Intraoperative 2D/3D Image Registration via Differentiable X-ray Rendering

Vivek Gopalakrishnan[1,2] Neel Dey[2] Polina Golland[1,2]
[1]Harvard-MIT Health Sciences and Technology [2]MIT Computer Science and Artificial Intelligence Laboratory
{vivekg,dey,polina}@csail.mit.edu

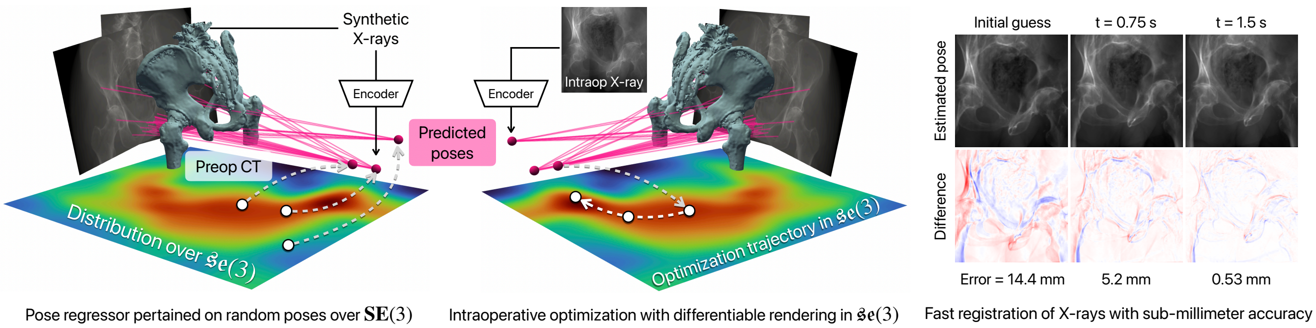

Figure 1. **We present DiffPose, a self-supervised framework for differentiable 2D/3D registration.** Trained exclusively on synthetic X-rays rendered from a patient-specific preoperative CT scan, DiffPose aligns intraoperative X-rays with sub-millimeter accuracy. DiffPose does not require manually annotated training data, performs consistently across subjects, and registers images at clinically relevant speeds.

Abstract

*Surgical decisions are informed by aligning rapid portable 2D intraoperative images (*e.g.* X-rays) to a high-fidelity 3D preoperative reference scan (*e.g.* CT). However, 2D/3D registration can often fail in practice: conventional optimization methods are prohibitively slow and susceptible to local minima, while neural networks trained on small datasets fail on new patients or require impractical landmark supervision. We present DiffPose, a self-supervised approach that leverages patient-specific simulation and differentiable physics-based rendering to achieve accurate 2D/3D registration without relying on manually labeled data. Preoperatively, a CNN is trained to regress the pose of a randomly oriented synthetic X-ray rendered from the preoperative CT. The CNN then initializes rapid intraoperative test-time optimization that uses the differentiable X-ray renderer to refine the solution. Our work further proposes several geometrically principled methods for sampling camera poses from* $\mathbf{SE}(3)$*, for sparse differentiable rendering, and for driving registration in the tangent space* $\mathfrak{se}(3)$ *with geodesic and multiscale locality-sensitive losses. DiffPose achieves sub-millimeter accuracy across surgical datasets at intraoperative speeds, improving upon existing unsupervised methods by an order of magnitude and even outperforming supervised baselines. Our implementation is at* https://github.com/eigenvivek/DiffPose.

1. Introduction

Many high-stakes surgical procedures use intraoperative X-ray image guidance to visualize surgical instruments and patient anatomy [42]. While X-ray imaging is both rapid and portable, it lacks the spatial detail of volumetric modalities such as CT. The advantages of 3D image guidance can be emulated by registering intraoperative X-rays to routinely acquired 3D preoperative scans, *i.e.* estimating the intraoperative X-ray scanner's pose relative to the patient's preoperative CT scan. Beyond localization of instruments relative to patient anatomy, accurate 2D/3D rigid registration is critical to cutting-edge surgical innovations incorporating robotic assistance and augmented reality [2, 3, 20].

Classical 2D/3D rigid registration methods apply ray tracing to CT volumes to render synthetic X-rays, referred to as digitally reconstructed radiographs (DRRs) in medical imaging [4, 53]. Iterative optimization is used to find the camera pose that generates a synthetic X-ray that most closely matches the real X-ray as quantified by an image similarity score. A major limitation of intensity-based methods is their limited capture range and sensitivity to the initial pose estimate [52]. If the initial pose is even a few millimeters from the true pose, these methods can converge to a wrong solution. To this end, two methods for initial pose estimation are commonly deployed: landmark-based localization and CNN-based pose regression.

In landmark-based localization, feature extractors find correspondences between anatomical landmarks in 2D and 3D images, which are then used by a Perspective-n-Point (PnP) solver to estimate the camera pose [30, 31]. Such methods require expert knowledge of landmarks that are visible on X-ray for the specific surgical application and manual annotation of 3D landmarks for every new CT volume [25]. While landmark localizers can be learned, only a few small labeled datasets exist for training deep networks [24, 41, 50], preventing the generalization of these methods to new patients, procedures, and pathologies.

In parallel, several approaches similar to PoseNet [29] have trained CNNs to directly regress the pose of an intraoperative X-ray [13, 35]. However, landmark-based localization consistently outperforms pose regression [10, 54] due to CNNs performing pose regression via image retrieval instead of leveraging 3D structure [44]. Therefore, generalizing supervised pose regression across all patients would require unattainably large datasets. Instead, we train a *patient-specific* pose regression CNN on potentially infinite synthetic X-rays rendered from the patient's own preoperative CT. We sample poses from a continuous distribution on $\mathbf{SE}(3)$ whose support includes all intraoperative views.

Even when supervised training data is augmented with synthetic X-rays, landmark-based localization and pose regression fail to achieve consistent sub-millimeter accuracy [17, 35, 56]. Therefore, state-of-the-art 2D/3D registration methods refine initial pose estimates with intensity-based iterative optimization at test time (*i.e*. intraoperatively). Unfortunately, this refinement can itself fail due to limitations with standard intensity losses (*e.g*. MSE, SSIM, NCC). We improve this test-time optimization procedure in two ways: We use differentiable X-ray rendering to optimize camera poses in the Lie algebra $\mathfrak{se}(3)$, and we develop a sparse multiscale variant of local normalized cross correlation [6, 7] that is robust to local minima and faster to evaluate.

Contributions. We present DiffPose, a self-supervised framework for differentiable 2D/3D registration (Figure 1). First, we pretrain a patient-specific CNN with our self-supervised pose regression task, removing the need for manually annotated training data. Through a combination of image similarity and $\mathbf{SE}(3)$-geodesic losses, the CNN learns to predict accurate initial pose estimates. Then, using a differentiable X-ray renderer for test-time optimization, we quickly refine the estimated camera pose by maximizing multiscale image similarity computed over sparse image patches. Our framework requires only a single preoperative CT scan, which is routinely acquired in the clinical standard of care. We evaluate our method on two publicly available datasets from different surgical specialties. While previous methods fail to generalize to new patients or procedures, DiffPose consistently achieves sub-millimeter registration accuracy across populations and anatomical structures.

