WACV 2026
Tucson, Arizona • Mar 6 - 10, 2026

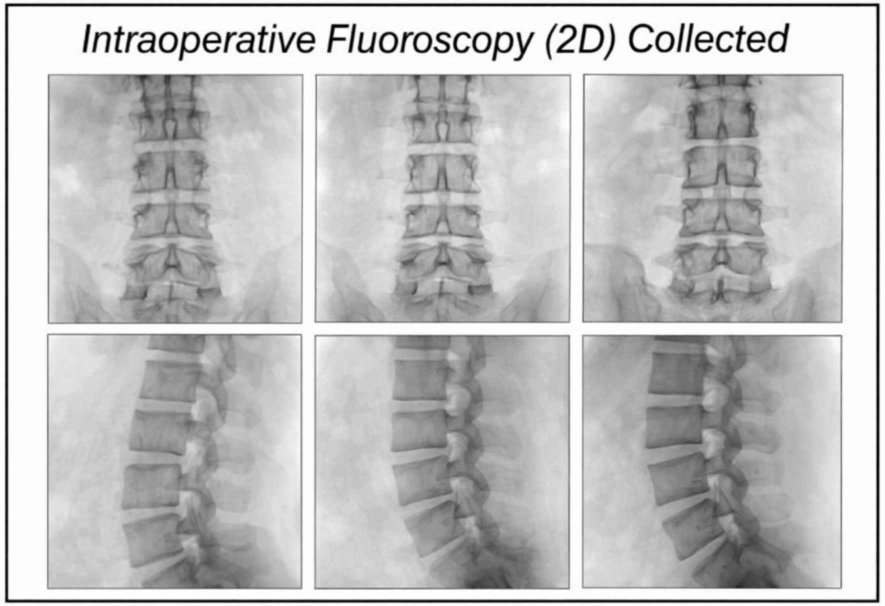
Intraoperative 2D/3D Registration via Spherical Similarity Learning and Differentiable Levenberg-Marquardt Optimization

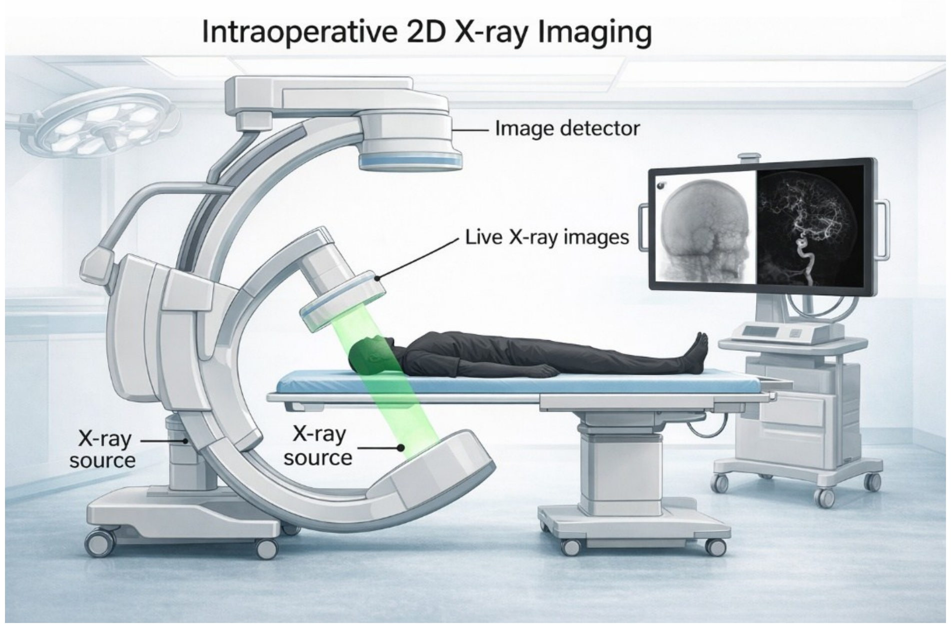
Minheng Chen[1,2], Youyong Kong[1]

[1]School of Computer Science and Engineering, Southeast University, China

[2]Department of Computer Science and Engineering, University of Texas at Arlington, USA

Background

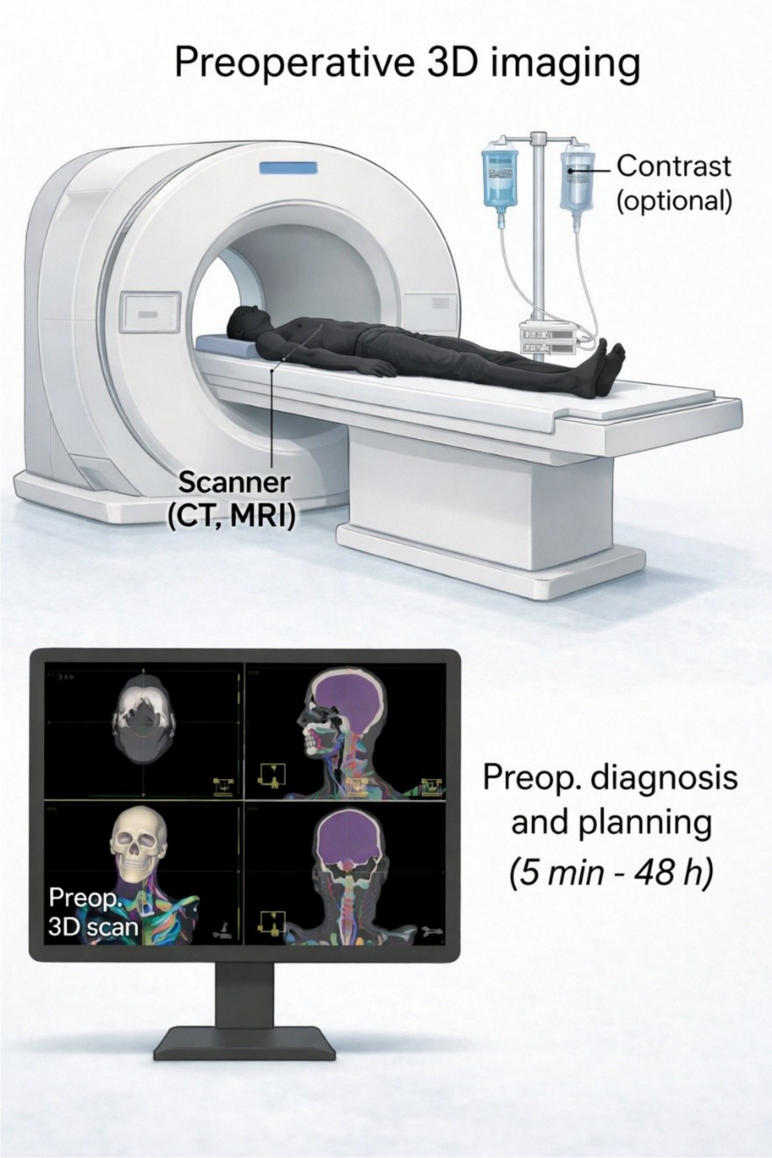

Intraoperative Fluoroscopy (2D) Collected


Intraoperative 2D X-ray Imaging

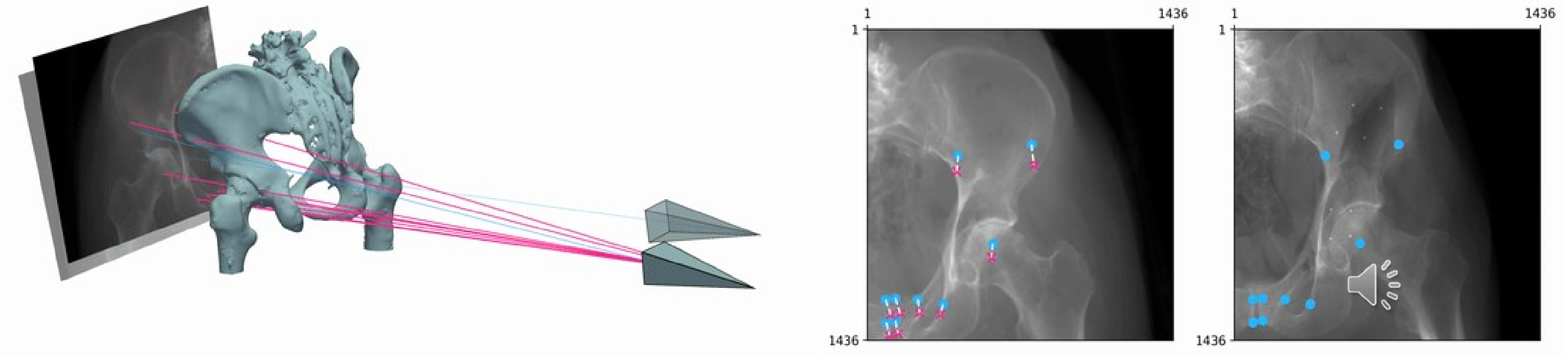
- The extensive application of intraoperative fluoroscopy across specialties has significantly improved patient outcomes by **minimizing invasiveness**, **shortening postoperative recovery times**, and **expanding access to lifesaving treatments** for patients considered too high risk for open surgery.
- 2D X-rays do not provide explicit depth information. This spatial ambiguity encumbers the navigation of medical devices within 3D anatomical structures, increasing the risks of suboptimal device deployment and intraoperative complications.
- Due to the difficulty in differentiating individual vertebrae on X-ray, nearly **50%** of spinal neurosurgeons have reported operating on the wrong vertebra at least once in their careers.

