Distributional Estimation of 3D Object Orientation

Minjoon Kim **Chunghyun Park** **Minsu Cho***

Pohang University of Science and Technology (POSTECH)
https://mjtrace.github.io/DORE

Abstract

The canonical orientation of a 3D object is often not unique, yet inherently ambiguous – symmetric objects, for instance, admit multiple equally valid canonical frames that no single-output regressor can faithfully represent. Existing methods struggle to handle this ambiguity: they either predict a single orientation, enforce symmetry only under pre-defined group assumptions, or rely on costly test-time augmentation to make their predictions robust. We propose a distributional approach that models 3D object orientation as a distribution over SO(3), accommodating both finite and continuous groups without any group-specific architectural bias. To this end, we develop an SO(3)-equivariant framework in which a truncated Wigner-D expansion encodes multiple valid orientations as a continuous, multi-modal distribution. On the ShapeNet benchmark, our method surpasses the prior state of the art across all categories under single-pass inference. Furthermore, even without test-time rotation augmentation, it matches the accuracy of a 50-vote ensemble baseline to within 0.5 percentage points at roughly $12\times$ lower latency, demonstrating that explicit distributional modeling is both accurate and efficient.

1 Introduction

Many important tasks in 3D computer vision and robotics – including shape recognition [46, 37, 38, 23, 42], CAD model registration [2, 17], and robotic manipulation [48, 44, 43, 3] – require objects to be placed in a canonical coordinate frame, a process known as canonicalization. The task of 3D object orientation estimation aims to identify an object's canonical orientation, defined by three principal axes (up, front, and side), from which the shape can be rotated into the canonical coordinate frame, where each axis aligns with its corresponding coordinate direction. The canonical orientations of objects are not intrinsic geometric quantities; rather, they reflect human judgments grounded in physical and functional considerations, which makes purely geometric approaches insufficient [39]. When a large dataset of canonically oriented shapes is available, however, orientation estimation can be naturally framed as a supervised learning problem [31, 39]. Despite progress in 3D deep learning, estimating stable, consistent orientations across diverse categories remains an open problem.

Recent advances in equivariant learning have led to more reliable and consistent orientation estimation. $SO(3)$-equivariant orientation estimators [20, 21] build equivariance directly into the architecture, ensuring that predictions transform consistently under arbitrary rotations of the input. This structural guarantee eliminates the need for rotation augmentation to learn equivariance and yields more reliable estimates across diverse input poses. Yet these methods assign only a single orientation per shape by regression, which proves insufficient when the input shape admits multiple valid orientations. Many real-world objects are geometrically identical under several distinct rotations, giving rise to a set of equivalent orientations rather than a unique one. When this ambiguity stems from rotational symmetry, the set of equivalent orientations forms an orbit of the symmetry group. Recent work [39] tackles this problem by assuming a fixed symmetry group and approximating equivariance through

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

test-time augmentation. However, its fixed group and discrete output cannot represent shapes with symmetry groups outside the octahedral group or near-continuous rotational ambiguity.

To address the aforementioned limitations, we introduce *Distributional Orientation Estimator* (DORE), a general framework for ambiguity-aware orientation estimation that is at once $SO(3)$-equivariant by construction and distributional in its output, covering finite and continuous symmetry groups within a single representation. It builds $SO(3)$-equivariance into its architecture rather than approximating it, and thus inference does not require rotation augmentation. It represents the orientation output as a continuous distribution over $SO(3)$ via its truncated Wigner-D expansion. Since a band-limited Wigner-D expansion can approximate any square-integrable function on $SO(3)$, the representation fixes no symmetry group; rotational symmetry is the structured instance studied in this work. To provide orbit supervision during training without manual annotation, we develop an automatic symmetry-labeling procedure that infers each shape's symmetry group from its geometry alone. This supervision currently covers cyclic and dihedral groups C_n and D_n with $n \leq 6$.

Our main contributions can be summarized as follows.

- We introduce DORE, a general $SO(3)$-equivariant framework for ambiguity-aware orientation estimation that effectively predicts multiple valid orientations as a continuous distribution over $SO(3)$ with no pre-defined symmetry group and no rotation augmentation at inference.
- We construct the supervision target by automatically labeling each shape's symmetry group and smoothing the orbit on $SO(3)$ with a heat kernel, which attenuates spectral truncation artifacts, and use the label statistics to choose each category's band limit $L_{\max}$ before training.
- On ShapeNet, under per-category training and single-pass inference, DORE outperforms the state-of-the-art fixed-quotient method in every category. Without rotation augmentation, full-density inference comes within 0.5 pp of the 50-vote accuracy of Scarvelis et al. [39] at $12\times$ lower latency.

2 Related work

3D orientation and symmetry analysis. Canonical 3D orientation estimation must account for the multiple valid orientations of symmetric objects. Upright-axis estimators [47, 31] recover only an up-axis from an image or point cloud. Equivariant regressors such as VN-SPD [20] and CHOIR [21], both built on Vector Neurons [8], predict a full orientation but still return a single pose and thus cannot represent this ambiguity. To address it, Scarvelis et al. [39] introduce a quotient-regression cascade tied to the octahedral group O, with equivariance approximated through random input rotations at test time. This construction handles octahedrally consistent shapes but cannot represent C_n or D_n symmetry with $n > 4$, icosahedral symmetry, or continuous rotational symmetry. Complementary work [28, 33, 12, 40] detects symmetry planes, axes, or rotation centers through transformation voting or learned prediction. For these methods, symmetry is the output rather than a means of defining the equally valid orientations an estimator should predict. DORE addresses both representation and supervision: it is $SO(3)$-equivariant by construction, represents orientation without fixing a symmetry group, and automatically supervises orbits of C_n and D_n with $n \leq 6$.

$SO(3)$-equivariant networks. Equivariant networks provide the architectural foundation for predicting orientations that transform consistently with the input. Tensor Field Networks [41] and $SE(3)$-Transformers [11] construct equivariant message passing using spherical harmonics and Clebsch–Gordan coefficients. Vector Neurons [8] extend point-cloud backbones to 3D vector-valued features, supporting rotation-aware shape representation learning [32]. e3nn [13] supports $O(3)$-equivariant modeling through steerable convolutions over irreducible representations. Frame averaging [36] takes a different approach, obtaining equivariance by symmetrizing a backbone over canonicalization frames. Building on Vector Neurons and e3nn, DORE uses an equivariant backbone and spectral head to predict a distribution over $SO(3)$.

Probability estimation on $SO(3)$. Predicting a distribution over $SO(3)$ allows orientation estimators to express multiple valid poses. Existing approaches differ in how they represent this distribution. *Histogram* methods [34, 4] classify over fixed discrete sets of rotations or rotation angles, trading angular resolution against the size of the label space. AQuaMaM [1] instead factorizes the density autoregressively over unit-quaternion components, enabling likelihood-based training at the cost

of sequential decoding. *Implicit* methods [30, 18, 25] represent flexible densities using MLPs, hypernetworks, or normalizing flows, but require per-query evaluation, making dense evaluation and distribution-level losses expensive. *Parametric* methods [35, 29, 14, 9, 50] use fixed-form distributions, including von Mises mixtures, matrix Fisher, Bingham, and Laplace-inspired forms. A fixed number of components can limit the representation of high-order symmetries, while the infinite-mixture variant of Prokudin et al. [35] relies on sampling-based likelihood approximations. *Explicit Fourier* methods represent the density through Wigner-D coefficients predicted in a single forward pass. Within this family, Image-to-Sphere [22] projects features of a non-equivariant 2D image encoder onto the sphere and applies $SO(3)$-equivariant layers, so the full model is not equivariant by construction. Lee and Cho [24] build on a similar architecture and train its Wigner-D coefficients against a single ground-truth pose in the frequency domain. DORE extends this distributional perspective to 3D point clouds, where symmetry yields an entire orbit of valid orientations. Accordingly, supervision targets a heat-kernel-smoothed orbit distribution rather than a single pose, while both the backbone and spectral head remain $SO(3)$-equivariant.

3 Preliminaries

Rotation group and finite symmetry subgroups. The special orthogonal group $SO(3)$ is the Lie group of orientation-preserving rotations of $\mathbb{R}^3$, equipped with a unique bi-invariant probability measure, the Haar measure μ. For a finite symmetry group $G \leq SO(3)$ and a reference orientation $R_0 \in SO(3)$, the orbit of R_0 under G is $\{gR_0 \mid g \in G\}$. Every finite subgroup of $SO(3)$ is isomorphic to one of the cyclic C_n, dihedral D_n, tetrahedral T, octahedral O, or icosahedral I groups. Throughout the paper, $\mathbf{P} = \{p_i\}_{i=1}^N \subset \mathbb{R}^3$ denotes an N-point cloud (treated as an $N \times 3$ matrix), $|G|$ is the order of G, and $g.\mathbf{P}$ is the action of $g \in SO(3)$ on $\mathbf{P}$.

Harmonic analysis on $SO(3)$. The irreducible unitary representations of $SO(3)$ are indexed by non-negative integers $l \in \{0, 1, 2, \ldots\}$ and realized by the Wigner-D matrices $D^l : SO(3) \to \mathbb{R}^{(2l+1)\times(2l+1)}$, which satisfy the homomorphism property $D^l(gh) = D^l(g)D^l(h)$ [41, 13]. The Peter–Weyl theorem [7] gives the Fourier expansion of any square-integrable $f \in L^2(SO(3))$ as

$$f(R) = \sum_{l=0}^{\infty} (2l+1)\,\mathrm{tr}\big(H^l D^l(R)^\top\big), \quad H^l = \int_{SO(3)} f(R)\, D^l(R)\, d\mu(R), \tag{1}$$

where $H^l \in \mathbb{R}^{(2l+1)\times(2l+1)}$ is the degree-l Fourier coefficient matrix. We use $\hat{H}^l$ for predicted coefficients throughout, and write a band-limited expansion truncated at $L_{\max}$ as the block-diagonal matrix $H = \bigoplus_{l=0}^{L_{\max}} H^l$ with $\sum_{l=0}^{L_{\max}} (2l+1)^2$ coefficients in total.

$SO(3)$-equivariance. A map $h : \mathcal{X} \to \mathcal{Y}$ is $SO(3)$-equivariant if $h(R \cdot x) = R \cdot h(x)$ for all $R \in SO(3)$ and $x \in \mathcal{X}$, under group actions defined appropriately for each space. For a network whose output is a band-limited expansion $\{\hat{H}^l\}$, equivariance with respect to an input rotation by $R \in SO(3)$ amounts to $\hat{H}^l(\mathbf{P}R) = \hat{H}^l(\mathbf{P})\, D^l(R)$ holding at every degree $l \in \{0, 1, \ldots, L_{\max}\}$.

4 Proposed method

Distributional Orientation Estimator (DORE) predicts a continuous orientation distribution with an $SO(3)$-equivariant network. To represent the multiple valid orientations of symmetric objects, we supervise this distribution with an orbit-based target. We first define the spectral representation (Sec. 4.1), then describe the network (Sec. 4.2), orientation inference (Sec. 4.3), and training objective (Sec. 4.4). Finally, we explain how automatic symmetry labels provide orbit supervision (Sec. 4.5) and determine the representation's band limit (Sec. 4.6).

4.1 Spectral orientation representation

Let $\mathbf{P} \in \mathbb{R}^{N\times 3}$ be a 3D point cloud with unknown orientation in $SO(3)$. For an object with symmetry group $G \leq SO(3)$, any rotation $R \in SO(3)$ is geometrically indistinguishable from every element of its orbit $\{gR \mid g \in G\}$, so predicting a single rotation is ill-posed. We therefore model orientation as a density f on $SO(3)$ concentrated around this orbit. We represent f through its Fourier coefficients

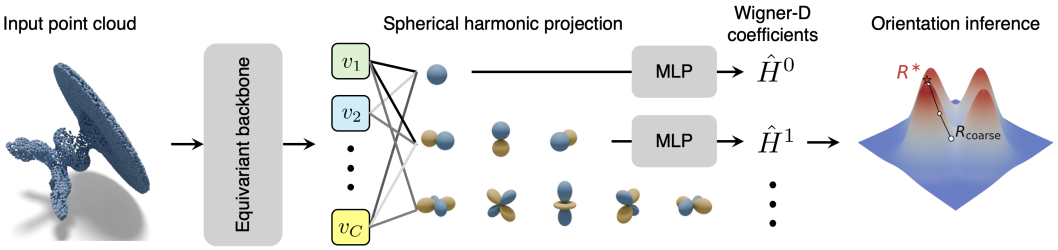

Figure 1: **Overview of DORE.** A Vector Neurons [8] backbone extracts C equivariant features $\{v_c\}_{c=1}^{C}$. An e3nn [13] spectral head projects them through $Y^l(v_c)$ and block-wise MLPs to predict Wigner-D coefficients $\hat{H}^l$ (Sec. 4.2). A heat-kernel spectral bound guides the output projection (Sec. 4.1), and a Laplace–Beltrami-weighted loss supervises the coefficients (Sec. 4.4). Grid search and Newton–Raphson refinement then recover the highest-density mode R^* of $f(R)$ (Sec. 4.3).

(Eq. (1)), truncated at degree $L_{\max}$. This truncation preserves unit mass but allows negative values, a trade-off quantified in Appendix C.2. Concretely, our network h outputs a block-diagonal coefficient matrix $h(\mathbf{P}) = \bigoplus_{l=0}^{L_{\max}} \hat{H}^l$ with $\sum_{l=0}^{L_{\max}} (2l+1)^2$ coefficients in total, e.g., 35, 165 and 455 at $L_{\max} = 2, 4, 6$, the bands used in Table 1. Thus, $L_{\max}$ controls the representation's capacity; Sec. 4.6 describes how we choose it from the symmetry labels of each class. For now, we assume that each training shape has a symmetry group G^*, which we estimate automatically from the input geometry when symmetry annotations are unavailable (Sec. 4.5).

Orbit-based target distribution. We assume that each shape has a canonical reference pose R_0 and that its orientation ambiguity arises solely from symmetry. Every element of $\{gR_0 \mid g \in G^*\}$ is then an equally valid canonical orientation. We encode these orientations as a normalized mixture of Dirac deltas over the orbit. In the canonical frame ($R_0 = I$),

$$f^*(R) = \frac{1}{|G^*|} \sum_{g \in G^*} \delta(g^{-1}R), \tag{2}$$

and for the randomly rotated point cloud $\mathbf{P}R$, the target transforms as $f^*(R') \mapsto f^*(R'R^{-1})$; in coefficient form, each block updates as $H^{*l} \leftarrow H^{*l} D^l(R)$. Truncating Eq. (2) introduces spectral ringing. To attenuate it, we convolve f^* with a heat kernel on $SO(3)$, attenuating each degree-l component by $\kappa_l = \exp(-l(l+1)\sigma^2)$. The smoothed target then satisfies $\|H^{*l}\|_F \le \sqrt{2l+1} \cdot \kappa_l$ (Appendix B.3). Motivated by this bound, we project predicted coefficients into a bounded range with a ceiling no tighter than the target bound and learnable multiplicative rescaling per degree. We conjecture that this output projection accelerates training convergence.

4.2 Equivariant coefficient prediction network

Our network combines a Vector Neurons (VN) backbone [8] with a spectral head built on e3nn [13], both $SO(3)$-equivariant by construction (Figure 1). The backbone produces C type-1 irrep features at each point. Mean pooling over points preserves equivariance and yields a global descriptor of C type-1 vectors. The spectral head predicts Wigner-D coefficients through spherical harmonics and two equivariant linear layers separated by a Gate nonlinearity, with fewer channels at higher degrees.

Block-diagonal equivariance. The head converts the backbone's type-1 features into coefficients at every degree of the expansion. Each pooled vector v_c, which transforms as $v_c \mapsto R^\top v_c$, is lifted by spherical harmonic projection, $\mathbf{z}_c^l = Y^l(v_c) \in \mathbb{R}^{2l+1}$ with $Y^l(Rv_c) = D^l(R)\, Y^l(v_c)$. Within each degree, the linear layers act only on channel indices, whereas $D^l(R)$ acts only on irrep indices; the two operations therefore commute. The Gate also preserves equivariance by rescaling each $l{\ge}1$ channel with the sigmoid of a rotation-invariant $l{=}0$ channel. The second layer outputs $2l{+}1$ degree-l channels, which form the rows of the $(2l{+}1) \times (2l{+}1)$ block $\hat{H}^l$. Together, these operations ensure that the output transforms as Wigner-D coefficients: $\hat{H}^l(\mathbf{P}R) = \hat{H}^l(\mathbf{P})\, D^l(R)$ for all R and l. Appendix A.3 gives the layer structure and the equivariance proof.

4.3 Orientation inference

Given the coefficients $\{\hat{H}^l\}_{l=0}^{L_{\max}}$, we reconstruct the predicted distribution as

$$f(R) = \sum_{l=0}^{L_{\max}} (2l+1)\ \mathrm{tr}\big(\hat{H}^l\, D^l(R)^\top\big), \quad \hat{H}^l \in \mathbb{R}^{(2l+1)\times(2l+1)}. \tag{3}$$

For non-trivial symmetries, the orbit target is multi-modal, so we seek the local maxima of f,

$$\mathcal{M}_f = \big\{R \in SO(3) : f(R) \geq f(R') \text{ for all } R' \text{ near } R\big\}. \tag{4}$$

For ideal C_n and D_n orbit targets ($n \geq 2$), this set contains the valid orientations. At $L_{\max} \geq n$, they are isolated maxima satisfying $\nabla_R f(R) = 0$ and $\nabla_R^2 f(R) \prec 0$ on the tangent space. Below this band, for $L_{\max} \geq 2$, the maxima form a continuous ridge (Appendix B.1). For top-1 evaluation, we seek the highest-density mode $R^* = \mathrm{argmax}_{R\in SO(3)} f(R)$. We approximate it with a coarse grid search, then refine the best grid point using Newton–Raphson iterations on $SO(3)$.

Coarse mode search. Since $f(R)$ is linear in the flattened coefficient vector $c \in \mathbb{R}^D$, $D = \sum_{l=0}^{L_{\max}}(2l+1)^2$, evaluating f at any rotation R_i reduces to a dot product $f(R_i) = w(R_i) \cdot c$, where $w(R_i) \in \mathbb{R}^D$ stacks the scaled Wigner-D matrices at R_i. We precompute a scoring matrix $W \in \mathbb{R}^{N_{\mathrm{grid}}\times D}$ over a Hopf fibration grid of $N_{\mathrm{grid}} = 6{,}912$ points (3,456 distinct rotations). A single matrix-vector product $\mathbf{f} = Wc$ thus evaluates all grid scores, and the highest-scoring candidate $R_{\mathrm{coarse}} = \mathrm{argmax}_{R\in\{R_i\}} f(R)$ initializes the Newton–Raphson refinement described below. Precomputing and sharing W across samples makes this initialization inexpensive.

Newton–Raphson refinement. Starting from the highest-scoring grid point R_{coarse}, Newton–Raphson iterations on $SO(3)$ aim to converge to a nearby local maximum of f. At each iterate R_t, we parameterize local perturbations through the exponential map $R(\omega) = R_t \exp(\hat{\omega})$, with $\omega \in \mathbb{R}^3 \cong \mathfrak{so}(3)$. The gradient and Hessian of f at $\omega = 0$ are estimated by finite differences along the Lie-algebra basis, and each Newton step $\omega^* = -(\nabla^2 f)^{-1}\nabla f$ is clamped to a trust region $\|\omega^*\| \leq 0.1$ rad per iteration. Appendix B.1 gives the update equations and finite-difference formulas.

Our default inference procedure combines one forward pass with this mode search. To reduce latency, we pipeline the GPU forward pass and CPU search across batches, searching one batch while predicting coefficients for the next. The full density also supports inference beyond this top-1 selection: a verifier can select among its peaks, and the linear dependence of f on its coefficients allows densities from several inputs to be combined by coefficient averaging (Sec. 5.4; Appendix B.2).

4.4 Training objective

For each training shape we sample two random rotations R_1, R_2 and form views $\mathbf{P}_i = \mathbf{P}R_i$, whose targets $H_i^{*l} = H_{\mathrm{can}}^{*l}\, D^l(R_i)$ rotate the canonical target. This closed-form transformation lets every rotated view inherit the symmetry label computed once before training (Sec. 4.5). Training-time rotation augmentation therefore requires neither repeated symmetry detection nor per-view annotation. We regress the predicted blocks against their heat-kernel-smoothed targets with weights given by the corresponding Laplace–Beltrami eigenvalues on $SO(3)$:

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \alpha\,\mathcal{L}_{\mathrm{equiv}}, \qquad \mathcal{L}_{\mathrm{sup}} = \frac{1}{Z}\sum_{l=0}^{L_{\max}} w_l\,\mathcal{L}_{\mathrm{mse}}^{(l)}, \quad \mathcal{L}_{\mathrm{mse}}^{(l)} = \frac{1}{2}\sum_{i=1}^{2}\|\hat{H}_i^l - H_i^{*l}\|_{\mathrm{rms}}^2, \tag{5}$$

where $w_l = l(l+1)$ for $l \geq 1$ and $w_0 = 0$, $Z = \sum_l w_l$, and $\|\cdot\|_{\mathrm{rms}}$ is the RMS over block entries. These weights emphasize the higher-frequency coefficients that encode the orbit target's multi-modal structure. Appendix C.1 analyzes the resulting distance and its sensitivity to the weights. We also apply a soft equivariance-consistency regularizer $\mathcal{L}_{\mathrm{equiv}}$ with $\alpha = 0.05$ (Appendix B.3).

4.5 Automatic symmetry labeling

The training objective requires a symmetry group G^* for each shape, but ShapeNet [5] and ModelNet40 [46] provide no such annotations. We therefore estimate G^* from geometry alone by measuring how nearly each candidate group G leaves $\mathbf{P}$ unchanged. We use a one-sided, unsquared ($p = 1$)

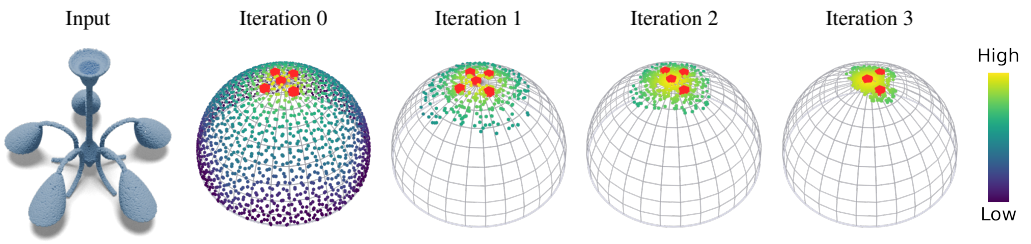

Figure 2: **Beam search over rotation axes.** Colors show the invariance score $S(C_n^a)$ for one order n on the upper hemisphere (Eq. (6)); yellow indicates higher scores. Starting from 2000 Fibonacci-lattice axes, the search repeatedly contracts around the top-k candidates marked in red.

Chamfer distance [10], which gives large residuals less relative weight than squared distances and is equivalent here to the symmetric distance at half the nearest-neighbor cost (Appendix A.4):

$$S(G) = \mathbb{E}_{g\sim\mathrm{Unif}(G)}\left[-d_{\mathbf{P}}^{(1)}\big(g.\mathbf{P}\big)\right], \quad d_X^{(p)}(Y) = \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y}\|x-y\|_2^p, \tag{6}$$

so larger scores favor G. The candidate set comprises cyclic and dihedral groups,

$$\mathcal{G} = \bigcup_{n=2}^{6}\{C_n^a \mid a\in\mathbb{RP}^2\} \cup \bigcup_{n=2}^{6}\{D_n^a \mid a\in\mathbb{RP}^2\}, \tag{7}$$

where C_n^a is the cyclic group of order n about axis $a\in\mathbb{RP}^2$ (a unit vector up to sign). The dihedral group D_n^a has order $2n$ and additionally includes a flip about a direction $b\perp a$, identified in the post-check below. We exclude the identity $\{e\}$ from $\mathcal{G}$ and assign it separately when no non-trivial symmetry passes the existence test based on the raw geometric scores.

Beam search over rotation axes. For each cyclic order $n\in\{2,\dots,6\}$ we maximize $S(C_n^a)$ over a via a beam search, illustrated in Figure 2. We initialize the axes uniformly on a Fibonacci sphere lattice [16], then iteratively shrink the angular sampling radius around the best candidates to obtain $\hat{a}_n$. To test whether this cyclic symmetry extends to a dihedral group, we evaluate 180° flips about the n candidate C_2 generators perpendicular to $\hat{a}_n$. Detected flips add D_n as a candidate.

Group selection with a size prior. We first test whether any non-trivial symmetry has sufficient geometric support. If $\max_{G\in\mathcal{G}} S(G) < \tau$, no candidate passes the threshold, so we assign $G^* = \{e\}$ using the raw scores alone. Otherwise, we select among the non-trivial candidates with a log-size prior. This prior favors larger groups because they impose more geometric constraints, making an equal score stronger evidence for a larger group. With weight $\beta > 0$, the selection rule is

$$G^* = \operatorname*{argmax}_{G\in\mathcal{G}}\big[\, S(G) + \beta\log|G|\,\big]. \tag{8}$$

Full pseudocode is in Appendix E.1; hyperparameters and implementation details are in Appendix B.3. We audit the assigned symmetry labels on 200 random shapes in each of 13 classes, checking threshold calibration and geometric consistency, and report the full label census (Appendices E.2 and E.3).

4.6 Symmetry-informed band selection

The symmetry labels also guide the choice of $L_{\max}$, which balances expressivity and sample efficiency. The n-fold orbit structure of C_n or D_n first becomes expressible at degree $l = n$ (Proposition 1). However, each additional degree introduces $(2l+1)^2$ coefficients, which small categories may not adequately constrain (Appendix D.6). Using the labels from Sec. 4.5, we therefore choose each class's smallest band *before training* that represents all groups with a share of at least 2%:

$$L_{\mathrm{sym}} = \max\big\{\, l(G) \;:\; p_G \geq 2\% \,\big\}, \qquad l(C_n) = l(D_n) = n, \tag{9}$$

where $l(G)$ is the smallest band limit that expresses the orbit of G, and p_G is the fraction of training shapes in the class labeled G (Table 24). The rule presumes no polyhedral symmetry, which our scan does not find (Appendix E.3). If no non-trivial group meets the threshold, we use the floor $L_{\mathrm{sym}} = 2$. Appendix D.3 compares this label-based choice with an exhaustive band sweep.

