Weakly-Supervised 2D/3D Image Registration via Differentiable X-ray Rendering and ROI Segmentation

Yuxin Cui[1], Max Q.-H. Meng[2,3], and Zhe Min[4,5](✉)

[1] School of Information Science and Engineering, Shandong University, Qingdao, China
[2] Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China
[3] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China
[4] School of Control Science and Engineering, Shandong University, Jinan, China
minzhe@sdu.edu.cn
[5] UCL Hawkes Institute, and Department of Medical Physics and Biomedical Engineering, University College London, London, UK

Abstract. Accurate registration between intraoperative 2D images and preoperative 3D anatomical structures is a prerequisite for image-guided minimally invasive surgery. Existing approaches for 2D/3D rigid registration, particularly those for X-ray to CT image registration, primarily rely on grayscale-based image similarity metrics. However, such metrics often fail to capture the optimal projection transformation due to their limited contextual information. To address this issue, we propose a novel and intuitive correspondence representation: the overlap of multiple corresponding Regions of Interest (ROIs). By introducing the differentiable Dice coefficient computed on the projection image, we establish a direct link between segmentation and registration within our weakly supervised 2D/3D registration framework. This framework comprises two stagesa learning-based preoperative stage and an optimization-based intraoperative stageboth of which leverage the ROI-based Dice score as a differentiable supervision signal. Additionally, we integrate automatic segmentation methods (e.g., UNet) and prompt-based methods (e.g., MedSAM) into the framework to investigate the impact of different segmentation approaches on registration performances. Furthermore, we validate the generalization ability of the proposed framework by integrating the ROI-based similarity with various similarity measures. Extensive experiments conducted on the DeepFluoro dataset yielded an mTRE of 0.67±1.34 mm, with rotational and translational error values being 0.2±0.5° and 1.6±2.9 mm respectively, outperforming existing state-of-the-art methods. The codes are available at https://github.com/CYXYZ/WSReg.

Keywords: 2D/3D Registration · Segmentation-assisted Registration · Surgical Navigation.

1 Introduction

Minimally invasive surgeries, such as femoral osteoplasty [9], hip replacement [4], spinal needle injection [11], and vascular interventions [20, 16], require precise localization of target structures intra-operatively. While 2D fluoroscopic X-ray images enable rapid acquisitions, their projection nature leads to spatial information loss [6]. Combining them with pre-operative 3D computed tomography (CT) scans allows for accurate localization of surgical instruments and interested anatomical regions [32, 21, 30]. Accurate 2D/3D image registration thus plays an important role in enabling successful surgical navigation [22, 23].

Traditional registration often relies on digitally reconstructed radiographs (DRRs), which generates high-quality synthetic X-rays from given CT scans and sampled rigid transformations [1, 28, 13, 27]. Efficient DRR generators concentrate on optimization-based search strategies [13, 10], and new similarity metrics, such as weighted local mutual information [20] and contour image force summation [25]. However, these metrics primarily rely on image intensity, providing limited information. Learning-based methods have been proposed, among which self-supervised approaches directly regress spatial mappings to predict projection parameters of perspective images, followed by an optimization process to obtain the optimal solution [31, 12]. However, training pose regressors typically requires a large amount of synthetic X-ray image data. Other supervised methods rely on expert annotations and employ Perspective-n-Point (PnP) algorithms to solve the mapping relationship [3, 4, 7]. Although these supervised methods can automatically establish correspondences between landmarks and often outperforms unsupervised methods in terms of accuracy [29, 12], it still demands manual landmark annotations or incurs substantial computational costs. Moreover, limited landmark availability or inaccurate annotations may lead to failures in pose estimation [26, 15].

To alleviate this challenge, inspired by those 3D/3D registration counterparts [8], we explore to develop an accurate weakly supervised registration approach. It is well noted that few 2D/3D registration studies leverage segmentation information. In some scenarios, segmentation of region-of-interest (ROI) provides sufficient representational capacity, similar to other correspondence representations [17, 18]. Unlike 3D/3D registration, where spatial transformations can be directly applied to masks or images for alignment, direct correspondence between 3D and 2D data via geometric transformations is not feasible, making operations on fixed segmentation data unworkable. Furthermore, even considering the DRR process as a mapping between 2D and 3D, the non-differentiable Dice coefficient prevents gradient-based updates towards the optimal solutions. To address this issue, we propose an alternative representation using paired ROIs between the projection images to denote correspondences. Training with predefined ROI types integrates ROI pairs with known correspondences into the registration algorithm, providing weak supervision for the task.

