

Equivariant Symmetry-Aware Head Pose Estimation for Fetal MRI

Ramya Muthukrishnan¹, Borjan Gagoski^{2,3}, Aryn Lee², P. Ellen Grant^{2,3},
Elfar Adalsteinsson¹, Polina Golland¹, Benjamin Billot⁴

¹Massachusetts Institute of Technology

²Boston Children’s Hospital

³Department of Radiology, Harvard Medical School

⁴Inria, Université Côte d’Azur

¹{ramyamut, elfar, polina}@mit.edu,

^{2,3}{borjan.gagoski, aryn.lee, ellen.grant}@childrens.harvard.edu, ⁴bbillot@inria.fr

Abstract

We present E(3)-Pose, a novel fast pose estimation method that jointly and explicitly models rotation equivariance and object symmetry. Our work is motivated by the challenging problem of accounting for fetal head motion during a diagnostic MRI scan. We aim to enable automatic adaptive prescription of 2D diagnostic MRI slices with 6-DoF head pose estimation, supported by 3D MRI volumes rapidly acquired before each 2D slice. Existing methods struggle to generalize to clinical volumes, due to pose ambiguities induced by inherent anatomical symmetries, as well as low resolution, noise, and artifacts. In contrast, E(3)-Pose captures anatomical symmetries and rigid pose equivariance by construction, and yields robust estimates of the fetal head pose. Our experiments on publicly available and representative clinical fetal MRI datasets demonstrate the superior robustness and generalization of our method across domains. Crucially, E(3)-Pose achieves state-of-the-art accuracy on clinical MRI volumes, paving the way for clinical translation. Our implementation is available at github.com/ramyamut/E3-Pose.

1. Introduction

Real-time 6-DoF object pose estimation enables a wide range of critical navigation tasks in robotic manipulation [1], autonomous driving [7], and image-guided surgery [74]. Here, we present a novel method for rapid object pose estimation from 3D images based on an architecture that captures pose equivariance and object symmetries by construction. We are motivated by the problem of motion estimation in fetal MRI, where stacks of 2D diagnostic MRI slices are acquired

to assess fetal development and detect pathology [21, 22] (Appendix A). Due to inter-slice fetal motion, oblique (i.e., misaligned with the target anatomical orientation) slices and spatial coverage gaps in the acquired stack present challenges to radiological assessment and often necessitate re-acquisition of the entire stack (Fig. 1) [45, 68, 73]. One solution is to estimate the fetal head pose from rapid, low-resolution, low signal-to-noise ratio (SNR) 3D MRI “navigator” volumes inserted into the MRI sequence before every 2D slice. The pose estimated from each navigator volume can be used to adaptively adjust the imaging plane of the next 2D slice in real time (Fig. 1) [16, 28, 62]. Estimating fetal head pose from navigator volumes is particularly challenging. In addition to low resolution and low SNR, these volumes contain an imaging artifact that presents as a dark plane in the position of the previous (high energy) diagnostic 2D slice and may obstruct relevant anatomical features (e.g., eyes) that are crucial for pose disambiguation. Moreover, underdeveloped neuroanatomy coupled with low resolution creates a left-right symmetry, introducing pose ambiguities [75]. Finally, the lack of available navigator data necessitates training on higher resolution, higher SNR, artifact-free research volumes, thus creating cross-domain generalization challenges.

Existing pose estimation methods detect pre-specified anatomical landmarks to compute pose [31, 60], but landmark visibility is not guaranteed in clinical data. Another class of methods formulates the problem as rigid registration of a moving volume to a template [3, 49, 64]. However, in navigator volumes, artifacts perturb the intensity distributions of moving volumes relative to the template, leading to suboptimal registration. Canonical pose regression with convolutional neural networks (CNNs) can improve robustness in the presence of noise and reduced anatomical visibility [18, 56, 71]. Despite their benefits, these methods are prone to overfitting even with data augmentations, and are

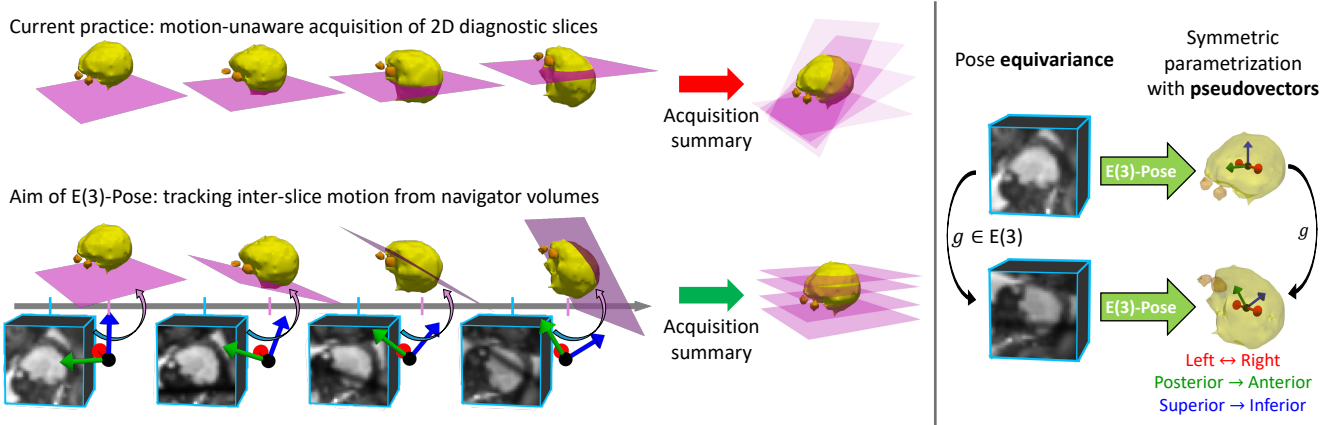


Figure 1. **E(3)-Pose is a rotation-equivariant and symmetry-aware framework for 6-DoF pose estimation.** *Left:* a rapid navigator volume is inserted between every two 2D diagnostic MRI slices. It is used to estimate the fetal head pose to adjust imaging plane prescription in real time. *Right:* To enable robust performance, the network architecture employs E(3)-equivariant convolutional filters to capture pose equivariance and pseudovectors to account for left-right head symmetry.

highly sensitive to pose ambiguities.

Contributions. We propose E(3)-Pose, a novel framework for accurate and generalizable head pose estimation from fetal brain volumes (Fig. 1). In contrast to standard convolutions that are by construction only equivariant to translations, we leverage an E(3)-equivariant CNN (E(3)-CNN) for pose regression to also account for rotation equivariance [19, 67]. Moreover, we use pseudovectors [27] in the learned equivariant rotation parametrization to model left-right head symmetry. E(3)-Pose achieves competitive performance on publicly available research-quality datasets, and state-of-the-art performance on *clinically* representative datasets. Crucially, our experiments demonstrate that constraining the network architecture to explicitly model physical symmetries provides better generalization to out-of-distribution data and stabilizes pose estimates in ambiguous cases with low anatomical visibility. Overall, while this paper focuses on fetal MRI, our method holds promise for object localization tasks that involve symmetries, limited training data, and real-world applications with noisy sensors.

2. Related work

Landmark-based pose estimation. Landmark registration involves algorithmic detection of pre-defined anatomical structures to estimate pose relative to a canonical anatomical frame [29, 31, 60]. To mitigate dependence on visibility of specific anatomy, recent approaches trained a neural network to learn an overcomplete set of landmarks [3, 63]. Nevertheless, learned landmarks display sensitivity to perturbations of volume intensities, posing problems when fast acquisitions introduce disruptive artifacts. In contrast, we directly regress pose, circumventing these challenges.

Pose regression. Early pose regression approaches relied on rotation parametrizations that suffer from discontinuities over $SO(3)$, such as axis-angle vectors or quaternions [24, 56]. It has been shown that high-dimensional, continuous parametrizations offer smoother loss landscapes and robustness to noise [20, 78]. Existing methods that adopt this strategy regress two basis directions [14, 78] or a deformation field [23] relative to the object frame. E(3)-Pose similarly regresses a continuous, overcomplete parametrization by using three basis directions of the canonical object frame.

Equivariant networks. A function is equivariant under a group of transforms if it commutes with every transform in the group. In the context of CNNs, standard convolutions are already equivariant to translations. There has been significant progress toward convolutions that are equivariant under a broader group, including $SE(3)$ and $E(3)$ [9–11, 19, 61, 66, 67, 70]. Specifically, these architectures restrict the space of learnable parameters to those that strictly adhere to problem symmetries, rather than learning equivariance through data augmentation. Theoretical analysis shows that equivariant architectures are more sample-efficient, reducing overfitting especially when training data is limited [4, 39, 57]. By leveraging an E(3)-equivariant CNN for pose estimation, E(3)-Pose robustly generalizes to noisy clinical data.

Existing equivariant methods for 6-DoF pose estimation either predict the relative transform between image pairs [3, 42, 49, 65, 76]; regress a canonical pose via a discontinuous rotation parametrization, such as quaternions [44, 50]; or formulate rotation estimation as a classification problem over a discretized set of 3D rotations [6, 33, 41]. Furthermore, these methods do not explicitly account for symmetric ambiguities in their network architectures. In contrast, E(3)-Pose

regresses a continuous output over $SO(3)$ and handles object symmetry in the construction of this parametrization.

Object symmetries. Rotational and reflectional object symmetries can lead to pose ambiguities. One class of solutions trains a network to implicitly learn object symmetries via probabilistic outputs, and heuristically selects the most probable pose during inference [30, 77]. Yet, these methods suffer from discontinuities in the network output over the object symmetry group [54]. Although training with symmetry-invariant losses partially alleviates this problem [71], these discontinuities are more robustly eliminated by incorporating object symmetry into the pose parametrization itself. Namely, this strategy regresses an output that remains *invariant* under the object symmetry group, while maintaining continuity over the quotient group of all non-symmetric input transformations [51, 54, 58]. To handle head pose ambiguities, E(3)-Pose models the basis direction along the axis of left-right symmetry as a pseudovector [27], which maintains both invariance under left-right reflections and continuity over $SO(3)$. While pseudovectors have been used to model rotation axes in pose estimation [40], our method is the first to use them to address symmetric ambiguities.

Automated slice prescription in fetal brain MRI. Existing implementations use the pose estimated from a single volume acquired at the start of the sequence to prescribe all diagnostic 2D slices in the stack [31, 60]. While this approach accurately prescribes slices according to the initial head pose, it does not account for inter-slice head motion. In contrast, E(3)-Pose promises to adjust the imaging plane for every slice in the stack, based on the head pose estimated from the preceding navigator volume.

3. Preliminaries

In this section, we provide context for E(3)-Pose by introducing key theoretical concepts in the design of E(3)-CNNs.

Tensor fields and E(3) representations. Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}^d$ be a d -dimensional tensor field. Consider 3D Euclidean group actions $g = g_t \circ g_r$, which comprise rotations/reflections $g_r \in O(3)$ and translations $g_t \in \mathbb{R}^3$. When the coordinate system is transformed by g , the tensor $f(x)$ at point $x \in \mathbb{R}^3$ is modified in two ways: it is moved to the location $g^{-1}x$ and the tensor is independently rotated/reflected by g_r [19, 67]. Formally, we can define an operator $\pi_f(\cdot)$ that transforms field f for any $g = g_t \circ g_r \in E(3)$:

$$[\pi_f(g)f](x) = \rho_f(g_r)f(g^{-1}x), \quad (1)$$

where $\rho_f : O(3) \rightarrow \mathbb{R}^{d \times d}$ is called a *representation*; it is a matrix acting linearly on the tensor space of $f(x)$ that describes how the tensor independently transforms under g_r .

Irreducibility. A representation ρ and the tensors it transforms are *irreducible* if there is no change of basis $Q \in \mathbb{R}^{d \times d}$

such that $Q\rho(g_r)Q^{-1}$ is block-diagonal for all $g_r \in O(3)$. Irreducible representations and tensors are characterized by order l and parity, defining how $f(x)$ transforms under proper rotations and inversions (i.e., reflections across all three planes) of the coordinate system, respectively [19]. Even parity indicates no change in $f(x)$ under inversion, while odd parity indicates a sign change in $f(x)$. While one can define tensors and representations of any order and parity, we focus here on three types: first, scalars $f(x) \in \mathbb{R}$ (order $l=0$, even parity) have representation $\rho_{\text{even}}^{l=0}(g_r) = 1$; second, vectors $f(x) \in \mathbb{R}^3$ ($l=1$, odd parity) have representation $\rho_{\text{odd}}^{l=1}(g_r) = M(g_r)$, where $M : O(3) \rightarrow \mathbb{R}^{3 \times 3}$ is the matrix representation of an orthogonal transformation; and last, pseudovectors $f(x) \in \mathbb{R}^3$ ($l=1$, even parity) have representation $\rho_{\text{even}}^{l=1}(g_r) = (\det M(g_r))M(g_r)$ with a positive determinant to account for invariance under inversions [27].

Equivariance. Let F be a mapping between two tensor fields f and h associated with operators π_f and π_h and representations ρ_f and ρ_h , respectively. F is equivariant under E(3) if and only if

$$F(\pi_f(g)f) = \pi_h(g)F(f), \text{ for all } g \in E(3). \quad (2)$$

A similar relation holds if F is a function mapping tensor field f to a single tensor h with representation ρ_h :

$$F(\pi_f(g)f) = \rho_h(g_r)F(f), \text{ for all } g = g_t \circ g_r \in E(3). \quad (3)$$

Equivariant networks. E(3)-CNNs can be built using successive equivariant layers, since their composition remains equivariant. Here we describe equivariant convolutional layers (see Appendix B for pooling and non-linearities). For enhanced computation and time efficiency, state-of-the-art implementations of such CNNs operate on irreducible fields [19, 67]. Let $\kappa_{l_f p_f, l_h p_h}(x) \in \mathbb{R}^{(2l_h+1) \times (2l_f+1)}$ be a 3D convolution kernel transforming field f (order l_f , parity p_f) to field h (order l_h , parity p_h). Then, κ must satisfy Eq. (2) to maintain equivariance, resulting in the solution

$$\text{vec}(\kappa_{l_f p_f, l_h p_h}(x)) = \sum_{m \in M} \sum_{j \in J} w_{jm} \varphi_m(\|x\|) Q_j Y_j(x/\|x\|), \quad (4)$$

where vec indicates vectorization, w_{jm} are coefficients, $\{\varphi_m\}$ are radially symmetric basis functions, $Q_j \in \mathbb{R}^{(2l_h+1)(2l_f+1) \times (2j+1)}$ is a change of basis, $Y_j : S^2 \rightarrow \mathbb{R}^{2j+1}$ are the degree- j spherical harmonics, and $J = \{j : |l_f - l_h| \leq j \leq |l_f + l_h|, j \text{ is even}\}$ if $p_f = p_h$ or $J = \{j : |l_f - l_h| \leq j \leq |l_f + l_h|, j \text{ is odd}\}$ if $p_f \neq p_h$ [19, 67]. E(3)-CNNs are thus trained by learning the coefficients w_{jm} rather than the kernel values themselves.

