This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

LXPose++: an Attention-Based Framework for Robust Real-Time CT to X-Ray Registration

Federica Facente[1], Wen Wei[2], Manasi Kattel[1], Hervé Delingette[1,*], Nicholas Ayache[1,*], Pierre Berthet-Rayne[2,*], and Benjamin Billot[1,*]

[1] Université Côte d'Azur, Inria, Epione Team, Sophia Antipolis, France
[2] Carvolix, Nice, France
federica.facente@inria.fr

Abstract. 2D X-Ray imaging provides real-time guidance during minimally invasive interventions. Nonetheless, it lacks depth and soft-tissue contrast. Clinicians must therefore mentally register it with preoperative 3D Computed Tomography (CT), increasing cognitive overload. LXPose, a recent two-stage CNN framework, automates this 3D/2D registration in real time (20ms) through a single forward pass, bypassing the much slower (7s) state-of-the-art optimization methods. However, its accuracy remains limited by the domain gap between synthetic training projections and real intraoperative X-Rays. Here, we propose to build on this method and present LXPose++, which introduces two targeted improvements: masking the CT to retain only rigid bones, preventing the networks from aligning anatomies that deform intraoperatively; and integrating a Convolutional Block Attention Module (CBAM) to refine feature extraction. On two datasets spanning different organs, these improvements nearly halve the projection error of LXPose with no added inference cost. Crucially, LXPose++ reaches the accuracy of state-of-the-art optimization methods, while being 40 times faster, a step toward online clinical applications. Code and models are available here.

Keywords: Image-guided interventions · 3D/2D registration · Attention mechanisms

1 Introduction

Minimally invasive interventions rely on image guidance to visualize patient anatomy and surgical tools, reducing invasiveness and improving procedural safety [29]. X-Ray fluoroscopy acquired with C-arm systems is the dominant modality in image-guided procedures due to its real-time capability. However, it provides only 2D projections with limited depth and poor soft-tissue contrast. For this reason, procedure planning typically relies on preoperative 3D Computed Tomography (CT). During intervention, clinicians must mentally project this preoperative 3D information onto intraoperative 2D X-Rays to interpret the spatial relationship between tools and anatomy, increasing the cognitive load and

[*] contributed equally

the risk of errors [44,35]. This can be addressed by aligning 3D CT with live 2D fluoroscopy, improving accuracy while reducing procedure time and radiation exposure [5,22]. However, beyond accuracy, clinical deployment imposes strict real-time constraints: a latency below ∼200ms is required to preserve hand–eye coordination, and is further decreased to 40ms per frame for robot-assisted guidance, where imaging must be tightly coupled to robotic control loops [36,39,34].

Aligning intraoperative X-Rays with a CT volume is typically formulated as a registration problem. Since both modalities share the same imaging physics, digitally reconstructed radiographs (DRRs) can be synthesized by simulating X-Ray propagation through the scan [28,24], and the goal is to find the C-arm pose for which these projections best match the intraoperative X-Rays. Classical approaches solve this by iteratively optimizing a similarity measure between DRRs and X-Rays [14,3,30]; their accuracy depends on the realism of the simulated projections, motivating advanced DRR frameworks such as DeepDRR [40,16,9]. While differentiable DRR rendering and learning-based initialization have improved robustness [11,13,12,45], optimization still requires several seconds to converge, making it unsuitable for real-time guidance. To remove this bottleneck, several works reformulate registration as direct pose regression, estimating the C-arm pose in a single forward pass [26,4,31]. As annotated clinical fluoroscopy is scarce, these networks are trained on synthetic DRRs, which is a major obstacle to clinical deployment due to the domain gap between training DRRs and real intraoperative X-Rays seen at test-time. This gap has two main components. The first is a morphological gap arising from deformations between preoperative imaging and intervention, partly rigid (bone motion) and partly non-rigid (soft-tissue deformation); the second is an intensity gap driven by differences in acquisition characteristics. Since DRRs are rendered from a single preoperative CT, they capture none of the morphological variation present in intraoperative X-Rays, and bridging this gap is therefore a prerequisite for accurate pose estimation. On the morphological side, one approach deforms the CT volume prior to DRR generation to simulate respiratory motion and other intraoperative changes [27], while xvr [13] sidesteps the problem entirely by masking the CT to retain only rigid bony structures, anchoring registration to anatomy that remains stable across acquisitions. The intensity gap, instead, is typically addressed through GAN-based image translation between the DRR and X-Ray domains [20,10,1,6] or through domain randomization, which has the advantage of requiring no intraoperative X-Rays [38,9].