2. Related Work

Intensity-based 2D/3D registration. A 2D intraoperative X-ray can be registered to a 3D preoperative CT volume by maximizing its similarity to a rendered X-ray with respect to the camera pose [8, 14, 34]. As traditional renderers are non-differentiable, pose estimation is commonly driven by gradient-free optimization [24]. This is both prohibitively slow for intraoperative use and often converges poorly in practice [58]. On synthetic data, gradient-based optimization with DiffDRR, a differentiable renderer, has been shown to recover the true pose faster and more robustly than gradient-free methods [1, 22]. We use DiffDRR to develop a self-supervised method for differentiable registration and evaluate on real clinical X-ray and CT images.

Landmark-based 2D/3D registration. Given supervision via anatomical landmarks, a camera pose can be estimated by applying a PnP solver [19] to corresponding landmarks extracted from 2D intraoperative and 3D preoperative imaging. Recent approaches train neural networks to segment intraoperative X-rays and localize predefined landmarks from the segmentation masks [17, 25, 27, 37, 39]. This approach requires expert knowledge of which landmarks to extract and manual annotation of a large 2D/3D dataset for the anatomy of interest, precluding application to new anatomical structures. These approaches often fail on patients with poses, anatomies, or pathologies that are substantially different from those in the training set. In contrast, our method does not use manual or estimated landmarks, operating directly on patient-specific image intensities.

Camera pose regression. Pose estimation can also be formulated as a supervised regression problem [13, 35]. Prior work has focused on regressing parameterizations of rotations that have been shown to introduce discontinuities in the representation space, *e.g*. Euler angles and quaternions [57]. In point cloud registration, higher-dimensional parameterizations of $\mathbf{SO}(3)$ have been proposed that do not suffer from this issue [32, 40, 57]. We investigate these parameterizations in the context of 2D/3D image registration.

Self-supervision. Augmenting training data with synthetically generated X-rays helps overcome the small clinical sample sizes when training landmark extractors [17, 28, 46] and pose regressors [35, 56]. While generating synthetic data for PnP registration pretraining still requires manual annotations of landmarks on each new CT, rendered X-rays come with ground truth camera poses for free. Typically, a finite number of synthetic X-rays is generated for pretraining [35, 56], failing to address the central limitation of CNN-based pose regression: as CNN-based pose regressors perform pose estimation through memorization of the training set [44], we use a fast differentiable X-ray renderer to generate unlimited synthetic training data.

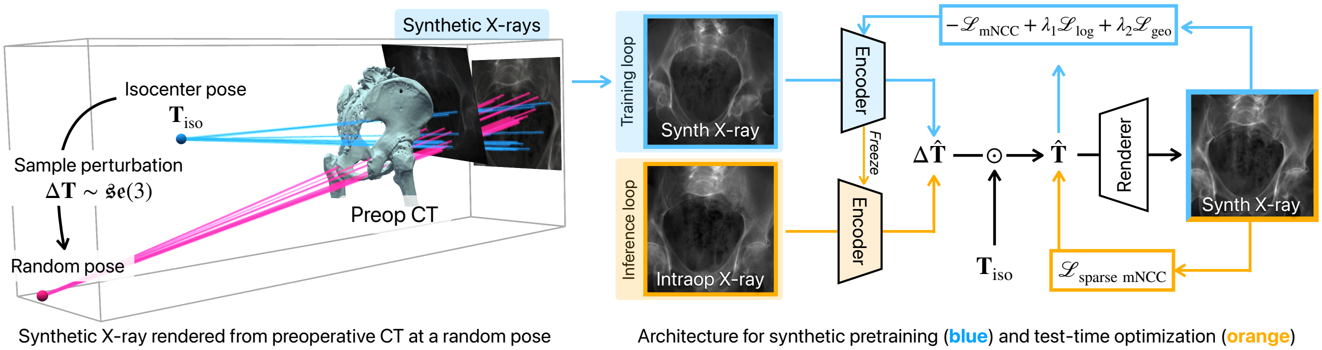

Figure 2. **DiffPose setup.** *Left:* Camera poses are sampled via random perturbations from the isocenter pose $\mathbf{T}_{\text{iso}}$. *Right:* An encoder is trained to regress the pose of a synthetic X-ray using a combination of image similarity and $\mathbf{SE}(3)$-geodesic losses. At inference, the pose of a real intraoperative X-ray is estimated by the encoder and iteratively refined using test-time optimization with differentiable rendering.

3. Preliminaries

Let $\mathbf{V} : \mathbb{R}^3 \to \mathbb{R}$ represent a 3D anatomical structure and $\mathbf{I} : \mathbb{R}^2 \to \mathbb{R}$ represent a 2D X-ray of $\mathbf{V}$ taken at an unknown camera pose $\mathbf{T} \in \mathbf{SE}(3)$. These functions are related by the projection operator $\mathcal{P}(\mathbf{T}) : \mathbf{V} \to \mathbf{I}$, which models X-ray image formation using a pinhole camera whose intrinsic parameters are known. Given $\mathbf{I}$ and $\mathbf{V}$, the goal of 2D/3D registration is to estimate the unknown pose $\mathbf{T}$.

Differentiable rendering of synthetic X-rays. We render synthetic X-rays via a physics-based simulation of the image formation model. Specifically, we model the linear attenuation of X-rays in tissue and ignore second-order effects such as scattering and beam hardening [48]. Let $\mathbf{r}(\alpha) = \mathbf{s} + \alpha(\mathbf{p} - \mathbf{s})$ be a ray cast from the radiation source $\mathbf{s} \in \mathbb{R}^3$ to a pixel $\mathbf{p} \in \mathbb{R}^3$ on the imaging plane, where the spatial location of $\mathbf{p}$ relative to $\mathbf{s}$ is given by the known intrinsic parameters of the imaging system (*e.g.* focal length). As $\mathbf{r}$ travels through the anatomic volume $\mathbf{V}$, it loses intensity proportional to the linear attenuation coefficient $\mathbf{V}(\mathbf{x})$ of every point $\mathbf{x} \in \mathbb{R}^3$ along its path. Assuming the ray has initial intensity I_0, its attenuated intensity once it has reached $\mathbf{p}$ is given by the Beer-Lambert law [49]:

$$
\begin{aligned}
I(\mathbf{p}) &= I_0 \exp\left(- \int_{\mathbf{x} \in \mathbf{r}} \mathbf{V}(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right) && (1) \\
&= I_0 \exp\left(- \int_0^1 \mathbf{V}\big(\mathbf{r}(\alpha)\big) \|\mathbf{r}'(\alpha)\| \, \mathrm{d}\alpha \right) && (2) \\
&= I_0 \exp\big(- I_\mu(\mathbf{p}) \big), && (3)
\end{aligned}
$$

where $I_\mu(\mathbf{p}) \triangleq \|\mathbf{p} - \mathbf{s}\| \int_0^1 \mathbf{V}\big(\mathbf{s} + \alpha(\mathbf{p} - \mathbf{s})\big) \, \mathrm{d}\alpha$ is proportional to the total energy absorbed by $\mathbf{V}$. We model the line integral in Eq. (3) by approximating $\mathbf{V}$ with a discrete 3D CT volume (*i.e.* a voxel grid of linear attenuation coefficients). The discrete line integral is computed as in [47]:

$$
\|\mathbf{p} - \mathbf{s}\| \sum_{m=1}^{M-1} \mathbf{V}\left[\mathbf{r}\left(\frac{\alpha_{m+1} + \alpha_m}{2} \right) \right] (\alpha_{m+1} - \alpha_m), \quad (4)
$$

where α_m parameterizes the locations where $\mathbf{r}$ intersects one of the orthogonal planes comprising the CT volume and M is the number of such intersections. For a pixel grid $\mathbf{P} \in \mathbb{R}^{n \times 3}$ that forms the imaging plane, we can transform $\mathbf{P}$ and $\mathbf{s}$ by a rigid transformation $\mathbf{T} \in \mathbf{SE}(3)$ and render synthetic X-rays from any camera pose, *i.e.* $\mathbf{I} = \mathcal{P}(\mathbf{T}) \circ \mathbf{V}$. The rendering equation (4) can be implemented as a series of vectorized tensor operations [22], enabling the rendering of synthetic X-rays that are differentiable with respect to $\mathbf{T}$.

