
Neural Isometries: Taming Transformations for Equivariant ML

Thomas W. Mitchel
PlayStation
tommy.mitchel@sony.com

Michael Taylor
PlayStation
mike.taylor@sony.com

Vincent Sitzmann
MIT
sitzmann@mit.edu

Abstract

Real-world geometry and 3D vision tasks are replete with challenging symmetries that defy tractable analytical expression. In this paper, we introduce Neural Isometries, an autoencoder framework which learns to map the observation space to a general-purpose latent space wherein encodings are related by *isometries* whenever their corresponding observations are geometrically related in world space. Specifically, we regularize the latent space such that maps between encodings preserve a learned inner product and commute with a learned functional operator, in the same manner as rigid-body transformations commute with the Laplacian. This approach forms an effective backbone for self-supervised representation learning, and we demonstrate that a simple off-the-shelf equivariant network operating in the pre-trained latent space can achieve results on par with meticulously-engineered, handcrafted networks designed to handle complex, nonlinear symmetries. Furthermore, isometric maps capture information about the respective transformations in world space, and we show that this allows us to regress camera poses directly from the coefficients of the maps between encodings of adjacent views of a scene.

1 Introduction

We constantly capture lossy observations of our world – images, for instance, are 2D projections of the 3D world. Observations captured consecutively in time are often related by transformations which are easily described in world space, but are intractable in the space of observations. For instance, video frames captured by a camera moving through a static scene are fully described by a combination of the 3D scene geometry and $SE(3)$ camera poses. In contrast, the image space transformations between these frames can only be characterized by optical flow, a high-dimensional vector field that does not itself have any easily tractable low-dimensional representation.

Geometric deep learning seeks to build neural network architectures that are provably robust to transformations acting on their inputs, such as rotations [1–3], dilations [4, 5], and projective transformations [6, 7]. However, such approaches are only tractable for transformations that have group structure, and, even in those cases, still require meticulously handcrafted and complex architectures. Yet, many real-world transformations of interest, for instance in vision and geometry processing, altogether lack identifiable group structure, such as the effect of camera motion in image space—see Fig. 1 to the right. Even when group-structured, they are often non-linear and non-compact, such as is the case of image

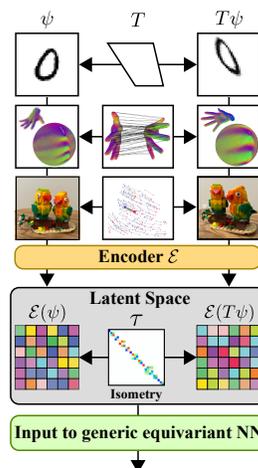


Figure 1: Neural Isometries find latent spaces where complex transformations become tractable.

homographies and non-rigid shape deformations where existing approaches can be prohibitively expensive.

In this paper, we propose Neural Isometries (**NIso**), an autoencoder framework which learns to map the observation space to a latent space in which encodings are related by tractable, highly-structured linear maps whenever their corresponding observations are geometrically related in world space.

Specifically, observations are encoded into latent functions by preserving their spatial dimensions. Images, for instance, can be encoded into latents defined over a lower resolution grid or the patch tokens of a ViT. For observations sharing some potentially unknown relationship in world space, we enforce their encodings be related by a functional map τ – a linear transformation on the space of latent functions. In particular, we require that τ is an *isometry* such that it preserves a learned inner product and commutes with a learned functional operator, in the sense that rigid body transformations commute with the Laplace operator in Euclidean space as do Lorentz transformations with the d’Alembert operator in Minkowski space.

Neural Isometries form an architecture-agnostic regime for self-supervised equivariant representation learning. To this end, we experimentally validate two principle claims regarding the efficacy and general applicability of our approach:

- Neural Isometries recover a general-purpose latent space in which challenging symmetries in the observation space can be reduced to compact, tractable maps in the latent space. We show that this can be exploited by simple isometry-equivariant networks to achieve results on par with leading hand-crafted equivariant networks in tasks with complex non-linear symmetries.
- The latent space constructed is geometrically informative in that it encodes information about the transformations in world space. We demonstrate that robust camera poses can be regressed directly from the isometric functional maps between encodings of adjacent views of a scene.

2 Related Work

Geometric Deep Learning. Geometric deep learning is generally concerned with hand-crafting architectures that are equivariant to (*i.e.* that commute with) *known* transformations acting on the data [1–4, 6–17]. To this end, many successful architectures exploit group representations by using established mappings, such as the Fourier or spherical harmonic transforms, to map features onto domains where the group actions manifest equivalently as linear transformations [1–3, 7]. In most cases, the representations considered are finite-dimensional and *irreducible* which, loosely speaking, means that the group action in observation space can be expressed exactly by frequency-preserving block-diagonal matrices acting on the transform coefficients. While finite-dimensional irreducible representations (**IRs**) are attractive building blocks for equivariance due to their computationally exploitable structure, they often don’t exist for non-compact groups, precluding generalizations to most non-linear symmetries, let alone those ill-modeled by groups. We instead avoid a heuristic choice of symmetry model and seek an approach that enables robustness to arbitrary transformations that may not even be group-structured.

Functional Maps. Essential to our approach is the parameterization of transformations between observations in the latent space not as group representations, but instead as *functional maps*. Introduced in the seminal work of Ovsjanikov *et al.* [18], functional maps (**FMs**) provide a powerful medium for interpreting, manipulating, and constructing otherwise intractable mappings between 3D shapes via their realization as linear transformations on spaces of functions on meshes, forming the basis for state-of-the-art pipelines in shape matching and correspondence [19–22]. Beyond 3D shapes, FMs can be seen as a tool to parameterize transformations between observations viewed as functions, in the sense that images, for instance, are functions that map 2D pixel coordinates to RGB colors. Integral to their study and implementation is the Laplace-Beltrami operator (the generalization of the Laplace operator to function spaces on manifolds), and in particular the expression of FMs in its eigenbasis which can expose specific geometric properties of deformations. In particular, isometries (distance-preserving transformations) manifest as highly-structured matrices not unlike IRs, being orthogonal and commuting with the diagonal matrix of eigenvalues and are thus approximately block-diagonal. That said, FMs are recovered through regularized linear solves and as such lack the immutability inherent in the analytical expressions for IRs of compact groups. However, freed from nagging theoretical constraints, approximately isometric FMs and their orthogonal relaxations have

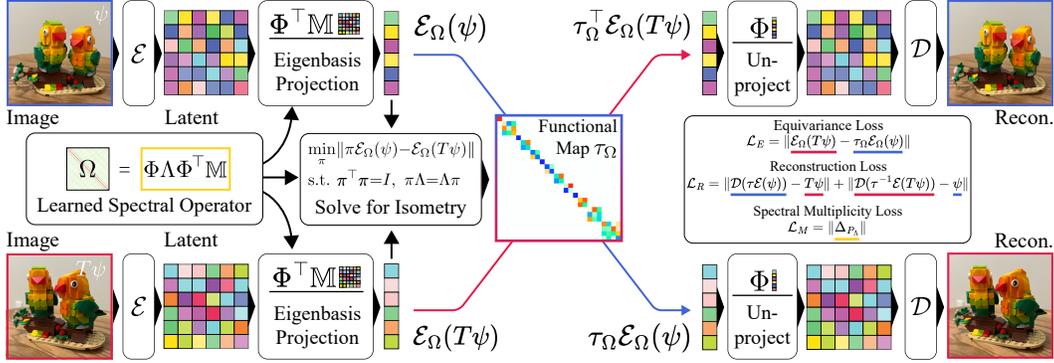


Figure 2: **Overview of Neural Isometries (Niso)**. NIsos learn a latent space where transformations of observations manifest as isometries, achieved by regularizing the functional maps τ between latents to commute with a learned operator Ω , parameterized via its spectral decomposition into a mass matrix \mathbb{M} , eigenfunctions Φ , and eigenvalues Λ (sec. 4.1). Given two observations ψ and $T\psi$ related by some unknown transformation T (in this case, camera motion in a 3D scene), they are first encoded into latent functions $\mathcal{E}(\psi)$ and $\mathcal{E}(T\psi)$ and projected into the operator eigenbasis. An isometric functional map τ_Ω is estimated between them, and used to map one to the other. Losses promote isometry-equivariance in the latent space, reconstruction of transformed latents, and distinct, low-multiplicity eigenvalues Λ , with the latter encouraging a diagonal as possible τ_Ω (sec. 4.2).

displayed remarkable representational capacity, well modeling a variety of highly-complex non-rigid deformations [23] including those without group structure, such as partial [24] and inter-genus correspondence [25]. Please see [26] for an outstanding introduction to functional maps.

Discovering Latent Symmetries. A number of recent approaches have proposed autoencoder frameworks wherein given symmetries in the base space manifest as simple operations in the latent space [27–33]. Perhaps most similar to our approach is recent work on a Neural Fourier Transform (NFT) that has sought to manifest group actions in observation space as IRs in the latent space [34, 35]. These methods offer impressive theoretical guarantees under the conditions that the observed transformations are either known or are a group action, although these assumptions may not hold for real-world data with complex symmetries. In contrast, our method is wholly unsupervised, and assumes no knowledge of the transformations in observation space nor that they even form a group. Furthermore, in lieu of an extensive theoretical justification we instead seek to validate the efficacy of our approach experimentally, including in geometry processing and real-world 3D vision tasks.

