# Intraoperative Fluoroscopy-Volume Registration Using Virtual-Supervision Skeleton Attention

Shurui Zhang[1,3], Ning Feng[1], Wenkang Fan[1], Hao Fang[1,2], and Xiongbiao Luo[1,2,3,*]

[1] Department of Computer Science and Engineering, Xiamen University, Xiamen 361102, China
[2] National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen 361102, China
[3] Discipline of Intelligent Instrument and Equipment, Xiamen University, Xiamen 361102, China
xiongbiao.luo@gmail.com, Asterisk indicates the corresponding author

**Abstract.** Endoscopic retrograde cholangiopancreatography with duodenoscopy and fluoroscopy is routinely performed to diagnose and treat biliary and pancreatic diseases. It still remains challenging to align 2D fluoroscopic images and 3D preoperative volumes for precisely navigating surgical tools and locating anatomical targets in the body. This work explores a new intraoperative fluoroscopy-volume (2D-3D) registration framework that combines virtual-supervision pose regression networks and pose optimization without manually annotating any 2D fluoroscopic images from 3D preoperative volumes. The regression networks with a dual-branch design, a skeleton-guided spatial attention mechanism, and an auxiliary bone mask decoder can successfully extract fluoroscopic image and its skeleton bone mask features to initially estimate the fluoroscope pose, while the optimization step combines bone-proportion-weighted local normalized cross correlation with the dice of the bone mask to suppress non-bone interference and enforce structural consistency for iterative registration. We evaluate our methods on public data (without duodenoscope occlusion) and in-house data (with duodenoscope occlusion), with the experimental results showing that the proposed framework significantly outperforms state-of-the-art registration methods, reducing the registration error from 66.07 to 8.25 mm, as well as enhancing the registration success rate from 12.8% to 79.4%.

**Keywords:** Fluoroscopy · Intraoperative 2D-3D Registration · Vision Transformer · Domain Adaptation · ERCP

## 1 Introduction

Minimally invasive surgery such as endoscopic retrograde cholangiopancreatography (ERCP) usually requires fluoroscopy (X-ray) imaging to obtain spatial information of target anatomical structures and surgical instruments [4].X-ray imaging suffers from inherent limitations, e.g., lack of depth information, radiation exposure, and limited soft-tissue visibility.Typically, ERCP is a surgical

procedure for biliary and pancreatic diseases but carries a high risk of complications [1]. Intraoperative 2D-3D registration, i.e., aligning preoperative acquired 3D computed tomography (CT) or magnetic resonance (MR) volumes with intraoperative 2D fluoroscopic (X-ray) images, is a promising way to deal with these limitations and complications [15], enhancing surgical safety and efficiency.

Current intraoperative registration methods include landmark-, intensity- and learning–based methods. Landmark-based methods annotate anatomical landmarks in 2D and 3D images and estimate the pose using techniques like perspective-n-point [12,13]. These methods are accurate but rely on expert annotations [2] and are sensitive to instrument occlusion. Intensity-based methods maximize the similarity between real and virtual fluoroscopic images [6,8]. They require no manual annotation and offer good generalization but are affected by occlusion and domain shift. Deep learning–based approaches are increasingly developed in recent years [5,7]. These approaches typically rely on data annotation and augmentation to improve the generalization, but still fail to align real and virtual images with occlusion and domain shift.

To address the limitations mentioned above, this work proposes a new intraoperative fluoroscopy-volume (2D-3D) registration framework with virtual fluoroscopic data generation and virtual-supervision pose regression networks (VPRN). The technical contributions or highlights are clarified as follows. First, a new architecture of VPRN is created with skeleton-guided spatial attention, dual-branch feature extraction, and gated fusion. Specifically, we incorporate automatically segmented skeleton bone masks into feature extraction and pose optimization to address the problems of occlusion and domain shift between virtual and real X-ray images. Skeleton bone structures are high-contrast and minimally affected by occlusion and domain shift, significantly improving robustness and registration accuracy. Next, a new pose optimization strategy to improve the registration accuracy and stability is established by the weighted local normalized cross correlation and dice similarity measures. Additionally, a differentiable skeleton bone render is created for both model training and pose optimization.