5 Experiments

5.1 Setup

Dataset and baselines. We train and evaluate on ShapeNet [5] following prior protocols [21, 39]. We extend the four standard categories (airplane, car, chair, and table) with lamp and cellphone to cover a broader range of finite symmetry orders. We compare against CHOIR [21] and the state-of-the-art method of Scarvelis et al. [39]. CHOIR predicts a single pose using a Vector Neurons backbone and an $SO(3)$-invariant residual module. The latter method combines quotient regression on $SO(3)/O$ (the orienter) with a 24-way coset classifier (the flipper). We retrain both methods per category using their public code and identical splits, and report our own measurements throughout.

To avoid an input distribution shift, each method retains its training-time input convention at inference. Scarvelis et al. [39] use $2{,}000$ points at radius 0.95 with normals from the ShapeNetCore v2 meshes, whereas DORE and CHOIR use $1{,}024$ points without normals at the dataset's scale (Appendix B.4); DORE enlarges only inputs substantially smaller than the class's training shapes. For a common evaluation, we score all predictions on the same rotations and $2{,}000$-point clouds at radius 0.95.

Evaluation metrics. Following Scarvelis et al. [39], our metrics compare canonicalized predictions to the reference cloud, treating every orientation in the same symmetry orbit as correct. For each test object with canonical cloud $\mathbf{P}$, we sample K random rotations $\{R_k\}_{k=1}^{K}$ and predict the highest-density mode $\hat{R}_k$ of the predicted density f on each rotated input $\mathbf{P}R_k$ (Appendix B.1). The *mean Chamfer distance* and *symmetry-aware orientation accuracy* at threshold ϵ are defined as

$$\overline{d}_{\mathrm{CD}} = \frac{1}{K}\sum_{k=1}^{K} d_{\mathrm{CD}}\big(\mathbf{P}R_k\hat{R}_k^{\top},\, \mathbf{P}\big), \tag{10}$$

$$\mathrm{Acc}_{\mathrm{sym}}@\epsilon = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}\Big[\, d_{\mathrm{CD}}\big(\mathbf{P}R_k\hat{R}_k^{\top},\, \mathbf{P}\big) < \epsilon_\theta(\mathbf{P})\,\Big], \tag{11}$$

with $d_{\mathrm{CD}}(X,Y) = d_X^{(2)}(Y) + d_Y^{(2)}(X)$ (Eq. (6)) and $\mathbf{1}[\cdot]$ the indicator. $\mathrm{Acc}_{\mathrm{sym}}@\epsilon$ extends the up-axis accuracy of Scarvelis et al. [39] from a single direction in S^2 to all three rotational degrees of freedom. We adapt the threshold to object geometry via the expected Chamfer error under rotations of angle θ,

$$\epsilon_\theta(\mathbf{P}) = \mathbb{E}_{\mathbf{u}\sim\mathcal{U}(S^2)}\big[d_{\mathrm{CD}}\big(\mathbf{P}R(\mathbf{u},\theta),\, \mathbf{P}\big)\big], \tag{12}$$

where $R(\mathbf{u},\theta)$ is the rotation by θ about the unit axis $\mathbf{u}$. We estimate it by Monte Carlo, $\epsilon_\theta(\mathbf{P}) \approx \frac{1}{J}\sum_{j=1}^{J} d_{\mathrm{CD}}(\mathbf{P}R(\mathbf{u}_j,\theta),\mathbf{P})$, over $J = 32$ axes $\mathbf{u}_j$ drawn uniformly on S^2 once per object; we use $\theta{=}10^\circ$. Each category's value averages over its test objects ($K{=}11$ rotations each; Appendix B.4).

Orientation calibration. CHOIR [21] learns its own canonical frame, which can differ from the dataset frame by a fixed rotation in each class. Without calibration, Eqs. (10)–(11) therefore include this offset in the measured error (Table 1, uncalibrated row). We fit one rotation R_{calib} per class and report the calibrated results as CHOIR†; Appendix B.4 describes the fitting procedure. DORE and Scarvelis et al. [39] train against the dataset's canonical frame and require no such calibration.

Implementation details. Our model is trained with the Adam optimizer at a learning rate of 2×10^{-3}, a batch size of 2×64 (two GPUs), and 1500 epochs, with $\sigma = 0.2$. Every model, ours and the baselines, is evaluated at the final checkpoint of its fixed training schedule, with no checkpoint selection. Before training, we fix each class's band limit using the L_{sym} rule (Sec. 4.6) and the training split's symmetry labels alone. Thus, test data does not influence band selection. Appendices D.3 and B.3 describe uniform bands $L_{\max} \in \{2,\ldots,6\}$ and the training setup.

5.2 Benchmark results

Quantitative comparison. With a single forward pass, DORE outperforms Scarvelis et al. [39] without test-time augmentation in every category on both metrics (Table 1; six-category mean $\mathrm{Acc}_{\mathrm{sym}}@\epsilon$ 82.8 vs. 73.7, geometric-mean Chamfer distance 3.76 vs. 6.80). Section 5.4 compares both methods with additional test-time computation.

Table 1: **Single-pass Chamfer distance and symmetry-aware accuracy on ShapeNet [5].** CHOIR† uses per-class frame calibration (Eq. (51)). DORE uses one forward pass at L_{sym} (Eq. (9)): $L = 2/2/2/4/6/2$ for airplane/car/chair/table/lamp/cellphone (plane/phone in headers). Additional test-time computation is compared in Table 3. Bold: best; underline: second best.

Method	Mean Chamfer distance ($\times 10^{-3}$) $\downarrow$						$\text{Acc}_{\text{sym}}@\epsilon$ (%) $\uparrow$					
	plane	car	chair	table	lamp	phone	plane	car	chair	table	lamp	phone
CHOIR [21]	198.1	103.0	92.3	95.5	238.7	183.2	0.0	0.0	0.0	0.0	0.0	0.0
CHOIR† [21]	16.4	4.6	5.8	44.3	69.5	140.1	54.2	86.6	80.8	13.1	16.0	0.5
Scarvelis et al. [39]	<u>4.6</u>	<u>3.4</u>	<u>4.3</u>	<u>5.5</u>	<u>29.5</u>	<u>9.1</u>	<u>81.9</u>	<u>90.1</u>	<u>85.5</u>	<u>80.5</u>	<u>36.9</u>	<u>67.6</u>
DORE (ours)	**1.5**	**1.2**	**2.1**	**5.1**	**28.2**	**5.0**	**92.6**	**98.3**	**93.8**	**84.2**	**45.6**	**82.4**

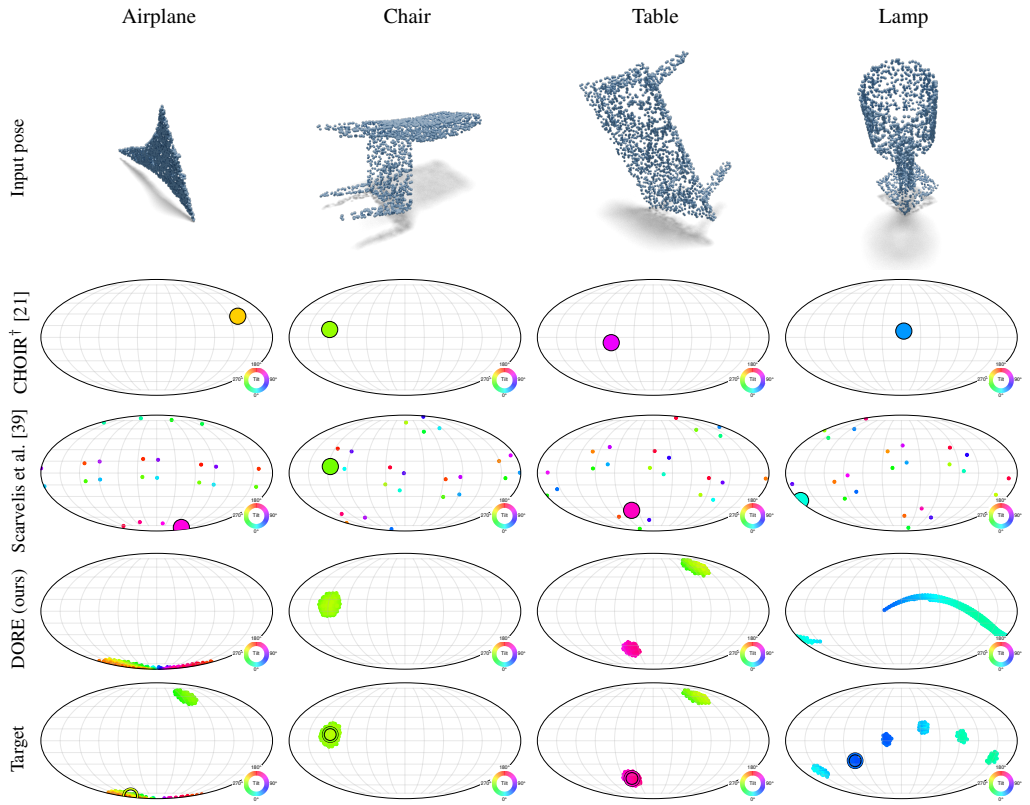

Figure 3: **Qualitative ShapeNet results [5].** Rotations use a Mollweide projection of S^2, with hue indicating in-plane tilt [30]. Filled circles show single predictions; for Scarvelis et al. [39], the flipper selects one of 24 quotient orientations (dots; 50 votes, seed 0). DORE and its target use L_{sym}; the outlined circle marks the applied rotation. On the nearly $SO(2)$-symmetric lamp, every point on the arc is valid, including the prediction of Scarvelis et al. [39]. DORE captures multiple target modes, although the second airplane mode (81% of the peak) falls below the display cutoff.

CHOIR's raw Chamfer (92.3–238.7) partly reflects the offset of its learned frame, which the per-class calibration (CHOIR†) removes. Calibrated accuracy is 86.6% on car, 80.8% on chair, and 54.2% on airplane, but remains low on symmetric categories (lamp 16.0%, table 13.1%, cellphone 0.5%), despite CHOIR's $SO(3)$-equivariance. Scarvelis et al. [39] absorbs ambiguity only within its octahedral quotient; lamp, rich in high-order C_n and continuous symmetries, is its weakest category without TTA (36.9 vs. 45.6). Together with the matched-head comparison (Appendix D.2), these results show that equivariance alone is insufficient: symmetric categories also require a representation of the multiple valid orientations in each orbit. The band rule, the effect of data volume, and parametric density heads are analyzed in Appendices D.3, D.4, and D.2.

Table 2: **Ablations on cellphone, table, and lamp** at $L_{\max} = 2/4/6$, their respective L_{sym} bands. The base runs match Table 1; the refinement ablation reuses those checkpoints with Newton–Raphson disabled. Accuracy is in %, and CD is scaled by 10^{-3}; bold marks the best value in each column. Appendix C.2 separately examines heat-kernel smoothing.

	cellphone ($L_{\max}$=2)		table ($L_{\max}$=4)		lamp ($L_{\max}$=6)	
Method	$\text{Acc}_{\text{sym}}@\epsilon \uparrow$	CD $\downarrow$	$\text{Acc}_{\text{sym}}@\epsilon \uparrow$	CD $\downarrow$	$\text{Acc}_{\text{sym}}@\epsilon \uparrow$	CD $\downarrow$
DORE (ours)	**82.4**	**5.0**	**84.2**	**5.1**	45.6	28.2
w/o higher-degree coefficients ($L_{\max}$=1)	2.2	120.8	15.3	37.4	12.9	53.9
w/o Laplace–Beltrami weighting	80.4	5.5	70.5	7.4	44.2	30.5
w/o equivariance loss (α=0)	80.7	5.4	81.1	5.5	**45.7**	**24.9**
w/o Newton–Raphson refinement (grid argmax)	62.7	7.2	62.6	7.4	39.5	29.1

Qualitative comparison. Figure 3 visualizes each method's predictions; Appendix F.6 shows randomly drawn test samples. DORE follows the target orbit: a single mode on chair, two modes for table's C_2 symmetry, and an arc through the six target modes of lamp's C_6 label. CHOIR's single-orientation prediction suffices on asymmetric chair, where the target collapses to a point, but on the symmetric categories it covers at most one orbit element. Scarvelis et al. [39] select one of the 24 orientations of their octahedral quotient. The lamp's broad arc reflects its near-continuous rotational symmetry, which the finite C_6 label represents with six discrete orientations (Appendix D.7).

5.3 Ablation study

Table 2 evaluates each component at the L_{sym} bands of cellphone, table, and lamp. These categories cover the smallest and largest symmetric training sets, together with the category richest in high-order C_n and continuous symmetries. Truncating the spectral head to $L_{\max} = 1$, below the first expressible degree of every C_n and D_n (Proposition 1), yields the largest drop ($84.2 \to 15.3$ Acc_{sym}, $5.1 \to 37.4$ CD on table), confirming that higher-degree irreps are needed for orbit ambiguity.

Removing the Laplace–Beltrami weighting costs 13.7 pp on table ($84.2 \to 70.5$) and less on cellphone and lamp (-2.0 and -1.4 pp), consistent with balancing each degree's contribution to the loss. The representation and target remain unchanged in this ablation; only the relative contribution of each degree to the loss changes. The table gap therefore shows that retaining higher-degree coefficients alone is insufficient: their weighting during training also matters for recovering the orbit structure.

Removing the equivariance term lowers accuracy on cellphone and table ($82.4 \to 80.7$, $84.2 \to 81.1$) and leaves lamp nearly unchanged ($45.6 \to 45.7$). These smaller changes are consistent with its role as a consistency regularizer. The backbone and spectral head already enforce the required transformation law, so the loss does not supply the model's equivariance. Instead, it encourages predictions from two rotated views of the same shape to agree after transforming their coefficients into a common frame (Appendix B.3). This complements orbit supervision by penalizing deviations introduced by numerical operations in the implementation.

Disabling Newton–Raphson refinement reads the argmax directly from the 6,912-point Hopf grid while keeping the trained predictor fixed. Accuracy falls by 19.7 pp on cellphone ($82.4 \to 62.7$), 21.6 pp on table ($84.2 \to 62.6$), and 6.1 pp on lamp ($45.6 \to 39.5$). Chamfer distance also increases in all three categories, from 5.0 to 7.2, 5.1 to 7.4, and 28.2 to 29.1, respectively. With predicted densities fixed, these drops reflect localization error from restricting predictions to the coarse grid. The coarse grid initializes the search, and Newton–Raphson refinement improves localization to meet the geometry-dependent 10° criterion without requiring an additional network evaluation.

5.4 Test-time compute scaling for enhanced performance

Test-time compute scaling offers an effective way to boost performance at the cost of increased latency. Scarvelis et al. [39] improve accuracy through rotation voting in testing; the orienter maps predictions from 50 randomly rotated inputs and selects the one minimizing summed octahedral-invariant loss over all votes, while the flipper also selects the most frequent prediction from 50 random octahedral flips. This raises mean accuracy from 73.7 to 86.0 at 353.2 ms per shape (Table 3). DORE also

Table 3: **Performance and latency with additional test-time computation.** Base models and protocol match Table 1. Best-of-N selects one of up to four candidate modes (Sec. 5.4). Coefficient averaging uses eight inputs (Appendix D.9); other DORE rows use one input. TTA uses 50 votes per network. Latency uses one RTX 3090 (Appendix B.4). Bold: best; underline: second best.

Method	Mean Chamfer distance ($\times 10^{-3}$) ↓						$\mathrm{Acc}_{\mathrm{sym}}@\epsilon$ (%) ↑						Latency (ms) ↓
	plane	car	chair	table	lamp	phone	plane	car	chair	table	lamp	phone	
Scarvelis et al. [39]	4.6	3.4	4.3	5.5	29.5	9.1	81.9	90.1	85.5	80.5	36.9	67.6	<u>7.1</u>
Scarvelis et al. [39] + TTA	**0.8**	1.6	<u>2.0</u>	**3.6**	**18.6**	4.5	**96.4**	**99.5**	<u>93.9</u>	**87.7**	**54.8**	83.6	353.2
DORE	1.5	<u>1.2</u>	2.1	5.1	28.2	5.0	92.6	98.3	93.8	84.2	45.6	82.4	**3.0**
DORE + BoN	1.5	<u>1.2</u>	2.1	4.9	27.5	<u>4.4</u>	93.1	98.5	93.8	85.1	50.6	<u>86.5</u>	7.8
DORE + BoN + CA	<u>1.2</u>	**1.1**	**1.7**	<u>4.5</u>	<u>26.5</u>	**4.1**	<u>94.9</u>	<u>99.1</u>	**95.2**	<u>86.1</u>	<u>51.3</u>	**86.6**	28.5

significantly improves mean accuracy from 82.8 to 85.5 at 28.5 ms by using two complementary test-time scaling methods: best-of-N selection and coefficient averaging, described below.

Best-of-N selection (BoN): Following the verifier-based selection strategy of Cobbe et al. [6], we score candidate orientations at predicted density peaks and select the highest-scoring one. To filter out low peaks, including potential ringing artifacts, we retain up to four grid maxima at or above a fixed relative-height threshold (we use 0.7 of the highest peak), then apply Newton–Raphson refinement to each (Sec. 4.3). The DGCNN verifier [45] assigns a scalar canonical-alignment score to the point cloud canonicalized by each candidate. It is trained on the training split to assign higher scores to valid orientations (Appendix D.9). The verifier is used only when at least two peaks qualify, which occurs for 0.2%–100% of test inputs depending on the category. Canonicalizing the input with each candidate makes verification rotation-invariant, up to finite-grid errors (Appendix B.2).

Coefficient averaging (CA): we form 8 inputs of $1{,}024$ points from the same $2{,}000$-point cloud using subsampling and small spatial perturbations (Appendix D.9). We average their predicted coefficients and run one peak search; this averages the predicted densities f by linearity, aiming to reduce prediction variance across input samples. Rotation augmentation is redundant under exact $SO(3)$ equivariance: after undoing each input rotation, the predicted density is unchanged.

Best-of-N raises the six-category mean accuracy from 82.8 to 84.6, mostly on lamp ($+5.0$ pp), cellphone ($+4.1$ pp) and table ($+0.9$ pp). Combined, they outperform 50-vote TTA in accuracy on chair and cellphone at about $12\times$ lower latency. At matched latency (four votes per network), Scarvelis et al. [39] reach 81.0 versus our 85.5 mean accuracy (Appendix F.1).

6 Conclusion

We have introduced DORE, an equivariant framework that represents 3D object orientation as a continuous distribution over $SO(3)$. This formulation captures the full orbit of valid orientations induced by symmetry, allowing multiple correct poses to coexist in a single prediction. The equivariant backbone and spectral head predict a truncated Wigner-D expansion that transforms consistently with the input. On ShapeNet, under per-category training and single-pass inference, DORE outperforms the previous state of the art in every category. Best-of-N selection and coefficient averaging achieve mean accuracy within 0.5 percentage points of the 50-vote result of Scarvelis et al. [39], at roughly $12\times$ lower latency, without using rotation augmentation. These results support a distributional view that preserves geometric ambiguity while enabling task-specific pose selection.

7 Limitations

The output has unit mass but may contain negative side lobes; heat-kernel smoothing reduces them at the cost of localization accuracy (Appendix C.2). We predict orientations from density peaks; interpreting the output as a probability density additionally requires nonnegativity. Our supervision covers symmetry ambiguity in complete shapes, but not partial observations, appearance ambiguity, or unequally weighted modes. Although the spectral representation can express these cases, extending supervision to them remains untested. The labeler is limited to C_n and D_n with $n \leq 6$, excluding polyhedral groups, and the threshold τ and smoothing scale σ are chosen heuristically. Finally, experiments use per-class models; a single model across all 55 categories remains untested.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korean government (MSIT) (RS-2022-II220290: Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense (50%), RS-2022-II220113: Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework (45%), RS-2019-II191906: AI Graduate School Program at POSTECH (5%)).

References

[1] M. A. Alcorn. AQuaMaM: An autoregressive, quaternion manifold model for rapidly estimating complex SO(3) distributions. *arXiv preprint arXiv:2301.08838*, 2023.

[2] A. Avetisyan, M. Dahnert, A. Dai, M. Savva, A. X. Chang, and M. Nießner. Scan2CAD: Learning CAD model alignment in RGB-D scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. S. Ryoo, G. Salazar, P. R. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. H. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[4] R. Cai, B. Hariharan, N. Snavely, and H. Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[6] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[7] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling. Spherical CNNs. In *International Conference on Learning Representations (ICLR)*, 2018.

[8] C. Deng, O. Litany, Y. Duan, A. Poulenard, A. Tagliasacchi, and L. J. Guibas. Vector neurons: A general framework for SO(3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[9] H. Deng, M. Bui, N. Navab, L. Guibas, S. Ilic, and T. Birdal. Deep Bingham networks: Dealing with uncertainty and ambiguity in pose estimation. *International Journal of Computer Vision*, 2022.

[10] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] F. B. Fuchs, D. E. Worrall, V. Fischer, and M. Welling. SE(3)-Transformers: 3D roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[12] L. Gao, L.-X. Zhang, H.-Y. Meng, Y.-H. Ren, Y.-K. Lai, and L. Kobbelt. PRS-Net: Planar reflective symmetry detection net for 3D models. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[13] M. Geiger and T. Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

[14] I. Gilitschenski, R. Sahoo, W. Schwarting, A. Amini, S. Karaman, and D. Rus. Deep orientation uncertainty learning based on a Bingham loss. In *International Conference on Learning Representations (ICLR)*, 2020.

[15] M. Golubitsky, I. Stewart, and D. G. Schaeffer. *Singularities and Groups in Bifurcation Theory, Volume II*. Applied Mathematical Sciences. Springer, 1988.

[16] Á. González. Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices. *Mathematical Geosciences*, 2010.

[17] C. Gümeli, A. Dai, and M. Nießner. ROCA: Robust CAD model retrieval and alignment from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[18] T. Höfer, B. Kiefer, M. Messmer, and A. Zell. HyperPosePDF - hypernetworks predicting the probability distribution on SO(3). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[19] E. Ihrig and M. Golubitsky. Pattern selection with O(3) symmetry. *Physica D: Nonlinear Phenomena*, 1984.

[20] O. Katzir, D. Lischinski, and D. Cohen-Or. Shape-pose disentanglement using SE(3)-equivariant vector neurons. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[21] S. Kim, C. Park, Y. Jeong, J. Park, and M. Cho. Stable and consistent prediction of 3D characteristic orientation via invariant residual learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[22] D. M. Klee, O. Biza, R. Platt, and R. Walters. Image to sphere: Learning equivariant features for efficient pose prediction. In *International Conference on Learning Representations (ICLR)*, 2023.

[23] W. Kuo, A. Angelova, T.-Y. Lin, and A. Dai. Mask2CAD: 3D shape prediction by learning to segment and retrieve. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[24] J. Lee and M. Cho. 3D equivariant pose regression via direct Wigner-D harmonics prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[25] Y. Liu, H. Liu, Y. Yin, Y. Wang, B. Chen, and H. Wang. Delving into discrete normalizing flows on SO(3) manifold for probabilistic rotation modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[26] T. Luo, C. Rockwell, H. Lee, and J. Johnson. Scalable 3D captioning with pretrained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[27] A. Martinet, C. Soler, N. Holzschuch, and F. X. Sillion. Accurate detection of symmetries in 3D shapes. *ACM Transactions on Graphics*, 2006.

[28] N. J. Mitra, L. J. Guibas, and M. Pauly. Partial and approximate symmetry detection for 3D geometry. *ACM Transactions on Graphics*, 2006.

[29] D. Mohlin, J. Sullivan, and G. Bianchi. Probabilistic orientation estimation with matrix Fisher distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[30] K. A. Murphy, C. Esteves, V. Jampani, S. Ramalingam, and A. Makadia. Implicit-PDF: Non-parametric representation of probability distributions on the rotation manifold. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[31] X. Pang, F. Li, N. Ding, and X. Zhong. Upright-Net: Learning upright orientation for 3D point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[32] C. Park, S. Kim, J. Park, and M. Cho. Learning SO(3)-invariant semantic correspondence via local shape transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[33] M. Pauly, N. J. Mitra, J. Wallner, H. Pottmann, and L. J. Guibas. Discovering structural regularity in 3D geometry. *ACM Transactions on Graphics*, 2008.

[34] O. Poursaeed, T. Jiang, H. Qiao, N. Xu, and V. G. Kim. Self-supervised learning of point clouds via orientation estimation. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020.

[35] S. Prokudin, P. Gehler, and S. Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[36] O. Puny, M. Atzmon, E. J. Smith, I. Misra, A. Grover, H. Ben-Hamu, and Y. Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations (ICLR)*, 2022.

[37] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[38] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[39] C. Scarvelis, D. Benhaim, and P. Zhang. Symmetry-robust 3D orientation estimation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[40] A. Seo, B. Kim, S. Kwak, and M. Cho. Reflection and rotation symmetry detection via equivariant learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[41] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

[42] M. A. Uy, V. G. Kim, M. Sung, N. Aigerman, S. Chaudhuri, and L. J. Guibas. Joint learning of 3D shape retrieval and deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[43] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese. DenseFusion: 6D object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[44] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas. Normalized object coordinate space for category-level 6D object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[45] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, 2019.

[46] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[47] W. Xian, Z. Li, M. Fisher, J. Eisenmann, E. Shechtman, and N. Snavely. UprightNet: Geometry-aware camera orientation estimation from single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[48] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018.

[49] A. Yershova, S. Jain, S. M. LaValle, and J. C. Mitchell. Generating uniform incremental grids on SO(3) using the Hopf fibration. *The International Journal of Robotics Research*, 2010.

[50] Y. Yin, Y. Wang, H. Wang, and B. Chen. A Laplace-inspired distribution on SO(3) for probabilistic rotation estimation. In *International Conference on Learning Representations (ICLR)*, 2023.

Appendix

Appendix contents

A. Theory and proofs **16**

A.1 Harmonic analysis on $SO(3)$ via the Peter–Weyl theorem 16

A.2 Equivariance proof 18

A.3 Equivariance of the spectral head 19

A.4 Equivalence of one-sided and symmetric Chamfer score in $S(G)$ 20

B. Inference and implementation **21**

B.1 Inference: one-shot mode search and Newton–Raphson refinement 21

B.2 Using the full density at test time 23

B.3 Implementation details: model and training 23

B.4 Implementation details: data and evaluation protocol 25

C. Analysis of the objective and the target **27**

C.1 The training objective as a Sobolev metric 27

C.2 Properness and the smoothing trade-off. 29

D. Additional experiments and analyses **30**

D.1 Metric agreement and the success threshold 30

D.2 Distributional heads under a matched setup 31

D.3 Band-limit sweep and the L_{sym} rule 32

D.4 Sample-size sweep 32

D.5 Training convergence across bands 33

D.6 Run-to-run variance 33

D.7 Predictions broader than the orbit target 34

D.8 Motivation for best-of-N 35

D.9 Settings for using the full density at test time. 36

E. Symmetry labeling **38**

E.1 Symmetry group labeling algorithm 38

E.2 Symmetry group labeling validation 38

E.3 Symmetry labeling statistics 40

F. Further evaluation and visualization **42**

F.1 Baseline at matched latency 42

F.2 Up-axis metrics. 42

F.3 Zero-shot transfer to ModelNet-40 43

F.4 Band-limit gallery 44

F.5 Band-limit gallery of the polyhedral groups 48

F.6 Gallery of random test samples 49

A Theory and proofs

A.1 Harmonic analysis on $SO(3)$ via the Peter–Weyl theorem

We first review why Wigner-D coefficients can represent orientation distributions, then relate the retained degrees to the symmetry groups they can express.