Lie theory of $\mathbf{SE}(3)$. $\mathbf{SE}(3)$ is the Lie group of all rigid transformations in 3D. $\mathbf{T} \in \mathbf{SE}(3)$ comprises a rotation $\mathbf{R}(\boldsymbol{\varphi}) \in \mathbf{SO}(3)$ and a translation $\mathbf{t} \in \mathbb{R}^3$, where $\boldsymbol{\varphi} \in \mathbb{R}^d$ is a Euclidean parameterization of $\mathbf{SO}(3)$. Classical parameterizations such as the axis-angle representation ($d = 3$), Euler angles ($d = 3$), and quaternions ($d = 4$) suffer from degeneracies that can make pose estimation difficult, such as Gimbal lock or discontinuities in the representation space [57]. This has motivated the development of several higher-dimensional ($d > 5$) alternatives with theoretical advantages [32, 40, 57]. Instead of parameterizing rotations and translations separately, we propose performing 2D/3D registration directly in the Lie algebra $\mathfrak{se}(3)$, which is isomorphic to $\mathbb{R}^3 \times \mathbb{R}^3$ and jointly represents an axis-angle rotation about a translation vector [38]. We find that pose estimation and gradient-based optimization are most accurate when performed in $\mathfrak{se}(3)$. Further, we leverage the Lie algebra to define geodesic losses for training pose regression CNNs and to parameterize random distributions over $\mathbf{SE}(3)$. Appendix A summarizes the relevant Lie theory.

4. Methods

Let $\mathcal{I}$ be the set of X-ray images (both synthetic and real). Let $\mathcal{E} : \mathcal{I} \to \mathbb{R}^d \times \mathbb{R}^3$ be an encoder network that maps an X-ray to a Euclidean parameterization of its camera pose. For every parameterization, there exists a surjective mapping $g : \mathbb{R}^d \times \mathbb{R}^3 \to \mathbf{SE}(3)$. *E.g.*, for $\mathfrak{se}(3)$, $g(\cdot) = \exp(\cdot)$.

Let $\mathcal{P}$ be a differentiable renderer that generates synthetic X-rays from a CT volume at any camera pose $\mathbf{T} \in \mathbf{SE}(3)$. We implement $\mathcal{E}$ as a convolutional neural network and $\mathcal{P}$ using DiffDRR [22] such that the entire 2D/3D registration framework is end-to-end differentiable (Figure 2).

4.1. Training Pose Estimation Networks

Sampling synthetic X-ray poses. The dimensions and voxel spacing of a given CT volume depend on the patient's anatomy, the type of surgery, and the imaging equipment available at the given hospital. Therefore, instead of sampling random camera poses in absolute coordinates, we sample random perturbations relative to the patient's isocenter pose $\mathbf{T}_{\text{iso}}$. The isocenter is defined as the posteroanterior (PA) view where the camera is pointed directly at the patient (*i.e.* $\mathbf{R}_{\text{iso}} = \mathbf{I}$). The isocenter translation is $\mathbf{t}_{\text{iso}} = (b_x\Delta_x, b_y\Delta_y, b_z\Delta_z)/2$ where (b_x, b_y, b_z) are the number of voxels in each dimension of the CT volume and $(\Delta_x, \Delta_y, \Delta_z)$ are the spacings of each voxel in millimeters per voxel. We sample the three rotational and three translational parameters of $\mathfrak{se}(3)$ from normal distributions defined with sufficient variance to capture wide perturbations from the isocenter. Applying the exponential map to samples from the Lie algebra yields the perturbation $\Delta\mathbf{T} \in \mathbf{SE}(3)$. The random camera pose is then $\mathbf{T} = \Delta\mathbf{T} \circ \mathbf{T}_{\text{iso}}$.

Synthetic pose regression pretraining. After a preoperative CT scan has been obtained for a patient who will undergo surgery, we begin training $\mathcal{E}$. The training data consists solely of synthetic X-rays generated from the patient's preoperative CT at random camera poses. Given a randomly sampled camera pose $\mathbf{T} \in \mathbf{SE}(3)$, we render the associated synthetic X-ray $\mathbf{I} = \mathcal{P}(\mathbf{T}) \circ \mathbf{V}$. From this image, we estimate the perturbation $\Delta\hat{\mathbf{T}} \triangleq \mathcal{E}(\mathbf{I})$, construct the estimated pose $\hat{\mathbf{T}} = \Delta\hat{\mathbf{T}} \circ \mathbf{T}_{\text{iso}}$, and render the predicted X-ray $\hat{\mathbf{I}} = \mathcal{P}(\hat{\mathbf{T}}) \circ \mathbf{V}$. We optimize the weights of $\mathcal{E}$ using a combination of geodesic losses on $\mathbf{SE}(3)$ between $\hat{\mathbf{T}}$ and $\mathbf{T}$ and image-based losses on $\hat{\mathbf{I}}$ and $\mathbf{I}$.

4.2. Registration Losses

Geodesic pose regression losses. We use the geodesic distance between the estimated and ground truth camera poses as a regression loss when training $\mathcal{E}$. Given two rotation matrices $\mathbf{R}_A, \mathbf{R}_B \in \mathbf{SO}(3)$, the angular distance between their axes of rotation is

$$d_\theta(\mathbf{R}_A, \mathbf{R}_B) = \arccos\left(\frac{\operatorname{trace}(\mathbf{R}_A^T\mathbf{R}_B) - 1}{2}\right) \tag{5}$$
$$= \|\log(\mathbf{R}_A^T\mathbf{R}_B)\|, \tag{6}$$

where $\log(\cdot)$ is the logarithm map on $\mathbf{SO}(3)$ [26]. Using the logarithm map on $\mathbf{SE}(3)$, this generalizes to a geodesic loss function on camera poses $\mathbf{T}_A, \mathbf{T}_B \in \mathbf{SE}(3)$:

$$\mathcal{L}_{\log}(\mathbf{T}_A, \mathbf{T}_B) = \|\log(\mathbf{T}_A^{-1}\mathbf{T}_B)\|. \tag{7}$$

We can also formulate a geodesic distance on $\mathbf{SE}(3)$ with units of length. Using the camera's focal length f, we convert the angular distance in Eq. (5) to an arc length:

$$d_\theta(\mathbf{R}_A, \mathbf{R}_B; f) = \frac{f}{2} d_\theta(\mathbf{R}_A, \mathbf{R}_B). \tag{8}$$