3 Method Overview

Neural Isometries are an architecture-agnostic autoencoder framework which learns to map pairs of observations related in world space to latents which are related by an approximately isometric FM τ – see Fig. 2. We first formulate encoding and decoding, and define a FM in the latent space. Next, we show how a functional operator Ω and mass matrix \mathbb{M} can be learned in the latent space to regularize τ by requiring it to be an isometry. Specifically, for observations sharing some potentially unknown relationship in world space, we enforce there exist a FM τ between their encodings satisfying two key properties: 1) That τ preserves the functional inner product in the latent space determined by \mathbb{M} ; and 2). τ commutes with the functional operator Ω . Subsequently, we show that such maps can be recovered analytically through a differentiable, closed-form least squares solve in the operator eigenbasis. Last, we formulate NIsos as an optimization problem incorporating both the strictness of the isometric correspondence between latents and the eigenvalue multiplicity of the operator, the latter of which controls the structure of the maps.

In experiments, we demonstrate that NIsos are capable of both discovering approximations of known operators and constructing latent spaces where complex, non-linear symmetries in the observation space manifest equivalently as isometries. We show how the latter property can be exploited by demonstrating that a simple vector neuron MLP [36] acting in our pre-trained latent space can achieve results on par with state-of-the-art handcrafted equivariant networks operating in observation space. Subsequently, we consider the task of pose estimation, and demonstrate that robust SE(3) camera

poses can be extracted from latent transformations, serving as evidence that NISO encourages models to encode information about transformations in world space.

4 Neural Isometries

We consider an observation space $O \subset L^2(M, \mathbb{R}^n)$ consisting of functions defined over some domain M (e.g. with M the plane and $n = 3$ for RGB images). Here, elements of O are in fact captures from some world space W with $\sigma : W \rightarrow O$ representing the mechanism from which O is formed from W . Furthermore, we assume there is a potentially unknown collection of phenomena $\{T\}$ acting on the world space that relates observations. That is, for some $w \in W$, $\psi = \sigma(w) \in O$, and denoting $T\psi \equiv \sigma(Tw)$, we assume that $T\psi$ is also in O and that we are able to associate it with ψ .

Additionally, we consider an autoencoder consisting of an encoder and decoder

$$\mathcal{E} : L^2(M, \mathbb{R}^n) \rightarrow L^2(N, \mathbb{R}^d) \quad \text{and} \quad \mathcal{D} : L^2(N, \mathbb{R}^d) \rightarrow L^2(M, \mathbb{R}^n) \quad (1)$$

mapping between observation space and a space of latent functions over some domain N . In practice, we operate over discretizations of M and N , with $\psi \in O$ and $\mathcal{E}(\psi) \in L^2(N, \mathbb{R}^d)$ represented as tensors $\psi \in \mathbb{R}^{|M| \times n}$ and $\mathcal{E}(\psi) \in \mathbb{R}^{|N| \times d}$. For example, if M consists of the pixel indices of an image, then N could be grid or token indices if \mathcal{E} is a ConvNet or ViT, respectively.

Goal: Equivariance of Latent Functions. Our aim is to train the autoencoder such that for any T acting in the world space and corresponding observations $\psi, T\psi \in O$, there exists a linear map $\tau : L^2(N, \mathbb{R}^d) \rightarrow L^2(N, \mathbb{R}^d)$ such that

$$\mathcal{E}(T\psi) \approx \tau \mathcal{E}(\psi). \quad (2)$$

In other words, we desire our latent space to be *equivariant* under world space transformations. As our problem is discrete, τ is a *functional map* – an $|N| \times |N|$ matrix representation of maps on $L^2(N, \mathbb{R}^d)$. In the case of latents with $|N| = H \times W$ pixels, τ is a matrix whose rows express each pixel in $\mathcal{E}(T\psi)$ as a linear combination of pixels in $\mathcal{E}(\psi)$, similar to the weight matrix one might obtain from a cross-attention operation.

4.1 Regularization Through Isometries

We will find τ by solving a least-squares problem of the form $\tau = \min_{\pi} \|\mathcal{E}(T\psi) - \pi \mathcal{E}(\psi)\|$. Unfortunately, as we will show in experiments, a direct solve without additional regularization leads to uninformative maps that capture little information about the actual world-space transformations T . To add structure, we might ask that τ be *orthogonal* with $\tau^\top \tau = I_{|N|}$, generating gradients promoting latent codes having the property $\|\mathcal{E}(\psi)\| \approx \|\mathcal{E}(T\psi)\|$.

However, we can obtain more structure yet. We propose to learn a representation of the latent geometry by jointly regressing a diagonal mass matrix \mathbb{M} and positive semi-definite (PSD) operator $\Omega \in \mathbb{R}^{|N| \times |N|}$ such that τ manifests as an *isometry*. That is, τ preserves the functional inner product defined by $\mathbb{M} - \langle f, g \rangle_{\mathbb{M}} = f^\top \mathbb{M} g$ for $f, g \in L^2(N, \mathbb{R}^d)$ – and is Ω -commutative with

$$\tau^\top \mathbb{M} \tau = \mathbb{M} \quad \text{and} \quad \tau \Omega = \Omega \tau. \quad (3)$$

Together, the conditions in Equation (3) form a strong regularizer, the effects of which are best seen in the expression of τ in the eigenspectrum of Ω . As Ω is a PSD matrix with respect to the inner product defined by \mathbb{M} , it can be expressed in terms of its spectral decomposition as

$$\Omega = \Phi \Lambda \Phi^\top \mathbb{M} \quad \text{with} \quad \Phi^\top \mathbb{M} \Phi = I_{|N|}, \quad (4)$$

where $\Lambda = \text{diag}(\{\lambda_i\}_{1 \leq i \leq |N|})$ is the diagonal matrix of (non-negative) eigenvalues and $\Phi \in \mathbb{R}^{|N| \times |N|}$ is the matrix whose columns are the \mathbb{M} -orthogonal eigenfunctions of Ω . Denoting

$$\tau_\Omega \equiv \Phi^\top \mathbb{M} \tau \Phi \quad (5)$$

as the projection of τ into the eigenbasis, it can be shown that the conditions in Eq. (3) reduce to τ_Ω being orthogonal and Λ -commutative [26]. This is equivalent to asking that τ_Ω be both *sparse* and *condensed* in that it forms an *orthogonal, block-diagonal matrix*, with the size of each block determined by the multiplicity of the eigenvalues in Λ .

4.2 Estimating τ and End-to-End Optimization

Fig. 2 visualizes Neural Isometries’s training loop. First, τ is estimated between pairs of encoded T -related observations such that it approximately satisfies the conditions for an Ω -isometry as in Eq. (3). Second, the weights of the autoencoder, \mathbb{M} , and Ω are jointly updated with respect to a combined loss term, promoting: a) latent equivariance as in Eq. (2), b) the ability of the decoder to reconstruct observations, and c) distinct eigenvalues Λ which encourage a diagonal-as-possible τ_Ω .

Recovering τ Between Latents. Instead of estimating τ directly, we equivalently estimate τ_Ω in the eigenbasis of Ω , motivated by the corresponding simplification of the conditions in Eq. (3). Let

$$\mathcal{E}_\Omega \equiv \Phi^\top \mathbb{M} \circ \mathcal{E} \quad (6)$$

be the map given by encoding followed by projection into the eigenbasis of Ω . Then, given observations ψ and $T\psi$, we define τ_Ω to be the solution to the least squares problem

$$\tau_\Omega = \underset{\pi^\top \pi = I, \pi \Lambda = \Lambda \pi}{\text{minimum}} \|\pi \mathcal{E}_\Omega(\psi) - \mathcal{E}_\Omega(T\psi)\|. \quad (7)$$

While Eq. (7) has an exact analytical solution [37], we instead approximate τ_Ω with a fuzzy analogue which we find better facilitates backwards gradient flow to the parameters of Ω .



Specifically, letting $\varkappa : \mathbb{R}^{|N| \times |N|} \rightarrow \text{O}(|N|)$ denote the Procrustes projection to the nearest orthogonal matrix (e.g. through the SVD), we recover τ_Ω via the approximation

$$\tau_\Omega \approx \varkappa (P_\Lambda \odot \mathcal{E}_\Omega(T\psi) [\mathcal{E}_\Omega(\psi)]^\top), \quad (8)$$

where $[P_\Lambda]_{ij} = \exp(-|\lambda_i - \lambda_j|)$ is a smooth multiplicity mask over the eigenvalues Λ applied element-wise. See the supplement for details.

To facilitate the recovery of τ_Ω , we parameterize \mathbb{M} directly by its diagonal elements and Ω in terms of its spectral decomposition, learning a \mathbb{M} -orthogonal matrix of eigenfunctions Φ and non-negative eigenvalues Λ . This has the added benefit of enabling a low-rank approximation of Ω by parameterizing only the first k eigenvalues and eigenfunctions, i.e. $\Phi \in \mathbb{R}^{|N| \times k}$ and $\Lambda = \text{diag}(\{\lambda_i\}_{1 \leq i \leq k})$ with $k \leq |N|$, mirroring similar approaches in SoTA FM pipelines [19, 38, 39]. This reduces the complexity of the orthogonal projection in Eq. (8) from $|N| \times |N|$ to $k \times k$.