Peter–Weyl theorem. Let G be a compact real Lie group. By the Peter–Weyl theorem, $L^2(G)$ is a Hilbert space direct sum indexed by the equivalence classes of irreducible unitary representations (ρ, V_ρ) of G:

$$L^2(G) \;\cong\; \widehat{\bigoplus}_{\rho\in\hat{G}} V_\rho \otimes V_\rho^*, \tag{13}$$

where $\hat{G}$ denotes the unitary dual of G. By the Schur orthogonality relations, the matrix coefficients $\rho_{ij}(g) = \langle e_i, \rho(g)e_j\rangle$ are mutually orthogonal with $\|\rho_{ij}\|_{L^2}^2 = 1/\dim\rho$; the rescaled collection $\{\sqrt{\dim\rho}\;\rho_{ij}\}$ is a complete orthonormal Fourier basis for $L^2(G)$.

Application to $SO(3)$. $SO(3)$ is a compact real Lie group, so the Peter–Weyl theorem applies. Its irreducible unitary representations are indexed by non-negative integers $l \in \{0, 1, 2, \dots\}$, with each irrep V^l having dimension $2l+1$. The corresponding decomposition is

$$L^2(SO(3)) \;\cong\; \widehat{\bigoplus}_{l=0}^{\infty} V^l \otimes (V^l)^* \;\cong\; \widehat{\bigoplus}_{l=0}^{\infty} \mathrm{End}(V^l), \tag{14}$$

where $\mathrm{End}(V^l) \cong \mathbb{R}^{(2l+1)\times(2l+1)}$ is the space of linear maps from V^l to itself.

Wigner-D matrices as basis elements. The real spherical harmonics $\{Y_m^l\}_{m=-l}^{l}$ form an orthonormal basis for V^l. In this basis, the group action of $R \in SO(3)$ on V^l is represented by the degree-l Wigner-D matrix $D^l(R) \in \mathbb{R}^{(2l+1)\times(2l+1)}$, whose (i,j)-entry is the matrix coefficient $D^l_{ij}(R) = \langle Y_i^l, \rho^l(R) Y_j^l\rangle$. These matrix coefficients satisfy the Schur orthogonality relations,

$$\int_{SO(3)} D^l_{ij}(R)\, D^{l'}_{i'j'}(R)\,\mathrm{d}R = \frac{1}{2l+1}\,\delta_{ll'}\,\delta_{ii'}\,\delta_{jj'}, \tag{15}$$

so $\{\sqrt{2l+1}\,D^l_{ij}\}$ over all l is a complete orthonormal basis for $L^2(SO(3))$.

Fourier expansion. Every square-integrable function $f \in L^2(SO(3))$ admits the expansion

$$f(R) = \sum_{l=0}^{\infty} (2l+1)\;\mathrm{tr}\big(H^l\, D^l(R)^\top\big), \quad H^l \in \mathbb{R}^{(2l+1)\times(2l+1)}, \tag{16}$$

where $H^l = \int_{SO(3)} f(R)\, D^l(R)\,\mathrm{d}R$ is the degree-l Fourier coefficient matrix and $\mathrm{d}R$ denotes the normalized Haar measure (total mass 1) on $SO(3)$. Truncating at degree $L_{\max}$ gives the band-limited approximation used in our method, with a coefficient vector of dimension $\sum_{l=0}^{L_{\max}}(2l+1)^2$.

Plancherel identity. For a compact group G, the Fourier transform $f \mapsto \{\hat{f}(\rho)\}_{\rho\in\hat{G}}$ with $\hat{f}(\rho) = \int_G f(g)\,\rho(g)\,\mathrm{d}g$, with $\mathrm{d}g$ the normalized Haar measure, satisfies

$$\|f\|^2_{L^2(G)} = \sum_{\rho\in\hat{G}} (\dim\rho)\,\|\hat{f}(\rho)\|_F^2. \tag{17}$$

Specializing to $SO(3)$, where $\dim\rho^l = 2l+1$ and $\hat{f}(\rho^l) = H^l$,

$$\|f\|^2_{L^2(SO(3))} = \sum_{l=0}^{\infty} (2l+1)\,\|H^l\|_F^2, \tag{18}$$

which follows directly from the orthogonality relations of Eq. (15). Up to the per-degree factors $2l+1$, the Fourier transform is thus an isometry between $L^2(SO(3))$ and the space of coefficient blocks, enabling coefficient-based distances between band-limited densities (Appendix C.1).

Which symmetry groups a band can express. Our labels are heat-damped group averages. The multiplicity of the trivial representation at each degree determines which groups a truncated label can distinguish and the first degree at which they become distinguishable. These multiplicities also identify groups directly from stored coefficients, allowing us to analyze the labels (Appendix E.3) and reconstruct the Bingham mixture's orbit components without additional information (Appendix D.2).

Lemma 1 (Group identification)**.** *Let $G \leq SO(3)$ be a closed subgroup and let $F_G^l = \frac{1}{|G|}\sum_{g\in G} D^l(g)$, with the sum replaced by the Haar average when G is continuous. Then:*

(i) F_G^l is the orthogonal projector onto the subspace of G-invariant vectors of V^l; in particular $\|F_G^l\|_F^2 = \operatorname{tr} F_G^l = m_l(G) \in \mathbb{Z}_{\geq 0}$, the multiplicity of the trivial representation of G in V^l.

(ii) $m_l(C_n) = 1 + 2\lfloor l/n \rfloor$ and $m_l(D_n) = \lfloor l/n \rfloor + \mathbf{1}[l \text{ even}]$.

(iii) The full sequence $(m_l(G))_{l\geq 0}$ depends only on the conjugacy class of G and determines it completely: distinct conjugacy classes of closed subgroups have distinct multiplicity sequences. Truncation to $l \leq L$ is what loses information: the tuple $\mathbf{m}_L(G) = (m_0, \ldots, m_L)$ still separates every pair whose separating degree lies within the band (for C_n and D_n, the degree separating them from $SO(2)$ and $O(2)$ is n), and, for every band $L \geq 2$, merges the remainder into the band-limit fibers $\{C_n : n > L\} \cup \{SO(2)\}$, $\{D_n : n > L\} \cup \{O(2)\}$, and a polyhedral fiber that thins with the band ($\{T, O, I, SO(3)\}$ for $L < 3$, $\{O, I, SO(3)\}$ at $L = 3$, and $\{I, SO(3)\}$ for $4 \leq L < 6$), following the first-invariant degrees 3, 4, 6 of T, O, I.

Part (i) follows from the standard projector onto the invariant subspace. Part (ii), together with the first-invariant degrees 3, 4, and 6 of T, O, and I, follows from Theorem 3.2 of Ihrig and Golubitsky [19]: with the axis along z, a degree-l vector is C_n-invariant iff its azimuthal index satisfies $n \mid m$, and the flip of D_n keeps one combination of each pair $\pm m$ and the $m = 0$ line only for even l. Part (iii) follows from (ii) and the classification of the closed subgroups of $SO(3)$ [15]: m_1 and m_2 split the classes, the first degree at which m_l departs from the $SO(2)$ or $O(2)$ pattern recovers n, and below that degree the tuples of C_n and D_n coincide with those of $SO(2)$ and $O(2)$, as the polyhedral groups coincide with $SO(3)$ below their first invariants. The proposition below turns the count into the density-level claim that Sec. 4.6 uses; the same condition $n \mid m$ underlies the spherical-harmonic symmetry-detection criteria of Martinet et al. [27].

Proposition 1 (First expressible degree)**.** *Fix the common axis of C_n or D_n ($n \geq 2$), and let f_L^G denote the band-limited orbit density of G with any positive per-degree damping. Then: (i) for every band $L < n$, $f_L^{C_n} = f_L^{SO(2)}$ and $f_L^{D_n} = f_L^{O(2)}$ identically on $SO(3)$, so no band below n distinguishes either discrete orbit from its corresponding continuous orbit; and (ii) at $L = n$ the densities separate, and along the axial fiber $\{R_z(\phi)\}$ the maxima of $f_n^{C_n}$ are exactly the n orbit elements $\phi = 2\pi j/n$.*

Proof. Both parts follow from the condition $n \mid m$ of Lemma 1(ii). (i) For $l < n$ the only admissible index is $m = 0$, so the degree-l projectors of C_n and $SO(2)$ coincide (with the flip, those of D_n and $O(2)$); damping scales each block by the same positive factor and the density is linear in the blocks. (ii) At $l = n$ the pair $m = \pm n$ enters (for D_n, its flip-invariant combination), so the densities separate, and on the axial fiber the density becomes $c_0 + a\cos(n\phi)$ with $a > 0$, the contribution of the $m = \pm n$ entries, so it peaks exactly at $\phi = 2\pi j/n$. □

A stored label satisfies $H^{*l} = \kappa_l F_G^l$, so its multiplicities are recovered as $m_l = \kappa_l^{-2}\,\|H^{*l}\|_F^2$ without accessing the shape or repeating symmetry detection. We use this reconstruction to supply label-derived components to the Bingham baseline (Appendix D.2). At $L = 5$, we reconstruct C_6 and D_6 labels as C_{12} and D_{12} because their retained coefficients coincide. Regenerating the coefficients checks each reconstruction, although at this band the check cannot resolve the flip-axis phase of a D_6 label, a limitation affecting 4 of the 8,436 table shapes in this reconstruction.

The fibers in (iii) are Proposition 1(i) at the vocabulary boundary: a C_6 label first separates from a continuous one at $l = 6$, the same $l \geq n$ threshold that governs peak-search resolution and motivates the L_{sym} rule (Sec. 4.6). At $L = 5$, however, these fibers are unambiguous within the labeler's vocabulary: it stops at $n = 6$ (Sec. 4.5), leaving one possible output group in each relevant fiber and none in $\{I, SO(3)\}$. Finally, every C_n and D_n has $m_2 \geq 1$ while T, O, I have $m_2 = 0$, so a polyhedral label could not be silently misread as cyclic or dihedral.

Table 4: **Multiplicity signatures of closed subgroups of** $SO(3)$ (Lemma 1). Stored labels extend to $l = 6$, where every row is distinct. Cyclic and dihedral orders separate from their continuous limits at $l = n$, so $C_{n\geq 7}$ and $D_{n\geq 7}$ share the displayed signatures with $SO(2)$ and $O(2)$, respectively.

	m_0	m_1	m_2	m_3	m_4	m_5	m_6
E	1	3	5	7	9	11	13
C_2	1	1	3	3	5	5	7
C_3	1	1	1	3	3	3	5
C_4	1	1	1	1	3	3	3
C_5	1	1	1	1	1	3	3
C_6	1	1	1	1	1	1	3
$C_{n\geq 7}$, $SO(2)$	1	1	1	1	1	1	1
D_2	1	0	2	1	3	2	4
D_3	1	0	1	1	2	1	3
D_4	1	0	1	0	2	1	2
D_5	1	0	1	0	1	1	2
D_6	1	0	1	0	1	0	2
$D_{n\geq 7}$, $O(2)$	1	0	1	0	1	0	1
T	1	0	0	1	1	0	2
O	1	0	0	0	1	0	1
I	1	0	0	0	0	0	1
$SO(3)$	1	0	0	0	0	0	0

A.2 Equivariance proof

Our network organizes the predicted Fourier coefficients in a block-diagonal matrix

$$h(\mathbf{P}) = \bigoplus_{l=0}^{L_{\max}} \hat{H}^l, \tag{19}$$

whose blocks define the truncated Fourier expansion of the predicted orientation density,

$$h(\mathbf{P})(S) = \sum_{l=0}^{L_{\max}} (2l+1) \sum_{i,j=-l}^{l} \hat{h}^l_{ij} D^l_{ij}(S). \tag{20}$$

Transport of the density with the input. A valid orientation S is a rotation that undoes the input: $\mathbf{P}S^\top$ is canonical, as in the metrics of Sec. 5. Rotating the input to $\mathbf{P}R^\top$ (by $R^\top$ in the convention of Sec. 3) gives the undoing rotation $SR^\top$, since $(\mathbf{P}R^\top)(SR^\top)^\top = \mathbf{P}S^\top$. Thus, the predicted density f transforms as $(\rho(R)f)(S) = f(SR)$, or $h(\mathbf{P}R^\top) = \rho(R)\, h(\mathbf{P})$, with the representation law

$$\rho(R_1R_2) = \rho(R_1)\rho(R_2), \qquad \big(\rho(R_1)\rho(R_2)f\big)(S) = \big(\rho(R_2)f\big)(SR_1) = f(SR_1R_2). \tag{21}$$

This matches the input composition law, $(\mathbf{P}R_2^\top)R_1^\top = \mathbf{P}(R_1R_2)^\top$. Thus, successive input rotations transport the density in the same way as their product. Replacing R by R^{-1} gives the input convention used throughout the method: $h(\mathbf{P}R)(S) = h(\mathbf{P})(SR^{-1})$.

Proposition 2 (Coefficient condition for equivariance)**.** *Let $R \in SO(3)$. The predicted coefficients satisfy $\hat{H}^l(\mathbf{P}R) = \hat{H}^l(\mathbf{P})\, D^l(R)$ for every $l \leq L_{\max}$ if and only if $h(\mathbf{P}R)(S) = h(\mathbf{P})(SR^{-1})$ for all $S \in SO(3)$. The latter states that the predicted density transports with the input: rotating the input by R moves every mode of $h(\mathbf{P})$ from S^* to S^*R, on the same side on which R acts on the input, so repeated rotations keep the density aligned with the point cloud.*

Proof. Using the homomorphism property of Wigner-D matrices, $D^l(gh) = D^l(g)D^l(h)$, we expand

$$h(\mathbf{P})(SR^{-1}) = \sum_{l=0}^{L_{\max}} (2l+1) \sum_{i,j=-l}^{l} \hat{h}^l_{ij} D^l_{ij}(SR^{-1}) \tag{22}$$

$$= \sum_{l=0}^{L_{\max}} (2l+1) \sum_{i,j=-l}^{l} \hat{h}^l_{ij} \left(\sum_{k=-l}^{l} D^l_{ik}(S) D^l_{kj}(R^{-1}) \right) \tag{23}$$

$$= \sum_{l=0}^{L_{\max}} (2l+1) \sum_{i,k=-l}^{l} \left(\sum_{j=-l}^{l} \hat{h}^l_{ij}\, D^l_{kj}(R^{-1}) \right) D^l_{ik}(S). \tag{24}$$

The inner sum contracts the *column* index of both factors, so its matrix form carries a transpose:

$$\sum_{j=-l}^{l} \hat{h}^l_{ij}\, D^l_{kj}(R^{-1}) = \left[\hat{H}^l\, D^l(R^{-1})^\top\right]_{ik} = \left[\hat{H}^l\, D^l(R)\right]_{ik}, \tag{25}$$

where the last step uses the orthogonality of the real Wigner-D matrices, $D^l(R^{-1}) = D^l(R)^\top$. Hence the function $S \mapsto h(\mathbf{P})(SR^{-1})$ is the expansion with coefficient blocks $\hat{H}^l(\mathbf{P})\, D^l(R)$. If $\hat{H}^l(\mathbf{P}R) = \hat{H}^l(\mathbf{P})D^l(R)$ for every l, the functions $h(\mathbf{P}R)$ and $S \mapsto h(\mathbf{P})(SR^{-1})$ have identical coefficients and therefore coincide. Conversely, the matrix entries $\{D^l_{ik}\}$ are linearly independent functions on $SO(3)$ (Peter–Weyl), so equality of the two functions forces equality of the coefficients degree by degree. The condition is therefore both necessary and sufficient. □

A.3 Equivariance of the spectral head

Proposition 3 (Equivariance of the spectral head)**.** *In exact arithmetic, the spectral head satisfies*

$$\hat{H}^l(\mathbf{P}R) = \hat{H}^l(\mathbf{P})\, D^l(R) \quad \forall\, R \in SO(3), \quad l = 0, \ldots, L_{\max}. \tag{26}$$

Proof. We track each degree-l feature block under an input rotation R.

Notation. Let $\{v_c(\mathbf{P})\}_{c=1}^{C}$ denote the C type-1 equivariant vectors produced by the VN backbone after mean pooling on input $\mathbf{P}$. Our input convention is $\mathbf{P}R$ where $\mathbf{P} \in \mathbb{R}^{N\times 3}$ and $R \in SO(3)$, so each point transforms as $p_i \mapsto p_i R$, which in column-vector form is $p_i^\top \mapsto R^\top p_i^\top$. The VN backbone operates on the column-form 3-axis, so under this convention:

$$v_c(\mathbf{P}R) = R^\top v_c(\mathbf{P}) = R^{-1} v_c(\mathbf{P}) \quad \forall\, c. \tag{27}$$

Mean pooling commutes with the linear action of R^{-1} and therefore preserves equivariance.

Spherical harmonic projection. The degree-l SH feature for channel c is $\mathbf{z}^l_c(\mathbf{P}) = Y^l(v_c(\mathbf{P})) \in \mathbb{R}^{2l+1}$, with Y^l evaluated at $v/\|v\|$ and $Y^l(0) = 0$ for $l \geq 1$. By the defining property of real spherical harmonics, $Y^l(Qv) = D^l(Q)\, Y^l(v)$ for all $Q \in SO(3)$. Combining this identity with the backbone transformation in Eq. (27) gives

$$\mathbf{z}^l_c(\mathbf{P}R) = Y^l(R^{-1} v_c(\mathbf{P})) = D^l(R^{-1})\, Y^l(v_c(\mathbf{P})) = D^l(R)^{-1}\, \mathbf{z}^l_c(\mathbf{P}). \tag{28}$$

Stacking over channels, the full degree-l block transforms as

$$\mathbf{z}^l(\mathbf{P}R) = D^l(R)^{-1}\, \mathbf{z}^l(\mathbf{P}), \tag{29}$$

where $D^l(R)^{-1}$ acts on the $(2l+1)$ irrep dimension and leaves the channel index unchanged. Writing the block as a $(2l+1) \times C$ matrix makes this separation explicit: rotations act from the left, while channel mixing acts from the right, so the two commute. The following steps preserve equivariance through this separation and scalar gates that are invariant to rotation.

Equivariant linear layers. Each linear layer mixes channels with $W^l \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}}}$, applying the same channel weights independently to every coordinate of the degree-l representation:

$$[W^l \mathbf{z}^l]_{c'} = \sum_c W^l_{c'c}\, \mathbf{z}^l_c \in \mathbb{R}^{2l+1}. \tag{30}$$

The channel action of W^l commutes with the irrep action of $D^l(R)^{-1}$ on $\mathbb{R}^{(2l+1)\times C}$:

$$\begin{aligned}
[W^l(D^l(R)^{-1}\,\mathbf{z}^l)]_{c'} &= \sum_c W^l_{c'c}\, D^l(R)^{-1}\,\mathbf{z}^l_c \\
&= D^l(R)^{-1} \sum_c W^l_{c'c}\,\mathbf{z}^l_c \\
&= D^l(R)^{-1}\,[W^l\,\mathbf{z}^l]_{c'}.
\end{aligned} \tag{31}$$

Hence if $\mathbf{z}^l(\mathbf{P}R) = D^l(R)^{-1}\,\mathbf{z}^l(\mathbf{P})$, then after the linear layer the same relation holds.

Gate nonlinearity. After the first linear layer, the $l{=}0$ output splits into a self-path, passed through $\tanh$ to the second layer, and a gate-path $\tilde{\mathbf{z}}^0_{\text{gate}}$ with one entry per $l{\geq}1$ channel. The gate scalar for channel c is $s_c = \sigma([\tilde{\mathbf{z}}^0_{\text{gate}}]_c)$, derived solely from the $l{=}0$ channel. Since $D^0(R) = 1$ for all $R \in SO(3)$, the $l{=}0$ channel is rotation-invariant:

$$\tilde{\mathbf{z}}^0_{\text{gate}}(\mathbf{P}R) = D^0(R)^{-1}\,\tilde{\mathbf{z}}^0_{\text{gate}}(\mathbf{P}) = \tilde{\mathbf{z}}^0_{\text{gate}}(\mathbf{P}). \tag{32}$$

Therefore $s_c(\mathbf{P}R) = s_c(\mathbf{P})$, so gating preserves equivariance for every $l \geq 1$:

$$s_c(\mathbf{P}R)\cdot \mathbf{z}^l_c(\mathbf{P}R) = s_c(\mathbf{P})\cdot D^l(R)^{-1}\,\mathbf{z}^l_c(\mathbf{P}) = D^l(R)^{-1}\,[s_c(\mathbf{P})\cdot \mathbf{z}^l_c(\mathbf{P})]. \tag{33}$$

Composition and conclusion. The spectral head is the composition $\text{Linear}_2 \circ \text{Gate} \circ \text{Linear}_1 \circ \text{SH}$. The three preceding paragraphs show that each constituent map ϕ satisfies

$$\phi(\mathbf{z}^l(\mathbf{P}R)) = D^l(R)^{-1}\,\phi(\mathbf{z}^l(\mathbf{P})), \tag{34}$$

so after the composition every output channel, a column vector in $\mathbb{R}^{2l+1}$, transforms by $D^l(R)^{-1}$. The degree-l coefficient block stacks the $2l+1$ output channels as the rows of $\hat{H}^l$, and transposing the column relation turns the left action into a right one: each row obeys $\left(D^l(R)^{-1}\mathbf{z}\right)^\top = \mathbf{z}^\top D^l(R)$ by orthogonality of the real Wigner-D matrices. The final output projection (Sec. 4.2) rescales each degree block using only its Frobenius norm. Because $D^l(R)$ is orthogonal, this norm is rotation-invariant, so the projection also preserves equivariance. Consequently,

$$\hat{H}^l(\mathbf{P}R) = \hat{H}^l(\mathbf{P})\,D^l(R) \quad \forall\, R \in SO(3), \tag{35}$$

which is precisely the coefficient condition of Proposition 2. □

Expressivity on exactly symmetric inputs. Equivariance does not imply that the network can realize every band-limited orbit target. For a point set exactly invariant under C_n ($n \geq 2$), permutation-invariant pooling and rotation equivariance constrain every pooled type-1 vector to the symmetry axis. The head then produces a density invariant under all rotations about that axis, rather than resolving the discrete orbit. For exact D_n symmetry ($n \geq 2$), all pooled vectors vanish; every output block with $l \geq 1$ is then invariant under all rotations and hence zero, so the density is constant. These statements concern exact input symmetries; they do not preclude the multi-modal predictions observed on sampled point clouds.

A.4 Equivalence of one-sided and symmetric Chamfer score in $S(G)$

Lemma 2 (One-sided and symmetric Chamfer)**.** *Let*

$$d_X^{(p)}(Y) = \frac{1}{|X|}\sum_{x\in X} \min_{y\in Y} \|x-y\|_2^p$$

with distance exponent $p > 0$, as in Eq. (6), where $S(G)$ uses $p = 1$ and d_{CD} (Sec. 5) uses $p = 2$. For any nonempty finite set $\Gamma \subseteq SO(3)$ closed under inversion, in particular any finite group, and any nonempty finite cloud $\mathbf{P}$, averaging the one-sided distance over Γ gives

$$\mathbb{E}_{g\sim\text{Unif}(\Gamma)}\left[d^{(p)}_{\mathbf{P}}(g.\mathbf{P})\right] = \mathbb{E}_{g\sim\text{Unif}(\Gamma)}\left[d^{(p)}_{g.\mathbf{P}}(\mathbf{P})\right]. \tag{36}$$

The symmetric sum $d_{\mathbf{P}}^{(p)}(g.\mathbf{P}) + d_{g.\mathbf{P}}^{(p)}(\mathbf{P})$ therefore averages over Γ to twice the mean one-sided distance. Both $S(G)$ and the implemented score of Appendix B.3 thus require only one-sided nearest-neighbor searches.

Proof. Every point of $g.\mathbf{P}$ is uniquely gx for some $x \in \mathbf{P}$, so $|g.\mathbf{P}| = |\mathbf{P}|$. Rotations preserve distance: $\|gx - z\|_2 = \|x - g^{-1}z\|_2$. Reindexing the source points by x and the target points by $y = g^{-1}z$ therefore gives

$$\begin{aligned} d_{g.\mathbf{P}}^{(p)}(\mathbf{P}) &= \frac{1}{|\mathbf{P}|} \sum_{x \in \mathbf{P}} \min_{z \in \mathbf{P}} \|gx - z\|_2^p \\ &= \frac{1}{|\mathbf{P}|} \sum_{x \in \mathbf{P}} \min_{y \in g^{-1}.\mathbf{P}} \|x - y\|_2^p = d_{\mathbf{P}}^{(p)}(g^{-1}.\mathbf{P}). \end{aligned} \tag{37}$$

Thus reversing the one-sided distance is equivalent to replacing g by g^{-1}. Because Γ is closed under inversion, $g \mapsto g^{-1}$ permutes its elements; hence g^{-1} is uniform on Γ whenever g is. Averaging the identity over $g \sim \mathrm{Unif}(\Gamma)$ proves the claim. □

B Inference and implementation

B.1 Inference: one-shot mode search and Newton–Raphson refinement

We reconstruct the density f from the predicted Wigner-D coefficients and seek its highest mode,

$$f(R) = \sum_{l=0}^{L_{\max}} (2l+1)\, \mathrm{tr}\big(\hat{H}^l\, D^l(R)^\top\big), \tag{38}$$

whose local maxima form $\mathcal{M}_f$ (Eq. (4)). For top-1 evaluation, we seek $R^* = \mathrm{argmax}_{R \in SO(3)} f(R)$, which exists by continuity on compact $SO(3)$.

Differential characterization. Wigner-D entries are polynomial in R, so a band-limited f is real-analytic. At critical points, its Hessian is coordinate-independent. A closed connected submanifold $M \subseteq SO(3)$ with constant f is a *nondegenerate critical manifold of maxima* in the Morse–Bott sense if, at every $R \in M$,

$$\nabla_R f = 0, \qquad \ker \mathrm{Hess}_R f = T_R M, \qquad \mathrm{Hess}_R f \prec 0 \text{ on a complement of } T_R M, \tag{39}$$

The isolated mode has $\dim M = 0$. Along a ridge, differentiating the constant f twice gives $v^\top \mathrm{Hess}_R f\, v = 0$ for $v \in T_R M$, since $\nabla_R f = 0$. As $\mathrm{Hess}_R f \preceq 0$, this forces $T_R M \subseteq \ker \mathrm{Hess}_R f$, ruling out strict negative definiteness.