When combined with the Euclidean distance on the translations $d_t(\mathbf{t}_A, \mathbf{t}_B) = \|\mathbf{t}_A - \mathbf{t}_B\|$, this yields the *double geodesic* loss on $\mathbf{SE}(3)$ [15]:

$$\mathcal{L}_{\text{geo}}(\mathbf{T}_A, \mathbf{T}_B; f) = \sqrt{d_\theta^2(\mathbf{R}_A, \mathbf{R}_B; f) + d_t^2(\mathbf{t}_A, \mathbf{t}_B)}. \tag{9}$$

Multiscale NCC. Global normalized cross correlation (NCC) is a widely-used metric used to quantify the similarity between two images $\mathbf{I}_A$ and $\mathbf{I}_B$:

$$\text{NCC}(\mathbf{I}_A, \mathbf{I}_B) = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \mathbf{Z}_A[i,j]\mathbf{Z}_B[i,j], \tag{10}$$

where $\mathbf{Z} = (\mathbf{I} - \mu(\mathbf{I}))/\sigma(\mathbf{I})$ is an $N \times M$ image normalized by its pixel-wise mean and standard deviation. Alternatively, Eq. (10) can be evaluated on overlapping patches of $\mathbf{I}_A$ and $\mathbf{I}_B$ to capture local correlations in small regions [6]. When it successfully converges, we find the best registration accuracy achieved by local NCC is an order of magnitude better than global NCC. However, the metric is also more unstable and frequently guides the test-time optimizer to a poor quality solution. Instead, we find averaging NCC over multiple patch sizes, a generalization known as multiscale NCC ($\mathcal{L}_{\text{mNCC}}$), to be both more accurate and more numerically stable. This image similarity metric consistently achieves sub-millimeter registration accuracy while successfully converging more reliably than local NCC. Section 5 and Appendix B provide evaluations and visualizations of multiple image losses for this application.

Composite pretraining loss. We train $\mathcal{E}$ using the following loss function (λ_1 and λ_2 are hyperparameters):

$$-\mathcal{L}_{\text{mNCC}}(\mathbf{I}, \hat{\mathbf{I}}) + \lambda_1 \mathcal{L}_{\log}(\mathbf{T}, \hat{\mathbf{T}}) + \lambda_2 \mathcal{L}_{\text{geo}}(\mathbf{T}, \hat{\mathbf{T}}). \tag{11}$$

Sparse differentiable rendering. Evaluating NCC across multiple scales is computationally expensive. We accelerate this calculation by estimating multiscale NCC with sparse rendering. Using the activations of the encoder $\mathcal{E}$, we identify important regions of the image. During test-time optimization, we use the activation map to sample random subsets of image patches to render. In addition to estimating local NCC with this small number of patches, we can also estimate global NCC without additional rendering cost. This is achieved by evaluating Eq. (10) with the sparse subset of rendered pixels. Visualizations and speed benchmarks for sparse multiscale NCC are provided in the Appendix C.

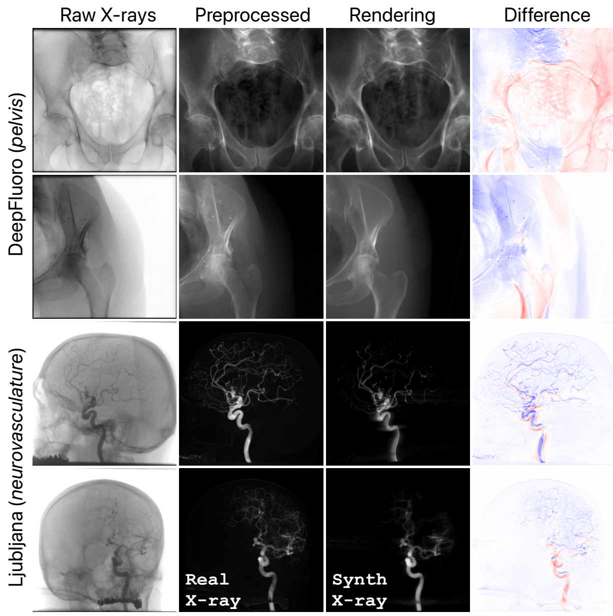

Figure 3. **Sample renders.** Raw X-rays are preprocessed to match the image formation model in [22]. Difference maps between intraoperative X-rays and renderings from a preoperative CT visualize domain shift between real and synthetic images. In Row 2, the left femur moves between acquisition of preoperative and intraoperative images; in Rows 3 and 4, 3D volumes do not capture the smallest cranial blood vessels [33], so they cannot be rendered.

4.3. Test-Time Optimization

At test time (*i.e.* during the surgery), the model is applied to real X-rays for which the ground truth pose $\mathbf{T}$ is unknown. Given a real X-ray $\mathbf{I}$ acquired intraoperatively, the encoder estimates the underlying pose $\hat{\mathbf{T}}$. Despite training on a massive amount of synthetic data, this pose estimate often does not achieve optimal alignment on real images. Further refining this initial pose estimate with intensity-based registration decreases registration error by orders of magnitude. To perform test-time optimization, we render the synthetic X-ray $\hat{\mathbf{I}} = \mathcal{P}(\hat{\mathbf{T}}) \circ \mathbf{V}$ and compute an image-based similarity metric between the real and synthetic X-rays. Since $\mathcal{P}$ is differentiable, we can use gradient-based optimization to update $\hat{\mathbf{T}}$ such that the similarity between $\mathbf{I}$ and $\hat{\mathbf{I}}$ is increased. Note that gradient updates occur in the Euclidean representation space $\mathbb{R}^3 \times \mathbb{R}^3$ such that the estimated pose remains on the $\mathbf{SE}(3)$ manifold.

4.4. Landmark-Based Evaluation

We evaluate performance with mean Target Registration Error (mTRE), an independent measurement of registration accuracy at important anatomical landmarks. For each CT volume, the open-source datasets we use in our experiments predefine a set of m 3D anatomical landmarks $\mathbf{M} \in \mathbb{R}^{3\times m}$. Since the intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3\times 3}$ of each imaging system used in these studies is known, we can calculate the perspective projection of $\mathbf{M}$ for any camera pose. mTRE is defined as the average distance between the projections given by the ground truth and estimated camera poses:

$$\mathrm{mTRE}(\mathbf{T}, \hat{\mathbf{T}}) = \frac{1}{m} \|\mathbf{K}([\mathbf{R} \mid \mathbf{t}] - [\hat{\mathbf{R}} \mid \hat{\mathbf{t}}])\mathbf{M}\|_F. \quad (12)$$

Since we never use landmark supervision for training or test-time optimization, mTRE is an independent metric of registration accuracy. Following guidelines from the American Association of Physicists in Medicine (AAPM) [11], a registration is successful if mTRE $\leq$ 1 mm. We report a sub-millimeter success rate (**SMSR**) for all experiments.

4.5. Implementation Details

Pretraining. We used a ResNet18 backbone to implement the encoder $\mathcal{E}$. Extracted features were processed by two fully connected layers, which regressed the rotational ($\mathbb{R}^3$) and translational ($\mathbb{R}^3$) components of the pose, respectively. For each CT scan, a patient-specific $\mathcal{E}$ was trained from scratch using synthetic X-ray images rendered on the fly. The Adam optimizer with warm-up was used with a maximal learning rate of 1×10^{-3} and a cosine learning rate scheduler. As both the rendering of synthetic X-rays and CNN regression were performed on the same device, we were limited to batch sizes of eight 256×256 images on an A6000 GPU, leading us to replace batch normalization with group normalization [55]. $\mathcal{E}$ was trained on 1,000,000 synthetic images, which took approximately 12 hours. Lastly, we set $\lambda_1 = \lambda_2 = 10^{-2}$ in Eq. (11) and used multiscale NCC with patch sizes 13 and 256 (*i.e.* the whole image).