In this study, we propose a novel weakly-supervised 2D/3D medical image registration framework that leverages ROI segmentation information from both 2D input and rendered projected images. We incorporate the differentiable Dice

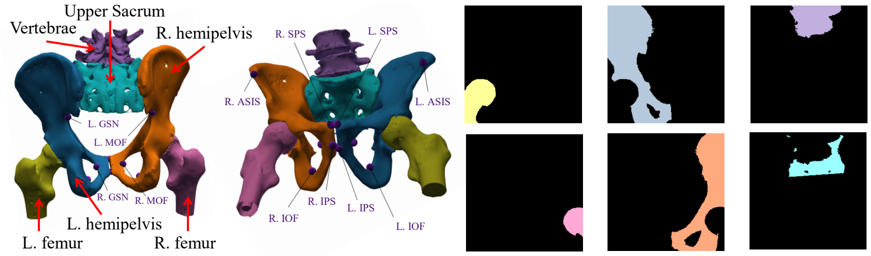

Fig. 1: **Anatomical Segmentations and Landmarks of Human Hip Bone.** The left panel shows the 3D landmarks(in Dark Purple) and 3D mask; the right panel shows the corresponding 2D mask.

coefficient as a correspondence representation of the registration within both learning-based and optimization-based methods. Furthermore, we investigate the impact of segmentation models, such as MedSAM [19] and U-Net [15], on registration performances, analyzing their respective strengths and limitations in the framework. We combined commonly used similarity measures in image registration with ROI scores to validate the generalization capability of the proposed method. The proposed method achieved an mTRE of 0.67±1.34 mm, with rotational and translational errors of 0.2±0.5° and 1.6±2.9 mm in extensive experiments on DeepFluoro [15], confirming its effectiveness and clinical potential.

Our contributions of this paper are summarised as follows. **1)** We have established a weakly supervised 2D/3D registration framework, which consists of two stages including the preoperative training and the intraoperative optimization. The effectiveness of weak supervision signal of segmentation was demonstrated in both stages. **2)** We have proposed a differentiable search method for optimal projection parameters based on ROI scores by establishing correspondences between ROI pairs in different perspective images, acquired with automatic segmentation methods (e.g., U-Net) and prompt-based methods (e.g., MedSAM). **3)** We have introduced ROI scores into different similarity measures, demonstrating the strong generalization capability of the weak supervision approach.

2 Method

2.1 Weakly-Supervised 2D/3D Image Registration

ROI Segmentation in X-ray Images. To identify corresponding ROI pairs, we utilized the DeepFluoro dataset[6] and applied leave-one-out cross-validation to train X-ray segmentation models (U-Net [15] and MedSAM [19]). The hip skeletal structure is divided into six sub-regions, as shown in Fig. 1, with corresponding X-ray images annotated for each.

[6] https://github.com/rg2/DeepFluoroLabeling-IPCAI2020

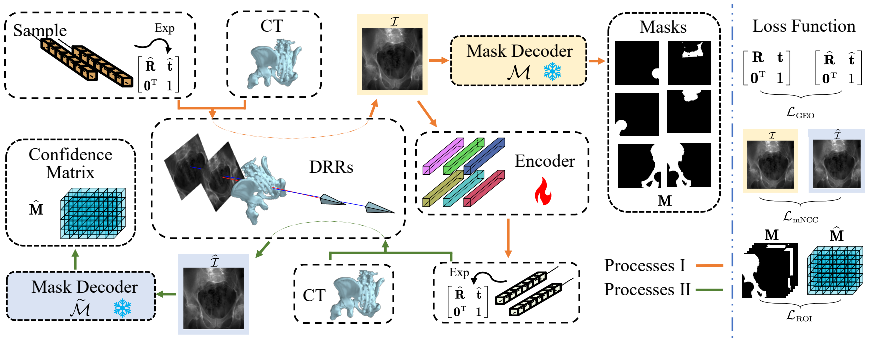

Fig. 2: **Preoperative Training Framework.** The Orange arrows and Teal arrows represent the processes using sampled and estimated poses, respectively. The right panel depicts the input used for computing the loss function.

Pose-Sampled X-ray Synthesis. We use the rapid DRR generator in [13] to synthesize X-ray images through vectorized tensor operations. For each patients hip, multiple rotation vectors $\mathbf{r} \in \mathbb{R}^3$ and translation vectors $\mathbf{t} \in \mathbb{R}^3$ are sampled in the posterior-anterior (PA) view, generating projection perturbations via the Lie algebra exponential map $\text{Exp}([\mathbf{r},\ \mathbf{t}]) \to \Delta\mathbf{T} \in \text{SE}(3)$. Using preoperative imaging, the PA view transformation $\mathbf{T}_{\text{PA}} \in \text{SE}(3)$ is pre-established, and the perturbation-induced pose is given by $\mathbf{T} = \Delta\mathbf{T} \cdot \mathbf{T}_{\text{PA}}$ with $\mathbf{T} \in \text{SE}(3)$. The DRR-generated dataset serves as training data for the registration framework, where only the perturbation needs to be estimated. The DRR process is formally defined as $\mathcal{I} = \mathcal{R}(\mathbf{V}, \mathbf{T})$, where $\mathcal{I} : \mathbb{R}^2 \to \mathbb{R}$ is the rendered X-ray image, $\mathbf{V} : \mathbb{R}^3 \to \mathbb{R}$ is the patients CT volume, and $\mathcal{R}(\cdot)$ denotes the rendering operator[7].