Unberath, Mathias, et al. "The impact of machine learning on 2d/3d registration for image-guided interventions: A systematic review and perspective." Frontiers in Robotics and AI 8 (2021):716007.
Gopalakrishnan, Vivek, et al. "Rapid patient-specific neural networks for intraoperative X-ray to volume registration." *ArXiv* (2025): arXiv-2503.

Background



- Volumetric imaging modalities, offer high-resolution 3D anatomical and functional visualization. While these 3D modalities are routinely acquired preoperatively, they are often unavailable during procedures due to their high radiation dose or incompatibility with surgical equipment and workflows.
- 3D imaging has lengthy acquisition and reconstruction times, which diminishes its utility in real-time surgical navigation.
- live 3D spatial information is inaccessible during most interventions, and mono- or biplane C-arm fluoroscopy remains the intraoperative standard for image guidance.

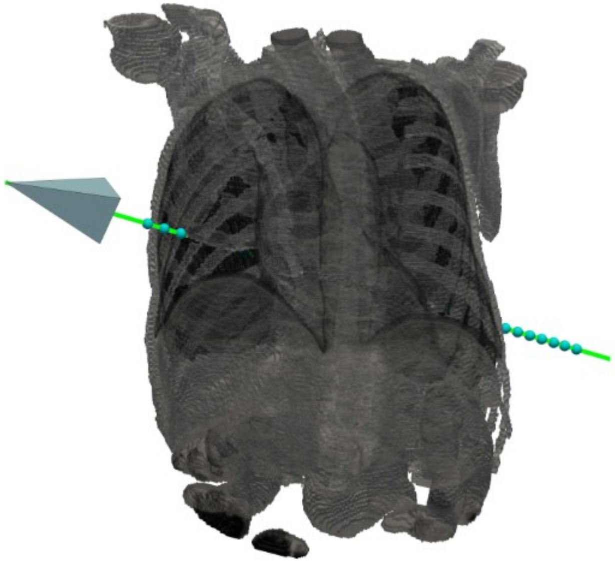


Unberath, Mathias, et al. "The impact of machine learning on 2d/3d registration for image-guided interventions: A systematic review and perspective." Frontiers in Robotics and AI 8 (2021):716007.
Gopalakrishnan, Vivek et al. "Intraoperative 2D/3D image registration via differentiable X-ray rendering." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

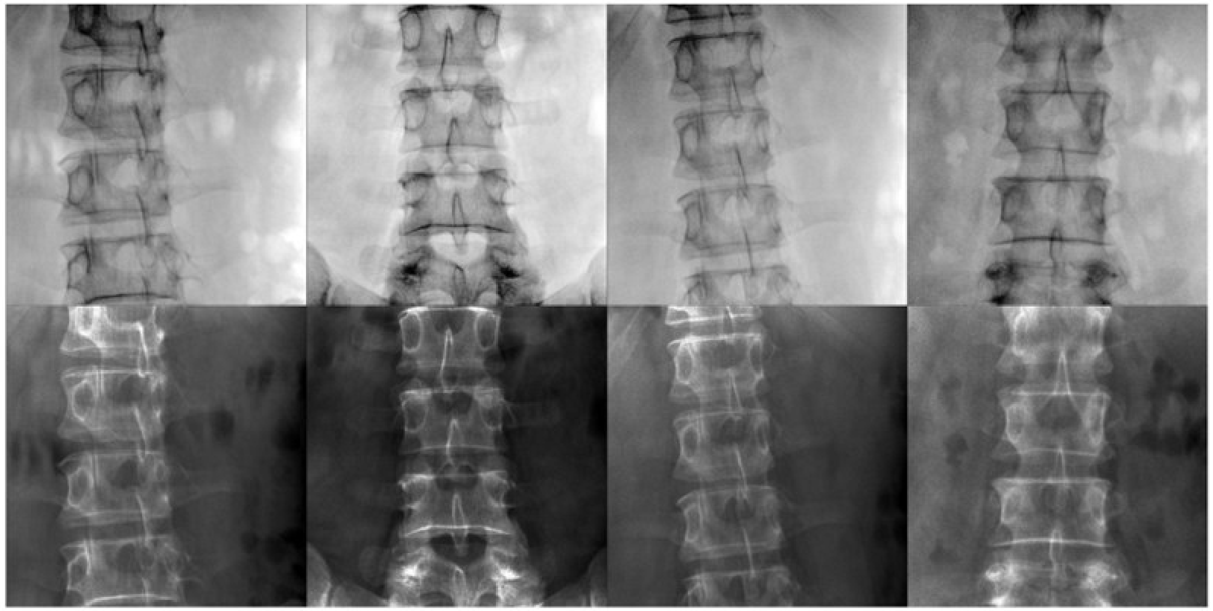
Background

Input: Preoperative 3D volume $V \in \mathbb{R}^3$

2D radiograph $I \in \mathbb{R}^2$



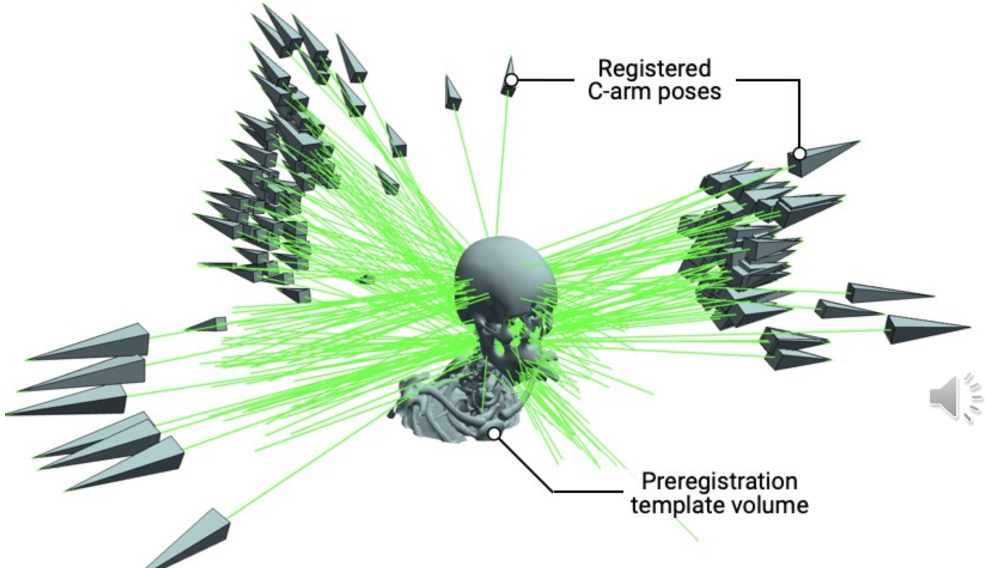
Preoperative volume



Output:

$$\mathbf{T}^* = \arg \min_{\mathbf{T}\in\mathrm{SE}(3)} \mathcal{S}(I, \mathcal{P}(\mathbf{T}) \circ V)$$

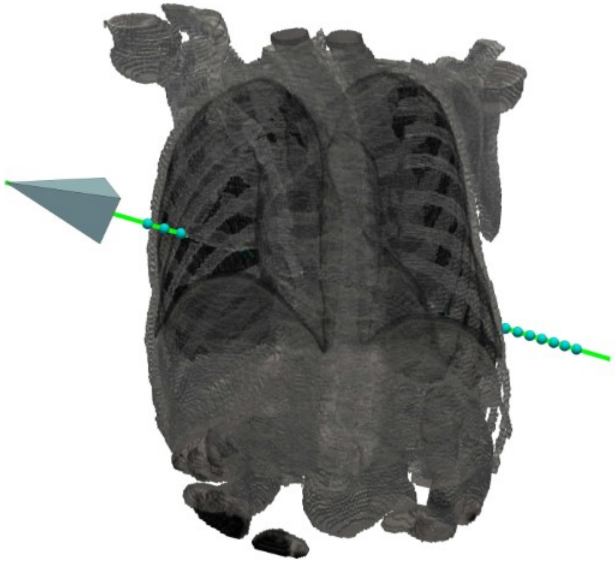
$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \in \mathrm{SE}(3)$$