Test-time optimization. Test-time optimization was performed to refine the initial pose estimate. We used sparse multiscale NCC with 100 patches and a patch size of 13 as the image-based loss function and parameterized the space of poses with $\mathfrak{se}(3)$. Gradient-based pose updates were performed using the Adam optimizer with a learning rate of 7.5×10^{-3} on the rotational components of $\mathfrak{se}(3)$ and a learning rate of 7.5×10^{0} on the translational components of $\mathfrak{se}(3)$ for 250 iterations. Additionally, a step learning rate decay at a factor of 0.9 was applied every 25 iterations.

5. Experiments

5.1. Datasets

We perform an in-depth analysis of two public datasets that provide calibrated camera poses and expert annotations used for evaluation. We first evaluate our method using the DeepFluoro dataset, an open-source collection of pelvic X-rays and CTs from six cadavers [25]. For each subject, there is one CT scan and between 24-111 X-rays for a total of six CTs and 366 X-rays. For each X-ray, the ground

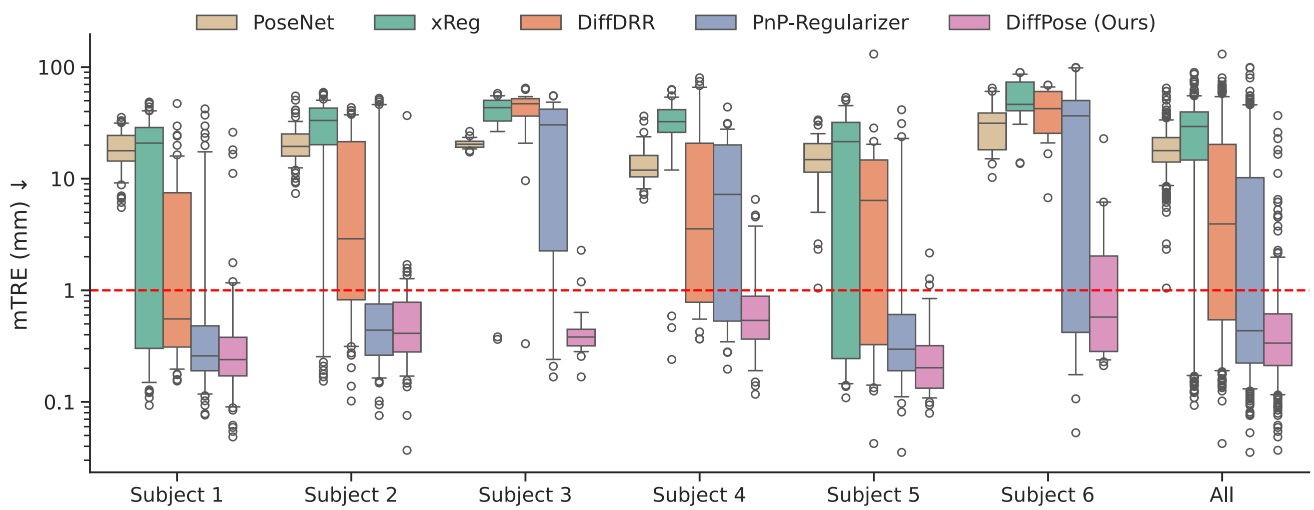

Figure 4. **Quantitative evaluation.** Evaluation of different registration methods on the DeepFluoro dataset via mTRE. A method successfully registered an X-ray if the final mTRE was less than one millimeter (red line). DiffPose is the only method that consistently achieves sub-millimeter mTRE, outperforming fully supervised methods (PoseNet and PnP-Regularizer). Note that the y-axis is on a log-scale.

truth extrinsic matrix is provided. The X-ray imaging system is also calibrated, *i.e.* its intrinsic matrix is known. To demonstrate the difficulty in adapting existing registration methods to new surgical procedures, we also perform evaluations on the Ljubljana dataset [41], a clinical dataset consisting of 2D and 3D X-ray angiography images from 10 patients undergoing neurovascular surgery. Each patient has two 2D X-rays and one 3D volume with accompanying ground truth extrinsic and intrinsic camera matrices. To demonstrate generality across surgical domains, we focus our ablations on the DeepFluoro dataset and do not change hyperparameters or modeling decisions for the Ljubljana dataset. Importantly, we note that sample sizes are generally low in real 2D/3D surgical datasets due to the difficulty of acquiring expert annotations with calibrated poses in surgical settings, limiting the utility of supervised methods.

Raw images acquired intraoperatively measure X-ray attenuation. To match our physics-based differentiable renderer, we converted these to X-ray absorption images by inverting Eq. (1), yielding $I_\mu(\mathbf{p}) = \log I_0 - \log I(\mathbf{p})$. We estimate the rays' initial intensity as $I_0 = \max I(\mathbf{r})$, where the max is taken over all images in the dataset (*i.e.* the pixel with maximum intensity corresponds to an X-ray that intersected no anatomy and therefore experienced zero attenuation). Additionally, we cropped the attenuation images by 50 pixels on each side to remove the effect of the collimator. Finally, log transforming yields the absorption image. Raw attenuation X-rays, preprocessed absorption X-rays, and synthetic X-rays rendered from a preoperative CT at the ground truth camera pose are shown in Figure 3.

Unlike many datasets used to evaluate previous 2D/3D registration methods, the CTs in the DeepFluoro and Ljubljana datasets are not reconstructed from the acquired X-rays. This makes our registration problem more challenging because the subject can move between the acquisition of the preoperative and intraoperative images, meaning the 2D X-rays are no longer directly embedded in the 3D CT scan. Testing on datasets with independent and physically-acquired 2D and 3D modalities avoids test set leakage during evaluation and directly simulates the clinical use case.

5.2. Baseline Methods

An evaluation of multiple existing 2D/3D registration approaches was conducted. We first compare the proposed method against PoseNet [29], a supervised pose regression algorithm. We evaluate PoseNet using leave-one-subject-out cross-validation, training a new model for each subset of five subjects. We also include two unsupervised intensity-based registration methods: xReg [24] and DiffDRR [22]. Initialized at a PA pose, xReg uses a multi-stage gradient-free optimization routine and a patch-based image-gradient NCC similarity metric [23]. DiffDRR was initialized at the same PA pose as xReg. We evaluated DiffDRR using the same learning rates and image-based loss function as our test-time optimization. Finally, we compare against two PnP registration methods [25]. In the first approach (PnP), a U-Net is trained to extract landmarks from 2D X-rays, again using a leave-one-subject-out approach. These landmarks are used with a PnP solver to predict the camera pose, which can be subject to poor initialization especially if less than four landmarks are visible in the 2D image [25]. A more robust approach, PnP-Regularizer, uses the detected landmarks to formulate a regularizer on potential camera poses, effectively optimizing estimated mTRE (12).