Proposition 4 (Modes of the damped orbit target)**.** *Let $G = C_n$ or D_n, $n \geq 2$, and let f have coefficients $H^l = \kappa_l F_G^l$ for $0 \leq l \leq L_{\max}$, with $\kappa_l > 0$ and F_G^l as in Lemma 1. For C_n, the maximum set is an $SO(2)$ ridge when $1 \leq L_{\max} < n$. For D_n, f is constant when $L_{\max} \leq 1$; when $2 \leq L_{\max} < n$, its maximum set is the embedded $O(2)$, whose two circles are nondegenerate ridges. For either group, every orbit element is an isolated nondegenerate maximum once $L_{\max} \geq n$.*

Proof. Take the principal axis to be z. At $R = I$, let $X_v^{(l)}$ denote the infinitesimal degree-l representation of $v \in \mathfrak{so}(3)$. For symmetric idempotent F and skew X, $\mathrm{tr}(FX) = 0$ and $\mathrm{tr}(FX^2) = -\|XF\|_F^2$. Thus I is critical and

$$v^\top \mathrm{Hess}_I f\, v = -\sum_{l=1}^{L_{\max}} (2l+1)\, \kappa_l \left\| X_v^{(l)} F_G^l \right\|_F^2 \;\leq\; 0, \tag{40}$$

For $v \neq 0$, equality holds iff every retained invariant subspace lies in $\ker X_v^{(l)}$, the one-dimensional v-zonal line for $l \geq 1$. Below degree n, the C_n invariants are exactly the z-zonal lines; degree 1 forces $v \parallel z$, giving $\ker \mathrm{Hess}_I f = \mathbb{R} e_z$. At degree n, the pair $m = \pm n$ enters, so no zonal line contains the invariant subspace and the Hessian is negative definite.

For D_n, only even zonal lines survive below degree n. Hence $L_{\max} \leq 1$ gives a constant; for $2 \leq L_{\max} < n$, degree 2 restricts the kernel to $\mathbb{R}e_z$. At degree n, the even zonal component and the flip-invariant $m = \pm n$ component together rule out any nonzero kernel direction (for odd n, use the zonal component at degree 2). Right invariance, $F_G^l D^l(g) = F_G^l$ for $g \in G$, extends these conclusions along the orbit. Globally, since $D^l(R)$ is orthogonal and F_G^l is an orthogonal projector,

$$f(I) - f(R) = \frac{1}{2} \sum_{l=0}^{L_{\max}} (2l+1)\kappa_l \left\| (D^l(R) - I) F_G^l \right\|_F^2 \geq 0.$$

In the ridge regimes, equality requires $Rz = z$ for C_n (degree 1) or $Rz = \pm z$ for D_n (degree 2). These conditions also fix every retained zonal line pointwise, giving exactly the maximum sets $SO(2)$ and $O(2)$, whose tangent directions are the Hessian kernels. □

These results describe the orbit elements of the ideal target. Away from the orbit, a truncated density can retain additional local maxima (for C_2 at $L_{\max} = 2$ the circle of horizontal flips survives as a ridge of local maxima at 14% of the mode height); such lower maxima do not affect the selected orientation as long as the global grid argmax of Sec. 4.3 lies near the highest mode, since refinement starts from it. When the global maximum is a ridge, the mode search can return any point along that ridge.

Coarse mode search. We evaluate the density on a Hopf grid [49] with $N_{\text{grid}} = 6{,}912$ points ($3{,}456$ distinct rotations, each listed under both quaternion signs). The grid uses the Hopf fibration, which parametrizes $SO(3) \cong S^3/\mathbb{Z}_2$ as a circle bundle over S^2, to provide uniform coverage. The precomputed scoring matrix $W \in \mathbb{R}^{N_{\text{grid}} \times 455}$ requires approximately 12 MB and is built once for the largest band ($L_{\max} = 6$); a smaller band uses its leading columns. A single matrix-vector multiplication $\mathbf{f} = Wc$ evaluates all N_{grid} scores without computing Wigner-D matrices at inference time. The highest-scoring grid point, $R_{\text{coarse}} = \text{argmax}_{R \in \{R_i\}} f(R)$, then initializes refinement.

Newton–Raphson refinement. Starting from R_{coarse}, we apply Newton–Raphson iterations on the $SO(3)$ manifold to converge to a nearby local maximum of f. At each step we parameterize a local perturbation around the current estimate R_t via the exponential map,

$$R(\omega) = R_t \cdot \exp(\hat{\omega}), \quad \omega \in \mathbb{R}^3 \cong \mathfrak{so}(3), \tag{41}$$

and estimate the gradient $\nabla f \in \mathbb{R}^3$ and Hessian $\nabla^2 f \in \mathbb{R}^{3\times 3}$ of f with respect to ω at $\omega = 0$ via central finite differences with step $\eta = 0.01$ along the Lie-algebra basis $\{\hat{e}_k\}_{k=0}^2$:

$$(\nabla f)_k = \frac{f(R_t \exp(\eta \hat{e}_k)) - f(R_t \exp(-\eta \hat{e}_k))}{2\eta}, \tag{42}$$

$$(\nabla^2 f)_{kk} = \frac{f(R_t \exp(\eta \hat{e}_k)) + f(R_t \exp(-\eta \hat{e}_k)) - 2f(R_t)}{\eta^2}. \tag{43}$$

For the off-diagonal entries, letting $\Delta_{kj}^{\pm\pm} = f(R_t \exp(\eta(\pm \hat{e}_k \pm \hat{e}_j)))$,

$$(\nabla^2 f)_{kj} = \frac{\Delta_{kj}^{++} - \Delta_{kj}^{+-} - \Delta_{kj}^{-+} + \Delta_{kj}^{--}}{4\eta^2}. \tag{44}$$

The gradient and Hessian then define the Newton update on the rotation group,

$$\omega^* = -(\nabla^2 f)^{-1} \nabla f, \qquad R_{t+1} = R_t \cdot \exp(\widehat{\omega^*}), \tag{45}$$

clamped to a trust region $\|\omega^*\| \leq 0.1$ rad per iteration. On a ridge, the Hessian is singular along the tangent, so $(\nabla^2 f)^{-1}$ does not exist and numerical steps can diverge along the near-null direction. The trust region bounds these steps. We also replace non-ascending steps with short gradient steps and skip the update if the linear solve fails. These safeguards keep the steps bounded; they are meant to move the estimate toward a ridge without constraining its position along it, not to guarantee convergence. We use $n_{\text{iter}} = 2$ iterations during training-time monitoring for speed and $n_{\text{iter}} = 3$ at evaluation: in our measurements, further steps change the metrics by less than run-to-run noise. We seed monitoring rotations by sample index to isolate model changes from evaluation randomness.

B.2 Using the full density at test time

Representing the prediction as a density retains information beyond a single rotation. This supports two complementary inference procedures: mode selection and coefficient averaging.

Best-of-N selection. The density keeps multiple modes of a near-symmetric shape, so when several modes have similar height, selecting the highest peak may yield an incorrect orientation. Best-of-N takes the local maxima $S_1, \ldots, S_N$ of $h(\mathbf{P})$ whose height is at least a fixed fraction of the highest, and a separate verifier chooses among the canonicalized clouds $\mathbf{P}S_k^\top$, given also the density ratios $h(\mathbf{P})(S_k)/h(\mathbf{P})(S_1)$. Because the candidates are modes of the predicted density, selection is restricted to orientations that the predictor supports. These modes rotate with the input (Appendix A.2), making the resulting canonicalized clouds invariant to input rotation, so the verifier needs no training on rotated inputs. This holds for exact local maxima; since the search finds them on a finite grid, near-tied modes can be selected differently under rotation. The verifier is described in Appendix D.9 and evaluated in Table 3; Fig. 4 shows a cellphone for which the verifier corrects an incorrect top-1 prediction.

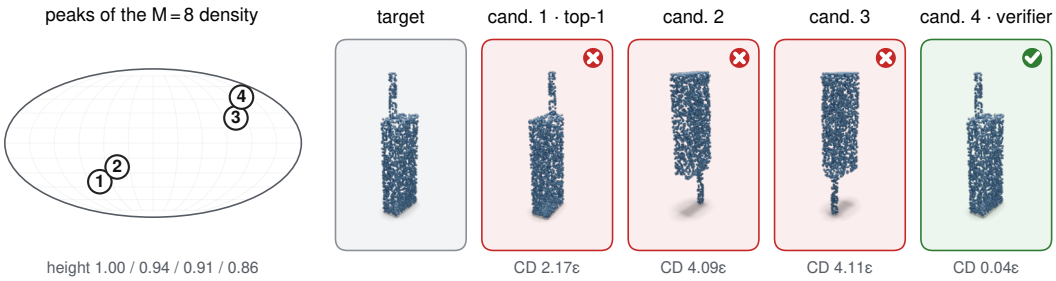

Figure 4: **Best-of-N recovers a cellphone orientation.** Left: the four peaks above 0.7 of the maximum after coefficient averaging, shown with their heights using the projection of Fig. 3. Right: the cloud aligned by each candidate, with Chamfer distance in units of ϵ. Candidate 1 flips front and back, while candidates 2 and 3 invert the up axis. The verifier selects the correct candidate 4.

Coefficient averaging. Coefficient averaging combines multiple predicted densities by averaging their coefficient vectors. This is possible because the density depends linearly on its coefficients. This combines the full distributions instead of reducing each to a single rotation. In contrast, an arithmetic average of rotation matrices generally leaves $SO(3)$, and a mixture of matrix-Fisher distributions is generally not itself a single matrix-Fisher distribution. Coefficient averaging can combine predictions from subsamples, augmentations, or different models; here, we study subsamples of one input. With more points available than one pass uses, we average M subsample predictions ($M = 8$ by default) before one peak search to improve accuracy without retraining (Appendix D.9).

B.3 Implementation details: model and training

Architecture. The VN backbone used by DORE and CHOIR [21] processes the input point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$ and produces point-wise equivariant features of dimension $C \times 3$ per point. After global mean pooling, the spectral head lifts these features via spherical harmonic projection to a multi-degree irrep space and applies two equivariant linear layers separated by a Gate nonlinearity. The spectral head uses hidden channel multiplicities $C_l = (128, 64, 32, 16, 8, 4, 4)$ for $l = 0, \ldots, 6$. Since the second linear layer only mixes channels, each output block satisfies $\operatorname{rank} \hat{H}^l \le \min(C_l, 2l+1)$, at most 8, 4, and 4 at $l = 4, 5, 6$. A target block has rank $m_l(G^*)$ (Lemma 1), which exceeds this cap only for E ($l \ge 4$), C_2 ($l \ge 5$), and C_3 ($l = 6$), groups already distinguished at lower degrees; these blocks are fit only approximately, a trade-off of the tapering between cost and coefficient expressivity. At the bands used for the CG variant in Appendix D.2 (cellphone L=2, table L=4), its final channel-mixing layer imposes no rank bound below the block dimension (12 hidden channels for the 9×9 block at $l = 4$). This variant yields no clear empirical gain, although it also changes cross-degree interactions and therefore does not isolate the effect of the rank constraint. The final output is the flattened block-diagonal coefficient matrix $\bigoplus_{l=0}^{L_{\max}} \hat{H}^l$.

Bound-aware output projection. We shape the output toward the smoothing bound $\|\hat{H}^l\|_F \leq \sqrt{2l+1} \cdot \kappa_l$ via three stages applied independently to each degree-l block; stages (i)–(ii) keep each block within a ceiling no tighter than the bound at the default bandwidth $\sigma = 0.2$, and stage (iii) is a learnable per-degree rescaling of that ceiling. The degree-0 block is exempt: it is fixed to 1 rather than predicted, so every prediction integrates to exactly 1 over $SO(3)$ (Appendix C.2).

(i) Direction normalization. Let $\mathbf{b}^l$ be the flattened degree-l output before projection. We normalize it as $\hat{\mathbf{d}}^l = \mathbf{b}^l / \max(\|\mathbf{b}^l\|_2, \epsilon_0)$ with $\epsilon_0 = 10^{-12}$.

(ii) Magnitude gating. The magnitude is mapped into a dynamic range via tanh gating, $m^l = r^l_{\min} + (r^l_{\max} - r^l_{\min}) \tanh(\|\mathbf{b}^l\|_2)$, where $r^l_{\max} = \sqrt{2l+1}\,\exp(-l(l+1)\sigma_p^2/2)$ with $\sigma_p = 0.1$, a ceiling no tighter than $\sqrt{2l+1} \cdot \kappa_l$ at the default $\sigma = 0.2$, since $\sigma_p^2/2 < \sigma^2$ (Lemma 3; see the note after it), and $r^l_{\min} = 0.8 \cdot r^l_{\max}/\sqrt{K}$ with $K = 6$ fixed, the vocabulary's largest cyclic order. Only the ceiling follows from the bound (Lemma 3); the floor is a heuristic for keeping early gradients nonzero. It cannot depend on the unknown per-shape $|G^*|$: dihedral orbits can have $|G| = 12$, and blocks without invariants have zero target norm (Table 4). The learnable multiplicative rescaling below therefore allows trained magnitudes to fall beneath this initial floor.

(iii) Learnable multiplicative rescaling. A per-degree scalar γ_l, initialized to 0, rescales the window, so trained magnitudes can settle above or below it, $\tilde{m}^l = m^l \cdot \exp(\gamma_l)$. The projected block combines the learned magnitude and unit-normalized coefficient direction as $\hat{H}^l = \tilde{m}^l\,\hat{\mathbf{d}}^l$.

Derivation of the spectral bound.

Lemma 3 (Spectral bound of the smoothed target)**.** *For each retained degree $l \leq L_{\max}$, the heat-kernel-smoothed target obeys the coefficient bound $\|H^{*l}\|_F \leq \sqrt{2l+1}\,\kappa_l$.*

Proof. In the canonical frame, the degree-l coefficient of the target f^* after smoothing is

$$H^{*l} = \kappa_l \cdot \frac{1}{|G^*|} \sum_{g \in G^*} D^l(g). \tag{46}$$

Since $D^l(g)$ is a $(2l+1) \times (2l+1)$ unitary matrix for every $g \in SO(3)$, its Frobenius norm is exactly $\|D^l(g)\|_F = \sqrt{\mathrm{tr}(D^l(g)^\top D^l(g))} = \sqrt{\mathrm{tr}(I_{2l+1})} = \sqrt{2l+1}$ for all g. The triangle inequality therefore bounds the norm of the averaged target coefficients:

$$\|H^{*l}\|_F \leq \kappa_l \cdot \frac{1}{|G^*|} \sum_{g \in G^*} \|D^l(g)\|_F = \kappa_l \cdot \sqrt{2l+1}. \tag{47}$$

□

The bound constrains only the maximum target norm, so any ceiling at or above it remains valid.

Training objective. The total loss combines supervised coefficient regression with a regularizer that encourages consistency between rotated views.

The equivariance loss enforces consistency between predictions from two randomly rotated views $\mathbf{P}_1 = \mathbf{P}R_1$ and $\mathbf{P}_2 = \mathbf{P}R_2$. Under perfect equivariance $\hat{H}^l_{2,b}$ equals $\tilde{H}^l_{2,b} := \hat{H}^l_{1,b}\,D^l(R_1^\top R_2)$. Since the naive normalized MSE of this constraint becomes unstable when coefficient norms approach zero, we instead measure consistency via block-wise cosine similarity:

$$c_l = \frac{1}{B} \sum_{b=1}^{B} \frac{\langle \tilde{H}^l_{2,b}, \hat{H}^l_{2,b} \rangle}{\max(\|\tilde{H}^l_{2,b}\|, \epsilon_1)\ \max(\|\hat{H}^l_{2,b}\|, \epsilon_1)}, \tag{48}$$

$$\mathcal{L}_{\text{equiv}} = 1 - \frac{1}{Z} \sum_{l=1}^{L_{\max}} w_l c_l, \tag{49}$$

with $\epsilon_1 = 10^{-6}$ and w_l and Z as in Sec. 4.4, where $l = 0$ is excluded since degree-0 coefficients are rotation-invariant scalars for which the equivariance constraint is trivially satisfied. Although our architecture is equivariant in exact arithmetic, floating-point evaluation deviates from it, *e.g.*, where max pooling selects among near-tied features. $\mathcal{L}_{\text{equiv}}$ acts as a soft regularizer that encourages consistency between the predictions for rotated views.

Combining supervision with this equivariance regularizer gives the final training objective,

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \alpha \mathcal{L}_{\text{equiv}}, \tag{50}$$

where α balances equivariance consistency and supervised coefficient regression.

Training. We train with the Adam optimizer, learning rate 2×10^{-3}, weight decay 10^{-4}, batch size 64 per GPU (128 effective with DDP), for 1500 epochs. We fix this schedule before the experiments, use it for every DORE run, and evaluate the final checkpoint. The baselines keep their own fixed schedules (3,000 epochs for CHOIR; 1,919 orienter and 3,719 flipper epochs for Scarvelis et al. [39]). The loss is multiplied by a constant 20 to keep its small values away from floating-point underflow. Training uses a 10-epoch linear warm-up with cosine decay, gradient-norm clipping at 1, and Gaussian jitter of 0.01 on each view; the band limit and σ are as in Sec. 5. We additionally apply an equivariance consistency regularizer between predictions on two randomly rotated views with weight $\alpha = 0.05$ (the $\mathcal{L}_{\text{equiv}}$ term of Eq. (5)); removing it reduces cellphone and table accuracy by 1.7 and 3.1 points, respectively, while leaving lamp accuracy nearly unchanged (Table 2).

Symmetry labeling. Our symmetry labeler uses beam search (Algorithm 2) with $k = 20$, $m = 5$, $L = 80$, $K = 5$, initial sample count $M = 2000$, and angular radius schedule $(25^\circ, 15^\circ, 9^\circ, 5^\circ, 3^\circ)$, keeping selected axes at least 10° apart. The implementation replaces $S(G)$ with $\tilde{S}(G) = -\frac{1}{|\Gamma_G|} \sum_{g \in \Gamma_G} d^{\text{norm}}_{\mathbf{P}}(g.\mathbf{P})$, where $d^{\text{norm}}_{\mathbf{P}} = d^{(1)}_{\mathbf{P}} / \rho_{\mathbf{P}}$ is the unsquared distance of Lemma 2 divided by the RMS radius $\rho_{\mathbf{P}} > 0$ of $\mathbf{P}$, and Γ_G is $C_n \setminus \{e\}$ for C_n and $D_n \setminus C_n$, the n half-turns about axes perpendicular to the principal axis, for D_n; both sets are closed under inversion, so Lemma 2 applies. The identity group threshold $\tau = -0.075$, the model-size prior weight $\beta = 10^{-2}$, and the dihedral tolerances below refer to $\tilde{S}$, which Algorithm 2 uses in place of S. The dihedral check accepts D_n when its error is at most $r = 1.2$ times the cyclic error plus $\epsilon_D = 0.01$. Because ShapeNet models are axis-aligned, the implementation does not search the phase of the flip axes about $\hat{a}_n$: it places them at multiples of π/n from the projection of the x-axis (the y-axis when $\hat{a}_n$ is near x). For 95% of the D_n labels of the six benchmark categories, a sweep of this phase finds its best value within 2.1° of the fixed phase and lowers the flip error by less than 1%. If some order $n \geq 3$ has cyclic error within the same tolerance of the smallest cyclic error, only such orders are kept (removing C_2 and D_2); the existence test and the group selection are both taken over this set $\mathcal{G}'$. Labeling uses 512 centered points per shape, with scale handled solely by RMS-radius normalization.

Inference acceleration. At test time the inputs are processed in batches of 64. The coarse search scores a whole batch with one matrix product of its coefficients with W, and the Newton–Raphson refinement shares its Wigner-D evaluations across the batch, so the search runs once per batch rather than once per shape. The search runs on the CPU and the forward pass on the GPU, so the two are pipelined across batches: the search of one batch overlaps the forward pass of the next. With best-of-N, the candidate extraction joins the CPU stage and the verifier of one batch overlaps the search of the next.

B.4 Implementation details: data and evaluation protocol

Input point clouds. All three methods compared in Table 1 use the Cap3D release of ShapeNet [26]: 16,384 points per shape on the outer surface, the clouds on which the symmetry labels are computed. DORE uses 1,024 points, matching the input size of CHOIR [21], on whose Vector Neurons backbone it builds. Both methods use zero-centered points at their original scale, without normals, during training and evaluation. Scarvelis et al. [39] take 2,000 points with normals, as its released code expects, with each cloud normalized once to mean 0 and largest radius 0.95. The baseline's default input requires normals, which a range scan does not measure directly; DORE's default inference procedure uses point positions alone.

The normals used by Scarvelis et al. [39] come from the ShapeNetCore v2 mesh of the same model. The mesh is brought into the frame of the cloud by a fixed rotation of $+90^\circ$ about the x-axis, centering of its bounding box and scaling of its largest extent to one, and a per-object scale and translation fitted to the cloud (rotation fixed, ten trimmed closest-point iterations). Each point takes the face normal of its closest triangle. Mesh winding does not reliably determine normal direction because many faces are duplicated with opposite winding. We therefore orient each normal by visibility: from the closest

surface point, we cast 16 random rays into each hemisphere and select the side from which more rays leave the mesh. Rare ties use $\mathbf{n}^\top(\mathbf{p} - \bar{\mathbf{p}}) \geq 0$. Every training iteration of the baseline draws 2,000 of the 16,384 points with their normals, and its test inputs are fixed draws. Four car training clouds of the release are degenerate (collapsed to a small cluster) and are excluded from the baseline's training.

Input scale check. About 0.2–2.6% of the Cap3D clouds in each category are much smaller than the rest while otherwise intact. Since DORE reads its input at the original scale, each test input is compared with its class: if the largest side of its PCA-aligned bounding box is below 0.75 times the median of that quantity over the class's training clouds, the input is scaled about its centroid by $(4/3)^n$ with the smallest n that brings the ratio to at least 0.75. The statistic uses training data only and is rotation invariant when the covariance eigenvalues are distinct; inputs above the threshold are untouched. The check raises the six-category mean $\mathrm{Acc}_{\mathrm{sym}}@\epsilon$ from 82.6 to 82.8. The baseline normalizes every cloud and needs no such check.

Why we use 1,024 input points. Increasing DORE's input from 1,024 to 2,000 points changes the spatial extent of its neighborhoods when $k = 40$ is held fixed. Its only local aggregation uses one k-nearest-neighbor graph on the input positions. A denser cloud therefore shrinks the region each neighborhood covers, from 3.9% of the points at 1,024 to 2.0% at 2,000 (a radius about $\sqrt{1{,}024/2{,}000} \approx 0.72$ times as large). Keeping the receptive field the layers were designed for would require scaling k, and with it the architecture, with the input.

Baseline retraining protocol. The baseline results in Tables 1 and 3 come from our own evaluations; none is transcribed from Scarvelis et al. [39], whose paper reports multi-category results only. Both of its networks are trained from scratch with the authors' public code and settings for their full schedule, with no early stopping (1,919 orienter and 3,719 flipper epochs, scored at the final saved checkpoints, epochs 1,910 and 3,710), once per category on our 80:20 split (the first 80% of each category's shapes in ShapeNet model-ID order for training) instead of one model across categories, on the inputs described above. CHOIR [21] is retrained per category on the same split with its default settings (1,024 points, no normals) and calibrated as below.

Evaluation protocol. We evaluate every run at its final checkpoint under the same frozen test protocol. Each test shape is evaluated under eleven fixed rotations (one drawn from a seeded generator, ten stored with the dataset). All Chamfer distances are computed on a frozen 2,000-point cloud per view, normalized once per object (mean 0, largest radius 0.95), as the symmetric $d_{\mathrm{CD}}(X, Y) = d_X^{(2)}(Y) + d_Y^{(2)}(X)$ of the one-sided term of Eq. (6); $\mathrm{Acc}_{\mathrm{sym}}$ is thresholded per object by $\epsilon_\theta(\mathbf{P})$ of Eq. (12). All methods are scored using these same rotations, clouds, and thresholds, regardless of their model inputs. We release the frozen evaluation set with the code. Scarvelis et al. [39] retain their official unseeded flipper voting, so TTA results vary slightly across runs.

Orientation calibration for CHOIR. We calibrate each class using unrotated test clouds (Sec. 5),

$$R_{\mathrm{calib}} = \operatorname*{argmin}_{R \in SO(3)} \sum_{i=1}^{N} d_{\mathrm{CD}}\big(\mathbf{P}_i \hat{R}_i^\top R,\, \mathbf{P}_i\big), \tag{51}$$

where $\hat{R}_i$ is the prediction on the unrotated input $\mathbf{P}_i$, and the metrics of Eqs. (10)–(11) are reported with $\hat{R}_k$ replaced by the calibrated rotation $\hat{R}_k^{\mathrm{calib}} = R_{\mathrm{calib}}^\top \hat{R}_k$ (the CHOIR† rows). A deterministic search finds one rotation per class using 65 initializations and coordinate refinement down to 0.125°. Because calibration uses the test shapes' canonical poses, this adjustment benefits CHOIR† alone; the other methods receive neither this adjustment nor any test-set calibration.

Latency breakdown. We measure latency for every method on the same RTX 3090, 12 physical cores of an Intel Xeon Gold 6342 CPU (2.8 GHz), and software stack, without kernel-level optimizations (Table 3). For each category, we use the first 23 frozen test batches, discard the first 3, and divide the remaining wall time by their 1,280 inputs, including result transfer to the host. We report the median of three runs, averaged over the six categories.

Table 5 breaks down DORE's per-shape latency into stages, run sequentially with a synchronization after each, under the protocol above. A single pass spends 2.8 ms in the forward pass and 1.0 ms in

the peak search (3.8 ms); pipelining overlaps the two, giving 3.0 ms. The search grows with the band, from 0.5–0.6 ms at L=2 to 1.4 ms on table (L=4) and 2.5 ms on lamp (L=6), while the forward pass stays at 2.7–2.8 ms. With best-of-N, candidate extraction raises the search to 2.6 ms and the verifier adds 4.6 ms on average, from under 0.01 ms on car and chair, where it is called on at most 0.08% of the timed inputs, to 13.0 ms on cellphone, where it is called on all of them; overlapping these stages gives a pipelined best-of-N latency of 7.8 ms.

Table 5: **Per-shape latency of DORE by stage (ms).** We time its inference stages sequentially with synchronization after each, while the final column reports the pipelined latency used in Table 3. Values average the six benchmark categories using the same hardware and timing protocol.

	forward	search	verifier	sequential	pipelined
DORE	2.8	1.0	–	3.8	3.0
DORE + BoN	2.8	2.6	4.6	10.0	7.8

Hardware. All models trained in our pipeline (DORE and the distributional heads of Appendix D.2) use two GPUs (Titan RTX or RTX 3090; the two have the same 24 GB of memory, so every run shares one configuration, including the per-GPU batch size) with Distributed Data Parallel training under the configuration of Appendix B.3. Scarvelis et al. [39] follows its own schedule and settings (Baseline retraining protocol above).