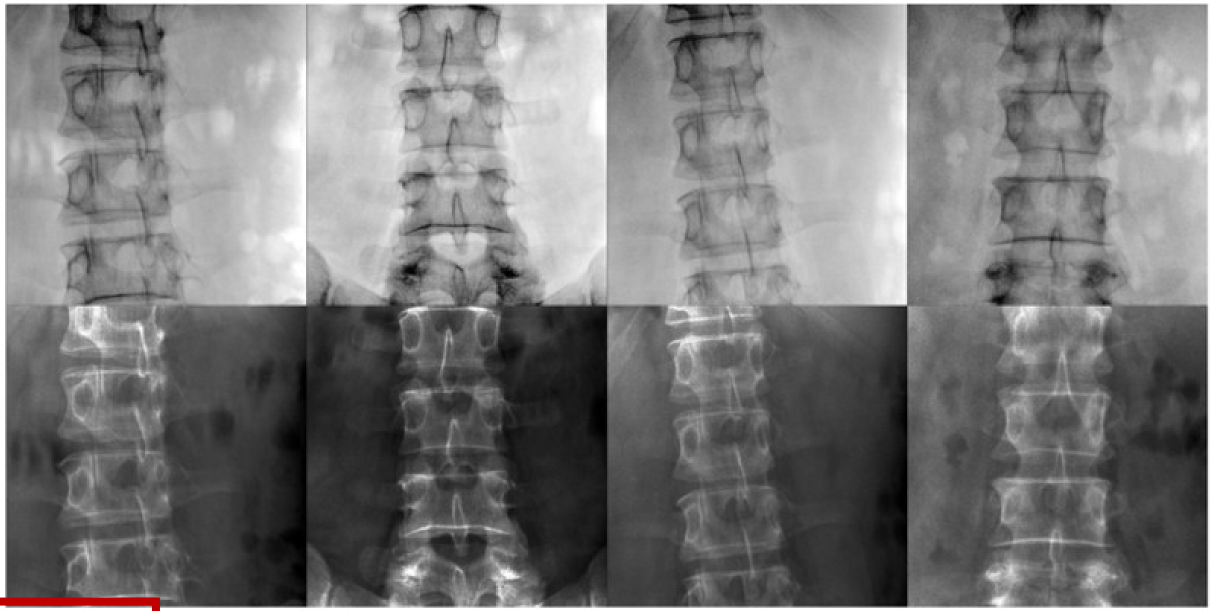
Gopalakrishnan, Vivek, Neel Dey, and Polina Golland. "Intraoperative 2D/3D image registration via differentiable X-ray rendering." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Background

Input: Preoperative 3D volume $V \in \mathbb{R}^3$ 2D radiograph $I \in \mathbb{R}^2$



Preoperative volume



Input: Preoperative 3D volume $V \in \mathbb{R}^3$ 2D radiograph $I \in \mathbb{R}^2$

Output:

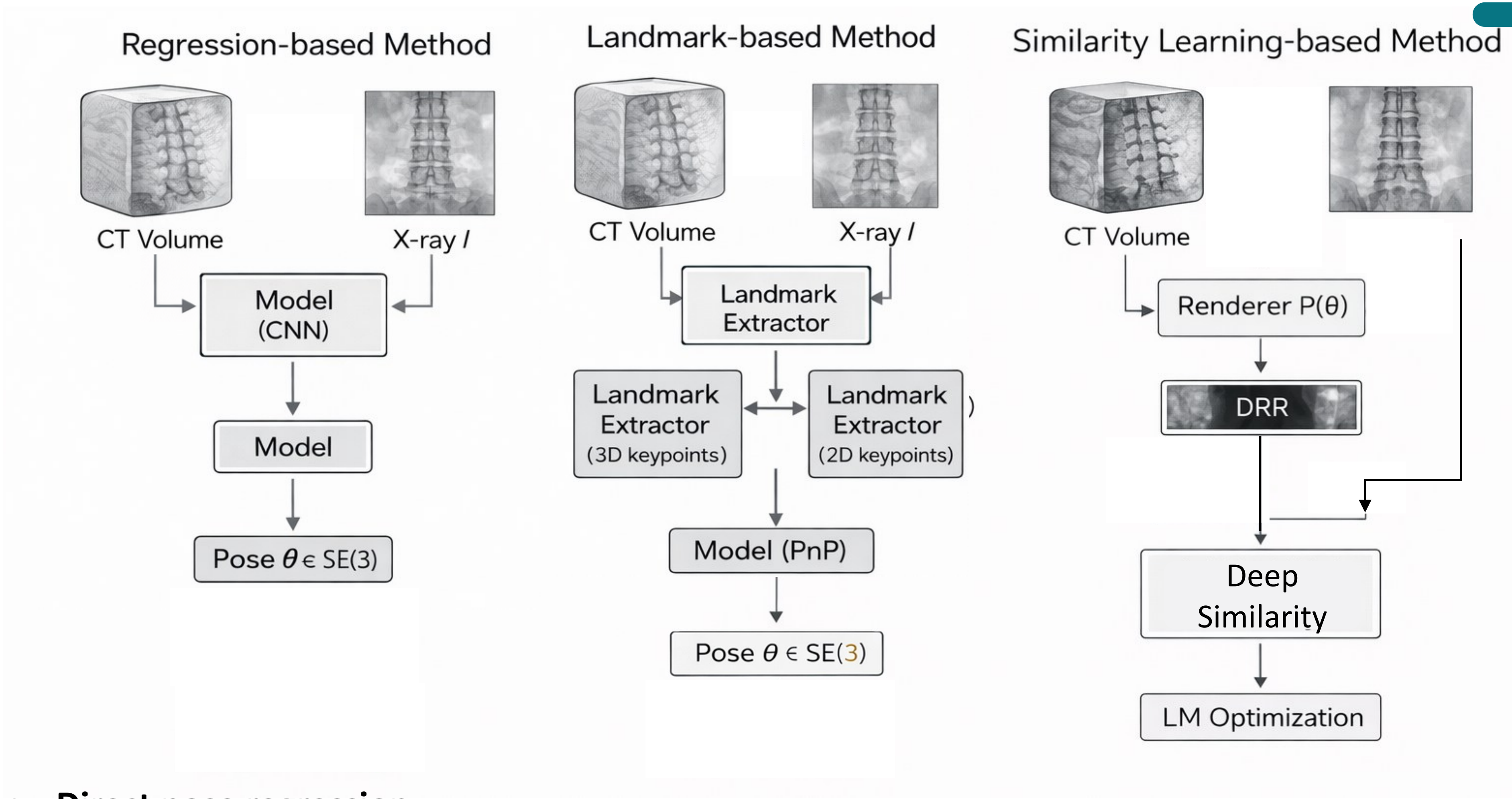
$$\boxed{\mathbf{T}^* = \arg\min_{\mathbf{T}\in\mathrm{SE}(3)} \mathcal{S}(I, \mathcal{P}(\mathbf{T}) \circ V)}$$

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \in \mathrm{SE}(3)$$

Output:

Gopalakrishnan, Vivek, Neel Dey, and Polina Golland. "Intraoperative 2D/3D image registration via differentiable X-ray rendering." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Related Work



Regression-based Method

CT Volume

X-ray I

Model (CNN)

Model

Pose $\theta \in SE(3)$

Landmark-based Method

CT Volume

X-ray I

Landmark Extractor

Landmark Extractor (3D keypoints)

Landmark Extractor (2D keypoints)

Model (PnP)

Pose $\theta \in SE(3)$

Similarity Learning-based Method

CT Volume

Renderer P(θ)

