on both seen and held-out scenes among the evaluated models (Table 1). The gain comes from fewer large errors rather than better median accuracy. Only 1.05% of our predictions have error at least 0.25, against 3.91% for HLW-CNN (Figure 14).

6 Conclusions

Architectural priors shape how evidence is carried, but none of them has to name the object. A readout does. It is the one component that must return an element of the label space itself, so it cannot stay agnostic to how that space is shaped, and an operator that averages coordinates is only as correct as the chart it averages in. What VSmax changes is therefore not the operation but the space handed to it. That exchange asks for three things. An embedding under which the identifications of the label space become equalities, a linear space to average in, and an inverse in closed form. None of the three introduces a trainable parameter, so whether the readout is well-posed is decided by design, before training and independently of any validation error. Our bounds are of that kind. They hold for the operator, not for a particular fit.

The same exchange settles the objective, since a loss is a function on the label space as well, and the controls show that the two consequences are separable. A readout that respects the identification improves a fixed heatmap, an objective that respects it improves the heatmap the network learns, and neither substitutes for the other. Quotients by finite groups admit polynomial invariant embeddings of the same kind, and the choice of embedding fixes both the operator and the metric its loss measures in. The reach of a differentiable argmax is thus bounded not by flat domains but by the embeddings available for the space one wants to read out. We prove the bounds for real projective spaces of any dimension, while the evidence here concerns the real projective plane.

Outlook. Just as soft-argmax estimates one coordinate per channel, VSmax is a single-target readout per heatmap. It extends directly to labelled multi-channel outputs $\hat{\ell}_k = \mathrm{VSmax}(P_k)$, and, since the bounds hold on $\mathbb{RP}^{m-1}$, to any projective label space with a heatmap over candidates.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing the text. We did not use generative AI tools to generate research ideas, theoretical results, or experimental data. All AI-assisted text was reviewed by the authors. We take responsibility for the final content of this work, including all text, claims and artifacts.

ETHICS STATEMENT

This work does not involve human subjects or the collection of new data. Our real-image evaluation uses a filtered subset of the publicly released Horizon Lines in the Wild dataset (Workman et al., 2016). We redistribute no imagery and report only aggregate geometric error statistics. The filtering criteria and the resulting sample counts are documented in Appendix D.5. We are not aware of harmful applications specific to geometric line readout beyond those already associated with camera calibration and scene understanding.

REPRODUCIBILITY STATEMENT

Full proofs of Proposition 1 and of Theorem 1 are given in Appendix B. Architecture, Hough accumulation, readout implementation, optimiser settings and evaluation conventions are specified in Appendix C. The HLW-visible protocol, configuration-selection rule, full query sweeps and the remaining protocol checks we were unable to close are documented in Appendix D.5.

REFERENCES

Syed Ammar Abbas and Andrew Zisserman. A Geometric Approach to Obtain a Bird's Eye View From an Image. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4095–4104, October 2019. doi: 10.1109/ICCVW.2019.00504.

Magdalena Bachmaier, Maximilian Rohleder, Benedict Swartman, Maxim Privalov, Andreas Maier, and Holger Kunze. Robust Hough and Spatial-To-Angular Transform Based Rotation Estimation for Orthopedic X-Ray Images. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 446–455, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43990-2. doi: 10.1007/978-3-031-43990-2_42.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics*, 31(1):1–29, February 2003. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1046294456.

Taco Cohen and Max Welling. Group Equivariant Convolutional Networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 2990–2999. PMLR, June 2016.

John H. Conway, Ronald H. Hardin, and Neil J. A. Sloane. Packing Lines, Planes, etc.: Packings in Grassmannian Spaces. *Experimental Mathematics*, 5(2):139–159, January 1996. ISSN 1058-6458. doi: 10.1080/10586458.1996.10504585.

Chandler Davis and W. M. Kahan. The Rotation of Eigenvectors by a Perturbation. III. *SIAM J. Numer. Anal.*, 7(1):1–46, March 1970. ISSN 0036-1429. doi: 10.1137/0707001.

Richard O. Duda and Peter E. Hart. Use of the Hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, January 1972. ISSN 0001-0782. doi: 10.1145/361237.361242.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José M. F. Moura. Adversarial Geometry-Aware Human Motion Prediction. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 823–842, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01225-0. doi: 10.1007/978-3-030-01225-0_48.

Gyu-Sung Ham and Kanghan Oh. Learning Spatial Configuration Feature for Landmark Localization in Hand X-rays. *Electronics*, 12(19):4038, January 2023. ISSN 2079-9292. doi: 10.3390/electronics12194038.

Jihun Hamm and Daniel D. Lee. Grassmann discriminant analysis: A unifying view on subspace-based learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 376–383, New York, NY, USA, July 2008. Association for Computing Machinery. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390204.

Mehrtash T. Harandi, Mathieu Salzmann, Sadeep Jayasumana, Richard Hartley, and Hongdong Li. Expanding the Family of Grassmannian Kernels: An Embedding Perspective. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 408–423, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10584-0. doi: 10.1007/978-3-319-10584-0_27.

Joe Harris. *Algebraic Geometry*, volume 133 of *Graduate Texts in Mathematics*. Springer New York, New York, NY, 1992. ISBN 978-1-4419-3099-6 978-1-4757-2189-8. doi: 10.1007/978-1-4757-2189-8.

Sebastian Janampa and Marios Pattichis. Linea: Fast and Accurate Line Detection using Scalable Transformers. In *2025 IEEE International Conference on Image Processing (ICIP)*, pp. 2115–2120, September 2025. doi: 10.1109/ICIP55913.2025.11084517.

Sunghun Joung, Seungryong Kim, Hanjae Kim, Minsu Kim, Ig-Jae Kim, Junghyun Cho, and Kwanghoon Sohn. Cylindrical Convolutional Networks for Joint Object Detection and Viewpoint Estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14151–14160, June 2020. doi: 10.1109/CVPR42600.2020.01417.

Jinpyo Kim, Wooekun Jung, Hyungmo Kim, and Jaejin Lee. CyCNN: A Rotation Invariant CNN using Polar Mapping and Cylindrical Convolution Layers, July 2020.

Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An Analysis of SVD for Deep Rotation Estimation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22554–22565. Curran Associates, Inc., 2020.

Lizhen Lin, Brian St Thomas, Hongtu Zhu, and David B. Dunson. Extrinsic local regression on manifold-valued data. *J Am Stat Assoc*, 112(519):1261–1273, 2017. ISSN 0162-1459. doi: 10.1080/01621459.2016.1208615.

Yancong Lin, Silvia L. Pintea, and Jan C. van Gemert. Deep Hough-Transform Line Priors, July 2020.

Diogo C. Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, December 2019. ISSN 00978493. doi: 10.1016/j.cag.2019.09.002.

Andreas K. Maier, Christopher Syben, Bernhard Stimpel, Tobias Würfl, Mathis Hoffmann, Frank Schebesch, Weilin Fu, Leonid Mill, Lasse Kling, and Silke Christiansen. Learning with known operators reduces maximum error bounds. *Nat Mach Intell*, 1(8):373–380, August 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0077-5.

K. V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. J. Wiley, Chichester ; New York, 2000. ISBN 978-0-471-95333-3.

F. Landis Markley, Yang Cheng, John L. Crassidis, and Yaakov Oshman. Averaging Quaternions. *Journal of Guidance, Control, and Dynamics*, 30(4):1193–1197, July 2007. ISSN 0731-5090, 1533-3884. doi: 10.2514/1.28949.

Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical Coordinate Regression with Convolutional Neural Networks, May 2018.

Valentin Peretroukhin, Matthew Giamou, David M. Rosen, W. Nicholas Greene, Nicholas Roy, and Jonathan Kelly. A Smooth Representation of Belief over SO(3) for Deep Rotation Learning with Uncertainty. In *Robotics: Science and Systems XVI*, July 2020. doi: 10.15607/RSS.2020.XVI.007.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

Lalith Sharan, Gabriele Romano, Julian Brand, Halvar Kelm, Matthias Karck, Raffaele De Simone, and Sandy Engelhardt. Point detection through multi-instance deep heatmap regression for sutures in endoscopy. *Int J CARS*, 16(12):2107–2117, December 2021. ISSN 1861-6429. doi: 10.1007/s11548-021-02523-w.

Max Steiger, Mohammad Rezapourian, Marko Rak, and Christian Hansen. Dynamic uncertainty-level assessment framework for real-time needle tracking in CT-guided surgical environments. *Int J CARS*, June 2026. ISSN 1861-6429. doi: 10.1007/s11548-026-03725-w.

John Stillwell. *Geometry of Surfaces*. Springer Science & Business Media, February 1995. ISBN 978-0-387-97743-0.

Maurice Weiler, Patrick Forré, Erik Verlinde, and Max Welling. Coordinate Independent Convolutional Networks – Isometry and Gauge Equivariant Convolutions on Riemannian Manifolds, June 2021.

Hermann Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann.*, 71 (4):441–479, December 1912. ISSN 1432-1807. doi: 10.1007/BF01456804.

Agata M. Wijata, Bartłomiej Pyciński, and Jakub Nalepa. A Needle In A (Medical) Haystack: Detecting A Biopsy Needle In Ultrasound Images Using Vision Transformers. In *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 3017–3023, October 2024. doi: 10.1109/ICIP51287.2024.10647846.

Scott Workman, Menghua Zhai, and Nathan Jacobs. Horizon lines in the wild: 27th British Machine Vision Conference, BMVC 2016. *British Machine Vision Conference 2016, BMVC 2016*, pp. 20.1–20.12, 2016. ISSN 1901725596. doi: 10.5244/C.30.20.

Yifan Xu, Weijian Xu, David Cheung, and Zhuowen Tu. Line Segment Detection Using Transformers Without Edges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4257–4266, 2021.

Xue Yang and Junchi Yan. Arbitrary-Oriented Object Detection with Circular Smooth Label. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 677–694, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58598-3. doi: 10.1007/978-3-030-58598-3_40.

Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11830–11841. PMLR, July 2021.

Kai Zhao, Qi Han, Chang-Bin Zhang, Jun Xu, and Ming-Ming Cheng. Deep Hough Transform for Semantic Line Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (9):4793–4806, September 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3077129.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5738–5746, June 2019. doi: 10.1109/CVPR.2019.00589.

APPENDIX CONTENTS

A Numerical example of geometric proximity across the angular boundary **15**

B Mathematical foundations and proofs **16**

B.1 General Veronese map 16

B.2 Projective distance and embedding identities 16

B.3 Canonicalisation and representative invariance 17

B.4 Failure of representative-vector averaging 17

B.5 Failure of coordinate readouts 18

B.6 Proof of the VSmax guarantees 19

C Implementation and reproducibility **21**

C.1 Hough accumulation and line decoding 21

C.2 Double-cover fold 21

C.3 Gauge-equivariant interpretation 21

C.4 Training setup and baseline configurations 22

C.5 EA score and evaluation conventions 22

D Additional experiments **24**

D.1 Readouts on constructed heatmaps 24

D.2 Empirical verification of the recovery guarantees 25

D.3 Classical and heatmap-supervised Hough comparisons 26

D.4 Broader pipeline ablations 27

D.5 HLW-visible horizon estimation 30

A NUMERICAL EXAMPLE OF GEOMETRIC PROXIMITY ACROSS THE ANGULAR BOUNDARY

With the image origin at its centre and half-width as the unit of length, the lines $(\theta_1, \rho_1) = (5°, 0.4)$ and $(\theta_2, \rho_2) = (175°, -0.4)$ have chordal distance approximately 0.161, but Euclidean coordinate distance 3.073 (angles in radians), as illustrated in Figure 6.

Value ranges. Neither magnitude means anything on its own, so we state both ranges formally. Write $p_i = (\theta_i, \rho_i)$ for a point of the parameter domain $\mathcal{D} = [0, \pi) \times [-1, 1]$ of Figure 6, and $\ell(p_i)$ for its unit coefficient vector (Equation 1). The two quantities compared above are

$$\begin{aligned} d\big([\ell(p_1)], [\ell(p_2)]\big) &= \sqrt{1 - \big(\ell(p_1)^\top \ell(p_2)\big)^2}, \\ d_{\text{euc}}(p_1, p_2) &= \|p_1 - p_2\|_2 = \sqrt{(\theta_1 - \theta_2)^2 + (\rho_1 - \rho_2)^2}, \end{aligned} \tag{17}$$

with angles in radians. Over all $p_1, p_2 \in \mathcal{D}$ their ranges are exactly

$$d \in [0, 1], \qquad d_{\text{euc}} \in \big[0, \sqrt{\pi^2 + 4}\,\big), \qquad \sqrt{\pi^2 + 4} \approx 3.724. \tag{18}$$

Both endpoints of the d range are attained. $d = 0$ holds for $p_1 = p_2$, and $d = 1$ whenever the two coefficient vectors are orthogonal, as for $\ell(0, 0)$ and $\ell(\pi/2, 0)$, the vertical and horizontal lines through the image centre. The upper bound on d_{euc} is the diagonal of $\mathcal{D}$. It is a supremum only, approached as $|\theta_1 - \theta_2| \to \pi$ with $|\rho_1 - \rho_2| = 2$, but never attained because $\theta < \pi$. Both ranges are full intervals, since each quantity is continuous on the connected domain $\mathcal{D} \times \mathcal{D}$.