In this paper, we present LXPose++, an enhanced version of LXPose [8], a two-stage CNN that progressively refines pose estimates in a single forward pass for real-time deployment. Beyond the extended data augmentation used to mitigate the domain gap, we introduce two complementary strategies to bridge the discrepancy between simulated training data and clinical intraoperative X-Rays. First, we mask the preoperative CT to isolate rigid bony structures [13], preventing the network from aligning anatomies subject to intraoperative deformation. Second, we integrate a Convolutional Block Attention Module (CBAM) [42] into both stages, applying channel-wise and spatial attention to steer the network to-

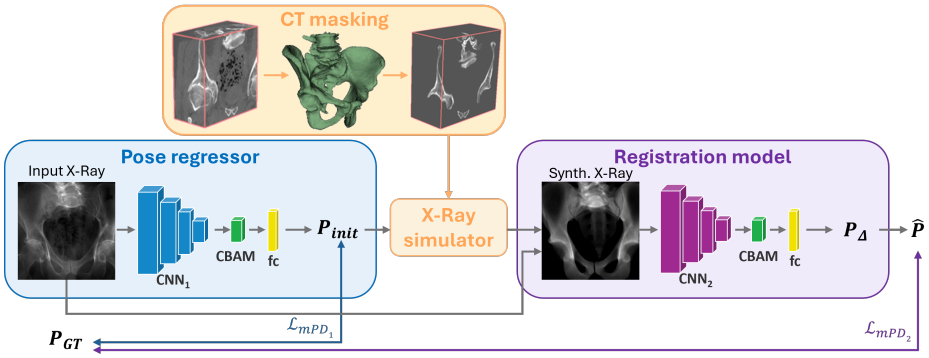


Fig. 1. Method overview. Given an input X-Ray, CNN_1 first regresses an initial pose P_{init} using a ResNet-18 backbone [18], a CBAM attention module, and fully connected (fc) layers, trained with a projection loss $\mathcal{L}_{mPD_1}$. A DRR is then rendered at P_{init} and fed alongside the input X-Ray to CNN_2, which predicts a correction ΔP, yielding the final pose $\hat{P} = \Delta P \cdot P_{\text{init}}$, optimized with $\mathcal{L}_{mPD_2}$. Crucially, the DRR used in this refinement stage is simulated from a CT masked to retain only rigid bone structures.

ward informative regions. While performing competitively with the state-of-the-art methods on two public datasets, our method operates 40× faster, thereby providing real-time inference speed necessary for clinical deployment.

2 Methods

Our method builds on LXPose [8], a real-time multi-stage CNN for 3D/2D registration, and introduces two contributions to tackle the real-to-synthetic domain gap: we mask the CT used in the registration model to restrict alignment to rigid bony anatomy [13], and we integrate a CBAM [42] attention module at each stage to refine feature extraction. Our method is illustrated in Figure 1.

2.1 Projection geometry

Our goal is to estimate the pose of the X-Ray source (i.e., C-arm) relatively to the 3D preoperative CT. The projection geometry of a C-arm is commonly formulated using the pinhole camera model [17], describing how 3D CT voxels are projected onto a 2D X-Ray image plane. This model relies on a 3×4 projection matrix $\Pi = KP$ that maps 3D world points to 2D image pixels via $X^{2D} = KP\,X^{3D}$. The intrinsic matrix K encodes the camera parameters, while the extrinsic pose matrix $P = [R\,|\,t] \in \mathbb{R}^{3\times4}$ captures the rigid transform from world to camera coordinates, with $R \in \mathbb{R}^{3\times3}$ the rotation matrix and $t \in \mathbb{R}^3$ the translation vector. In practice, t encodes the in-plane position (t_x, t_y) and magnification t_z, and R is parametrised by three angles (α, β, γ) comprising the orbital and secondary gantry angles and the detector normal angle, respectively.