## 2 Intraoperative Fluoroscopy-Volume Registration

This section details our virtual data generator, VPRN and pose optimization for intraoperative 2D-3D registration. Virtual fluoroscopic (X-ray) images and skeleton bone masks are automatically generated to train VPRN for initial fluoroscope pose estimation. In the optimization stage, the VPRN-estimated fluoroscope pose is refined for accurate registration. Fig. 1 illustrates the workflow of virtual fluoroscopic data generation and the architecture of VPRN.

### 2.1 Virtual Fluoroscopic Data Generation

As shown in Fig. 1, for a given pose $\mathbf{T}$, 2D virtual fluoroscopic images are generated using differential digitally reconstructed radiography (DiffDRR) [7], while

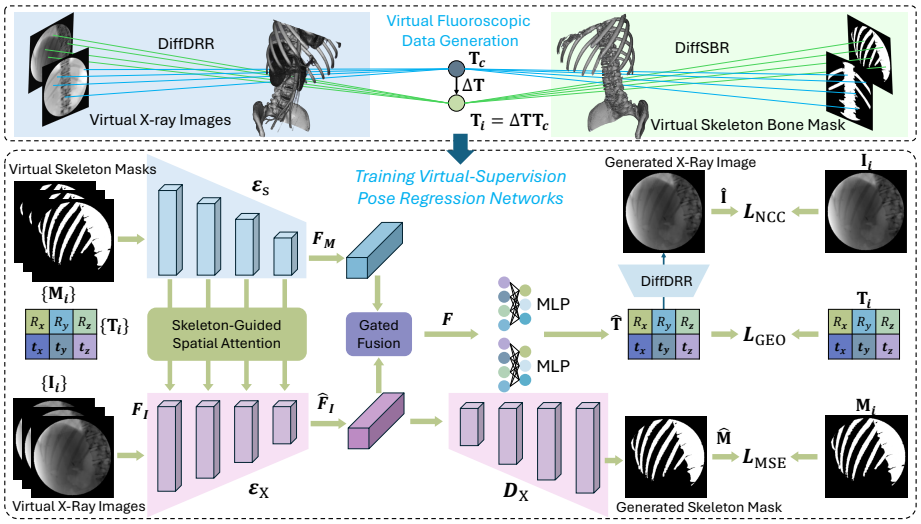



Fig. 1: The Intraoperative fluoroscopy-volume registration with virtual fluoroscopic data generation and virtual-supervision pose regression networks

their corresponding skeleton bone masks are generated by a differential skeleton bone render (DiffSBR). Their differentiability enables end-to-end training and gradient-based optimization. Here, DiffSBR is designed based on DiffDRR to avoid additional model complexity and computational overhead. Specifically, non-bone tissues are first removed from CT images via thresholding, retaining only the attenuation coefficients of skeletal structures. The shared differentiable rendering process then produces both virtual X-ray image $\mathbf{I}$ and bone-only projection image $\mathbf{I}_M$. Finally, $\mathbf{I}_M$ is normalized and passed through a tanh function to generate a skeleton bone mask $\mathbf{M}$ that approximates a binary distribution:

$$\mathbf{M} = \tanh\left(\frac{\alpha(\mathbf{I}_M - \min(\mathbf{I}_M))}{\max(\mathbf{I}_M) - \min(\mathbf{I}_M)}\right), \tag{1}$$

where $\alpha$ is a scaling factor that expands the normalized values of $\mathbf{I}_M$. To automatically generate training samples or data, random perturbations $\{\Delta\mathbf{T}_i\}_{i=1}^N$ are sampled around the central pose $\mathbf{T}_c$ to generate a set of poses $\{\mathbf{T}_i\}_{i=1}^N$ defined as $\mathbf{T}_i = \Delta\mathbf{T}_i\mathbf{T}_c$. All the poses are used to render virtual fluoroscopic $\{\mathbf{I}_i\}_{i=1}^N$ and their corresponding skeleton bone masks $\{\mathbf{M}_i\}_{i=1}^N$ for training.