Measured against these ranges, the two lines above are $0.161/1 = 16\%$ apart geometrically and $3.073/3.724 = 83\%$ apart in coordinates. Lines that nearly coincide sit almost as far apart as the parametrisation allows.

For the family $(\varepsilon, a), (\pi - \varepsilon, -a)$, the former vanishes as $\varepsilon \to 0$, while the latter tends to $\sqrt{\pi^2 + 4a^2} > 0$. At $a = 0.4$ that limit is 3.242, or 87% of the d_{euc} range, and at $|a| = 1$ it exhausts that range. The contrast concerns this limiting behaviour and the fraction of each range it occupies, not the ratio of the two raw magnitudes. The same family exposes the readout failure in Proposition 1.

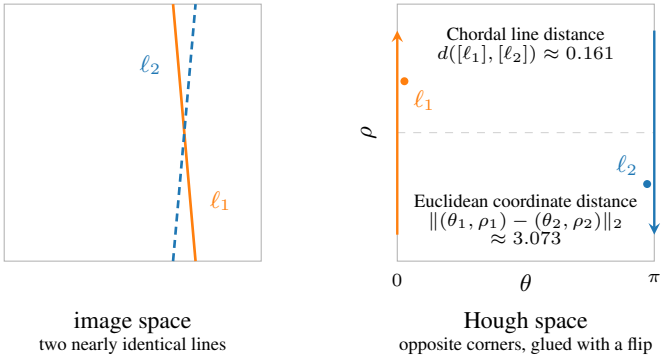


Figure 6: Nearby undirected lines (left) occupy opposite ends of the parameter domain (right). The angular edges are identified with a sign flip on ρ. Both panels use $(\theta_1, \rho_1) = (\varepsilon, a)$ and $(\theta_2, \rho_2) = (\pi - \varepsilon, -a)$, with $\varepsilon = 5°$ and $a = 0.4$. The image origin is at its centre, with half-width as the unit of length, and the parameter domain is $[0, \pi) \times [-1, 1]$. The annotations compare chordal line separation with Euclidean coordinate distance (angles in radians). As $\varepsilon \to 0$, only the former vanishes. The raw magnitudes are on different scales. Over this domain the two metrics have maxima 1 and $\sqrt{\pi^2 + 4} \approx 3.724$, so the annotated values are 16% and 83% of their respective ranges.

B MATHEMATICAL FOUNDATIONS AND PROOFS

B.1 GENERAL VERONESE MAP

The Veronese map ν_d (Harris, 1992) represents a projective point by all its degree-d monomials:

$$\nu_d : \mathbb{P}^n \longrightarrow \mathbb{P}^N, \qquad [x_0 : \cdots : x_n] \longmapsto [x_0^d : x_0^{d-1}x_1 : \cdots : x_n^d]. \tag{19}$$

Here d is the polynomial degree, n the source dimension, and $N = \binom{n+d}{d} - 1$ the target dimension. For our undirected-line targets, $n = 2$ and $d = 2$ give the six distinct entries of $xx^\top$. The projective map specifies these entries only up to scale. Choosing a unit representative $x \in S^2$ fixes that scale, and the even degree removes the remaining sign ambiguity. The resulting vector-space realisation is $\nu_2(x) = \mathrm{vech}(xx^\top)$ from Section 3.1.

B.2 PROJECTIVE DISTANCE AND EMBEDDING IDENTITIES

The following identity compares two line projectors, not an arbitrary embedding average with a decoded line. Write $h(\mathbf{u}) = \mathrm{vech}(\mathbf{u}\mathbf{u}^\top)$ for the unweighted half-vectorisation, so the features of Equation 10 are $v_2(\mathbf{u}) = Wh(\mathbf{u})$. The unweighted vector norm counts the six distinct entries once:

$$\|h(\mathbf{u}) - h(\mathbf{v})\|_2^2 = \sum_{i \le j} (u_i u_j - v_i v_j)^2, \tag{20}$$

whereas the Frobenius norm runs over all nine matrix entries,

$$\|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F^2 = \sum_{i,j} (u_i u_j - v_i v_j)^2, \tag{21}$$

so the difference is the multiplicity of off-diagonal entries. The weighting $W = \mathrm{diag}(1, 1, 1, \sqrt{2}, \sqrt{2}, \sqrt{2})$ gives $\|v_2(\mathbf{u}) - v_2(\mathbf{v})\|_2^2 = \|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F^2$. The unweighted loss is coordinate-dependent and is not, in general, a function of projective angle alone.

The second step identifies this Frobenius distance geometrically. For unit vectors $\mathbf{u}, \mathbf{v} \in S^n$, expanding the trace gives

$$\begin{aligned}
\|\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top\|_F^2 &= \mathrm{tr}\big((\mathbf{u}\mathbf{u}^\top - \mathbf{v}\mathbf{v}^\top)^2\big) \\
&= \mathrm{tr}(\mathbf{u}\mathbf{u}^\top\mathbf{u}\mathbf{u}^\top) - 2\,\mathrm{tr}(\mathbf{u}\mathbf{u}^\top\mathbf{v}\mathbf{v}^\top) + \mathrm{tr}(\mathbf{v}\mathbf{v}^\top\mathbf{v}\mathbf{v}^\top) \\
&= \|\mathbf{u}\|^4 - 2(\mathbf{u}^\top\mathbf{v})^2 + \|\mathbf{v}\|^4 \\
&= 2 - 2(\mathbf{u}^\top\mathbf{v})^2 \\
&= 2\big(1 - \cos^2\alpha\big) \\
&= 2\sin^2\alpha,
\end{aligned} \tag{22}$$

where $\alpha = \arccos|\mathbf{u}^\top\mathbf{v}| \in [0, \frac{\pi}{2}]$ is the angle between the projective points $[\mathbf{u}], [\mathbf{v}] \in \mathbb{RP}^n$. The right-hand side is twice $d([\mathbf{u}], [\mathbf{v}])^2$ under the normalisation in Equation 2. It is invariant to the signs of $\mathbf{u}$ and $\mathbf{v}$, hence well defined on undirected lines ($n = 2$). Chaining the two steps yields Equation 11, so the squared Euclidean distance between weighted embeddings equals twice the squared chordal distance d. This is the projection form of the Grassmannian chordal distance (Conway et al., 1996; Hamm & Lee, 2008).

Why the chordal metric. The geodesic angle α is also a metric on $\mathbb{RP}^n$. The squared chordal distance is the one to train against, for four reasons.

1. **Euclidean pullback.** $d^2 = 1 - (\mathbf{u}^\top\mathbf{v})^2$ is a polynomial in the representative, so it is the pullback of a Euclidean distance under ν_2 by Equation 11. By contrast, $\alpha = \arccos|\mathbf{u}^\top\mathbf{v}|$ admits no finite-dimensional isometric Euclidean embedding, since an isometry would map the two halves of a closed geodesic, which join the same two points at maximal distance, onto the same straight segment.

2. **Extension to mixtures.** Being quadratic, it extends linearly, $\sum_i p_i\, d([x_i],[x^\star])^2 = 1 - x^{\star\top} M x^\star$, so the objective is defined on the aggregate, is what recovery minimises (Section 3.3) and is read off as $\bar{d}_2$ (Appendix D.2). The geodesic analogue $\sum_i p_i \alpha_i^2$ is not a function of M.
3. **Bounded gradient.** $|\partial d^2/\partial\alpha| = |\sin 2\alpha| \le 1$ everywhere, whereas $\partial\alpha/\partial(\mathbf{u}^\top\mathbf{v}) = -1/\sin\alpha$ diverges at the target and α^2 is non-smooth at $\alpha = \pi/2$, the cut locus, through the absolute value. The same behaviour is documented for the geodesic loss on $SO(3)$ (Gui et al., 2018).
4. **Price.** The gradient vanishes at $\alpha = \pi/2$ alone, an isolated unstable maximum at which the prediction is orthogonal to the target.

Near the target the two metrics agree to second order, $d = \sin\alpha = \alpha + O(\alpha^3)$, which is the regime the bounds of Theorem 1 address.

B.3 Canonicalisation and representative invariance

Applied directly to noncanonical coordinates, the coordinate readouts need not be representative-invariant. All mass at $(0, a)$ and equal mass at $(0, a), (\pi, -a)$ represent the same line, but yield offsets a and 0. Canonicalisation removes this dependence on the chosen representatives.

Proposition B.1 (Canonicalisation restores representative invariance)**.** *Let $s\colon \mathbb{RP}^2 \to S^2$ select one unit representative per target, so $[s(q)] = q$. Such a map is called a* section*. Define*

$$C_s(P) = \sum_{q \in \operatorname{supp} \bar{P}} \bar{P}(q)\, \delta_{s(q)}, \tag{23}$$

where $\delta_{s(q)}$ assigns probability one to $s(q)$. For any fixed deterministic readout G from canonicalised distributions to projective predictions, $\mathcal{R}(P) = G(C_s(P))$ is representative-invariant.

Proof. The distribution $C_s(P)$ depends only on $\bar{P}$, so both the output of G and, for a partially defined G, whether it exists depend only on $\bar{P}$. □

For finite lines, folding antipodal probabilities into $\theta \in [0, \pi)$ implements this construction before the coordinate readout. No continuity of the section is asserted, and invariance alone does not imply concentration-consistency. For partially defined readouts, representative invariance includes the domain. Whether a prediction exists must depend only on the class distribution $\bar{P}$. Convergence in Definition 2 requires predictions to be defined for all sufficiently large indices of every concentrating sequence, not merely along a subsequence on which the readout happens to be defined.

B.4 Failure of representative-vector averaging

Canonicalisation restores representative invariance, but no canonicalisation is continuous. Averaging the selected unit vectors therefore fails in a second way, independent of any coordinate chart. This concerns selected unit vectors, rather than the coordinate averages analysed in the main text.

Proposition B.2 (Section-based vector averaging is inconsistent)**.** *Let $s\colon \mathbb{RP}^{m-1} \to S^{m-1}$, $m \ge 2$, be any section, so $[s(q)] = q$, and let*

$$\mathcal{R}_s(P) = \left[\textstyle\sum_i p_i s([x_i])\right] \tag{24}$$

whenever the sum is nonzero. Then $\mathcal{R}_s$ is representative-invariant but not concentration-consistent. There are a target $[x^\star]$ and equal-weight two-point distributions P_k with $\bar{d}(P_k, x^\star) \to 0$ on which $\mathcal{R}_s$ is defined and $d(\mathcal{R}_s(P_k), [x^\star]) \to 1$.

Proof. Representative invariance is Proposition B.1 with G the normalised vector average.

For the second claim, no section s of the antipodal quotient can be continuous. Otherwise $f(x) = x^\top s([x])$ would be a continuous map from the connected sphere S^{m-1}, $m \ge 2$, to $\{-1, 1\}$, and hence constant. But $f(-x) = -f(x)$, a contradiction.

Choose a discontinuity $q^\star = [x^\star]$ of s, and set $u = s(q^\star)$. There is a sequence of points $q_k \neq q^\star$ with $q_k \to q^\star$ for which $s(q_k)$ stays a positive distance from u. By compactness of the sphere, pass to a subsequence with $v_k = s(q_k) \to v$. Continuity of the quotient map gives $[v] = q^\star$, so $v = \pm u$, and the positive distance from u forces $v = -u$.

Let $P_k = \frac{1}{2}\delta_{q^\star} + \frac{1}{2}\delta_{q_k}$. Then $\bar{d}(P_k, u) = \frac{1}{2}d(q_k, q^\star) \to 0$. Since $q_k \neq q^\star$, we have $v_k \neq -u$, so the average is nonzero and its normalised direction is $r_k = (u + v_k)/\|u + v_k\|_2$. Unit norms give

$$(u^\top r_k)^2 = \frac{(1 + u^\top v_k)^2}{2 + 2u^\top v_k} = \frac{1 + u^\top v_k}{2} \longrightarrow 0. \tag{25}$$

Consequently $d([r_k], q^\star) \to 1$, even though the evidence concentrates on $q^\star$. This proves the assertion for an arbitrary section. □

The obstruction also holds on finite affine line space. Restrict any section to the lines through the origin. Their unit representatives lie in the equatorial $S^1 \subset S^2$, so the same argument applies with $m = 2$ and never uses the line at infinity.

B.5 Failure of coordinate readouts

Oriented object detection encounters a related π-periodic boundary discontinuity for box angles and addresses it with smoothed angular labels or distribution-level losses (Yang & Yan, 2020; Yang et al., 2021). Here the obstruction is at the readout. Changing the training objective while retaining linear signed-offset averaging leaves the cancellation in Proposition 1 unchanged.