Segmentation-Assisted Rigid Transformation Estimation. Fig. 2 illustrates the proposed U-Net-based training process. Given $\mathbf{V}$, pose sampling is performed to obtain a set of poses $\mathbf{T}$ and rendered images $\mathcal{I}$. The goal is to train an encoder $\mathcal{E}$ that takes an X-ray (either real or synthetic) as input and outputs the projection transformation parameters. To integrate segmentation information, $\mathcal{I}$ inputs $\mathcal{E}$ and the mask decoder $\mathcal{M}$, which generate a six-dimensional vector $[\widehat{\mathbf{r}},\ \widehat{\mathbf{t}}] \in \mathbb{R}^6$ and segmentation mask $\mathbf{M} \in \{0,1\}^{m \times n \times 6}$ respectively, where $m \in \mathbb{N}^+$ and $n \in \mathbb{N}^+$ represent the mask dimensions. $[\widehat{\mathbf{r}},\ \widehat{\mathbf{t}}]$ is then mapped through the exponential map $\text{Exp}([\widehat{\mathbf{r}},\ \widehat{\mathbf{t}}]) \to \Delta\widehat{\mathbf{T}} \in \text{SE}(3)$ to compute the estimated perturbation of the projection transformation $\widehat{\mathbf{T}} \in \text{SE}(3)$ as $\widehat{\mathbf{T}} = \Delta\widehat{\mathbf{T}} \cdot \mathbf{T}_{\text{PA}}$. $\widehat{\mathbf{T}}$ is then used in the rendering operator to obtain rendered projective image $\widehat{\mathcal{I}} = \mathcal{R}(\mathbf{V}, \widehat{\mathbf{T}})$. The other mask decoder $\widetilde{\mathcal{M}}$ receives $\widehat{\mathcal{I}}$ and outputs a confidence matrix $\widetilde{\mathbf{M}} \in [0,1]^{m \times n \times 6}$ denotes a normalized probability map, where each element lies in the range $[0,1]$. The encoder $\mathcal{E}$ is trained using a combination of $\mathcal{L}_{\text{GEO}}$ between $\mathbf{T}$ and $\widehat{\mathbf{T}}$, $\mathcal{L}_{\text{mNCC}}$ between $\mathcal{I}$ and $\widehat{\mathcal{I}}$, and $\mathcal{L}_{\text{ROI}}$ between $\mathbf{M}$ and $\widetilde{\mathbf{M}}$.

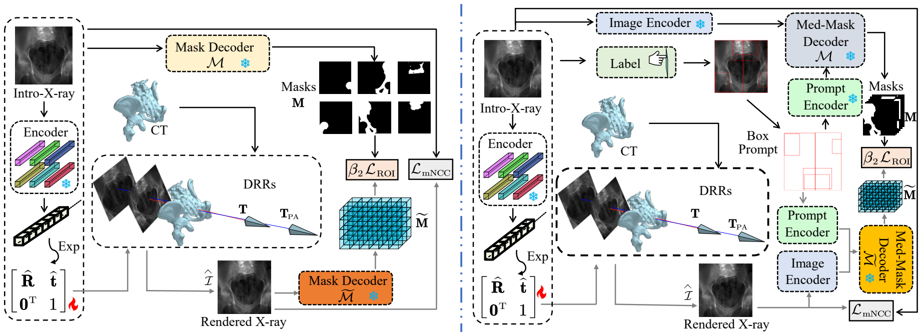
Intro-X-ray
Encoder
Exp
$\widehat{\mathbf{R}}$ $\widehat{\mathbf{t}}$
$\mathbf{0}^\top$ 1
Mask Decoder $\mathcal{M}$
CT
DRRs
$\mathbf{T}$
$\mathbf{T}_{\text{PA}}$
Masks
$\mathbf{M}$
$\beta_2\mathcal{L}_{\text{ROI}}$
$\mathcal{L}_{\text{mNCC}}$
$\widetilde{\mathbf{M}}$
$\widehat{\mathcal{I}}$
Mask Decoder $\widetilde{\mathcal{M}}$
Rendered X-ray
Image Encoder
Med-Mask Decoder $\mathcal{M}$
Label
Prompt Encoder
Box Prompt
Prompt Encoder
Image Encoder
Med-Mask Decoder $\widetilde{\mathcal{M}}$

Fig. 3: **Intraoperative Registration Framework.** The left and right sides show the registration frameworks based on U-Net and MedSAM. The Black and Grey arrows indicate constant inputs and parameter updates during the process.