### 2.2 Virtual-Supervision Pose Regression Networks

VPRN adopts a symmetric dual-branch architecture with encoders $\mathcal{E}_\mathrm{X}$ and $\mathcal{E}_\mathrm{S}$ extracting input image feature $F_I$ and skeleton bone mask feature $F_M$ from $\mathbf{I}$ and $\mathbf{M}$. Encoders $\mathcal{E}_\mathrm{X}$ and $\mathcal{E}_\mathrm{S}$ are the architecture of ResNet-18 [10].

**Skeleton-Guided Spatial Attention.** Fig. 2 shows the structure of the skeleton-guide spatial attention. At each encoding layer $l$ ($l = 1, 2, 3, 4$), bone

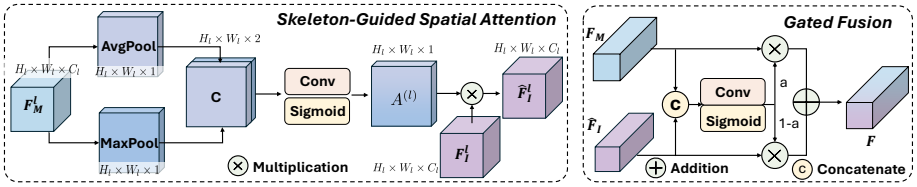



Fig. 2: Structures of the skeleton-guided spatial attention and gated fusion

mask feature $F_M^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ generates a spatial attention map $A^l$, which modulates the image branch feature $F_I^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ to produce feature $\hat{F}_I^l$:

$$A^l = \text{Sigmoid}\Big(\text{Conv}_{3\times3}([\text{AvgPool}(F_M^l); \text{MaxPool}(F_M^l)])\Big), \hat{F}_I^l = F_I^l \otimes A^l, \quad (2)$$

where $A^l \in \mathbb{R}^{H_l \times W_l \times 1}$, AvgPool and MaxPool are channel-wise average and maximum operations, and $\otimes$ denotes the element-wise multiplication. The skeleton-guided spatial attention focuses the image features on stable bone regions, suppressing interference from non-bone regions and enhancing robustness under complex intraoperative fluoroscopic imaging conditions.

**Gated Fusion.** Fig. 2 shows the structure of the gated fusion. The image feature $\hat{F}_I$ and skeletal mask feature $F_M$ are first concatenated along the channel dimension. A convolutional layer followed by a sigmoid activation function learns a coefficient $a$ for adaptive fusion:

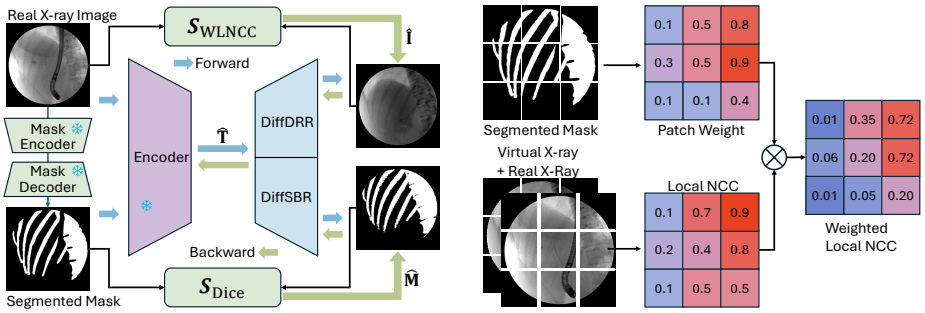
$$a = \text{Sigmoid}\Big(\text{Conv}([\hat{F}_I; F_M])\Big) \quad (3)$$

$$F = a \otimes F_M + (1 - a) \otimes \hat{F}_I \quad (4)$$

This module adaptively balances image and skeletal semantic information for pose regression. The fused feature $F$ is then fed into independent multi-layer perceptron (MLP) to regress rotation and translation parameters.