C Analysis of the objective and the target

C.1 The training objective as a Sobolev metric

Eq. (5) weights the per-entry-averaged block errors, $\mathcal{L}_{\text{mse}}^{(l)} = \frac{1}{2(2l+1)^2}\sum_{i=1}^{2}\|\hat{H}_i^l - H_i^{*l}\|_F^2$ with the standard Frobenius norm, by $w_l = l(l+1)$. Since the degree-l block has $(2l+1)^2$ entries, the effective weight per coefficient entry is

$$\frac{l(l+1)}{(2l+1)^2} = \frac{1}{4}\Big(1 - \frac{1}{(2l+1)^2}\Big), \tag{52}$$

which is bounded above and below by absolute constants for all $l \geq 1$, independently of the band limit. The supervised objective is thus equivalent, up to constants, to flat coefficient ℓ^2 error.

This approximately uniform per-entry weighting is consistent with the high-frequency emphasis described in Sec. 4.4: the two statements use different reference weights. Relative to uniform per-degree averaging (the "w/o Laplace–Beltrami weighting" row of Table 2), our weights increase each entry's contribution by $l(l+1)$, a tenfold difference between the l=1 and l=4 blocks. The corresponding density-space metric follows by comparison with the L^2 norm.

For band-limited densities f and g, the Plancherel identity (Eq. (18)) gives

$$\|f - g\|_{L^2(SO(3))}^2 = \sum_{l=0}^{L_{\max}} (2l+1)\,\|H_f^l - H_g^l\|_F^2, \tag{53}$$

so the flat ℓ^2 scales the degree-l term of the L^2 distance by $1/(2l+1) \asymp (1+l(l+1))^{-1/2}$: it is equivalent to the squared Sobolev distance $H^{-1/2}$. By the same accounting, removing the Laplace–Beltrami weighting is equivalent to the squared $H^{-3/2}$ distance, and the squared Parseval L^2 distance corresponds to $w_l \propto (2l+1)^3$. These are members of the one-parameter family $w_l \propto (2l+1)^{2s+3}$ over the Sobolev exponent s; negative orders compare smoothed densities and are thus consistent with our heat-kernel targets.

We now derive this correspondence from the derivative-based definition of Sobolev norms. Let X_1, X_2, X_3 be an orthonormal basis of $\mathfrak{so}(3)$ with respect to $\langle A, B\rangle = -\frac{1}{2}\operatorname{tr}(AB)$, i.e., half the negative Killing form, acting on $C^\infty(SO(3))$ as the associated left-invariant vector fields, and define the bi-invariant Laplace–Beltrami operator $\Delta = \sum_i X_i^2$. Its degree-l eigenvalues have magnitude $\lambda_l = l(l+1)$, which determines the spectral weights below.

Lemma 4 (Casimir spectrum and the Dirichlet energy)**.** *Each X_i is skew-symmetric with respect to the Haar L^2 inner product on smooth functions, and $\Delta D^l_{ij} = -\lambda_l D^l_{ij}$ for every l and every entry. For any finite expansion u with coefficient blocks H^l_u, this yields*

$$\sum_{i=1}^{3} \|X_i u\|^2_{L^2} = \langle -\Delta u,\, u\rangle = \sum_l (2l+1)\,\lambda_l\,\|H^l_u\|^2_F. \tag{54}$$

Proof. The flow of a left-invariant vector field is right translation, which preserves the Haar measure, so $\int (X_i u)v = -\int u(X_i v)$ and $X_i^* = -X_i$; the first equality in Eq. (54) follows. For the spectrum, the homomorphism property gives $D^l(R\exp(tX)) = D^l(R)\,D^l(\exp(tX))$, so Δ acts on the degree-l block by the Casimir element of the degree-l irrep, which is the scalar $-l(l+1)$. The last equality is the polarized (inner-product) form of Eq. (18) applied to the pair $(-\Delta u, u)$, whose respective degree-l Fourier coefficient matrices are $\lambda_l H^l_u$ and H^l_u. □

Proposition 5 (Derivative norms are degree weights)**.** *For every integer $k \geq 0$, with the polynomial $p_k(\lambda) = \sum_{m=0}^{k} \lambda^m$ of degree k, positive on $[0,\infty)$,*

$$\sum_{|\alpha|\leq k} \|X^\alpha u\|^2_{L^2} = \sum_l (2l+1)\,p_k(\lambda_l)\,\|H^l_u\|^2_F, \tag{55}$$

the sum running over all 3^m ordered products $X^\alpha = X_{i_1}\cdots X_{i_m}$ of each order $m \leq k$; in particular $p_1(\lambda) = 1+\lambda$. Hence the squared classical Sobolev norm $\|u\|^2_{W^{k,2}}$ is equivalent to $\sum_l (2l+1)(1+\lambda_l)^k \|H^l_u\|^2_F$, the form that defines $\|u\|^2_{H^s}$ for real s through $(I-\Delta)^{s/2}$.

Proof. The sum over all ordered words of one order obeys an exact identity. Write, bilinearly, $Q_m(v,u) := \sum_{i_1,\ldots,i_m} \langle X_{i_1}\cdots X_{i_m} v,\; X_{i_1}\cdots X_{i_m} u\rangle$. Skew-symmetry gives $Q_1(v,u) = \langle -\Delta v,\, u\rangle$ (Lemma 4), and since the Casimir Δ is central (it commutes with every X_i), we group terms by their outermost factor to obtain the following recurrence:

$$Q_m(v,u) = \sum_{|\alpha|=m-1} \big\langle -\Delta X^\alpha v,\; X^\alpha u\big\rangle = Q_{m-1}(-\Delta v,\, u), \tag{56}$$

so induction on m yields $Q_m(u,u) = \langle (-\Delta)^m u,\, u\rangle = \sum_l (2l+1)\,\lambda_l^m\,\|H^l_u\|^2_F$, the last step by the polarized form of Eq. (18). Summing over $m \leq k$ gives the statement with $p_k(\lambda) = \sum_{m=0}^{k}\lambda^m$ exactly, which is positive on $[0,\infty)$ and satisfies $p_k(\lambda) \asymp (1+\lambda)^k$, the stated equivalence. □

Corollary 1 (Locating Eq. (5): $s = k - 3/2$)**.** *Let the difference u have $H^0_u = 0$ and set $w_0 = 0$, as in training. Because $\mathcal{L}^{(l)}_{\mathrm{mse}}$ averages over the $(2l+1)^2$ entries of its block, the weight Eq. (5) places on $\|\hat{H}^l - H^{*l}\|^2_F$ is $\omega_l = w_l/(2l+1)^2$. Since $(2l+1)^2/4 \leq 1+\lambda_l \leq (2l+1)^2$ for all $l \geq 0$, the H^s weight satisfies $4^{-|s|}(2l+1)^{2s+1} \leq (2l+1)(1+\lambda_l)^s \leq 4^{|s|}(2l+1)^{2s+1}$, so taking $w_l = \lambda_l^k$ with k real (the derivative reading of Proposition 5 applying at integer k) gives $\omega_l \asymp (2l+1)^{2k-2}$ and hence an objective equivalent to $\|u\|^2_{H^s}$ with*

$$s = k - \tfrac{3}{2}. \tag{57}$$

The statement concerns $Z\,\mathcal{L}_{\mathrm{sup}}$, averaged over the two views; $Z = \sum_l w_l$ grows with the band and only rescales the objective at a fixed band. The $-3/2$ comes from two normalizations: averaging over the $(2l+1)^2$ entries and omitting the Plancherel factor $(2l+1)$. The four settings of Table 6 are $k=0$ (uniform averaging, $s=-3/2$), the default $k=1$ (the Dirichlet energy of Eq. (54), $s=-1/2$), $k=3/2$ (Parseval L^2, $s=0$), and $k=2$ ($s=+1/2$), the last two run with the equivalent weights $w_l = (2l+1)^{2k}$.

Two details concern the implemented loss. First, the default weights are $\omega_l = \lambda_l/(2l+1)^2$ rather than a constant, the flat weight equivalent to $s=-1/2$, and the two agree to within 11% for every $l \geq 1$, since $\omega_l = \frac{1}{4}(1-(2l+1)^{-2}) \in [\frac{2}{9}, \frac{1}{4})$. Second, $w_0 = 0$ makes Eq. (5) a squared seminorm whose kernel is the degree-0 component, but the kernel is never exercised: the prediction pins that block to 1 (Appendix B.3) and the smoothed orbit target has $H^{*0} = 1$ for every group, so the loss acts on differences with zero degree-0 coefficients, where it is a squared norm.

Table 6: **Sobolev-weight sensitivity at L_{sym}.** Cellphone, table, and lamp use $L_{\max} = 2/4/6$. The $s = -\frac{3}{2}$ and $s = -\frac{1}{2}$ rows report the same results as the unweighted and default rows in Table 2, respectively. Each cell uses the training setup of Appendix B.3 and one run. Accuracy is in %, and CD is scaled by 10^{-3}. Bold marks the best result per class and metric, without establishing an optimal exponent.

		table		cellphone		lamp	
s (metric)	$w_l \propto$	$\mathrm{Acc}_{\mathrm{sym}} \uparrow$	CD $\downarrow$	$\mathrm{Acc}_{\mathrm{sym}} \uparrow$	CD $\downarrow$	$\mathrm{Acc}_{\mathrm{sym}} \uparrow$	CD $\downarrow$
$-3/2$ (uniform averaging)	1	70.5	7.4	80.4	5.5	44.2	30.5
$-1/2$ (default, Eq. (5))	$l(l+1)$	**84.2**	5.1	82.4	**5.0**	**45.6**	**28.2**
0 (Parseval L^2)	$(2l+1)^3$	84.1	5.0	**83.0**	5.1	40.6	33.3
$+1/2$	$(2l+1)^4$	84.1	**4.6**	81.7	**5.0**	34.3	41.5

This correspondence characterizes the existing objective rather than establishing its optimality. Table 6 compares the default with other Sobolev weights at the three assigned L_{sym} bands ($L_{\max} = 2/4/6$).

The default is close to the best setting in every column: it has the highest accuracy on table and lamp and is within 0.6 pp of the best on cellphone. Since $w_l \propto (2l+1)^{2s+3}$, raising s shifts the loss weight toward the highest degrees of the band and away from the lower ones, and lowering it does the reverse: over $l \geq 1$, the degrees $l \leq 2$ receive 7.1% of the weight at $s = -1/2$, 3.2% at $s = 0$, and 1.3% at $s = +1/2$ when $L_{\max} = 6$, and at $L_{\max} = 4$, the top degree receives 50% under the default weights but 25% under uniform weighting.

The L_{sym} rule chooses the lowest degree that expresses every group above a 2% label share (Sec. 4.6). Its top degree is therefore needed for the highest-order group, while lower degrees encode the remaining target structure. Uniform weighting lowers the weight of the top degree and costs 13.7 pp on table, whose $L_{\mathrm{sym}} = 4$ comes from its C_4 labels; sharper weighting lowers that of the low degrees and costs 5.0 and 11.3 pp on lamp, the widest band. On cellphone, with only two degrees above $l = 0$, the exponent has little weight to redistribute, and the four settings span 80.4–83.0.

Each cell is a single run, so the differences on cellphone and among the settings $s = -1/2, 0, +1/2$ on table are not resolved against the seed spread of Table 15. These single-run comparisons therefore characterize sensitivity rather than choose the benchmark exponent.

C.2 Properness and the smoothing trade-off

Truncating the orbit density at $L_{\max}$ leaves negative side lobes; the heat-kernel bandwidth σ of the target damps them but widens the mode. For a single peak ($|G| = 1$) both effects are closed-form, since the smoothed, truncated kernel depends only on the geodesic angle θ from the peak,

$$f_\sigma(\theta) = \sum_{l=0}^{L_{\max}} (2l+1)\, e^{-l(l+1)\sigma^2} \, \frac{\sin\big((l+\frac{1}{2})\theta\big)}{\sin(\theta/2)}. \tag{58}$$

The degree-0 term fixes the integral at 1, since every higher-degree character integrates to zero under normalized Haar measure. This unit-mass constraint survives truncation but does not ensure nonnegativity. In particular, negative mass is not a fraction of probability: if its magnitude is a, the positive part integrates to $1 + a$. Thus, the negative mass 0.94 reported below coexists with positive mass 1.94, whose difference is 1.

Table 7 reports, at $L_{\max} = 4$, its undershoot $\min_\theta f_\sigma / f_\sigma(0)$, its full width at half maximum, and its negative mass $|\int f^-|$ (against a total of $\int f = 1$), the same quantity for the predictions of models trained on chair at each σ, and their accuracy. We sweep at $L_{\max} = 4$ rather than chair's $L_{\mathrm{sym}} = 2$ because only there does this range of σ span the trade-off: at $L_{\max} = 2$ the undershoot shrinks only from -12.0% to -5.4% by $\sigma = 0.4$, around a mode already 100° wide. At $L_{\max} = 6$, the undershoot is already only -2.3% at the default bandwidth $\sigma = 0.2$.

The negative lobes are shallow but cover about half of $SO(3)$, so at $\sigma = 0.2$ they integrate to -0.94 at a depth of only 5.8%. Smoothing therefore trades precision for properness: a larger σ removes more of the negative mass but widens the mode (58° at $\sigma = 0$, 81° at 0.4), and the models trained on

the smoother targets lose accuracy ($\mathrm{Acc}_{\mathrm{sym}}@\epsilon$ on chair 96.3 at $\sigma = 0$ and 93.4 at 0.4; mean Chamfer distance 1.75 and 2.19). Because the trade-off depends on the category and band, we use $\sigma = 0.2$ throughout without claiming it is optimal. Because orientations are selected from the maxima of f, the remaining negative side lobes do not enter orientation selection.

Table 7: **Smoothing trade-off at $L_{\max} = 4$.** Target columns describe the single-peak kernel ($|G| = 1$; Eq. (58)): relative undershoot $\min f/f(0)$, full width at half maximum, and negative mass $|\int f^-|$ with $\int f = 1$. Trained columns use identically trained chair models; $\sigma = 0.2$ reuses Appendix D.3. Predicted negative mass averages 128 test shapes on a 30,314-rotation grid. Each setting uses one run, with CD scaled by 10^{-3} under the evaluation protocol of Table 1.

	target (closed form)			trained (chair)		
σ	undershoot	FWHM	$\lvert\int f^-\rvert$	$\lvert\int f^-\rvert$	$\mathrm{Acc}_{\mathrm{sym}}@\epsilon\uparrow$	CD $\downarrow$
0 (no smoothing)	-9.7%	58°	2.62	2.50	96.3	1.75
0.10	-8.7%	59°	2.05	1.97	96.6	1.71
0.20 (ours)	-5.8%	62°	0.94	0.94	95.0	1.95
0.30	-1.5%	69°	0.18	0.23	95.0	2.21
0.40	-0.3%	81°	0.02	0.06	93.4	2.19

D Additional experiments and analyses

This appendix expands the evaluation in Sec. 5, connecting the benchmark results to metric choice, representation capacity, and inference design. Appendix D.1 examines where the two metrics of Tables 1 and 3 agree or disagree and compares the per-object threshold with a class-wide one, Appendix D.2 compares parametric density heads under a matched setup, Appendix D.7 analyzes predictions broader than their targets, Appendix D.8 motivates best-of-N from failures of the top-1 prediction, and Appendix D.9 gives the settings of the test-time procedures.

Appendices D.3–D.6 examine how band limit and data size affect stability and convergence. Together, they identify two requirements for choosing a band. First, the band must be wide enough to represent the category's symmetry groups: on table, $L_{\max} \in \{2, 3\}$ stay below $L{=}4$ in accuracy, and across the 13-category sweep the uniform $L{=}4$ is the strongest single band (Appendix D.3).

Second, its capacity must remain compatible with the available data, since each degree adds $(2l{+}1)^2$ coefficients: the collapse at $L{=}5$ is confined to the two categories with the fewest training shapes, cellphone (871) and monitor (874); on cellphone it does not converge within the 1,500-epoch schedule and occurs in two of five seeds, while at its L_{sym} band ($L{=}2$) three seeds give 81.53 ± 1.14 with no collapse (Appendix D.6); on car, with $3.2\times$ cellphone's training data, every band reaches the same floor; and at a fixed band, accuracy grows with the number of training shapes by a class-dependent amount (Appendix D.4). The L_{sym} rule balances these requirements using labels alone (Sec. 4.6; Appendix D.3): the smallest band covering groups above a 2% share assigns cellphone $L = 2$.

D.1 Metric agreement and the success threshold

Agreement of the two metrics. Against Scarvelis et al. [39] with test-time augmentation (Table 3), the two metrics rank methods differently only on car: their method achieves higher accuracy, while DORE achieves lower mean Chamfer distance. The rankings agree across all categories in the single-pass comparison (Table 1). Accuracy counts predictions below the per-object error threshold, whereas mean Chamfer distance also reflects the magnitude of each error. Among successful predictions on car, DORE achieves a much lower median error relative to that threshold (0.02–0.03 versus 0.16; Table 8), revealing a difference in precision that accuracy does not capture.

Per-object and class-wide threshold. Table 1 thresholds each sample by its own object's $\epsilon_\theta(\mathbf{P})$ (Eq. (12)). Table 9 repeats its accuracy with this per-object threshold and with one threshold per class, $\bar{\epsilon}_\theta$, the mean of $\epsilon_\theta(\mathbf{P})$ over the class's test objects, on the same predictions and per-sample Chamfer distances. The two thresholds give accuracies within 0.8 pp of each other on every category except lamp, where the class-wide threshold raises accuracy by 2.9–5.1 pp for every method except CHOIR. The order of the methods is the same in every column under both thresholds.

Table 8: **Accuracy and prediction errors on car.** We compare accuracy, mean Chamfer distance, and the median error of successful predictions on the car category, using the settings of Table 3. The final column reports the median of $d_{\mathrm{CD}}/\epsilon_\theta(\mathbf{P})$ over successful predictions. CD is scaled by 10^{-3}. Bold: best; underline: second best.

	$\mathrm{Acc}_{\mathrm{sym}}@\epsilon \uparrow$	CD $\downarrow$	median error of successes
Scarvelis et al. [39] + TTA	**99.5**	1.6	0.16ϵ
DORE	98.3	<u>1.2</u>	<u>0.03ϵ</u>
DORE + BoN + CA	<u>99.1</u>	**1.1**	$\mathbf{0.02\epsilon}$

Table 9: **Effect of threshold choice on accuracy (%).** Left: each object's $\epsilon_\theta(\mathbf{P})$ from Eq. (12). Right: the class mean $\bar{\epsilon}_\theta$ over test objects. Both use $\theta = 10^\circ$ and the predictions and distances of Table 1, isolating the threshold's effect. Bold: best; underline: second best.

Method	per-object $\epsilon_\theta(\mathbf{P})$						class-wide $\bar{\epsilon}_\theta$					
	plane	car	chair	table	lamp	phone	plane	car	chair	table	lamp	phone
CHOIR [21]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.5
CHOIR† [21]	54.2	86.6	80.8	13.1	16.0	0.5	54.9	87.3	80.4	12.4	20.1	0.5
Scarvelis et al. [39]	<u>81.9</u>	<u>90.1</u>	<u>85.5</u>	<u>80.5</u>	<u>36.9</u>	<u>67.6</u>	<u>82.5</u>	<u>90.0</u>	<u>85.7</u>	<u>80.0</u>	<u>42.0</u>	<u>67.3</u>
DORE (ours)	**92.6**	**98.3**	**93.8**	**84.2**	**45.6**	**82.4**	**93.1**	**98.5**	**93.8**	**84.1**	**48.5**	**83.2**

D.2 Distributional heads under a matched setup

We compare distributional heads under a matched training setup by training matrix-Fisher [29] and Bingham-mixture [14] heads from scratch with the same backbone, data, schedule, evaluation protocol, and metric code. The heads differ in their density representation, supervision, and orientation inference procedure, and every head is evaluated at the final checkpoint. The Bingham mixture uses orbit components recovered from the $L{=}5$ labels stored when this baseline was trained, normally with $K = |G^*|$. It shares a single concentration matrix across components, with the recovered orbit supplying the component modes when computing the loss. At $L{=}5$, however, C_6 and D_6 have the same retained coefficients as C_{12} and D_{12}, respectively. These cases use the latter orbits as band-equivalent surrogates that differ from their true orbits ($K = 12$ and 24), affecting 131 of the 6,748 table training shapes and none on cellphone. Table 1 instead benchmarks the full single-pass pipelines.

Orientation inference also differs: the parametric heads use closed-form solutions. The matrix-Fisher head uses the analytic mode $U\,\mathrm{diag}(1, 1, \det UV^\top)V^\top$ of $\arg\max_R \mathrm{tr}(F^\top R)$, exact and free of discretization, and the mixture uses the same mode of its component F, one element of the predicted orbit rather than the maximum of the full mixture, while DORE incurs a grid search plus refinement (Appendix B.1).

Table 10: **Density heads under a matched training setup.** Cellphone and table use the same backbone, data, and schedule across heads. Both DORE rows use L_{sym}. We report $\mathrm{Acc}_{\mathrm{sym}}@\epsilon$ in % and CD scaled by 10^{-3}.

	cellphone		table	
Head	$\mathrm{Acc}_{\mathrm{sym}}@\epsilon \uparrow$	CD $\downarrow$	$\mathrm{Acc}_{\mathrm{sym}}@\epsilon \uparrow$	CD $\downarrow$
matrix-Fisher [29]	2.7	114.3	21.2	42.0
Bingham mixture (label-derived components) [14]	8.6	65.7	22.5	37.1
DORE (Wigner-D density, CG tensor-product head)	**82.5**	5.6	83.3	5.3
DORE (Wigner-D density, default head)	82.4	**5.0**	**84.2**	**5.1**

Table 10 shows a large gap on both symmetric categories. The matrix-Fisher statistic degenerates on an orbit and leaves the azimuth unconstrained. The default-head row reuses the corresponding runs of Table 1; the CG row is described below. Both exceed the parametric heads on every cell, and the default-head row puts the gap at 62–80 pp $\mathrm{Acc}_{\mathrm{sym}}@\epsilon$, measured between the spectral head and these parametric heads, each with its own supervision and orientation inference procedure, on symmetric categories under the same setup.

The CG row augments the head with a Clebsch–Gordan tensor-product channel that lets the degree-l blocks exchange information across degrees, projected to a reduced multiplicity so that head parameters grow only $2.35\times$, trained at the same bands (cellphone L=2, table L=4). Zeroing the added channel recovers the original head, but the extension gives no clear gain: accuracy changes by $+0.1$ pp on cellphone and -0.9 pp on table, while mean Chamfer rises (5.6 vs. 5.0 and 5.3 vs. 5.1).

D.3 Band-limit sweep and the L_{sym} rule

Table 11 sweeps the band limit $L_{\max} \in \{2, \ldots, 6\}$ on all 13 categories, and Table 12 gives both metrics on the six benchmark categories at the uniform bands $L \in \{4, 5, 6\}$, the L_{sym} bands, and the per-class best band. Among uniform bands, L=4 is the strongest (mean 76.3), followed by L=6 (74.3). The collapse at L=5 is confined to the two categories with the fewest training shapes, cellphone (47.9) and monitor (15.6; Appendix D.6). Both recover at $L = 6$ (75.7, 66.2), showing that the degradation is not monotonic in bandwidth.

Table 11: **Accuracy across band limits on 13 categories.** Each $L_{\max} \in \{2, \ldots, 6\}$ uses one run evaluated at the final checkpoint. Bold marks the best swept band per category; the final columns compare that band with the label-based L_{sym} choice (Eq. (9)).

Category	L=2	L=3	L=4	L=5	L=6	best	L_{sym}
Airplane	92.6	94.0	94.3	93.9	**94.4**	6	2
Bench	88.7	88.2	**88.9**	86.1	85.6	4	2
Cabinet	21.5	14.9	**48.3**	42.0	35.3	4	2
Car	98.3	**98.9**	98.7	98.2	98.7	3	2
Chair	93.8	94.5	**95.0**	94.4	**95.0**	4, 6	2
Couch	91.6	92.0	**92.6**	92.3	92.5	4	2
Firearm	88.0	89.0	89.4	**91.1**	90.8	5	6
Lamp	37.0	36.6	45.5	41.5	**45.6**	6	6
Monitor	58.5	60.9	**67.0**	15.6	66.2	4	2
Cellphone	**82.4**	77.4	80.6	47.9	75.7	2	2
Speaker	7.6	7.0	**20.3**	20.1	15.4	4	2
Table	69.7	73.0	**84.2**	82.9	82.4	4	4
Watercraft	86.6	**88.0**	87.3	86.9	87.6	3	6
Mean	70.5	70.3	**76.3**	68.7	74.3		

Table 12: **Uniform and per-class band selection.** Metrics follow Table 1; latency follows Appendix B.4. The L_{sym} row repeats its results; the gray per-class best band is an oracle reference. The cellphone result at $L = 5$ is a single run in the unstable regime analyzed in Appendix D.6.

Configuration	Mean Chamfer distance ($\times 10^{-3}$) ↓						$\mathrm{Acc}_{\mathrm{sym}}$@$\epsilon$ (%) ↑						Latency (ms) ↓
	plane	car	chair	table	lamp	phone	plane	car	chair	table	lamp	phone	
DORE (uniform L=4)	1.2	1.1	1.9	5.1	25.9	5.8	94.3	98.7	95.0	84.2	45.5	80.6	2.9
DORE (uniform L=5)	1.2	1.2	2.0	5.2	28.0	14.0	93.9	98.2	94.4	82.9	41.5	47.9	3.1
DORE (uniform L=6)	1.1	1.1	2.0	6.0	28.2	6.8	94.4	98.7	95.0	82.4	45.6	75.7	3.5
DORE (L_{sym}, 2/2/2/4/6/2)	1.5	1.2	2.1	5.1	28.2	5.0	92.6	98.3	93.8	84.2	45.6	82.4	3.0
DORE (best L, oracle; 6/3/4/4/6/2)	1.1	1.2	1.9	5.1	28.2	5.0	94.4	98.9	95.0	84.2	45.6	82.4	3.1

To assess the training-free rule in Eq. (9), we define each category's regret as the accuracy gap between its L_{sym} setting and its best swept band. Over the 13 categories the median regret is 0.6 pp, nine lie within 1.5 pp, and the rule selects the best swept band on lamp, cellphone, and table. The three large regrets, cabinet (26.8 pp), speaker (12.7), and monitor (8.5), belong to categories whose labels leave them at L=2 (Table 24) while their accuracy rises at L=4, a gain the label statistics the rule reads do not show. At one run per cell, gaps of about a point are comparable to some of the seed standard deviations we measured at the selected bands of six categories (Table 15). Appendix F.4 shows each label truncated at every band.