2.2 Multi-stage CNN and learning

LXPose++ employs a two-stage cascaded architecture that progressively refines the pose estimate [8]. First, CNN_1 directly regresses the C-arm pose P_{init} from the input image, parametrizing rotation as an axis-angle vector converted to a rotation matrix via the Rodrigues formula [17]. A DRR is then rendered at P_{init} using the differentiable diffDRR X-Ray simulator [11]. Finally, CNN_2 takes both this synthetic image and the input X-Ray to predict a residual correction P_Δ, yielding the final pose $\hat{P} = P_\Delta P_{\text{init}}$. Each stage employs a ResNet-18 backbone [18], a CBAM [42] attention module (Section 2.4), and a fully connected layer that regresses the rotation (α, β, γ) and translation (t_x, t_y, t_z) parameters.

Training is performed end-to-end with a mean Projection Distance (mPD) loss for each CNN [7]. Following LXPose and to avoid manual annotation, landmarks used to compute mPD during training are extracted automatically from the CT surface meshes [8]. Crucially, CNN_2 can be applied iteratively, each time re-rendering a DRR at the current pose estimate and predicting a further correction, enabling progressive refinement.

2.3 Anatomical CT masking

Building on xvr [13], we propose to mask the CT used for DRR rendering for CNN_2, with the aim of retaining only rigid bony structures (e.g., pelvis, sacrum, and lumbar vertebrae). This is achieved by using an automated bone segmentation obtained with TotalSegmentator [41], and zeroing all voxels outside the skeletal mask, so that ray-casting produces a DRR only containing rigid bone. As a result, the registration CNN focuses exclusively on rigid anatomy.

2.4 Convolutional Block Attention Module

To help focusing networks' attention on key features for pose estimation, we integrate a CBAM [42] module after the final block of each CNN, only bringing negligible computation overhead. In contrast to recent segmentation-oriented attention modules targeting large receptive fields [2,19,43], we favor lightweight local reweighting suited to registration. Given a feature map $F \in \mathbb{R}^{C\times H\times W}$, CBAM sequentially computes channel attention $M_c(F) \in \mathbb{R}^{C\times 1\times 1}$ and spatial attention $M_s(F') \in \mathbb{R}^{1\times H\times W}$, refining F as $F' = M_c(F) \otimes F$ and $F'' = M_s(F') \otimes F'$, where $\otimes$ is the element-wise multiplication. The channel map is computed via a shared multi-layer perceptron (MLP) and sigmoid function σ as $M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F)))$, amplifying discriminative channels and suppressing the rest. The spatial map is then computed with a 7×7 convolution f on channel-pooled features $M_s(F') = \sigma(f([\text{AvgPool}(F'); \text{MaxPool}(F')]))$, with the aim of highlighting the most discriminative image regions [42].

2.5 Implementation details

DRR simulation: Training DRRs are simulated on the fly using diffDRR [11] at a resolution of 256×256, with C-arm pose parameters sampled from normal distributions (potentially multimodal, depending on the clinical acquisition

pattern), whose standard deviations are calibrated to each dataset. Aggressive data augmentation is applied as in [8] (Gaussian noise, Gaussian blur, sharpening, gamma correction, salt-and-pepper noise, and plasma intensity variations) to mitigate the domain gap between synthetic DRRs and real X-Rays.
Training: Training is conducted for 1500 epochs with a batch size of 16, using the Adam optimizer [23] with an initial learning rate of 10^{-3} and a warmup cosine scheduler. All networks are implemented in PyTorch [33]. Training converges within 8–15 hours on an NVIDIA A100 (80 GB), depending on anatomical complexity and pose parameter ranges. At inference, a single forward pass through both stages completes in 20ms on a consumer-grade 12 GB NVIDIA 1080 Ti, satisfying the 40ms real-time control constraint.