DRR

Deep Similarity

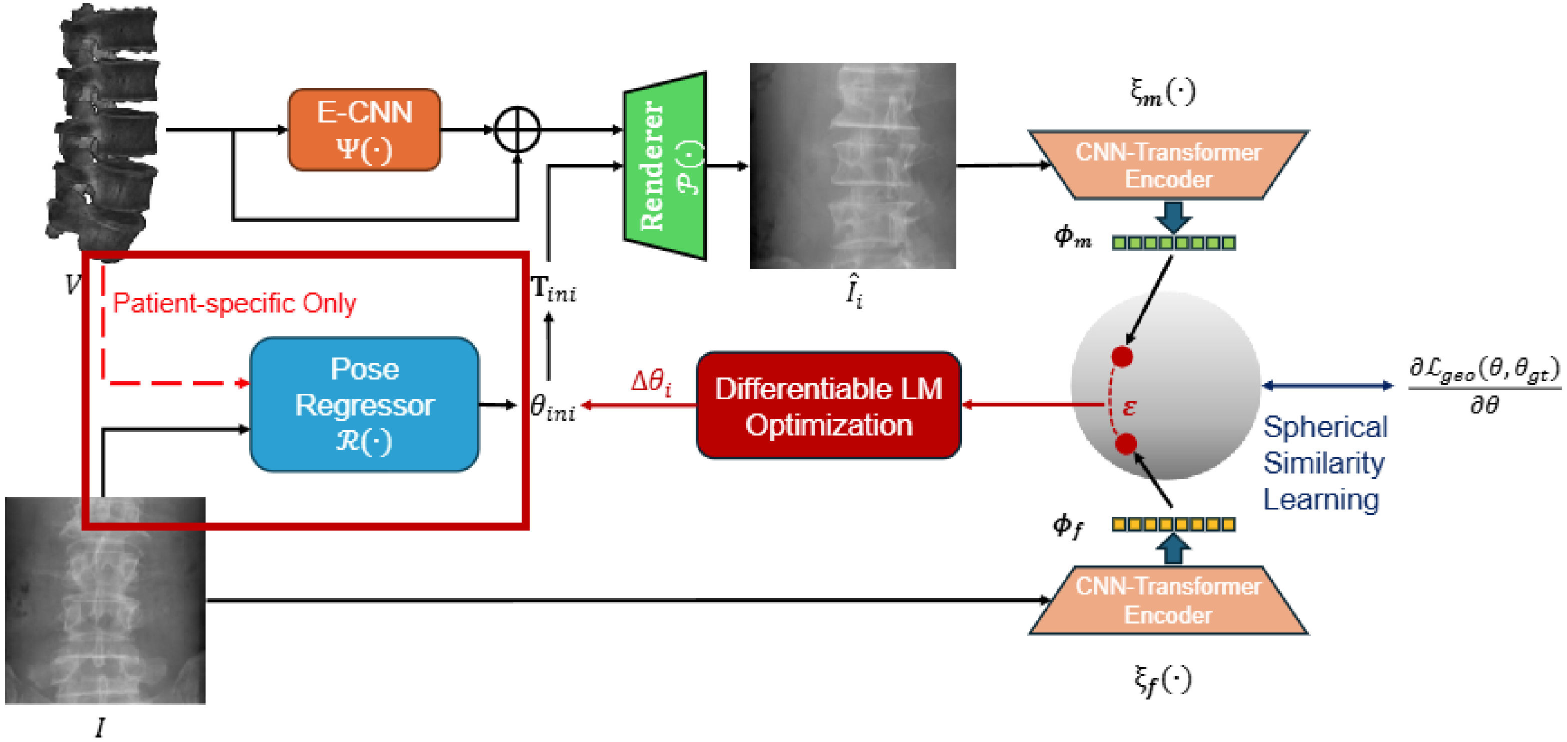
LM Optimization

- Direct pose regression
- Fast
- Low robustness to large offsets
- No explicit geometric constraint

- Requires annotated landmarks
- Sensitive to occlusion
- Depends on keypoint visibility

- Robust to large offsets & occlusion
- Needs no landmarks
- Smoother Optimization Landscape

Methodology



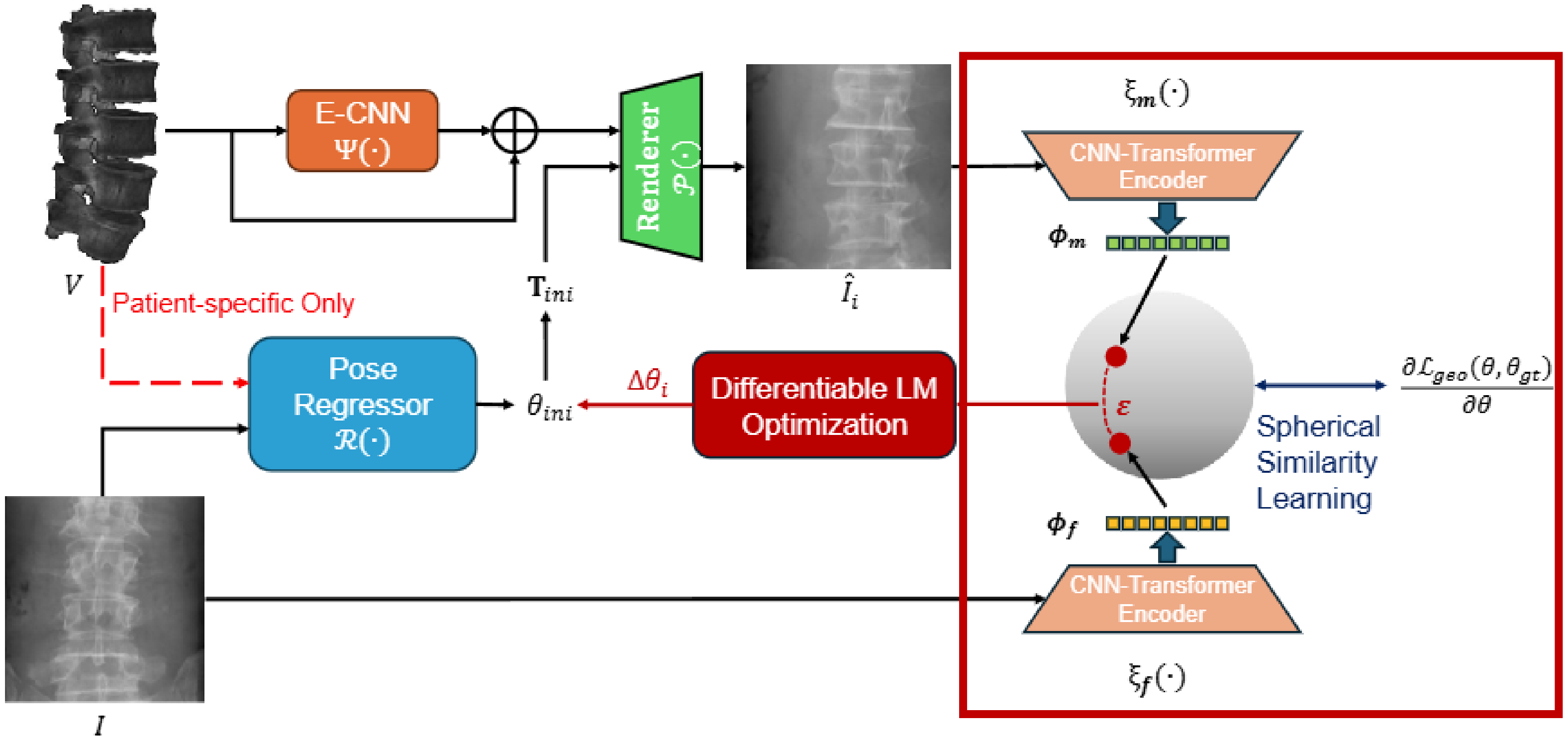
a) A pose regressor for initial pose estimation.

b) A neural network-based deep similarity model in hypersphere space.

c) We introduce a differentiable Levenberg-Marquardt (LM) optimization strategy as an alternative to the gradient descent method to accelerate the convergence of registration.

Chen, Minheng, and Kong, Youyong. "Intraoperative 2D/3D Registration via Spherical Similarity Learning and Inference-Time Differentiable Levenberg-Marquardt Optimization." 2026 IEEE/CVF Winter Conference on Applications of Computer Vision.