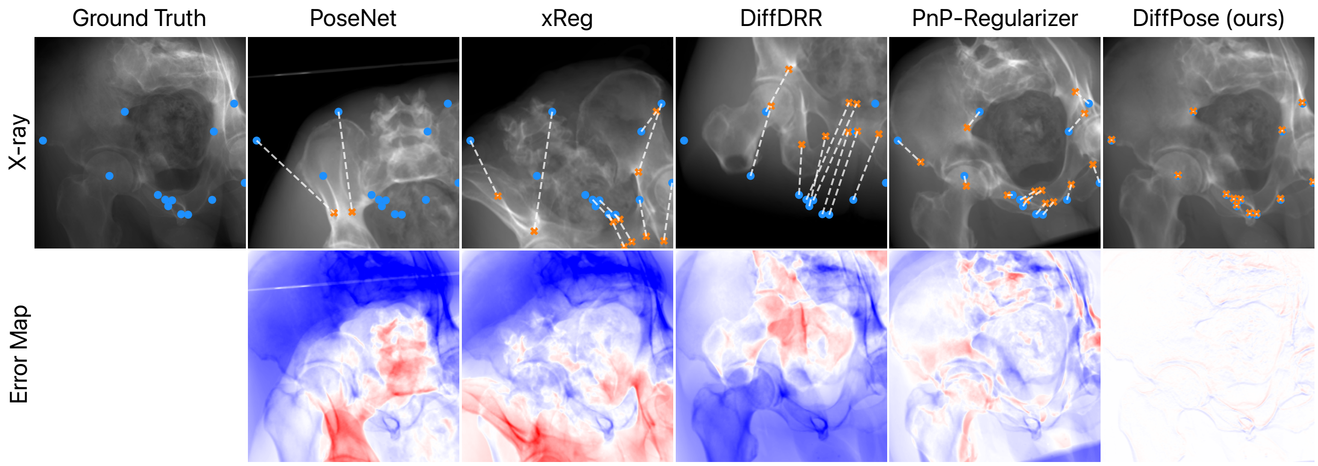

Figure 5. **Qualitative visualizations.** *Top:* Renderings at the final pose estimates produced by different registration methods. Correspondences are drawn between true landmarks (blue) and estimated landmarks (orange). *Bottom:* To compare geometric alignment and not appearance, error maps are computed as the difference between the X-rays rendered at the ground truth pose and the final pose estimate.

5.3. Results

DeepFluoro dataset. Figure 4 reports the mTRE statistics for each subject and in aggregate for DeepFluoro. Sub-millimeter success rates across baselines are summarized in Table 1. Our encoder's estimates were within 10 mm of the true pose for 80% of test cases, a commonly used cutoff for successful registration [21], but only achieved *sub*-millimeter mTRE on 1% of test cases (Table 2). Refining the initial pose with test-time optimization was thus critical to achieving successful registration, leading to 87% of test cases achieving sub-millimeter mTRE across all subjects.

As subjects are not registered to a common reference frame and there is a limited number of X-rays, a purely supervised pose regressor (PoseNet [29]) incurs large error. Traditional gradient-free optimization (xReg [24]) is able to successfully recover the poses of some X-rays but frequently converges to suboptimal minima. The gradient-based approach (DiffDRR [22]) is more robust than gradient-free methods, but not to a surgically viable degree. Taken together, xReg and DiffDRR show that test-time optimization alone is insufficient for accurate intraoperative 2D/3D registration. PnP methods provide better initial pose estimates on average, but their accuracy diminishes when only a few landmarks are present in the 2D image. When combined with xReg, PnP-Regularizer successfully registered 69% of test X-rays with high inter-subject variability: on Subjects 3, 4, and 6, DiffPose consistently achieves sub-millimeter accuracy while PnP-Regularizer's median mTRE is around 10 mm. Inter-subject variability arises from out-of-distribution failures when the pose or appearance of a particular X-ray does not match the training set, demonstrating the need for patient-specific methods. Visual comparisons are shown in Figure 5.

Table 1. **Baseline comparisons.** Sub-millimeter success rate (SMSR) and runtime statistics for DeepFluoro subjects. Methods are classified by registration strategy: unsupervised intensity-based (I), supervised landmark-based (L), and pose regression (R).

	I	L	R	SMSR ↑	Time (s) ↓
PoseNet [29]			✓	0%	N/A
xReg [24]	✓			20%	1.3 ± 1.9
DiffDRR (mNCC) [22]	✓			37%	7.5 ± 5.9
PnP [25]		✓		1%	N/A
PnP-Regularizer [25]	✓	✓		69%	$\mathbf{1.1 \pm 1.6}$
DiffPose (Ours)	✓		✓	**87%**	2.2 ± 1.2

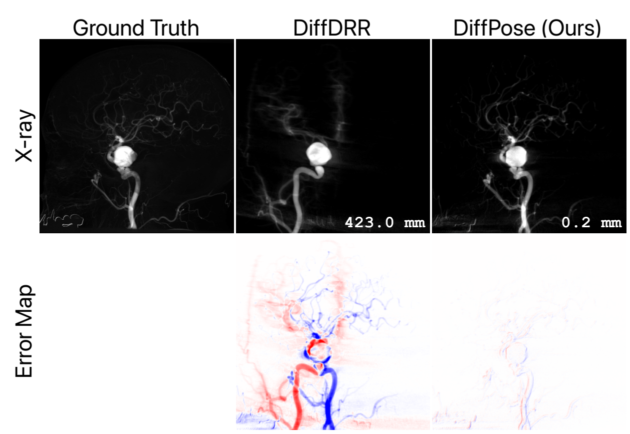

Figure 6. **External dataset validation.** Using modeling decisions from a pelvic dataset, DiffPose demonstrates high registration accuracy on blood vessels in the brain. *Top:* Renders at final pose estimates with associated mTRE. *Bottom:* Error maps between X-rays rendered at the estimated and ground truth poses.

Table 2. Test time-optimization and pose regression loss ablations.

	SMSR ↑	mTRE (mm) ↓
DiffPose (Ours)	**87%**	**0.9 ± 2.8**
Remove test-time opt. (TTOpt)	1%	5.4 ± 4.3
Remove TTOpt and $\mathcal{L}_{\text{mNCC}}$	0%	8.3 ± 4.7
Remove TTOpt, $\mathcal{L}_{\text{geo}}$, and $\mathcal{L}_{\text{log}}$	0%	18.1 ± 4.9
Remove TTOpt and use L_2	0%	35.5 ± 10.4

Ljubljana dataset. Certain baselines are not applicable to the Ljubljana dataset. PnP-based methods cannot be trained as there are no segmentation masks with which to estimate landmarks, and the small sample size (20 X-rays) precludes training an accurate PoseNet. However, as our method does not require any labeled training data, we train patient-specific pose regression networks. We compared to DiffDRR [22] initialized at a PA pose and found that DiffPose achieved a median mTRE of 0.2 ± 10.6 mm (85% SMSR) as compared to DiffDRR's 263.7 ± 193.9 mm (25% SMSR). Qualitative results are visualized in Figure 6. The high performance of DiffPose on the neurovasculature, a highly distinct surgical domain, demonstrates its generalizability across patients and anatomical structures.

5.4. Ablation Studies

Choice of pretraining loss. Table 2 compares our composite loss (11) to isolated image (Row 3) and geodesic losses (Row 4), as well as an L2 loss on Euler angles (Row 5). All models were trained with the same number of synthetic images. We find that combining image and geodesic losses produces the most accurate model (Row 2).

Choice of $\mathbf{SE}(3)$ parameterization. Table 3 compares optimizing camera poses directly in $\mathfrak{se}(3)$ to more conventional parameterizations that treat rotational and translational components as independent. Despite previously reported deficiencies of Euler angles and quaternions in other contexts [40, 57], they are the most performant parameterizations after the proposed $\mathfrak{se}(3)$ for 2D/3D registration.