Optimization. During training, the autoencoder is given pairs of T -related observations $(\psi, T\psi)$, which are mapped to the latent space and τ_Ω is estimated as in Eq. (8) giving $\tau = \Phi \tau_\Omega \Phi^\top \mathbb{M}$. First, an **equivariance loss** is formed between the eigenspace projections of the encodings,

$$\mathcal{L}_E = \|\tau_\Omega \mathcal{E}_\Omega(\psi) - \mathcal{E}_\Omega(T\psi)\|. \quad (9)$$

We note that for full rank Ω , this loss is equivalent to measuring the degree to which the equivariance condition in Eq. (2) holds due to the orthogonality of Φ . Next we compute a **reconstruction loss**,

$$\mathcal{L}_R = \|\mathcal{D}(\tau \mathcal{E}(\psi)) - T\psi\| + \|\mathcal{D}(\tau^{-1} \mathcal{E}(T\psi)) - \psi\|, \quad (10)$$

with $\tau^{-1} = \Phi \tau_\Omega^\top \Phi^\top \mathbb{M}$, forcing the decoder to map the transformed latents to the corresponding T -related observations. Last, we formulate a **multiplicity loss** which promotes distinct eigenvalues and ensures Ω is “interesting” by preventing it from regressing to the identity. We observe that the eigenvalue mask P_Λ can be viewed as a graph in which the number of connected components (i.e. the number of distinct eigenvalues) is equivalent to the dimension of the nullspace of the graph Laplacian Δ_{P_Λ} formed from the mask [40]. As a measure of the nullspace dimension, we use the norm of the eigenvalues of Δ_{P_Λ} , given by $\mathcal{L}_M = \|\Delta_{P_\Lambda}\|$. The NFT [34, 35] takes a similar approach, wherein a diagonalization loss is imposed on estimated transformations themselves, though our experiments show it be far less effective in enforcing structure. The total loss is the sum of aforementioned terms

$$\mathcal{L} = \mathcal{L}_R + \alpha \mathcal{L}_E + \beta \mathcal{L}_M, \quad (11)$$

with $\alpha, \beta \geq 0$ weighting the contributions of the equivariance and multiplicity losses.

In experiments, we also consider a similar triplet regime as proposed in [34, 35], where the autoencoder is given triples of T -related observations $(\psi, T\psi, T^2\psi)$ (assuming T is composable) and the estimated map τ between the encodings of $(\psi, T\psi)$ is used to form equivariance and reconstruction losses between $(T\psi, T^2\psi)$ and vice-versa. This works to prevent τ from “cheating” by encoding

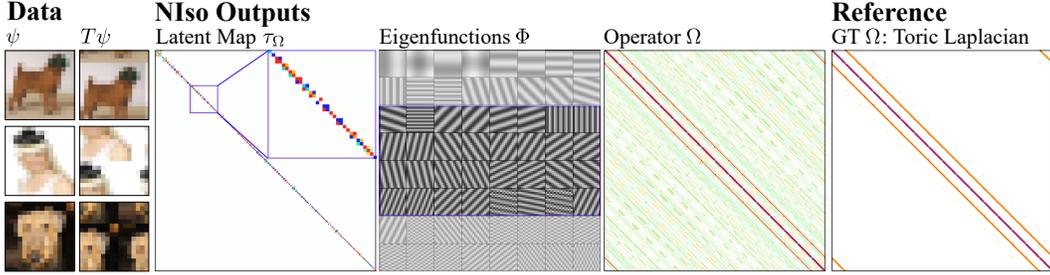


Figure 3: **Approximating the Toric Laplacian from Shifted Images.** Forced to map between shifted images (left), NIso regresses an operator (center right) structurally similar to the toric Laplacian (right). Maps τ_Ω between projected images are strongly diagonal (center left), with individual blocks (inset) preserving the subspaces spanned by eigenfunctions (center) sharing nearly the same eigenvalues.

privileged information about the relationship between pairs beyond T , a property we show to be critical in enforcing a useful notion of latent equivariance. However, triples of T -related observations are rare in practical settings, and we show in experiments that a major benefit of our isometric regularization is that our multiplicity loss (promoting a diagonal-as-possible and thus sparse and condensed τ_Ω) can serve as an effective substitute for access to triples.

4.3 A Simple Example: Approximating the Toric Laplacian

We demonstrate that NIsos are able to learn a compact representation of isometries that reflect the dynamics of transformations in world space. To do so, we consider pairs of observations formed by 16×16 images from ImageNet [41] perturbed by random circular shifts along both spatial axes. Thus pairs are related by the isometries of the torus which commute with the Laplacian. Taking the encoder and decoder to be the identity map (making the equivariance and reconstruction losses equivalent) we optimize for $\mathbb{M}, \Omega \in \mathbb{R}^{256 \times 256}$ via the pairwise training procedure described in sec. 4.2. As seen in Fig 3, NIso regresses an operator with significant structural similarities to the Laplacian matrix formed by the standard 3×3 stencil and an eigenspace that diagonalizes τ_Ω between shifted images. As our approach is data-driven and our estimated maps are only *approximately* isometric, our learned operator and its eigenspectrum do not exactly correspond to the ground-truth Laplacian. Instead, we are able to characterize similar, non-trivial spatial relationships that are preserved under shifts.

4.4 Representation Learning with NIso

Viewed in terms of representation learning, NIso can be seen as a recipe for the self-supervised pre-training of a network backbone \mathcal{E} satisfying the equivariance condition in Eq. (2) such that transformations T in the world space manifest as isometries τ in the sense of Eq. (3).

Exploiting Equivariance in Latent Space As we demonstrate in experiments, a simple off-the-shelf isometry-equivariant head can be appended to the pre-trained backbone and fine-tuned to achieve competitive results in tasks with challenging symmetries. We employ a simple strategy wherein a low rank k approximation of Ω is learned during the pre-training stage. Thus, the eigenspace projections of the encodings of T -related observations are $k \times d$ tensors that are nearly equivalent up to an *orthogonal* transformation τ_Ω . As such, we pass the projected encodings to a head consisting of an $O(k)$ -equivariant vector neuron (VN) MLP [36]. We note that for large-scale tasks NIso is potentially well-suited to pair with DiffusionNet [42], which can make use of the learned eigenbasis to perform accelerated operations in the latent space, though we do not consider this regime here.

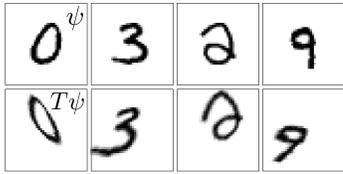
Pose Extraction from Latent Isometries. We propose that the recovered functional maps τ encode information about world-space transformations T . To test this, we consider a simple pose estimation paradigm consisting of a pre-training phase and fine-tuning phase. In the first phase, a NIso autoencoder is trained using T -related pairs of observations consisting of adjacent frames from video sequences. Subsequently, the decoder is discarded and the same pairs of observations are considered during fine-tuning. In the second phase, isometries τ_Ω are estimated between the eigenspace projections of the encoded observations, vectorized, and passed directly to an MLP which predicts the parameters of the $SE(3)$ transformation corresponding to the relative camera motion in world space between the adjacent frames. In our experiments, the weights of the NIso backbone

are frozen during fine-tuning to better evaluate the information about world space transformations encoded during the unsupervised pre-training phase. At evaluation, trajectories are recovered by composing estimated frame-to-frame poses over the length of the sequence.

5 Experiments

In this section, we provide empirical evidence through experiments that NIso 1) recovers a general-purpose latent space that can be exploited by isometry-equivariant networks to handle challenging symmetries (5.1, 5.2); and 2) NIso encodes information about transformations in world space through the construction of isometric maps in the latent space from which geometric quantities such as camera poses can be directly regressed (5.3). We provide reproducibility details in the supplement. We note that a consistent theme in our experiments are comparisons against the unsupervised variant of the NFT [35] (the semi-supervised variants cannot be applied because the symmetries we consider have no finite-dimensional IRs). Like our approach, the NFT seeks to relate latents via linear transformations though, it differs fundamentally in that maps are guaranteed additional structure only if the world space transformations are a compact group action. While not originally proposed by the authors, we evaluate it in place of our approach in the same self-supervised representation learning regimes discussed in sec. 4.4. Thus the role of these comparisons is to show that our proposed *isometric* regularization better and more consistently provides a tractable and informative latent space.

5.1 Homography-Perturbed MNIST



	Acc.
NIso	92.52 (\pm 0.91)
w/ triplet	97.38 (\pm 0.23)
w/o \mathcal{L}_E	77.30 (\pm 2.56)
w/o \mathcal{L}_M	45.27 (\pm 1.20)
NFT [35]	41.93 (\pm 0.84)
w/ triplet	67.15 (\pm 1.10)
AE Baseline	46.64 (\pm 0.41)
homConv [6]	95.71 (\pm 0.09)
LieDecomp [43]	98.30 (\pm 0.10)

Table 1: **Hom. MNIST.**

approach versus an NFT backbone pre-trained in the same manner. Last, we pre-train and evaluate a baseline backbone which considers only a reconstruction loss without respect to T -related pairs.

Results are shown in Tab. 1, averaged over five randomly initialized pre-training and fine-tuning runs with standard errors. Also included are those reported by homConv [6] and LieDecomp [43], top-performing homography-equivariant networks which serve as a handcrafted baseline. While NIso pre-trained with the triplet regime produces results on par with the handcrafted baselines, pre-training with the pairwise regime—which reflects a real-world scenario—achieves a classification accuracy above 90%, significantly better than all but the three aforementioned approaches. Critically, performance drops when the multiplicity loss \mathcal{L}_M is omitted, which corresponds to a regime where τ must only preserve the inner product and Ω converges to a multiple of the identity operator. This suggests that sparsifying τ_Ω by filtering it through a low-multiplicity eigenspace (*i.e.* enforcing that it is an “interesting” isometry) is fundamental in forcing the network to disentangle the structure of the observed transformations from the content of the observations themselves. In the same vein, the equivariance loss \mathcal{L}_M is also clearly instrumental, as the reconstruction loss alone does not explicitly enforce that $\mathcal{E}(\psi)$ is in fact mapped to $\mathcal{E}(T\psi)$ under the estimated τ . Furthermore, while the authors report that latent maps tend to converge to orthogonal maps for compact group actions in world

space [34, 35], both NFT regimes perform poorly, implying that the learned maps do not replicate the properties of finite-dimensional IRs when the group is non-compact.