Methodology

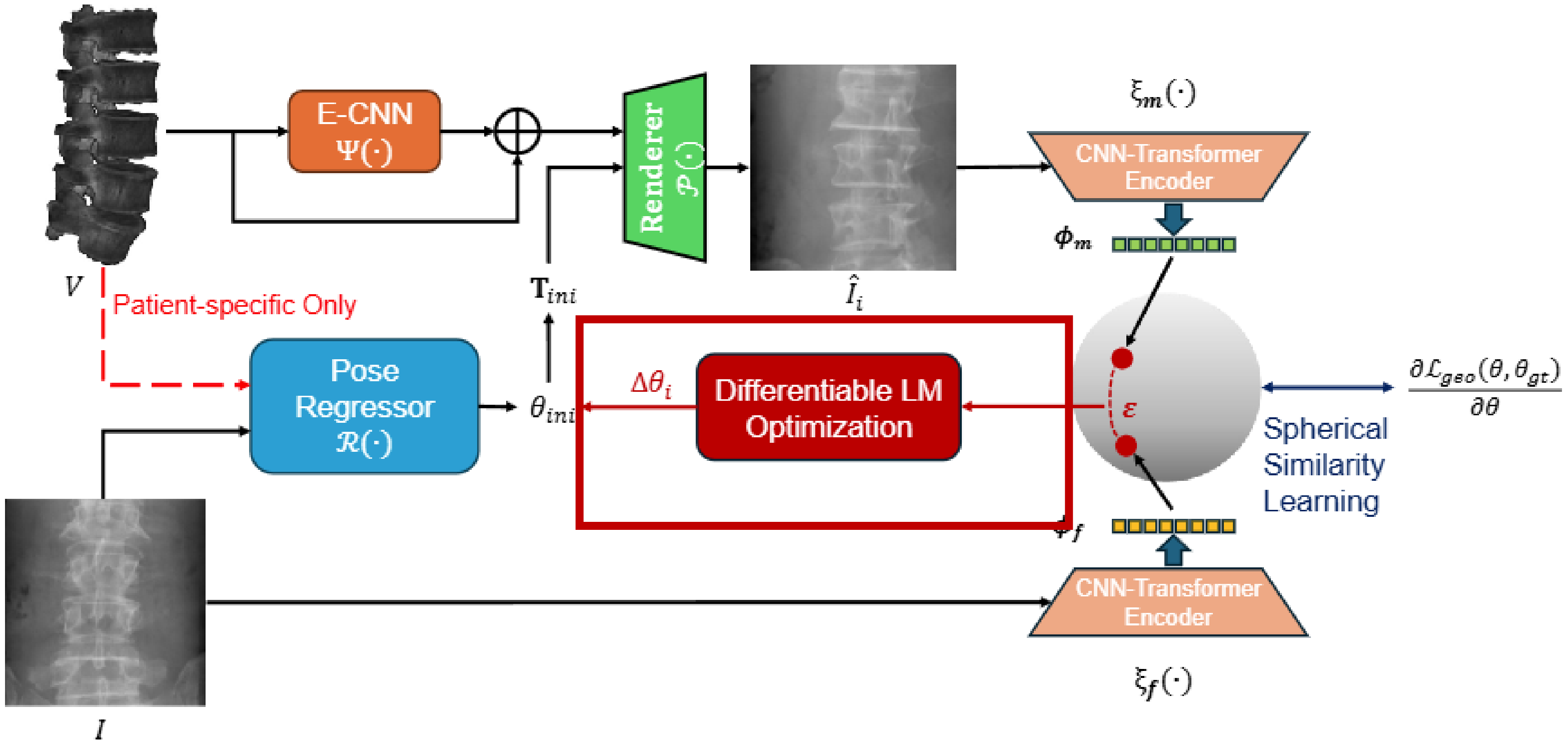


a) A pose regressor for initial pose estimation.

b) A neural network-based deep similarity model in hypersphere space.

c) We introduce a differentiable Levenberg-Marquardt (LM) optimization strategy as an alternative to the gradient descent method to accelerate the convergence of registration.

Methodology



a) A pose regressor for initial pose estimation.

b) A neural network-based deep similarity model in hypersphere space.

c) We introduce a differentiable Levenberg-Marquardt (LM) optimization strategy as an alternative to the gradient descent method to accelerate the convergence of registration.

Spherical Similarity Learning

- Many learning-based registration methods extract image features using neural networks and then measure the similarity between two feature vectors using Euclidean distance.
- Rigid body transformations belong to the Lie group **SE(3)**, which forms a curved Riemannian manifold rather than a flat vector space.
- Euclidean distance only provides a local approximation of the true geodesic distance on this manifold. As a consequence, the resulting optimization landscape may become irregular and non-smooth, increasing the risk of instability and potentially causing the algorithm to converge to incorrect solutions during the search process.

$$\Phi_m = \mathrm{EXP}(\phi_m), \quad \Phi_f = \mathrm{EXP}(\phi_f)$$

Riemannian mapping from Euclidean space to the spherical manifold

$$\mathrm{EXP}(\boldsymbol{\phi}) = \mathbf{N} \cos \|\bar{\phi}\| + \bar{\phi} \frac{\sin \|\bar{\phi}\|}{\|\bar{\phi}\|}$$

$$d(\Phi_m, \Phi_f) = \arccos\Big(\langle \Phi_m, \ \Phi_f \rangle\Big)$$

Spherical distance between two points lie on the sphere

$$\varepsilon = S(\phi_m, \phi_f) = \sum_{i=1}^{H} \sum_{j=1}^{W} (1 - \Phi_m[i,j,:]^T \Phi_f[i,j,:])$$

Bi-variant distance on SO(4) manifold

- SE(3) is ***Left-variance***, affecting the symmetry of objective function and the stability of optimization landscape.
- Map the pose representation to the bi-invariant group SO(4).
- A bi-invariant metric ensures the distance between two elements remains unchanged regardless of the reference frame.
- Independent of coordinate choices, leading to a more geometrically consistent metric.

$$\begin{aligned}\mathcal{L}^{\mathfrak{se}(3)}_{\text{geo}}(\theta_A,\theta_B) &= \left\|\mathcal{LOG}\left(\mathbf{T}_A^{-1}\mathbf{T}_B\right)\right\| \qquad (12)\\ &= \left\|\mathcal{LOG}\left(\mathcal{EXP}(-\theta_A)\,\mathcal{EXP}(\theta_B)\right)\right\|\end{aligned}$$

Geodesic distance on $\mathfrak{se}(3)$ Riemannian manifold.

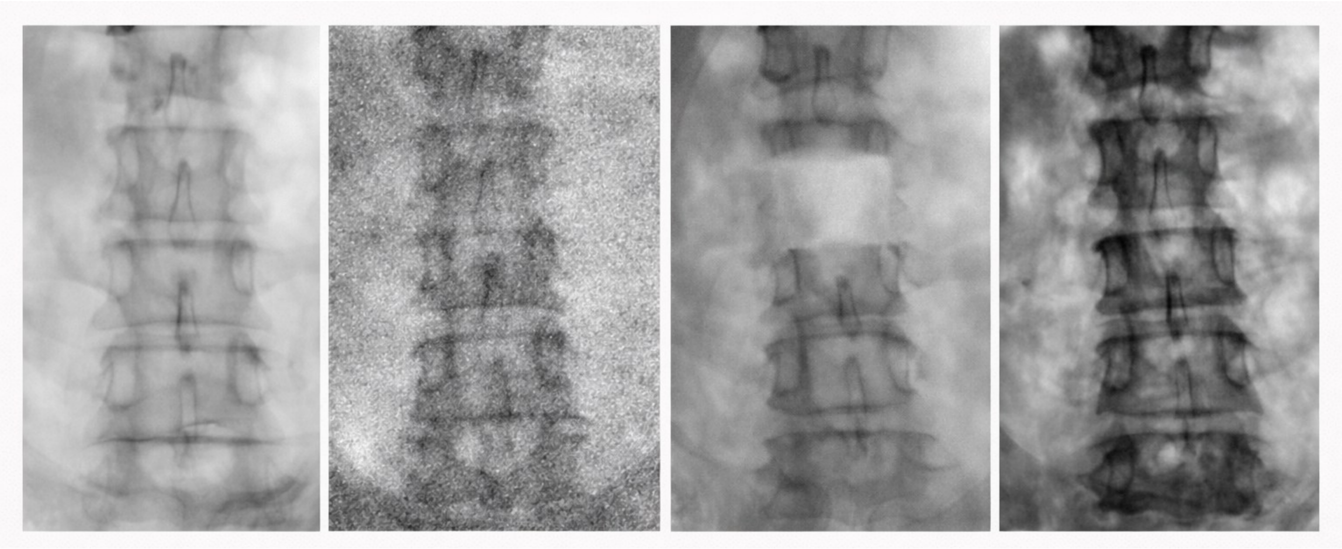
$$\mathcal{L}^{\mathrm{SO}(4)}_{\text{geo}}(\theta_A,\theta_B) = \|\mathbf{H}_A - \mathbf{H}_B\|_F = \|\mathcal{M}(\theta_A), \mathcal{M}(\theta_B)\|_F$$

Geodesic distance on SO(4) manifold.

$$\mathbf{H} = \mathcal{M}(\theta) = \begin{bmatrix} \mathbf{R} & \dfrac{\mathbf{t}}{2f} \\ -\mathbf{t}^\top\mathbf{R} & 1 \end{bmatrix}$$

$$\mathcal{L} = \mathcal{L}_{\text{geo}}\left(\frac{\partial\mathcal{L}_{\text{net}}(\theta)}{\partial\theta}, \frac{\partial\mathcal{L}_{\text{geo}}(\theta,\theta_{gt})}{\partial\theta}\right) \quad \text{where} \quad \mathcal{L}_{\text{net}} = \varepsilon$$

