SO(3)-Equivariant Representation Learning in 2D Images

Darnell Granberry
Simons Machine Learning Center
New York Structural Biology Center
New York, NY 10027
dgranberry@nysbc.org

Alireza Nasiri
Simons Machine Learning Center
New York Structural Biology Center
New York, NY 10027
anasiri@nysbc.org

Jiayi Shou
Simons Machine Learning Center
New York Structural Biology Center
New York, NY 10027
jshou@nysbc.org

Alex J. Noble
Simons Machine Learning Center
New York Structural Biology Center
New York, NY 10027
anoble@nysbc.org

Tristan Bepler
Simons Machine Learning Center
New York Structural Biology Center
New York, NY 10027
tbepler@nysbc.org

Abstract

Imaging physical objects that are free to rotate and translate in 3D is challenging. While an object's pose and location do not change its nature, varying them presents problems for current vision models. Equivariant models account for these nuisance transformations, but current architectures only model either 2D transformations of 2D signals or 3D transformations of 3D signals. Here, we propose a novel convolutional layer consisting of 2D projections of 3D filters that models 3D equivariances of 2D signals—critical for capturing the full space of spatial transformations of objects in imaging domains such as cryo-EM. We additionally present methods for aggregating our rotation-specific outputs. We demonstrate improvement on several tasks, including particle picking and pose estimation.

1 Introduction

Rotation and translation introduce challenges to many computer vision tasks including face and eye tracking [Liu, 2022], galactic imaging [Lintott et al., 2008], and cryogenic electron microscopy (cryo-EM) [Cheng et al., 2015, Sigworth, 2015]. In each case, the perceived object is free to move in three dimensions before being projected onto a 2D image plane. The object's identity remains the same and thus the information content of the image should be somewhat conserved. We aim to capture this symmetry with models incorporating more expressive 3D equivariances that still act on images alone.

Related Work Recent works have developed numerous techniques to achieve equivariance to transformations such as 2D rotation and scaling [Nasiri and Bepler, 2022, Marcos et al., 2017, Cohen and Welling, 2016]. Similarly, a wide variety of methods have been introduced to incorporate

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

equivariance to rotations in three dimensions [Thomas et al., 2018, Worrall and Brostow, 2018, Kondor et al., 2018]. However, none of these function directly in the 2D image domain.

2 Methods

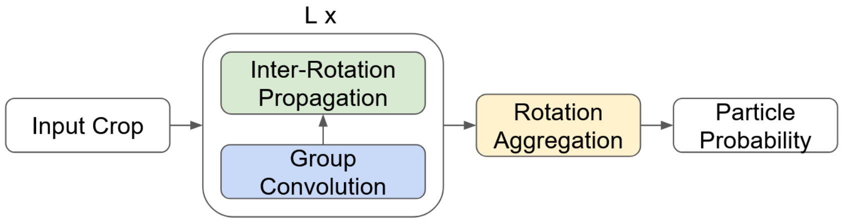

Figure 1: Particle picking model architecture, consisting of group convolutional modules to extract rotation-specific features, modules to propagate information between the corresponding rotations, and an aggregation module to synthesize the final rotation-specific feature maps into overall particle probabilities.

2.1 SO(3)-Equivariant Convolutional Layers

We begin by describing the cryo-EM image formation process. Each particle's volume V first adopts some orientation ϕ corresponding to a rotation $R \in SO(3)$ and a translation $t \in \mathbb{R}^3$. The volume's density is then projected along the Z-axis (according to our convention) via summation by P_Z into the 2D image plane. This projection removes information specific to Z-positions, so in typical formulations $t \in \mathbb{R}^2$. The image is finally subject to modulation by the microscope contrast transfer function C and the addition of noise W. Thus, the final observed 2D image becomes

$$I = (C \circ P_Z)(R(V) + t) + W \tag{1}$$

. To make our model equivariant to these rotation and projection operations, we generate convolutional filters by applying the same operations to 3D model weights. For each channel, a 3D weight is initialized. We then sample rotations from $SO(3)$, as described below, and apply them to the weight. Given these newly-rotated 3D arrays, we then project them by taking their means (which was more stable than summation) along the Z-axis. This produces 2D images of the weight in various orientations, which we use as filters in traditional 2D convolution. The 2D convolutions are grouped to ensure that filters and feature maps remain matched according to their orientations. This results in filters and layers that are equivariant to rotation and projection, in addition to possessing the translational equivariance of traditional CNNs. Our architecture, by virtue of the linear projection P_Z, is invariant to Z-axis translation. We can vary this property, and others, by varying our choice of projection P, as we discuss below. We provide a schematic of a typical particle picking model in Figure 1, of our layer's mechanics in Figure 2, and an illustration of particle and the corresponding filter projections in Figure 3.