5.2 Conformal Shape Classification

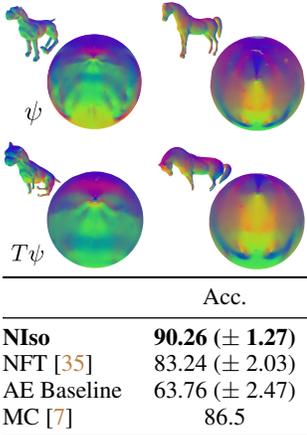


Table 2: **Conf. SHREC ‘11.**

Next, we apply NIso to classify conformally-related 3D shapes from the augmented SHREC ‘11 dataset [7, 44]. We follow [7] by mapping each mesh to the sphere and subsequently rasterizing to a grid. During pre-training, T -related pairs are selected from the sets of conformally-augmented meshes derived from the same base shape in the train split. In the fine-tuning phase, the encoder weights are unfrozen and are jointly optimized with the equivariant head, representing the practical implementation of our proposed approach for equivariant tasks.

Results are shown in Tab. 2, averaged over five randomly initialized pre-training and fine-tuning runs with standard errors. NIso outperforms the NFT and the autoencoder baseline in addition to Möbius Convolutions (MC) [7], a SoTA handcrafted spherical network equivariant to Möbius transformations. We consider this dataset to present a significant challenge as shape classes are roughly conformally-related and thus the maps between their spherical parameterizations are only approximated by Möbius transformations. This makes the performance of NIso particularly notable as it suggests that our framework

has the potential to offer a more flexible and effective alternative than specialized, handcrafted equivariant networks whose underlying group-based architectures, though elegant, can only approximate inexact symmetries. Furthermore, while the NFT is relatively more competitive with the encoder backbone unfrozen, its performance and that of the baseline indicate that a lack of existing equivariant structure cannot easily be overcome in the fine-tuning phase.

5.3 Camera Pose Estimation from Real-World Video

Last, we apply NIso to the task of camera pose estimation from real-world video on the CO3Dv2 dataset [45] following the procedure described in sec. 4.4. Due to the varying quality of the ground-truth trajectories in the dataset, we create train and test sets from the top 25% of sequences in the dataset as ranked by the provided pose quality scores. In particular, we are interested in the general ability of our method to encode information about both small and large scale world space transformations. Thus we train and evaluate in a variable baseline regime, randomly skipping between 0 and 9 frames between pairs during the pre-training and fine-tuning phases. Evaluation is performed by computing the mean absolute trajectory error (ATE) between the ground truth trajectories and those recovered by our method over the sequences in the evaluation set. To understand the efficacy of our approach at different scales, we report the mean ATE over six splits consisting of the same sequences in the evaluation set with frame skips of 0, 1, 3, 5, 7, and 9.

We compare against the NFT using the same pre-training and fine-tuning procedure along with a transformer baseline. The latter is inspired by the recent success of such models in 3D vision tasks [46, 47], and uses the same ViT-based architecture proposed in DUS3R [47], with the decoder modified to directly predict the parameters of the relative camera pose. As the weights of the NIso and NFT backbones are frozen during the fine-tuning phase, the transformer baseline is trained from scratch in the fine-tuning phase to more directly compare the descriptiveness of the latent representations relative to the information contained in the raw observations.

We additionally train and evaluate an ablativ version of NIso which does not learn an operator. Here τ is predicted directly between encodings and is required only to be orthogonal with the equivariance and reconstruction losses alone enforced during training. We note that computing a block-diagonalization loss on τ directly lacks justification as it would promote maps that preserve contiguous chunks of spatial indices in the latent tensors which are ordered arbitrarily. Thus, this regime serves to evaluate the degree to which the regularization through learned Ω -commutativity forces the latent transformations to reflect geometric relationships in world space.

Results are shown in Tab. 3, averaged over five randomly initialized pre-training and fine-tuning runs with standard errors. All methods perform similarly with 0 frame skip but diverge afterwards, with

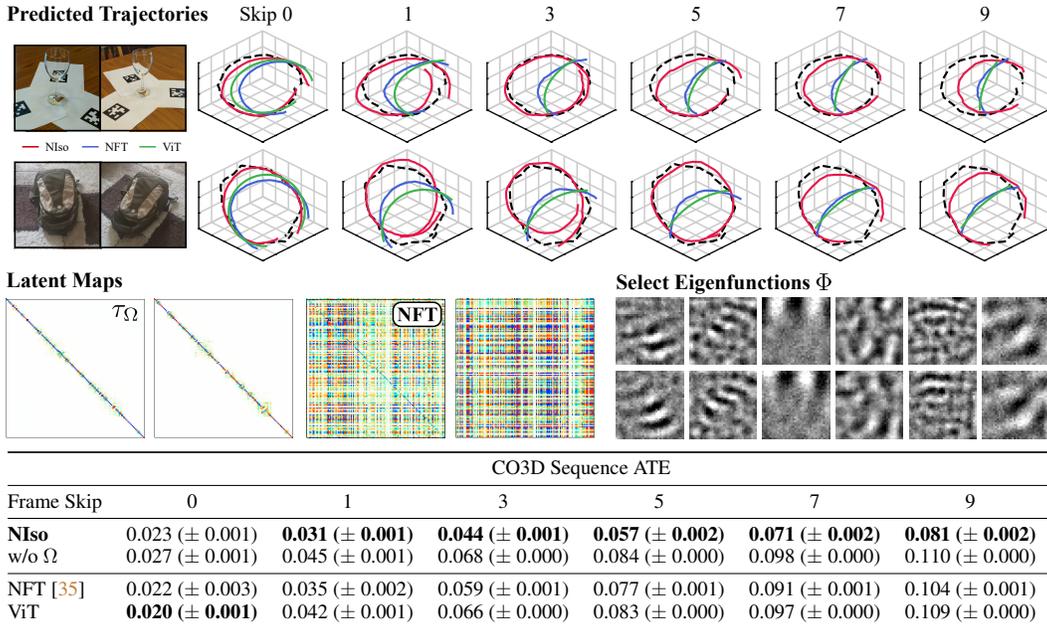


Table 3: **Pose Estimation Comparison.** Mean ATE for each method across all frame skips on CO3Dv2 evaluation sequences. *Top:* Examples of predicted trajectories, with ground truth in black, NIso in red, the NFT in blue, and the transformer baseline in green. Additional examples are shown in the supplement. *Center Left:* Representative examples of latent maps formed by NIso and the NFT, with the latter shown after applying the authors’ proposed diagonalization procedure. NIso is able to form highly sparse, block diagonal maps. In contrast, the NFT struggles to form similarly condensed maps. *Center Right:* Selected learned eigenfunctions. Each column forms a subspace, reflecting fundamental modes of symmetry identified by NIso.

NIso achieving significantly lower ATE values as the skip length increases. Notably, the ablate version of our method is consistently the worst performer, suggesting that isometric regularization is also critical to encode information about world space transformations. Overall the NFT achieves the second best performance, slightly outperforming NIso at 0 frame skip but diverging thereafter. However, the latent maps it recovers are neither sparse nor exhibit condensed structure (Tab. 3, center left), and we hypothesize that its inability to effectively regularize maps beyond linearity makes it difficult for the network to discover a consistent transformation model that applies across scales. This could cause the network to focus on a specific regime at the expense of others, which may explain its relatively strong performance at 0 frame skip. The transformer baseline is also highly flexible, and an analogous line of reasoning could explain its similar error profile.

6 Discussion

Limitations. A key factor limiting the broader applicability of our approach to geometry processing and graph-based tasks is an inability to learn and transfer an operator between domains with varying connectivity. This could potentially be overcome by parameterizing Ω by a network over the simplices of a mesh or graph following [48], and we seek to address generalization between data representations in future work. Furthermore, we observe that the eigenvalues learned in our parameterization of Ω are not necessarily interpretable as frequencies in the classical sense as visibly smooth eigenfunctions may correspond to large eigenvalues, and those with fine details to small eigenvalues.

Conclusion. In this paper we introduce Neural Isometries, a method which converts challenging observed symmetries into isometries in the latent space. Our approach forms an effective backbone for self-supervised representation learning, enabling simple off-the-shelf equivariant networks to achieve strong results in tasks with complex, non-linear symmetries. Furthermore, isometric regularization produces latent representations that are geometrically informative by encoding information about transformations in world space, and we demonstrate that robust camera poses can be extracted from the isometric maps between latents in a general baseline setting.

Acknowledgements. Use of the CO3Dv2 dataset is solely for benchmarking purposes and limited exclusively to the experiments described in sec. 5.3. We thank Ishaan Chandratreya for helpful discussions and David Charatan for his aesthetic oversight and for granting us permission to use his images captured for FlowMap [49] in Fig. 1-2.