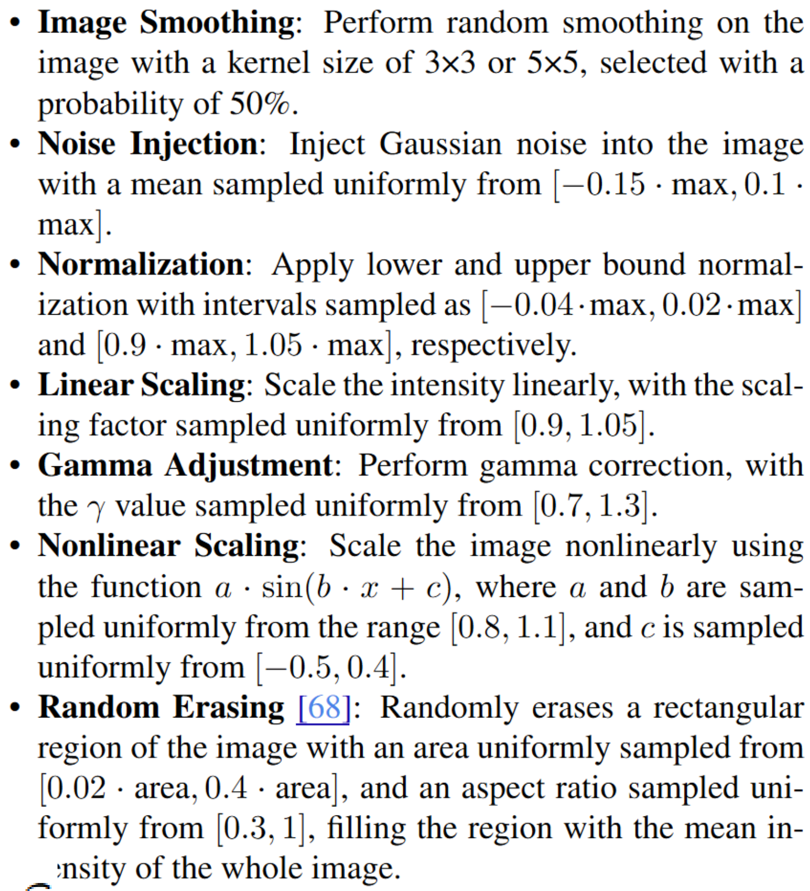
Training and Inference



$$\mathcal{L}^{\mathfrak{se}(3)}_{\text{geo}}(\theta_A, \theta_B) = \left\| \mathcal{LOG}\left(\mathbf{T}_A^{-1}\mathbf{T}_B\right) \right\| \qquad (12)$$

$$= \left\| \mathcal{LOG}\left(\mathcal{EXP}(-\theta_A)\,\mathcal{EXP}(\theta_B)\right) \right\|$$

Domain Randomization



- **Image Smoothing**: Perform random smoothing on the image with a kernel size of 3×3 or 5×5, selected with a probability of 50%.
- **Noise Injection**: Inject Gaussian noise into the image with a mean sampled uniformly from $[-0.15 \cdot \max, 0.1 \cdot \max]$.
- **Normalization**: Apply lower and upper bound normalization with intervals sampled as $[-0.04 \cdot \max, 0.02 \cdot \max]$ and $[0.9 \cdot \max, 1.05 \cdot \max]$, respectively.
- **Linear Scaling**: Scale the intensity linearly, with the scaling factor sampled uniformly from $[0.9, 1.05]$.
- **Gamma Adjustment**: Perform gamma correction, with the γ value sampled uniformly from $[0.7, 1.3]$.
- **Nonlinear Scaling**: Scale the image nonlinearly using the function $a \cdot \sin(b \cdot x + c)$, where a and b are sampled uniformly from the range $[0.8, 1.1]$, and c is sampled uniformly from $[-0.5, 0.4]$.
- **Random Erasing** [68]: Randomly erases a rectangular region of the image with an area uniformly sampled from $[0.02 \cdot \text{area}, 0.4 \cdot \text{area}]$, and an aspect ratio sampled uniformly from $[0.3, 1]$, filling the region with the mean in- nsity of the whole image.

Pose regressor loss

$$\mathcal{L} = \mathcal{L}_{\text{geo}}\left(\frac{\partial \mathcal{L}_{\text{net}}(\theta)}{\partial \theta}, \frac{\partial \mathcal{L}_{\text{geo}}(\theta, \theta_{gt})}{\partial \theta}\right) \quad \text{where} \quad \mathcal{L}_{\text{net}} = \mathcal{E}$$

Similarity network loss

Training and Inference

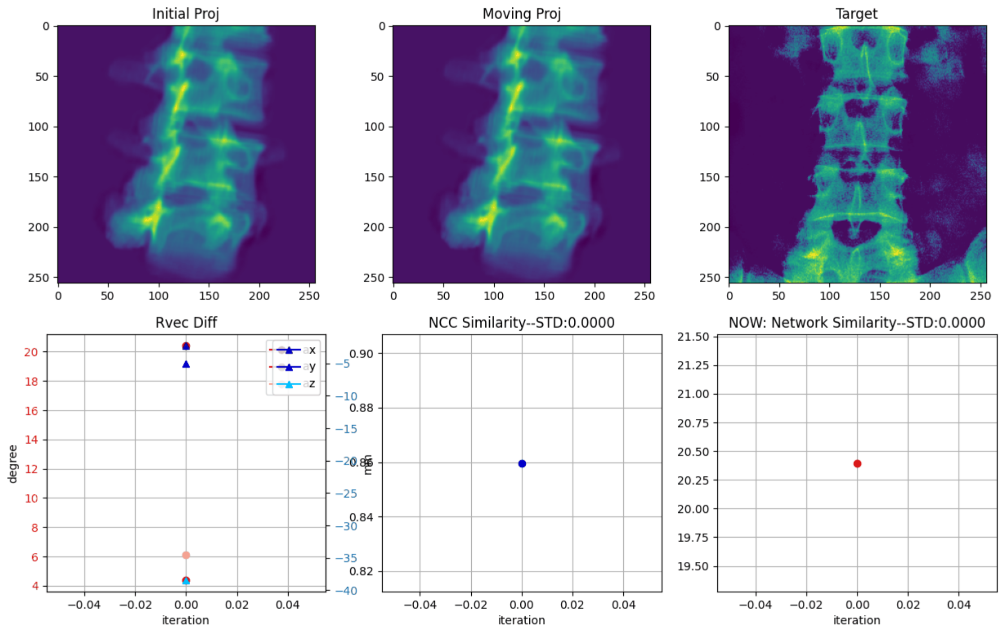
During the inference phase, the estimated initial pose is first obtained through the regressor, followed by differentiable Levenberg-Marquardt (LM) optimization based on spherical similarity learning.

The pose refinement is then formulated as an optimization problem:

$$\hat{\theta} \;=\; \arg\min_{\theta} \mathcal{L}_{net}(\theta)$$

At each LM iteration, starting from the previous estimate pose, the left-multiplied pose increment is computed as:

$$\Delta\theta_i = (J^T W J + \lambda I)^{-1} J^T W \mathbf{r}(\theta_{i-1})$$



Results

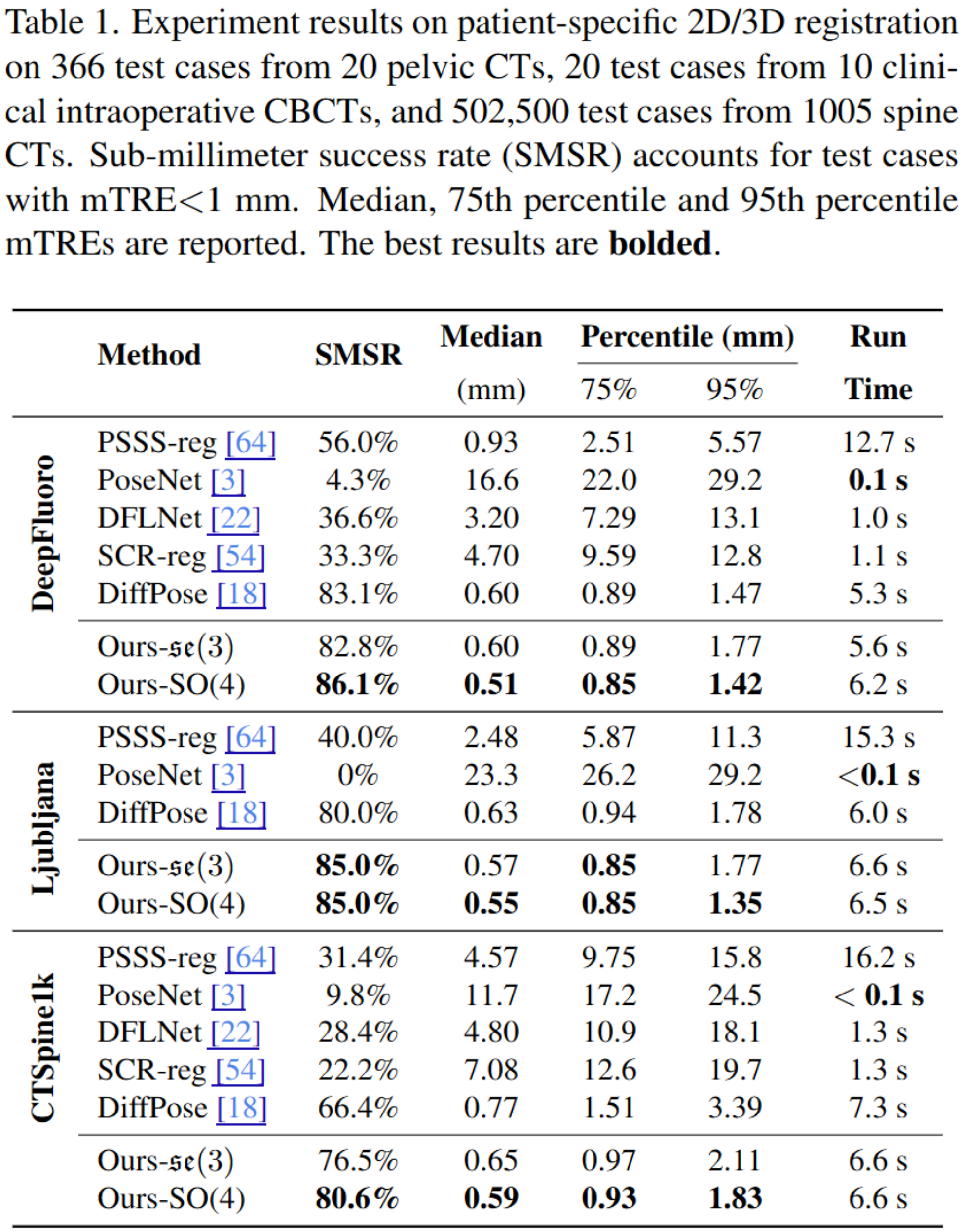


Table 1. Experiment results on patient-specific 2D/3D registration on 366 test cases from 20 pelvic CTs, 20 test cases from 10 clinical intraoperative CBCTs, and 502,500 test cases from 1005 spine CTs. Sub-millimeter success rate (SMSR) accounts for test cases with mTRE<1 mm. Median, 75th percentile and 95th percentile mTREs are reported. The best results are **bolded**.

