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.

Explainable AI for Image-based 2D/3D Registration Quality Assessment

Sue Min Cho, Alexander Do, Russell H. Taylor, and Mathias Unberath

Johns Hopkins University, Baltimore MD, USA
scho72@jhu.edu

Abstract. 2D/3D registration is a key enabling algorithm in image-guided surgery, yet even state-of-the-art algorithms occasionally produce clinically significant misalignments. Because even millimeter-scale errors can lead to surgical complications, robust quality assurance is essential. However, current strategies are insufficient for reliable human verification, and existing uncertainty quantification methods do not robustly correlate with registration error. To address these gaps, we present a learning-based framework for 2D/3D registration quality assessment, augmented with explainability mechanisms to support human oversight. Our approach employs early fusion of fluoroscopic images and digitally reconstructed radiographs (DRRs), processed through a convolutional backbone with spatial cross-attention to capture both local alignment features and global misalignment patterns. The model predicts registration success, defined as mean target registration error (mTRE) below 2 mm. To provide interpretable rationale, we integrate Grad-CAM for spatial attribution and conformal prediction for calibrated confidence estimates. Evaluated using leave-one-specimen-out cross-validation on a pelvic fluoroscopy dataset, our proposed approach achieves an accuracy of 0.801 ± 0.030, outperforming a baseline CNN without cross-attention while exhibiting reduced variance across specimens. This framework establishes a foundation for explainable, AI-assisted quality assurance in 2D/3D registration. Our code is available at: https://github.com/jasminecho1008/XAI-2D3D-RegQuality.

Keywords: 2D/3D registration · quality assessment · explainability · cross-attention · image-guided surgery

1 Introduction

2D/3D registration is an enabling technology in image-guided surgery, allowing preoperative volumetric data to be aligned with intraoperative fluoroscopic images for surgical navigation [10]. Despite significant advances in registration algorithms, even state-of-the-art methods occasionally produce clinically significant misalignments. In orthopedic procedures, where millimeter-scale accuracy is essential, undetected registration failures can lead to malpositioned implants, neurovascular injury, or suboptimal surgical outcomes [5]. Consequently, robust quality assurance mechanisms are critical for safe clinical deployment.

Current approaches to 2D/3D registration quality assurance face limitations. Human-based verification methods have been proposed to involve operators in the assessment of 2D/3D registration outcomes [3,4]. Studies using both existing and novel visualization paradigms showed that, on average, users can discern different levels of registration error; however, consistently and reliably detecting misalignments remains insufficient. Beyond human verification, non-learning-based probabilistic methods have sought to quantify registration uncertainty and correlate it with geometric accuracy [2]. Although this approach provides valuable confidence estimates, the relationship between uncertainty and true registration error is nonlinear and insufficient for robust validation.

Learning-based quality assessment has been explored for various registration contexts, primarily in 3D/3D scenarios. For 3D/3D registration, methods include automated quality control for rigid and affine MRI registration [13] and CNN-based approaches for quality assurance of deformable registration on lung CT [6]. However, these 3D-to-3D alignment scenarios address geometrically and perceptually distinct problems from 2D-to-3D registration, where a 2D projection must be compared against a 3D volume.

For 2D/3D registration specifically, prior work has explored automated verification using handcrafted features with classical machine learning [15]. However, such methods rely on anatomy-specific assumptions and do not incorporate deep learning or explainability mechanisms.

Together, these studies have established foundations for quality assurance in 2D/3D registration, but indicate that it remains an open challenge. To address these gaps, we present a learning-based framework for 2D/3D registration quality assessment augmented with explainability mechanisms. Our approach employs early fusion of fluoroscopic images and digitally reconstructed radiographs (DRRs), processed through a convolutional backbone with spatial cross-attention to capture both local alignment features and global misalignment patterns. The model predicts registration success, defined as mean target registration error (mTRE) below 2mm. To provide interpretable rationale supporting human oversight, we integrate Grad-CAM for spatial attribution and conformal prediction for calibrated confidence estimates.

Our contributions are as follows: (1) an architecture combining early fusion with spatial cross-attention, adapted for 2D/3D registration quality assessment, (2) systematic ablation demonstrating complementary benefits of cross-attention and simulated data augmentation, and (3) integration of explainability mechanisms to support human-AI collaboration in surgical quality assurance.