**Training Loss.** During encoding, the multi-scale skeleton bone mask feature modulates the image feature through the skeleton-guided spatial attention, enhancing the network's focus on skeletal structures. After the feature extraction, the image feature $\hat{F}_I$ is fed into two paths. One path is processed by decoder $\mathcal{D}_\text{X}$, which predicts the bone mask $\hat{\mathbf{M}}$ from the input X-ray image $\mathbf{I}$. The mean squared error loss $\mathcal{L}_\text{MSE}$ is then computed by comparing $\hat{\mathbf{M}}$ with ground truth mask $\mathbf{M}$. This auxiliary task helps encoder $\mathcal{E}_\text{X}$ focus on bone-related discriminative features. Note that $\mathcal{D}_\text{X}$ is only employed during training to provide this structural supervision, and is removed during intraoperative inference.

Another path is adaptively fused with bone mask features $F_M$ via a gated fusion module. The fused features are then fed into two independent MLPs to regress rotation and translation parameters, yielding the final pose estimate $\hat{\mathbf{T}}$. Subsequently, the predicted pose $\hat{\mathbf{T}}$ is compared with the ground-truth pose $\mathbf{T}$




(a) Pose optimization (b) WLNCC computation

Fig. 3: The pose optimization stage with the WLNCC computation

to compute the geometric loss $\mathcal{L}_{\text{GEO}}$:

$$\mathcal{L}_{\text{GEO}}(\mathbf{T}, \hat{\mathbf{T}}; f) = \sqrt{\left(\frac{f}{2}\|\log(\mathbf{R}^{\text{T}}\hat{\mathbf{R}})\|\right)^2 + \|\mathbf{t} - \hat{\mathbf{t}}\|^2} + \|\log(\mathbf{T}^{-1}\hat{\mathbf{T}})\| \tag{5}$$

where $\log(\cdot)$ denotes the logarithm and $f$ is the focal length. Using $\hat{\mathbf{T}}$, a synthetic X-ray image $\hat{\mathbf{I}}$ is generated through DiffDRR. The normalized cross-correlation (NCC) loss $\mathcal{L}_{\text{NCC}}$ is then computed to compare $\hat{\mathbf{I}}$ with the input X-ray image $\mathbf{I}$. The overall loss is the combination of the three components:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{GEO}} - \mathcal{L}_{\text{NCC}}, \tag{6}$$

where $\lambda_1$ and $\lambda_2$ are hyperparameters that are experimentally determined.

### 2.3 Intraoperative Pose Optimization

During the VPRN-based intraoperative 2D-3D registration, a pre-trained segmentation network (this work uses nnU-Net [11]) is first employed to extract skeleton bone mask $\mathbf{M}$ from 2D real fluoroscopic $\mathbf{I}$. Real X-ray image and segmented mask $\mathbf{I}$ and $\mathbf{M}$ are then fed into the trained virtual-supervision pose regression networks to estimate the initial pose of the current real fluoroscopic image. This initial pose is refined by the optimization procedure, as shown in Fig. 3(a).