D.4 Sample-size sweep

To isolate the effect of training-set size from class identity, each of the 13 categories of Table 11 is subsampled at the fixed default band $L_{\max} = 4$ under the training setup of Appendix B.3, over the

sizes $n \in \{871, 2000, 5000\}$ plus its full split: 871 is cellphone's full training-set size, and a category is subsampled to a size only if its training split exceeds it (Table 13); monitor retains all 874 training shapes because its full split exceeds 871 by only three.

Table 13: **Accuracy (%) versus training-set size at $L_{\max} = 4$.** Subsets are deterministic and nested, with the test split fixed. Dashes indicate unavailable sizes, including monitor at $n = 871$ because its 874-shape split is not subsampled. Cellphone's full split has 871 shapes. Full-split results reuse Table 11; all cells use seed 0 and the evaluation protocol of Table 1.

Category	n_{full}	n=871	n=2000	n=5000	full
plane	3236	88.0	93.9	—	94.3
car	2811	97.0	98.8	—	98.7
chair	5422	84.0	89.8	95.0	95.0
table	6748	61.5	77.5	82.4	84.2
lamp	1854	41.8	—	—	45.5
phone	871	80.6	—	—	80.6
bench	1450	88.8	—	—	88.9
cabinet	1256	39.1	—	—	48.3
couch	2538	85.5	90.9	—	92.6
firearm	1898	89.8	—	—	89.4
monitor	874	—	—	—	67.0
speaker	1277	18.1	—	—	20.3
watercraft	1551	84.3	—	—	87.3

At 871 shapes the change against the full split ranges from $+0.4$ pp (firearm) to -22.7 pp (table): the dependence on data volume is real but strongly class-dependent. Table, chair (-11.0 pp), cabinet (-9.2 pp), and couch (-7.1 pp) lose the most; bench and firearm stay within 0.4 pp of their full split and car within 1.7 pp, near saturation. Only two changes break monotonicity, both near saturation and within 0.4 pp: car at $n = 2000$ (98.8 vs. 98.7) and firearm at $n = 871$ (89.8 vs. 89.4).

D.5 Training convergence across bands

Table 14 tracks cellphone training at each band, reporting the first epoch at which the monitored Chamfer distance $\text{CD}_{\text{monitored}}$ reaches each target and its lowest value. Lower bands reach each target earlier, and $L_{\max}$=5 reaches none of them within the 1,500-epoch schedule (lowest 5.16).

Table 14: **Cellphone convergence across bands.** Entries give the first epoch at which the monitored Chamfer distance $\text{CD}_{\text{monitored}}$ ($\times 10^{-3}$) reaches each target, and its minimum. $\text{CD}_{\text{monitored}}$ is computed during training every 20 epochs on 50 fixed test shapes with 1,024-point clouds and does not select checkpoints; it is not on the scale of the CD in Table 1. The $L_{\max} = 5$ run is the degraded seed in Table 16.

	Epoch to $\text{CD}_{\text{monitored}} \leq$			Lowest
$L_{\max}$	4.0	3.0	2.5	$\text{CD}_{\text{monitored}}$
2	259	299	319	1.56
3	399	439	599	1.90
4	539	579	759	2.17
5	never	never	never	5.16

D.6 Run-to-run variance

We repeat each of the six per-class L_{sym} configurations with three random seeds while holding all other settings fixed. Cellphone has the largest accuracy standard deviation, 1.14 pp (Table 15).

Larger instabilities appear outside the L_{sym} bands, at a high band on the categories with the fewest training shapes: in the 13-category band sweep (Table 11) only cellphone (871 training shapes) and monitor (874) collapse at $L = 5$. Table 16 repeats cellphone at $L = 5$ under five seeds: two degrade while three reach 78.1–78.3. Damping the degree-4 and degree-5 target blocks ($\times 0.5$, $\times 0.25$) does not prevent degradation.

Table 15: **Seed variability at the selected bands.** Each of the six L_{sym} configurations (Sec. 4.6) is retrained with three seeds and otherwise identical settings. Values are mean $\pm$ sample standard deviation at epoch 1,500; seed 0 supplies Table 1. Accuracy is in %, and CD is scaled by 10^{-3}.

Category	band	$\mathrm{Acc}_{\mathrm{sym}}@\epsilon \uparrow$	CD $\downarrow$
plane	L=2	92.88 ± 0.31	1.46 ± 0.06
car	L=2	98.38 ± 0.09	1.21 ± 0.05
chair	L=2	94.59 ± 0.85	2.19 ± 0.30
table	L=4	83.26 ± 0.97	5.08 ± 0.16
lamp	L=6	44.85 ± 0.99	29.15 ± 0.90
phone	L=2	81.53 ± 1.14	5.32 ± 0.58

The degraded runs fit their own training set less well: over all training shapes, without test data, their training loss is 0.354–0.368 against 0.323–0.326 for the converged runs, and 0.256–0.282 against 0.252–0.253 with damping, so a degraded run is identifiable without the test split. The gap lies in the degree-2 and degree-4 blocks (with damping, which suppresses degree 4, mainly in degree 2), while degrees 1, 3, and 5 are fit equally by every run. The L_{sym} rule avoids this regime by assigning cellphone $L = 2$, the lowest band that expresses its groups above a 2% share, where all three seeds converge reliably without the high-band degradation.

Table 16: **Effect of target damping on DORE at cellphone $L_{\max} = 5$.** We train with and without damping using the same five random seeds, keeping all other settings fixed. Without damping, two runs degrade and three reach 78.1–78.3 accuracy; with damping, three reach 43.3–63.2 and two reach 77.1. Seed 0 degrades in both settings. CD is scaled by 10^{-3}. Training $\mathcal{L}_{\mathrm{sup}}$ (Eq. (5), scaled as in Appendix B.3) uses the final checkpoint, all training shapes, seeded rotations, and each run's own target; $l = 2$ isolates its degree-2 term.

	no damping				damping			
Seed	$\mathrm{Acc}_{\mathrm{sym}}@\epsilon \uparrow$	CD $\downarrow$	train $\mathcal{L}_{\mathrm{sup}} \downarrow$	l=2	$\mathrm{Acc}_{\mathrm{sym}}@\epsilon \uparrow$	CD $\downarrow$	train $\mathcal{L}_{\mathrm{sup}} \downarrow$	l=2
0	47.9	13.96	0.368	0.089	43.3	18.19	0.282	0.089
1	78.3	7.43	0.323	0.066	49.4	14.85	0.271	0.080
2	78.2	6.49	0.326	0.064	77.1	6.42	0.253	0.065
3	62.4	12.69	0.354	0.076	77.1	6.75	0.252	0.064
4	78.1	6.18	0.325	0.064	63.2	7.91	0.256	0.066

D.7 Predictions broader than the orbit target

In Fig. 3, the lamp prediction is a broad arc about the shape's vertical axis, whereas its target has six isolated modes; lamps with other labels behave alike (Fig. 5). To measure how close a density is to a ridge about the vertical axis, we use *azimuthal uniformity* v. For an input $\mathbf{P}R$ with top mode $\hat{R}$, we turn the shape once about that axis and divide the lowest density on this circle by the highest:

$$v = \frac{\min_\phi f(\hat{R}A_\phi)}{\max_\phi f(\hat{R}A_\phi)}, \qquad A_\phi = R^\top R(\mathbf{a}, \phi)\, R,$$

where $\mathbf{a}$ is the vertical axis of $\mathbf{P}$. A flat ridge gives $v = 1$, and deeper valleys between modes give smaller values. Truncation leaves negative side lobes (Appendix C.2), so v can be slightly negative. In Table 17, the C_6 targets have a larger v than the C_2–C_4 targets, and on lamp the C_2–C_4 predictions have a larger v than their targets. These two gaps reflect two limitations.

Finite label vocabulary. By Proposition 1, a C_6 target truncated below degree six is a flat ridge, so $v = 1.00$ on table ($L_{\mathrm{sym}} = 4$); on lamp ($L_{\mathrm{sym}} = 6$), only the degree-6 block separates its six modes, and $v = 0.64$. The lamp predictions are even closer to a ridge (median $v = 0.90$; 0.94 for the lamp of Fig. 3). The C_6-labeled lamps are nearly $SO(2)$-symmetric, and the labeling vocabulary, which ends at order six, represents this continuous symmetry with six discrete orientations (Sec. 7). The ridge therefore follows the shape's own symmetry rather than its six-fold label, and accuracy is the highest among the lamp labels (81.8%, against 14.0–60.1% on C_2–C_4).

Undetermined azimuth. The C_2–C_4 lamp predictions have median v of 0.75–0.88, against $v \leq 0.05$ for their targets. The ridge contains the target orbit but also orientations that the shape's symmetry does not allow, so the top-1 prediction takes an arbitrary azimuth. Accuracy is 60.1% on C_4, 37.2% on C_2, and 14.0% on C_3; over all test lamps it is 45.6% (Table 1). We hypothesize that the training data do not determine the azimuth of these shapes. They are few (189 of 1,854 training lamps), vary widely in form, and may not share a canonical azimuth. On table, only the C_3 predictions are broader than their targets ($v = 0.73$; 64 training shapes).

Table 17: **Azimuthal uniformity v about the vertical axis, and accuracy.** v_{target} is the target's value, v_{pred} the median over DORE's predictions, and $\text{Acc}_{\text{sym}}@\epsilon$ is in %, over the test shapes labeled C_n about the vertical axis (11–532 shapes per label), with the L_{sym} models and inputs of Table 1.

	table			lamp		
Label	v_{target}	v_{pred}	$\text{Acc}_{\text{sym}}@\epsilon$	v_{target}	v_{pred}	$\text{Acc}_{\text{sym}}@\epsilon$
C_2	−0.03	−0.03	98.8	−0.02	0.84	37.2
C_3	−0.02	0.73	72.7	−0.03	0.75	14.0
C_4	0.33	0.33	93.7	0.05	0.88	60.1
C_6	1.00	0.79	90.0	0.64	0.90	81.8

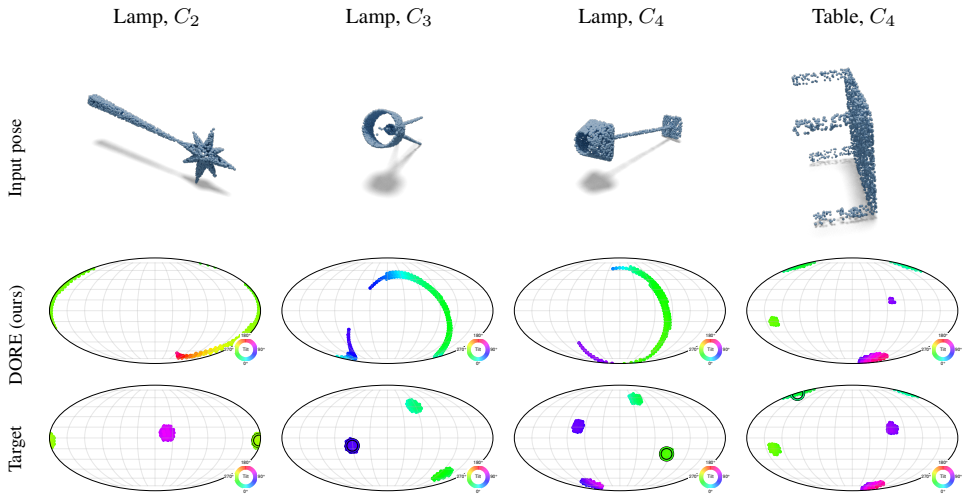

Figure 5: **Predictions about the vertical axis.** Test shapes under a random rotation, drawn as in Fig. 3. The lamp predictions are ridges, of which the display keeps the top 0.1% of grid rotations; the table prediction separates the four modes of its C_4 target. The C_2 and C_4 lamps are predicted correctly in all eleven evaluation views, and the C_3 lamp in eight, not including the one shown.

D.8 Motivation for best-of-N

The correct pose is often a lower peak. To assess whether lower peaks contain valid orientations missed by the top-1 prediction, we measure pass@k: a sample is counted as correct if any of its first k candidates is correct. We obtain the candidates using Algorithm 1, which selects up to four grid maxima at least 0.7 of the highest and applies Newton–Raphson refinement. Table 18 reports the results; pass@1 equals the accuracy in Table 1. On the symmetric categories pass@k rises with k: from 84.2 to 87.6 on table, from 45.6 to 57.4 on lamp, and from 82.4 to 89.4 on cellphone. Airplane and car gain under one point; chair gains none, as a second peak rarely qualifies.

Thus, a top-1 failure does not always mean the density missed the correct pose: among the top-1 failures, a lower peak is correct for 40% on cellphone, 22% on table and lamp, and 8–9% on airplane and car (Table 19). Most of these are half-turn flips: the first correct candidate lies at least 150° from the top-1 prediction on every such airplane, car, and cellphone sample, and on 58% and 33% of those on table and lamp. For $k > 1$, pass@k gives the oracle upper bound on selection accuracy.

Algorithm 1 Extracting density modes for best-of-N selection and pass@k evaluation

Input: coefficients $\hat{H}$, Hopf grid $\{R_i\}_{i=1}^{6,912}$ of the peak search, k $(= 4)$, ρ $(= 0.7)$;
$d(R, R') = \arccos\big((\mathrm{tr}(R^\top R') - 1)/2\big)$, the rotation angle between R and R'
Output: candidates $\hat{R}_1, \ldots, \hat{R}_n$, $n \le k$
$s_i \leftarrow f(R_i)$ for all i ▷ one matrix–vector product, as in the coarse search
$\mathcal{L} \leftarrow \{\, i : s_i \ge s_j$ for all j with $d(R_i, R_j) < 20^\circ \,\}$, sorted by decreasing s_i ▷ local maxima
$\mathcal{C} \leftarrow (\mathrm{argmax}_i\, s_i)$ ▷ the searcher's seed: $\hat{R}_1$ is the top-1 prediction
for $i \in \mathcal{L}$ **do**
  if $|\mathcal{C}| = k$ **then break**
  end if
  if $d(R_i, R_c) \ge 60^\circ$ for all $c \in \mathcal{C}$ **then** append i to $\mathcal{C}$
  end if
end for
$\mathcal{C} \leftarrow (c \in \mathcal{C} : s_c \ge \rho\, s_{\mathcal{C}_1})$ ▷ drop peaks below ρ of the highest
$\hat{R}_j \leftarrow$ Newton–Raphson refinement of $R_{\mathcal{C}_j}$ (3 steps, Appendix B.1), $j = 1, \ldots, |\mathcal{C}|$
return $\hat{R}_1, \ldots, \hat{R}_{|\mathcal{C}|}$

Best-of-N. This motivates the verifier of Sec. 5.4 (Appendix B.2), which chooses among the same candidates. It recovers 55–82% of the failures whose correct pose is among the candidates on airplane, table, lamp, and cellphone, and all of them on car, including 171 of 370 flips on table and 123 of 169 on cellphone, while turning at most 1.2% of the top-1 successes into failures, except 3.9% on lamp. It is called only when at least two peaks qualify: on at most 5% of the inputs of airplane, car, and chair, whose shapes carry at most 6.1% symmetric labels, and on 69–100% of those of table, lamp, and cellphone, whose shapes carry 40–64% non-trivial symmetry labels (Table 19).

Table 18: **Candidate coverage measured by pass@k.** A frozen test sample is correct if any of its first k candidates has $d_{\mathrm{CD}} \le \epsilon_\theta(\mathbf{P})$. Algorithm 1 selects grid maxima at least 60° apart and above 0.7 of the maximum, then refines them; samples with fewer candidates use those available. Pass@1 equals the single-pass L_{sym} accuracy in Table 1, while $k > 1$ assumes an oracle selector.

k	plane	car	chair	table	lamp	phone
1	92.6	98.3	93.8	84.2	45.6	82.4
2	93.2	98.5	93.8	86.8	51.5	85.9
3	93.2	98.5	93.8	87.5	55.3	87.8
4	93.2	98.5	93.8	87.6	57.4	89.4

Table 19: **Single-pass failures and best-of-N recovery.** All values are percentages. Failures are incorrect top-1 predictions; recoverable is the fraction of failures with a correct lower candidate (Table 18). Among recoverable failures, flips have their first correct candidate at least 150° from the top-1 prediction, and recovered are corrected by the verifier. Broken is the fraction of top-1 successes turned into failures. Called is the fraction of inputs with at least two candidates; symmetric is the fraction of shapes with non-identity labels (Table 24).

	failures	recoverable	flips	recovered	broken	called	symmetric
plane	7.4	9.2	100	82	0.0	4.8	6.1
car	1.7	7.8	100	100	0.0	0.5	0.6
chair	6.2	0.0	–	–	0.0	0.2	0.3
table	15.8	21.6	58	55	1.2	68.8	43.1
lamp	54.4	21.7	33	57	3.9	81.7	39.7
phone	17.6	40.0	100	73	1.2	100.0	64.1

D.9 Settings for using the full density at test time

We detail the extraction, verification, and averaging procedures used in Sec. 5.4.

Best-of-N: candidates. Algorithm 1 extracts candidate orientations from the predicted density. On the 6,912-point Hopf grid, a local maximum must score at least as high as every grid point within 20°. We greedily retain up to four maxima in descending score order, requiring at least 60° separation and a score of at least 0.7 times the highest. Each retained candidate receives three Newton–Raphson steps. When only one candidate remains, we return it without verification.

Best-of-N: verifier. Each candidate R_k aligns the 2,000-point cloud, which we center and scale to unit radius. The DGCNN [45] backbone from the flipper of Scarvelis et al. [39] scores this canonicalized cloud. We add $\alpha \log(f_k/f_1)$, where f_k is the density at R_k and α is learned, then select the candidate with the highest combined score.

Best-of-N: training. The verifier of each category is trained on DORE's candidates for the training shapes, built like test inputs, and on synthetic candidates: the shape in its canonical pose and three rotated versions, using rotations about the up axis or half-turns about horizontal axes, sampled more often from transform types the verifier gets wrong. The target is uniform over the candidates within $\epsilon_\theta(\mathbf{P})$ of the canonical shape, and the loss is its cross-entropy with the softmax of the scores, plus a term preferring the candidate within 20° of the canonical pose. Half of the real training samples contain both correct and incorrect candidates, so verifier selection can change the outcome. Because such samples are scarce, we set the training duration by how often each is sampled. These settings were chosen on a validation split of 10% of the training shapes, and the final verifiers are trained on all training shapes. The category-specific training settings appear in Table 20.

Table 20: **Training settings for the best-of-N verifier.** Each category has its own verifier, trained to select among the orientation candidates produced by DORE.

Backbone	EdgeConv with k=20 and widths $64, 64, 128, 256$; a 1×1 convolution to 1,024; max and mean pooling; an MLP $2048 \to 512 \to 256 \to 1$; 1.8M parameters
Real samples	a 2,000-point pool drawn from the 16,384-point cloud, a random rotation, and DORE's candidates from its first 1,024 points (M=1) or from M=8 averaged inputs; four draws per shape (eight for cellphone); at every step the 2,000 points are redrawn and a rotation of at most 3° is applied to all candidates of a sample
Synthetic samples	the canonical pose and three transforms of it, rotations by $360^\circ k/n$ about the up axis and 180° flips about horizontal axes ($n = 2, \ldots, 6$), each perturbed by at most 3°
Batch	32 samples; M=1, M=8 and synthetic samples in the ratio 40:40:20; each M=1 or M=8 sample has, with probability 0.5 each, both correct and wrong candidates or only correct ones
Loss	over samples with at least two candidates and at least one with $r_k = d_{\mathrm{CD}}/\epsilon_\theta(\mathbf{P}) \le 1$; weight 0.2 on the term preferring the candidate within 20° of the canonical pose
Shape augmentation	none for airplane, car, and chair; cellphone: per-axis scaling in $[0.9, 1.1]$; lamp: one scale for x, y and one for z in $[0.8, 1.25]$, tapering ($\lvert t\rvert \le 0.2$), and a smooth local deformation; table: scaling of z in $[0.8, 1.25]$ and tapering; ϵ_θ is rescaled with the shape
Optimization	AdamW (learning rate 10^{-3}, weight decay 10^{-4}), cosine schedule with 5% warm-up, final checkpoint; the length draws each sample with both correct and wrong candidates about 22 times, with at least 4,000 steps and epochs of at least 100 steps, rounded up to a multiple of 5 epochs: 40 for airplane, car, and chair, 15 for table, 55 for lamp, 20 for cellphone

Coefficient averaging. We average predicted coefficients from M input subsamples before a single peak search. By linearity, this averages the corresponding densities (Appendix B.2). We construct the inputs from a shared pool of 2,000 points from the same object as follows.

Input 1 is the model's own input; inputs 2 and 3 are other random 1,024-point subsets of these points; from input 4 on, to vary the inputs beyond such subsets, a growing fraction of the subset (0.1, 0.2, …, at most 1) is moved toward one of its 8 nearest neighbors among the 2,000 points, by a random fraction of at most 0.5 of their distance. The neighbor is restricted to the same side of the surface, which the normals decide; the normals are used for nothing else. Table 21 reports $M = 1, 2, 4, 8$:

the six-category mean accuracy rises from 82.8 to 83.5, 84.0, and 84.3, while the geometric mean of Chamfer distance across the six categories falls from 3.76 to 3.33.

Table 21: **Effect of the number of inputs on coefficient averaging.** We average predicted coefficients across M inputs per object using the models and evaluation protocol of Table 1. Inputs are constructed from the same 2,000-point cloud (Sec. 5.4). The mean column uses the geometric mean for Chamfer distance and the arithmetic mean for accuracy.

M	Mean Chamfer distance ($\times 10^{-3}$) $\downarrow$							$\mathrm{Acc}_{\mathrm{sym}}@\epsilon$ (%) $\uparrow$						
	plane	car	chair	table	lamp	phone	mean	plane	car	chair	table	lamp	phone	mean
1	1.54	1.20	2.14	5.09	28.23	4.99	3.76	92.6	98.3	93.8	84.2	45.6	82.4	82.8
2	1.43	1.12	1.88	4.89	27.71	4.82	3.53	93.6	98.8	94.6	84.7	46.5	83.0	83.5
4	1.30	1.08	1.76	4.76	27.62	4.76	3.40	94.0	98.9	95.1	85.0	47.7	83.6	84.0
8	1.23	1.07	1.71	4.74	27.16	4.69	3.33	94.5	99.0	95.2	85.2	48.4	83.5	84.3

E Symmetry labeling

E.1 Symmetry group labeling algorithm

Algorithm 2 combines axis search, dihedral checks, and group selection into the labeling procedure described in Sec. 4.5. Appendix B.3 lists its hyperparameters and implementation settings.

Algorithm 2 Automatic symmetry group labeling through axis search and geometric verification

Input: $\mathbf{P}$, orders $\mathcal{N}$, params $(k, m, L, K, M, \{\Theta_t\}, \tau, \beta, r, \epsilon_D)$
Output: G^*
Center $\mathbf{P}$; build KD-tree $\mathcal{T}$
for $n \in \mathcal{N}$ **do**
    $\mathcal{B}_0 \leftarrow$ top-k diverse axes from M Fibonacci samples
    for $t = 1, \ldots, K$ **do**
        Retain top-m elites from $\mathcal{B}_{t-1}$
        Sample L axes within radius Θ_t per member
        $\mathcal{B}_t \leftarrow$ elites $\cup$ top-$(k-m)$ new candidates
    end for
    $\hat{a}_n \leftarrow \mathrm{argmax}_{a \in \mathcal{B}_K} S(a, n)$
end for
for $n \in \mathcal{N}$ **do** ▷ Dihedral post-check
    Evaluate 180° flip scores on n perpendicular C_2 axes
    Mark D_n plausible if flip score $\geq S(\hat{a}_n, n) \cdot r - \epsilon_D$
end for
$e_n \leftarrow -S(\hat{a}_n, n)$; $\mathcal{N}' \leftarrow \{n \in \mathcal{N} : n \geq 3,\ e_n \leq r \min_m e_m + \epsilon_D\}$, or $\mathcal{N}$ if empty
$\mathcal{G}' \leftarrow \{C_n$ and plausible $D_n : n \in \mathcal{N}'\}$
if $\max_{G \in \mathcal{G}'} S(G) < \tau$ **then**
    return $G^* = \{e\}$
end if
return $G^* = \mathrm{argmax}_{G \in \mathcal{G}'} \big[S(G) + \beta \log |G|\big]$

E.2 Symmetry group labeling validation

We audit the automatic labels on 200 randomly sampled ShapeNet point clouds per class. The first step checks the global threshold; the second tests the geometric consistency of the assigned groups.

Step 1: Threshold calibration. We compare the global threshold $\tau = -0.075$ with an object-specific reference $\tau_\theta(\mathbf{P})$. Like $\epsilon_\theta(\mathbf{P})$ in Eq. (12), this reference measures the expected effect of a θ misalignment. We estimate its normalized one-sided Chamfer score over 50 axes,

$$\tau_\theta(\mathbf{P}) = \mathbb{E}_{\mathbf{u} \sim \mathcal{U}(S^2)} \left[-d_{\mathbf{P}}^{\mathrm{norm}}\big(\mathbf{P} R(\mathbf{u}, \theta)\big) \right], \tag{59}$$

where $d_{\mathbf{P}}^{\text{norm}}$ is the normalized unsquared distance of Appendix B.3.

For each class and $\theta \in \{10^\circ, 20^\circ\}$, Table 22 reports the mean and standard deviation of $\tau_\theta(\mathbf{P})$ and the fraction of objects for which the global threshold is stricter (less negative).

At $\theta = 20^\circ$, the global threshold $\tau = -0.075$ is stricter than the per-object reference for every object of every class. At $\theta = 10^\circ$, it remains stricter for 96–100% of samples in twelve classes, but only 61.5% on lamp, whose reference scores lie near τ (-0.078 ± 0.010). Thus, it is generally, but not uniformly, conservative relative to these references and may reject genuine symmetries.

Table 22: **Global versus object-specific labeling thresholds.** We compare $\tau = -0.075$ with $\tau_\theta(\mathbf{P})$ using scores $-d_{\mathbf{P}}^{\text{norm}}(\cdot)$ (Appendix B.3). The *%stricter* column counts objects with $\tau > \tau_\theta(\mathbf{P})$. Because the existence test accepts a non-trivial symmetry only when the best candidate score exceeds the threshold, a less negative threshold is more conservative.