2.2 Segmentation-Assisted Intraoperative Registration

During surgery, the trained encoder $\mathcal{E}$ initializes the estimation of real X-ray projection transformation matrix. The intraoperative registration framework, shown in Fig. 3, employs two segmentation-assisted schemes, i.e., U-Net-based and MedSAM-based. In the UNet-based method (left side of Fig. 3), the intraoperative X-ray image $\mathbf{I} : \mathbb{R}^2 \to \mathbb{R}$ serves as the fixed image, while the rendered image from the estimated pose serves as the moving image. $\mathbf{I}$ is passed through two branches: 1) the Mask Decoder $\mathcal{M}$ to generate $\mathbf{M}$, and 2) the encoder $\mathcal{E}$ and exponential transformation Exp to obtain the estimated pose $\widehat{\mathbf{T}}$. The rendered image $\widehat{\mathcal{I}} = \mathcal{R}(\mathbf{V}, \widehat{\mathbf{T}})$ is then processed by the Mask Encoder $\widetilde{\mathcal{M}}$ to obtain $\widetilde{\mathbf{M}}$. Pose parameters are updated by maximizing $\mathcal{L}_{\text{mNCC}}$ between $\mathbf{I}$ and $\widehat{\mathcal{I}}$, and $\mathcal{L}_{\text{ROI}}$ between $\mathbf{M}$ and $\widetilde{\mathbf{M}}$. In the MedSAM-based approach, the only difference with the UNet-based counterpart is the addition of a bounding box prompt for segmenting $\mathbf{I}$ and $\widehat{\mathcal{I}}$. Crucially, the trained encoder often brings the initial pose estimation close to a suboptimal pose, making the use of a fixed Bounding Box as a prompt for both $\mathbf{I}$ and $\widetilde{\mathcal{I}}$ during optimization a reasonable approach.

2.3 Loss Function

As shown in Fig. 2, the overall loss function $\mathcal{L}$ consists of $\mathcal{L}_{\text{GEO}}$, $\mathcal{L}_{\text{ROI}}$, and $\mathcal{L}_{\text{NCC}}$. $\mathcal{L}_{\text{GEO}}$ is defined as the geodesic distance between $\mathbf{T}$ and $\widehat{\mathbf{T}}$, i.e., $\mathcal{L}_{\text{GEO}} = \frac{f}{2}\sqrt{\|\text{Log}\big(\mathbf{R}^\top\widehat{\mathbf{R}}\big)\|^2 + \|(\mathbf{t}-\widehat{\mathbf{t}})\|^2 + \|\text{Log}\big(\mathbf{T}^{-1}\widehat{\mathbf{T}}\big)\|}$, where $f \in \mathbb{R}^+$ represents the camera focal length, and $|| \cdot ||$ indicates the L1 norm of a vector. For the X-ray image $\mathbf{I}$ (either real or simulated) under a given camera pose, and the rendered X-ray image $\widehat{\mathcal{I}}$ generated using the predicted pose, $\mathcal{L}_{\text{mNCC}}$ represents the mean of the local and global NCC loss between $\mathbf{I}$ and $\widehat{\mathcal{I}}$: $\mathcal{L}_{\text{mNCC}} = \frac{1}{2K}\sum_{k=1}^{K}\rho(\mathbf{I}_{W_k}) \cdot \rho(\widehat{\mathcal{I}}_{W_k}) + \rho(\mathbf{I}) \cdot \rho(\widehat{\mathcal{I}})$, where $W_k : \mathbb{R}^2 \to \mathbb{R}$ denotes the k-th sliding window of $\mathbf{I}$ and $\widehat{\mathcal{I}}$, and $K \in \mathbb{N}^+$ is the number of windows into which the image is divided.

Table 1: Comparison with advanced medical image registration methods.

Baseline	mTRE(mm)	SRR	RE(°)	TE(mm)
DiffDRR [13]	25.08 ± 24.01	26%	10.8 ± 9.6	72.5 ± 63.2
PSSS [31]	4.40 ± 4.63	26%	1.8± 2.0	10.1± 10.0
DiffPose [12]	1.01 ± 3.24	83%	0.4 ± 1.3	2.1± 5.8
Ours	**0.67 ± 1.34** ↑	**87%** ↑	**0.2 ± 0.5** ↑	**1.6 ± 2.9** ↑

Table 2: Comparison of dice scores for different segmentation methods on real and synthetic data.