We use the Hesse coordinates of Equation 1 and the angular averages of Equation 5, with the common signed-offset average $\hat{\rho} = \sum_i p_i \rho_i$. All distributions in the counterexample have finite support and finite offsets, so the averages exist. The sinusoidal angular sums do not vanish simultaneously on this sequence.

Proof of Proposition 1. The two cells in Equation 1 are

$$x_1 = \frac{(\cos\varepsilon, \sin\varepsilon, -a)^\top}{\sqrt{1 + a^2}}, \qquad x_2 = \frac{(-\cos\varepsilon, \sin\varepsilon, a)^\top}{\sqrt{1 + a^2}}, \tag{26}$$

and the target is $x^\star = \ell(0, a) = (1, 0, -a)^\top/\sqrt{1 + a^2}$. Both x_1 and $-x_2 = (\cos\varepsilon, -\sin\varepsilon, -a)^\top/\sqrt{1 + a^2}$ tend to $x^\star$ as $\varepsilon \to 0$, so both lines approach $[x^\star]$ and $\bar{d}(P_\varepsilon, x^\star) \to 0$.

For the lower bound, any readout retaining the signed-offset average returns $\hat{\rho} = \frac{1}{2}a + \frac{1}{2}(-a) = 0$, so its predicted unit vector is $\hat{x}_\varepsilon = (\cos\hat{\theta}_\varepsilon, \sin\hat{\theta}_\varepsilon, 0)^\top$ for whatever angle $\hat{\theta}_\varepsilon \in [0, \pi)$ it produces. Then $\hat{x}_\varepsilon^\top x^\star = \cos\hat{\theta}_\varepsilon/\sqrt{1 + a^2}$, hence

$$d([\hat{x}_\varepsilon], [x^\star])^2 = 1 - \frac{\cos^2\hat{\theta}_\varepsilon}{1 + a^2} \geq 1 - \frac{1}{1 + a^2} = \frac{a^2}{1 + a^2}, \tag{27}$$

uniformly in ε. Taking square roots gives $d([\hat{x}_\varepsilon], [x^\star]) \geq |a|/\sqrt{1 + a^2}$, which is positive because $a \neq 0$. This proves Equation 6. □

Remark (Baseline values and tightness). Standard soft-argmax averages the orientations to $\frac{1}{2}\varepsilon + \frac{1}{2}(\pi - \varepsilon) = \pi/2$ and returns $(\pi/2, 0)$, whose predicted unit vector $(0, 1, 0)^\top$ is orthogonal to $x^\star$, giving chordal distance one. For the sinusoidal variant, the doubled-angle features average to $\frac{1}{2}\cos 2\varepsilon + \frac{1}{2}\cos(2\pi - 2\varepsilon) = \cos 2\varepsilon > 0$ and $\frac{1}{2}\sin 2\varepsilon + \frac{1}{2}\sin(2\pi - 2\varepsilon) = 0$, so $\frac{1}{2}\operatorname{atan2}(0, \cos 2\varepsilon) = 0$ and the readout returns $(0, 0)$. Its orientation is exact, and substituting $\hat{\theta}_\varepsilon = 0$ above gives error exactly $|a|/\sqrt{1 + a^2}$, attaining the bound.

Comparison with eigenvector recovery. The VSmax second-moment matrix is

$$M_\varepsilon = \frac{1}{1 + a^2}\begin{pmatrix} \cos^2\varepsilon & 0 & -a\cos\varepsilon \\ 0 & \sin^2\varepsilon & 0 \\ -a\cos\varepsilon & 0 & a^2 \end{pmatrix}. \tag{28}$$

Its eigenvalues are $(\cos^2\varepsilon + a^2)/(1+a^2)$, $\sin^2\varepsilon/(1+a^2)$, and zero. Because $0 < \varepsilon < \pi/4$, the first is strictly largest, with unit eigenvector $(\cos\varepsilon, 0, -a)^\top/\sqrt{\cos^2\varepsilon + a^2}$. Since $\cos\varepsilon > 0$, its coordinates are $(\theta, \rho) = (0, a/\cos\varepsilon)$ and converge to $(0, a)$. Moreover, $\cos\varepsilon \to 1$ and $\sin\varepsilon \to 0$ imply $M_\varepsilon \to Q(x^\star)$, so $\|M_\varepsilon - Q(x^\star)\|_F^2 \to 0$ on this same family of input distributions. The baseline errors above nevertheless stay bounded away from zero, because the moment convergence concerns the input evidence, not the projectors of the coordinate-averaged predictions.

Two parts of this argument have different status. The lower bound Equation 6 is non-asymptotic and grid-independent. It holds for every $\varepsilon \in (0, \pi/4)$, so any grid containing the two cells exhibits the exact cancellation $\hat\rho = 0$ and the stated error. Only the concentration certificate $\bar d(P_\varepsilon, x^\star) \to 0$ is asymptotic, and it is what promotes the bound to a failure of concentration-consistency. On the fixed experimental grid we therefore assert the cancellation, not the limit.

B.6 Proof of the VSmax guarantees

We prove Theorem 1 in the more general space $\mathbb{RP}^{m-1}$, $m \geq 2$. The main-text statement is the case $m = 3$. Let P assign masses $p_i \geq 0$, $\sum_i p_i = 1$, to cells $x_i \in S^{m-1}$, let $x^\star \in S^{m-1}$, and put $Q(x) = xx^\top$ and

$$M = M(P) = \mathbb{E}_{X\sim P}\left[XX^\top\right] = \sum_i p_i Q(x_i).$$

Let $\hat x$ be any unit leading eigenvector of M and $\lambda_1(M) \geq \cdots \geq \lambda_m(M)$ its eigenvalues. Define d and $\bar d$ by the same formulas as Equation 2 and Equation 4, respectively. Extend v_2 to the $m(m+1)/2$ distinct entries[1] of $Q(x)$ with weight $\sqrt{2}$ on off-diagonal entries, so $v_2(x) = W\,\mathrm{vech}(Q(x))$ and $\hat v = \sum_i p_i v_2(x_i)$, and define $\mathcal{L}_{\mathrm{vs}} = \|\hat v - v_2(x^\star)\|_2^2$ as in Equation 12.

Lemma 1 (Frobenius form of Veronese supervision)**.** *$\hat v = W\,\mathrm{vech}(M)$ and $\mathcal{L}_{\mathrm{vs}} = \|M - Q(x^\star)\|_F^2$.*

Proof. Since vech is linear, $\hat v = \sum_i p_i W\,\mathrm{vech}(Q(x_i)) = W\,\mathrm{vech}(M)$, and $\hat v - v_2(x^\star) = W\,\mathrm{vech}(M - Q(x^\star))$. For a symmetric matrix A, the Frobenius norm counts each diagonal entry once and each off-diagonal entry twice, so $\|W\,\mathrm{vech}(A)\|_2 = \|A\|_F$ as in Appendix B.2. Taking $A = M - Q(x^\star)$ gives the claim. □

Proof of Theorem 1(i). Since $Q(-x) = Q(x)$, regrouping the cells by line gives

$$M(P) = \sum_i p_i Q(x_i) = \sum_{[x]\in \mathrm{supp}\,\bar P} \bar P([x])\, Q(x),$$

where either unit representative can be used for each line. Restricting to the support drops lines with zero mass, whose terms would vanish anyway. Hence M depends on P only through $\bar P$. Every quantity computed from M alone inherits this property. This covers $\hat v = W\,\mathrm{vech}(M)$ (Lemma 1), the eigenvalues of M and its leading eigenspace. If $\lambda_1(M) > \lambda_2(M)$, the leading eigenspace is the span of $\hat x$, whose unit vectors are $\pm\hat x$. It determines one line $[\hat x]$, whatever sign the eigensolver returns, so the prediction is a function of $\bar P$. This is the classical uniqueness condition for the extrinsic mean on $\mathbb{RP}^{m-1}$ (Bhattacharya & Patrangenaru, 2003), whose leading eigenvector is the mean axis of axial data (Mardia & Jupp, 2000). If $\lambda_1(M) = \lambda_2(M)$, the leading eigenspace still depends only on $\bar P$, but it contains more than one line. Any selection rule that depends only on M, such as a deterministic eigensolver, again yields a function of $\bar P$. Such a rule need not be continuous at ties. Finally, every symmetric matrix has a leading eigenvector. VSmax is therefore defined for every heatmap, so its domain trivially depends only on $\bar P$. Both conditions of Definition 1 hold. □

Proof of Theorem 1(ii). By Section 3.3 the prediction is a leading eigenvector $\hat x$ of M, unique as a projective point if and only if $\lambda_1(M) > \lambda_2(M)$. It therefore suffices to bound the gap and the error. We use the operator norm $\|A\|_{\mathrm{op}} = \max_{\|y\|_2=1}\|Ay\|_2$, which for symmetric A is the largest absolute eigenvalue.

[1] A symmetric $m \times m$ matrix has m diagonal entries and $m(m-1)/2$ distinct off-diagonal entries, so $m(m+1)/2$ in total.

For unit x, y the symmetric matrix $Q(x) - Q(y)$ has trace zero, rank at most two and $\|Q(x) - Q(y)\|_F^2 = 2d([x],[y])^2$ by Equation 22. Its nonzero eigenvalues are therefore $\pm d([x],[y])$, so $\|Q(x) - Q(y)\|_{\mathrm{op}} = d([x],[y])$. With the triangle inequality,

$$\|M - Q(x^\star)\|_{\mathrm{op}} = \Big\|\sum_i p_i\big(Q(x_i) - Q(x^\star)\big)\Big\|_{\mathrm{op}} \le \sum_i p_i\, d([x_i],[x^\star]) = \bar{d}.$$

$Q(x^\star)$ has eigenvalues $1, 0, \dots, 0$. By Weyl's inequality (Weyl, 1912), each eigenvalue of M lies within $\|M - Q(x^\star)\|_{\mathrm{op}} \le \bar{d}$ of the corresponding one. Hence

$$\lambda_1(M) \ge 1 - \bar{d}, \qquad \lambda_2(M) \le \bar{d}, \qquad \lambda_1(M) - \lambda_2(M) \ge 1 - 2\bar{d} > 0$$

for $\bar{d} < 1/2$, and the prediction is unique.

For the error, apply the projection $I - Q(x^\star)$ onto the orthogonal complement of $x^\star$ to $M\hat{x} = \lambda_1(M)\hat{x}$. Since $(I - Q(x^\star))Q(x^\star) = 0$,

$$\lambda_1(M)\,\big(I - Q(x^\star)\big)\hat{x} = \big(I - Q(x^\star)\big)\big(M - Q(x^\star)\big)\hat{x}.$$

The left side has norm $\lambda_1(M)\, d([\hat{x}],[x^\star])$, because $\|(I - Q(x^\star))\hat{x}\|_2^2 = 1 - (\hat{x}^\top x^\star)^2$. The right side has norm at most $\|M - Q(x^\star)\|_{\mathrm{op}} \le \bar{d}$. With $\lambda_1(M) \ge 1 - \bar{d} > 0$ from the gap step,

$$d([\hat{x}],[x^\star]) \le \frac{\bar{d}}{\lambda_1(M)} \le \frac{\bar{d}}{1 - \bar{d}}.$$

This is the one-dimensional case of the $\sin\Theta$ theorem of Davis & Kahan (1970), with $Q(x^\star)$ as the reference operator, whose remaining spectrum $\{0\}$ is separated from $\lambda_1(M)$ by $\lambda_1(M)$. □

Embedding-supervision bound.

Proof of Theorem 1(iii). By Lemma 1, $\mathcal{L}_{\mathrm{vs}} = \|M - Q(x^\star)\|_F^2$, and $\|A\|_F^2 = \mathrm{tr}(A^2)$ for symmetric A. For a unit vector y we expand

$$\|M - Q(y)\|_F^2 = \mathrm{tr}(M^2) - 2\,\mathrm{tr}\big(M\,yy^\top\big) + \mathrm{tr}\big(yy^\top yy^\top\big) = \mathrm{tr}(M^2) + 1 - 2\,y^\top My,$$

using $\mathrm{tr}(M\,yy^\top) = y^\top My$ by cyclicity of the trace and $\mathrm{tr}(yy^\top yy^\top) = (y^\top y)^2 = 1$. Only the last term depends on y. By Section 3.3, $\hat{x}$ maximises $y^\top My$. Hence $Q(\hat{x})$ is at least as close to M as the projector of any other line, for every choice of leading eigenvector, also at ties. Taking $y = x^\star$ in this comparison gives $\|M - Q(\hat{x})\|_F \le \|M - Q(x^\star)\|_F = \sqrt{\mathcal{L}_{\mathrm{vs}}}$, and the triangle inequality gives

$$\|Q(\hat{x}) - Q(x^\star)\|_F \le \|Q(\hat{x}) - M\|_F + \|M - Q(x^\star)\|_F \le 2\sqrt{\mathcal{L}_{\mathrm{vs}}}.$$

By Equation 22 the left side equals $\sqrt{2}\, d([\hat{x}],[x^\star])$. Squaring yields $d([\hat{x}],[x^\star])^2 \le 2\mathcal{L}_{\mathrm{vs}}$.