Class	θ	N	Mean $\tau_\theta(\mathbf{P})$	Std $\tau_\theta(\mathbf{P})$	%stricter	%looser
Airplane	10°	200	-0.091	0.003	100.0%	0.0%
	20°	200	-0.163	0.006	100.0%	0.0%
Bench	10°	200	-0.085	0.006	**96.0%**	4.0%
	20°	200	-0.147	0.009	100.0%	0.0%
Cabinet	10°	200	-0.091	0.003	100.0%	0.0%
	20°	200	-0.138	0.006	100.0%	0.0%
Car	10°	200	-0.089	0.002	100.0%	0.0%
	20°	200	-0.135	0.004	100.0%	0.0%
Chair	10°	200	-0.092	0.003	100.0%	0.0%
	20°	200	-0.152	0.011	100.0%	0.0%
Monitor	10°	200	-0.091	0.003	100.0%	0.0%
	20°	200	-0.150	0.010	100.0%	0.0%
Lamp	10°	200	-0.078	0.010	**61.5%**	38.5%
	20°	200	-0.136	0.019	100.0%	0.0%
Speaker	10°	200	-0.090	0.004	**99.0%**	1.0%
	20°	200	-0.134	0.008	100.0%	0.0%
Firearm	10°	200	-0.083	0.004	**99.0%**	1.0%
	20°	200	-0.161	0.010	100.0%	0.0%
Couch	10°	200	-0.089	0.004	100.0%	0.0%
	20°	200	-0.136	0.005	100.0%	0.0%
Table	10°	200	-0.090	0.005	**98.5%**	1.5%
	20°	200	-0.149	0.013	100.0%	0.0%
Cellphone	10°	200	-0.089	0.003	100.0%	0.0%
	20°	200	-0.147	0.007	100.0%	0.0%
Watercraft	10°	200	-0.087	0.004	100.0%	0.0%
	20°	200	-0.145	0.008	100.0%	0.0%

Step 2: Geometric consistency of assigned groups. For all samples labeled with a non-trivial symmetry group ($G^* \neq \{e\}$), we verify that the mean orbit Chamfer score exceeds the reference score obtained by rotating the same object by 10° about uniformly sampled axes,

$$\frac{1}{|G^*|} \sum_{g \in G^*} d_{\mathbf{P}}^{\text{norm}}(g.\mathbf{P}) \;<\; |\tau_{10^\circ}(\mathbf{P})|. \tag{60}$$

All 593 samples labeled symmetric across the 13 audited classes pass this check (Table 23). The mean orbit score (-0.040) is about half the mean $\tau_{10^\circ}(\mathbf{P})$ (-0.084) in magnitude. Without human labels, the audit establishes consistency with an independent geometric threshold only.

Table 23: **Geometric consistency of assigned symmetry groups.** For each class, we report the number of symmetric samples, the fraction with orbit score above $\tau_{10^\circ}(\mathbf{P})$, and the mean orbit score and reference threshold. This checks geometric consistency without human symmetry annotations.

Class	#total	#sym	%sym	%pass	Mean orbit	Mean $\tau_{10^\circ}(\mathbf{P})$
Airplane	200	14	7.0%	100.0%	−0.037	−0.088
Bench	200	69	34.5%	100.0%	−0.031	−0.082
Cabinet	200	10	5.0%	100.0%	−0.040	−0.088
Car	200	2	1.0%	100.0%	−0.051	−0.084
Chair	200	0	0.0%	N/A	N/A	N/A
Monitor	200	37	18.5%	100.0%	−0.042	−0.091
Lamp	200	89	44.5%	100.0%	−0.040	−0.074
Speaker	200	21	10.5%	100.0%	−0.045	−0.084
Firearm	200	77	38.5%	100.0%	−0.040	−0.083
Couch	200	0	0.0%	N/A	N/A	N/A
Table	200	98	49.0%	100.0%	−0.036	−0.089
Cellphone	200	133	66.5%	100.0%	−0.046	−0.088
Watercraft	200	43	21.5%	100.0%	−0.049	−0.084
Total	2600	593	22.8%	100.0%	−0.040	−0.084

E.3 Symmetry labeling statistics

Table 24 reports the per-class, per-group breakdown of inferred labels over every shape of all 13 categories ($n = 39{,}739$; train and test splits). Symmetry is strongly category-dependent: cellphone (64.1% non-trivial), table (43.1%), lamp (39.7%), and firearm (35.5%) are the most symmetric, while car, chair, and couch are nearly asymmetric. Among symmetric shapes, C_2 dominates in airplane and bench (an approximate half-turn); cellphone D_2 labels come from handsets symmetric front-back and left-right; and the high-order labels (C_6 in lamp and firearm, D_6 in watercraft) set $L_{\mathrm{sym}} = 6$ for these three categories (Eq. (9)). The table category mixes C_2 (rectangular tops), C_4 (square tops), and C_6; lamp carries C_6 on 624 of its 2,318 shapes, consistent with circular shades. Rare high-order cyclic labels such as C_6 in airplane correspond to near-axisymmetric, rocket-shaped outliers whose continuous rotational symmetry is well-approximated by a discrete cyclic group. In the audit of Appendix E.2, 593 of the 2,600 sampled shapes (22.8%; 336 of 1,200 over the six benchmark categories) receive a non-trivial symmetry label, and all pass its geometric consistency check.

Table 24: **Symmetry labels across 13 ShapeNet categories.** The census covers all 39,739 shapes in the training and test splits. Band selection uses training fractions p_G only (Eq. (9)); these choose the same L_{sym} bands as the full-census fractions shown here.

Class	n	$\{e\}$	C_2	C_3	C_4	C_5	C_6	D_2	D_3	D_4	D_5	D_6
Airplane	4045	3797	240	1	2	–	2	3	–	–	–	–
Bench	1813	1204	587	1	3	–	3	12	1	1	–	1
Cabinet	1571	1501	40	–	1	–	–	24	–	5	–	–
Car	3514	3493	18	–	–	–	–	1	–	1	–	1
Chair	6778	6760	14	1	3	–	–	–	–	–	–	–
Couch	3173	3157	13	–	–	–	1	1	–	1	–	–
Firearm	2373	1531	410	–	1	–	380	30	–	–	–	21
Lamp	2318	1398	117	40	94	12	624	2	1	9	2	19
Monitor	1093	881	115	–	–	–	–	95	–	2	–	–
Cellphone	1089	391	278	–	–	–	–	417	–	3	–	–
Speaker	1597	1482	48	1	12	2	24	15	–	11	–	2
Table	8436	4796	2774	82	577	10	157	30	–	6	–	4
Watercraft	1939	1550	184	–	–	–	23	15	–	7	–	160

Absence of exact polyhedral symmetry. Our C_n/D_n labeler (Sec. 4.5) may omit polyhedral groups, so we test all 7,798 non-trivially symmetric shapes for T, O, and I (Table 25). For each shape and group, we optimize the frame and divide the full-orbit mean Chamfer distance by a random-

Table 25: **Polyhedral-symmetry scan with optimized group frames.** Every non-trivially symmetric benchmark shape is tested. Cells give the lowest r/r_0 per category and group, with $n_{\neq e}$ counting the shapes not labeled $\{e\}$ (Table 24). Exact symmetry approaches the sampling floor, $r/r_0 \approx 1$; synthetic solids yield 0.94–1.16.

	Airpl.	Bench	Cab.	Car	Cell.	Chair	Couch	Fire.	Lamp	Mon.	Spk.	Table	Wat.	All
$n_{\neq e}$	248	609	70	21	698	18	16	842	920	212	115	3640	389	7798
best T	5.12	4.79	4.88	6.11	5.98	4.42	5.63	8.18	1.31	5.70	4.61	2.22	5.85	1.31
best O	4.34	5.22	5.08	6.32	6.29	4.53	5.85	8.40	1.35	5.46	5.20	2.48	6.01	1.35
best I	6.15	5.36	5.20	6.39	6.46	4.76	5.93	8.62	1.38	6.05	5.29	4.02	6.89	1.38

Table 26: **Closest polyhedral candidates from Table 25.** Shapes are ranked by their lowest r/r_0 over T, O, and I, then rescanned at 1,024 points (three-subsample mean) and 8,192 points. Genuine symmetry remains near one at both densities, as illustrated by the first row.

	Shape	Group	Scan	1,024 pts	8,192 pts
–	synthetic T, O, I solids	own	–	0.94–1.16	0.97–1.13
1	lamp, spherical shade	T	1.31	1.14	2.11
2	lamp, icosahedral pendant	T	1.43	1.07	1.71
3–20	13 tables, 5 lamps		2.22–3.69	2.19–4.36	4.66–9.20

rotation baseline to obtain r. We compare it with the sampling floor r_0, the mean nearest-neighbor spacing divided by the same baseline. Exact symmetry should give $r/r_0 \approx 1$: synthetic T, O, and I solids sampled as the benchmark yield 0.94–1.16 at 1,024 points and 0.97–1.13 at 8,192 points.

The benchmark medians are 9.2, 9.5, and 9.8 for T, O, and I; all categories except lamp and table stay above 4. At 8,192 points, the next 18 candidates rise from 2.2–3.7 to 4.7–9.2, away from the sampling floor (Table 26).

Only a wire-frame icosahedral pendant and a spherical shade score at or below 1.5, rising to 1.7–2.1 at 8,192 points. Their cords restrict symmetry to rotations about them. The labeler assigns C_5, the pendant's group type, and C_6, the largest finite stand-in for the shade's continuous symmetry, but about axes other than the cords. The scan finds no exact polyhedral symmetry among these shapes, although the band expresses T, O, and I (Appendix F.5).

A near-icosahedral pendant. Figure 6 compares the pendant's predicted density with its C_5 target and an icosahedral target. The prediction uses the input scale check of Appendix B.4. Under three random input rotations, the aligned predictions differ from the unrotated prediction by at most 1.4% in relative coefficient ℓ_2 norm ($l \geq 1$), but form a ridge matching neither the C_5 target (coefficient cosine 0.06 over $l \geq 1$) nor the icosahedral orbit (0.00). This is a training shape; successful test lamps reach a median cosine of 0.70 (interquartile range 0.44–0.92). Its C_5 target uses another icosahedral vertex axis rather than the cord, leaving little input cue for that axis.

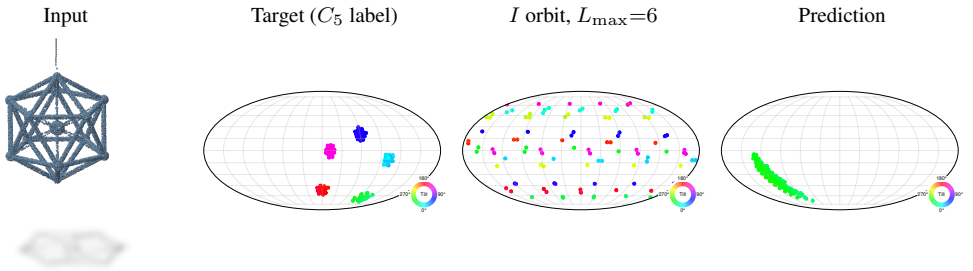

Figure 6: **Failure on a near-icosahedral pendant.** From left: a training lamp, its C_5 target, the icosahedral target in the same frame at $L_{\max} = 6$, and the lamp model's prediction (Table 1). All densities are shown in the canonical frame. The predicted ridge matches neither target.

F Further evaluation and visualization

F.1 Baseline at matched latency

Table 3 compares best-of-N and coefficient averaging against the default 50-vote test-time augmentation of Scarvelis et al. [39], which takes about $12\times$ longer. The baseline's latency grows with the number of votes (7.1 ms for one, 353.2 ms for 50), so we also evaluate it with four votes, whose latency matches that of best-of-N and coefficient averaging, and with eight (Table 27). In both settings, the orienter and flipper each use the stated number of votes. With four votes, the baseline's mean accuracy is 81.0% against 85.5%, lower on every category; its Chamfer distance is higher on four of the six categories, the exceptions being table and lamp, where it is also lower with 50 votes. To assess whether the matched-latency gap stems from the allocation of votes between the orienter and the flipper, we additionally evaluate eight votes for each network. Their per-pass inference costs are similar (3.6 and 3.5 ms), so this setting increases the computation available to both networks rather than favoring either through reallocation. At about twice our latency, the baseline reaches 83.2% mean accuracy, below our 85.5%. This result suggests that the gap is not simply due to an unfavorable allocation of the matched-latency budget. Both settings use the checkpoints, inputs, and protocol of Table 1.

Table 27: **Baseline at matched latency.** Scarvelis et al. [39] with four and eight votes against best-of-N with coefficient averaging (Table 3). Evaluation and timing follow Appendix B.4. Bold: best; underline: second best.

Method	Mean Chamfer distance ($\times 10^{-3}$) ↓						$\mathrm{Acc}_{\mathrm{sym}}@\epsilon$ (%) ↑						Latency (ms) ↓
	plane	car	chair	table	lamp	phone	plane	car	chair	table	lamp	phone	
Scarvelis et al. [39] (4 votes)	2.2	2.2	2.8	4.2	24.6	6.1	91.4	97.5	91.3	84.9	43.3	77.5	**28.3**
Scarvelis et al. [39] (8 votes)	1.6	2.0	2.5	**3.9**	**21.5**	5.4	94.0	98.9	92.2	**86.1**	47.8	80.3	56.6
DORE + BoN + CA	**1.2**	**1.1**	**1.7**	4.5	26.5	**4.1**	**94.9**	**99.1**	**95.2**	**86.1**	**51.3**	**86.6**	28.5

F.2 Up-axis metrics

Scarvelis et al. [39] report a strict up-axis accuracy, following Pang et al. [31]; we additionally report an *antipodal* variant that also accepts up/down flips. For rotation k, the predicted and true up axes form the angle θ_k,

$$\cos\theta_k = \left\langle \hat{R}_k^\top e_z,\ R_k^\top e_z \right\rangle = \left(R_k \hat{R}_k^\top\right)_{33}, \tag{61}$$

with R_k the applied rotation, $\hat{R}_k$ the predicted rotation, and e_z the canonical up axis (the benchmark frame is z-up), so that $R_k^\top e_z$ is the up direction of the rotated cloud $\mathbf{P}R_k$. Using the K test rotations of Eq. (11), we compute the strict and antipodal accuracies as follows:

$$\mathrm{up}^{\mathrm{strict}}@10 = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}\left[\theta_k < 10^\circ\right], \qquad \mathrm{up}^{\mathrm{anti}}@10 = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}\left[\min(\theta_k,\ 180^\circ - \theta_k) < 10^\circ\right], \tag{62}$$

which differ only on predictions within 10° of the flipped up axis $-R_k^\top e_z$. Unlike $\mathrm{Acc}_{\mathrm{sym}}@\epsilon$, neither accounts for the object's full symmetry group (Table 28).

Table 28: **Strict and antipodal up-axis accuracy@10°.** All rows use the runs, shapes, and frozen rotations of Table 1. Results with TTA or with best-of-N and coefficient averaging follow Table 3, and latency follows Appendix B.4. Bold: best; underline: second best.

Method	up@10° strict (%) ↑						up@10° antipodal (%) ↑						Latency (ms) ↓
	plane	car	chair	table	lamp	phone	plane	car	chair	table	lamp	phone	
CHOIR [21]	0.0	0.0	0.0	0.2	0.4	1.3	0.0	0.0	0.1	0.2	0.4	2.4	**2.8**
CHOIR† [21]	56.1	95.7	89.5	75.4	30.6	1.4	56.5	95.7	89.5	75.7	33.7	3.5	**2.8**
Scarvelis et al. [39]	85.4	90.5	91.6	95.4	61.1	53.0	87.2	90.5	91.8	96.2	67.2	71.6	7.1
Scarvelis et al. [39] + TTA	**97.1**	**99.9**	96.4	**98.6**	**75.4**	**69.0**	**98.3**	**99.9**	96.6	**99.1**	**82.4**	88.2	353.2
DORE	93.6	99.4	96.7	94.3	63.5	40.9	97.3	99.4	96.9	96.0	74.3	**89.2**	3.0
DORE + BoN + CA	95.0	99.5	**97.1**	95.0	66.9	56.8	98.2	99.5	**97.4**	96.7	77.4	88.8	28.5

On the antipodal criterion, the six-category mean of DORE is 92.2, against 84.1 for Scarvelis et al. [39] without TTA and 94.1 with it. On the strict criterion the baseline with TTA leads by more (89.4

vs. 81.4), mostly on cellphone (69.0 vs. 40.9). DORE keeps the orbit-equivalent modes and reads out the highest, so more of its samples invert the up axis (strict failure, antipodal success): 48.2% on cellphone against 19.1% for the baseline with TTA, whose flipper is trained to reproduce the dataset's up direction. Best-of-N with coefficient averaging chooses among DORE's peaks and cuts these to 32.0%, raising the strict mean to 85.1 (cellphone 56.8).

Neither criterion accounts for each shape's symmetry, which decides whether an up/down flip is valid. On 187 of the 218 cellphone test shapes, a half-turn about a horizontal axis changes the cloud by less than $\epsilon_\theta(\mathbf{P})$, so 87% of the inverted cellphone samples of either method are correct poses under $\mathrm{Acc}_{\mathrm{sym}}@\epsilon$. An upside-down chair is a wrong pose, yet the antipodal variant counts it as a success. We report both criteria for comparison with the baseline; our main metric respects each shape's geometric symmetry.

F.3 Zero-shot transfer to ModelNet-40

We apply the ShapeNet models from Table 1 to the five shared ModelNet-40 [46] categories without retraining. ModelNet-40 aligns mesh up axes but not azimuths, so we report up-axis accuracy (Appendix F.2) instead of the Chamfer-based metrics; it evaluates two of the three rotational degrees of freedom.

Table 29: **Zero-shot transfer to ModelNet-40 [46].** ShapeNet models from Table 1 are evaluated without retraining. TTA uses 50 votes per network; best-of-N with coefficient averaging follows Table 3. Each test shape uses 16 fixed rotations, with 100 shapes per category except 20 for lamp. Latency follows Appendix B.4 on these inputs. Bold: best; underline: second best.

Method	up@10° strict (%) ↑					up@10° antipodal (%) ↑					Latency (ms) ↓
	plane	car	chair	table	lamp	plane	car	chair	table	lamp	
CHOIR [21]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**2.8**
CHOIR† [21]	78.0	96.0	89.0	95.0	25.0	78.0	96.0	89.0	95.0	25.0	**2.8**
Scarvelis et al. [39]	89.4	86.3	88.3	96.2	75.6	89.6	86.3	88.8	96.3	80.6	7.1
Scarvelis et al. [39] + TTA	**98.0**	**99.9**	**95.3**	99.0	**84.7**	**98.0**	**99.9**	**95.3**	99.0	**90.0**	353.6
DORE	**98.0**	98.0	94.0	**100.0**	73.1	**98.0**	98.0	94.0	**100.0**	84.1	3.1
DORE + BoN + CA	**98.0**	98.5	94.1	**100.0**	74.4	**98.0**	98.5	94.1	**100.0**	85.0	27.1

DORE places the up axis within 10° on 94–100% of the airplane, car, chair, and table samples and on 73.1% of lamp, also its weakest category in-distribution (Tables 29 and 1). Only on lamp do DORE's strict and antipodal accuracies differ (73.1 vs. 84.1), so about two fifths of its errors invert a correctly placed axis. On the other four categories, adding best-of-N and coefficient averaging changes DORE's accuracy by at most 0.5 pp. CHOIR's raw predictions never place the up axis (0.0%); with its ShapeNet anchor R_{calib}, not refitted on ModelNet, CHOIR† reaches 78.0–96.0% on airplane, car, chair, and table and 25.0% on lamp.

We evaluate each method in its learned coordinate convention. For Scarvelis et al. [39], inputs are rotated by $R_x(-90°)$ and the predicted up axis (third column) is negated; this single global convention, not a per-object correction, reduces its median up-error from 175–179° to 1–5°. The baseline reaches 75.6–96.2% (strict) without TTA and 84.7–99.9% with it; on lamp it leads both without (75.6 vs. 73.1) and with additional test-time computation (84.7 vs. 74.4).

F.4 Band-limit gallery

Figures 7–10 show, for every symmetry type in the labeling vocabulary, one representative shape and its target density truncated at each band limit $L_{\max} \in \{1, \ldots, 6\}$; $L_{\max} = 1$ is the setting of the "w/o higher-degree coefficients" row of Table 2. Below a group's first expressible degree, the discrete orbit becomes indistinguishable from a continuous symmetry fiber. The density then equals the $SO(2)$ target for C_n or its flip-symmetrized $O(2)$ counterpart for D_n (Proposition 1). For consistent comparison across bands, every example is a lamp aligned with its symmetry axis along $+z$.

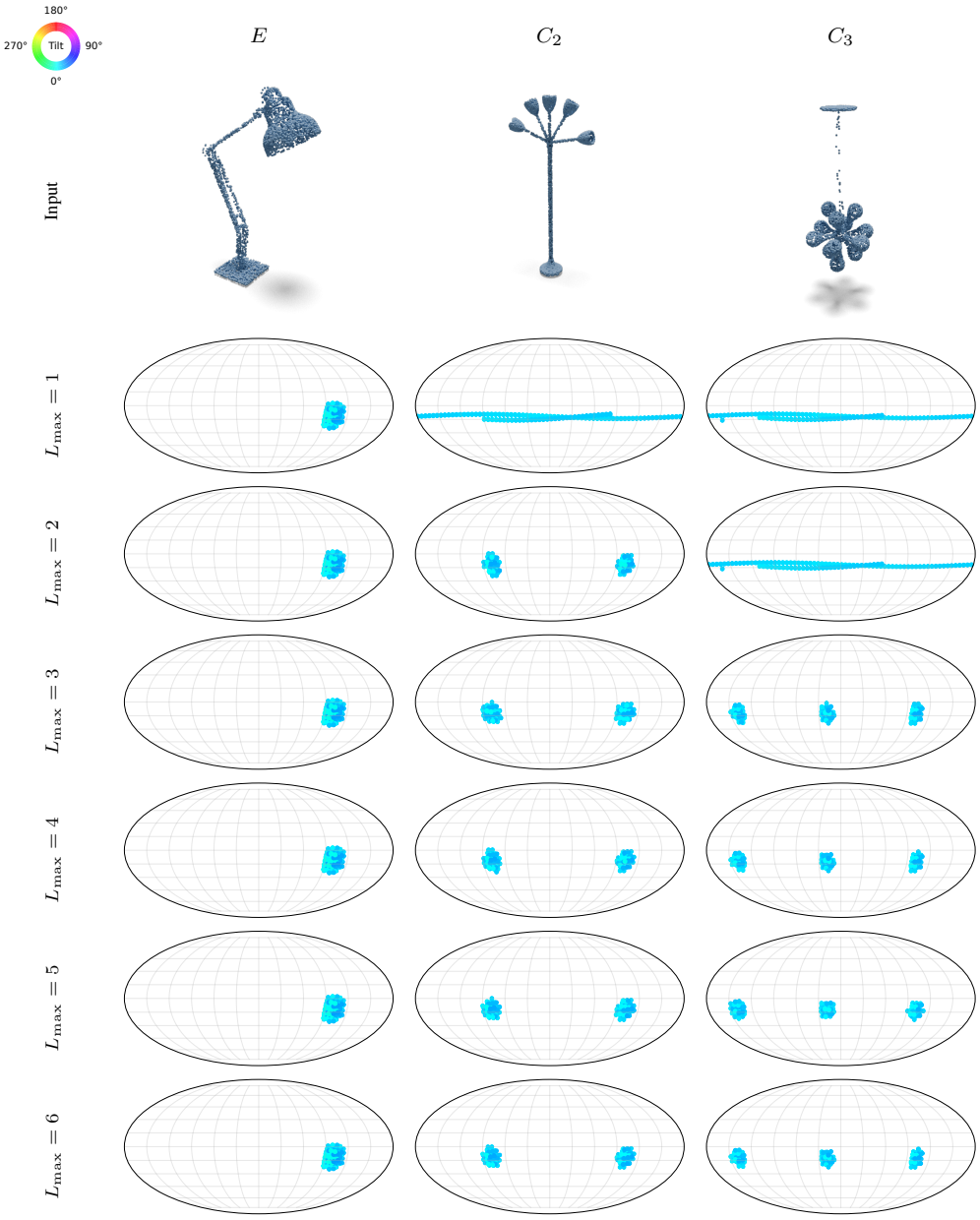

Figure 7: **Band-limit gallery (1/4): E, C_2, and C_3.** The rendering follows Fig. 3. Higher bands sharpen the single mode of E; C_2 and C_3 resolve at $L_{\max} = 2$ and 3.

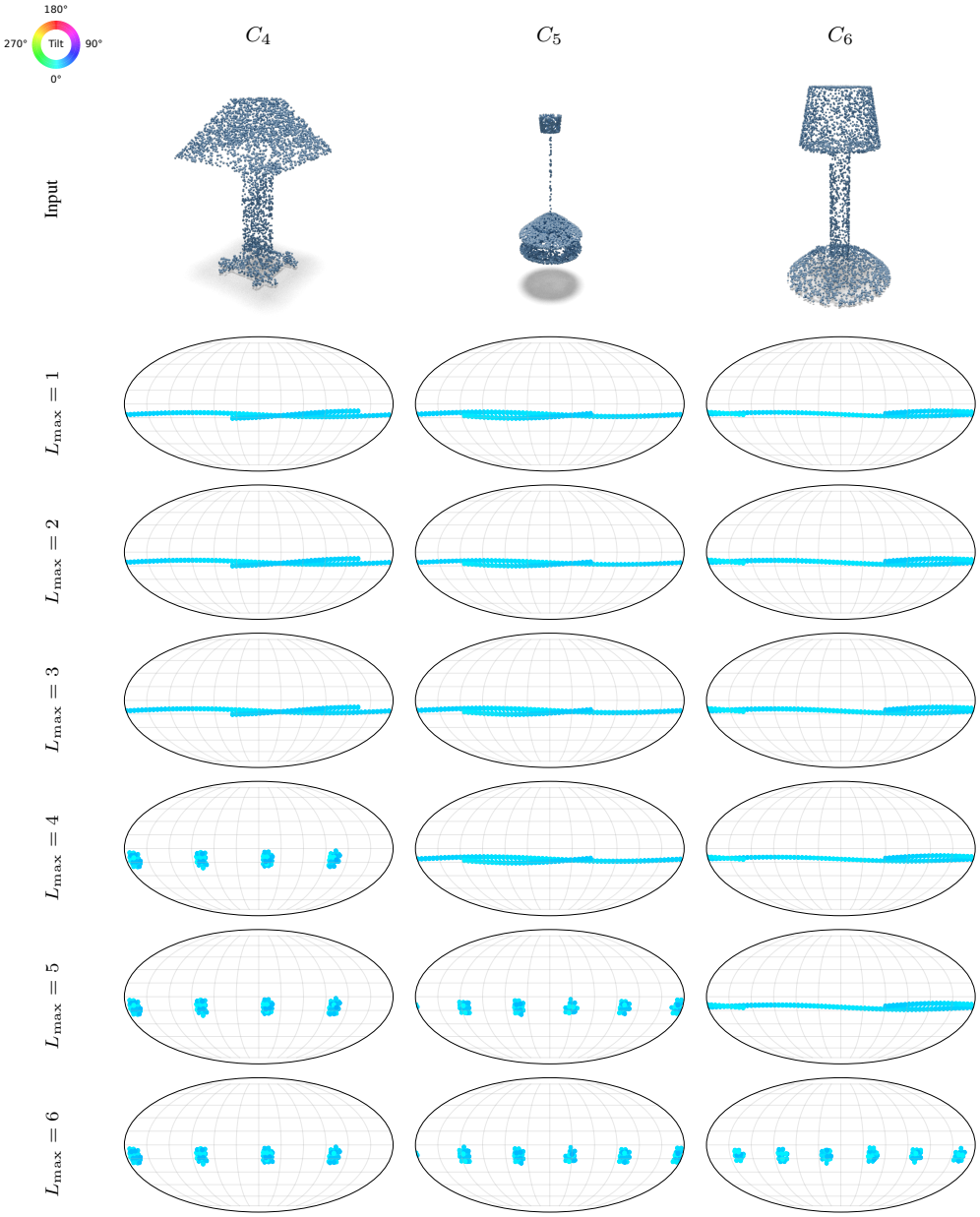

Figure 8: **Band-limit gallery (2/4): C_4, C_5, and C_6.** Each C_n orbit first resolves into n peaks at $L_{\max} = n$. Thus, the six modes of C_6 appear only in the final row.