X-rays	U-Net	Med-SAM
Real	0.81 ± 0.07	0.85 ± 0.07
Synth	0.80 ± 0.07	0.85 ± 0.06

In general, for any image $\mathbf{A} : \mathbb{R}^2 \rightarrow \mathbb{R}$, $\rho(\mathbf{A}) = (\mathbf{A} - \mu(\mathbf{A}))/\sigma(\mathbf{A})$ represents the normalized cross-correlation of the image, where $\mu(\mathbf{A}) = \frac{1}{N}\sum_{(x,y)} \mathbf{A}(x, y)$ and $\sigma(\mathbf{A}) = \sqrt{\frac{1}{N}\sum_{(x,y)} \left(\mathbf{A}(x, y) - \mu(\mathbf{A})\right)^2}$ are the mean and standard deviation of the image's pixel values, respectively. Here, $x \in \mathbb{R}$ and $y \in \mathbb{R}$ denote the pixel coordinates, $N \in \mathbb{N}^+$ represents the number of pixels. Finally, the ROI loss $\mathcal{L}_{\text{ROI}}$ using the Dice score of the image is defined as $= \frac{1}{P}\sum_{p=1}^{N} \frac{2\sum\limits_{x,y} \mathbf{M}_p(x,y) \cdot \widetilde{\mathbf{M}}_n(x,y)}{\sum\limits_{x,y}\left(\mathbf{M}_p(x,y) + \widetilde{\mathbf{M}}_p(x,y)\right)}$, where $\mathbf{M}_p \in \mathbb{R}^{2\times 2}$ represents the one-hot encoding of the p-th channel of $\mathbf{M}$, and $\widetilde{\mathbf{M}}_p \in \mathbb{R}^2$ is the predicted probability map of the p-th ROI extracted from $\widetilde{\mathbf{M}}$, $P \in \mathbb{N}^+$ represents the number of ROI categories. These three loss components together form the training loss function: $\mathcal{L} = \beta_1\mathcal{L}_{\text{GEO}} + \beta_2(1 - \mathcal{L}_{\text{ROI}}) + 1 - \mathcal{L}_{\text{mNCC}}$, where β_1, β_2 are hyperparameters. For the optimization process in Sec. 2.2, the objective function to be minimized is: $\mathcal{L}_{\text{mNCC}} + \beta_2\mathcal{L}_{\text{ROI}}$.

3 Experiments and Results

3.1 Datasets, Implementation Details and Evaluation Metrics

Datasets. We evaluate our method on the DeepFluoro dataset, which contains pelvic CT scans and X-ray images from six cadavers (three males and three females, aged 5794 years) [15]. Each subject undergoes one CT scan with 24 to 111 X-ray fluoroscopy images, resulting in a total of six CT scans and 366 X-ray images. The dataset provides intrinsic and extrinsic parameters for each X-ray system, with annotated landmarks and masks in the CT scans (cf. Fig. 1).
Implementation Details. For $\mathcal{E}$, we use a ResNet18 backbone combined with a two-layer MLP [12], and train it with the Adam optimizer at a learning rate of 10^{-3} for 1300 epochs. The Mask Encoder (cf. Fig. 2 and Fig. 3) follows a six-level U-Net design [24], with 2×2 convolution for downsampling and transposed convolution for upsampling, and is trained with the Adam optimizer at 10^{-1} learning rate for 500 epochs. The Image Encoder, Prompt Encoder, and Med-Mask Decoder in Fig. 3 follow the MedSAM design [19]. The Box Prompt is simulated by adding up to 10 pixels of random translation noise to mimic manual annotation errors. These are trained using the AdamW optimizer at 10^{-3} for 400 epochs. Intraoperative registration (cf. Sec. 2.2) $\mathbf{r}$ and $\mathbf{t}$ is optimised using the Adam optimizer at learning rates of 7.5×10^{-3} and 7.5, respectively, over