2 Methods

2.1 Problem Formulation

We formulate 2D/3D registration quality assessment as a binary classification task. Given a fluoroscopic X-ray image I_{xray} and a digitally reconstructed radiograph (DRR) I_{drr} generated from a 3D volume at an estimated pose, the goal

is to predict whether the registration is successful. Following established clinical standards in orthopedic surgery [9], we define success as achieving a mean target registration error (mTRE) below 2mm, computed from ground-truth anatomical landmarks.

2.2 Dataset and Preprocessing

Fluoroscopic images of the pelvis were obtained from a public dataset [7], which included five unique cadaveric specimens. This dataset contained real fluoroscopic projections, CT scans, and the corresponding ground truth poses. Using DeepDRR [14], additional simulated projections were generated from the CT scans, resulting in a total of 200 projections (real + simulated) per specimen.

To create training data reflecting realistic registration outcomes, we uniformly sampled 6 degrees of freedom (6DoF) for registration initialization and used an open-source 2D/3D single-view registration algorithm [8] to generate 100 registration results per projection. For each result, the offset and the corresponding DRR were saved, yielding a dataset of 100,000 DRRs across 1,000 fluoroscopic projections. Registration accuracy was evaluated by computing the mTRE based on ground-truth 3D landmarks and the 3D points transformed by the offsets. Registrations were classified as successful if mTRE was below 2mm; otherwise, they were considered failed.

We employed leave-one-specimen-out cross-validation (LOOCV). In each fold, three specimens were used for training, one for validation, and one for testing. Test evaluation used only real projections from the held-out specimen.

To address class imbalance where there were far more failed registrations, non-geometric data augmentations were applied to increase the number of successful registrations during training. Specifically, random additive Gaussian noise ($p = 0.6$), Gaussian blur ($p = 0.4$), and brightness/contrast adjustment ($p = 0.8$) were applied. Projections with >90% failed registrations were removed. The ratio of successful to failed registrations was adjusted from approximately 1:2 to nearly 1:1, effectively reducing class imbalance in the training set.

2.3 Model Architecture

An overview of the model architecture is shown in Fig. 1. We describe both the baseline and proposed models.

Early Fusion and Convolutional Backbone. Both models utilize an early fusion approach where X-ray and DRR images are concatenated along the channel dimension, doubling the first convolution block's channel size. Since the X-ray and DRR are registered onto the same projection plane, early fusion allows the model to learn local cross-modal features from the first layer. A convolutional backbone learns spatially localized alignment features between the X-ray and its corresponding DRR. Each convolution block consists of two convolutional layers (3×3 kernels) with GELU activation, max pooling (2×2 kernels with stride 2),

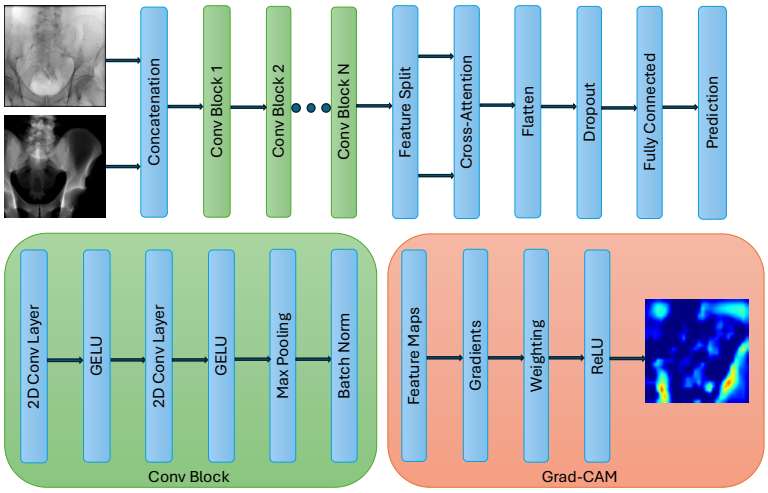


Fig. 1. Overview of proposed model architecture. X-ray and DRR images undergo early fusion (channel concatenation) and processing through a convolutional backbone. Feature channels are split into two groups for bidirectional spatial cross-attention. Grad-CAM visualizes influential regions for interpretability.

and batch normalization. The backbone comprises five convolution blocks in total (one operating on the channel-concatenated input followed by four standard blocks).

Baseline Model. The baseline processes backbone features by flattening, followed by dropout (0.3) and a fully connected layer for classification. While this preserves spatial information, it lacks explicit mechanisms for modeling long-range dependencies between distant spatial locations.