4. Methods

Our goal is to estimate 6-DoF pose T from a 3D volume, relative to a canonical object coordinate frame. As detailed

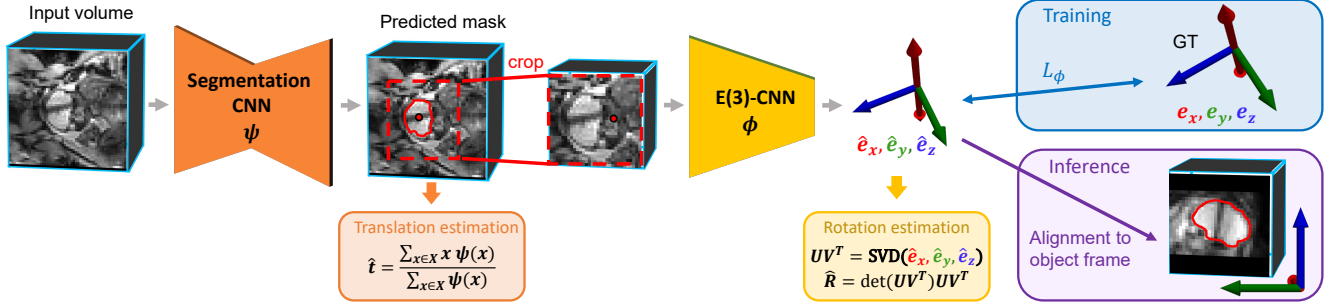


Figure 2. **Overview of E(3)-Pose.** We first train a CNN ψ to segment the object. We estimate translation based on the center-of-mass of the predicted mask and then crop the 3D volumes around this mask. The cropped volumes are fed to an E(3)-CNN ϕ trained independently to regress the orthonormal basis of the object frame, parametrized as one pseudovector (red) and two vectors (blue and green). The output is later constrained to represent a rotation matrix by applying SVD and then choosing the pseudovector \hat{e}_x direction that results in a proper rotation without reflection (i.e., $\det(M(\hat{R})) = 1$).

in Fig. 2, we estimate pose by separating the rigid transform $T = t \circ R$, into its rotation R and translation t components.

4.1. Translation estimation

In rigid pose estimation, it is customary to place the origin of the object coordinate frame at the object center [71]. If $\psi : \mathbb{R}^3 \rightarrow \{0, 1\}$ is a binary object occupancy function, then t can be computed from a set X of sampled points (i.e., voxel coordinates):

$$t = \frac{\sum_{x \in X} x \psi(x)}{\sum_{x \in X} \psi(x)}. \quad (5)$$

See Appendix C for our implementation of ψ using a network trained to segment the fetal brain segmentation from whole-uterus MRI volumes.

4.2. Rotation estimation

Function equivariance. We first define $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ as the scalar field representing the input volume, with transformation operator $[\pi_f(g)f](x) = f(g^{-1}x)$; $h(R) \in \mathbb{R}^d$ as some d -dimensional tensor parametrization of the output rotation R , with representation ρ_h ; and $\phi : f \rightarrow h$ as the rotation estimation function. Rotating/reflecting the input volume rotates/reflects its pose equivalently, requiring function equivariance detailed in Eq. (3):

$$\phi(f \circ g^{-1}) = \rho_h(g_r) \phi(f), \quad \text{for all } g = g_t \circ g_r \in E(3). \quad (6)$$

Although in practice we require equivariance under SE(3) rather than the full group E(3), we show below that using an E(3)-CNN to implement ϕ enables us to jointly formulate $h(R)$ as equivariant under SE(3) and invariant under left-right reflections. Since E(3)-CNNs must operate on irreducible tensors, it suffices to parametrize $h(R)$ with irreducible tensors to guarantee the equivariance of ϕ .

Rotation parametrization. Let $G_{\text{symm}} = \{g_s \in O(3) : f \circ g_s^{-1} = f\}$ be the object symmetry group [58]. To handle

object symmetries, we aim to parametrize $h(R)$ such that $\rho_h(g_s)h(R) = h(R)$ for all $g_s \in G_{\text{symm}}$ and $R \in SO(3)$; and h remains continuous over the quotient group $H = O(3)/G_{\text{symm}}$. Here, we consider the case $S = \{g_{\leftrightarrow}\}$, where g_{\leftrightarrow} denotes left-right reflection, defined by the axis of left-right symmetry e_{\leftrightarrow} . In this case, $H = O(3)/\{g_{\leftrightarrow}\} = SO(3)$.

Formally, we define $h(R)$ based on three orthonormal basis directions $\{e_x, e_y, e_z\}$ of the object frame ($d = 9$), a continuous function over $SO(3)$:

$$\begin{aligned} h(R) &= e_x \oplus e_y \oplus e_z, \\ \text{s.t. } e_x &= e_{\leftrightarrow} \quad \text{and} \quad e_y, e_z \perp e_{\leftrightarrow}, \\ \rho_h(g_r) &= (\det M(g_r))M(g_r) \oplus M(g_r) \oplus M(g_r) \\ &= \rho_{\text{even}}^{l=1}(g_r) \oplus \rho_{\text{odd}}^{l=1}(g_r) \oplus \rho_{\text{odd}}^{l=1}(g_r), \end{aligned} \quad (7)$$

where \oplus denotes columnwise and block-diagonal concatenation for vectors and matrices, respectively. Specifically, we decompose $h(R)$ into three irreducible tensors: we model the basis direction parallel to the left-right axis as a *pseudovector*, and the two basis directions orthogonal to the left-right axis as vectors. The invariance of $h(R)$ under G_{symm} is straightforward:

$$\begin{aligned} \rho_h(g_{\leftrightarrow})h(R) &= (\det M(g_{\leftrightarrow}))M(g_{\leftrightarrow})e_x \\ &\quad \oplus M(g_{\leftrightarrow})e_y \oplus M(g_{\leftrightarrow})e_z, \\ &= (\det M(g_{\leftrightarrow}))(-e_x) \oplus e_y \oplus e_z \\ &= e_x \oplus e_y \oplus e_z = h(R), \end{aligned} \quad (8)$$

where the second equality follows from the fact that vectors orthogonal and parallel to the axis of reflection remain unchanged and become inverted following the reflection, respectively. By using an E(3)-equivariant function to estimate $h(R)$ and further imposing invariance of $h(R)$ under G_{symm} , we ensure that $h(R)$ is equivariant under $E(3)/G_{\text{symm}} = SE(3)$ and continuous over $SO(3)$.

E(3)-CNN regressor. To handle both rotational equivariance and reflectional symmetry, we implement ϕ as an E(3)-CNN

regressor taking a scalar field as input, and returning an output tensor that decomposes into one pseudovector \hat{e}_x and two vectors \hat{e}_y, \hat{e}_z , normalized to unit length. Appendix B provides a detailed description of the network architecture. During training, we minimize the objective function

$$\mathcal{L}_\phi = |\sin(\theta_x)| + \beta(|\sin(\theta_y/2)| + |\sin(\theta_z/2)|), \quad (9)$$

where (e_x, e_y, e_z) is the ground-truth (GT) orthonormal basis, $\theta_k = \arccos(\hat{e}_k \cdot e_k)$ for $k \in \{x, y, z\}$, and β is a weight hyperparameter. We use $\sin(\theta/2)$ for the standard vectors since it is monotonically related to the cosine similarity loss. We modify this term to $\sin(\theta)$ for the pseudovector in order to guarantee invariance under inversions (i.e., left-right reflections) of e_x . During inference, we obtain an orthonormal output matrix by applying SVD to the predicted basis [43]. We then choose between \hat{e}_x and $-\hat{e}_x$ in order to guarantee a positive determinant matrix, i.e., a proper rotation. This step is related to prior work that heuristically selects the output pose from a set of symmetrically plausible poses [30, 51, 77]. Here, the task is straightforward since we are choosing between a right-handed and left-handed rotation.

4.3. Implementation details

Object frame. We place the origin of the target object frame at the center-of-mass (CoM) of the brain. We define the GT orthonormal basis e_x, e_y , and e_z as the unit vectors pointing along the left→right (L→R), posterior→anterior (P→A), and inferior→superior (I→S) anatomical orientations of the brain, respectively (Fig. 2). This follows Eq. (7), where e_x is parallel and e_y, e_z are orthogonal to the left-right axis, respectively.

Data augmentation. We augment training volumes with commonly used spatial (rigid transformations, scaling, flips) and intensity (bias field, gamma, Gaussian noise, low resolution) transforms [2, 52] in volumetric brain MRI. To enable our method to generalize to navigator volumes interleaved with diagnostic slices, we additionally simulate a spin history artifact from the preceding diagnostic slice [16, 47], which appears as dark shading along the slice imaging plane (blue arrows in Fig. 3). We model this artifact with a Gaussian approximation of the slice profile:

$$\tilde{f}(x) = f(x)(1 - \mathcal{N}((x - c_{\text{slice}})^T n_{\text{slice}}; 0, \sigma^2)), \quad (10)$$

where f and \tilde{f} are the original and augmented volumes, respectively; c_{slice} and n_{slice} are the center and normal to the slice imaging plane, sampled uniformly at random; and σ^2 is determined by slice thickness and the rate of spin decay [38] (see Appendix D for more details).

5. Experiments and results

5.1. Data

We train two variants of E(3)-Pose on two different large, high-quality *research* datasets. We first evaluate on in-distribution held-out test data. Then, we evaluate on two *clinical* datasets to assess domain generalization and the potential for clinical translation. All volumes are manually annotated for ground-truth pose and brain/eyes segmentations (see Appendix E for details).

Research. We train one instance of E(3)-Pose on **Research-Fetal**, a dataset of 3D whole-uterus MRI volumes in 153 pregnant women acquired at the Boston Children’s Hospital. We train another instance of our method on **dHCP**, a publicly available dataset of fetal brain MRI scans in 245 subjects [35]. Subject gestational age (GA) is 18-38 and 20-38 weeks in Research-Fetal and dHCP, respectively. All volumes have 2-3mm isotropic voxels. They are not interleaved with 2D diagnostic slices and do not contain spin history artifacts. We use training/validation/testing splits of 114/15/25 and 148/48/49 subjects for Research-Fetal and dHCP, respectively. dHCP consists of higher quality volumes, presenting a more challenging clinical domain generalization task compared to Research-Fetal. Additionally, dHCP volumes are neatly cropped around the fetal head. Therefore, segmentation networks trained on dHCP fail to accurately predict the brain mask in volumes where the field of view is larger and includes uterine regions outside the head and maternal tissues (Fig. 3). For this reason, we train our segmentation network only on Research-Fetal.

Clinical. We evaluate the methods on **Clinical-Young**, a dataset of routine clinical volumes in 60 *younger* fetuses (GA 18-23 weeks). This dataset presents two challenges: clinical volumes exhibit a substantial domain shift from high-quality research data, and underdeveloped anatomy complicates pose estimation [75]. The voxel size is 1.76-3.5mm, and volumes are not interleaved with diagnostic slices. We also test the models on **Navigators**, a dataset containing 47 time-series (1210 total volumes) of 3D MRI navigator volumes interleaved with 2D diagnostic slices in 9 subjects (GA 26-36 weeks). Due to fast and interleaved acquisitions, these volumes have voxel size 4-6mm and contain real spin history artifacts. This dataset represents the largest domain gap from the research data, and volumes are representative of the intended clinical application, i.e. adaptive slice prescription. Both clinical datasets are acquired at the Boston Children’s Hospital.

5.2. Evaluation metrics

We assess performance with the geodesic rotation error $\arccos(\frac{1}{2}[\text{trace}(R\hat{R}^{-1}) - 1])$ and the average absolute distance (AAD) for voxels on the brain surface [25] (see Ap-

Table 1. **Baseline comparisons.** Mean \pm standard deviation for rotation error ($^\circ$) and average absolute error (AAD, mm). Best score is in bold. * indicates statistical significance compared to E(3)-Pose ($p < 0.05$, pairwise Wilcoxon test). FireANTs and Fetal-Align rely only on segmentation networks trained on Research-Fetal and are not re-trained on dHCP, and therefore, we omit the corresponding rows for these methods. FireANTs runtime is 1.8s, prohibiting real-time deployment. EquiTrack runtime is 0.7s, and the remaining methods run in 0.3s.