	Method	SMSR	Median	Percentile (mm)		Run
			(mm)	75%	95%	Time
DeepFluoro	PSSS-reg [64]	56.0%	0.93	2.51	5.57	12.7 s
	PoseNet [3]	4.3%	16.6	22.0	29.2	**0.1 s**
	DFLNet [22]	36.6%	3.20	7.29	13.1	1.0 s
	SCR-reg [54]	33.3%	4.70	9.59	12.8	1.1 s
	DiffPose [18]	83.1%	0.60	0.89	1.47	5.3 s
	Ours-$\mathfrak{se}(3)$	82.8%	0.60	0.89	1.77	5.6 s
	Ours-SO(4)	**86.1%**	**0.51**	**0.85**	**1.42**	6.2 s
Ljubljana	PSSS-reg [64]	40.0%	2.48	5.87	11.3	15.3 s
	PoseNet [3]	0%	23.3	26.2	29.2	<**0.1 s**
	DiffPose [18]	80.0%	0.63	0.94	1.78	6.0 s
	Ours-$\mathfrak{se}(3)$	**85.0%**	0.57	**0.85**	1.77	6.6 s
	Ours-SO(4)	**85.0%**	**0.55**	**0.85**	**1.35**	6.5 s
CTSpine1k	PSSS-reg [64]	31.4%	4.57	9.75	15.8	16.2 s
	PoseNet [3]	9.8%	11.7	17.2	24.5	< **0.1 s**
	DFLNet [22]	28.4%	4.80	10.9	18.1	1.3 s
	SCR-reg [54]	22.2%	7.08	12.6	19.7	1.3 s
	DiffPose [18]	66.4%	0.77	1.51	3.39	7.3 s
	Ours-$\mathfrak{se}(3)$	76.5%	0.65	0.97	2.11	6.6 s
	Ours-SO(4)	**80.6%**	**0.59**	**0.93**	**1.83**	6.6 s

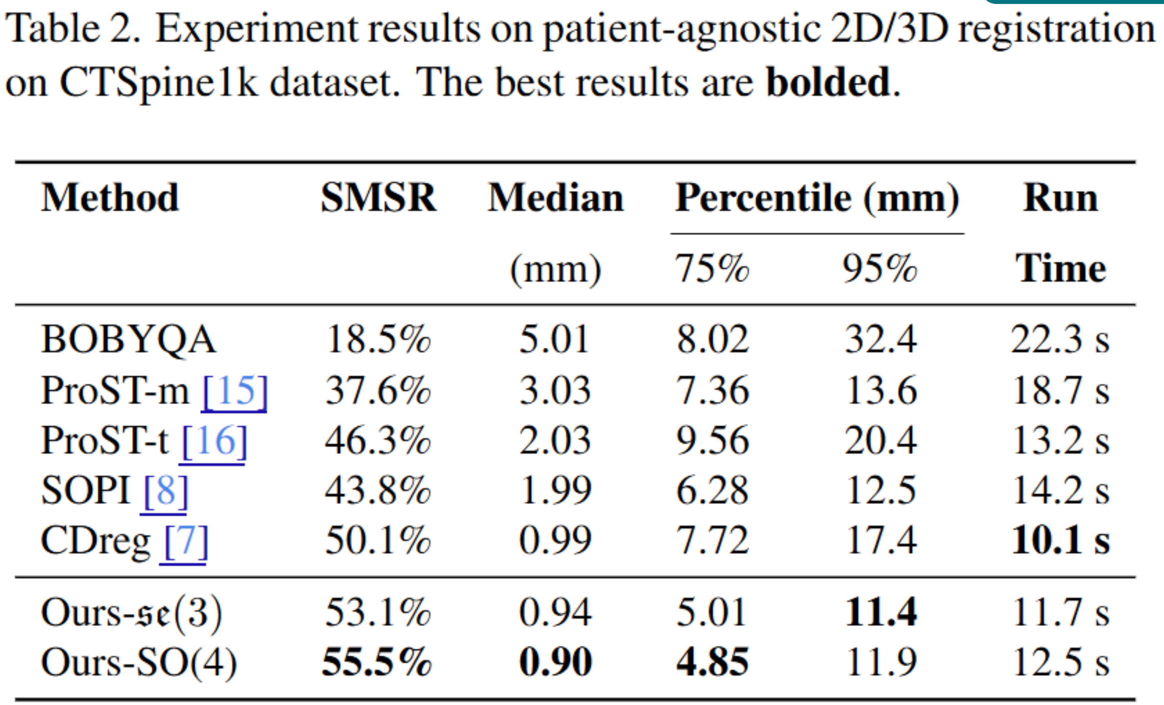


Table 2. Experiment results on patient-agnostic 2D/3D registration on CTSpine1k dataset. The best results are **bolded**.

Method	SMSR	Median	Percentile (mm)		Run
		(mm)	75%	95%	Time
BOBYQA	18.5%	5.01	8.02	32.4	22.3 s
ProST-m [15]	37.6%	3.03	7.36	13.6	18.7 s
ProST-t [16]	46.3%	2.03	9.56	20.4	13.2 s
SOPI [8]	43.8%	1.99	6.28	12.5	14.2 s
CDreg [7]	50.1%	0.99	7.72	17.4	**10.1 s**
Ours-$\mathfrak{se}(3)$	53.1%	0.94	5.01	**11.4**	11.7 s
Ours-SO(4)	**55.5%**	**0.90**	**4.85**	11.9	12.5 s