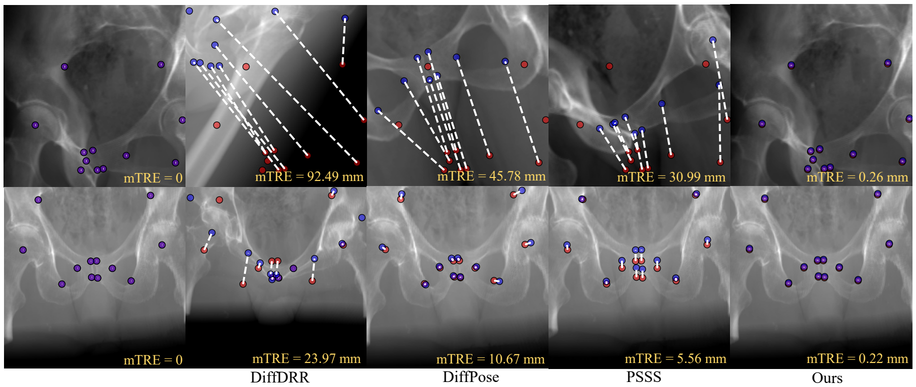

Fig. 4: **Comparison with several baseline methods.** The Red and Blue markers represent the 2D projections of 3D landmarks under the estimated and ground truth poses, respectively. The Purple area indicates the overlap region, whose larger area indicates better registration accuracy.

Table 3: Comparison of the performance of different objective functions with ('+') and without ('-') $\mathcal{L}_{\text{ROI}}$.

	mTRE(mm)		**SRR**		**RE(°)**		**TE(mm)**	
	−	+	−	+	−	+	−	+
$\text{Local}_{\text{NCC}}$	2.83 ± 6.33	**2.15 ± 6.38**	72%	**78%**	1.1 ± 2.4	**0.7 ± 1.9**	6.9 ± 14.3	**5.4 ± 13.3**
$\text{Global}_{\text{NCC}}$	4.39 ± 4.71	**3.51 ± 3.68**	24%	**27%**	1.9 ± 2.0	**1.5 ± 1.5**	10.1 ± 10.1	**8.5 ± 9.0**
$\text{Gradient}_{\text{NCC}}$	11.12 ± 7.35	**9.95 ± 7.39**	4	**9%**	4.0 ± 2.2	**3.5 ± 2.3**	46.8 ± 25.6	**40.1 ± 28.2**
SSIM	13.62 ± 8.63	**5.55 ± 5.14**	2%	**10%**	5.6 ± 3.2	**2.1 ± 1.7**	48.7 ± 26.2	**14.7 ± 13.3**
MSE	9.84 ± 6.46	**9.23 ± 6.12**	0%	**1%**	4.2 ± 2.1	**3.9 ± 2.2**	31.1 ± 19.9	**29.7 ± 19.1**
MAE	10.38 ± 6.61	**9.85 ± 5.92**	1%	1%	4.3 ± 2.2	**4.1 ± 2.0**	34.8 ± 21.2	**32.3 ± 19.3**
PSNR	10.21 ± 6.50	**10.13 ± 6.75**	0%	0%	4.3 ± 2.3	**4.2 ± 2.2**	**33.9 ± 20.2**	33.9 ± 21.3

250 epochs. The hyperparameters appeared are $\beta_1 = 10^{-2}$ and $\beta_2 = 10^{-1}$. All experiments are conducted on an RTX 4090 Ti GPU.

Evaluation Metrics. Three evaluation metrics are utilised: (1) mean Target Registration Error (mTRE) defined as the average Euclidean distance between ground truth (GT) and estimated 3D landmark projections, (2) rotational/translational errors (RE/TE) defined as vector norms of the differences between GT and estimated rotation and translation vectors, and (3) Submillimeter Registration Success Rate (SRR) defined as the ratio of registration trials with mTRE $<$ 1mm [12, 5].

3.2 Experimental Results

Analysis of Real-to-sim Segmentation. As shown in Sec. 2.1, while the segmentation model is trained on real X-ray images, the registration model uses DRRs, making this a real-to-simulation task. A key question is whether the

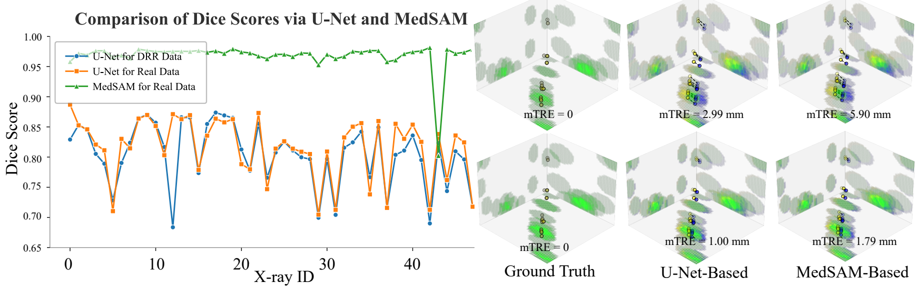

Fig. 5: **Segmentation Dice Scores (Left) and Registration Results (Right) for Patient 18-2799.** Blue and Yellow spheres denote landmarks under ground-truth and estimated poses. These are projected onto three orthogonal 2D planes where greater overlap in Green indicates better registration accuracy.

segmentation model can perform the real-to-sim X-ray image segmentation task effectively. Table 2 shows the Dice scores for segmentation on real and synthetic X-ray images using two segmentation methods, evaluated via leave-one-out cross-validation. The synthetic X-rays are rendered based on the real X-ray projection poses. Column-wise results indicate that the impact of using synthetic X-rays on ROI extraction performance is negligible.