3 Experiments and results

3.1 Dataset and preprocessing

We evaluate LXPose++ on two public datasets covering distinct anatomies and X-Ray modalities.
DeepFluoro [15]: this dataset contains six cadaveric subjects, each with a pelvic CT and a series of [24–111] X-Rays (366 total), with known intrinsics, ground truth C-arm poses, and annotated landmarks for evaluation. DRRs are generated by sampling C-arm poses from zero-centered Gaussians with standard deviations: 15° (α, β), 5° (γ), 50mm (t_x, t_y), and 100mm (t_z). Raw attenuation images I are converted to absorption images $I_\mu = \log[\max(I)] - \log I$, where $\max(I)$ corresponds to rays through air only. Images are cropped to remove the collimator effect and resampled to 256×256.
Ljubljana [32]: this dataset comprises cerebral digital subtraction angiographies (DSA) from 10 endovascular neurointervention patients. Each patient has one 3D DSA and two 2D DSAs (lateral and frontal views), available with ground truth intrinsics, C-arm poses, and 3D landmarks for evaluation. The 3D DSA images are thresholded at 1000HU to suppress background noise. DRRs are generated by sampling poses from zero-centered Gaussians with standard deviations of 0.05° (α, γ), 10mm (t_x), and 20mm (t_z). To match the lateral/frontal acquisition protocol, β and t_y are sampled from a bimodal Gaussian mixtures: $\mathcal{N}(\mu = -0.16, \sigma^2 = 0.18)$ and $\mathcal{N}(\mu = 1.56, \sigma^2 = 0.03)$ for β; $\mathcal{N}(\mu = -23, \sigma^2 = 2.5)$ and $\mathcal{N}(\mu = 0.4, \sigma^2 = 1)$ for t_y. Fixed intrinsics are used during training, whereas homographies are applied at test time to correct for intrinsics variations [8]. Images are resampled to 256×256 and converted to absorption images.

3.2 Baselines

We compare LXPose++ against the following methods. First, **PoseNet** [21] is a supervised method that regresses C-arm poses directly from real X-Rays. Next, **xvr** [13] is the state-of-the-art method in CT/X-ray registration, and uses optimization to refine initial poses estimated from a regression CNN. We also evaluate **xvr-reg**, a variant of xvr with only its regression CNN. We then compare

Table 1. Mean (standard deviation) scores for all methods. Within the strict real-time control scenario (<40ms), bold values indicate the best performance and * marks statistical significance (Bonferroni-corrected Wilcoxon signed-rank test, 1% level). Results for the real-time visualization (<200ms) and offline scenarios are also reported. mPD, mTRE, t_x, t_y, t_z are reported in mm; α, β, γ in degrees; time in seconds.

	Real-time control 40ms				**Real-time vis. 200ms**	**Offline**
metric	PoseNet	xvr-reg	LXPose	LXPose++	LXPose++ 12-iter	xvr
				DeepFluoro dataset		
mPD	77.7 (22.6)	4.07 (2.06)	0.98 (1.22)	**0.51 (1.20)***	0.45 (1.20)	0.41 (1.38)
mTRE	187 (22.6)	32.4 (18.3)	11.3 (15.8)	**5.75 (14.9)***	4.92 (14.9)	3.81 (14.9)
α	10.4 (7.31)	1.35 (1.18)	0.35 (0.54)	**0.22 (0.50)***	0.19 (0.52)	0.16 (0.52)
β	5.90 (4.12)	1.55 (1.30)	0.55 (1.07)	**0.29 (1.04)***	0.25 (1.04)	0.18 (1.03)
γ	7.39 (3.30)	0.98 (0.73)	0.26 (0.30)	**0.09 (0.19)***	0.08 (0.19)	0.06 (0.19)
t_x	15.9 (11.8)	1.20 (0.93)	0.28 (0.44)	**0.14 (0.32)***	0.12 (0.30)	0.12 (0.29)
t_y	48.0 (16.4)	1.03 (0.78)	0.34 (0.30)	**0.13 (0.19)***	0.11 (0.16)	0.08 (0.31)
t_z	62.9 (37.6)	12.7 (9.56)	4.89 (6.39)	**1.83 (3.69)***	1.54 (3.52)	1.36 (3.54)
time	0.01 (0.03)	0.01 (0.03)	0.02 (0.03)	0.02 (0.03)	0.19 (0.02)	7.78 (2.09)
				Ljubljana dataset		
mPD	-	2.43 (1.02)	0.87 (0.28)	**0.75 (0.21)***	0.68 (0.17)	0.63 (0.14)
mTRE	-	28.3 (16.1)	11.7 (6.90)	**6.86 (4.02)***	4.75 (2.94)	1.69 (1.03)
α	-	1.29 (1.13)	0.53 (0.42)	**0.29 (0.28)***	0.23 (0.22)	0.06 (0.06)
β	-	1.19 (1.03)	0.31 (0.37)	**0.23 (0.16)***	0.16 (0.13)	0.05 (0.06)
γ	-	0.43 (0.41)	0.23 (0.17)	**0.19 (0.15)***	0.14 (0.13)	0.06 (0.05)
t_x	-	0.60 (0.48)	0.29 (0.20)	**0.23 (0.19)***	0.27 (0.16)	0.20 (0.10)
t_y	-	0.84 (0.73)	0.46 (0.29)	**0.42 (0.22)***	0.41 (0.19)	0.34 (0.10)
t_z	-	8.07 (6.69)	6.55 (5.09)	**3.27 (2.78)***	1.70 (0.92)	0.66 (0.50)
time	-	0.01 (0.03)	0.02 (0.03)	0.02 (0.03)	0.19 (0.02)	8.25 (1.12)