For the gap, write $\lambda_j = \lambda_j(M)$. Since M is positive semidefinite with unit trace, the eigenvalues are non-negative and sum to one. In particular $\lambda_1 \le 1$. The expansion above at $y = x^\star$ reads $\mathcal{L}_{\mathrm{vs}} = \mathrm{tr}(M^2) + 1 - 2x^{\star\top} Mx^\star$. We have $\mathrm{tr}(M^2) = \sum_j \lambda_j^2$, and $x^{\star\top} Mx^\star \le \lambda_1$ by the Rayleigh principle. Therefore

$$\mathcal{L}_{\mathrm{vs}} \ge \sum_j \lambda_j^2 + 1 - 2\lambda_1 = (1 - \lambda_1)^2 + \sum_{j \ge 2} \lambda_j^2 \ge (1 - \lambda_1)^2 + \lambda_2^2 \ge \tfrac{1}{2}(1 - \lambda_1 + \lambda_2)^2,$$

where the last step uses $a^2 + b^2 \ge \frac{1}{2}(a + b)^2$. Since $\lambda_1 \le 1$ and $\lambda_2 \ge 0$, the term $1 - \lambda_1 + \lambda_2$ is non-negative. Taking square roots gives $\lambda_1 - \lambda_2 \ge 1 - \sqrt{2\mathcal{L}_{\mathrm{vs}}}$. For $\mathcal{L}_{\mathrm{vs}} < 1/2$ the gap is positive and the prediction is unique. □

The sufficient loss threshold cannot be increased. For orthogonal unit e_1, e_2, the mixture $M = \frac{1}{2}Q(e_1) + \frac{1}{2}Q(e_2)$ and target $x^\star = e_1$ have $\mathcal{L}_{\mathrm{vs}} = 1/2$ and a zero top eigengap. At the other extreme, $\mathcal{L}_{\mathrm{vs}} = 0$ forces $M = Q(x^\star)$, giving exact recovery and a unit gap.

Scope and degeneracy. These statements hold in the full projective space for a fixed homogeneous coordinate normalisation. For the line parametrisation, that metric is not invariant to arbitrary changes of image origin or scale, and is not the EA score. Decoding to finite Hesse coordinates requires a nonzero affine normal in addition to spectral uniqueness.

C IMPLEMENTATION AND REPRODUCIBILITY

C.1 HOUGH ACCUMULATION AND LINE DECODING

The Hough transform stores voting weights between image locations (u, v) and cells (θ_j, ρ_k) in a fixed correspondence tensor $\mathcal{A}$. For an image or feature map F,

$$H_{jk} = \sum_{u,v} \mathcal{A}_{jkuv} F_{uv}. \tag{29}$$

The entries encode $\rho_k = u \cos\theta_j + v \sin\theta_j$, assigning pixel votes to the corresponding offset bin or neighbouring bins. This contraction is linear and differentiable with respect to F, even with hard voting bins. The readout aggregates the normalised accumulator with the cache V of Equation 8. Whereas $\mathcal{A}$ maps image evidence to the accumulator, V maps accumulator probabilities to embedding features.

Decoding to Hough coordinates. For the recovered unit representative $\hat{\ell}$, let $r = \sqrt{\hat{\ell}_1^2 + \hat{\ell}_2^2} > 0$. Angle and offset must use the same sign:

$$\begin{aligned} &\phi = \operatorname{atan2}(\hat{\ell}_2, \hat{\ell}_1) \in (-\pi, \pi], \quad \hat{\theta} = \phi \bmod \pi \in [0, \pi), \\ &s = \begin{cases} 1, & 0 \le \phi < \pi, \\ -1, & \text{otherwise}, \end{cases} \qquad \hat{\rho} = -\frac{s\hat{\ell}_3}{r}. \end{aligned} \tag{30}$$

If $r = 0$, the prediction is the line at infinity and has no finite offset.

C.2 DOUBLE-COVER FOLD

For the Hough implementation, the readout's antipodal invariance lets the refinement network retain both representatives. Two crossings of the angular boundary restore the offset sign, giving the cylinder $S^1 \times \mathbb{R}$ on which ordinary circular angular padding can be used (Kim et al., 2020). Averaging on the double cover is equivalent to first summing each antipodal pair:

$$\hat{\nu} = \sum_{\substack{0 \le \theta < \pi \\ \rho}} \big[P(\theta, \rho) + P(\theta + \pi, -\rho)\big]\, \nu_2(\ell(\theta, \rho)). \tag{31}$$

The paired activations need not be equal. The following gauge interpretation relates this implementation to gauge-equivariant convolutions (Weiler et al., 2021) and states the conditions on the discrete grid and boundary handling.

C.3 GAUGE-EQUIVARIANT INTERPRETATION

The identification across the angular boundary is a glide reflection (Stillwell, 1995). Its relation to symmetry-aware convolutions (Cohen & Welling, 2016; Weiler et al., 2021) motivates the following interpretation of the double-cover implementation.

The cylinder does not remove the Möbius reflection. It retains both choices of local orientation. The two activations at an undirected line can therefore differ, provided they transform as a pair when the local frame changes. This is the regular-feature construction of Weiler et al. (2021), with the orientation index stored on the cover rather than in separate channels.

To see this, write $x = (\theta, \rho)$ and $\tau x = (\theta + \pi, -\rho)$ for the two representatives. Choosing τx instead of x reflects the local Hough frame by $J = \operatorname{diag}(1, -1)$ and exchanges the entries of $\big(f(x), f(\tau x)\big)$. This exchange is the regular representation of $\mathbb{Z}_2$, and we denote its matrix by S. A displacement u near x corresponds to Ju near τx, since $\tau(x + u) = \tau x + Ju$. To make the kernel correspondence explicit, write a local cylindrical convolution as

$$(T_K f)(x) = \int K(u) f(x + u)\, \mathrm{d}u.$$

In a quotient neighbourhood whose chosen lift is x, define $F_+(u) = f(x+u)$ and $F_-(u) = f(\tau(x+u)) = f(\tau x + Ju)$. Changing variables $v = Ju$, with $|\det J| = 1$, gives

$$\begin{pmatrix} (T_K f)(x) \\ (T_K f)(\tau x) \end{pmatrix} = \int \begin{pmatrix} K(u) & 0 \\ 0 & K(Ju) \end{pmatrix} \begin{pmatrix} F_+(u) \\ F_-(u) \end{pmatrix} \mathrm{d}u.$$

The discrete version replaces the integral by a stencil sum and requires the reflection to permute the stencil. Thus the same cylindrical kernel at both lifts induces, in one quotient frame,

$$\mathcal{K}(u) = \begin{pmatrix} K(u) & 0 \\ 0 & K(Ju) \end{pmatrix}, \qquad \mathcal{K}(Ju) = S_{\mathrm{out}}\, \mathcal{K}(u)\, S_{\mathrm{in}}^{-1}, \tag{32}$$

where S_{in} and S_{out} exchange the input and output blocks. This is precisely the reflection-steerability condition of Weiler et al. (2021). The cover supplies the reflected block automatically, so the learned kernel K need not itself be reflection symmetric. Following each lift through its convolution neighbourhood also supplies the required transport between local frames, including across the angular wrap. The resulting paired-feature convolution is therefore $\mathbb{Z}_2$ gauge equivariant under these frame changes. This is a restricted, block-diagonal family of gauge-steerable kernels, not the full family permitting learned mixing between the two orientation components.

Gauge covariance versus scalar deck equivariance. This interpretation does not assert that an arbitrary cylindrical kernel preserves scalar fields constant on antipodal pairs. If $(\tau^* f)(x) = f(\tau x)$, then $T_K \tau^* = \tau^* T_{K \circ J}$, and commutation with τ^* as a scalar operator generally requires $K(u) = K(Ju)$. No such restriction is imposed here. Instead, a change of local orientation both reflects coordinates and exchanges the two feature components, as in Equation 32. Their values may differ without making the paired field dependent on the chosen local frame.

Shared biases and pointwise nonlinearities commute with the exchange. Batch normalisation does too when its statistics use the full cover and its parameters are shared between lifts. Global softmax likewise only permutes paired entries. Finally, the Veronese readout depends on $P(x) + P(\tau x)$, not on their ordering or equality, because $\nu_2(\ell(x)) = \nu_2(\ell(\tau x))$ (Equation 31). The refinement thus retains orientation-dependent responses, while the readout is gauge invariant.

This argument uses the flat Hough geometry, a consistently doubled input, and local stride-one convolutions on rectangular stencils. Discretely, the half-turn and offset reflection must be exact bin permutations, with circular angular padding throughout and reflection-compatible offset boundaries and masks. The claim concerns the $\mathbb{Z}_2$ frame ambiguity of the Möbius strip, not equivariance to arbitrary image or projective transformations. Circular padding alone is not a general guarantee of gauge equivariance for additional operations such as striding, pooling or position-dependent masks, whose compatibility would need to be checked separately. Independently of these architectural conditions, Equation 31 and the readout guarantees hold for any normalised input probability map. They do not imply that refinement preserves or improves geometric concentration.

C.4 Training setup and baseline configurations

Table 2 lists the optimisation, batching and baseline settings shared by all synthetic runs. The Hesse objective $\mathcal{L}_{\mathrm{hs}}$ of Equation 12 takes angles in radians and offsets normalised by $\rho_{\max} = \sqrt{H^2 + W^2}/2$ for image height H and width W in pixels, with $(\hat{\theta}, \hat{\rho})$ the decoded prediction and $(\theta^\star, \rho^\star)$ the Hough-native target. Veronese supervision uses the target embedding and weighted Euclidean metric of Section 3.2. VSmax supervises its embedding average, whereas coordinate readouts supervise the embedding of their decoded line.

C.5 EA score and evaluation conventions

For a predicted line $\hat{\ell}$ and ground-truth line $\ell^\star$, the EA-score of Zhao et al. (2022) combines angular and positional similarity. Let $\Delta\theta \in [0, \pi/2]$ be the angle between the two undirected lines and d_ℓ the Euclidean distance between the midpoints of the segments in which they intersect the image domain, normalised by the longer image side (our images are square, so this is the side length). Then

$$S_\theta = 1 - \frac{\Delta\theta}{\pi/2}, \qquad S_d = \max(0,\, 1 - d_\ell), \qquad \mathrm{EA}(\hat{\ell}, \ell^\star) = (S_\theta S_d)^2 \in [0, 1]. \tag{33}$$

Table 2: Optimisation, batching and baseline configurations. The optimisation and batching blocks apply to the synthetic experiments. HLW-visible settings are reported in Appendix D.5.

Optimisation	
Optimiser	AdamW
Learning rate	10^{-4}
Weight decay	10^{-4}
Betas	$(0.9,\ 0.999)$
Scheduler	ReduceLROnPlateau (factor 0.1, patience 10)
Early stopping	on val. loss (patience 20, min. delta 10^{-4})
Max. epochs	1000
Model selection	best validation loss
Batching	
Batch size	64 (train / validation / test)
Seed	42, deterministic
Baseline	
CNN+MLP	ResNet-18 (ImageNet-initialised, fine-tuned) $+$ MLP head with 256 hidden units, 11.3M parameters

Table 3: Mean EA-score of Figure 4 with the standard deviation over the $14{,}383$ paired test lines.

	$\sigma = 0.5$	$\sigma = 0.6$	$\sigma = 0.7$	$\sigma = 0.8$
(a) Frozen heatmaps, readout changed				
standard soft-argmax	0.904 ± 0.190	0.836 ± 0.262	0.680 ± 0.332	0.475 ± 0.340
sinusoidal soft-argmax	0.906 ± 0.182	0.853 ± 0.245	0.739 ± 0.317	0.577 ± 0.350
VSmax	0.918 ± 0.170	0.878 ± 0.233	0.787 ± 0.315	0.637 ± 0.378
(b) VSmax retrained under each objective				
VSmax, $\mathcal{L}_{\mathrm{hs}}$	0.875 ± 0.225	0.692 ± 0.295	0.480 ± 0.303	0.333 ± 0.272
VSmax, $\mathcal{L}_{\mathrm{vs}}$	0.918 ± 0.170	0.878 ± 0.233	0.787 ± 0.315	0.637 ± 0.378
(c) Separately trained pipelines				
standard soft-argmax	0.906 ± 0.172	0.749 ± 0.298	0.472 ± 0.344	0.272 ± 0.285
sinusoidal soft-argmax	0.908 ± 0.212	0.749 ± 0.284	0.522 ± 0.305	0.339 ± 0.272
VSmax	0.918 ± 0.170	0.878 ± 0.233	0.787 ± 0.315	0.637 ± 0.378

Since $\Delta\theta \le \pi/2$, S_θ is already non-negative and needs no clamp. Both midpoints lie in the image domain, so $d_\ell \in [0, \sqrt{2}]$ and $1 - d_\ell$ can be negative. The clamp sets $S_d = 0$ for $d_\ell \ge 1$. Both factors therefore lie in $[0, 1]$ and EA is non-increasing in $\Delta\theta$ and in d_ℓ. Equivalently, $\mathrm{EA} = \big(\max(0, 1 - \frac{\Delta\theta}{\pi/2})\big)^2 \big(\max(0, 1 - d_\ell)\big)^2$, matching the reference implementation of Zhao et al. (2022). Larger values indicate better agreement.