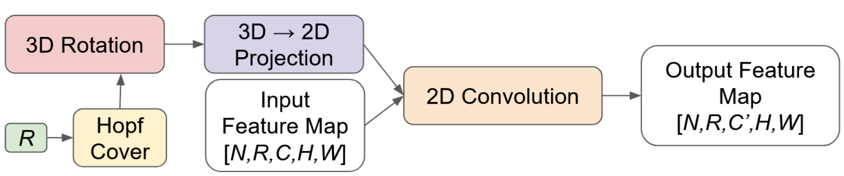

Figure 2: SO(3)-rotation-and- projection-equivariant convolutional module. Each module first rotates R copies of a 3D weight to unique orientations in SO(3). Each 3D weight is then projected into a 2D plane, which is finally convolved with the 2D input image or feature maps.

2.2 Rotation Sampling

In order to generate a finite number of projections about the group $SO(3)$, we need to first discretize it. We begin by finding points on the sphere–defined by two angles–with which to align the Z-axis

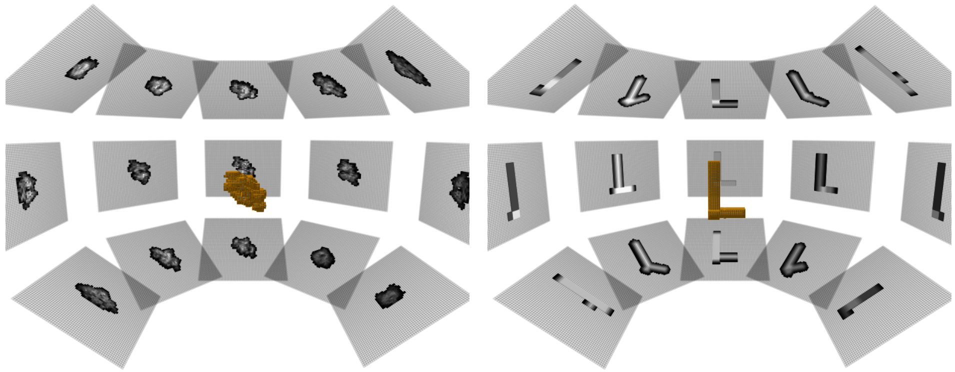

Figure 3: Left: an example protein (PBD 7QTQ) and its projections at various angles. Right: a example 3D filter that our model might store, with it's corresponding projections. In our module's forward pass, the particle projections will be convolved with the filter projections matching their orientations.

and then rotate by a third angle about these newly-aligned Z'-axes. We do so using a modification of the Hopf fibration [Yershova et al., 2010]. For simplicity, our models generate the first two angles using the Fibonacci sphere, a simple and relatively accurate sampling method [González, 2010]. We use the chosen angles to create 3D affine flow field grids, on which we sample the weights we are rotating, as used by Jaderberg et al. [2015].

2.3 Rotation Aggregation

For tasks that don't need rotation-specific features, we synthesize our feature vector into a single, invariant output. The first method we consider for this is max-pooling over rotations, corresponding to taking the score of the most likely rotation. Our second method–a novel approach–combines the probabilities associated with each orientation. Let Y be a Bernoulli random variable indicating whether an object is present in the given image, and let $\{y_i, i \in [1, R]\}$ be Bernoulli random variables indicating whether an object is present with rotation i. Thus:

$$P(Y = 1) = 1 - P(Y = 0) = 1 - \prod_{i=1}^{R} P(y_i = 0) \tag{2}$$

We deem this the "at least one" (AL1) aggregation. To maintain numerical stability, we compute the above in the log domain. We also add to each output logit a (negative) bias of $\log(2^{\frac{1}{R}} - 1)$ to counteract the increasing false-positive rate associated with increasing R [Benjamini and Hochberg, 1995].

3 Experiments

3.1 Image Classification

We evaluate our architectures on a variety of cryo-EM datasets, the common image dataset CIFAR-10 [Krizhevsky, 2009], Galaxy Zoo 2–a subset of the Galaxy Zoo/Sloan Digital Sky Survey dataset [Willett et al., 2013]. We compare a variety of models with a range of equivariances: a linear model, a CNN, a ResNet, an SO(2)-equivariant model (in-plane rotations only), and our SO(3)-equivariant models. Comparing various levels of equivariance allows us to observe the performance gains associated with explicitly modeling each type of symmetry.Equivariant models are tested with our AL1 aggregation. We also evaluate a model using a single SO(3)-equivariant layer on the cryo-EM datasets in an attempt to generate rough 3D models of the underlying particles.