against **LXPose** [8] in order to isolate the contribution of our two improvements (masked CT and CBAM). Finally, we test **LXPose++** with a single pass (satisfying the 40ms real-time control constraint), and a variant **LXPose++ 12-iter** that applies the registration CNN 12 times (satisfying the 200ms real-time visualization constraint). All methods are retrained using original implementations.

3.3 Results

Registration performance: In this experiment, we compare all methods on the two datasets, except for PoseNet on Ljubljana due to the limited availability of real X-Rays necessary to train PoseNet. Since image guidance is performed on the 2D X-Ray plane, our primary metric is the mean Projection Distance (mPD). For completeness we also report the 3D mean Target Registration Error (mTRE), the per-parameter absolute errors (α, β, γ in degrees; t_x, t_y, t_z in mm), and the inference time per image. Results are grouped by deployment regime: strict real-time control (<40ms), real-time visualization (<200ms), and offline.

Remarkably, Table 1 shows that our method LXPose++ outperforms every other real-time-compliant baseline on both anatomies, for all metrics, and with statistical significance. This superiority is particularly pronounced for the direct

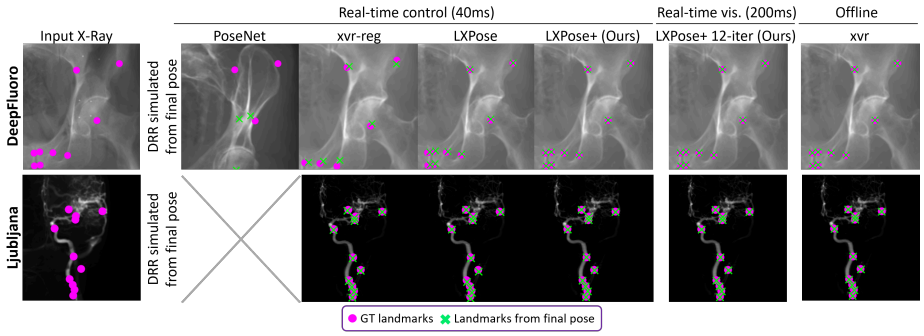


Fig. 2. CT to X-Ray registration results for all methods visualized as: the input X-Ray with 3D-derived landmarks projected from the ground truth pose, DRR with the landmarks projected from the predicted pose.

regression baselines with improvements of 77.2mm and 3.66mm in mPD for LXPose++ over PoseNet and xvr-reg, respectively. These results highlight the difficulty of tackling the 3D CT to 2D X-Ray registration problem with direct regression alone, and confirms the need for further refinement, as in LXPose++.