Proposed Model: Spatial Cross-Attention. To capture global patterns of misalignment, we incorporate spatial cross-attention after the convolutional backbone. The feature map is split channel-wise into two groups, each reshaped into a sequence of spatial tokens. Bidirectional attention enables each spatial location in one group to attend to all locations in the other, capturing long-range spatial dependencies. The cross-attended features are averaged, flattened, and processed through dropout (0.3) and a fully connected layer for classification.

2.4 Explainability Mechanisms

Grad-CAM. To visualize the spatial locations of the X-ray images that are most important to the model's predictions, Grad-CAM [12] is used (Fig. 1). Hooks are set at the last convolutional layer in the backbone to capture feature maps and gradients flowing back to this layer. The last convolutional layer is

used since it is the final point in the model that preserves spatial structure while containing high-level features most relevant to the model's prediction. During inference time, for each X-ray and DRR image input, we spatially average the gradients to obtain channel weights, compute the weighted sum of the feature maps, and apply ReLU to show only positive contributions.

Conformal Prediction. Conformal prediction is used to obtain statistical insights about the model's prediction. For each LOOCV fold, we compute nonconformity scores on the validation (calibration) set measuring the difference between prediction and ground truth and choose the $1 - \alpha$ quantile as the threshold. We set $\alpha = 0.1$ which amounts to a 90% quantile. Prediction sets are constructed on test examples by including all possible labels (accepted or rejected registration) that yield nonconformity scores less than or equal to the threshold. A single label indicates the model's certainty whereas two labels indicates uncertainty.

2.5 Training Details

The training procedure was optimized using Optuna [1], which determined the optimal hyperparameters: learning rate 2×10^{-4}, weight decay 5×10^{-5}, and batch size 16. Binary cross-entropy was used as the loss function. Training used early stopping (patience 10) on validation loss. The model was implemented in PyTorch 2.3.1 and trained on an NVIDIA Titan RTX GPU.

3 Experiments and Results

3.1 Experimental Setup

We conducted an ablation study examining two factors:

- **Architecture**: Baseline (early fusion CNN) vs. Proposed (with spatial cross-attention)
- **Training data**: Real projections only vs. Real + Simulated projections

This 2×2 design isolates contributions of each component. We report accuracy, sensitivity, precision, and specificity as mean ± SD across five LOOCV folds. Test sets retain natural class distributions (25–48% successful registrations), reflecting specimen-specific variation in registration difficulty.

3.2 Classification Performance

Table 1 presents the ablation study examining contributions of spatial cross-attention and simulated training data.

Adding simulated projections improved accuracy for both architectures: from 0.758 to 0.777 for the baseline (+2.5%) and from 0.772 to 0.801 for the proposed model (+3.8%). The larger improvement for the proposed model suggests that

Table 1. Ablation study: effect of spatial cross-attention and simulated training data. Mean ± SD across 5 LOOCV folds. Best results per metric in **bold**.

Model	Accuracy	Sensitivity	Precision	Specificity
$\text{Baseline}_{\text{real}}$	0.758 ± 0.092	0.839 ± 0.189	0.611 ± 0.158	0.715 ± 0.175
$\text{Baseline}_{\text{real+sim}}$	0.777 ± 0.135	$\mathbf{0.880 \pm 0.109}$	0.649 ± 0.186	0.747 ± 0.211
$\text{Proposed}_{\text{real}}$	0.772 ± 0.071	0.683 ± 0.277	0.658 ± 0.144	0.802 ± 0.163
$\text{Proposed}_{\text{real+sim}}$	$\mathbf{0.801 \pm 0.030}$	0.766 ± 0.258	$\mathbf{0.688 \pm 0.133}$	$\mathbf{0.806 \pm 0.130}$

spatial cross-attention benefits more from increased training data diversity, enabling it to learn more effective long-range spatial relationships. Cross-attention showed minimal impact with real data alone (accuracy: 0.772 vs. 0.758), but provided consistent improvement when combined with simulated data (accuracy: 0.801 vs. 0.777). This interaction suggests that cross-attention requires sufficient training data to learn effective global spatial dependencies.