Our cryo-EM datasets consisted of EMPIAR-10025 [Campbell et al., 2015], EMPIAR-10028 [Wong et al., 2014], and EMPIAR-11076 [Ehrenbolger et al., 2020]. We used Topaz [Bepler et al., 2019] to downsample all micrographs to 8 Å/pixel and extract 200,000 45x45 crops with 10% positive labels, mimicking the sparsity of cryo-EM labels. Galaxy Zoo 2 (GZ2) images were 424x424 pixels, so they were center-cropped to 180 pixels (minimally clipping the galaxies pictured), resized to 45 pixels. CIFAR-10 images are 32x32 pixels, so they were simply normalized, as were the other datasets. The cryo-EM datasets were split into 70% training, 15% validation, and 15% testing. The CIFAR and GZ2 datasets already contain train/test splits, so the training images were split to yield similar validation sets.

All classification models were trained to convergence on one 80GB NVIDIA A100 GPU for a maximum of 200 epochs using a batch size of 512. Early stopping was used with a patience of 5 epochs. We reduced the learning rate upon validation loss plateaus with patience of three epochs, the Adam optimizer [Kingma and Ba, 2014], and an initial learning rate of 10^{-3}. Binary classification models were evaluated on the area under the precision-recall curve and trained using binary cross-entropy loss, while multi-class classification models used categorical cross-entropy.

3.2 Pose Estimation

Here, we train models to predict an object's orientation from its 2D projection. Our model uses equivariant layers to output weights over r discretized 3D-rotations and offsets ϕ associated with each one of these rotations. In each r_i, the filters are rotated by angle θ_i, so to get the angles for each r_i dimension we use $\theta_i + \phi_i$. We train these models by minimizing

$$-logP(\theta_{pred}|r) = -\Sigma_i(logP(\theta_{pred}|r_i) + logP_i) \tag{3}$$

The first part of this loss function is calculated based on the quaternion distance between the predicted angles for r_k and θ_k, where k is the ground-truth rotation dimension. For calculating the second part of this loss function which is $logP_i$, we use cross-entropy loss. We identify the class assignments for both the ground-truth of samples using $argmin_i(1 - q\theta_i)^2$, where q is the ground-truth rotations in quaternions. We use this class assignment along with the weights over r, which is outputted by the model to calculate the $logP_i$.

We train our models on the projections of the (arbitrarily chosen) volume from [Imada et al., 1998]. For this volume, we have generated two datasets: one is based on the uniform sampling of the projection angles over SO(3), and the other one is based on the preferred orientation, where the projection angles are sampled from a Gaussian distribution with a standard deviation of 0.1 around a randomly sampled angle (the preferred orientation). Each training dataset has 10,000 samples, and we use a separate dataset of 1000 samples for testing.

We compare models using convolutional layers, SO(2)-equivariant layers, and multiple variants of our SO(3) model. In each, we compare the arc distance between the predicted and ground-truth quaternion components. We further evaluate how well each model can generalize to angles not seen in the training distribution, inspired by the important preferred orientation problem in cryo-EM [Cheng et al., 2015]. We are using the same number of filters in all these models. They are trained with the Adam optimizer, learning rate of 10^{-3} which is decayed by 0.9 after 10 iterations with no improvement in the loss, and batch size of 100 samples. We train all our models for 100 iterations and save the one with the best performance over the validation set.

4 Results

4.1 Image Classification

Our results in Tables 1 and 2 demonstrate that our models perform similarly or better than those without SO(3) equivariance, all while using fewer parameters. Models consisting of a single equivariant filter (2D or 3D), which are analogous to the linear baseline, do not perform as well as that baseline. This is likely due to the corners of the square/cubic weights being clipped during rotation, which results in the associated parameters being used less frequently and, therefore, less effectively. We could compensate for this by padding the filters before rotation then cropping the result to the desired size, which vastly increases the computational resources required to train.

Table 1: Binary cross-entropy loss, area under the precision-recall curve, and accuracy for various models on our cyro-EM datasets. Due to the imbalanced nature of our datasets, AUPR is our primary measure of model performance. Equivariant models generally outperform non-equivariant models. Our model does so using fewer parameters.