More detailed comparison with the original LXPose reveals that, on DeepFluoro, LXPose++ reduces the mPD metric from 0.98mm to 0.51mm (a 48% relative improvement) while preserving the 20ms inference time, thus remaining within the strict 40ms real-time control budget. The gain is consistent across every pose parameter: the depth translation t_z, in particular, drops from 4.89mm to 1.83mm. When the time budget is relaxed to the 200ms real-time visualization regime, applying the registration network for 12 iterations further lowers the mPD to 0.45mm and the mTRE to 4.92mm. Crucially, these performances match the accuracy of the state-of-the-art offline method xvr while running 40× faster, demonstrating that real-time performance can be achieved without compromising registration accuracy. This behaviour is also illustrated qualitatively in Figure 2: the base LXPose produces a visibly large landmark misalignment, whereas LXPose++ reaches a quality that is visually indistinguishable from xvr.

A similar trend is observed on the Ljubljana dataset, where adding CBAM attention alone in LXPose++ improves the mPD of LXPose from 0.87mm to 0.75mm and roughly halves the mTRE (11.7mm to 6.86mm) at no additional inference cost. With 12 iterations in LXPose++, the mPD further drops to 0.68mm and the mTRE to 4.75mm. While the in-plane mPD is closely comparable to the state-of-the-art offline xvr, a residual 3D gap remains in mTRE (1.69mm), driven primarily by the depth translation t_z; LXPose++ nonetheless runs over 43× faster. This outcome further demonstrates that accurate CT to X-Ray alignment is achievable under a real-time deployment budget.

Effect of attention and CT masking: We now conduct ablations to study the impact of each new component of LXPose++ over LXPose (first two columns of Table 2) on the DeepFluoro dataset. CT masking and attention modules both

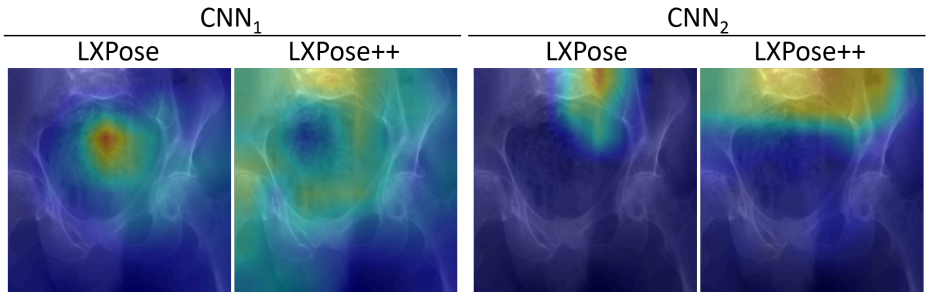


Fig. 3. Grad-CAM [37] maps for both CNN stages in LXPose and LXPose++ on a representative subject of the DeepFluoro dataset.

Table 2. Ablation scores reported for the DeepFluoro real X-Rays as well as on equivalent DRRs generated using the ground truth poses. Best scores are in bold and * marks statistical significance (Bonferroni-corrected Wilcoxon signed-rank test, 1% level).

	Real X-Rays		DRRs	
Ablations	mPD	mTRE	mPD	mTRE
LXPose	0.98 (1.22)	11.3 (15.8)	0.22 (0.28)	2.54 (3.32)
LXPose ⊕ attention	0.87 (1.26)	9.18 (15.37)	0.13 (0.15)	**1.09 (1.12)**
LXPose ⊕ masked CT	0.90 (1.59)	8.73 (15.82)	0.35 (0.55)	3.15 (3.65)
LXPose++	**0.51 (1.20)***	**5.75 (14.94)***	**0.12 (0.10)***	1.16 (1.14)

improve performance by enabling the registration network to focus its attention on the rigid bony structures, which prevents the network from being influenced by non-rigid features. Importantly, combining both improvements within LXPose++ roughly halves the error of LXPose, reaching 0.45 mm mPD and 4.92mm mTRE on real X-Rays, significantly below either component in isolation.

In order to precisely quantify the remaining performance gap between the real X-Rays and the training DRRs, we additionally report results obtained on DRRs generated using diffDRR on the CTs from the ground truth poses. Interestingly, this residual gap is markedly reduced by LXPose++: a ∼57% reduction in both metrics with respect to the base LXPose.