	Research-Fetal test		dHCP test		Clinical-Young		Navigators	
	Rot. error	AAD	Rot. error	AAD	Rot. error	AAD	Rot. error	AAD
Trained on Research-Fetal								
FireANTs [34]	10.6 \pm 24.5*	5.7 \pm 9.3*	0.4 \pm 1.6*	0.6 \pm 1.0*	17.3 \pm 40.6	4.1 \pm 7.4	44.6 \pm 55.*	19.1 \pm 20.3*
EquiTrack [3]	11.3 \pm 13.1*	6.2 \pm 5.9*	7.1 \pm 5.0	3.6 \pm 2.6	18.9 \pm 29.0*	5.4 \pm 5.7*	44.7 \pm 48.7*	20.3 \pm 18.3*
Fetal-Align [31]	5.0 \pm 3.0	3.0 \pm 1.7	8.5 \pm 24.7*	3.5 \pm 6.8*	10.3 \pm 21.0*	3.2 \pm 5.0*	56.2 \pm 57.6*	24.1 \pm 23.3*
3DPose-Net [56]	34.0 \pm 36.0*	16.9 \pm 15.5*	29.5 \pm 35.6*	13.0 \pm 14.1*	35.7 \pm 35.4*	10.2 \pm 8.1*	48.1 \pm 40.6*	22.1 \pm 15.3*
6DRep [14]	11.8 \pm 6.8*	6.4 \pm 3.5*	12.5 \pm 7.1*	6.0 \pm 3.4*	15.7 \pm 9.1*	4.9 \pm 2.6*	33.2 \pm 9.5*	15.1 \pm 13.8*
RbR [23]	7.6 \pm 3.9*	4.3 \pm 2.2*	8.6 \pm 5.0	4.3 \pm 2.6	11.2 \pm 4.6*	3.7 \pm 1.4*	22.6 \pm 24.1*	11.5 \pm 9.9*
E(3)-Pose (ours)	5.1 \pm 2.6	3.0 \pm 1.7	7.4 \pm 3.6	3.7 \pm 1.8	9.1 \pm 4.7	3.0 \pm 1.4	9.4 \pm 7.5	6.3 \pm 4.0
Trained on dHCP								
EquiTrack [3]	29.5 \pm 43.5*	14.1 \pm 19.5*	13.6 \pm 36.7*	5.4 \pm 11.6*	48.3 \pm 62.1*	11.6 \pm 13.2*	59.0 \pm 57.5*	25.7 \pm 22.1*
3DPose-Net [56]	64.1 \pm 44.6*	29.9 \pm 18.7*	38.4 \pm 34.0*	17.0 \pm 13.2*	74.9 \pm 52.9*	18.9 \pm 10.3*	87.3 \pm 45.9*	36.1 \pm 16.4*
6DRep [14]	19.2 \pm 9.9*	10.4 \pm 5.5*	10.8 \pm 5.5*	5.2 \pm 2.6*	45.1 \pm 44.9*	11.9 \pm 9.8*	84.9 \pm 57.5*	33.0 \pm 17.4*
RbR [23]	15.4 \pm 18.3*	8.3 \pm 8.5*	8.0 \pm 4.0*	3.9 \pm 1.9	50.9 \pm 48.4*	12.7 \pm 9.9*	77.2 \pm 53.6*	30.8 \pm 17.3*
E(3)-Pose (ours)	5.7 \pm 3.6	3.4 \pm 2.0	7.3 \pm 3.4	3.7 \pm 1.7	12.1 \pm 6.8	3.9 \pm 1.9	13.9 \pm 13.2	7.9 \pm 5.6

pendix G for additional metrics).

5.3. Baseline methods

Template-based. We evaluate two baseline methods that perform rigid registration to a template volume. FireANTs [34] is a state-of-the-art method for fast, optimization-based 3D image registration. EquiTrack [3] uses a SE(3)-equivariant network that learns matching landmarks to compute the optimal rigid transform from the input to the template via SVD [43]. To enable comparison with E(3)-Pose, we generate subject-specific template volumes using the GT pose. For accurate registration, we mask input volumes with the brain masks predicted by our segmentation network.

Template-free. We evaluate four baseline methods that estimate the canonical pose from only the input volume. Fetal-Align [31] implements landmark-based pose estimation, which segments the brain and eyes as landmarks and uses brain shape to resolve left-right ambiguity. For fair comparison on our datasets, we use our trained segmentation network, which we train to segment both the brain and eyes, to perform landmark detection. The remaining three baselines directly regress pose with CNNs. 3DPose-Net [56] and 6DRep [14] regress the axis-angle vector and two basis directions of the object frame, respectively. The latter method applies Gram-Schmidt orthonormalization [79] to the network output to obtain the final rotation. Registration-by-Regression (RbR) [23] trains a U-Net to regress a deformation field to the object frame at the same resolution as the input volume. The final rotation is computed from the network output with SVD [32]. We train all three networks on the research datasets and tune augmentation parameters for optimal performance. We evaluate rotation only and use

our CoM translation estimation for all methods. Appendix F provides further details on the baseline methods.

5.4. Results

We first evaluate performance on in-domain research data, and then assess generalization to target clinical domains. Table 1 reports the performance statistics on the test volumes from each dataset, and Fig. 3 shows example alignments. Since FireANTs and Fetal-Align do not involve network training beyond our segmentation network, which is only trained on Research-Fetal, we exclude these baseline methods from our training experiments on dHCP.

Research. E(3)-Pose accurately predicts pose and significantly outperforms other direct pose regression methods, showcasing the advantage of modeling rotational equivariance. The relative rankings of 3DPose-Net, 6DRep, and RbR highlight the stability gains of estimating continuous and higher-dimensional rotation parametrizations [78]. E(3)-Pose takes advantage of this concept by regressing a continuous, overcomplete output. On dHCP test data, E(3)-Pose is outperformed by template-based methods. This strategy is able to effectively exploit the information advantage of a template volume coming from the same subject when the intensity distributions of both volumes are highly detailed and similar, eliminating alignment ambiguities. Furthermore, the 1.8s inference time of FireANTs (cf. 0.3s for E(3)-Pose) is prohibitive for real-time deployment. Finally, Fetal-Align also yields accurate pose estimates, demonstrating the reliability of landmark detection methods in volumes where landmarks are clearly visible.

Clinical-Young. E(3)-Pose yields accurate pose estimates in clinical acquisitions of younger ages, outperforming all

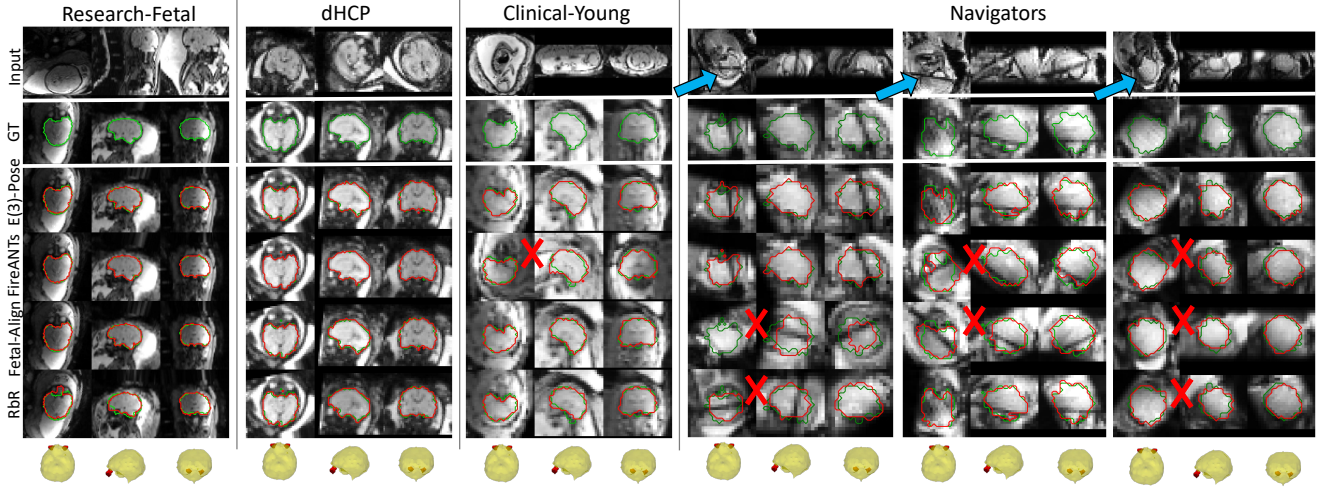


Figure 3. **Example results for methods trained on Research-Fetal.** Volumes are displayed before (row 1) and after (rows 2-6) alignment to the canonical object frame. The brain mask is also aligned to the GT (green outline) and predicted (red outline) frames. Navigator volumes include spin history artifacts (blue arrows) and low resolution/SNR, posing challenges for pose estimation. While baseline methods struggle (red Xs, rotation error $> 60^\circ$) in younger fetuses (column 3) and navigator volumes (columns 4-6), E(3)-Pose correctly predicts pose in all cases. E(3)-Pose remains accurate under significant pose ambiguity, e.g. when the artifact intersects both eyes, large voxel size obscures brain structure, and the fetal brain is close to a sphere (column 6).

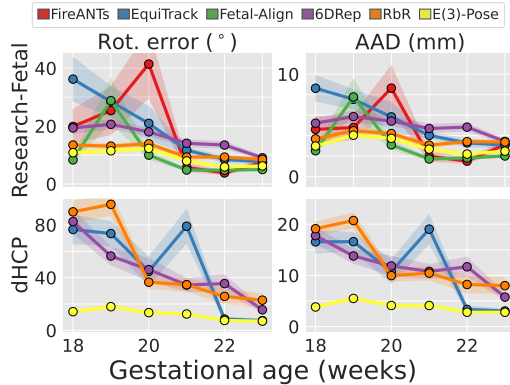


Figure 4. **Clinical-Young results.** Performance statistics are shown as a function of gestational age for all well-performing methods trained on Research-Fetal (*top*) and dHCP (*bottom*).

baseline methods. This difference is more pronounced when training on dHCP, which consists of higher-SNR volumes and presents an even larger distribution shift from clinical data compared to Research-Fetal. E(3)-Pose is the only method that maintains its performance in both training settings, highlighting the robust generalization advantages of our method. For further insight, we study the performance as a function of GA. Fig. 4 reveals that template-based methods (FireANTs and EquiTrack) break down on younger fetuses, where underdeveloped, spherical brain shapes and the lack of structure differentiation within the brain [75] lead to alignment ambiguities (Fig. 3). 6DRep and RbR remain relatively stable across GAs when trained on Research-Fetal: by re-

moving dependence on specific anatomical markers, direct pose regression is less sensitive to the lack of fine-grained details. However, pose regression with standard CNNs struggles to generalize to younger GAs when trained on dHCP, suggesting that these methods are prone to overfitting to the training distribution. In contrast, E(3)-Pose provides accurate results across GAs and training datasets.

Navigators. E(3)-Pose achieves accurate pose estimation in navigator volumes and significantly outperforms all baselines (Table 1). While all baseline methods struggle with the challenging task of generalizing from dHCP to Navigators, E(3)-Pose remains robust to the substantial domain gap. Importantly, E(3)-Pose continues to maintain accuracy in the most challenging cases, which occur when the effects of low resolution, low SNR, and artifact obstruction combine to eliminate most asymmetric cues (Fig. 3, columns 4-6). Nevertheless, our method successfully recovers head pose, even in volumes with high uncertainty. In contrast, template registration methods are prone to errors when spin history artifacts disrupt the similarity between input and template intensity distributions, and Fetal-Align remains highly sensitive to eye visibility, which is often limited by artifact obstruction or blurring due to voxel size. Moreover, E(3)-Pose generally yields more stable estimates compared to CNN-based pose regression, highlighting that enforcing pose equivariance and anatomical symmetry on the network parameters is an effective regularization technique that enables generalization to out-of-distribution clinical volumes and stability under pose ambiguity.

Table 2. **Ablation study.** Mean \pm standard deviation statistics for rotation error ($^\circ$) and average absolute error (AAD, mm). Best score is in bold. * indicates statistical significance compared to E(3)-Pose ($p < 0.05$, pairwise Wilcoxon test).

	Trained on Research-Fetal				Trained on dHCP			
	Clinical-Young		Navigators		Clinical-Young		Navigators	
	Rot. error	AAD	Rot. error	AAD	Rot. error	AAD	Rot. error	AAD
E(3)-Pose (ours)	9.1 ± 4.7	3.0 ± 1.4	9.4 ± 7.5	6.3 ± 4.0	12.1 ± 6.8	3.9 ± 1.9	13.9 ± 13.2	7.9 ± 5.6
Standard CNN	$10.3 \pm 5.4^*$	$3.4 \pm 1.5^*$	$18.0 \pm 19.3^*$	$9.7 \pm 7.6^*$	$53.3 \pm 55.8^*$	$12.9 \pm 11.9^*$	$80.2 \pm 52.7^*$	$31.9 \pm 17.3^*$
no pseudovector	$10.4 \pm 4.0^*$	$3.4 \pm 1.3^*$	$10.9 \pm 11.1^*$	$7.0 \pm 5.0^*$	12.1 ± 7.7	3.9 ± 2.2	$20.6 \pm 26.1^*$	$10.6 \pm 10.1^*$
$h(R) = e_y \oplus e_z$	$10.3 \pm 4.2^*$	$3.4 \pm 1.3^*$	$11.9 \pm 12.9^*$	$7.3 \pm 5.5^*$	11.6 ± 6.5	3.7 ± 1.8	$26.1 \pm 35.6^*$	$12.3 \pm 12.2^*$
$\mathcal{L}_\phi = \sum_{k \in \{x,y,z\}} \sin(\theta_k/2) $	8.8 ± 3.6	3.0 ± 1.3	$12.1 \pm 11.1^*$	$7.2 \pm 5.1^*$	11.1 ± 5.3	3.6 ± 1.5	18.5 ± 27.3	9.4 ± 9.4
geodesic loss	9.2 ± 4.7	3.1 ± 1.5	$13.5 \pm 20.0^*$	$9.7 \pm 7.6^*$	30.9 ± 53.4	7.7 ± 10.9	$27.8 \pm 42.5^*$	$12.6 \pm 14.0^*$
no artifact augmentation	12.0 ± 21.1	3.5 ± 3.8	$16.2 \pm 18.6^*$	$7.7 \pm 7.6^*$	$18.4 \pm 21.1^*$	$5.4 \pm 4.5^*$	$31.9 \pm 36.7^*$	$15.1 \pm 13.9^*$

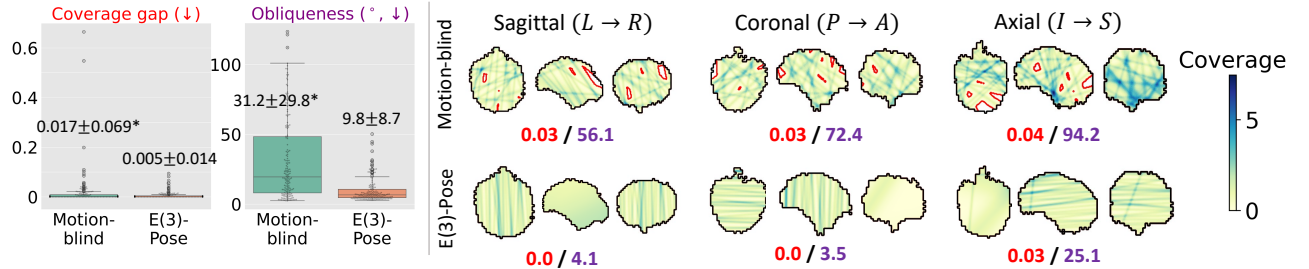


Figure 5. **Simulation results.** *Left:* Quantitative comparison of diagnostic slice stacks obtained using motion-blind prescription and E(3)-Pose. Mean \pm standard deviation statistics are displayed. * indicates statistical significance ($p < 0.05$, pairwise Wilcoxon). *Right:* Brain coverage of the diagnostic slice stacks prescribed by each method, for three different example subjects and target anatomical orientations. Coverage gap (red) and obliqueness (purple, $^\circ$) metrics are respectively displayed. Spatial coverage gap regions are outlined in red.