Dataset	**Method**	**Parameters**	**Loss**↓	**AUPR**↑	**Accuracy**↑
	Linear	2026	0.196	0.7982	0.9519
	Convolutional	74054	0.4301	0.8698	0.9529
	ResNet	73598	0.4218	0.8701	0.9549
	SO(2)-AL1	74137	0.0804	0.8765	0.9674
10025	SO(3)-AL1 1-filter	91126	0.6281	0.7851	0.733
	SO(3)-AL1	63369	0.0828	0.8817	0.9676
	Linear	2026	0.1929	0.8127	0.9546
	Convolutional	74054	0.4269	0.8656	0.9473
	Resnet	73598	0.4307	0.8608	0.9519
	SO(2)-AL1	74137	0.0882	0.8734	0.9644
10028	SO(3)-AL1 1-filter	91126	0.6226	0.8181	0.719
	SO(3)-AL1	63369	0.0899	0.8732	0.9641
	Linear	2026	0.1514	0.8014	0.9558
	Convolutional	74054	0.4357	0.8693	0.9516
	ResNet	73598	0.4178	0.8693	0.956
	SO(2)-AL1	74137	0.0783	0.8812	0.9673
11076	SO(3)-AL1 1-filter	91126	0.6281	0.7861	0.732
	SO(3)-AL1	63369	0.0836	0.8814	0.9676

Table 2: Classification statistics for generic image classification. Here, we use the cross-entropy loss as our primary indicator of performance. Equivariant models again outperform non-equivariant models. Our model does so using fewer parameters.

Dataset	**Method**	**Parameters**	**Loss**↓	**Accuracy**↑
	Linear	32680	1.9376	0.3388
	Convolutional	51104	1.3517	0.5568
	ResNet	89240	1.4501	0.5195
	SO(2)-AL1	51178	1.3276	0.5526
CIFAR-10	SO(3)-AL1	48842	1.2816	0.5573
	Linear	48608	1.3288	0.4771
	Convolutional	560176	0.7583	0.7666
	ResNet	639216	0.7863	0.7575
	SO(2)-AL1	802568	0.622	0.7764
Galaxy Zoo 2	SO(3)-AL1	691784	0.6412	0.7691

4.2 Pose Estimation

Our results in Table 3 show that our 3D group convolutional models outperform the other methods in correctly predicting rotation angles of projections. Furthermore, our models display greater generalizability; after being trained on narrowly-distributed data, they perform better than the others in predicting rotations of data sampled from outside of the training distribution.

Table 3: Test losses for various models on cryo-EM particle pose estimation. Evaluation scenarios are labeled with their training distribution/testing distribution pairs, each either uniform or adopting a preferred orientation (a randomly-centered Gaussian distribution). Our model performs best in all tasks. In particular, we show significant improvement over 2D and non-equivariant models when the underlying volume can adopt any orientation (the most general task) and when the training views are constrained (displaying enhanced generalizability).

Method	**R**	**Parameters**	**Unif/Unif**	**Pref/Pref**	**Pref/Unif**
Conv2D	–	2.74M	0.508	0.018	1.939
SO(2)	9	2.75M	0.231	0.029	1.926
SO(3)-Unimodal	256	2.80M	0.288	0.015	1.776
SO(3)	256	2.74M	0.14	0.103	1.156

5 Discussion

The models we present here demonstrate similar or better performance, greater generalizability, and improved parameter efficiency than non- and SO(2)-equivariant models in image classification and cryo-EM pose estimation. Additionally, our models generalize significantly better from data adopting preferred orientations. In generic image classification, it is unclear why our models' are more efficient than others. Though such images lack a single projected volume, their subjects still undergo rotation and projection; therefore, we hypothesize that modeling these operations provides some weaker inductive bias than for the comparatively restricted cryo-EM environment.

In the future, we will continue to examine our relationship between the discrete cover's density and performance, data efficiency, and deeper models with richer features. We also aim to explore the structure of our models' SO(3)-equivariant feature space, and applications like projection alignment. As we've demonstrated the usefulness of including richer symmetries, there are numerous areas for future work to explore. One area is formulating more complex projection operators that include properties like perspective and occlusion. Another area is adapting this approach to lower- or higher-dimensional signals. For example, architectures that model 4D transformations in 3D signals. Due to the curse of dimensionality, higher-dimensional applications will likely require even more sophisticated sampling and aggregation methods.