Finally, to qualitatively assess how the employed attention mechanisms impact the learning process, we compute Grad-CAM [37] activation maps for the base LXPose and LXPose++ for both CNN_1 and CNN_2 (Figure 3). For the base LXPose, the activations are small and spatially restricted, relying on only a few isolated spots. In contrast, LXPose++ produces more spread-out activations that extend over the registration-relevant anatomy (e.g., the pelvis).

4 Conclusion

We presented LXPose++, a multi-stage registration framework that narrows the gap between synthetic training data and real intraoperative X-Rays through

targeted input masking and attention-based mechanisms. Their combination reduces the projection error of the base LXPose by nearly half, without any increase in inference time. With iterative refinement under the 200ms visualization budget, accuracy becomes competitive with offline state-of-the-art optimization method xvr [13], while running 40 times faster. Future work will target deployment in live clinical and robotic settings by bridging the residual DRR/X-Ray domain gap, validating on datasets with surgical instruments, and extending to patient-agnostic and non-rigid models [25]. Overall, LXPose++ paves the way for real-time registration in image-guided and robot-assisted interventions.

Acknowledgments. This work has been supported by the French government, through the 3IA Cote d'Azur Investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001. Further support was provided with computer resources by GENCI at CINES/IDRIS thanks to the grant 2026 AD010316779 and A0200315107 on the supercomputer Jean Zay's V100/A100 Partition.

Disclosure of Interests. Pierre Berthet-Rayne and Wen Wei are employed by Carvolix. The remaining authors declare no competing interests.

References

1. Aubert, B., et al.: X-ray to DRR images translation for efficient multiple objects similarity measures in deformable model 3D/2D registration. IEEE Transactions on Medical Imaging **42**(4), 897–909 (2023)
2. Azad, R., Niggemeier, L., Hüttemann, M., Kazerouni, A., Aghdam, E.K., Velichko, Y., Bagci, U., Merhof, D.: Beyond self-attention: Deformable large kernel attention for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1287–1297 (2024)
3. Berger, M., et al.: Marker-free motion correction in weight-bearing cone-beam CT of the knee joint. Medical Physics **43**(3), 1235–1248 (2016)
4. Bui, M., et al.: X-Ray PoseNet: 6 DoF pose estimation for mobile X-ray devices. In: WACV. pp. 1036–1044 (2017)
5. Chahla, B., Ozen, M.: Fluoroscopy and cone beam CT guidance in robotic interventions. Journal of Interventional Radiology (2024)
6. Dai, J., et al.: Volumetric tumor tracking from a single cone-beam X-ray projection image enabled by deep learning. Medical Image Analysis **91** (2024)
7. Facente, F., Billot, B., Gopalakrishnan, V., Kattel, M., Wei, W., Golland, P., Delingette, H., Ayache, N., Berthet-Rayne, P.: Multistage cnn for fast registration of 3d preoperative cts to 2d intraoperative x-rays. In: COLAS, MICCAI 2025 (September 2025)
8. Facente, F., et al.: Towards real-time alignment of 3D CT and 2D X-ray with multi-stage CNNs. Computer Assisted Surgery (2026)
9. Gao, C., et al.: Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. Nature Machine Intelligence **5**, 294–308 (2023)
10. Geng, H., et al.: CT2X-IRA: CT to x-ray image registration agent using domain-cross multi-scale-stride deep reinforcement learning. Physics in Medicine & Biology **68**(17) (2023)