Degenerate predictions are scored by an explicit convention. Every ground-truth line is visible by construction of the test grid, so only predictions can degenerate. If a predicted line does not meet the image domain, or is the line at infinity of Equation 30, its midpoint is undefined and we set $\mathrm{EA} = 0$ regardless of S_θ. Global scores average all evaluated test lines, retaining these zeros rather than discarding them. Heatmaps place each score at its ground-truth (θ, ρ) bin. EA is an image-space evaluation metric, not the chordal distance used in the theoretical guarantees.

Spread of the score. Table 3 lists the means of Figure 4 with the standard deviation of EA over the $14{,}383$ test lines. The standard deviations are large in relation to the differences between methods because EA is a per-line success score rather than an error around a typical value. By Equation 33 it is the squared product of an angular and a positional factor, so it stays near one only when both are accurate, is driven towards zero by a single wrong parameter, and is exactly zero for a degenerate prediction. The per-line distribution therefore concentrates at both ends, and a mean is read as the fraction of lines recovered accurately rather than as a typical accuracy. This also explains why the

best method has the largest spread. At $\sigma = 0.8$ in the complete pipelines, VSmax scores above 0.9 on 42.8% of the lines and below 0.1 on 20.8%, the latter including 6.8% degenerate predictions at extreme offsets (Figure 5). Standard and sinusoidal soft-argmax score above 0.9 on only 3.7% and 2.0% of the lines and below 0.1 on 40.8% and 26.7%, with the remaining mass in between. A bimodal distribution with most mass at the top has a larger standard deviation than a unimodal one in the middle, so the ± 0.38 of VSmax against ± 0.28 and ± 0.27 for the baselines reflects more exact recoveries, not less reliable ones.

D Additional experiments

Table 4 records what each comparison holds fixed and changes. Readouts are first tested on constructed heatmaps without a network. The main text then verifies the recovery bounds on trained predictions, presents the readout and supervision controls, and finally the real-image transfer.

Table 4: Experimental controls and the questions they address, in the order the main text presents them. Common embedding targets use the same target features and metric, but supervise embeddings of decoded lines for coordinate readouts and the embedding average for VSmax.

Comparison	Held fixed	Changed	Question
Constructed heatmaps	Heatmap given exactly, no network	Readout, target position and width	Does the readout alone fail at the boundary?
Frozen heatmaps	Checkpoint and probabilities	Inference readout	Effect of aggregation?
VSmax objectives	Architecture and readout	Training loss	Benefit of Veronese supervision?
Synthetic pipelines	Data and evaluation	Readout–objective pair	Complete-pipeline accuracy?
Common embedding targets	Refinement, targets and optimisation	Readout and supervised prediction	Effect under geometric supervision?
Real-image models	Data and evaluation	Complete model	Transfer to natural images?

D.1 Readouts on constructed heatmaps

Proposition 1 and Theorem 1(ii) concern the readout alone, applied to a given heatmap. This subsection evaluates them on heatmaps of prescribed concentration, with no network and no training, so that the dispersion $\bar{d}$ is controlled exactly.

Construction. We use a canonical 127×127 grid at the resolution of the synthetic experiments, with cell centres $\theta_j = (j + \frac{1}{2})\pi/127$ and normalised offsets $\rho_k = -1 + (2k+1)/127$. A target $[x^\star] = [\ell(\theta^\star, \rho^\star)]$ receives a Gaussian heatmap in the chordal distance,

$$p_i \propto \exp\bigl(-d([x_i], [x^\star])^2 / 2s^2\bigr), \tag{34}$$

of width s. It depends on each cell only through its distance to the target line, so both representatives of a line receive the same mass. The heatmap is therefore defined directly on the canonical grid, and the coordinate readouts of Equation 5 and VSmax see the same probabilities. The width takes 61 log-spaced values $s \in [0.002, 0.7]$, and $\rho^\star = 0.504$ is the offset cell centre nearest $1/2$. The interior target $\theta^\star = 63\pi/127 \approx 89.3^\circ$ and the target $\theta^\star = 0$ on the angular boundary both lie midway between two angular cell centres. A shift by 63 angular cells, combined with the identification $(\theta, \rho) \sim (\theta + \pi, -\rho)$, maps the grid onto itself and one target onto the other, so the two differ only in their position relative to the boundary.

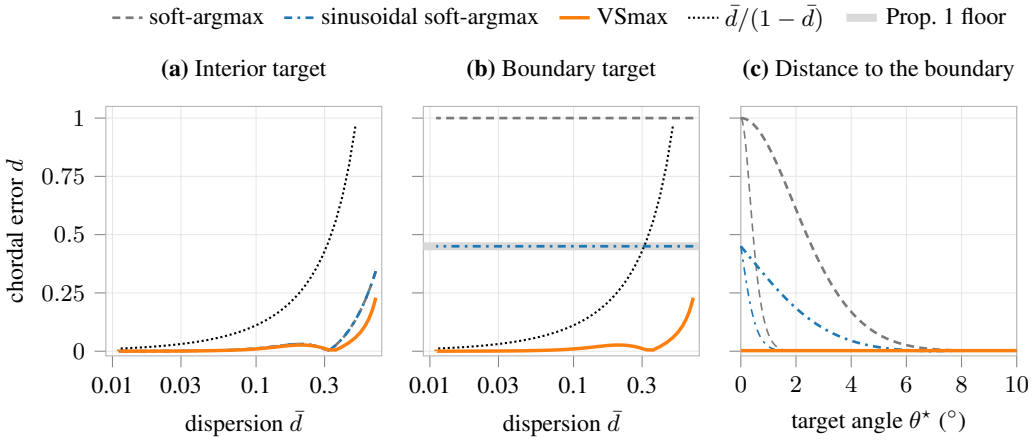


Figure 7: Readouts on constructed Gaussian heatmaps (Equation 34) over the 127×127 canonical grid, without a network. (a,b) Chordal error against the dispersion $\bar{d}$ of the heatmap, for an interior target and a target on the angular boundary. The black dotted curve is the bound of Theorem 1(ii) for $\bar{d} < 1/2$, the grey band the error floor of Proposition 1. The coordinate readouts coincide in (a), and the VSmax curves in (a) and (b) coincide. (c) Error against the target angle at width $s = 0.04$ (thick) and $s = 0.01$ (thin). VSmax is shown at $s = 0.04$.

Concentration at the boundary. The heatmap of the boundary target is symmetric under $(\theta, \rho) \mapsto (\pi - \theta, -\rho)$. Every cell is paired with one of equal mass and opposite offset, so $\hat{\rho} = 0$ exactly, at every concentration. This is the configuration of Proposition 1, realised by the whole heatmap rather than by two cells. Standard soft-argmax returns $(\pi/2, 0)$ with error 1, and the sinusoidal variant returns $(0, 0)$ with error exactly $|\rho^\star|/\sqrt{1+\rho^{\star 2}} = 0.450$, while $\bar{d}$ falls to 0.011 (Figure 7b). At the interior target all three readouts are accurate, and their errors fall with $\bar{d}$ (Figure 7a).

VSmax gives the same error at both targets to numerical precision, because the targets are related by a symmetry of the grid and the readout depends only on the classes. Its error falls to 3.1×10^{-5} and stays below $\bar{d}/(1-\bar{d})$ throughout. Across all 512 constructed heatmaps in this subsection with $\bar{d} < 1/2$, neither inequality of Equation 15 is violated. Since the two heatmaps are the same up to a symmetry of the grid, everything except the position of the target relative to the boundary is held fixed, and the failure in Figure 7b is a property of the coordinate readouts alone.

Distance to the boundary. Figure 7c moves the target from the boundary into the chart, $\theta^\star \in [0°, 10°]$, at fixed concentration. The coordinate readouts fail only while the heatmap reaches across the boundary. Their error exceeds 0.05 up to 5.35° from the boundary at $s = 0.04$, and up to 1.05° at $s = 0.01$. The band narrows with concentration but does not close at the boundary itself. On a fixed grid, the cancellation therefore holds at every concentration for heatmaps split across the boundary, while the set of targets it affects shrinks as the evidence concentrates. VSmax stays accurate across the band.

D.2 Empirical verification of the recovery guarantees

Theorem 1 is stated in $\bar{d}$, $\mathcal{L}_{\text{vs}}$, the eigengap and the chordal distance d, whereas the main-text comparisons report the EA-score against image noise σ. This subsection reports the theorem's own quantities on trained predictions. It introduces no new training and no new evaluation run. Every number is recomputed offline from the stored inference records of the VSmax–Veronese checkpoint used in Figure 4, which saves the six-component moment $W\,\text{vech}(M)$ for each test image.

For each of the 129,447 stored evaluations we rebuild M, form the target projector $Q^\star$ from the recorded ground truth, and compute $\mathcal{L}_{\text{vs}} = \|M - Q^\star\|_F^2$, the eigengap $\lambda_1(M) - \lambda_2(M)$, the decoded chordal error $d([\hat{x}], [x^\star])$, and the root-mean-square dispersion

$$\bar{d}_2 = \left(\sum_i p_i\, d([x_i], [x^\star])^2\right)^{1/2} = \sqrt{1 - x^{\star\top} M x^\star} \;\geq\; \bar{d} \tag{35}$$

(Table 5). The quantity $\bar{d}_2^2$ is the recovery objective $1-y^\top My$ of Section 3.3 evaluated at $y = x^\star$, so it is read directly from the stored aggregate. It dominates $\bar{d}$ by Jensen's inequality, and $t \mapsto t/(1-t)$ is increasing, so $\bar{d}_2 < 1/2$ certifies the hypothesis of Theorem 1(ii) and its bound holds at $\bar{d}_2$. Part (iii) is checked exactly.

Two consistency gates precede any reported number. The reconstructed moment must satisfy $|\operatorname{tr}(M) - 1| \le 10^{-4}$, since only an average of unit projectors lies on that simplex. The worst observed deviation is 2.2×10^{-6}. The leading eigenvector of M must reproduce the stored $(\hat{\theta}, \hat{\rho})$ prediction. The worst observed chordal disagreement is 5.8×10^{-6}. A run failing either gate is not a moment average and is excluded.

Table 5: Theorem 1 quantities for the VSmax–Veronese checkpoint, averaged over the 14,383 test lines at each noise level. The last column is the percentage of images with $\mathcal{L}_{\mathrm{vs}} < 1/2$, for which part (iii) certifies a unique projective prediction. EA is repeated for comparison and is not the metric the theorem controls.

σ	$\mathcal{L}_{\mathrm{vs}}$	$\lambda_1 - \lambda_2$	$\bar{d}_2$	d	EA	unique (%)
0.0	0.0024	0.9985	0.0271	0.0253	0.9507	100.0
0.1	0.0029	0.9984	0.0277	0.0255	0.9506	100.0
0.2	0.0039	0.9977	0.0290	0.0266	0.9491	99.9
0.3	0.0053	0.9952	0.0327	0.0280	0.9463	99.8
0.4	0.0113	0.9852	0.0452	0.0325	0.9363	99.3
0.5	0.0278	0.9644	0.0770	0.0463	0.9179	97.9
0.6	0.0617	0.9114	0.1616	0.0754	0.8778	95.1
0.7	0.1422	0.7900	0.3173	0.1435	0.7859	89.1
0.8	0.2859	0.6219	0.4955	0.2602	0.6360	78.0

Mean $\bar{d}_2$ rises from 0.027 at $\sigma = 0$ to 0.4955 at $\sigma = 0.8$, while the mean eigengap falls from 0.998 to 0.622. Uniqueness is certified by part (iii) for every image at $\sigma = 0$ and for 78.0% of images at $\sigma = 0.8$. The certificate is evaluated per image. A mean dispersion below $1/2$ does not establish the condition for every prediction. For images outside the sufficient loss condition, uniqueness is determined by the actual eigengap, and lack of a certificate alone does not imply ill-conditioning.

This check characterises the checkpoint used in the main comparison. The theorem holds for every admissible moment, while the observed bound ratios and certificate coverage describe this trained model. The two evaluation metrics also answer different questions. At $\sigma = 0.8$ mean chordal error is 0.260 and mean EA is 0.636. Chordal distance compares normalised homogeneous coefficients, while EA combines angular agreement with image-space displacement of the visible segment.