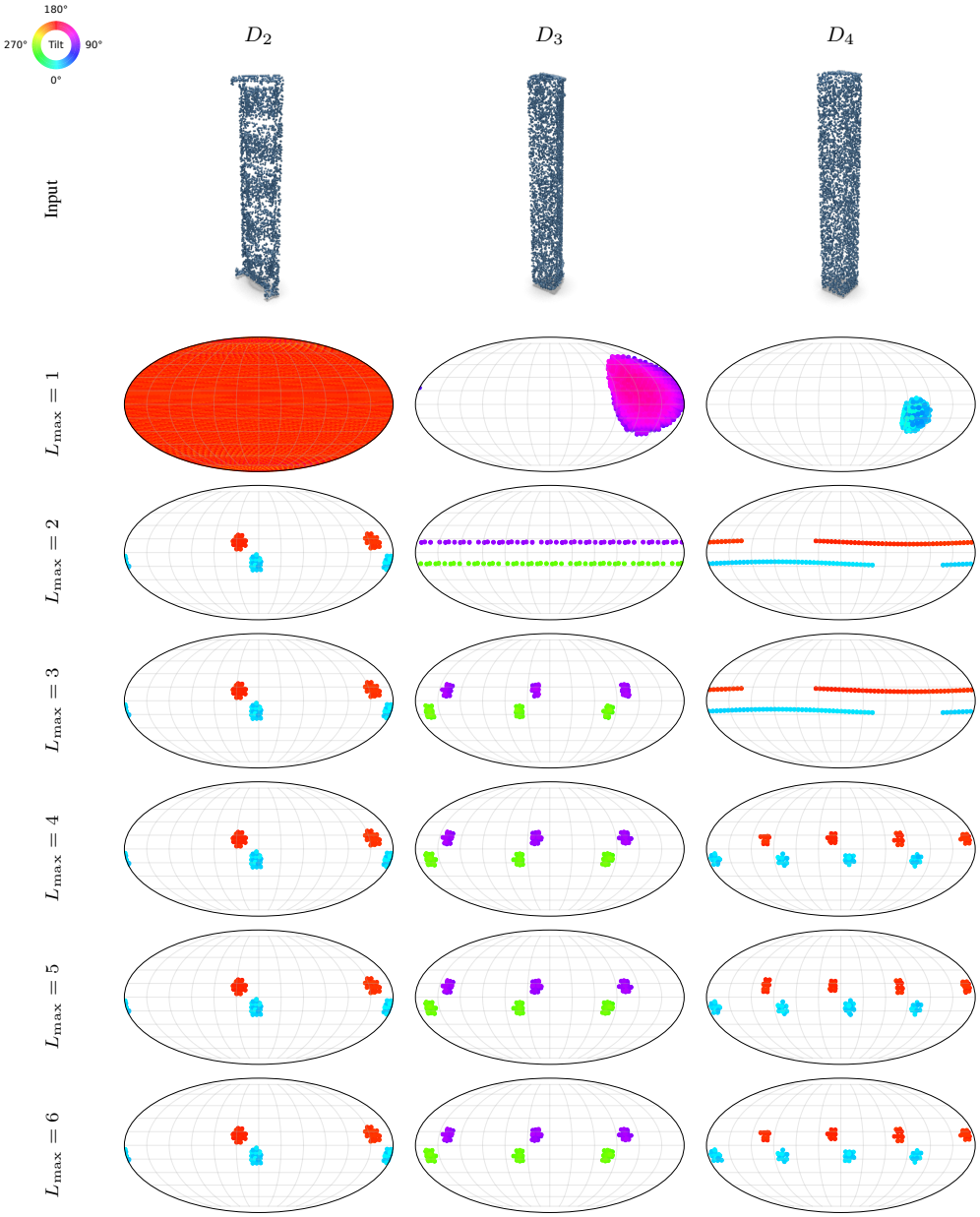

Figure 9: **Band-limit gallery (3/4): D_2, D_3, and D_4.** Adding a flip doubles the cyclic orbit. Below $L_{\max} = n$, the target equals the truncated $O(2)$ density; at $L_{\max} = 1$, it is uniform because $m_1 = 0$ (Table 4). Figure 11 explains the rendering of uniform targets.

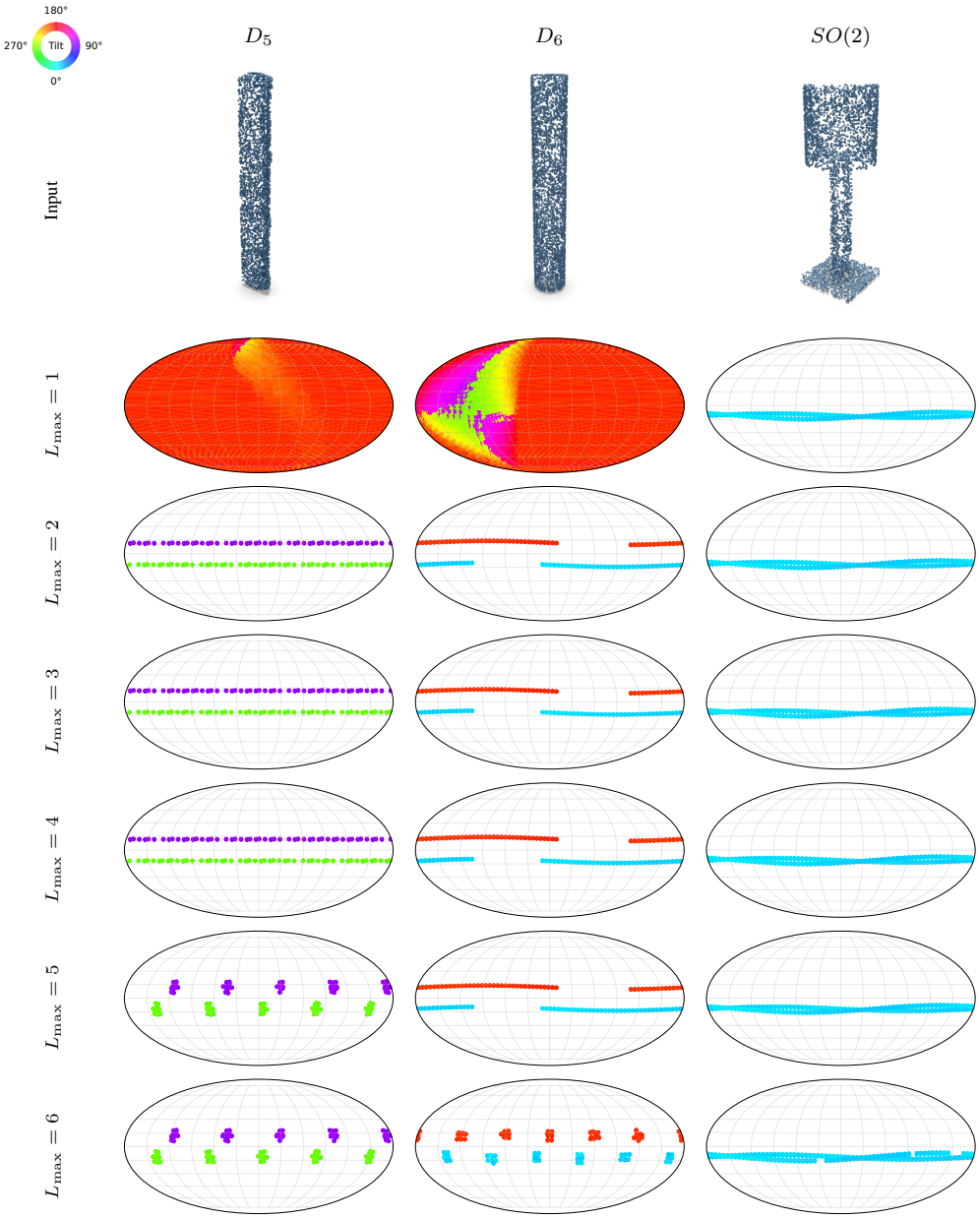

Figure 10: **Band-limit gallery (4/4): D_5, D_6, and $SO(2)$.** The final column shows the analytic $SO(2)$ target using the round-shade lamp from Fig. 3. Our finite vocabulary assigns this shape C_6, whose truncated target equals $SO(2)$ below degree 6. More generally, C_n and D_n targets coincide with their continuous counterparts below degree n. At $L_{\max} = 1$, the D_n targets are uniform ($m_1 = 0$; Table 4); Fig. 11 explains the artifacts that arise when rendering these uniform densities.

F.5 Band-limit gallery of the polyhedral groups

Although T, O, and I lie outside the labeling vocabulary and are not exact symmetries of the scanned shapes (Appendix E.3), the representation can express them. Targets constructed directly from these groups first resolve T at $L_{\max} = 3$, O at $L_{\max} = 4$, and I at $L_{\max} = 6$.

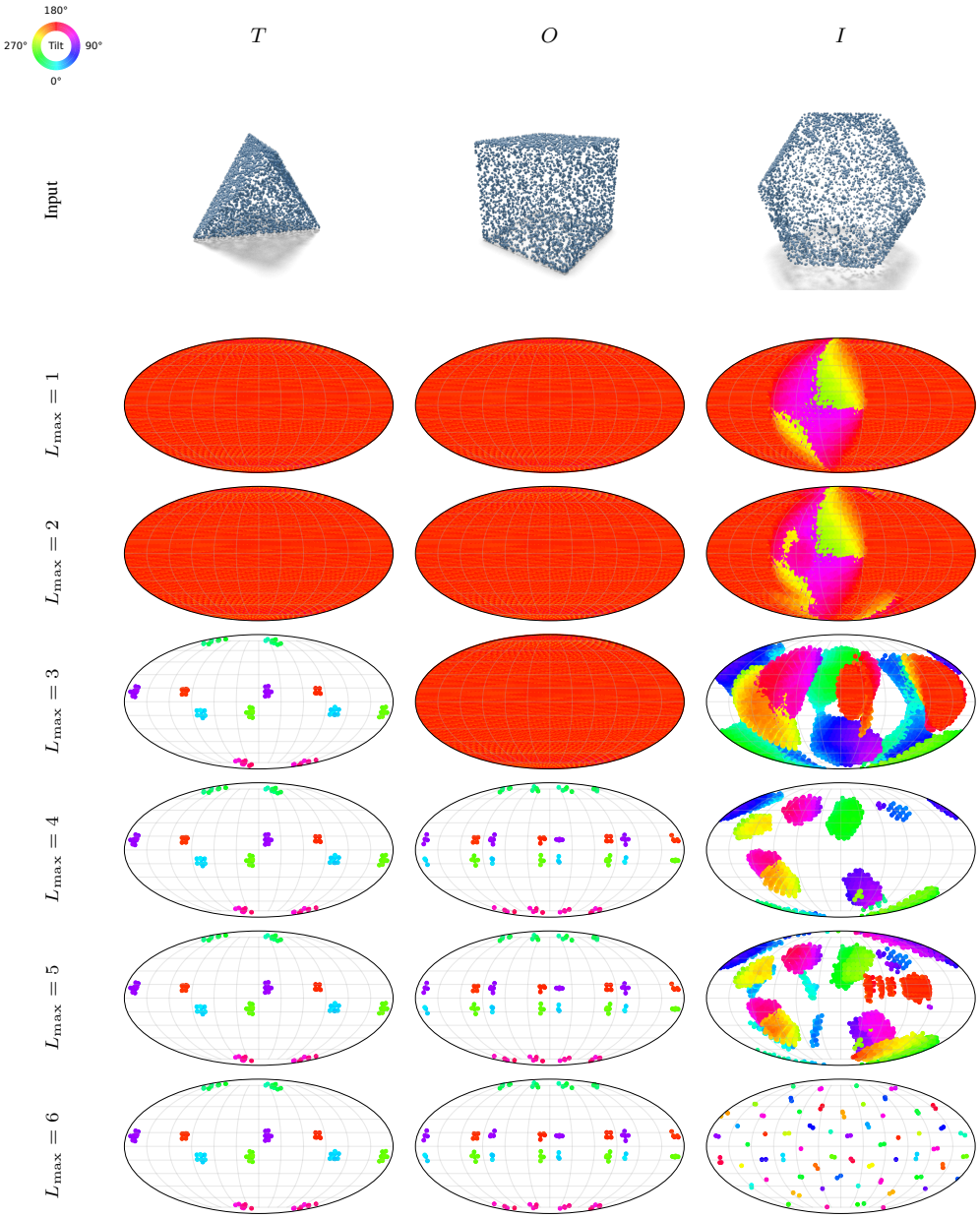

Figure 11: **Polyhedral band-limit gallery: T, O, and I.** Synthetic tetrahedral, cubic, and icosahedral solids use group-derived targets with $\sigma = 0.2$, rendered as in Fig. 3. Their orbits resolve at $L_{\max} = 3$, 4, and 6, respectively (Table 4). Below these degrees, the targets are uniform. The renderer's top-0.1% cutoff then retains either all overlapping points, showing the last-drawn tilt color, or points selected by numerical noise. The resulting blobs and patches are rendering artifacts, not modes.

F.6 Gallery of random test samples

Figures 12 and 13 show test samples of every category: the randomly rotated input and the cloud DORE canonicalizes from it in a single pass. Within each category, we sample successes and failures randomly, using the observed failure rate to determine their proportions and including at least one failure. For table, lamp, and cellphone, the third row additionally shows best-of-N selection on the same single-pass predictions. The final two columns randomly sample corrected flips from the verifier's recoveries (Appendix D.8).

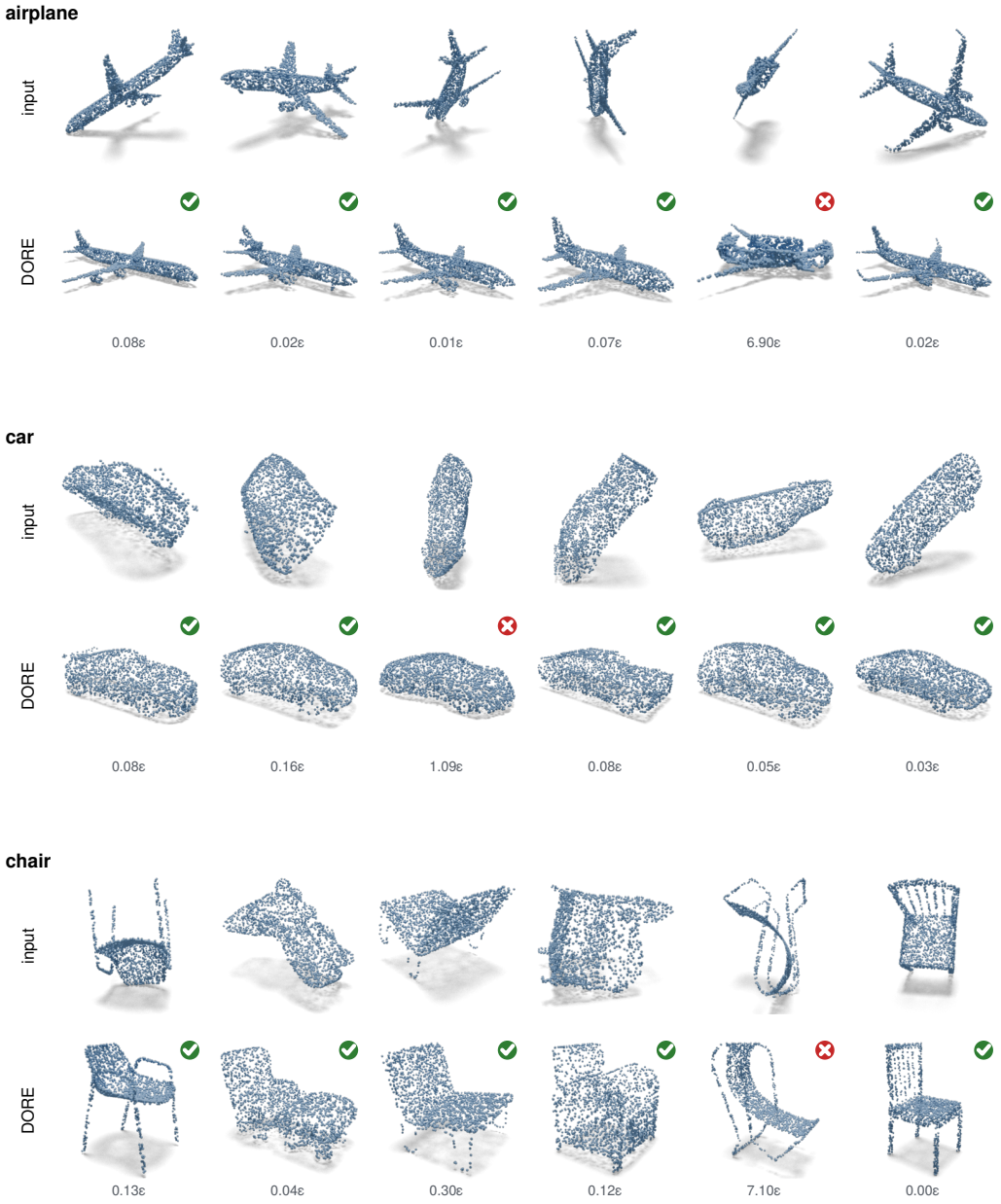

Figure 12: **Random test-sample gallery (1/2): airplane, car, and chair.** Each block shows the rotated 2,000-point evaluation cloud above its single-pass canonicalization by DORE at L_{sym}. Errors are in units of $\epsilon_\theta(\mathbf{P})$, with values at most one marked correct (Table 1). Successes and failures are sampled randomly in proportion to the category's failure rate, with at least one failure included.

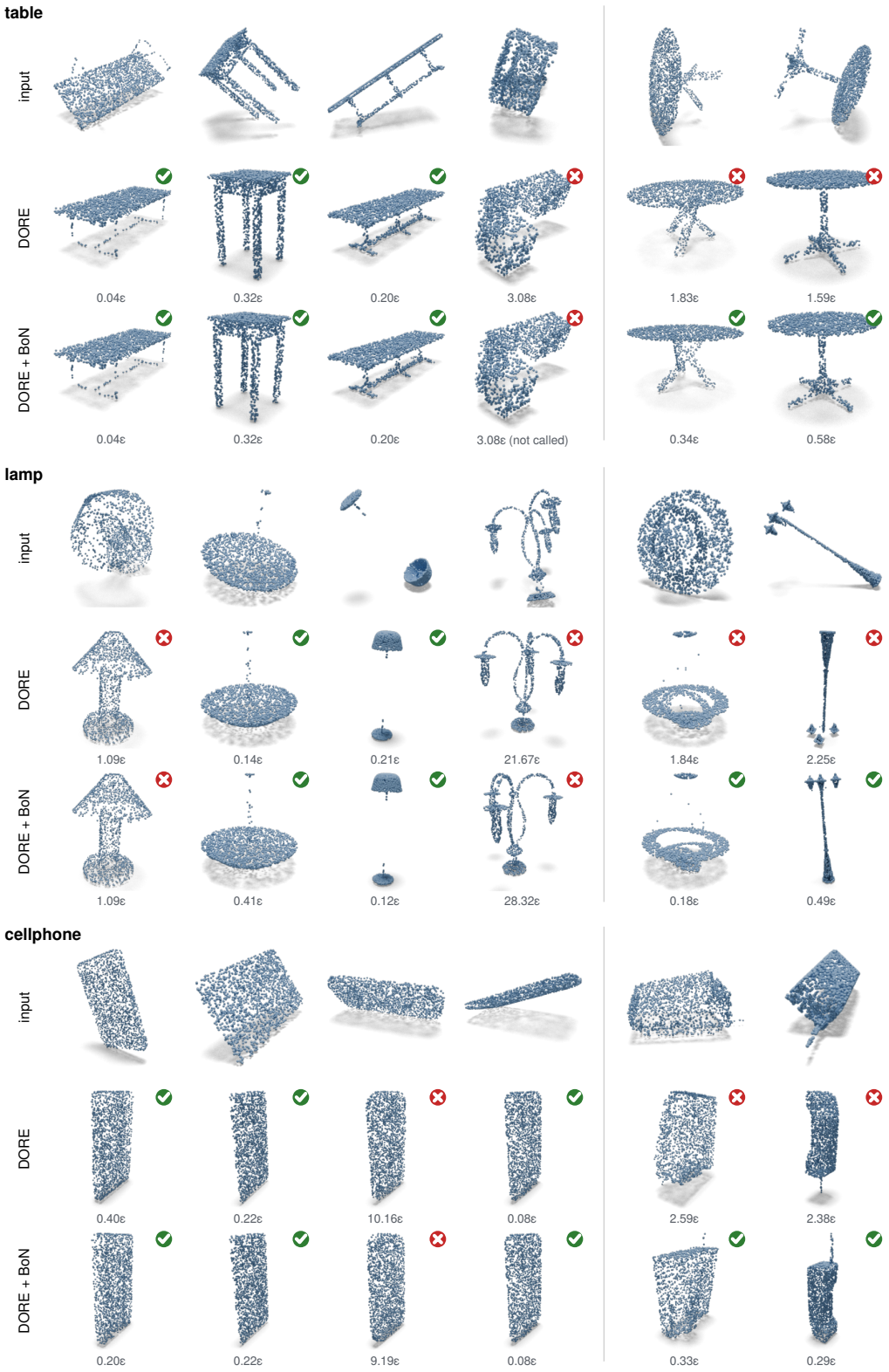

Figure 13: **Random test-sample gallery (2/2): table, lamp, and cellphone.** Sampling and rendering follow Fig. 12. The third row shows best-of-N, retaining the top-1 prediction when only one peak qualifies. The last two columns show randomly sampled recoveries in which the verifier corrects a top-1 prediction at least 150° from the valid peak (Appendix D.8).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the contributions (an $SO(3)$-equivariant estimator that predicts a distribution over orientations as a truncated Wigner-D expansion, trained on symmetry-orbit targets from an automatic labeler that covers the cyclic and dihedral groups C_n, D_n with $n \leq 6$) and their scope, and are supported by the method (Sec. 4.2, 4.1), theory (Appendix), and experiments (Sec. 5).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 7 states that the output can have negative side lobes and is used through its modes rather than as a calibrated probability; that supervision covers symmetry ambiguity in complete shapes but not partial observations, appearance ambiguity, or unequally weighted modes; that the labeler covers only C_n and D_n with $n \leq 6$, with heuristic τ and σ; and that all experiments use per-class models. Appendix C.2 analyzes the first; Appendices A.3 and B.3 state the head's limits on exactly symmetric inputs and its rank cap at high degrees; Appendices E.3 and F.5 cover the labeler's scope.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical results (Lemma 1, Proposition 1, Propositions 2 and 3, Proposition 5 with Corollary 1, and Lemma 2) are stated with their assumptions and proved in Appendices A and C; the mode analysis of Appendix B.1 states the conditions under which it holds.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The architecture (Sec. 4.2), loss (Sec. 4.4), and inference (Sec. 4.3) are described in detail, with hyperparameters and dataset preparation in the appendix; all datasets used are publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code is not yet public. The datasets used (ShapeNet and ModelNet40) are publicly available; the architecture, training procedure, and hyperparameters are described in the main paper and appendix in sufficient detail to reproduce the results, and the code and the frozen evaluation set (Appendix B.4) will be made publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Train/test splits, optimizer, learning rate, batch size, $L_{\max}$, smoothing bandwidth, and other hyperparameters are reported in Sec. 5 and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Main-text tables report single runs at the final checkpoint of a fixed schedule, without error bars. Appendix D.6 repeats two settings over seeds: the six per-class L_{sym} configurations of the main results (mean $\pm$ sample standard deviation over three seeds, Table 15) and cellphone at $L_{\max} = 5$ (five paired seeds, Table 16).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All DORE models are trained on two 24 GB GPUs (Titan RTX or RTX 3090) for 1,500 epochs, and the baselines on their own schedules (Appendices B.3 and B.4); a single-category run takes roughly 5 to 44 hours depending on the category's size and band; the best-of-N verifier trains on one RTX 3090 in 43 to 182 minutes per category. Reproducing the main results (the six per-category models of Table 1 and their verifiers) takes at most about 190 RTX 3090 GPU-hours, including periodic validation. The full project also included preliminary and discarded runs not reported here.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research uses publicly available 3D shape datasets and involves no human subjects, no scraping, and no deployment with foreseeable harm.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is foundational research on geometric representation learning over $SO(3)$. We do not foresee a direct path to negative societal impact; potential applications (e.g., robotics, graphics) are general-purpose.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The proposed model predicts 3D orientations from point clouds and poses no high-risk misuse profile.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The dataset and codebases used are cited and used under their respective licenses: ShapeNet under the ShapeNet Terms of Use (non-commercial research; `https://shapenet.org/terms`), read through the Cap3D release [26] under ODC-By 1.0, Vector Neurons [8] under the MIT License, and e3nn [13] under the MIT License; the public code of Scarvelis et al. [39], used to retrain the baseline, is under GPL-3.0. ModelNet40 [46] is distributed for academic research without an explicit license, and the public CHOIR code [21], whose authors include two of ours, specifies no license. Our use is consistent with each license and these terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets are released with this version; the code and the frozen evaluation set (Appendix B.4) will be made publicly available with documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects research; no IRB review is applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the core methodology; any use was limited to writing, editing, and coding assistance, which does not require declaration per the policy.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.