5.5. Ablation study

To study performance on challenging out-of-distribution cases, we evaluate ablations on clinical datasets (Table 2). Appendix I provides results on research datasets.

Equivariance. We ablate the modeling of rotational pose equivariance by replacing all E(3)-equivariant convolutions with standard convolutions equivariant to only translations. The results show that rotation equivariance is critical for generalization to clinical data, especially when the domain gap between training and testing data is significant. This outcome is in line with prior work [4, 39, 57] and our earlier comparison of E(3)-Pose to non-equivariant RbR, 6DRep, and 3DPose-Net.

Rotation parametrization. We consider alternative equivariant rotation parametrizations. First, we study the impact of estimating a symmetry-invariant parametrization by replacing the pseudovector with a standard vector. Table 2 shows that E(3)-Pose significantly outperforms this ablation, which struggles under pose ambiguities. Second, we reduce the dimensionality of the rotation parametrization by predicting only two basis directions: e_y and e_z . We observe that estimating a third basis direction significantly reduces errors in challenging clinical volumes.

Loss function and data augmentation. We now examine the effect of two alternative loss functions that are not invari-

ant under left-right reflections. We first train our network by replacing the pseudovector loss term in Eq. (9) with that of the standard vectors. Next, we replace our objective with the widely used geodesic loss [56]. Table 2 shows that both loss functions significantly reduce performance in Navigators, emphasizing the importance of combining both symmetry-invariant architectures and loss functions for stable performance under high ambiguity. Lastly, ablating the simulated spin history artifact simulations during training shows that this augmentation critically enables generalization to Navigators, where volumes contain real spin history artifacts.

5.6. Simulation study

We further establish the clinical utility of E(3)-Pose through simulations of real-time diagnostic slice prescription, using test volumes from Research-Fetal. We simulate stacks of diagnostic slices in the sagittal (L \rightarrow R), coronal (P \rightarrow A), and axial (I \rightarrow S) anatomical orientations in all subjects. We compare our method to a “motion-blind” strategy, which prescribes every slice based on the fetal head pose in the first navigator volume only. We assess the quality of the simulated stacks by computing the coverage gap (i.e., the ratio of brain volume that is not captured); and slice obliqueness (i.e., the rotation error between prescribed and target slice anatomical orientations). Appendix J provides implemen-

tation details and additional evaluations. Fig. 5 shows that E(3)-Pose yields significant improvements in both metrics. Under motion-blinds prescription, large inter-slice motion leads to oblique slices that frequently intersect, producing highly uneven spatial coverage and, in the worst cases, complete coverage gaps. In contrast, E(3)-Pose enables adjustment of the imaging plane to follow the movements of the head. As a result, E(3)-Pose prescribes slices closely aligned with the target anatomical orientation, producing parallel slices that improve the uniformity of coverage and thus reduce the likelihood of coverage gaps in our simulations.

6. Discussion and Conclusion

We present E(3)-Pose, a novel framework for 6-DoF pose estimation. By jointly modeling rigid pose equivariance and reflectional object symmetry, E(3)-Pose enables robust generalization to challenging domains with pose ambiguities.

E(3)-Pose does not explicitly model motion history or identify high-uncertainty cases, both of which could help to mitigate errors in noisy clinical volumes. Thus, future work will investigate equivariant approaches for estimating probability distributions over $SO(3)$, conditioned on motion trajectories [37, 53]. Beyond fetal imaging, it is interesting to investigate applications of our method when *approximate* reflectional and/or rigid symmetries [65] are present, such as cardiac, lung, and adult brain imaging. Lastly, we expect that our method could be adapted to objects with more complex (e.g., n -fold rotations) symmetries (Appendix K).

In our experiments, we provide the first evaluation on data representative of clinical applications in fetal brain MRI, bringing the promise of clinical deployment closer to reality.

References

- [1] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446):eaat8414, 2019. 1
- [2] Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. SynthSeg: segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*, 86:102789, 2023. 5
- [3] Benjamin Billot, Neel Dey, Daniel Moyer, Malte Hoffmann, Esra Abaci Turk, Borjan Gagoski, P. Ellen Grant, and Polina Golland. SE(3)-equivariant and noise-invariant 3D rigid motion tracking in brain MRI. *IEEE Transactions on Medical Imaging*, 43(11):4029–4040, 2024. 1, 2, 6, 4, 5
- [4] Srinath Bulusu, Matteo Favoni, Andreas Ipp, David I. Müller, and Daniel Schuh. Equivariance and generalization in neural networks. *EPJ Web of Conferences*, 258:09001, 2022. 2, 8
- [5] Shulei Cai, Guofu Zhang, He Wang, and Jing Wang. Normative linear and volumetric biometric measurements of fetal brain development in magnetic resonance imaging. *Child’s Nervous System*, 36:2997–3005, 2020. 1
- [6] Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, and Randall Hill. Equivariant point network for 3D point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14514–14523, 2021. 2
- [7] Hsu-Kuang Chiu, Jie Li, Rareş Ambruş, and Jeannette Bohg. Probabilistic 3D multi-modal, multi-object tracking for autonomous driving. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14227–14233, 2021. 1
- [8] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv: Learning*, 2015. 3
- [9] Taco S. Cohen and Max Welling. Group equivariant convolutional networks. *Proc. Int. Conf. Mach. Learn.*, pages 2990–2999, 2016. 2
- [10] Taco S. Cohen and Max Welling. Steerable CNNs. In *Int. Conf. Learn. Rep.*, 2017.
- [11] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *Int. Conf. Learn. Rep.*, 2018. 2
- [12] Ivan Diaz, Mario Geiger, and Richard Iain McKinley. Leveraging $SO(3)$ -steerable convolutions for pose-robust semantic segmentation in 3D medical data. *Machine Learning for Biomedical Imaging*, 2:834–855, 2024. 3
- [13] Sebastian Diaz, Benjamin Billot, Neel Dey, Molin Zhang, Esra Abaci Turk, P. Ellen Grant, Polina Golland, and Elfar Adalsteinsson. Robust fetal pose estimation across gestational ages via cross-population augmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23–27, 2025, Proceedings, Part VII*, page 549–559, Berlin, Heidelberg, 2025. Springer-Verlag. 6
- [14] Razieh Faghhipirayesh, Davood Karimi, Deniz Erdogmus, and Ali Gholipour. Automatic brain pose estimation in fetal MRI. In *Medical Imaging 2023: Image Processing*, page 124640T. International Society for Optics and Photonics, SPIE, 2023. 2, 6, 4
- [15] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 4
- [16] Borjan Gagoski, Patrick McDaniel, André J. W. van der Kouwe, Himanshu Bhat, Lawrence L. Wald, Elfar Adalsteinsson, P. Ellen Grant, and M. Dylan Tisdall. HASTE imaging with EPI volumetric navigators for real-time fetal head motion detection. *Proc. Intl. Soc. Magn. Reson. Med.*, 2016. 1, 5
- [17] Fei Gao, Hyunsoo Yoon, Teresa Wu, and Xianghua Chu. A feature transfer enabled multi-task deep learning model on medical imaging. *Expert Systems with Applications*, 143:112957, 2020. 3
- [18] Ge Gao, Mikko Lauri, Yulong Wang, Xiaolin Hu, Jianwei Zhang, and Simone Frntrop. 6D object pose regression via supervised learning on point clouds. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3643–3649, 2020. 1
- [19] Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks, 2022. 2, 3, 8
- [20] A. René Geist, Jonas Frey, Mikel Zhobro, Anna Levina, and Georg Martius. Learning with 3D rotations: a hitchhiker’s

- guide to $SO(3)$. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. [2](#)
- [21] Ali Gholipour, Judith A. Estroff, Carol E. Barnewolt, Richard L. Robertson, P. Ellen Grant, Borjan Gagoski, Simon K. Warfield, Onur Afacan, Susan A. Connolly, Jeffrey J. Neil, Adam Wolfberg, and Robert V. Mulkern. Fetal MRI: a technical update with educational aspirations, 2014. [1](#)
- [22] Orit A. Glenn. MR imaging of the fetal brain. *Pediatr Radiol*, 40:68–81, 2009. [1](#)
- [23] Karthik Gopinath, Xiaoling Hu, Malte Hoffmann, Oula Puonti, and Juan Eugenio Iglesias. Registration by regression (RbR): a framework for interpretable and flexible atlas registration. *Biomed. Image Regist. Proc.*, 90:205–215, 2025. [2](#), [6](#), [4](#)
- [24] F. Sebastin Grassia. Practical parameterization of rotations using the exponential map. *J. Graph. Tools*, 3(3):29–48, 1998. [2](#)
- [25] Douglas N. Greve and Bruce Fischl. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72, 2009. [5](#)
- [26] K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. HEALPix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759, 2005. [8](#)
- [27] Walter Hauser. Vector products and pseudovectors. *Am. J. Phys.*, 54:168–172, 1986. [2](#), [3](#)
- [28] Aaron T. Hess, M. Dylan Tisdall, Ovidiu C. Andronesi, Ernesta M. Meintjes, and André J. W. van der Kouwe. Real-time motion and B0 corrected single voxel spectroscopy using volumetric navigators. *Magnetic Resonance in Medicine*, 66(2):314–323, 2011. [1](#)
- [29] Derek L. G. Hill, Philipp G. Batchelor, Mark Holden, and David J. Hawkes. Medical image registration. *Phys. Med. Biol.*, 46(3), 2001. [2](#)
- [30] Tomáš Hodaň, Dániel Baráth, and Jiří Matas. EPOS: Estimating 6D pose of objects with symmetries. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [3](#), [5](#)
- [31] Malte Hoffmann, Esra Abaci Turk, Borjan Gagoski, Leah Morgan, Paul Wighton, M. Dylan Tisdall, Martin Reuter, Elfar Adalsteinsson, P. Ellen Grant, Lawrence L. Wald, and André J W van der Kouwe. Rapid head-pose detection for automated slice prescription of fetal-brain MRI. *Int. J. Imaging. Syst. Technol.*, 31(3):1136–1154, 2021. [1](#), [2](#), [3](#), [6](#), [4](#)
- [32] Berthold Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society*, 4: 629–642, 1987. [6](#)
- [33] Owen Howell, David Klee, Ondrej Biza, Linfeng Zhao, and Robin Walters. Equivariant single view pose prediction via induced and restriction representations. In *Advances in Neural Information Processing Systems*, pages 47251–47263. Curran Associates, Inc., 2023. [2](#)
- [34] Rohit Jena, Pratik Chaudhari, and James C Gee. FireANTs: Adaptive riemannian optimization for multi-scale diffeomorphic registration. *arXiv preprint arXiv:2404.01249*, 2024. [6](#), [4](#), [5](#)
- [35] Vyacheslav R. Karolis, Lucilio Cordero-Grande, Anthony N. Price, Emer Hughes, Sean P. Fitzgibbon, Vanessa Kyriakopoulou, Alena Uus, Nicholas Harper, Denis Prokopenko, Devi Bridglal, Jucha Willers Moore, Sian Wilson, Maximilian Pietsch, Daan Christiaens, Maria Deprez, Logan Z.J. Williams, Emma C. Robinson, Antonis Makropoulos, Seyedeh-Rezvan Farahibozorg, Jonathan O’Muircheartaigh, Mary A. Rutherford, Daniel Rueckert, A. David Edwards, Tomoki Arichi, Stephen M. Smith, Eugene Duff, and Joseph V. Hajnal. The developing human connectome project fetal functional MRI release: methods and data structures. *Imaging Neuroscience*, 3, 2025. [5](#), [3](#)
- [36] Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *Int. Conf. Learn. Rep.*, 2015. [3](#)
- [37] David M Klee, Ondrej Biza, Robert Platt, and Robin Walters. Image to sphere: learning equivariant features for efficient pose prediction. In *International Conference on Learning Representations*, 2023. [9](#)
- [38] Maria Kuklisova-Murgasova, Gerardine Quaghebeur, Mary A. Rutherford, Joseph V. Hajnal, and Julia A. Schnabel. Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. *Medical Image Analysis*, 16(8): 1550–1564, 2012. [5](#), [6](#)
- [39] Hannah Lawrence, Kristian Georgiev, Andrew Dienes, and Bobak T Kiani. Implicit bias of linear equivariant networks. *Int. Conf. Mach. Learn*, 2022. [2](#), [8](#)
- [40] Byung-Hun Lee, Sung-Mo Kang, and Hyo-Sung Ahn. Distributed orientation estimation in $SO(d)$ and applications to formation control and network localization. *IEEE Transactions on Control of Network Systems*, 6(4):1302–1312, 2019. [3](#)
- [41] Jongmin Lee and Minsu Cho. 3D equivariant pose regression via direct wigner-D harmonics prediction. In *Advances in Neural Information Processing Systems*, pages 127765–127798. Curran Associates, Inc., 2024. [2](#), [8](#)
- [42] Jongmin Lee, Byungjin Kim, and Minsu Cho. Self-supervised equivariant learning for oriented keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4847–4857, 2022. [2](#)
- [43] Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snaveley, Afshin Rostamizadeh Angjoo Kanazawa, and Ameesh Makadia. An analysis of SVD for deep rotation estimation. In *Advances in Neural Information Processing Systems 34*, 2020. To appear in. [5](#), [6](#)
- [44] Xiaolong Li, Yijia Weng, Li Yi, Leonidas J Guibas, A. Abbott, Shuran Song, and He Wang. Leveraging $SE(3)$ equivariance for self-supervised category-level object pose estimation from point clouds. In *Advances in Neural Information Processing Systems*, pages 15370–15381. Curran Associates, Inc., 2021. [2](#)
- [45] C. Malamateniou, S.J. Malik, S.J. Counsell, J.M. Allsop, A.K. McGuinness, T. Hayat, K. Broadhouse, R.G. Nunes, A.M. Ederies, J.V. Hajnal, and M.A. Rutherford. Motion-compensation techniques in neonatal and fetal MR imaging. *American Journal of Neuroradiology*, 34(6):1124–1136, 2013. [1](#)
- [46] J Matas, O Chum, M Urban, and T Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Im-*