D.3 Classical and heatmap-supervised Hough comparisons

We want to show why end-to-end training is beneficial in this setup. For that, we compare our method to two Hough-variants that do not rely on it. The plain Hough transform (no refinement), read out by non-differentiable (argmax) peak selection, and the Reverse mapping of Hough transform (RHT) (Zhao et al., 2022), a learned heatmap-matching baseline (with refinement) from the multi-line detection literature with a non-differentiable readout. Figure 8 sweeps the test noise. The classical transform is exact when clean (≈ 0.98) but decays towards 0.5 as noise grows. This is exactly the regime a learned network fixes, by cleaning the accumulator before the readout. The RHT does not transfer to single-line regression. At every hyperparameter configuration we tried it collapses under mild noise (we sweep these RHT-specifics in Appendix D.3.1). Its recipe is built to detect an unknown number of lines rather than regress a single one end-to-end. Only our differentiable VSmax pipeline degrades gracefully and stays robust across the sweep.

D.3.1 Target-width sweep for the learned Hough baseline

The RHT is a learned Hough baseline from the multi-line detection literature. A network regresses a heatmap over the (θ, ρ) accumulator whose peaks mark the lines, trained to match Gaussian targets of width σ_t. The lines are then recovered by a separate, non-differentiable peak-extraction step. The method is built to detect an unknown number of lines rather than to regress a single line end-to-end,

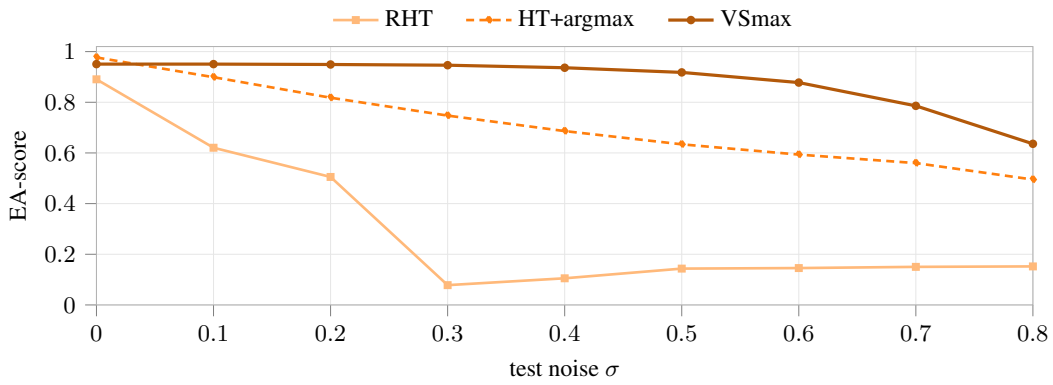


Figure 8: Global EA-score on the full test set vs. test noise. The classical Hough readout is exact when clean but brittle. The RHT (best target width, Appendix D.3.1) collapses under noise. Our end-to-end differentiable Veronese soft-argmax degrades gracefully.

and its readout cannot be trained through. We evaluate it here as a Hough-based reference. Because the target width σ_t is a free hyperparameter uncoupled from the line error, we sweep it widely. Figure 9 reports global EA-score on the full test set for $\sigma_t \in \{0.5, 1.1, 1.5, 2.0, 3.0\}$. On clean data a wider target helps, and EA-score rises from 0.79 at σ_t=0.5 to 0.89 at σ_t=3.0. Under even mild noise, however, every width collapses to EA $\approx$ 0.1 by σ=0.3, and no configuration we tried recovers. We conclude the recipe does not transfer to single-line regression, and adopt its strongest member (σ_t=3.0) as the RHT reference in Figure 8.

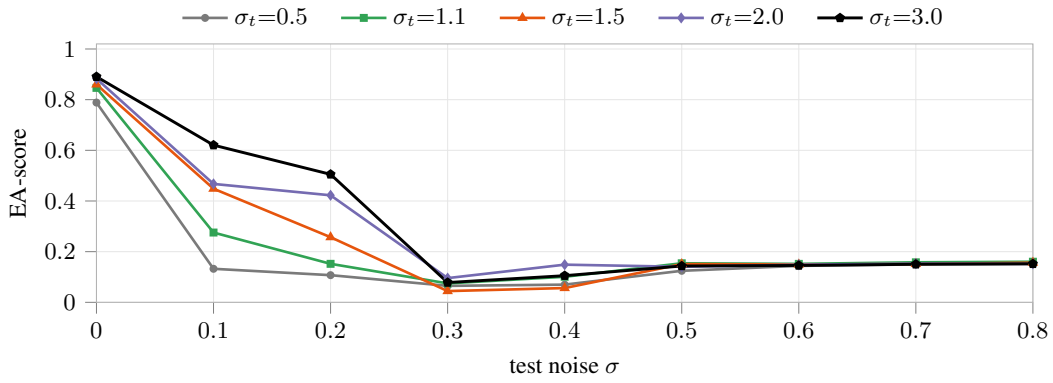


Figure 9: RHT global EA-score on the full test set vs. test noise σ, per target-heatmap width σ_t. A wider target helps when clean, but every width collapses once noise appears and no target width recovers.

D.4 Broader pipeline ablations

Figure 10 shows the direct CNN–MLP regressor used in these ablations. The known-operator pipeline is shown in Figure 2.

D.4.1 Direct-regression representations and spatial errors

To examine representation-related errors without spatial averaging, we compare the two direct-regression variants in Figure 11. An ImageNet-initialised, fine-tuned ResNet-18 backbone with an MLP head, abbreviated MLP, implements Figure 10. The variants share the backbone and hidden-layer design but predict either two Hesse coordinates or six Veronese components, with the corresponding supervision and line decoding. This changes the head dimension as well as the output

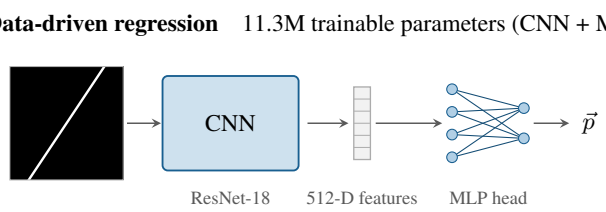


Figure 10: Data-driven line regression. A ResNet-18 backbone produces a 512-dimensional feature vector, and an MLP head predicts Hesse or Veronese parameters $\vec{p}$. The CNN and MLP are trainable, with 11.3M parameters in total. Feature cells and MLP nodes are schematic, not literal counts.

representation and supervision, so it is not a pure loss ablation. Canonical Hesse targets can jump between distant coordinates for geometrically nearby lines across the angular boundary, whereas their Veronese targets remain nearby. Direct regression can therefore inherit a chart-boundary difficulty without using spatial soft-argmax. At $\sigma = 0.6$, the Hesse variant scores lowest in the columns adjacent to the angular boundary, and the Veronese variant lifts these columns most while the interior changes little (Figure 11). This localised improvement is consistent with removing the target discontinuity.

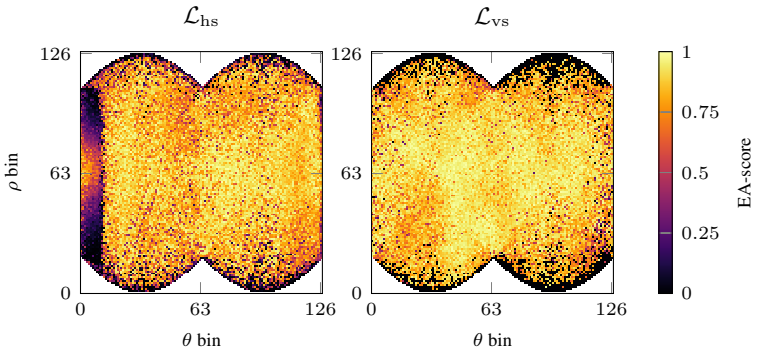


Figure 11: Per-line EA-scores for direct MLP regression at $\sigma = 0.6$ with Hesse (left) and Veronese (right) output representations. Ground-truth Hough bins are evaluation coordinates, not a network operation. Colour scale and masking follow Figure 5.

D.4.2 Supervision under image corruption

Figure 12 compares the separately trained Hesse and Veronese MLP variants over $\sigma \in [0.5, 0.8]$, retaining 0.5 as a training-boundary reference. The Veronese variant has higher global mean EA-score at $\sigma = 0.5$ and 0.6, but the ordering reverses at the tested levels 0.7 and 0.8. The reversal is a property of recovery without the operator. The Veronese head emits six unconstrained numbers, so under unseen noise its output drifts away from the projector mixtures that eigenvector recovery is designed for and the recovery becomes ill-conditioned, whereas the Hesse head decodes its two coordinates directly. In the known-operator pipeline the aggregate is a convex combination of projectors by construction, and Theorem 1(ii) keeps its eigengap bounded by the concentration of the heatmap. Beyond the reversal, both direct regressors lose most of their accuracy once the noise leaves the training range, whichever output representation they use, and both remain far below the known-operator VSmax pipeline at the same noise levels (Figure 4c), which has a small fraction of their parameters. The representation removes the boundary discontinuity of the targets, but generalisation to unseen noise comes from the pipeline, in which a fixed Hough transform accumulates the evidence and the readout aggregates it geometrically. A representation change alone does not reproduce that behaviour.

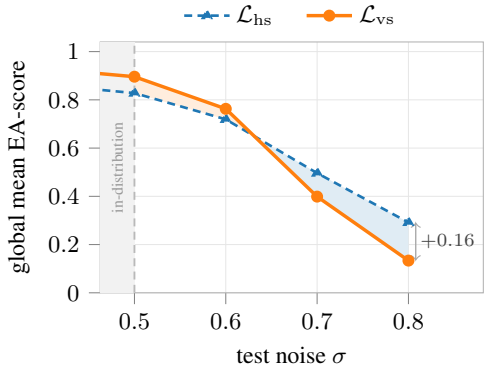


Figure 12: Global mean EA-score for separately trained MLP variants with Hesse (blue, dashed) and Veronese (orange, solid) outputs and supervision. The dashed vertical line marks the training maximum $\sigma = 0.5$. Shading and the bracket show the higher-scoring variant's absolute advantage, not uncertainty.

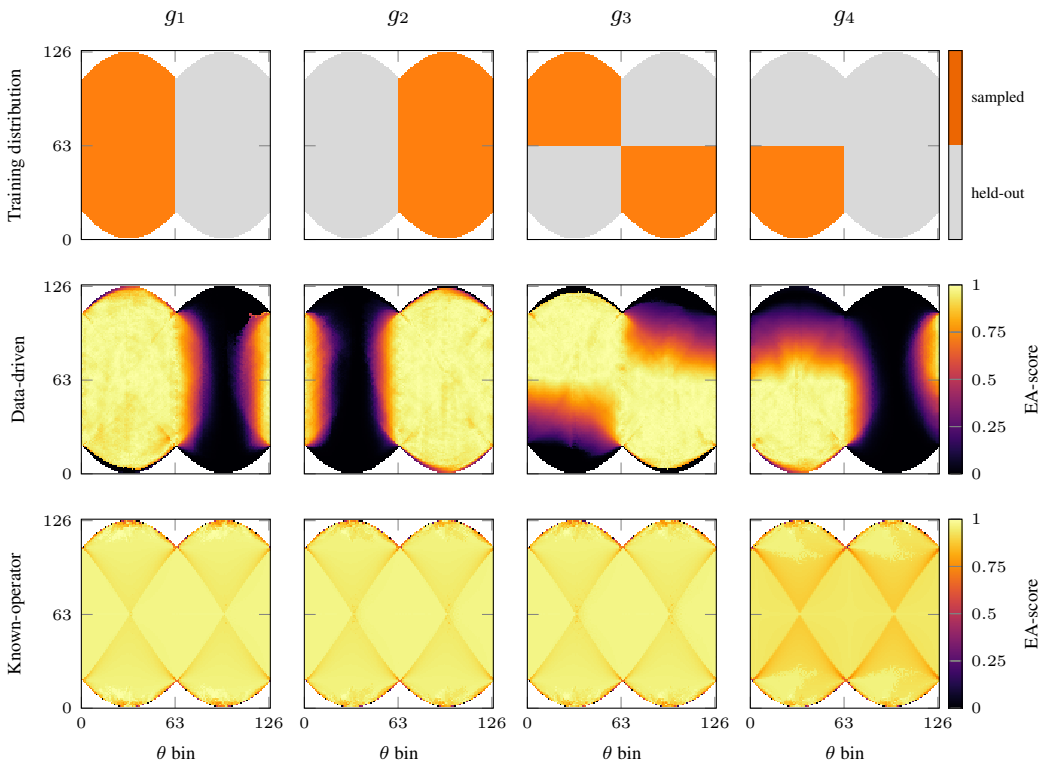


Figure 13: Generalisation across training-support settings g_1–g_4 (columns). The rows show the binary training-support region in (θ, ρ), the data-driven MLP's EA-score, and our known-operator pipeline's EA-score (clean data). In these settings, the MLP's accuracy drops in held-out regions, while the known-operator pipeline is more uniform. This compares complete pipelines, not the operator alone.