Basically, the pose optimization iteratively generate virtual fluoroscopic $\hat{\mathbf{I}}(\hat{\mathbf{T}})$ and skeleton bone mask $\hat{\mathbf{M}}(\hat{\mathbf{T}})$ via differentiable rendering and the initial pose $\hat{\mathbf{T}}$ to compute the skeleton bone mask dice $\mathcal{S}_{\text{Dice}}(\mathbf{M}, \hat{\mathbf{M}}(\hat{\mathbf{T}}))$ and the weighted local NCC (WLNCC) similarity $\mathcal{S}_{\text{WLNCC}}(\mathbf{I}, \hat{\mathbf{I}}(\hat{\mathbf{T}}), \mathbf{M})$

$$\mathcal{S}_{\text{WLNCC}}(\mathbf{I}, \hat{\mathbf{I}}(\hat{\mathbf{T}}), \mathbf{M}) = \frac{\sum_{k=1}^{K} \omega(\mathbf{M}_k)\text{NCC}(\mathbf{I}_k, \hat{\mathbf{I}}_k(\hat{\mathbf{T}}))}{\sum_{k=1}^{K} \omega(\mathbf{M}_k)}, \tag{7}$$

$$\omega(\mathbf{M}_k) = \frac{1}{UV}\sum_{u=1}^{U}\sum_{v=1}^{V}\mathbf{M}_k(u,v), \tag{8}$$

where $\mathbf{I}_k$ and $\hat{\mathbf{I}}_k(\hat{\mathbf{T}})$ are the $k$-th image patches in $\mathbf{I}$, $\hat{\mathbf{I}}(\hat{\mathbf{T}})$, and $\mathbf{M}_k$ is the $k$-th skeleton bone mask patch in $\mathbf{M}$.

Eventually, the pose optimization procedure is formulated as:

$$\mathbf{T}_* = \arg_{\hat{\mathbf{T}}}\max\left(\mathcal{S}_{\text{WLNCC}}(\mathbf{I},\hat{\mathbf{I}}(\hat{\mathbf{T}}),\mathbf{M}) + \mathcal{S}_{\text{Dice}}(\mathbf{M},\hat{\mathbf{M}}(\hat{\mathbf{T}}))\right), \tag{9}$$

which maximizes the total similarity to obtain the optimal pose $\mathbf{T}_*$.

A robust similarity measure is essential to deal with instrument occlusion and domain shift between virtual and real X-ray images for accurate 2D-3D registration. This work introduces a hybrid similarity measure with $\mathcal{S}_{\text{Dice}}(\mathbf{M},\hat{\mathbf{M}}(\hat{\mathbf{T}}))$ and $\mathcal{S}_{\text{WLNCC}}(\mathbf{I},\hat{\mathbf{I}}(\hat{\mathbf{T}}),\mathbf{M})$. WLNCC can independently compute the structural similarity in local windows or patches and be less sensitive to occlusion regions, and weight or suppress non-bone regions that are typically more severely corrupted. The skeleton bone dice similarity enforces global structural consistency of the skeletal regions. Therefore, the hybrid similarity measure can complementarily characterize local details and global structures during optimization, achieving an accurate alignment between virtual and real fluoroscopic images.

## 3 Validation

We collected clinical fluoroscopic images and preoperative CT scans from patients who underwent ERCP procedures. Our in-house database totally contains 320 real fluoroscopic (X-ray) images. Most real fluoroscopic images in the in-house database involve duodenoscope occlusion and inter-case variations. Additionally, we conducted experiments on a public database DeepFluoro [9] that includes 366 X-ray images. Both in-house and public databases were used to evaluate our proposed and state-of-the-art 2D-3D registration methods.

We used the Adam optimizer with a learning rate of $1 \times 10^{-3}$ and cosine annealing for 1000 epochs. Each epoch dynamically generates 800 virtual X-ray images with a batch size of 8. Segmenting skeleton bone mask from 2D real fluoroscopic images uses nnU-Net [11] that was trained on 524 annotated images. In the pose optimization, the WLNCC window size was 21, with initial learning rates of $7.5 \times 10^{-3}$ for rotation and 5 for translation, using Adam. The learning rate decays by 0.9 every 25 iterations in 250 iterations. The hyperparameters used in all experiments were experimentally determined: $\alpha = 50$, $\lambda_1 = 10$ and $\lambda_2 = 10^{-2}$. All experiments were conducted on an NVIDIA RTX 3090 GPU.