Vincent Sitzmann was supported by the National Science Foundation under Grant No. 2211259, by the Singapore DSTA under DST00OECI20300823 (New Representations for Vision and 3D Self-Supervised Learning for Label-Efficient Vision), by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) under 140D0423C0075, by the Amazon Science Hub, and by IBM.

References

- [1] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2990–2999. PMLR, 2016.
- [2] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5028–5037, 2017.
- [3] Carlos Esteves, Ameesh Makadia, and Kostas Daniilidis. Spin-weighted spherical cnns. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [4] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3165–3176. PMLR, 2020.
- [5] Daniel E Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. *arXiv preprint arXiv:1905.11697*, 2019.
- [6] Lachlan E MacDonald, Sameera Ramasinghe, and Simon Lucey. Enabling equivariance for arbitrary lie groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8183–8192, 2022.
- [7] Thomas W Mitchel, Noam Aigerman, Vladimir G Kim, and Michael Kazhdan. Möbius convolutions for spherical cnns. In *ACM Trans. Graph.*, pages 1–9, 2022.
- [8] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018.
- [9] Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In *International conference on machine learning*, pages 3318–3328. PMLR, 2021.
- [10] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.
- [11] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.
- [12] Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Geometric and physical quantities improve e (3) equivariant message passing. *arXiv preprint arXiv:2110.02905*, 2021.
- [13] Alexander Bogatskiy, Brandon Anderson, Jan Offermann, Marwah Roussi, David Miller, and Risi Kondor. Lorentz group equivariant neural network for particle physics. In *International Conference on Machine Learning*, pages 992–1002. PMLR, 2020.
- [14] Erik J Bekkers. B-spline cnns on lie groups. *arXiv preprint arXiv:1909.12057*, 2019.
- [15] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019.
- [16] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

- [17] Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. *arXiv preprint arXiv:2110.03336*, 2021.
- [18] Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (ToG)*, 31(4): 1–11, 2012.
- [19] Simone Melzi, Jing Ren, Emanuele Rodola, Abhishek Sharma, Peter Wonka, and Maks Ovsjanikov. Zoomout: Spectral upsampling for efficient shape correspondence. *arXiv preprint arXiv:1904.07865*, 2019.
- [20] Robin Magnet and Maks Ovsjanikov. Memory-scalable and simplified functional map learning. *arXiv preprint arXiv:2404.00330*, 2024.
- [21] Mingze Sun, Shiwei Mao, Puhua Jiang, Maks Ovsjanikov, and Ruqi Huang. Spatially and spectrally consistent deep functional maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14497–14507, 2023.
- [22] Ahmed Abdelreheem, Abdelrahman Eldesokey, Maks Ovsjanikov, and Peter Wonka. Zero-shot 3d shape correspondence. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.
- [23] Mikhail Panine, Maxime Kirgo, and Maks Ovsjanikov. Non-isometric shape matching via functional maps on landmark-adapted bases. In *Computer graphics forum*, volume 41, pages 394–417. Wiley Online Library, 2022.
- [24] Souhaib Attaiki, Gautam Pai, and Maks Ovsjanikov. Dpfm: Deep partial functional maps. In *2021 International Conference on 3D Vision (3DV)*, pages 175–185. IEEE, 2021.
- [25] Dongliang Cao, Paul Roetzer, and Florian Bernard. Revisiting map relations for unsupervised non-rigid shape matching. *arXiv preprint arXiv:2310.11420*, 2023.
- [26] Maks Ovsjanikov, Etienne Corman, Michael Bronstein, Emanuele Rodolà, Mirela Ben-Chen, Leonidas Guibas, Frederic Chazal, and Alex Bronstein. Computing and processing correspondences with functional maps. In *SIGGRAPH ASIA 2016 Courses*, pages 1–60. 2016.
- [27] Yilun Du, Katie Collins, Josh Tenenbaum, and Vincent Sitzmann. Learning signal-agnostic manifolds of neural fields. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8320–8331, 2021.
- [28] Sharut Gupta, Joshua Robinson, Derek Lim, Soledad Villar, and Stefanie Jegelka. Learning structured representations with equivariant contrastive learning. 2023.
- [29] Sangnie Bhardwaj, Willie McClinton, Tongzhou Wang, Guillaume Lajoie, Chen Sun, Phillip Isola, and Dilip Krishnan. Steerable equivariant representation learning. *arXiv preprint arXiv:2302.11349*, 2023.
- [30] Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with lie algebra convolutional network. *Advances in Neural Information Processing Systems*, 34:2503–2515, 2021.
- [31] Artem Moskalev, Anna Sepiarskaia, Ivan Sosnovik, and Arnold Smeulders. Liegg: Studying learned lie group generators. *Advances in Neural Information Processing Systems*, 35:25212–25223, 2022.
- [32] Jianke Yang, Robin Walters, Nima Dehmamy, and Rose Yu. Generative adversarial symmetry discovery. In *International Conference on Machine Learning*, pages 39488–39508. PMLR, 2023.
- [33] Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Latent space symmetry discovery. *arXiv preprint arXiv:2310.00105*, 2023.
- [34] Takeru Miyato, Masanori Koyama, and Kenji Fukumizu. Unsupervised learning of equivariant structure from sequences. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:768–781, 2022.
- [35] Masanori Koyama, Kenji Fukumizu, Kohei Hayashi, and Takeru Miyato. Neural fourier transform: A general approach to equivariant representation learning. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [36] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulencard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so(3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12200–12209, 2021.

- [37] user1551 (<https://math.stackexchange.com/users/1551/user1551>). Solve $\|xa - b\|$ subject to $xc = cx$. Mathematics Stack Exchange. URL <https://math.stackexchange.com/q/4886103>. URL:<https://math.stackexchange.com/q/4886103> (version: 2024-03-23).
- [38] Nicolas Donati, Abhishek Sharma, and Maks Ovsjanikov. Deep geometric functional maps: Robust feature learning for shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8592–8601, 2020.
- [39] Souhaib Attaiki and Maks Ovsjanikov. Understanding and improving features learned in deep functional maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1326, 2023.
- [40] Jiashi Feng, Zhouchen Lin, Huan Xu, and Shuicheng Yan. Robust subspace segmentation with block-diagonal prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3818–3825, 2014.
- [41] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [42] Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. Diffusionnet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics (TOG)*, 41(3):1–16, 2022.
- [43] Mircea Mironenco and Patrick Forré. Lie group decompositions for equivariant neural networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=p34fRKp8qA>.
- [44] Z Lian, A Godil, B Bustos, M Daoudi, J Hermans, S Kawamura, Y Kurita, G Lavoua, P Dp Suetens, et al. Shape retrieval on non-rigid 3d watertight meshes. In *Eurographics workshop on 3d object retrieval (3DOR)*. Citeseer, 2011.
- [45] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021.
- [46] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022.
- [47] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy, 2023.
- [48] Dmitriy Smirnov and Justin Solomon. Hodgenet: Learning spectral geometry on triangle meshes. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021.
- [49] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. *arXiv preprint arXiv:2404.15259*, 2024.
- [50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [51] Michael Kazhdan, Jake Solomon, and Mirela Ben-Chen. Can mean-curvature flow be modified to be non-singular? In *Computer Graphics Forum*, volume 31, pages 1745–1754. Wiley Online Library, 2012.
- [52] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009.

A Derivations and Implementation Details

A.1 Isometries in the Eigenbasis

Here we show that isometries manifest as orthogonal matrices that commute with the diagonal matrix of eigenvalues in the operator eigenbasis. Suppose we are given a PSD operator Ω and diagonal mass matrix \mathbb{M} , with the former expressed in terms of its eigendecomposition as in Equation (4) for eigenfunctions Φ and eigenvalues Λ .

Let τ be an isometric functional map satisfying the conditions in Equation (3) with τ_Ω its projection into the eigenbasis as in Equation (5). Noting that the condition $\Phi^\top \mathbb{M} \Phi = I_{|N|}$ implies that $\Phi^{-1} = \Phi^\top \mathbb{M}$ and $\Phi^{-\top} = \mathbb{M} \Phi$, it follows that

$$\begin{aligned} \tau_\Omega^\top \tau_\Omega &= [\Phi^\top \tau^\top \mathbb{M} \Phi] \Phi^\top \mathbb{M} \tau \Phi \\ &= \Phi^\top \tau^\top \mathbb{M} \tau \Phi \\ &= \Phi^\top \mathbb{M} \Phi = I_{|N|}, \end{aligned}$$

and thus τ_Ω is orthogonal. Furthermore, it follows that

$$\begin{aligned} \tau_\Omega \Lambda &= \Phi^\top \mathbb{M} \tau \Phi \Lambda \\ &= \Phi^\top \mathbb{M} \tau \Phi \Lambda \Phi^\top \mathbb{M} \Phi \\ &= \Phi^\top \mathbb{M} \tau \Omega \Phi \\ &= \Phi^\top \mathbb{M} \Omega \tau \Phi \\ &= \Lambda \Phi^\top \mathbb{M} \tau \Phi = \Lambda \tau_\Omega, \end{aligned}$$

so $\tau_\Omega \Lambda = \Lambda \tau_\Omega$.