Across all conditions, the baseline favored sensitivity (0.839–0.880) while the proposed model favored specificity (0.802–0.806) and precision (0.658–0.688). This reflects different operating characteristics: the baseline more readily accepts registrations, while the proposed model adopts a conservative decision boundary. For safety-critical surgical applications where false acceptance of misaligned registrations poses greater patient risk than false rejection, the proposed model's behavior is clinically preferable. The full proposed approach (spatial cross-attention + simulated data) achieved the highest accuracy (0.801), precision (0.688), and specificity (0.806) with the lowest variance, indicating complementary benefits from architectural innovation and data augmentation.

3.3 Explainability Analysis

To illustrate model behavior beyond aggregate metrics, Figure 2 presents representative cases with Grad-CAM visualizations and conformal prediction outputs. Grad-CAM highlights regions the model focuses on when making predictions. Conformal prediction outputs provide complementary information: "Certain" indicates single-label prediction sets, while "Uncertain" flags cases where the prediction set contains both labels.

4 Discussion

Our results demonstrate that learning-based quality assessment for 2D/3D registration benefits from both architectural innovations and data augmentation strategies. The spatial cross-attention mechanism, when combined with sufficient training data, captures global misalignment patterns by enabling each spatial location to attend to all others, going beyond the local receptive fields of convolutional layers alone.

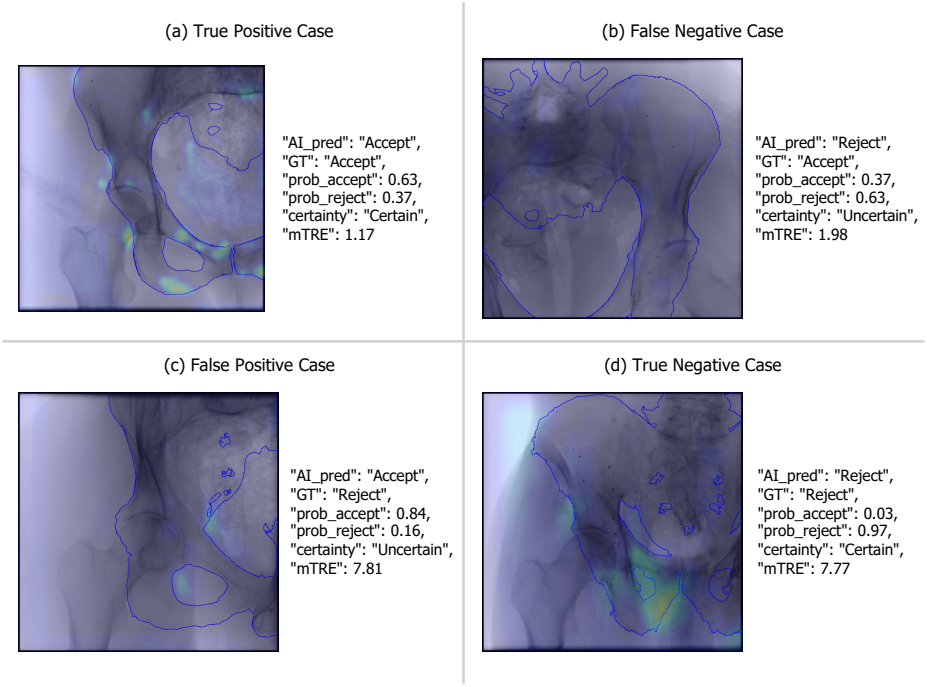


Fig. 2. Representative cases across prediction outcomes. Each section shows Grad-CAM heatmaps overlaid on the X-ray image, registration overlay, the model's prediction (Accept/Reject) with associated probabilities, conformal prediction certainty, and ground truth mTRE. (a) True Positive: correct acceptance of successful registration. (b) False Negative: incorrect rejection of successful registration. (c) False Positive: incorrect acceptance of failed registration. (d) True Negative: correct rejection of failed registration.

The sensitivity-specificity trade-off between architectures reflects fundamentally different operating characteristics. The baseline's higher sensitivity comes at the cost of more false acceptances, while the proposed model's conservative decision boundary prioritizes avoiding false acceptances. For safety-critical surgical applications where accepting a misaligned registration poses greater patient risk than rejecting a correct one, the proposed model's behavior is clinically preferable. The explainability outputs (Grad-CAM and conformal prediction) provide interpretable information that could support human oversight. Its higher specificity is particularly notable, as rejected cases can prompt re-registration while accepted misalignments can lead to surgical error.