Table 3. Comparison of different parameterizations of $\mathbf{SE}(3)$ for test-time optimization with identical initialization.

	SMSR ↑	mTRE (mm) ↓
$\mathfrak{se}(3)$ (Ours)	**87%**	**0.9 ± 2.8**
Axis-angle × $\mathbb{R}^3$	75%	1.1 ± 4.6
Euler angles × $\mathbb{R}^3$	83%	1.0 ± 3.2
Quaternion × $\mathbb{R}^3$	83%	1.0 ± 3.7
Rotation6D [57] × $\mathbb{R}^3$	83%	1.0 ± 3.2
Rotation10D [40] × $\mathbb{R}^3$	81%	1.5 ± 4.8
Quaternion Adjugate [32] × $\mathbb{R}^3$	77%	3.5 ± 9.7

Table 4. Image similarity loss comparisons. All rows correspond to test time optimization with identical pretrained initialization.

	SMSR ↑	mTRE (mm) ↓
Sparse mNCC (Ours)	**87%**	**0.9 ± 2.8**
Global NCC	27%	4.4 ± 5.2
Local NCC [6]	81%	1.7 ± 5.2
Gradient NCC [23]	9%	13.3 ± 5.2
SSIM	1%	13.4 ± 7.8
MSE	0%	30.8 ± 15.8
MAE	0%	28.4 ± 14.1

Choice of image similarity metric. Table 4 compares sparse multiscale NCC (mNCC) to previously described loss functions for 2D/3D rigid registration. Local NCC [6] performs competitively but has high variance, demonstrating the instability of the loss. Using mNCC, which averages global and local NCC, stabilizes the variance and results in the highest success rate (see Appendices B and C).

6. Discussion

Limitations and future work. DiffPose estimates a single rigid transformation between pre- and intraoperative scans, precluding direct application to deformable target structures. Fortunately, our core contributions can be extended to estimate piecewise rigid transformations (*e.g.* one transform per rigid anatomical component), which can be aggregated to model arbitrary deformations [5]. Further, DiffPose uses per-subject training (analogous to NeRF [36]), which may be too slow for emergency surgeries that do not allow for hours-long pretraining. We expect that our pose regression networks can be well-initialized by first training on a dataset of preregistered CT scans and synthetic X-rays, and then rapidly fine-tuning on a new subject in a few iterations of our patient-specific pretraining task (*e.g.* with MAML [18]).

Conclusion. Intraoperative 2D/3D registration holds immense promise that has as yet been unfulfilled due to a high rate of registration failure and dependence on expert supervision. We present DiffPose, the first intraoperative 2D/3D registration framework that is sub-millimeter accurate without using any surgically impractical supervision. DiffPose demonstrated strong performance relative to current methods while retaining surgically relevant runtime, thus enabling successful applications to intraoperative pelvic and neurovascular procedures and many more surgical domains.

Acknowledgements. The authors thank Andrew Abumoussa, Nalini Singh, and Bill Worstell for helpful feedback. This work was supported, in part, by NIH NIBIB NAC P41EB015902, NIH NINDS U19NS115388, NIH NIBIB 5T32EB001680-19, and a Takeda Fellowship.

References

[1] Andrew Abumoussa, Vivek Gopalakrishnan, Benjamin Succop, Michael Galgano, Sivakumar Jaikumar, Yueh Z Lee, and Deb A Bhowmick. Machine learning for automated and real-time two-dimensional to three-dimensional registration of the spine using a single radiograph. *Neurosurgical Focus*, 54(6):E16, 2023. 2

[2] John R Adler Jr, Martin J Murphy, Steven D Chang, and Steven L Hancock. Image-guided robotic radiosurgery. *Neurosurgery*, 44(6):1299–1306, 1999. 1

[3] Sebastian Andress, Alex Johnson, Mathias Unberath, Alexander Felix Winkler, Kevin Yu, Javad Fotouhi, Simon Weidert, Greg Osgood, and Nassir Navab. On-the-fly augmented reality for orthopedic surgery using a multimodal fiducial. *Journal of Medical Imaging*, 5(2):021209–021209, 2018. 1

[4] Souha Aouadi and Laurent Sarry. Accurate and precise 2d–3d registration based on x-ray intensity. *Computer vision and image understanding*, 110(1):134–151, 2008. 1

[5] Vincent Arsigny, Olivier Commowick, Nicholas Ayache, and Xavier Pennec. A fast and log-euclidean polyaffine framework for locally linear registration. *Journal of Mathematical Imaging and Vision*, 33:222–238, 2009. 8

[6] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011. 2, 4, 8

[7] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018. 2

[8] Martin Berger, K Müller, A Aichert, M Unberath, J Thies, J-H Choi, R Fahrig, and A Maier. Marker-free motion correction in weight-bearing cone-beam ct of the knee joint. *Medical physics*, 43(3):1235–1248, 2016. 2

[9] José Luis Blanco-Claraco. A tutorial on SE(3) transformation parameterizations and on-manifold optimization. *arXiv preprint arXiv:2103.15980*, 2021. 12

[10] Eric Brachmann and Carsten Rother. Learning less is more-6d camera localization via 3d surface regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4654–4662, 2018. 2

[11] Kristy K Brock, Sasa Mutic, Todd R McNutt, Hua Li, and Marc L Kessler. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the aapm radiation therapy committee task group no. 132. *Medical physics*, 44(7):e43–e76, 2017. 5

[12] William R Brody. Digital subtraction angiography. *IEEE Transactions on Nuclear Science*, 29(3):1176–1180, 1982. 18

[13] Mai Bui, Shadi Albarqouni, Michael Schrapp, Nassir Navab, and Slobodan Ilic. X-ray posenet: 6 dof pose estimation for mobile x-ray devices. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1036–1044. IEEE, 2017. 2

[14] Xin Chen, Jim Graham, Charles Hutchinson, and Lindsay Muir. Automatic inference and measurement of 3d carpal bone kinematics from single view fluoroscopic sequences. *IEEE transactions on medical imaging*, 32(2):317–328, 2012. 2

[15] Gregory S Chirikjian. Partial bi-invariance of SE(3) metrics. *Journal of Computing and Information Science in Engineering*, 15(1):011008, 2015. 4

[16] Ethan Eade. Lie groups for 2d and 3d transformations. *http://ethaneade.com/lie.pdf*, 117:118, 2013. 12

[17] Javier Esteban, Matthias Grimm, Mathias Unberath, Guillaume Zahnd, and Nassir Navab. Towards fully automatic x-ray to ct registration. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 631–639. Springer, 2019. 2

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 8

[19] 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. 2

[20] Cong Gao, Amirhossein Farvardin, Robert B Grupp, Mahsan Bakhtiarinejad, Liuhong Ma, Mareike Thies, Mathias Unberath, Russell H Taylor, and Mehran Armand. Fiducial-free 2d/3d registration for robot-assisted femoroplasty. *IEEE transactions on medical robotics and bionics*, 2(3):437–446, 2020. 1

[21] Cong Gao, Anqi Feng, Xingtong Liu, Russell H Taylor, Mehran Armand, and Mathias Unberath. A fully differentiable framework for 2d/3d registration and the projective spatial transformers. *IEEE transactions on medical imaging*, 2023. 7