D.4.3 GENERALISATION ACROSS TRAINING-SUPPORT REGIONS

Finally, we compare generalisation of our 37.6K-parameter known-operator pipeline and a data-driven CNN+MLP regressor (11.3M parameters, roughly $300\times$ more). The latter fine-tunes an ImageNet-initialised ResNet-18 and is not trained from scratch. Both use Veronese supervision, while we restrict *training* to sub-regions of line space and test on the full space. This is a broader pipeline comparison in which architecture, capacity and the recovery computation differ, so it does not isolate the operator. The four settings g_1–g_4 tighten the trained support from a half-space (g_1, g_2) through a diagonal split (g_3) to a single quadrant (g_4), pushing held-out lines progressively farther from anything seen. In-support the two are comparable, but their out-of-support accuracy differs sharply (Figure 13). The MLP degrades on held-out lines, whereas the known-operator pipeline stays almost uniform in- and out-of-support. These settings support the generalisation benefit of this pipeline with fewer parameters. They do not establish that direct regression cannot generalise under other training or architectural choices.

D.5 HLW-VISIBLE HORIZON ESTIMATION

The real-image comparison tests complete horizon-estimation pipelines rather than isolating the inference readout. All results here are recomputed from the imported per-image records under `hlw_visible_v1`, whose metadata explicitly marks them as not comparable to full-HLW evaluation.

D.5.1 DATA, VISIBILITY FILTERING AND SPLITS

The HLW source split (Workman et al., 2016) contains $1{,}300$ seen-scene test images and 718 held-out-scene images. The imported visible protocol retains 934 and 499, respectively, giving $1{,}433$ combined images. We keep an image when its annotated horizon is visible in the view the network receives (Table 6). Because all models consume a centre-square crop resized to 256×256, visibility is judged on that cropped view, so the criterion is stricter than discarding only the images whose horizon already falls outside the released frame. ID denotes held-out images from seen scenes, not training images. The retained ID set spans 13 scenes. The OOD set comprises Piazza del Popolo (337 images) and Yorkminster (162 images). Thus OOD here refers to scene membership, unlike the image-noise shift in the synthetic experiments.

Table 6: HLW-visible inclusion counts. Images are excluded when the annotated horizon is not visible in the cropped view.

Split	Original	Excluded	Retained
ID	1300	366	934
OOD	718	219	499
Combined	2018	585	1433

The configured training preprocessing uses random cropping with scale 0.85, horizontal flipping with probability 0.5, and resizing to 256×256 RGB. Validation and test preprocessing use a centre-square crop and the same resize. The visibility configuration permits up to eight training attempts, and the evaluation metadata records a deterministic centre-square/resize fallback. Training and validation use the supplied HLW training and validation H5 files. The imported records do not supply their retained sample counts.

D.5.2 MODELS AND TRAINING CONFIGURATIONS

Our real-image model adds an ImageNet-initialised EfficientNet-B0 U-Net before the fixed Hough transform, 37.6K-parameter refinement CNN and VSmax readout. Its total is approximately 6.3M parameters, not the 37.6K of the synthetic refinement-only pipeline. HLW-CNN uses a pretrained GoogLeNet-based horizon network with slope/offset classification heads and auxiliary heads (Workman et al., 2016). LETR and LINEA use the imported task-specific training wrappers for the architectures of Xu et al. (2021) and Janampa & Pattichis (2025).

The LETR configurations enable backbone pretraining but specify no DETR checkpoint. Their summaries contain one transformer of approximately 17.4M parameters. We therefore report the evaluated variants, not a reproduction claim for the original two-stage training procedure. LINEA configurations enable pretrained backbones but not pretrained full detectors. The model summaries contain set-prediction criteria and matching modules for LETR and LINEA. The `PolarLoss` label in their exported evaluation records is not evidence that it is their native training objective.

All runs configure 7,000 sampled training images per epoch, a maximum of 1,000 epochs and early stopping on validation loss with patience 20 and minimum improvement 0.001. Batch size is 64, except HLW-CNN uses 40 for training. These are configured budgets, not observed epoch counts. Unlike the synthetic comparisons, optimisers and schedules vary by model (Table 7). The imported artifacts contain no validation histories or selected-epoch records, so monitoring validation loss does not establish which checkpoint was restored for evaluation.

Table 7: Configured HLW-visible optimisation. Step schedules multiply the learning rate by 0.1, and plateau schedules use patience 10.

Model	Optimiser	Learning rate	Schedule
HLW-CNN	SGD	10^{-3}	Plateau
LETR-R18/R50	AdamW	10^{-4}	Step, every 200 epochs
LINEA-N	AdamW	$8 \cdot 10^{-4}$	Step at epoch 60
LINEA-S	AdamW	$2 \cdot 10^{-4}$	Step at epoch 25
LINEA-M	AdamW	$2 \cdot 10^{-4}$	Step at epoch 20
LINEA-L	AdamW	$2.5 \cdot 10^{-4}$	Step at epoch 11
Ours	AdamW	10^{-4}	Plateau

Weight decay is 10^{-4}, except LINEA-L uses $1.25 \cdot 10^{-4}$. LETR's backbone learning rate is 10^{-5}, and the LINEA backbone multipliers are $0.5, 0.5, 0.1, 0.05$ for N/S/M/L. HLW-CNN uses momentum 0.9 and Nesterov updates. AdamW uses betas $(0.9, 0.999)$. Parameter counts are rounded totals from each instantiated model summary, including non-trainable parameters and training-only auxiliary heads, so they are not inference-active counts. In particular, HLW-CNN's total includes its two auxiliary branches.

D.5.3 Metric and common prediction source

Let $e_i \geq 0$ denote the stored dimensionless horizon error (`scores.horizon_err`) for image i. We report the empirical cumulative distribution and its normalised integral at $T = 0.25$:

$$F(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{e_i \leq t\}, \qquad \mathrm{AUC}_T = \frac{1}{T} \int_0^T F(t)\,\mathrm{d}t = \frac{1}{N} \sum_{i=1}^{N} \max\left(0, 1 - \frac{e_i}{T}\right). \tag{36}$$

Errors above T remain in the denominator and contribute zero area. Curves use exact empirical steps, and scores use the closed-form integral, not numerical integration on a sparse plotting grid. No invalid or negative horizon-error values occur in the imported records.

All three reported splits use predictions from each run's combined inference record, `hlw_visible_test_all`. We partition it by recorded source H5 membership into ID and OOD. Consequently, combined AUC is exactly $(934\,\mathrm{AUC}_{\mathrm{ID}} + 499\,\mathrm{AUC}_{\mathrm{OOD}})/1433$ before rounding. Source-image identifiers, ground-truth parameters and native image dimensions match across all 50 runs. Separate ID/OOD records contain the same disjoint image sets but are used only for a reproducibility audit, not as alternative predictions for the table.

Our model has the largest $\mathrm{AUC}_{0.25}$ on each split, but does not dominate at every threshold (Figure 14). On the combined set, HLW-CNN has a lower median error (0.0335 versus 0.0355), whereas the fraction with error at least 0.25 is 3.91% for HLW-CNN and 1.05% for ours. The wider-range AUC gain therefore accompanies fewer large errors, not uniformly better localisation. This comparison concerns complete models with different architectures and training, not a causal readout ablation.

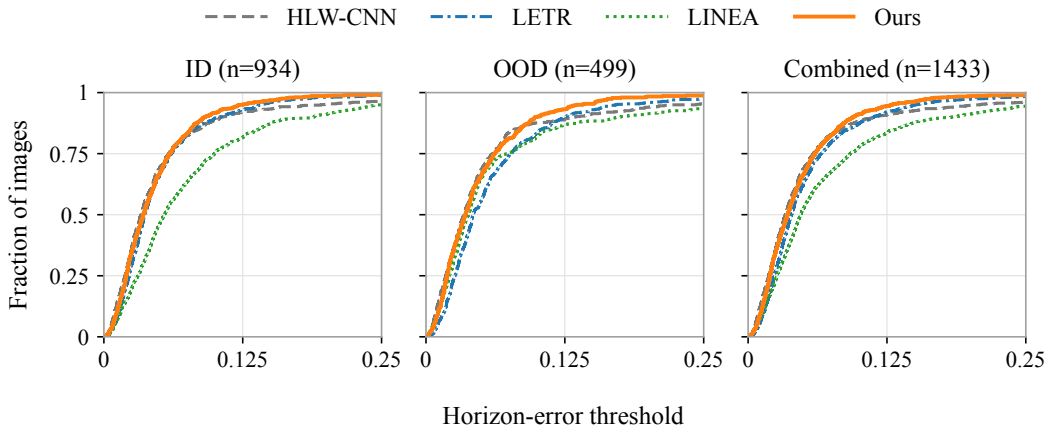


Figure 14: Empirical horizon-error distributions for the four models in Table 1. The same selected run is used in all panels. Higher curves indicate more images below a given error threshold, and their normalised areas over $[0, 0.25]$ give the reported AUCs.

D.5.4 Sweep coverage and configuration selection

The LETR sweep contains two backbones (R18 and R50), and the LINEA sweep four sizes (N/S/M/L), each evaluated with $20, 50, 100, 200, 350, 500, 750, 1000$ queries. All 48 configurations are included in Tables 8 and 9 and Figures 15–17. Within each variant, query count is the swept architecture field. Training settings also differ between model sizes as listed above.

Table 8: Complete LETR sweep on HLW-visible. Scores are $\text{AUC}_{0.25}$, and parameters are approximate totals in millions. Bold marks the single combined-test-selected configuration, not independent winners per split.

Variant	Queries	ID	OOD	Combined	Params
LETR-R18	20	0.7696	0.7388	0.7589	28.8
LETR-R18	50	0.7683	0.7263	0.7537	28.8
LETR-R18	100	0.5482	0.6593	0.5869	28.8
LETR-R18	200	0.3254	0.2349	0.2939	28.8
LETR-R18	350	0.2190	0.2728	0.2378	28.9
LETR-R18	500	0.0219	0.0329	0.0257	28.9
LETR-R18	750	0.2974	0.3670	0.3216	29.0
LETR-R18	1000	0.2183	0.2685	0.2358	29.1
LETR-R50	20	0.7804	0.7520	0.7705	41.5
LETR-R50	**50**	**0.7993**	**0.7574**	**0.7847**	41.5
LETR-R50	100	0.0635	0.0835	0.0705	41.5
LETR-R50	200	0.3225	0.2308	0.2906	41.5
LETR-R50	350	0.1855	0.2268	0.1999	41.6
LETR-R50	500	0.1316	0.1256	0.1295	41.6
LETR-R50	750	0.2133	0.2569	0.2285	41.7
LETR-R50	1000	0.3457	0.3129	0.3343	41.7

We select one run per swept family by maximum unrounded *combined-test* $\text{AUC}_{0.25}$, breaking exact ties deterministically. This selects LETR-R50 with 50 queries and LINEA-N with 500 queries. HLW-CNN and ours each have one imported run. The chosen run supplies every split score and its parameter count, and no split-specific winners are substituted. This is an exploratory best-of-tested-configurations comparison. The test set informs hyperparameter selection, so the selected baseline scores are not unbiased estimates for validation-selected configurations. Publishing the full sweep makes selection transparent but does not remove this limitation.

Table 9: Complete LINEA sweep on HLW-visible, with the conventions of Table 8. The selected row uses the same run on all splits.

Variant	Queries	ID	OOD	Combined	Params
LINEA-N	20	0.6728	0.6969	0.6812	3.8
LINEA-N	50	0.6129	0.7174	0.6493	3.8
LINEA-N	100	0.6778	0.7399	0.6994	3.8
LINEA-N	200	0.6222	0.7316	0.6603	3.8
LINEA-N	350	0.6066	0.6547	0.6233	3.9
LINEA-N	**500**	**0.6949**	**0.7496**	**0.7139**	3.9
LINEA-N	750	0.6968	0.7413	0.7123	3.9
LINEA-N	1000	0.6468	0.6668	0.6537	3.9
LINEA-S	20	0.6432	0.7187	0.6695	8.4
LINEA-S	50	0.6270	0.7127	0.6569	8.4
LINEA-S	100	0.6262	0.7012	0.6523	8.4
LINEA-S	200	0.6195	0.7024	0.6484	8.4
LINEA-S	350	0.6305	0.7239	0.6631	8.5
LINEA-S	500	0.6010	0.6937	0.6333	8.5
LINEA-S	750	0.6365	0.7079	0.6613	8.6
LINEA-S	1000	0.6266	0.7023	0.6529	8.6
LINEA-M	20	0.6084	0.6857	0.6353	13.4
LINEA-M	50	0.6035	0.6775	0.6293	13.4
LINEA-M	100	0.6326	0.7084	0.6590	13.4
LINEA-M	200	0.6050	0.6912	0.6350	13.4
LINEA-M	350	0.6024	0.6926	0.6338	13.5
LINEA-M	500	0.6001	0.6820	0.6286	13.5
LINEA-M	750	0.6040	0.6851	0.6323	13.6
LINEA-M	1000	0.5986	0.6683	0.6228	13.6
LINEA-L	20	0.5755	0.7040	0.6202	25.5
LINEA-L	50	0.5938	0.7193	0.6375	25.5
LINEA-L	100	0.5932	0.6873	0.6260	25.5
LINEA-L	200	0.5674	0.6137	0.5835	25.5
LINEA-L	350	0.5361	0.6498	0.5757	25.5
LINEA-L	500	0.5654	0.6724	0.6027	25.6
LINEA-L	750	0.5904	0.6679	0.6174	25.7
LINEA-L	1000	0.4871	0.6774	0.5534	25.7

LETR performs best with 20–50 queries in this setup, and several larger-query runs perform poorly. Without training histories and a verified decoding path, we do not attribute those failures to a specific mechanism. LINEA-N gives the highest observed combined AUC among the LINEA sizes. This does not establish a universal advantage for smaller networks. The held-out set spans only two scenes, limiting conclusions about generalisation to arbitrary new scenes.