- age and Vision Computing*, 22(10):761–767, 2004. British Machine Vision Computing 2002. 4
- [47] Karla L. Miller, Andreas J. Bartsch, and Stephen M. Smith. Simultaneous multi-slice imaging for resting-state fMRI. *Magnetom Flash*, 63(3):389–399, 2015. 5
- [48] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In *Fourth Int. Conf. 3D Vision*, pages 565–571, 2016. 3
- [49] Daniel Moyer, Esra Abaci Turk, P. Ellen Grant, William M. Wells, and Polina Golland. Equivariant filters for efficient tracking in 3D imaging. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 193–202, Cham, 2021. Springer International Publishing. 1, 2
- [50] Mohamed Adel Musallam, Vincent Gaudillière, Miguel Ortiz del Castillo, Kassem Al Ismaeil, and Djamila Aouada. Leveraging equivariant features for absolute pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6876–6886, 2022. 2
- [51] Giorgia Pitteri, Michaël Ramamonjisoa, Slobodan Ilic, and Vincent Lepetit. On object symmetries and 6D pose estimation from images. *CoRR*, abs/1908.07640, 2019. 3, 5
- [52] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. TorchIO: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236, 2021. 5, 3
- [53] Arun Srivatsan Rangaprasad, Gillian Rosen, Drqaid Hadi, and Howie Choset. Estimating SE(3) elements using a dual quaternion based linear kalman filter. In *Robotics: Science and Systems (RSS)*, 2016. 9
- [54] Jesse Richter-Klug and Udo Frese. Handling object symmetries in CNN-based pose estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13850–13856, 2021. 3
- [55] Francois Rousseau, Orit A. Glenn, Bistra Iordanova, Claudia Rodriguez-Carranza, Daniel B. Vigneron, James A. Barkovich, and Colin Studholme. Registration-based approach for reconstruction of high-resolution in utero fetal MR brain images. *Academic Radiology*, 13(9):1072–1081, 2006. 6
- [56] Seyed S.M. Salehi, Shadab Khan, Deniz Erdogmus, and Ali Gholipour. Real-time deep pose estimation with geodesic loss for image-to-template rigid registration. *IEEE Trans. Med. Imag.*, 38(2):470–481, 2019. 1, 2, 6, 8, 4
- [57] Akiyoshi Sannai, Masaaki Imaizumi, and Makoto Kawano. Improved generalization bounds of group invariant / equivariant deep networks via quotient feature spaces. In *Conference on Uncertainty in Artificial Intelligence*, 2019. 2, 8
- [58] Ashutosh Saxena, Justin Driemeyer, and Andrew Y. Ng. Learning 3-D object orientation from images. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*, page 4266–4272. IEEE Press, 2009. 3, 4
- [59] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, page 245–254, New York, NY, USA, 1985. Association for Computing Machinery. 6
- [60] Sara Neves Silva, Sarah McElroy, Jordina Aviles Verdera, Kathleen Colford, Kamilah St Clair, Raphael Tomi-Tricot, Alena Uus, Valéry Ozenne, Megan Hall, Lisa Story, Kuberan Pushparajah, Mary A. Rutherford, Joseph V. Hajnal, and Jana Hutter. Fully automated planning for anatomical fetal brain MRI on 0.55T. *Magn. Reson. Med.*, 92(3):1263–1276, 2024. 1, 2, 3
- [61] Nathaniel Thomas, Tess E. 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. *CoRR*, abs/1802.08219, 2018. 2
- [62] M. Dylan Tisdall, Aaron T. Hess, Martin Reuter, Ernesta M. Meintjes, Bruce Fischl, and André J. W. van der Kouwe. Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI. *Magn. Reson. Med.*, 68(2):389–399, 2012. 1
- [63] Alan Q. Wang, Evan M. Yu, Adrian V. Dalca, and Mert R. Sabuncu. A robust and interpretable deep learning framework for multi-modal registration via keypoints. *Medical Image Analysis*, 90:102962, 2023. 2
- [64] Jian Wang, Razieh Faghihpirayesh, Polina Golland, and Ali Gholipour. SpaER: learning spatio-temporal equivariant representations for fetal brain motion tracking. In *Perinatal, Preterm and Paediatric Image Analysis*, pages 3–13, Cham, 2025. Springer Nature Switzerland. 1
- [65] Rui Wang, Rose Yu, and Robin Walters. Approximately equivariant networks for imperfectly symmetric dynamics. In *Int. Conf. Mach. Learn.*, 2022. 2, 9
- [66] Maurice Weiler and Gabriele Cesa. General E(2)-Equivariant Steerable CNNs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 2
- [67] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3D steerable CNNs: learning rotationally equivariant features in volumetric data. *Proc. Adv. Neural Inf. Process. Systems*, 31, 2018. 2, 3, 8
- [68] Nathan White, Cooper Roddey, Ajit Shankaranarayanan, Eric Han, Dan Rettmann, Juan Santos, Josh Kuperman, and Anders Dale. PROMO: real-time prospective motion correction in MRI using image-based tracking. *Magn. Reson. Med.*, 61(3):91–105, 2010. 1
- [69] Paul Wightton, Oliver Hinds, Robert Frost, Malte Hoffmann, Borjan Gagoski, Divya Varadarajan, Sebastien Proulx, Martin Reuter, Jonathan R. Polimeni, Bruce Fischl, Satrajit Ghosh, and André J. W. van der Kouwe. MR software tools for real-time decision making and FOV prescription. *Proc. Intl. Soc. Magn. Reson. Med.*, 2024. 1
- [70] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: deep translation and rotation equivariance. In *IEEE Conf. Comp. Vis. and Pat. Recog.*, pages 7168–7177, 2017. 2
- [71] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes. *Robotics: Science and Systems (RSS)*, 2018. 1, 3, 4
- [72] Junshen Xu, Daniel Moyer, P Ellen Grant, Polina Golland, Juan Eugenio Iglesias, and Elfar Adalsteinsson. SVoRT:

- iterative transformer for slice-to-volume registration in fetal brain MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–13. Springer, 2022. [6](#)
- [73] Junshen Xu, Daniel Moyer, Borjan Gagoski, Juan Eugenio Iglesias, P. Ellen Grant, Polina Golland, and Elfar Adalsteinsson. NeSVoR: implicit neural representation for slice-to-volume reconstruction in MRI. *IEEE Transactions on Medical Imaging*, 42(6):1707–1719, 2023. [1](#), [4](#)
- [74] Ziv Yaniv and Kevin Cleary. Image-guided procedures: a review. *Computer Aided Interventions and Medical Robotics*, 3(1-63):7, 2006. [1](#)
- [75] Jinfeng Zhan, Ivo D. Dinov, Junning Li, Zhonghe Zhang, Sam Hobel, Yonggang Shi, Xiangtao Lin, Alen Zamanyan, Lei Feng, Gaojun Teng, Fang Fang, Yuchun Tang, Fengchao Zang, Arthur W. Toga, and Shuwei Liu. Spatial-temporal atlas of human fetal brain development during the early second trimester. *NeuroImage*, 82:115–126, 2013. [1](#), [5](#), [7](#)
- [76] Chao Zhang, Ignas Budvytis, Stephan Liwicki, and Roberto Cipolla. Rotation equivariant orientation estimation for omnidirectional localization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. [2](#)
- [77] Heng Zhao, Shenxing Wei, Dahu Shi, Wenming Tan, Zheyang Li, Ye Ren, Xing Wei, Yi Yang, and Shiliang Pu. Learning symmetry-aware geometry correspondences for 6D object pose estimation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13999–14008, 2023. [3](#), [5](#)
- [78] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746, 2019. [2](#), [6](#)
- [79] Åke Björck. Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT Numerical Mathematics*, 7: 1–21, 1967. [6](#)

Equivariant Symmetry-Aware Head Pose Estimation for Fetal MRI

Supplementary Material

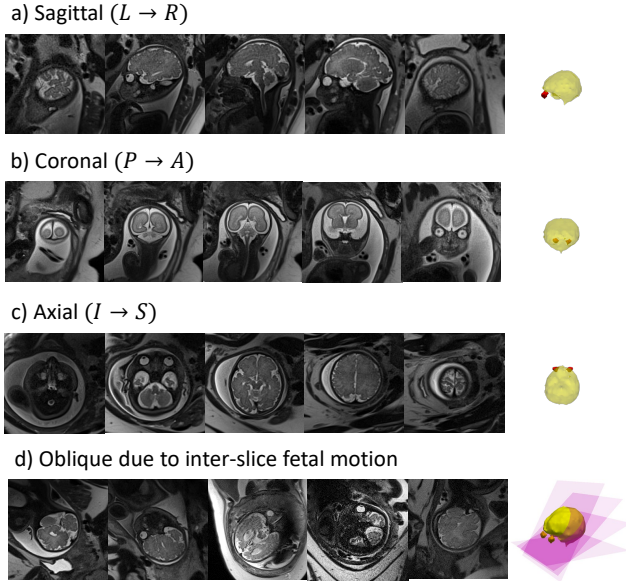


Figure 6. **Example diagnostic 2D slices.** Accurate radiological assessment of developmental abnormalities necessitates aligned with (a) sagittal (left→right), (b) coronal (posterior→anterior), and (c) axial (inferior→superior) anatomical orientations. Inter-slice fetal motion causes oblique slice orientations relative to the canonical anatomical directions. (d). This produces slices that are challenging for the radiologist to interpret and induces coverage gaps in the acquired stack of slices.

A. Automated slice prescription in fetal brain MRI

In this section, we describe our clinical application in further detail.

Overview. Radiological assessment of abnormal fetal brain development relies on volumetric fetal brain MRI, which provides greater contrast and resolution than standard ultrasound examinations. Due to sensitivity to fetal motion, current clinical practice does not involve direct acquisition of 3D MRI scans [22]. Instead, stacks of 2D MRI slices are acquired in the three target anatomical orientations (i.e., sagittal, coronal, and axial). Slices that closely align with the target anatomical orientation are critical for abnormality detection. For instance, non-oblique midline sagittal slices are necessary for detecting agenesis of the corpus callosum [21]. Furthermore, inter-slice fetal motion is often correlated with poor spatial coverage. Even small coverage gaps can substantially reduce diagnostic potential. For example, the proportion of the total fetal brain volume constituted by the cerebellum ranges from only 0.03-0.06 in second- and

third-trimester fetuses [5]. Adaptive slice prescription based on the fetal head motion thus holds potential to improve the quality of prenatal care.

Slices are T2-weighted and acquired with the half-Fourier acquisition single-shot turbo spin-echo (HASTE) MRI sequence (1.25×1.25 mm pixels, 3mm slice thickness, 320×320 mm FOV, $TR=2.5$ s, $TE=106$ ms, GRAPPA $R=2$, Partial Fourier= $5/8$, 90° flip angle). Fig. 6 shows examples of diagnostic 2D MRI slices in the target anatomical orientations, as well as *oblique* slices caused by fetal motion. To automatically prescribe the acquisition plane of each 2D slice according to the current fetal head pose, we rapidly acquire a 3D navigator volume before each slice. Navigator volumes are acquired with the echo-planar imaging (EPI) sequence (4-6mm isotropic voxels, $324 \times 324 \times 120$ mm FOV, $TR=29$ ms, $TE=14$ ms, 5° flip angle). Low-resolution, low-energy navigator volumes are not of diagnostic quality but suffice for head pose estimation.

Automation framework. We aim to rapidly estimate the 6-DoF head pose in each navigator volume and use it to adjust the imaging plane of the next slice (Fig. 7). Formally, at the start of the sequence, we define the target imaging planes $P_1, \dots, P_K \in SE(3)$ of all K slices in the stack, relative to the canonical fetal head frame. The rotational and translational components of P_k indicate the target anatomical orientation and the relative position of the slice within the stack, respectively. Before the k th slice in the stack, we acquire a navigator volume, from which we rapidly estimate the head pose $T_k = t_k \circ R_k$ and prescribe the next slice imaging plane to be $\tilde{P}_k = T_k P_k$. We also translate the field-of-view (FOV) of the next navigator volume around the fetal head based on t_k . We include this step to ensure full head capture for accurate pose estimation. Because the time interval between every navigator volume and the next slice is 1s, we require the runtime of pose estimation to be less than 1s.

In-utero implementation. We developed a real-time feedback loop system using existing software [69] that runs this sequence on a 3T Siemens scanner (Fig. 8). Specifically, we have implemented the translation estimation step of E(3)-Pose and used it to translate the FOV of every navigator volume on pregnant patients in real time. The videos in the linked project page demonstrate how the navigator volumes are automatically shifted to follow the translational movements of the fetal head. Future work will implement real-time prescription of 2D diagnostic slices, using the full pose T_k returned by E(3)-Pose in every navigator volume.