A.2 Recovering τ_Ω

Here we derive the approximate solution to the least squares problem in Equations (7–8). Specifically, let $\Lambda = \text{diag}(\{\lambda_i\}_{1 \leq i \leq k}) \in \mathbb{R}^{k \times k}$ be a diagonal matrix of non-negative eigenvalues with $\Lambda_{ii} \leq \Lambda_{i+1, i+1}$. Furthermore, suppose that there are q distinct eigenvalues $\{\tilde{\lambda}_i\}_{1 \leq i \leq q}$, $\tilde{\lambda}_i \leq \tilde{\lambda}_{i+1}$ each with multiplicity m_i . Thus, Λ can be expressed as the direct sum

$$\Lambda = \bigoplus_{i=1}^q \tilde{\lambda}_i I_{m_i}, \quad (12)$$

with I_{m_i} denoting the $m_i \times m_i$ identity matrix. It follows that any $k \times k$ matrix π satisfying $\pi \Lambda = \Lambda \pi$ must be of the form

$$\pi = \bigoplus_{i=1}^q \pi_i, \quad \pi_i \in \mathbb{R}^{m_i \times m_i}, \quad 1 \leq i \leq q. \quad (13)$$

That is, π must be a block diagonal matrix with the size of each block determined by the multiplicity of the eigenvalues of Λ .

Now, given any $A, B \in \mathbb{R}^{k \times d}$ consider the minimization problem

$$\pi^* = \underset{\pi^\top \pi = I, \pi \Lambda = \Lambda \pi}{\text{minimum}} \|\pi A - B\|, \quad (14)$$

consisting of finding an orthogonal, Λ -commuting map minimizing the distance between A and B under the Frobenius norm. Writing

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_q \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_1 \\ \vdots \\ B_q \end{bmatrix}, \quad (15)$$

the solution to the minimization problem in Equation (14) is equivalently expressed as the direct sum of the solutions to the orthogonal sub-problems corresponding to each block [37] with

$$\pi^* = \bigoplus_{i=1}^q \underset{\pi_i^\top \pi_i = I}{\text{minimum}} \|\pi_i A_i - B_i\|. \quad (16)$$

Now, for arbitrary n let $\varkappa : \mathbb{R}^{n \times n} \rightarrow \mathbf{O}(n)$ denote the Procrustes projection to the nearest orthogonal matrix such that for any $M \in \mathbb{R}^{n \times n}$ with SVD $M = U\Sigma V^\top$, $\varkappa(M) \equiv UV^\top$. Thus, the solution to the minimization problem is given by

$$\pi^* = \bigoplus_{i=1}^q \varkappa(B_i A_i^\top), \quad (17)$$

which, following from the properties of the SVD, can be expressed equivalently via a single $k \times k$ Procrustes projection such that

$$\pi^* = \varkappa \left(\bigoplus_{i=1}^q B_i A_i^\top \right). \quad (18)$$

Here we observe that by defining $P_\Lambda \in \mathbb{R}^{k \times k}$ to be the eigenvalue mask given by

$$[P_\Lambda]_{ij} = \begin{cases} 1, & \lambda_i = \lambda_j \\ 0, & \text{otherwise} \end{cases} \quad (19)$$

we have

$$P_\Lambda \odot BA^\top = \bigoplus_{i=1}^q B_i A_i^\top, \quad (20)$$

and thus

$$\pi^* = \varkappa(P_\Lambda \odot BA^\top). \quad (21)$$

In practice we substitute P_Λ in Equation (19) with the fuzzy analogue $[P_\Lambda]_{ij} = \exp(-|\lambda_i - \lambda_j|)$, which we find better facilitates the flow of backwards gradients to the parameters of Ω . Replacing A and B with $\mathcal{E}_\Omega(\psi)$ and $\mathcal{E}_\Omega(T\psi)$ we arrive at the approximate solution for τ_Ω in Equation (8).

A.3 Graph Laplacian of P_Λ

Given the $k \times k$ eigenvalue mask P_Λ with $[P_\Lambda]_{ij} = \exp(-|\lambda_i - \lambda_j|)$, we form its graph Laplacian as

$$\Delta_{P_\Lambda} \equiv \text{diag}(P_\Lambda \mathbf{1}) - P_\Lambda, \quad (22)$$

with $\mathbf{1} \in \mathbb{R}^k$ the vector of ones.

B Reproducibility

All code for replicating experiments will be made available at publication.

B.1 Hardware

All experiments were performed on a single NVIDIA A6000 GPU with 48 GB of memory.

B.2 Approximating the Toric Laplacian

The model consists of a single NISO layer, which learns a full-rank operator and mass matrix, with the former parameterized by an orthogonal matrix of eigenvectors $\Phi \in \mathbb{R}^{256 \times 256}$ and diagonal matrix of non-negative eigenvalues Λ . During training, input features are created by stacking 86 $16 \times 16 \times 3$ examples from ImageNet into a single $16 \times 16 \times 258$ tensor representing an observation ψ . Random circular shifts about both spatial axes are applied to form $T\psi$. As the encoder and decoders are the identity maps, the equivariance and reconstruction losses are equivalent, and the objective in Equation (11) is minimized with $\alpha = 0$ and $\beta = 0.5$. The model is trained for 100,000 iterations with a batch size of 1 using the AdamW optimizer [50] with a weight decay of 10^{-4} . The learning rate follows a schedule consisting of a 2,000 step warm up from 0.0 to 5×10^{-4} and afterwards decays to 5×10^{-5} via cosine annealing. Training takes approximately 6 hours.

B.3 Homography-Perturbed MNIST

Architecture. The encoder and decoder are mirrored 2-level ConvNets. Specifically, each level consists of three convolutional ResNet blocks with 128 and 256 channels at the finest and coarsest resolutions respectively. A mean pool (nearest neighbor unpool) layer halving (doubling) the spatial resolution bridges the two levels.

The latent space has 32 dimensions. A rank $k = 32$ operator Ω and mass matrix \mathbb{M} are learned, with the former parameterized by 32 eigenfunctions and non-negative eigenvalues.

The equivariant classification head consists of a 2-layer VN-MLP [36] with 128 channels, followed by an output layer which computes invariant features from the inner products between vectors which are passed to a standard linear layer to output class predictions.

Pre-Training. Following [6], MNIST digits are padded to 40×40 , and pairs are produced by warping digits with respect to homographies T sampled from the proposed distribution. In the triplet regimes, tuples are created by applying T^2 in addition to T . Both the pairwise and triplet regimes are trained with respect to the composite loss with $\alpha = 0.5$ and $\beta = 0.1$. The two ablative regimes are trained with $\alpha = 0$ and $\beta = 0$, respectively. In the triplet regime, denoting τ and σ to be the maps estimated between $(\mathcal{E}(\psi), \mathcal{E}(T\psi))$ and $(\mathcal{E}(T\psi), \mathcal{E}(T^2\psi))$ respectively, the equivariance and reconstruction losses are formulated as

$$\mathcal{L}_E = \|\sigma_\Omega \mathcal{E}_\Omega(\psi) - \mathcal{E}_\Omega(T\psi)\| + \|\tau_\Omega \mathcal{E}_\Omega(T\psi) - \mathcal{E}_\Omega(T^2\psi)\| \quad (23)$$

and

$$\mathcal{L}_R = \|\mathcal{D}(\sigma \mathcal{E}(\psi)) - T\psi\| + \|\mathcal{D}(\tau \mathcal{E}(T\psi)) - T^2\psi\|. \quad (24)$$

The autoencoders are trained for 50,000 steps with a batch size of 16 using the AdamW optimizer with a weight decay of 10^{-4} . The learning rate follows a schedule consisting of a 2,000 step warm up from 0.0 to 5×10^{-4} and afterwards decays to 5×10^{-5} via cosine annealing. Training takes approximately 30 minutes.

Fine-Tuning. During the fine tuning phase, the decoder is discarded, and the weights of the encoder, operator Ω , and mass \mathbb{M} are frozen. Examples ψ are encoded, and their eigenspace projections \mathcal{E}_Ω are passed to the classification head to form class predictions under a standard softmax cross-entropy loss. Training is performed for 10,000 iterations with a batch size of 16 using the AdamW optimizer the same weight decay and learning rate schedule as used previously. Training takes approximately one minute. Evaluation is performed on the fixed test set proposed in [6].

Baselines. As no code release exists for the NFT, we implement it following the procedure described in the paper [35]. While a ViT-based encoder is proposed in the original paper, we find that it produces worse evaluation results than the 2-layer ConvNet encoder described above and thus use the latter in our evaluations to provide a fair comparison. We follow the authors’ proposed implementation and consider only a reconstruction loss and learn a diagonalization transform in a post-processing step. Otherwise, the NFT is pre-trained, fine-tuned, and evaluated identically to our own approach, and uses the same equivariant classification head. We note that the dimension of the latent map constructed by the NFT is the same as τ_Ω (32×32) and that the tensor passed to the classification head is the same size for both our method and the NFT.

The autoencoder baseline consists of the same encoder and decoder, pre-trained identically but with respect to the standard reconstruction loss $\|\mathcal{D}(\mathcal{E}(\psi)) - \psi\|$ without considering T -related pairs. During the fine tuning phase, latents are passed directly to a standard non-equivariant 2-layer MLP with 128 channels to form class predictions. Otherwise, fine-tuning and evaluation are performed in the same manner.

B.4 Conformal Shape Classification

Dataset. Experiments are performed using the conformally extended SHREC ‘11 shape classification dataset proposed in [7], with 30 random conformal transformations are applied to each shape to extend the original dataset [44]. The augmented dataset contains 25 distinct shape classes. Each shape is mapped to the sphere using the method of [51]. Input features are taken to be the values of the heat kernel signature [52] computed on the original mesh at 16 timescales logarithmically distributed in the range $[-2, 0]$ and rasterized to 96×192 spherical grids. Pairs of T -related observations are formed

by randomly selecting two conformal augmentations of the same shape. For each set of training and fine-tuning runs, train and evaluation splits are randomly generated by selecting respectively 10 and 4 of the 20 sets of conformally augmented shapes per class.