Acknowledgments and Disclosure of Funding

This work was supported by Simons Foundation grant SF349247. We'd like to thank the other members of the Simons Machine Learning Center, the Simons Electron Microscopy Center, and those of the Flatiron Institute for insightful discussions and suggestions.

References

Delilah Liu. New nvidia maxine cloud-native architecture delivers breakthrough audio and video quality at scale, 2022.

Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389(3): 1179–1189, 09 2008. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2008.13689.x. URL `https://doi.org/10.1111/j.1365-2966.2008.13689.x`.

Yifan Cheng, Nikolaus Grigorieff, Pawel A. Penczek, and Thomas Walz. A primer to single-particle cryo-electron microscopy . *Cell*, 161(3):438–449, 2015. URL https://doi.org/10.1016/j.cell.2015.03.050.

Fred J. Sigworth. Principles of cryo-EM single-particle image processing. *Microscopy*, 65(1):57–67, 12 2015. ISSN 2050-5698. doi: 10.1093/jmicro/dfv370. URL https://doi.org/10.1093/jmicro/dfv370.

Alireza Nasiri and Tristan Bepler. Unsupervised object representation learning using translation and rotation group equivariant VAE. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=qmm__jMjMlL.

Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5058–5067, 2017. doi: 10.1109/ICCV.2017.540.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Daniel Worrall and Gabriel Brostow. Cubenet: Equivariance to 3d rotation and translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 567–584, 2018.

Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch–gordan nets: a fully fourier space spherical convolutional neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

Anna Yershova, Swati Jain, Steven M. LaValle, and Julie C. Mitchell. Generating uniform incremental grids on so(3) using the hopf fibration. *The International Journal of Robotics Research*, 29(7): 801–812, 2010. doi: 10.1177/0278364909352700.

Álvaro González. Measurement of areas on a sphere using fibonacci and latitude–longitude lattices. *Mathematical Geosciences*, 42:49–64, 2010.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2017–2025, Cambridge, MA, USA, 2015. MIT Press.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society series b-methodological*, 57: 289–300, 1995.

Alex Krizhevsky. Learning multiple layers of features from tiny images. , University of Toronto, 2009.

Kyle W. Willett, Chris J. Lintott, Steven P. Bamford, Karen L. Masters, Brooke D. Simmons, Kevin R. V. Casteels, Edward M. Edmondson, Lucy F. Fortson, Sugata Kaviraj, William C. Keel, Thomas Melvin, Robert C. Nichol, M. Jordan Raddick, Kevin Schawinski, Robert J. Simpson, Ramin A. Skibba, Arfon M. Smith, and Daniel Thomas. Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 09 2013. ISSN 0035-8711. doi: 10.1093/mnras/stt1458. URL https://doi.org/10.1093/mnras/stt1458.

Melody G Campbell, David Veesler, Anchi Cheng, Clinton S Potter, and Bridget Carragher. 2.8 å resolution reconstruction of the thermoplasma acidophilum 20s proteasome using cryo-electron microscopy. *Elife*, 4:e06380, 2015.

Wilson Wong, Xiao-chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors HW Scheres. Cryo-em structure of the *Plasmodium falciparum* 80s ribosome bound to the anti-protozoan drug emetine. *eLife*, 3:e03080, jun 2014. ISSN 2050-084X. doi: 10.7554/eLife.03080. URL https://doi.org/10.7554/eLife.03080.

Kai Ehrenbolger, Nathan Jespersen, Himanshu Sharma, Yuliya Y. Sokolova, Yuri S. Tokarev, Charles R. Vossbrinck, and Jonas Barandun. Differences in structure and hibernation mechanism highlight diversification of the microsporidian ribosome. *PLOS Biology*, 18(10):1–15, 10 2020. doi: 10.1371/journal.pbio.3000958. URL `https://doi.org/10.1371/journal.pbio.3000958`.

Tristan Bepler, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J. Noble, and Bonnie Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature Methods*, 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0575-8. URL `https://doi.org/10.1038/s41592-019-0575-8`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL `https://arxiv.org/abs/1412.6980`.

Katsumi Imada, Kenji Inagaki, Hideyuki Matsunami, Hiroshi Kawaguchi, Hidehiko Tanaka, Nobuo Tanaka, and Keiichi Namba. Structure of 3-isopropylmalate dehydrogenase in complex with 3-isopropylmalate at 2.0 å resolution: the role of glu88 in the unique substrate-recognition mechanism. *Structure*, 6(8):971–982, 1998.