We used four evaluation metrics: 3D mean target registration error (mTRE), 2D mean projection distance (mPD), registration success rate (SR) that is the proportion of cases with mTRE $< 10$ mm, and sub-millimeter success rate (SSR) that is the proportion of cases with mTRE $< 1$ mm.

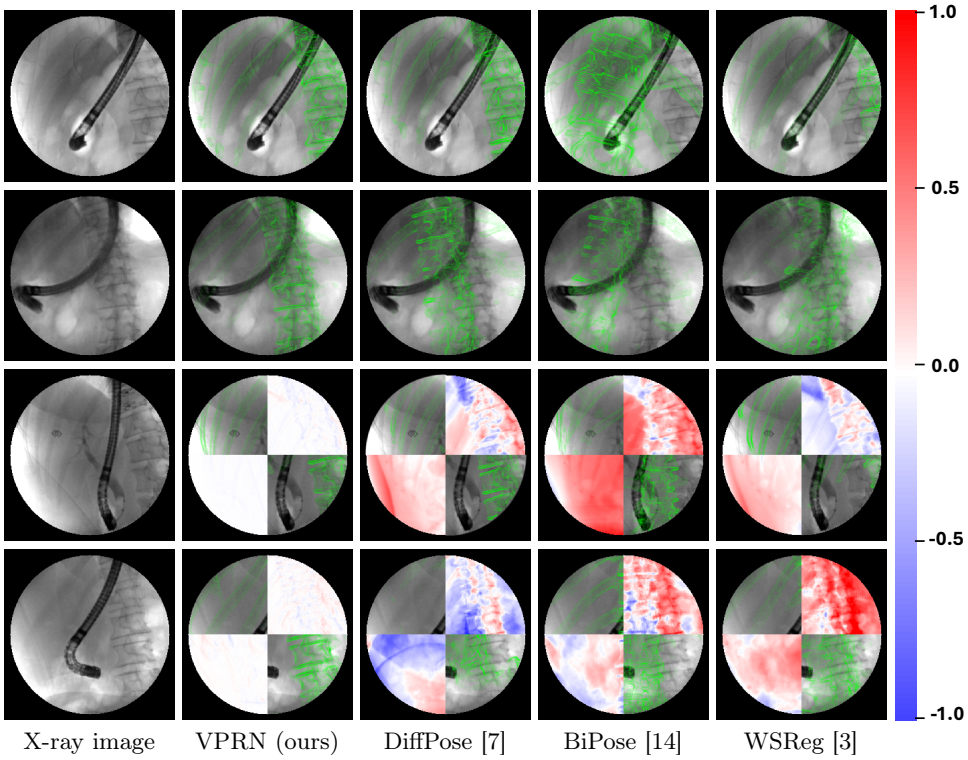



Fig. 4: The registration results of using different methods: Columns 1-5 correspond to real fluoroscopic images, our VPRN model, DiffPose [7], BiPose [14] and WSReg [3]. The registered structures (green contours) and error maps (Rows 3-4, blue and red indicate large errors) were overlaid on real X-ray images

## 4 Results and Discussion

**Methods Comparison.** We compare our VPRN-based intraoperative 2D-3D registration method with state-of-the-art approaches including DiffPose [7], BiPose [14], and WSReg [3]. Fig. 4 illustrates the registration results of using these compared methods tested on in-house and public data. Rows 1-2 in Fig. 4 demonstrate VPRN can maintain stable registration with strong structural (green regions – skeleton bones) consistency between real and virtual images under occlusion and domain shift. Rows 3-4 show the error maps (blue and red indicate large errors) of the methods. VPRN attains much lower error than the others.

Table 1 summarizes the quantitative results of all the methods. While all the methods achieve comparable results in DeepFluoro without endoscope occlusion, VPRN attains much better registration accuracy than the other methods. It can reduce mTRE, mPD from 66.07 mm, 33.25 mm to 8.25 mm, 4.34 mm, as well as improve SR from 12.8% to 79.4%. This implies that our method is much robust to occlusion and domain shift, while the others suffer from these problems.