Architecture. The encoder and decoder are mirrored 4-level ConvNets, with two ResNet blocks per level and 32, 64, 128, 256 channels per layer from finest to coarsest resolution. Mean pooling and nearest neighbor upsampling are used to halve and double the resolution between layers.

The latent space has 64 dimensions. A rank $k = 64$ operator Ω and mass matrix \mathbb{M} are learned, with the former parameterized by 64 eigenfunctions and non-negative eigenvalues.

The equivariant classification head consists of a 4-layer VN-MLP with 128 channels, using the same output layer as described in sec. (B.3) to produce class predictions.

Pre-Training. Training is performed with respect to the composite loss with $\alpha = 0.5$ and $\beta = 0.01$. The model is trained for 50,000 iterations with a batch size of 16 using the AdamW optimizer with a weight decay of 10^{-4} . The learning rate follows a schedule consisting of a 2,000 step warm up from 0.0 to 5×10^{-4} and afterwards decays to 5×10^{-5} via cosine annealing. Training takes approximately four hours.

Fine-Tuning. During the fine tuning phase, the decoder is discarded and the encoder, operator Ω , and mass matrix \mathbb{M} are jointly optimized with the equivariant classification head using the standard softmax cross-entropy loss. Training is performed for 10,000 iterations with a batch size of 16 using the AdamW optimizer, with the same weight decay and learning rate schedule as used previously. Training takes approximately one hour.

Baselines. We pre-train, fine-tune, and evaluate the NFT and autoencoder baseline using the same autoencoder architecture, classification head, and training regimes as our own method in the same manner as sec. (B.3). Again, the both the size of the latent map and size of the tensor passed to the classification are the same size for both our method and the NFT. Here, the weights of the encoder are left unfrozen, and are jointly optimized with the classification head during the fine-tuning phase.

B.5 Camera Pose Estimation from Real-World Video

Dataset. Experiments are performed with a subsection of the CO3Dv2 dataset [45]. Given the high-degree of variability in the quality of the ground truth trajectories, we consider only the top 25% of sequences in the dataset as ranked by the provided pose quality score, slightly under 9,000 trajectories in total. An evaluation set is formed by withholding 10% of said trajectories, with the rest used for training. The list sequences in the train and evaluation sets will be made available along with the code release. As the videos in the dataset are of different resolutions, we center crop each frame to the minimum of its height and width dimensions and resize to 144×144 .

Architecture. The encoder and decoder are mirrored 3-level ConvNets, with four ResNet blocks per level and 64, 128, 256 channels per layer from finest to coarsest resolution. Mean pooling and nearest neighbor resolution are used to halve and double the resolution between layers.

The latent space has 128 dimensions. A rank $k = 128$ operator Ω parameterized by and mass matrix \mathbb{M} are learned, with the former parameterized by 128 eigenvectors and eigenvalues.

The pose extraction network consists of a standard 5-layer MLP with 512 channels, which outputs 9-dimensional tensor representing the parameters of an SE(3) transformation (translation + two vectors used to form the rotation matrix through cross products).

Pre-Training. During pre-training, T -related pairs of observations are formed by randomly selecting adjacent frames from sequences with a frame skip between 0 and 10. Training is performed with respect to the composite loss with $\alpha = 0.5$ and $\beta = 0.025$. The model is trained for 200,000 iterations with a batch size of 8 using the AdamW optimizer with a weight decay of 10^{-4} . The learning rate follows a schedule consisting of a 2,000 step warm up from 0.0 to 5×10^{-4} and afterwards decays to 5×10^{-5} via cosine annealing. Training takes approximately 12 hours.

Fine-Tuning. During the fine-tuning phase, the weights of the encoder, operator Ω , and mass matrix \mathbb{M} are frozen. As in the pre-training phase, pairs of observations are passed to the decoder and $\tau_{\Omega} \in \mathbb{R}^{128 \times 128}$ is estimated between the eigenspace projection of the encodings. Then, τ_{Ω} is vectorized and passed to the MLP head to predict the parameters of the camera pose.

Training is performed with respect to a two term loss measuring the degree to which the predicted translation and rotation deviate from those of the ground truth pose. Denoting R_Ω and R to be the predicted and ground truth poses respectively, the rotational component of the loss is given by

$$\mathcal{L}_R = \|R - R_\Omega\|. \quad (25)$$

The translational component of the loss must be scale invariant as the depth scale factor is unknown between sequences. To handle this, we consider sub-batches formed from pairs of frames *all* belonging to the same sequence. Denoting t_Ω and t to be the matrices consisting of the sub-batched predicted and ground truth translations stacked column-wise, we form the translational component of the loss via

$$\mathcal{L}_T = \|t - s \cdot t_\Omega\|, \quad (26)$$

where

$$s = \frac{\text{tr}(t^\top t_\Omega)}{\|t_\Omega\|^2} \quad (27)$$

is the scale factor that minimizes \mathcal{L}_T over the pairs of frames in the sub-batch. The total loss is given by $\mathcal{L} = \mathcal{L}_R + \mathcal{L}_T$

The prediction head is trained for 20,000 iterations with a batch size of 16 (consisting of 4 sub-batches of pairs of frames from 4 sequences) using the AdamW optimizer with the same weight decay and learning rate schedule as used previously. Training takes approximately one 80 minutes.

Baselines. The NFT is pre-trained, fine-tuned, and evaluated using the same autoencoder architecture, prediction head, and training regime as our own method in the same manner as is done in the preceding experiments. We note that both our method and the NFT construct latent maps of the same dimension, which are passed to identical prediction heads.

The transformer baseline is based on the CroCo/DUS3R[46, 47] architectures. As with our approach, input frames are passed to the same decoder. Here the encoder consists of 12-layer ViTs with a patch size of 16, with 16 heads and 512 channels per layer. Adjacent frames are independently encoded, and then concatenated along the spatial dimension in addition to a single learned token. The resulting tensor is passed to a 12-layer transformer decoder which mirrors the encoder with 16 heads and 512 channels per layer. Subsequently, the learned token is extracted from the output and passed through a linear layer to recover the predicted pose parameters. The model is trained from scratch in the fine-tuning phase, using the same loss and optimization procedure as described above.

C Additional Results

Below we provide additional qualitative results including visualizations of the mass matrices \mathbb{M} learned in each experiment and additional comparisons of estimated camera trajectories recovered by each method in the pose prediction experiments. For the latter, examples of failure cases are also included.

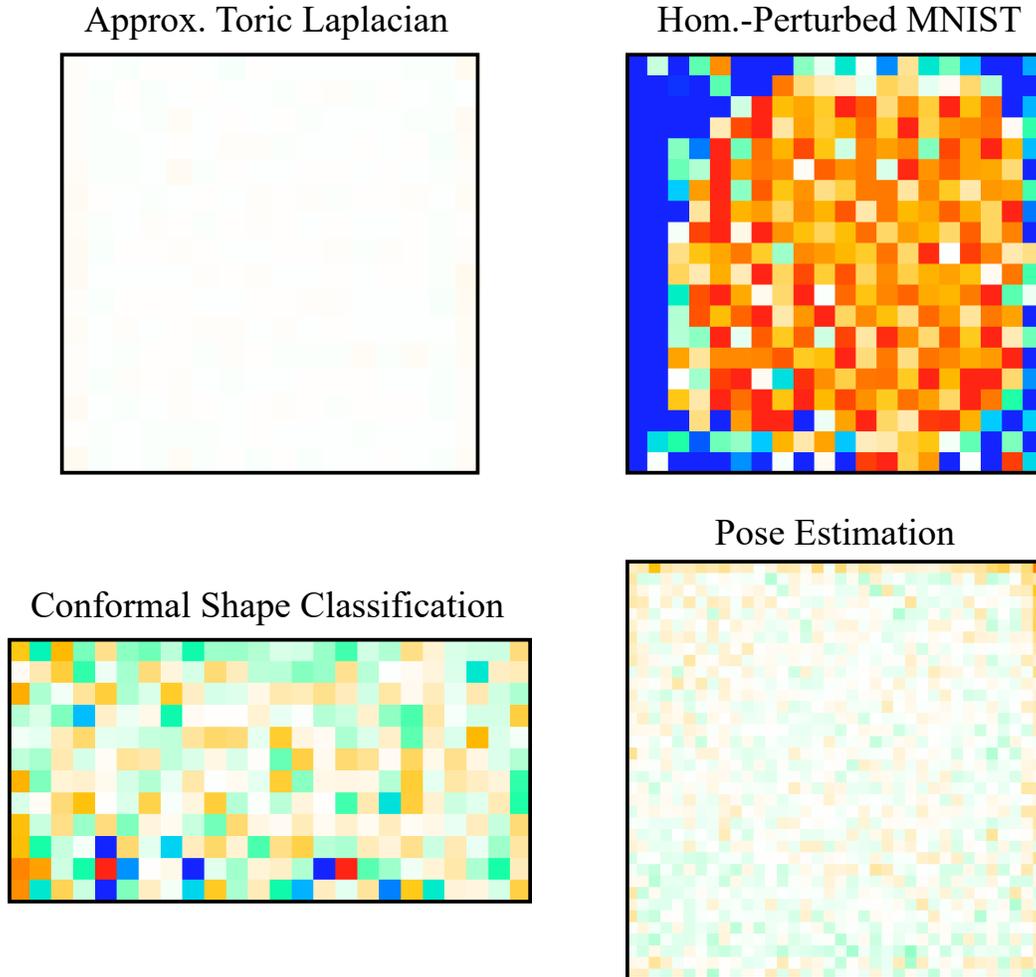


Figure 4: **Visualizing the Mass Matrix \mathbb{M} Elements at Grid Indices.** Visualizations of the elements of the diagonal mass matrices \mathbb{M} learned in each experiment, in terms of the magnitude of the deviation from the mean value at each grid index in the latent space. White indicates little deviation from the mean, with green-blue indicating mass values above the mean and orange-red indicating mass values below. When approximating the toric Laplacian (sec. 4.3), the learned mass matrix is approximately constant, mirroring the standard geometry of the toric grid, which assumes unit mass at each index. In the MNIST experiments (sec. 5.1), the distribution of mass appears to segment null space from the central region most often occupied by the digits. In the conformal shape classification experiments (sec. 5.2), the larger deviations from the mean values appear closer to the poles (the top-most and bottom-most rows of the spherical grid). For the post estimation experiments (sec. 5.3), larger deviations appear at the boundaries, with the lower half of the grid having slightly higher values.