11. Gopalakrishnan, V., Golland, P.: Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In: CLIP (2022)
12. Gopalakrishnan, V., et al.: Intraoperative 2D/3D image registration via differentiable X-ray rendering. CVPR pp. 11662–11672 (2024)
13. Gopalakrishnan, V., et al.: Rapid patient-specific neural networks for intraoperative X-ray to volume registration. arXiv:2503.16309 (2025)
14. Groher, M.: 2D-3D Registration of Vascular Images: Towards 3D-guided catheter interventions. Ph.D. thesis, Technische Universität München (2008)
15. Grupp, R., 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**(5), 759–769 (2020)
16. Gu, W., et al.: Extended capture range of rigid 2D/3D registration by estimating Riemannian pose gradients. In: MLMI. pp. 281–291 (2020)
17. Hartley, R.: Multiple view geometry in computer vision. Cambridge University Press (2003)
18. He, K., et al.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
19. Hu, J., Chen, S., Pan, Z., Zeng, S., Yang, W.: Perspective+ unet: Enhancing segmentation with bi-path fusion and efficient non-local attention for superior receptive fields. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). Lecture Notes in Computer Science, vol. 15009, pp. 499–509 (2024)
20. Jaganathan, S., et al.: Self-supervised 2D/3D registration for X-ray to CT image fusion. In: WACV. pp. 2788–2798 (2023)
21. Kendall, A., et al.: PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In: ICCV. pp. 2938–2946 (2015)
22. Key, J., et al.: Cone-beam CT with enhanced needle guidance and augmented fluoroscopy overlay: applications in interventional radiology. European Radiology **33** (2023)
23. Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv:1412.6980 (2014)
24. Knaan, D., Joskowicz, L.: Effective intensity-based 2D/3D rigid registration between fluoroscopic X-ray and CT. In: MICCAI. pp. 351–358 (2003)
25. Lecomte-Denis, F., et al.: Domain agnostic 2D-3D deformable registration application to fluoroscopic guidance without contrast agent. Medical Image Analysis **105** (2025)
26. Lee, B., et al.: Breathing-compensated neural networks for real time C-arm pose estimation in lung CT-fluoroscopy registration. In: ISBI (2022)
27. Lee, B., et al.: Breathing-compensated neural networks for real time C-Arm pose estimation in lung CT-fluoroscopy registration. In: ISBI. pp. 1–5 (2022)
28. Lemieux, L., et al.: A patient-to-computed-tomography image registration method based on digitally reconstructed radiographs. Medical Physics **21**, 1749–1760 (1994)
29. Makkar, R.R., et al.: Five-year outcomes of transcatheter or surgical aortic-valve replacement. New England Journal of Medicine **382**, 799–809 (2020)
30. Meng, C., et al.: 2D-3D registration with weighted local mutual information in vascular interventions. IEEE Access **7**, 629–638 (2019)
31. Miao, S., et al.: A CNN regression approach for real-time 2D/3D registration. IEEE Transactions on Medical Imaging **35**(5), 1352–1363 (2016)
32. Mitrovic, U., 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)

33. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019)
34. Pescio, et al.: Endovascular robotics: technical advances and future directions. Minimally Invasive Therapy & Allied Technologies (2025)
35. Pfandler, M., et al.: Technical and nontechnical skills in surgery: A simulated operating room environment study. Spine **44**(23), 1396–1400 (2019)
36. Ruijters, D., et al.: Latency optimization for autostereoscopic volumetric visualization in image-guided interventions. Neurocomputing **144** (2014)
37. Selvaraju, R.R., et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (2020)
38. Toth, D., et al.: Training deep networks on domain randomized synthetic X-ray data for cardiac interventions. In: MIDL. pp. 468–482 (2019)
39. Tzolov, D., et al.: Impact of laparoscopic training, haptic feedback and visual haptic latency on virtual fine motor performance. Scientific Reports **15** (2025)
40. Unberath, M., et al.: Enabling machine learning in X-ray-based procedures via realistic simulation of image formation. International Journal of Computer Assisted Radiology and Surgery **14**, 1517–1528 (2019)
41. Wasserthal, J., et al.: TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence (2023)
42. Woo, S., et al.: Cbam: Convolutional block attention module. In: ECCV. pp. 3–19 (2018)
43. Xiao, J., Wei, Y., Huang, W., Wang, J., Bi, X., Yang, X., et al.: Neurocognitive insights: Cognitive comprehension attention in multi-organ segmentation. IEEE Transactions on Big Data **11**(6), 3300–3313 (2025)
44. Zaffino, P., et al.: A review on advances in intra-operative imaging for surgery and therapy. Annals of Biomedical Engineering **48**(8), 2171–2191 (2020)
45. Zhang, B., et al.: A patient-specific self-supervised model for automatic X-Ray/CT registration. In: MICCAI (2023)