Table 1: Comparison of different methods DiffPose [7], BiPose [14], WSReg [3] and VPRN tested on in-house and public databases (mPD and mTRE unit: mm)

| Methods | In-house Data | | | Public DeepFluoro | | |
|---|---|---|---|---|---|---|
| | mTRE ↓ | mPD↓ | SR ↑ | mTRE ↓ | mPD↓ | SSR ↑ |
| DiffPose [7] | 69.13±46.61 | 49.86±48.47 | 2.9% | 0.78±1.49 | 2.04±1.32 | 83.6% |
| BiPose [14] | 136.9±59.92 | 109.1±80.49 | 0.0% | 0.84±2.27 | 2.13±3.25 | 83.9% |
| WSReg [3] | 66.07±56.41 | 33.25±23.43 | 12.8% | 0.73±1.38 | 1.97±1.24 | 84.4% |
| Our VPRN | **8.25±11.34** | **4.34±6.40** | **79.4%** | **0.63±1.35** | **1.91±1.09** | **90.4%** |

Table 2: Different modules of $O_1$, $O_2$,$O_3$, $O_4$ used in VPRN and different similarity measures of $Q_N$, $Q_L$, $Q_S$, $Q_W$, $Q_N^D$, $Q_L^D$, $Q_S^D$, $Q_W^D$ selected in the optimization

| Module | mTRE ↓ | SR ↑ | Measure | mTRE ↓ | SR ↑ | Measure | mTRE ↓ | SR ↑ |
|---|---|---|---|---|---|---|---|---|
| $O_1$ | 59.78 | 0.0% | $Q_N$ | 88.33 | 0.0% | $Q_N^D$ | 35.64 | 18.8% |
| $O_2$ | 39.12 | 0.5% | $Q_L$ | 33.56 | 24.5% | $Q_L^D$ | 27.01 | 47.1% |
| $O_3$ | 33.12 | 6.0% | $Q_S$ | 59.15 | 0.0% | $Q_S^D$ | 32.57 | 14.3% |
| $O_4$ | **32.04** | **14.5%** | $Q_W$ | **30.80** | **39.3%** | $Q_W^D$ | **8.25** | **79.4%** |

**Ablation Study.** We conduct an ablation study by progressively adding modules into VPRN: (1) $O_1$ denotes the ResNet-18 baseline using only X-ray images, (2) $O_2$: Adding the bone mask branch with feature fusion into $O_1$; (3) $O_3$: Adding the skeleton-guided spatial attention into $O_2$, and (4) $O_4$: Integrating the bone mask decoder into $O_3$. The ablation results were shown in the left of Table 2. Each module consistently improves pose regression. In particular, the baseline without using skeleton bone masks cannot accurately estimate fluoroscopic poses under occlusion and domain shift. On the other hand, different similarity measures were evaluated for the pose optimization procedure. We typically compare the similarity measures: (1) $Q_N$ is NCC, (2) $Q_L$ is local NCC, (3) $Q_S$ is the structural similarity index and (4) $Q_W$ is WLNCC. We also integrate the dice into these measures to obtain hybrid measures $Q_N^D$, $Q_L^D$, $Q_S^D$ and $Q_W^D$. The optimized results of these measures were summarized at the right of Table 2. WLNCC ($Q_W$ or $Q_W^D$) attains lower mTRE and higher SR than the others.

**Discussion.** The effectiveness of our proposed registration framework lies in several aspects. First, the skeleton-guided spatial attention and auxiliary skeleton bone mask decoder in VPRN can precisely capture anatomical structures (skeleton bones) in fluoroscopic or X-ray images, mitigating interference from non-structure regions that are sensitive to occlusion and appearance changes. During intraoperative optimization, WLNCC can suppress the influence of unreliable non-bone regions, enabling more precise iterative refinement. Moreover, the dice similarity measure can further enforce the skeleton-bone structural consistency, enhancing stability and robustness of the optimization.