Comparison with State-of-the-Art Methods. We perform a fair comparison with PSSS [31] and DiffPose [12], and an admittedly less fair comparison with the purely optimization-based method DiffDRR [13]. Both PSSS and DiffPose train a pose regression network to initialize pose parameters, followed by optimization-based fine-tuning. Our method outperforms all baselines across four evaluation metrics (cf. Table 1). Fig. 4 shows qualitative results of 3D landmark projections. Additionally, we compare two segmentation-assisted registration variants under the same pose initialization conditions. Although MedSAM performs better in ROI segmentation, it yields an mTRE of 0.95±3.01 mm, an SRR of 84%, an RE of 0.27±0.61°, and a TE of 3.15±1.70 mm across all data. In contrast, the U-Net-based method achieves better registration results, with an mTRE of 0.74±1.97 mm, an SRR of 85%, an RE of 0.25±0.50°, and a TE of 1.64±2.96 mm. This may be because MedSAM's box prompt for DRRs is based on real image input. While the poses of DRRs and the real image are similar, they are not identical, resulting in greater noise in the box prompt used for DRRs. In contrast, the U-Net-based method avoids this issue. To validate the broad applicability of the segmentation-assisted registration paradigm, we performed ablation experiments using commonly used similarity measures for image registration [2, 12, 14]. Table 3 confirms the effectiveness of incorporating this auxiliary information into the registration procedure.

A Case Study of 2D/3D Registration of the Pelvis. We selected the patient ID 18-2799 to evaluate the impact of segmentation on the registration. Images were segmented using U-Net and MedSAM. MedSAM outperforms U-

Net in segmentation (cf. Fig. 5). Based on these, we applied both MedSAM- and U-Net-based segmentations to the real input images, while using U-Net-based segmentation for the rendered images during optimization. The registration outcomes, shown on the right of Fig.5, revealed no improvement with MedSAM, despite its superior segmentation. This suggests that segmentation-assisted registration depends not only on segmentation accuracy but also on consistency.

4 Conclusion

We propose a novel weakly supervised 2D/3D registration framework leveraging segmentation from 2D projections. By incorporating a differentiable dice score into both learning-based and optimization-based pipelines, our method achieves accurate and generalizable registration. Furthermore, we demonstrate that segmentation models provide effective supervision and that combining ROI-based and conventional similarity metrics enhances registration accuracy. Experiments on the DeepFluoro dataset confirm our methods competitive, state-of-the-art performance and offer insights into 2D/3D registration.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62303275. This work was also supported by the National Natural Science Fund for Excellent Young Scientists Fund Program (Overseas) under Grant 22IAA01849, Jinan Science and Technology Bureau under Grant 202333011.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Aouadi, S., Sarry, L.: Accurate and precise 2d3d registration based on x-ray intensity. Computer Vision and Image Understanding **110**(1), 134–151 (2008)
2. Avants, B.B., Tustison, N.J., Song, G., et al.: A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage **54**(3), 2033–2044 (2011)
3. Bier, B., Unberath, M., Zaech, J.N., et al.: X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 55–63. Springer International Publishing, Cham (2018)
4. Bradley, M.P., Benson, J.R., Muir, J.M.: Accuracy of acetabular component positioning using computer-assisted navigation in direct anterior total hip arthroplasty. Cureus **11**(4) (2019)
5. Brock, K.K., Mutic, S., McNutt, T.R., Li, H.H., Kessler, M.L., Dong, L., Xia, P., Weiss, E., Meyer, J.L., Schreibmann, E., et al.: 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). https://doi.org/10.1002/mp.12256