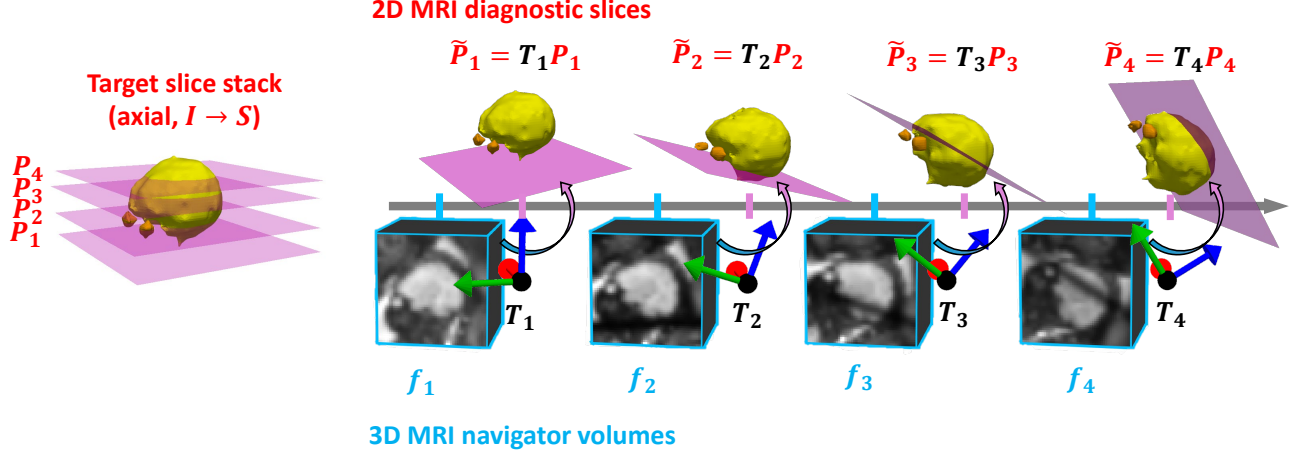


Figure 7. **Interleaved acquisition framework.** To correct for fetal head motion, we aim to automatically adjust the imaging plane P_k of every diagnostic 2D slice, using the current head pose T_k . We rapidly acquire a navigator volume f_k before every slice in order to accurately estimate T_k from f_k with E(3)-Pose.

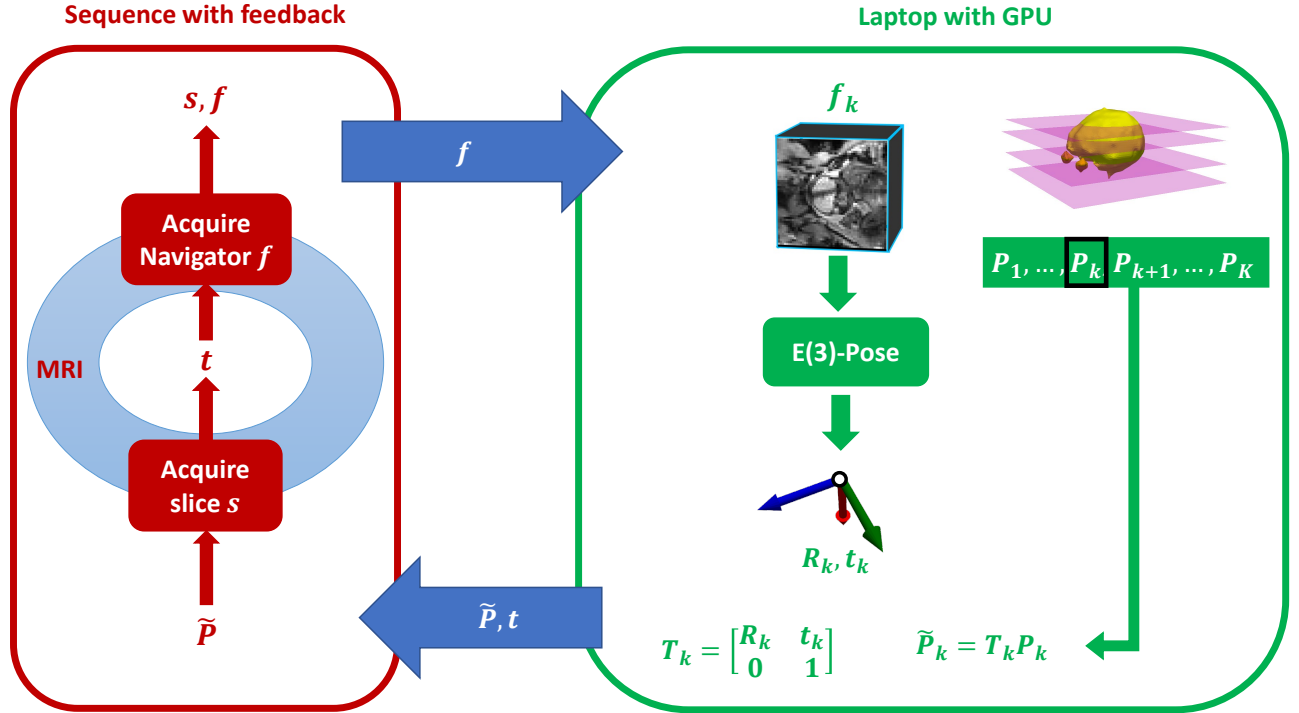


Figure 8. **Real-time feedback loop system.** Our implementation runs on a 3T Siemens scanner, connected to a server hosted on a GPU-enabled laptop. On the server (*right*), E(3)-Pose separately estimates the rotation R_k and translation t_k of the fetal head pose T_k , based on the navigator volume f_k sent by the scanner (*left*). We apply translation t_k to center the next navigator volume FOV around the fetal head. Future work will apply the full rigid transform T_k to prescribe the next 2D diagnostic slice to account for the current head pose. Finally, we send the prescription parameters to the scanner for the next acquisition.

B. E(3)-CNN implementation details

Equivariant pooling and non-linearities. E(3)-CNNs necessitate special poolings and non-linearities, since standard versions only satisfy Eq. (2) for zero-order fields. Specifically, higher order ($l > 0$) fields require (i) norm-based

pooling, and (ii) special non-linearities learned by the network as multiplicative scalar fields: $f(x)\sigma(\kappa_{l,p,0,0}*f(x))$, where σ is the sigmoid function, and $*$ is the convolution operator [67].

Network architecture. We crop input volumes to the E(3)-

CNN around the predicted brain mask to 64^3 voxels, scaled such that the brain occupies 60% of the cropped volume size. Our E(3)-CNN architecture has 4 levels, each consisting of 2 E(3)-equivariant convolutions [12, 19] with kernel size $5 \times 5 \times 5$, followed by instance normalization and equivariant max-pooling. Following [12], the first level has 8, 4, and 2 even- and odd-parity output features of orders $l = 0, 1, 2$, respectively, and the feature count is doubled in each level. We use ReLU and tanh activations for even and odd scalar features respectively, and equivariant sigmoid activations for higher-order features. For the radial basis functions φ_m in the k^{th} level, we use $\varphi_m(r) = 8.433573 \text{sus}(x + m - 1) \text{sus}(1 - m - x)$, where sus is the soft unit step function implemented in [12] and parametrized by $m \in \{0, r/4, r/2, 3r/4, r\}$, where $r = 2^{k-2}$.

C. Segmentation network architecture

Following [60], we implement the brain segmentation function ψ with a standard 3D U-Net. Our architecture has 4 levels with 16, 32, 64, and 128 output channels, respectively, each consisting of 2 convolutional layers with kernel size $3 \times 3 \times 3$, followed by batch normalization and ELU activation [8]. Inputs are padded to 128^3 voxels. While we only use the predicted brain mask for inference, we train the U-Net to additionally segment the eyes, with the aim of enhancing the overall robustness via multi-task learning [17].

D. Training details

Segmentation Network. We train the segmentation U-Net for 1000 epochs using Adam optimization [36] with batch size 4 and learning rate 10^{-4} . We use a weighted sum of the cross-entropy and Dice losses [48], with weights of 1 and 0.5, respectively. We re-weight both loss terms by 8, 2, and 1 for the brain, eyes, and background classes, respectively. During training and inference, we resample all input volumes to 3mm isotropic voxels before padding to 128^3 voxels.

We spatially augment training volumes with random rotations, translations, and scaling uniformly sampled from $\text{SO}(3)$, $[-30, 30]^3 \text{mm}$, and $[0.5, 1.3]$, respectively. We also simulate low resolution in 75% of training volumes, with isotropic voxel size sampled uniformly from $[3, 8] \text{mm}$. We add Gaussian noise with σ sampled uniformly from $[0, 0.03]$, random gamma correction with $\log \gamma$ sampled uniformly from $[-0.8, 0]$, and bias field artifacts. For the bias field simulation, we use existing software [52] to model the artifact as a linear combination of polynomial basis functions of order 3, with coefficients sampled uniformly from $[0, 0.5]$. For spin history artifact simulation, we sample n_{slice} uniformly from the unit sphere and σ uniformly from $[1.5, 2.3] \text{mm}$. We sample c_{slice} such that the likelihood of a point is inversely proportional to its distance from the brain boundary. In navigator volumes, when the artifact intersects regions near the

boundary of the brain, there is greater uncertainty in disambiguating the brain from external uterine tissue. Thus, we provide the network with more examples of these challenging cases during training.

E(3)-CNN. We train the E(3)-CNN regressor for 2500 epochs using stochastic gradient descent with batch size 1, learning rate 10^{-2} , weight decay 3×10^{-5} , and momentum 0.99. We spatially augment training volumes with random rotations uniformly sampled from $\text{SO}(3)$. We simulate low resolution in 90% of training volumes, with isotropic voxel size sampled uniformly from $[3, 7.5] \text{mm}$. We use gamma augmentations with $\log \gamma$ sampled uniformly from $[-2.0, 0.1]$ and the same bias field artifact as the segmentation network. For spin history artifact simulation, we sample n_{slice} uniformly from the unit sphere and σ uniformly from $[2.3, 4.6] \text{mm}$. We sample c_{slice} uniformly from voxels in the GT brain segmentation.

E. Datasets

In this section, we provide additional information on all datasets. For Research-Fetal, Clinical-Young, and Navigators, all subjects consented to data collection, and the approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of the Boston Children’s Hospital.

Research-Fetal. This dataset consists of whole-uterus EPI volumes in 153 pregnant patients (3T, Siemens, 3mm isotropic resolution, $\text{TR}=2.9\text{-}4\text{s}$, $\text{TE}=32\text{-}47 \text{ms}$, flip angle= 90°). Fetus GA ranges from 18 to 38 weeks, with mean GA 29.0 ± 5.0 weeks. The average FOV size is $300 \times 300 \times 156 \text{mm}$.

dHCP. This dataset consists of fetal brain fMRI scans in 245 pregnant patients (3T, Philips Achieva, 2.2mm isotropic resolution, $\text{TR}=2.2\text{s}$, $\text{TE}=60\text{ms}$, flip angle= 90°) [35]. Fetus GA ranges from 20 to 38 weeks, with mean GA 28.9 ± 3.8 weeks. The average FOV size is $143 \times 151 \times 108 \text{mm}$. Volumes are cropped around the fetal head and are not whole-uterus volumes. Our manually annotated brain/eyes segmentations and poses for the dHCP dataset are publicly available on the project website for full transparency and future use.

Clinical-Young. This dataset consists of whole-uterus EPI volumes in 60 pregnant patients (3T, Siemens, 1.76-3.5mm voxels, $\text{TR}=2.2\text{-}5\text{s}$, $\text{TE}=37\text{ms}$, flip angle= 90°). There are 10 subjects for each GA in the range of 18 to 23 weeks. The average FOV size is $301 \times 301 \times 129 \text{mm}$.

Navigators. This dataset consists of 47 stacks of 2D HASTE slices (3T, Siemens, $\text{TR}=3\text{-}3.5\text{s}$, $\text{TE}=100\text{ms}$, $1.25 \times 1.25 \times 3 \text{mm}$, FA= 90°) interleaved with 3D EPI navigator volumes (3T, Siemens, $\text{TR}=25\text{-}46\text{ms}$, 4-6mm, $\text{TE}=12\text{-}22\text{ms}$, flip angle= 5°) in 9 pregnant patients. Fetus GA ranges from 26 to 36 weeks, with mean GA 30.6 ± 2.9 weeks. The

average FOV size is $327 \times 327 \times 125$ mm.

Recall from Appendix A that slices are prescribed according to $\tilde{P}_k = T_k P_k$. To manually annotate GT poses in navigator volumes, we first algorithmically obtain P_k values for all slices, using optimization-based slice-to-volume registration [73]. Since all \tilde{P}_k values are available to us as part of the acquisition parameters, we then compute $T_k = \tilde{P}_k P_k^{-1}$ as the GT pose for every navigator volume preceding the k th slice in the stack. Lastly, we manually correct the poses in each navigator volume.

F. Baseline methods

We describe our implementations of the baseline methods and the modifications we make in order to optimize performance of baseline algorithms for our experiments.

FireANTs [34]. We initialize pose estimates with moments-of-inertia matching, followed by multi-resolution adaptive gradient optimization with learning rate 3×10^{-4} at downsampling factors of 4, 2, and 1 with 200, 100, and 50 iterations respectively. We find that the initialization step helps boost performance particularly in the case of large rotational misalignment between the input and template volumes.

EquiTrack [3]. This method first uses a “denoiser” CNN to remove intensity differences between input and template volumes, followed by a SE(3)-equivariant network that learns matching landmarks. Here, we re-implement EquiTrack to guarantee fair comparison with our method. First, we re-train the denoiser network to remove spin history artifacts in addition to standard intensity transformations in brain MRI volumes, such as bias field and noise. Second, we use RANSAC [15] to compute the optimal rigid transform to further boost robustness to disruptive intensity perturbations (e.g., spin history artifacts).