Unfortunately, our method still suffers from certain limitations. First, the registration depends on the skeleton bone segmentation. The registration accuracy can be deteriorated if severe skeleton bones are severely occluded or inaccurately

segmented in real fluoroscopic images. Next, tissue deformation (e.g., heart and respiratory motion) is still an open problem. Additionally, our registration is a time-consuming procedure. The registration took about 26 seconds per image.

In summary, this work proposes a new intraoperative fluoroscopy-volume registration framework established by virtual-supervision pose regression networks and pose refinement. The experimental results show that the proposed method significantly outperforms state-of-the-art approaches tested on public DeepFluoro and in-house databases, reducing the target registration error from 66.07 to 8.25 mm in real fluoroscopic images with occlusion and domain shift.

**Acknowledgments.** This work was supported in part by the National Natural Science Foundation of China under Grant 82272133, in part by the High-Quality Development Science and Technology Major Project of Xiamen Health Commission under Grant 2024GZL-ZD03, and in part by Ningbo 2035 Key Research and Development Program under Grant 2024Z127.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Barakat, M., Saumoy, M., Forbes, N., Elmunzer, B.J.: Complications of endoscopic retrograde cholangiopancreatography. Gastroenterology **169**(2), 230–243 (2025)
2. Bier, B., Goldmann, F., Zaech, J.N., Fotouhi, J., Hegeman, R., Grupp, R., Armand, M., Osgood, G., Navab, N., Maier, A., et al.: Learning to detect anatomical landmarks of the pelvis in x-rays from arbitrary views. International journal of computer assisted radiology and surgery **14**(9), 1463–1473 (2019)
3. Cui, Y., Meng, M.Q.H., Min, Z.: Weakly-supervised 2d/3d image registration via differentiable x-ray rendering and roi segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 648–658. Springer (2025)
4. Cui, Y., Song, R., Li, Y., Meng, M.Q.H., Min, Z.: Multi-view 2d/3d image registration via differentiable x-ray rendering. Procedia Computer Science **250**, 282–288 (2024)
5. Gao, C., Feng, A., Liu, X., Taylor, R.H., Armand, M., Unberath, M.: A fully differentiable framework for 2d/3d registration and the projective spatial transformers. IEEE transactions on medical imaging **43**(1), 275–285 (2023)
6. Gao, C., Liu, X., Gu, W., Killeen, B., Armand, M., Taylor, R., Unberath, M.: Generalizing spatial transformers to projective geometry with applications to 2d/3d registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 329–339. Springer (2020)
7. 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)
8. 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 (2022)

9. Grupp, R.B., Unberath, M., Gao, C., Hegeman, R.A., Murphy, R.J., Alexander, C.P., Otake, Y., McArthur, B.A., Armand, M., Taylor, R.H.: Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2d/3d registration. International journal of computer assisted radiology and surgery **15**(5), 759–769 (2020)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
12. Lepetit, V., Moreno-Noguer, F., Fua, P.: Ep n p: An accurate o (n) solution to the p n p problem. International journal of computer vision **81**(2), 155–166 (2009)
13. Li, S., Xu, C., Xie, M.: A robust o (n) solution to the perspective-n-point problem. IEEE transactions on pattern analysis and machine intelligence **34**(7), 1444–1450 (2012)
14. Meng, R., Chen, C., Lu, M., Fu, X., Xiao, Y., Wang, K., Zou, Y., Li, Y.: Parametric bi-invariant learning for improved precision in 2d/3d image registration. Biomedical Signal Processing and Control **105**, 107603 (2025)
15. Zhang, J., Yang, Z., Jiang, S., Zhou, Z.: A spatial registration method based on 2d–3d registration for an augmented reality spinal surgery navigation system. The International Journal of Medical Robotics and Computer Assisted Surgery **20**(1), e2612 (2024)