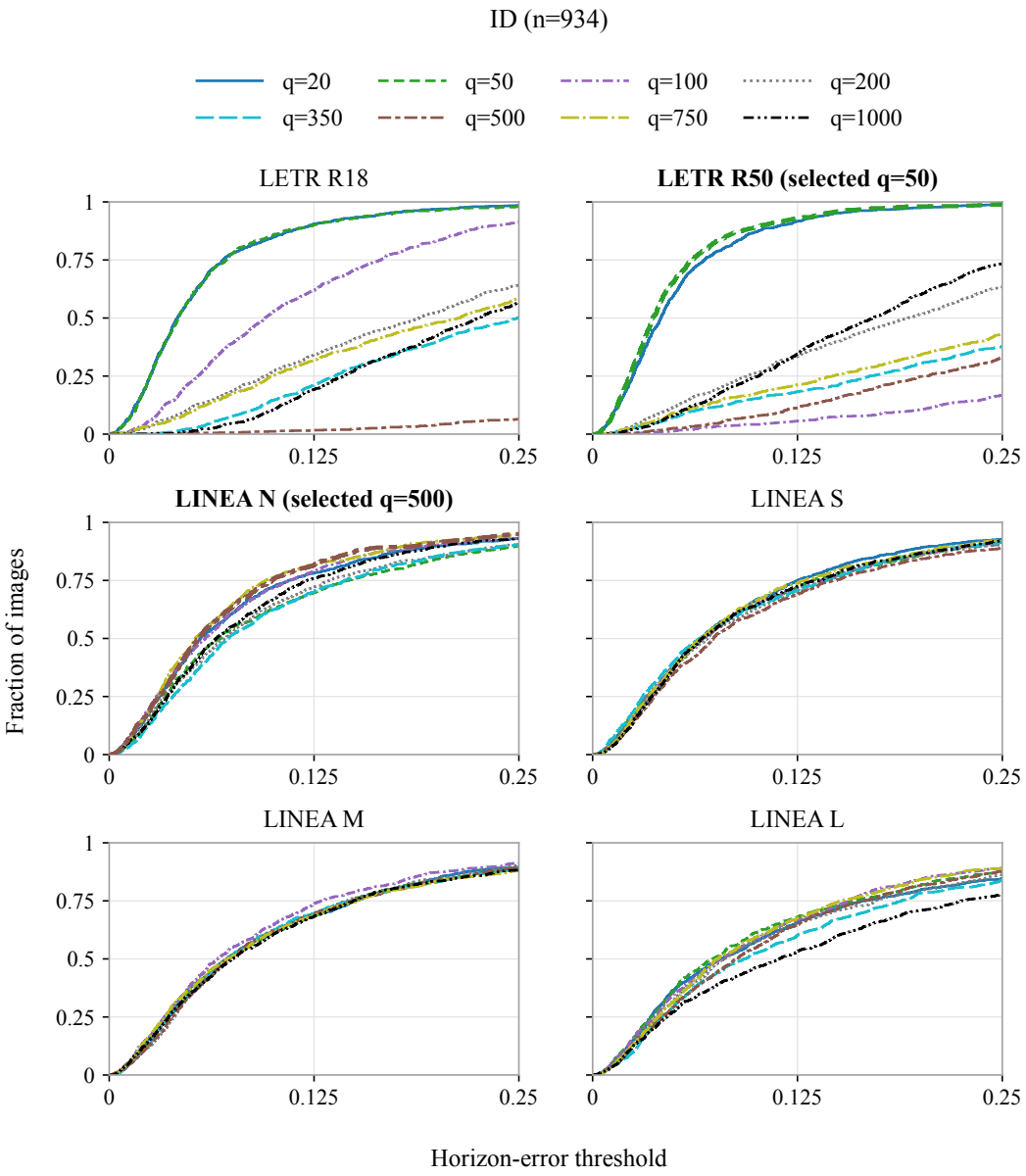


Figure 15: Full query-count sweep on the 934 ID images. Each panel fixes a backbone or model size and shows all eight query counts. The thicker traces mark the configurations selected for the main comparison by combined-test AUC, not by this panel's split. Curves are empirical CDFs, not uncertainty bands.

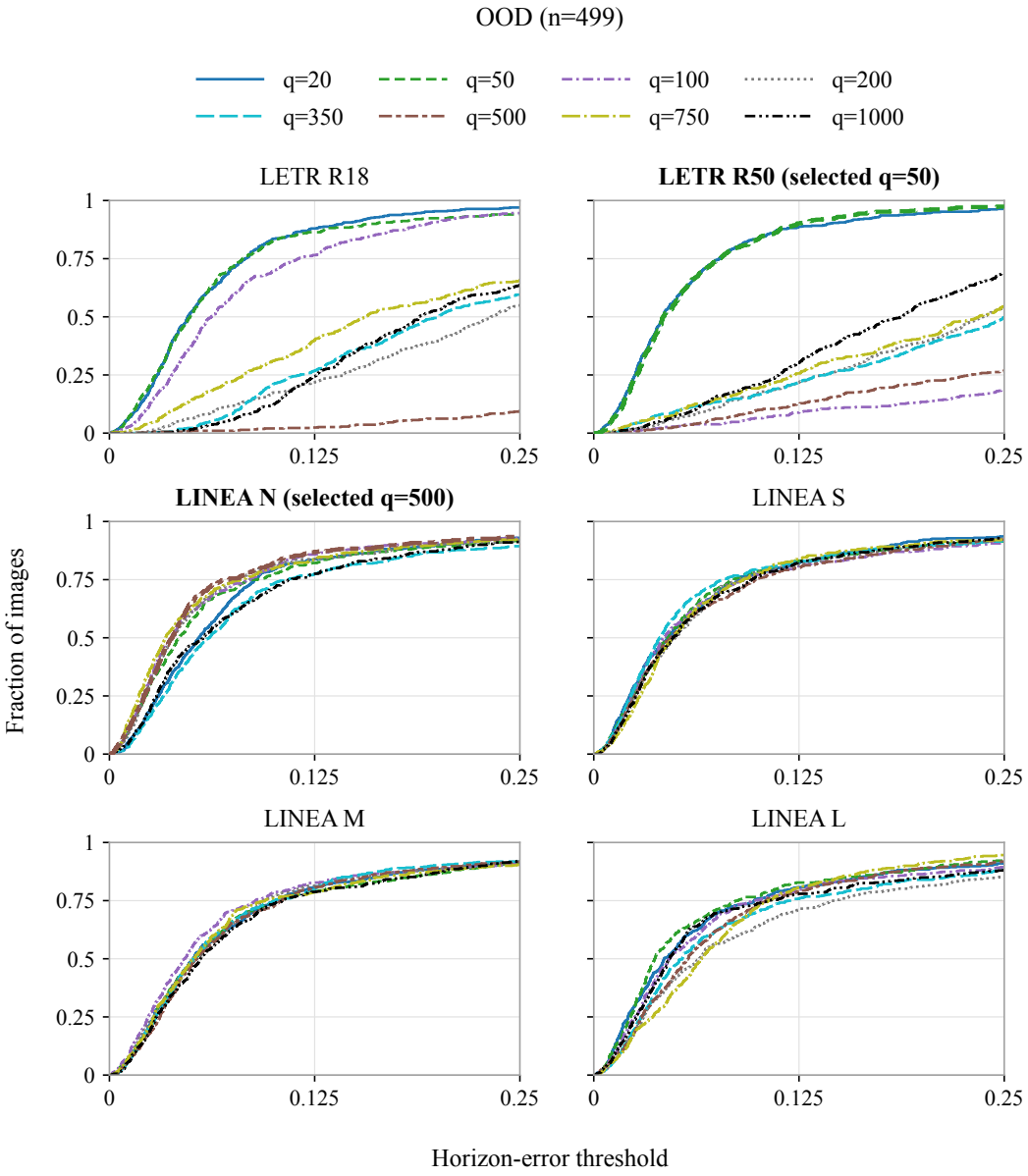


Figure 16: Full query-count sweep on the 499 OOD images, with the same configurations and styles as Figure 15.

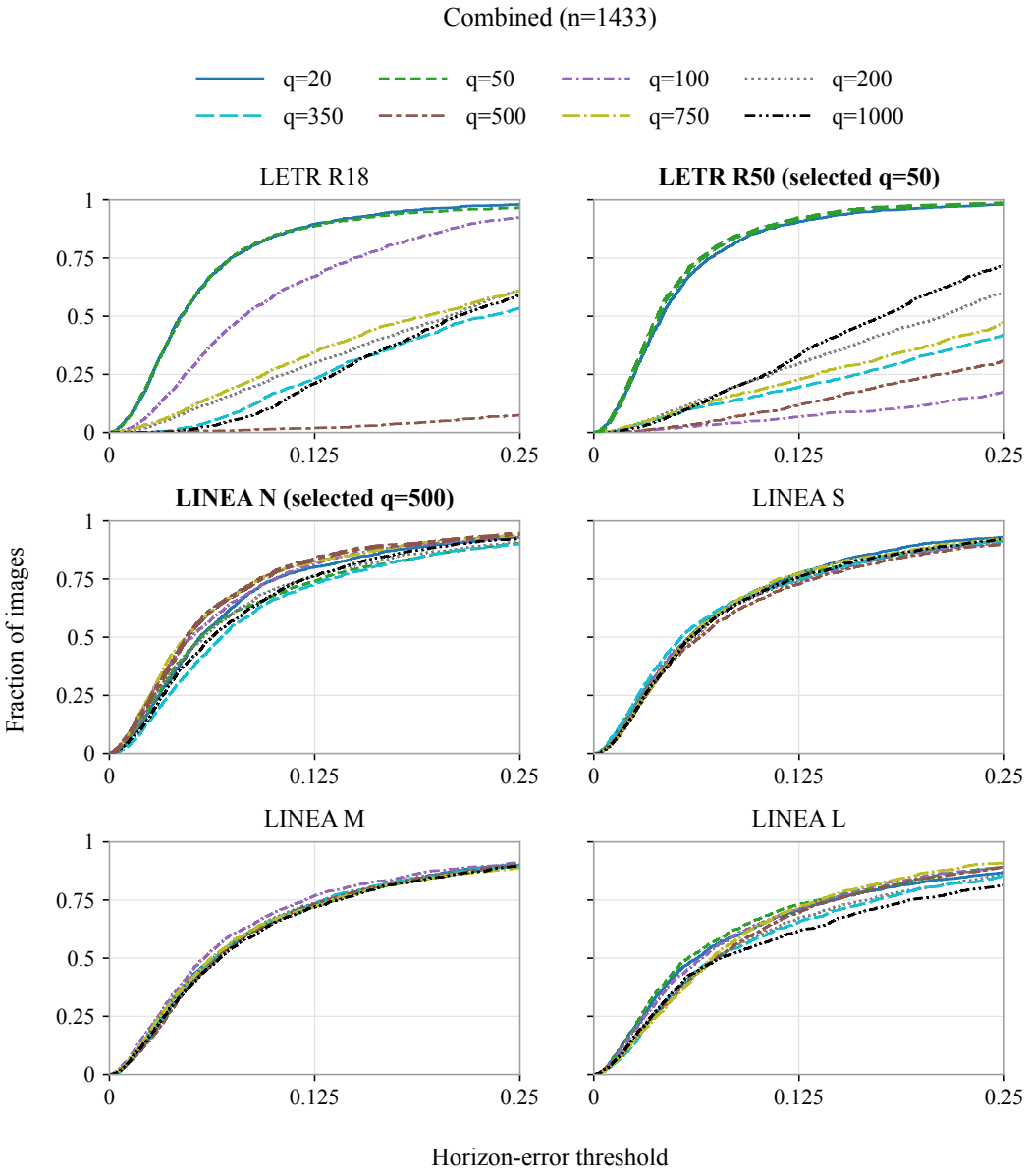


Figure 17: Full query-count sweep on all 1,433 HLW-visible test images. Normalised area over $[0, 0.25]$ is the configuration-selection criterion. All curves, including poorly performing configurations, are retained.