Fetal-Align [31]. The original implementation uses standard image processing techniques (i.e., detection of maximally stable extremal regions [46]) conditioned on GA, rather than deep learning, to estimate brain and eye masks in fetal brain volumes. We find that this approach to segmentation becomes unstable when applied to clinical volumes. For fair evaluation of landmark-based pose estimation on our data, we use the brain and eye masks predicted by our segmentation U-Net trained on simulated spin history artifacts. When eye detection fails (a common occurrence when the artifact obstructs the eyes in navigator volumes), we return the identity transformation as an estimate of rotation.

3DPose-Net [56]. We train this network with the \mathcal{L}_1 loss function instead of the geodesic loss proposed in the original work. We find that this training objective provides better performance on our datasets. For our training experiments on Research-Fetal, we use the same augmentations as that of E(3)-Pose. For our training experiments on dHCP, we tune

the data augmentation parameters such that $\log \gamma$ is sampled uniformly from $[-3, -1]$ and voxel size is sampled uniformly from $[3, 9]$ mm.

6DRep [14]. Similar to 3DPose-Net, we replace the geodesic loss function with the \mathcal{L}_1 loss function, which operates on the output and GT orthonormal basis directions. Again, we find that this modification improves the test-time performance on our data. We use the same training augmentations as 3DPose-Net.

RbR [23]. We find that this method provides more stable and accurate pose estimates on fetal brain volumes that are preprocessed with the same steps as our E(3)-CNN, i.e., inputs that are center-of-mass aligned and scaled based on the size of the brain. Since center-of-mass alignment of input volumes eliminates the need for translation estimation, we keep the translation estimation of E(3)-Pose and predict only rotation with the output of RbR. We use the same training augmentations as 3DPose-Net. We use Adam optimization with learning rate 10^{-4} to train the network.

G. Additional results

Evaluation of translation estimation. We separately evaluate the accuracy of translation estimation with our segmentation network by providing the translation error $\|\hat{t} - t\|_2$. We compare E(3)-Pose to baseline methods that estimate translation differently from E(3)-Pose, i.e., FireANTs and EquiTrack. Table 3 shows that E(3)-Pose outperforms both FireANTs and EquiTrack on navigator volumes. The presence of spin history artifacts breaks the assumption of similarity between the input and template volume intensity distributions. Template-free pose regression improves robustness to highly disruptive artifacts compared to the template-based alignment strategy.

Additional examples. In this section, we provide additional example alignments produced by E(3)-Pose and the baseline methods. Fig. 9 displays the same volumes aligned by the baseline methods that were omitted from Fig. 3 in the main paper, i.e., EquiTrack, 3DPose-Net, and 6DRep. Fig. 10 displays the same set of volumes aligned by all methods trained on dHCP. Fig. 11 replicates Fig. 3 for a different set of test volumes in each dataset.

H. Sensitivity study

Intensity perturbations. To understand performance gaps between Research-Fetal and Navigators, we assess robustness to simulated larger voxel size and spin history artifacts in test volumes from Research-Fetal (Fig. 12). E(3)-Pose remains much more consistent than all five tested baseline algorithms, which become increasingly error-prone as the severity of the perturbation increases. Particularly, E(3)-Pose

Table 3. **Performance statistics for translation estimation.** Mean \pm standard deviation for translation error (mm) are reported. Best score is in bold. * indicates statistical significance compared to E(3)-Pose ($p < 0.05$, pairwise Wilcoxon test). We compare E(3)-Pose only to methods that estimate translation differently from E(3)-Pose. We evaluate the instance of E(3)-Pose trained on Research-Fetal since translation estimation relies only on brain segmentation. FireANTs runtime is 1.8s, prohibiting real-time deployment.

	Research-Fetal test	dHCP test	Clinical-Young	Navigators
FireANTs [34]	$2.0 \pm 2.3^*$	$0.5 \pm 1.6^*$	$0.6 \pm 0.5^*$	$4.2 \pm 3.1^*$
EquiTrack [3]	$1.3 \pm 0.9^*$	$1.1 \pm 0.7^*$	$1.6 \pm 1.6^*$	$4.4 \pm 3.2^*$
E(3)-Pose (ours)	0.9 ± 0.8	0.8 ± 0.8	0.9 ± 1.1	3.8 ± 2.8

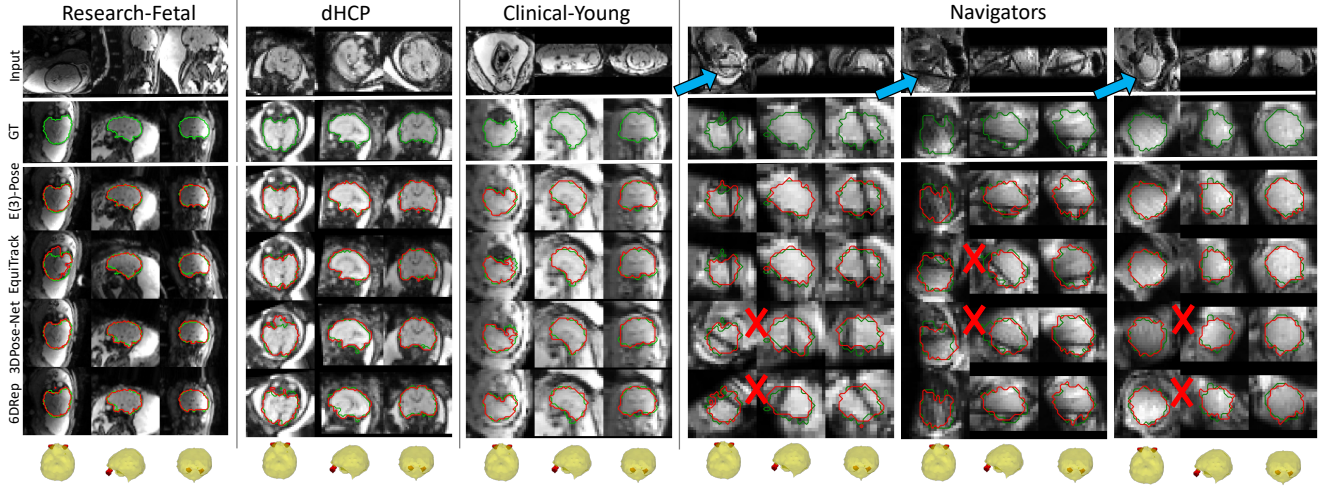


Figure 9. **Example results comparing E(3)-Pose to EquiTrack, 3DPose-Net, and 6DRep, trained on Research-Fetal.** See Fig. 3 in the main paper for descriptions.

exhibits the smallest increase in both error metrics from the original 3mm volumes to the 6mm volumes with simulated spin history artifacts. This is particularly the case for template-dependent methods, which are highly sensitive to the presence of disruptive artifacts. Furthermore, while E(3)-Pose, 6DRep and RbR are trained on volumes augmented with these intensity transformations, E(3)-Pose significantly outperforms 6DRep and RbR on 6mm volumes with artifacts ($p < 0.05$, pairwise Wilcoxon). This result demonstrates that E(3)-CNNs display more effective generalization than standard CNNs.

Spatial perturbations. To evaluate robustness of the method across input poses, we use test volumes from Research-Fetal to simulate a wide range of rotational misalignments of the canonical object frame relative to the input volume. E(3)-Pose displays the most stable consistency across rotational misalignments (Fig. 13). This result underscores the value of modeling rotation equivariance in pose estimation.

I. Ablation study

Table 4 reports the performance statistics on research data for the ablations corresponding to Table 2 in the main paper. For most experiments, there is no significant difference in performance metrics between E(3)-Pose and ablations.

These results suggest that modeling inherent symmetries by construction yields competitive accuracy but does not improve performance on in-distribution, high-SNR volumes. Specifically, methods that do not explicitly address pose ambiguities perform well when such ambiguities are eliminated by clearly visible anatomy. The design choices in E(3)-Pose provide the largest performance gains in cases with low visibility and high uncertainty, which is more common in clinical data.

J. Simulation study

Additional implementation details. We simulate stacks of 2D diagnostic slices (1×1 pixels, 3mm slice thickness) in a specific target anatomical orientation (i.e., left→right, posterior→anterior, or inferior→superior). We simulate navigator volumes rapidly acquired before each slice from test volumes in Research-Fetal, by augmenting with 6mm isotropic resolution and a simulated spin history artifact based on the imaging plane parameters of the preceding slice. We simulate one stack in each orientation for each test volume, resulting in 168 total stacks. The number of slices per stack is adjusted for brain size as in current clinical practice, resulting in 23-40 slices per stack.

Recall from Appendix A that every slice is prescribed ac-

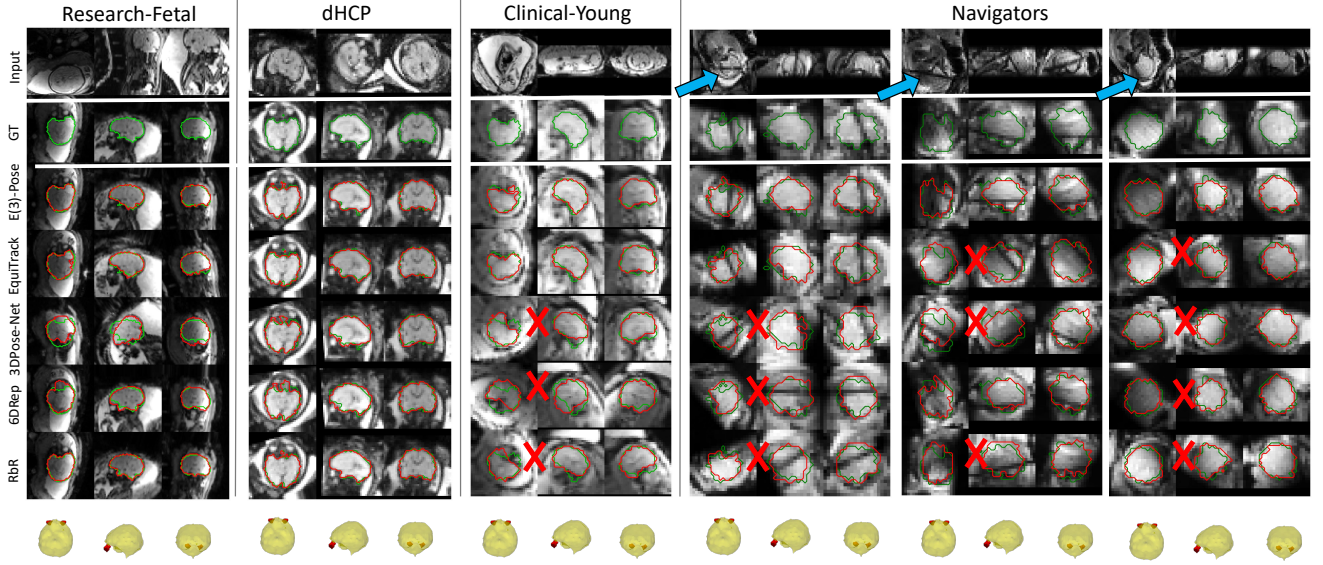


Figure 10. **Example results for methods trained on dHCP.** We omit FireANTs and Fetal-Align since these methods are not trained on dHCP. See Fig. 3 in the main paper for descriptions.

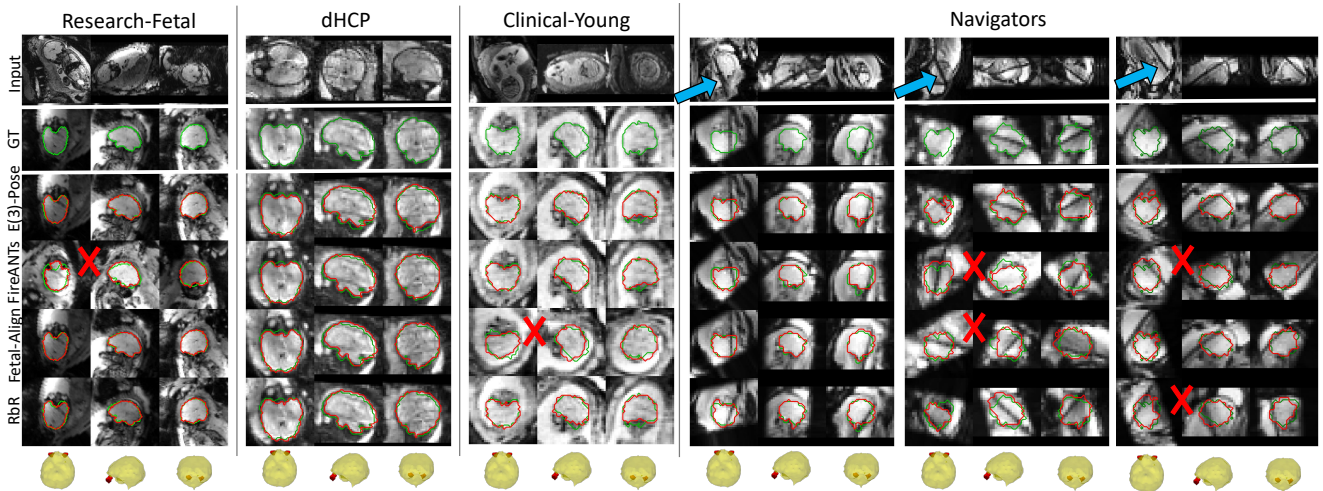


Figure 11. **Additional example results for methods trained on Research-Fetal.** See Fig. 3 in the main paper for descriptions.

cording to $\tilde{P}_k = \hat{T}_k P_k$, where \hat{T}_k is the estimated head pose in the preceding navigator volume. We compare \hat{T}_k returned by E(3)-Pose to that of “motion-blind” slice prescription, where we set $\hat{T}_k = T_{k=0}$ for all $k \geq 0$. To simulate inter-slice fetal brain motion for $k > 0$, we sample rigid motion $T_{k+1} T_k^{-1}$ from real fetal head motion trajectories, interpolated at time intervals of $TR=3s$ to match the time between navigator volume acquisitions in the clinical setting [59]. Our motion trajectory dataset consists of full time-series from subjects in Clinical-Young, since increased uterine mobility in younger fetuses produces wider ranges of motion for our analysis. We follow prior work [72] in annotating motion trajectories in this dataset for the purposes of sim-

ulation. Specifically, we apply an existing fetal landmark detection algorithm to detect the fetal eyes and shoulders in every frame [13], manually correct algorithmically estimated landmarks, and finally compute the rigid motion between consecutive frames using detected landmarks.