The achieved accuracy of 80.1% represents a meaningful step toward learning-based, automated quality assurance, suitable for decision support rather than fully autonomous deployment. We envision this framework providing a "second opinion" for operators, with Grad-CAM highlighting regions influencing the model's prediction and conformal prediction flagging uncertain cases. To-

gether, these explainability mechanisms offer complementary information that could support human oversight.

Finally, we used a 2mm threshold to define successful registration, consistent with standards in orthopedic surgery [9], but clinically acceptable margins are application-specific and may vary for different procedures [11]. Performance and conclusions may shift with different thresholds, and future work could explore threshold-adaptable methods. Similarly, while explainability mechanisms provide interpretable outputs, their utility for human decision-making remains to be validated through user studies.

5 Conclusion

In conclusion, we presented a learning-based framework for 2D/3D registration quality assessment that combines early fusion for local cross-modal features with spatial cross-attention for global context. Systematic ablation demonstrated complementary benefits from cross-attention and simulated data augmentation, with the combined approach achieving the highest accuracy (0.801) and precision (0.688) with the lowest variance. Integrated Grad-CAM and conformal prediction provide interpretable rationales supporting human oversight. Future work includes multi-anatomy evaluation, regression to continuous mTRE, integration with iterative registration refinement, and prospective clinical validation. This framework establishes a foundation for explainable, AI-assisted quality assurance in image-guided surgery.

Acknowledgments. This study was supported in part by Johns Hopkins Internal Funds and by a Google Research Scholar Award.

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

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)
2. Cho, S.M., Do, A., Grupp, R., Armand, M., Taylor, R., Unberath, M.: Uncertainty quantification in image-based 2d/3d registration and its relationship with accuracy. IJCARS pp. 1–9 (2025)
3. Cho, S.M., Grupp, R.B., Gomez, C., Gupta, I., Armand, M., Osgood, G., Taylor, R.H., Unberath, M.: Visualization in 2d/3d registration matters for assuring technology-assisted image-guided surgery. IJCARS **18**(6), 1017–1024 (2023)
4. Cho, S.M., Wu, W., Kilmer, E., Taylor, R.H., Unberath, M.: Feeling the stakes: Realism and ecological validity in user research for computer-assisted interventions. In: MICCAI. pp. 189–197. Springer (2025)

5. Fitzpatrick, J.M.: The role of registration in accurate surgical guidance. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine **224**(5), 607–622 (2010)
6. Galib, S.M., Lee, H.K., Guy, C.L., Riblett, M.J., Hugo, G.D.: A fast and scalable method for quality assurance of deformable image registration on lung ct scans using convolutional neural networks. Medical physics **47**(1), 99–109 (2020)
7. Grupp, R., Unberath, M., Gao, C., Hegeman, R., Murphy, R., Alexander, C., Otake, Y., McArthur, B., Armand, M., Taylor, R.: Data and code associated with the publication: Automatic Annotation of Hip Anatomy in Fluoroscopy for Robust and Efficient 2D/3D Registration (2020). https://doi.org/10.7281/T1/IFSXNV
8. 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. Springer (2018)
9. Van de Kraats, E.B., Penney, G.P., Tomaževič, D., van Walsum, T., Niessen, W.J.: Standardized evaluation of 2d-3d registration. In: MICCAI. pp. 574–581. Springer (2004)
10. Markelj, P., Tomaževič, D., Likar, B., Pernuš, F.: A review of 3d/2d registration methods for image-guided interventions. Medical image analysis **16**(3), 642–661 (2012)
11. Rampersaud, Y.R., Simon, D.A., Foley, K.T.: Accuracy requirements for image-guided spinal pedicle screw placement. Spine **26**(4), 352–359 (2001)
12. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
13. Tummala, S., Thadikemalla, V.S.G., Kreilkamp, B.A., Dam, E.B., Focke, N.K.: Fully automated quality control of rigid and affine registrations of t1w and t2w mri in big data using machine learning. Computers in biology and medicine **139**, 104997 (2021)
14. Unberath, M., Zaech, J.N., Lee, S.C., Bier, B., Fotouhi, J., Armand, M., Navab, N.: Deepdrr–a catalyst for machine learning in fluoroscopy-guided procedures. In: MICCAI. pp. 98–106. Springer (2018)
15. Varnavas, A., Carrell, T., Penney, G.: Fully automated 2d–3d registration and verification. Medical image analysis **26**(1), 108–119 (2015)