6. Cui, Y., Song, R., Li, Y., et al.: Multi-view 2d/3d image registration via differentiable x-ray rendering. Procedia Computer Science **250**, 282–288 (2024)
7. Cui, Y., Song, R., Li, Y., Meng, M.Q.H., Min, Z.: Robust and accurate multi-view 2d/3d image registration with differentiable x-ray rendering and dual cross-view constraints (2025), https://arxiv.org/abs/2506.22191
8. Fu, Y., Brown, N.M., Saeed, S.U., et al.: Deepreg: a deep learning toolkit for medical image registration. Journal of Open Source Software **5**(55), 2705 (2020)
9. Gao, C., Farvardin, A., Grupp, R.B., et al.: Fiducial-free 2d/3d registration for robot-assisted femoroplasty. IEEE Transactions on Medical Robotics and Bionics **2**(3), 437–446 (2020)
10. Gao, C., Liu, X., Gu, W., et al.: Generalizing spatial transformers to projective geometry with applications to 2d/3d registration. In: Medical Image Computing and Computer Assisted InterventionMICCAI 2020: 23rd International Conference, Lima, Peru, October 48, 2020, Proceedings, Part III 23. pp. 329–339. Springer International Publishing, Cham (2020)
11. Gao, C., Phalen, H., Margalit, A., et al.: Fluoroscopy-guided robotic system for transforaminal lumbar epidural injections. IEEE Transactions on Medical Robotics and Bionics **4**(4), 901–909 (2022)
12. Gopalakrishnan, V., Dey, N., Golland, P.: Intraoperative 2d/3d image registration via differentiable x-ray rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11662–11672 (2024)
13. Gopalakrishnan, V., Golland, P.: Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In: Workshop on Clinical Image-Based Procedures. pp. 1–11. Springer Nature Switzerland, Cham (2022)
14. Grupp, R.B., Armand, M., Taylor, R.H.: Patch-based image similarity for intraoperative 2d/3d pelvis registration during periacetabular osteotomy. In: International Workshop on Computer-Assisted and Robotic Endoscopy. pp. 153–163 (2018)
15. Grupp, R.B., Unberath, M., Gao, C., et al.: 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)
16. Guan, S., Wang, T., Sun, K., et al.: Transfer learning for nonrigid 2d/3d cardiovascular images registration. IEEE Journal of Biomedical and Health Informatics **25**(9), 3300–3309 (2020)
17. Huang, S., Xu, T., Shen, Z., et al.: One registration is worth two segmentations. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 665–675 (2024)
18. Huang, S., Xu, T., Shen, Z., et al.: Samreg: Sam-enabled image registration with roi-based correspondence. arXiv preprint arXiv:2410.14083 (2024)
19. Ma, J., He, Y., Li, F., et al.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
20. Meng, C., Wang, Q., Guan, S., et al.: 2d-3d registration with weighted local mutual information in vascular interventions. IEEE Access **7**, 162629–162638 (2019)
21. Mi, J., Yin, W., Zhao, L., et al.: Sgreg: segmentation guided 3d/2d rigid registration for orthogonal x-ray and ct images in spine surgery navigation. Physics in Medicine & Biology **68**(13), 135004 (2023)
22. Min, Z., Lai, J., Ren, H.: Innovating robot-assisted surgery through large vision models. Nature Reviews Electrical Engineering pp. 1–14 (2025)
23. Min, Z., Zhang, A., Zhang, Z., et al.: 3-d rigid point set registration for computer-assisted orthopedic surgery (caos): A review from the algorithmic perspective. IEEE Transactions on Medical Robotics and Bionics **5**(2), 156–169 (2023)

24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted InterventionMICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III. pp. 234–241. Springer International Publishing, Cham (2015)
25. Schmid, J., Chênes, C.: Segmentation of x-ray images by 3d-2d registration based on multibody physics. In: Asian Conference on Computer Vision. pp. 674–687. Springer International Publishing, Cham (2014)
26. Shrestha, P., Xie, C., Shishido, H., et al.: X-ray to ct rigid registration using scene coordinate regression. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 781–790. Springer Nature Switzerland, Cham (2023)
27. Unberath, M., Zaech, J.N., Lee, S.C., et al.: Deepdrra catalyst for machine learning in fluoroscopy-guided procedures. In: Medical Image Computing and Computer Assisted InterventionMICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11. pp. 98–106. Springer International Publishing, Cham (2018)
28. Van Der Bom, I.M.J., Klein, S., Staring, M., et al.: Evaluation of optimization methods for intensity-based 2d-3d registration in x-ray guided interventions. In: Medical Imaging 2011: Image Processing, SPIE. vol. 7962, pp. 657–671 (2011)
29. Walch, F., Hazirbas, C., Leal-Taixe, L., et al.: Image-based localization using lstms for structured feature correlation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 627–637 (2017)
30. Wang, J., Schaffert, R., Borsdorf, A., et al.: Dynamic 2-d/3-d rigid registration framework using point-to-plane correspondence model. IEEE Transactions on Medical Imaging **36**(9), 1939–1954 (2017)
31. Zhang, B., Faghihroohi, S., Azampour, M.F., et al.: A patient-specific self-supervised model for automatic x-ray/ct registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 515–524. Springer Nature Switzerland, Cham (2023)
32. Zhang, W., Zhao, L., Gou, H., et al.: Prscs-net: Progressive 3d/2d rigid registration network with the guidance of single-view cycle synthesis. Medical Image Analysis **97**, 103283 (2024)