We define the slice coverage distribution p_C over the underlying brain volume on a 1mm resolution grid as the sum of the Gaussian kernels representing the point spread function [38, 55, 72] of all slices in the stack. We then compute the coverage gap as the proportion of the brain volume where $p_C = 0$.

Additional evaluations. Automated prescription of diag-

Table 4. **Ablation performance statistics on research data.** Mean \pm standard deviation statistics for rotation error ($^\circ$) and average absolute error (AAD, mm). Best score is in bold. * indicates statistical significance compared to E(3)-Pose ($p < 0.05$, pairwise Wilcoxon test).

	Trained on Research-Fetal				Trained on dHCP			
	Research-Fetal		dHCP		Research-Fetal		dHCP	
	Rot. error	AAD	Rot. error	AAD	Rot. error	AAD	Rot. error	AAD
E(3)-Pose (ours)	5.1 ± 2.6	3.0 ± 1.7	7.4 ± 3.6	3.7 ± 1.8	5.7 ± 3.6	3.4 ± 2.0	7.3 ± 3.4	3.7 ± 1.7
Standard CNN	$6.9 \pm 3.5^*$	$3.8 \pm 1.9^*$	7.5 ± 4.2	3.7 ± 2.0	$11.2 \pm 4.9^*$	$6.2 \pm 3.1^*$	7.5 ± 3.4	3.7 ± 1.7
no pseudovector	5.6 ± 2.4	3.2 ± 1.6	8.6 ± 3.3	4.2 ± 1.7	5.9 ± 3.0	3.4 ± 1.8	6.8 ± 3.2	$3.4 \pm 1.6^*$
$h(R) = e_y \oplus e_z$	5.3 ± 2.5	3.1 ± 1.5	6.8 ± 3.5	3.4 ± 1.6	5.9 ± 2.7	3.4 ± 1.6	7.5 ± 3.4	3.7 ± 1.7
$\mathcal{L}_\phi = \sum_{k \in \{x,y,z\}} \sin(\theta_k/2)$	4.9 ± 2.4	2.9 ± 1.5	7.2 ± 3.6	3.5 ± 1.6	5.9 ± 3.2	3.4 ± 1.8	7.3 ± 3.2	3.6 ± 1.5
geodesic loss	4.9 ± 2.1	2.9 ± 1.2	$6.5 \pm 3.8^*$	$3.3 \pm 1.7^*$	14.7 ± 38.1	6.8 ± 14.9	$73.8 \pm 84.4^*$	$24.9 \pm 27.4^*$
no artifact augmentation	5.1 ± 2.2	3.0 ± 1.4	$6.6 \pm 3.2^*$	$3.2 \pm 1.5^*$	$7.2 \pm 3.9^*$	$4.1 \pm 2.4^*$	7.6 ± 3.4	3.8 ± 1.6

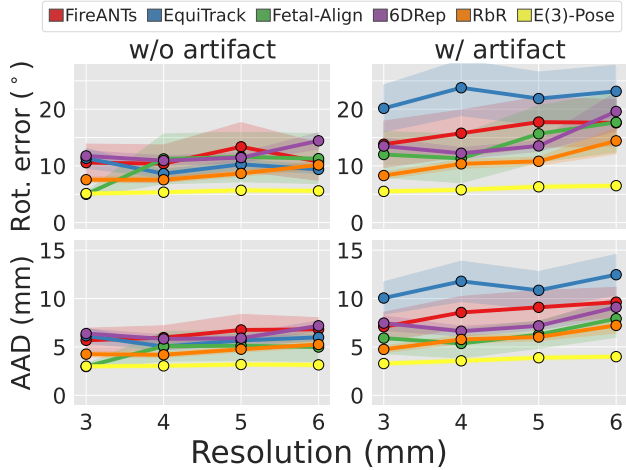


Figure 12. **Sensitivity to intensity perturbations.** We evaluate all well-performing methods trained on Research-Fetal on augmented test volumes in Research-Fetal. E(3)-Pose remains robust to progressively larger voxel sizes (*left*), even when simulating additional spin history artifacts (*right*).

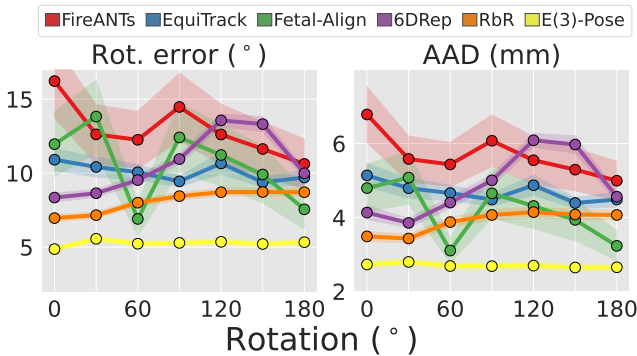


Figure 13. **Sensitivity to spatial perturbations.** We evaluate all well-performing methods trained on Research-Fetal on rotated test volumes in Research-Fetal. E(3)-Pose demonstrates stable pose estimates across a wide range of input poses.

nostic 2D slices in fetal brain MRI aims to improve spatial coverage of the brain and the radiological interpretability

of the acquired stack of slices. In the main paper, we partially quantify these qualities with the coverage gap and slice obliqueness, respectively. In this section, we include two additional metrics that further assess the diagnostic potential of the simulated slice stacks. First, because uneven spatial coverage is correlated with poor coverage quality and coverage gaps, we calculate the coverage irregularity $KL(p_C || p_U)$, where KL is the Kullback-Leibler divergence and p_U is the uniform distribution over the underlying brain volume. Second, since slice obliqueness only quantifies the rotational incoherence of slices within each stack, we also assess translational incoherence by computing the slice offset, i.e., the distance in mm between the center voxels of the GT and prescribed slices. Fig. 14 demonstrates that in our simulations, slice prescription with E(3)-Pose significantly improves both additional quantitative metrics of diagnostic potential, compared to motion-blind prescription.

K. Rotational object symmetries

In this work, we have demonstrated a rotation-equivariant framework for pose estimation of objects with reflectional symmetries, e.g., the fetal head. However, many real-world objects possess more complex symmetries. In this section, we sketch a method to estimate the pose of objects with N -fold rotational symmetries about an axis with an E(3)-CNN. We hypothesize that this will stabilize pose estimation in high-ambiguity situations, and will investigate the proposed method in future work.

Rotation parametrization. Recall that we denote the rotation of the object frame as R , the rotation parametrization as $h(R)$, and object symmetry group as G_{symm} . Here we consider the case where $G_{\text{symm}} = \{g_{\theta=2\pi(k/N)} | k = 1, \dots, N\}$ where $\theta = 2\pi(k/N)$ are the N -fold rotations in the plane of rotational symmetry. Let e_\odot be axis of rotational symmetry, and let $\text{parity}(N)$ be equal to the parity of the integer N . Then, we can train an E(3)-CNN to predict the following

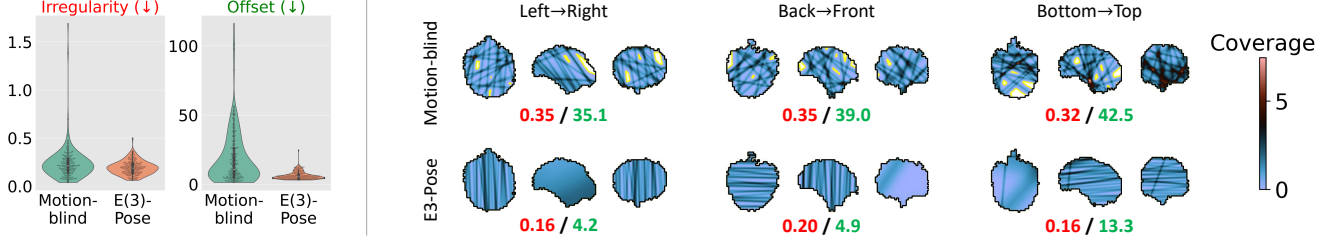


Figure 14. **Additional simulation results.** *Left:* Coverage irregularity and slice offset (mm) of diagnostic slice stacks obtained using motion-blind prescription and E(3)-Pose. Mean \pm standard deviation statistics are displayed. * indicates statistical significance ($p < 0.05$, pairwise Wilcoxon). *Right:* See Fig. 5 in the main paper for details. Here, coverage irregularity (blue) and slice offset (green, mm) metrics are respectively displayed.

rotation parametrization:

$$\begin{aligned}
 h(R) &= e_x \oplus e_y \oplus e_z, \\
 \text{s.t. } e_x &= D_{:, -N}^{l=N}(R), e_y = D_{:, N}^{l=N}(R) \text{ and } e_z = e_{\odot}, \\
 \rho_h(g_r) &= D^{l=N}(g_r) \oplus D^{l=N}(g_r) \oplus M(g_r) \\
 &= \rho_{\text{parity}(N)}^{l=N}(g_r) \oplus \rho_{\text{parity}(N)}^{l=N}(g_r) \oplus \rho_{\text{odd}}^{l=1}(g_r),
 \end{aligned} \tag{11}$$

where $D^l : \text{SO}(3) \rightarrow \mathbb{R}^{2l+1 \times 2l+1}$ is the Wigner-D matrix function of order l , and $D_{:,m}^l$ denotes its columns, indexed from $-l \leq m \leq l$. Columns of the Wigner-D matrices are irreducible tensors with irreducible Wigner-D matrix representations [19, 67]. Since $h(R)$ is parametrized in terms of irreducible tensors, it maintains equivariance under E(3). Therefore, the E(3)-CNN output is formulated as 1 vector and 2 irreducible tensors of order $l = N$ and parity $p = \text{parity}(N)$.

Invariance under object symmetry. Here we show that $h(R) = \rho_h(g_{\theta=2\pi(k/N)})h(R)$ for all $k = 1, \dots, N$. First, it is easy to see that $e_z = M(g_{\theta=2\pi(k/N)})e_z$ because in-plane rotations leave orthogonal vectors unchanged. Next, we show that $e_y = D^{l=N}(g_{\theta=2\pi(k/N)})e_y$ by showing that $\sum_{-N \leq m' \leq N} [e_y]_{m'} Y_{m=m'}^{l=N}(\theta, \phi) = \sum_{-N \leq m' \leq N} [D^{l=N}(g_{\theta=2\pi(k/N)})e_y]_{m'} Y_{m=m'}^{l=N}(\theta, \phi)$, where Y_m^l is the spherical harmonic function of degree l and order m , and the spherical coordinates (θ, ϕ) refer to the polar and azimuthal angles, respectively. The same reasoning can be applied to e_x .

Let $(\theta, \phi) \xrightarrow{R} (\theta', \phi')$ define the spherical coordinate transformation under rotation $R \in \text{SO}(3)$. We observe that

$$\begin{aligned}
 &\sum_{-N \leq m' \leq N} [e_y]_{m'} Y_{m=m'}^{l=N}(\theta, \phi) \\
 &= \sum_{-N \leq m' \leq N} D_{m', N}^{l=N}(R) Y_{m=m'}^{l=N}(\theta, \phi) \\
 &= Y_{m=N}^{l=N}(\theta', \phi')
 \end{aligned} \tag{12}$$

where we have used the property $Y_m(\theta', \phi') =$

$$\sum_{-l \leq m \leq l} D_{m', m}^l(R) Y_m^l(\theta, \phi).$$

Similarly, we have

$$\begin{aligned}
 &\sum_{-N \leq m' \leq N} [D^{l=N}(g_{\theta=2\pi(k/N)})e_y]_{m'} Y_{m=m'}^{l=N}(\theta, \phi) \\
 &= \sum_{-N \leq m' \leq N} [D^{l=N}(g_{\theta=2\pi(k/N)})D_{:, N}^{l=N}(R)]_{m'} Y_{m=m'}^{l=N}(\theta', \phi') \\
 &= \sum_{-N \leq m' \leq N} D_{m', N}^{l=N}(g_{\theta=2\pi(k/N)}R) Y_{m=m'}^{l=N}(\theta, \phi) \\
 &= Y_m^{l=N}(\theta' + \frac{2\pi k}{N}, \phi') = e^{iN2\pi(k/N)} Y_{m=N}^{l=N}(\theta', \phi') \\
 &= e^{i2\pi k} Y_{m=N}^{l=N}(\theta', \phi') = Y_{m=N}^{l=N}(\theta', \phi') \\
 &= \sum_{-N \leq m' \leq N} [e_y]_{m'} Y_{m=m'}^{l=N}(\theta, \phi)
 \end{aligned} \tag{13}$$

where we use the property $Y_l^m(\theta_1 + \theta_2, \phi) = e^{im\theta_2} Y_l^m(\theta_1, \phi)$.

Training. We aim to maintain symmetry invariance in our training objective. To this end, any standard regression loss (e.g., \mathcal{L}_1 or \mathcal{L}_2 norm) on the error term $h(R) - [\hat{e}_x \oplus \hat{e}_y \oplus \hat{e}_z]$ between the GT rotation parametrization and the E(3)-CNN output, respectively, satisfies this condition.

Inference. Since Wigner-D matrices are a Fourier basis for functions defined on SO(3), the E(3)-CNN output represents a likelihood function P over SO(3), where $P(R')$ is monotonically related to the inner product between $h(R')$ and $[\hat{e}_x \oplus \hat{e}_y \oplus \hat{e}_z]$ [41]. It is possible to sample from P by querying values on a predefined equivolumetric grid over SO(3) [26]. Indeed, if the E(3)-CNN output $[\hat{e}_x \oplus \hat{e}_y \oplus \hat{e}_z]$ is equal to $h(R)$, then P is nonzero and uniformly distributed over the set of poses that are symmetrically equivalent to R , and zero elsewhere.