[22] Vivek Gopalakrishnan and Polina Golland. Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In *Workshop on Clinical Image-Based Procedures*, pages 1–11. Springer, 2022. 2, 3, 4, 5, 6, 7, 8, 15, 16

[23] Robert B Grupp, Mehran Armand, and Russell H Taylor. Patch-based image similarity for intraoperative 2D/3D pelvis registration during periacetabular osteotomy. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 5*, pages 153–163. Springer, 2018. 6, 8, 13

[24] Robert B Grupp, Rachel A Hegeman, Ryan J Murphy, Clayton P Alexander, Yoshito Otake, Benjamin A McArthur, Mehran Armand, and Russell H Taylor. Pose estimation of periacetabular osteotomy fragments with intraoperative x-ray navigation. *IEEE Transactions on Biomedical Engineering*, 67(2):441–452, 2019. 2, 6, 7

[25] Robert B Grupp, Mathias Unberath, Cong Gao, Rachel A Hegeman, Ryan J Murphy, Clayton P Alexander, Yoshito Otake, Benjamin A McArthur, Mehran Armand, and Russell H Taylor. Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2D/3D registration. *International journal of computer assisted radiology and surgery*, 15:759–769, 2020. 2, 5, 6, 7

[26] Du Q Huynh. Metrics for 3D rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, 35: 155–164, 2009. 4

[27] Srikrishna Jaganathan, Jian Wang, Anja Borsdorf, Karthik Shetty, and Andreas Maier. Deep iterative 2d/3d registration. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pages 383–392. Springer, 2021. 2

[28] Srikrishna Jaganathan, Maximilian Kukla, Jian Wang, Karthik Shetty, and Andreas Maier. Self-supervised 2d/3d registration for x-ray to ct image fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2788–2798, 2023. 2

[29] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 2, 6, 7

[30] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate O(n) solution to the PnP problem. *International journal of computer vision*, 81:155–166, 2009. 2

[31] Shiqi Li, Chi Xu, and Ming Xie. A robust o (n) solution to the perspective-n-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1444–1450, 2012. 2

[32] Chen Lin, Andrew J Hanson, and Sonya M Hanson. Algebraically rigorous quaternion framework for the neural network pose estimation problem. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14097–14106, 2023. 2, 3, 8

[33] Song Luo, Long Jiang Zhang, Felix G Meinel, Chang Sheng Zhou, Li Qi, Andrew D McQuiston, U Joseph Schoepf, and Guang Ming Lu. Low tube voltage and low contrast material volume cerebral ct angiography. *European radiology*, 24: 1677–1685, 2014. 5

[34] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997. 2

[35] Shun Miao, Z Jane Wang, and Rui Liao. A CNN regression approach for real-time 2D/3D registration. *IEEE Transactions on Medical Imaging*, 35(5):1352–1363, 2016. 2

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 8

[37] Shuntaro Mizoe, Yoshito Otake, Takuma Miyamoto, Mazen Soufi, Satoko Nakao, Yasuhito Tanaka, and Yoshinobu Sato. 4d-foot: a fully automated pipeline of four-dimensional analysis of the foot bones using bi-plane x-ray video and ct. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pages 182–192. Springer, 2021. 2

[38] Richard M Murray, Zexiang Li, and S Shankar Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 1994. 3

[39] Van Nguyen, Luis F Alves Pereira, Zhihua Liang, Falk Mielke, Jeroen Van Houtte, Jan Sijbers, and Jan De Beenhouwer. Automatic landmark detection and mapping for 2d/3d registration with bonenet. *Frontiers in Veterinary Science*, 9:923449, 2022. 2

[40] Valentin Peretroukhin, Matthew Giamou, David M Rosen, W Nicholas Greene, Nicholas Roy, and Jonathan Kelly. A smooth representation of belief over SO(3) for deep rotation learning with uncertainty. *arXiv preprint arXiv:2006.01031*, 2020. 2, 3, 8

[41] Franjo Pernus et al. 3d-2d registration of cerebral angiograms: a method and evaluation on clinical images. *IEEE transactions on medical imaging*, 32(8):1550–1563, 2013. 2, 6

[42] Terry M Peters. Image-guided surgery: from x-rays to virtual reality. *Computer methods in biomechanics and biomedical engineering*, 4(1):27–57, 2001. 1

[43] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. *arXiv:2007.08501*, 2020. 12, 15

[44] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the limitations of cnn-based absolute camera pose regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3302–3312, 2019. 2

[45] Beth A Schueler, David F Kallmes, and Harry J Cloft. 3d cerebral angiography: radiation dose comparison with digital subtraction angiography. *American journal of neuroradiology*, 26(8):1898–1901, 2005. 18

[46] Pragyan Shrestha, Chun Xie, Hidehiko Shishido, Yuichi Yoshii, and Itaru Kitahara. X-ray to ct rigid registration using scene coordinate regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 781–790. Springer, 2023. 2

[47] Robert L Siddon. Fast calculation of the exact radiological path for a three-dimensional ct array. *Medical physics*, 12 (2):252–255, 1985. 3, 16

[48] David Staub and Martin J Murphy. A digitally reconstructed radiograph algorithm calculated from first principles. *Medical physics*, 40(1):011902, 2013. 3

[49] Donald F Swinehart. The beer-lambert law. *Journal of chemical education*, 39(7):333, 1962. 3, 15

[50] Dejan Tomaevič, Boštjan Likar, and Franjo Pernuš. “gold standard” 2d/3d registration of x-ray to ct and mr images. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002: 5th International Conference Tokyo, Japan, September 25–28, 2002 Proceedings, Part II 5*, pages 461–468. Springer, 2002. 2

[51] Mathias Unberath, Jan-Nico Zaech, Sing Chun Lee, Bastian Bier, Javad Fotouhi, Mehran Armand, and Nassir Navab. Deepdrr–a catalyst for machine learning in fluoroscopy-guided procedures. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, pages 98–106. Springer, 2018. 16

[52] Mathias Unberath, Cong Gao, Yicheng Hu, Max Judish, Russell H Taylor, Mehran Armand, and Robert Grupp. The impact of machine learning on 2d/3d registration for image-guided interventions: A systematic review and perspective. *Frontiers in Robotics and AI*, 8:716007, 2021. 1

[53] IMJ Van Der Bom, Stefan Klein, Marius Staring, R Homan, Lambertus W Bartels, and Josien PW Pluim. Evaluation of optimization methods for intensity-based 2d-3d registration in x-ray guided interventions. In *Medical Imaging 2011: Image Processing*, pages 657–671. SPIE, 2011. 1

[54] Florian Walch, Caner Hazirbas, Laura Leal-Taixe, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based localization using lstms for structured feature correlation. In *Proceedings of the IEEE international conference on computer vision*, pages 627–637, 2017. 2

[55] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 5

[56] Baochang Zhang, Shahrooz Faghihroohi, Mohammad Farid Azampour, Shuting Liu, Reza Ghotbi, Heribert Schunkert, and Nassir Navab. A patient-specific self-supervised model for automatic X-Ray/CT registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 515–524. Springer, 2023. 2

[57] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 2, 3, 8

[58] L Zollei, Eric Grimson, Alexander Norbash, and W Wells. 2d-3d rigid registration of x-ray fluoroscopy and ct images using mutual information and sparsely sampled histogram estimators. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages II–II. IEEE, 2001. 2, 14