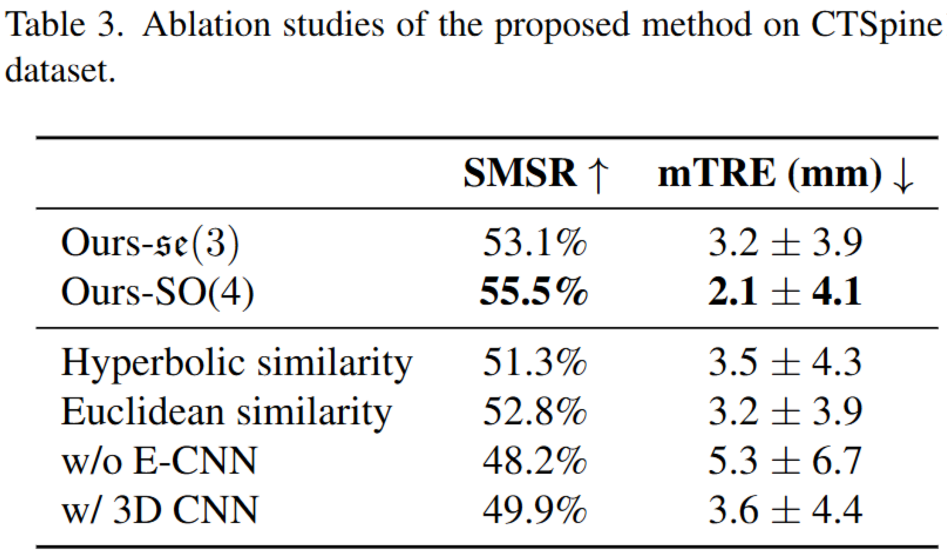
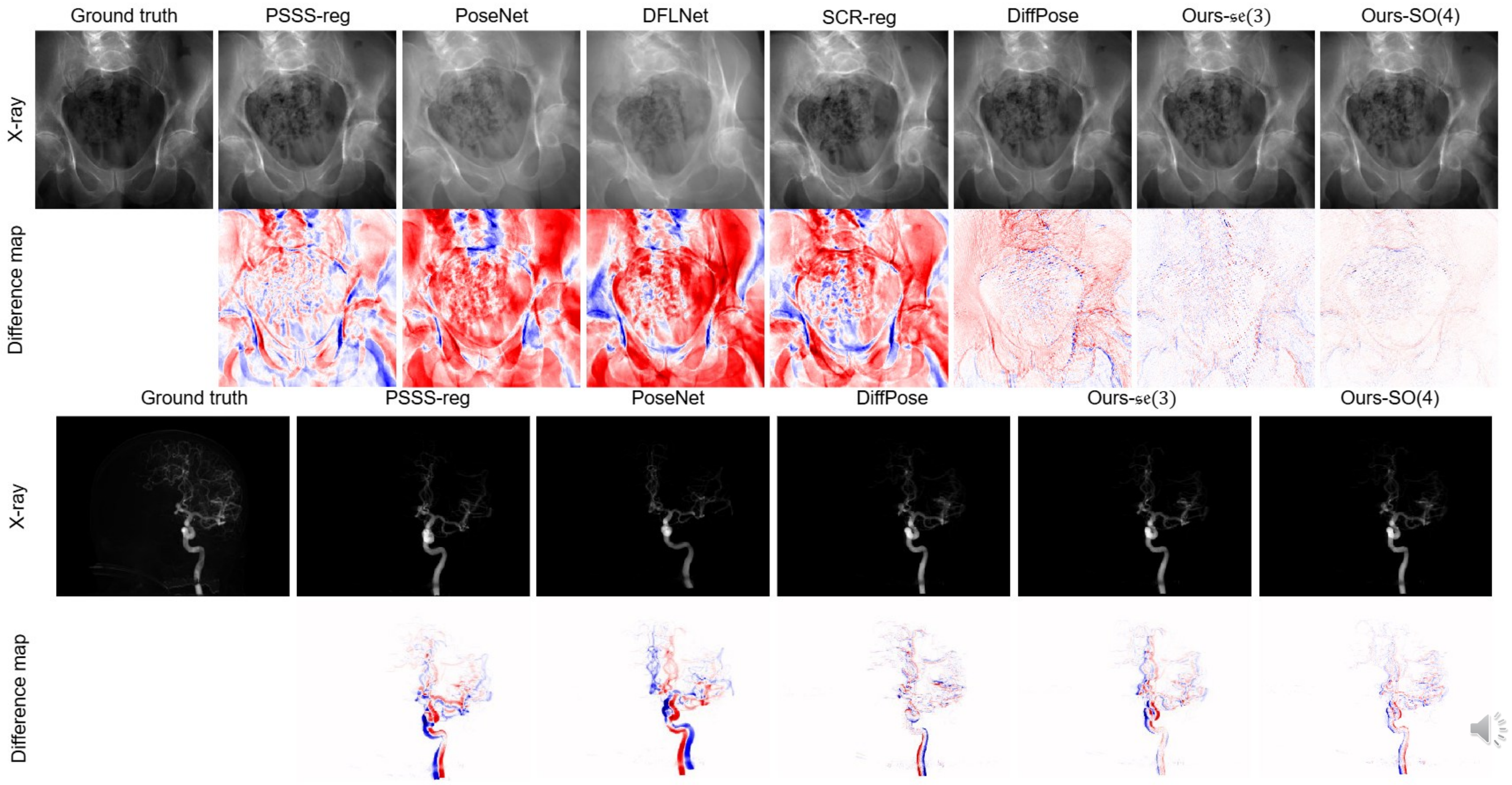


Table 3. Ablation studies of the proposed method on CTSpine1k dataset.

	SMSR ↑	mTRE (mm) ↓
Ours-$\mathfrak{se}(3)$	53.1%	3.2 ± 3.9
Ours-SO(4)	**55.5%**	**2.1** ± **4.1**
Hyperbolic similarity	51.3%	3.5 ± 4.3
Euclidean similarity	52.8%	3.2 ± 3.9
w/o E-CNN	48.2%	5.3 ± 6.7
w/ 3D CNN	49.9%	3.6 ± 4.4

Results


Ground truth
PSSS-reg
PoseNet
DFLNet
SCR-reg
DiffPose
Ours-$\mathfrak{se}(3)$
Ours-SO(4)
X-ray
Difference map
Ground truth
PSSS-reg
PoseNet
DiffPose
Ours-$\mathfrak{se}(3)$
Ours-SO(4)
X-ray
Difference map

Results

WACV 2026

Tucson, Arizona • Mar 6 - 10, 2026

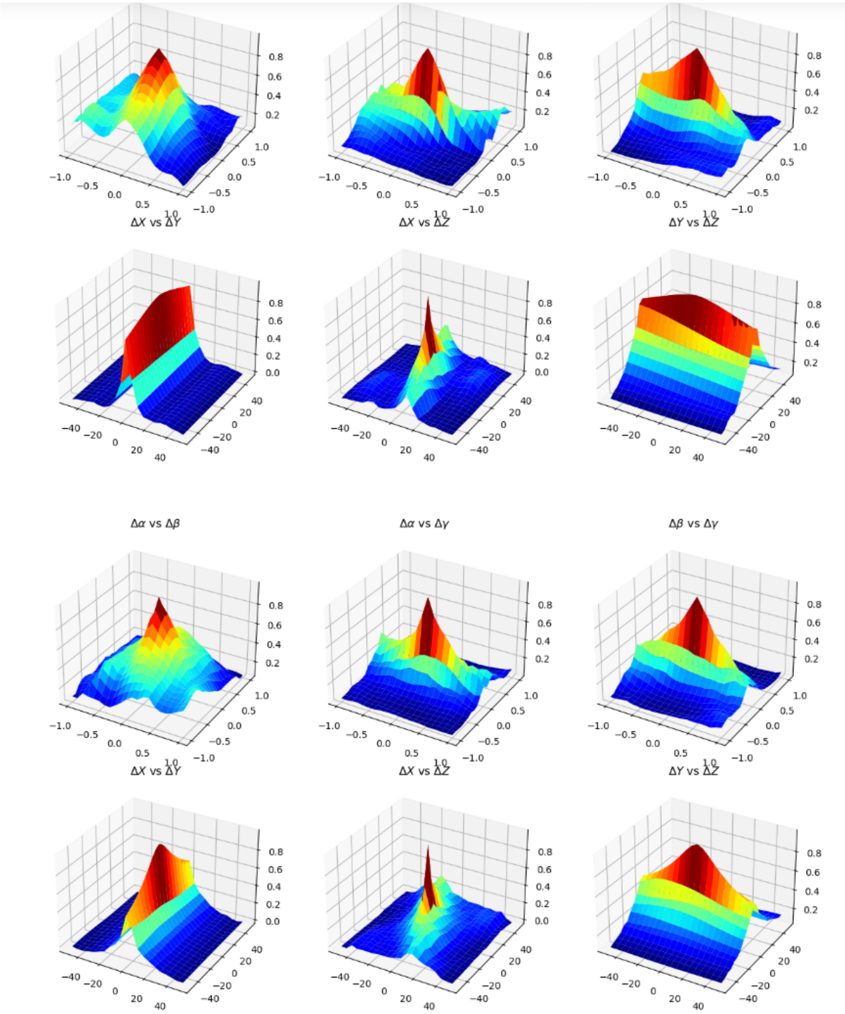


Figure 3. Visual comparison of the proposed spherical deep similarity landscape in $\mathfrak{se}(3)$ (top) and SO(4) (bottom). For clearer visualization, the deep similarity values are first normalized to the range [0,1], and then transformed by computing 1-ϵ, effectively inverting the scale to enhance contrast in the display.

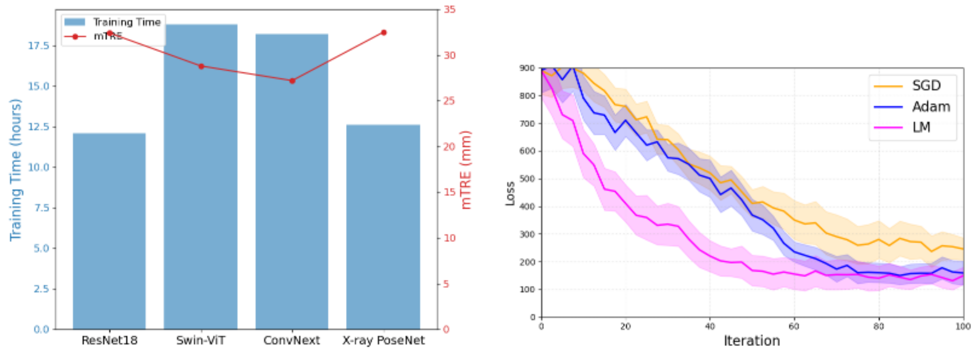


Figure 4. **Left:** comparison of training time and mTRE for position regressors with different backbone architectures.The training time reports the time when the standard deviation of the model's loss function is less than 10e-4 in the last ten epochs. **Right:** comparison of the convergence speed of the proposed framework using different gradient-based optimization methods in the inference phase.

Intraoperative 2D/3D Registration via Spherical Similarity Learning and Differentiable Levenberg-Marquardt Optimization

Thank You for Listening