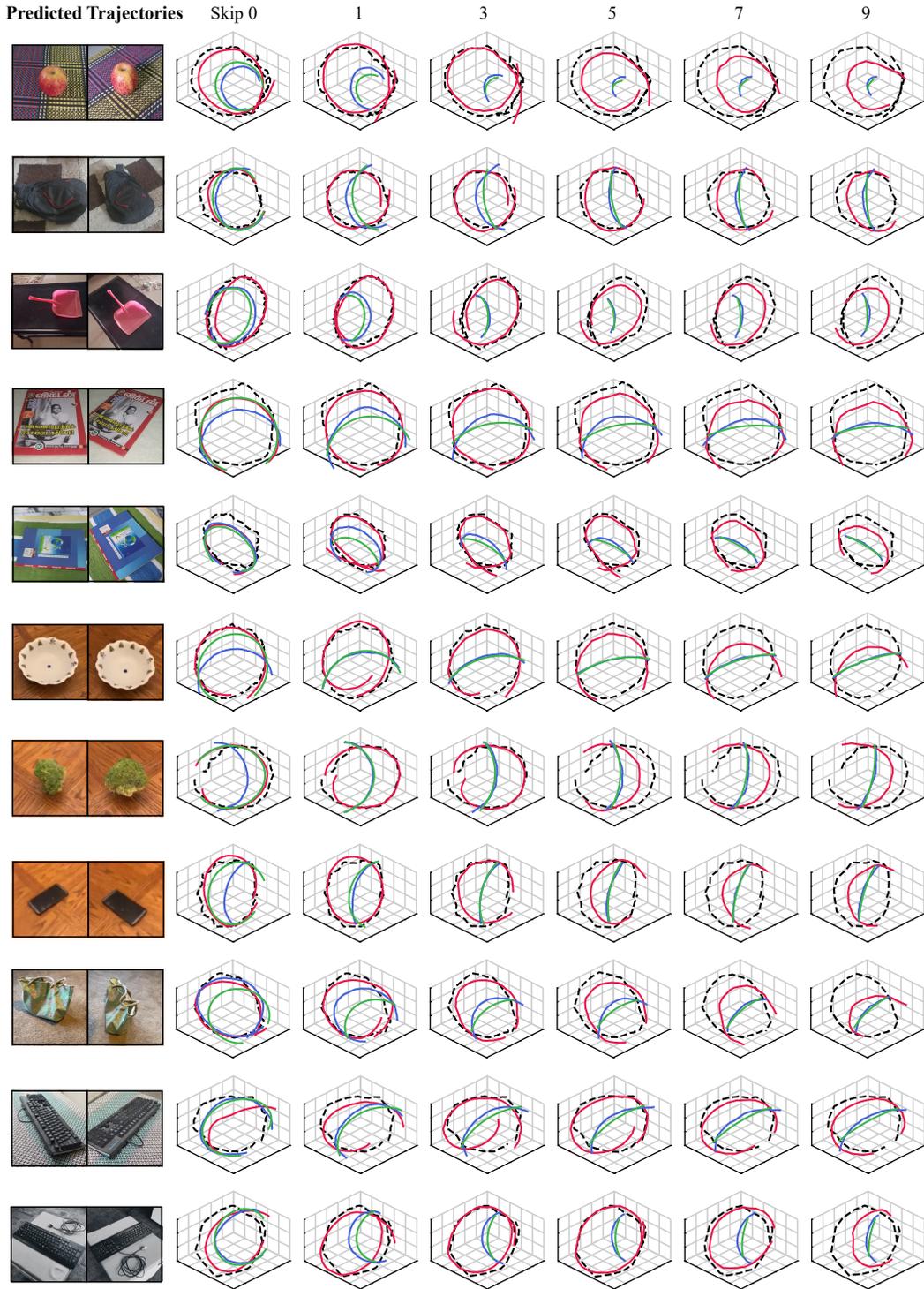


Figure 5: **Qualitative Pose Estimation Comparisons.** Example predicted trajectories for each method on select CO3Dv2 evaluation sequences. Ground truth is shown in black, NISO in red, the NFT in blue, and the transformer baseline in green. NISO appears to consistently better capture rotational (curvature) information about the world space transformations, helping it to better track the the camera motion across scales.

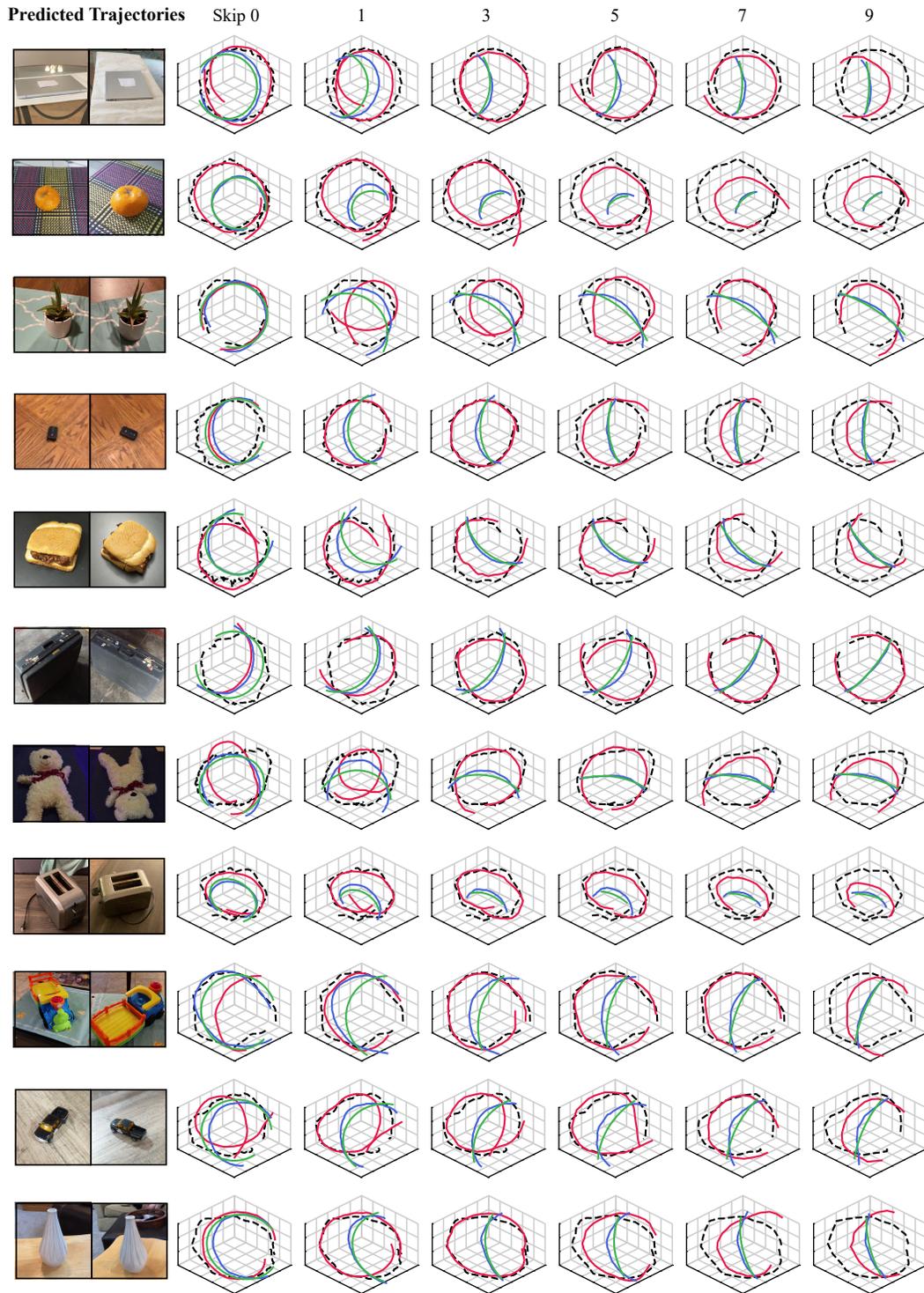


Figure 6: Qualitative Pose Estimation Comparisons Cont.



Figure 7: **Failure Cases.** Select failure cases. Interestingly, Niso appears to consistently fail when the background and foreground objects are respectively far from and close to the camera. In such cases, depth maps are often ill-defined, suggesting Niso may encourage models to reason about the underlying 3D structure of the scene.