Landmark Detection Uncertainty as a Reliability Weight for Robust Landmark-based 2D/3D Pelvic Pose Estimation

Yehyun Suh[1,2,3] YEHYUN.SUH@VANDERBILT.EDU
[1] *Department of Computer Science, Vanderbilt University, Nashville, TN, USA*
[2] *Vanderbilt Institute of Surgery and Engineering, Nashville, TN, USA*
[3] *Vanderbilt Lab for Immersive AI Translation, Nashville, TN, USA*

Brayden Schott[1,2,3] BRAYDEN.J.SCHOTT@VANDERBILT.EDU
Chuo Mo[4] ATHENAMO@G.UCLA.EDU
[4] *Department of Mathematics, University of California-Los Angeles, Los Angeles, CA, USA*

J. Ryan Martin[5] JOHN.MARTIN@VUMC.ORG
[5] *Department of Orthopaedic Surgery, Vanderbilt University Medical Center, Nashville, TN, USA*

Daniel Moyer*[1,2,3] DANIEL.MOYER@VANDERBILT.EDU

Editors: Under Review for MIDL 2026

Abstract

Landmark-based 2D/3D pelvis registration is vulnerable to noisy or ambiguous landmark detections in fluoroscopy, which can destabilize downstream pose estimation. We present an uncertainty-aware registration framework that models epistemic uncertainty in predicted landmarks and incorporates it directly into the Perspective-n-Point formulation. Using Monte Carlo dropout within a U-Net detector, we compute per-landmark reliability per sample using the variance of multiple stochastic forward passes. These reliability estimates guide two complementary strategies: continuous weighting, which integrates uncertainty into a weighted PnP optimization, and discrete selection, which removes the most uncertain landmarks during inference. We evaluate the framework on synthetic fluoroscopy derived from a public pelvic CT dataset. Our experiments show that uncertainty provides a principled mechanism for identifying unreliable landmarks and stabilizing pose estimation, enabling more robust 2D/3D registration and establishing a foundation for uncertainty-guided image-guided surgical workflows.

Keywords: Uncertainty-Weighted Pose Estimation, Landmark-based 2D/3D Registration, Monte Carlo Dropout, Epistemic Uncertainty Modeling

1. Introduction

2D/3D registration is an optimization task that aligns a 2D image with a 3D volume in a common spatial reference frame (Grupp et al., 2018; Unberath et al., 2021). In typical workflows, an X-ray or fluoroscopy image is registered to a CT scan so that 3D anatomical information compensates for the loss of depth and perspective in 2D projections. This paradigm is widely used in image-guided orthopaedic, spine, trauma, and vascular procedures that require precise localization of anatomy, tools, and implants, where fast but

* Corresponding Author
Code: https://github.com/yehyunsuh/Landmark-based-2D-3D-Registration-Uncertainty

Method	Mean Runtime (s)	Median Rot. Error (deg)	Median Trans. Error (mm)
Intensity (512×512 px^2)	95	54.98	27.58
Intensity (100×100 px^2)	6.2	70.41	33.49
Landmark+PnP [Baseline]	0.1	12.96	32.70
Weighted Landmark+PnP [**Prop.**]	0.9	2.73	6.97

Table 1: Comparison of intensity- and landmark-based methods, evaluated in terms of mean total registration time and median rotation and translation error. The Intensity method is the DiffDRR ((Gopalakrishnan and Golland, 2022)) projection metric with varying 2D image sizes, while Landmark+PnP use a U-Net landmark annotator fed into a direct pose optimization, either with or without our proposed weights (c.f. Finetune + Test Time CW in Table 2). The weights are estimated with MC dropout ($S = 100$).

depth-poor intra-operative fluoroscopy is complemented by registration to restore 3D context using standard operating room equipment (Cho et al., 2023).

Current methodology is divided into two broad classes of methods, image intensity matching (Unberath et al., 2018; Gao et al., 2020; Gopalakrishnan and Golland, 2022) and landmark matching (Gao et al., 2003; Lepetit et al., 2009; Li et al., 2012). The former uses a forward model of projection and iteratively updates beliefs about the detectors relative pose by matching that projection to the observed images. While this has the potential to have high accuracy and generality across anatomy, each forward pass is generally computationally expensive and thus often slow for bed-side applications. Landmark methods instead rely on anatomic knowledge of the target volume, and match pre-defined features or landmarks between the 2D and 3D sets. While this is much more computationally tractable, avoiding the reprojection steps of intensity matching, it is prone to higher errors due to the sensitivity of the point matching operation, and due to the intrinsic variability of anatomic landmarks.

In the present work we propose an uncertainty-aware framework that models the reliability of each anatomical point during the landmark identification phase, and includes that estimate as an optimization weight in the subsequent pose estimation phase. By integrating per-landmark uncertainty into a fully differentiable landmark detection and Perspective-n-Points (PnP) pipeline, our method stabilizes pose estimation by increasing the influence of trustworthy keypoints and suppressing unreliable ones. Our contributions are threefold: (1) we introduce a differentiable uncertainty-to-weight formulation that enables continuous weighting during training and inference of Landmark-PnP pose estimation schemes; (2) we show that our selection strategies improves robustness even without requiring retraining; and (3) we provide empirical evidence that uncertainty estimates landmark reliability, yielding substantially improved 2D/3D pelvis registration performance.

2. Related Work

There are multiple existing approaches for rigid 2D/3D registration of radiography or fluoroscopy to volumetric CT. Most relevant to our work are landmark- and feature-based

methods, which assume correspondence between 3D anatomical landmarks in CT and their 2D projections (Bier et al., 2018; Grupp et al., 2020). These are the x-ray/fluoroscopy case of the Perspective-n-Point problem from general imaging (Gao et al., 2003; Lepetit et al., 2009; Li et al., 2012). We choose to solve this optimization using gradient based methods due to their relative simplicity and apparent quality for our domain.

Another broad class of 2D/3D registration methods are based on matching image intensity. These intensity-based methods align digitally reconstructed radiographs (DRRs) (Unberath et al., 2018; Gao et al., 2020; Gopalakrishnan and Golland, 2022) generated from CT with intra-operative fluoroscopy by optimizing image similarity measures such as correlation or information based metrics (Gopalakrishnan et al., 2024). These methods are general in the sense that they do not need outside knowledge about the content of the images, but scale poorly in the size of the images, both 2D and 3D, leading to long run times and heavy computational costs (see Table 1).

Uncertainty estimation in deep learning and uncertainty-aware architectures are relatively well studied. In this paper we use one of the early approaches, Monte Carlo (MC) dropout, where dropout layers (Srivastava et al., 2014) are kept active at test time and multiple stochastic forward passes are used to approximate epistemic uncertainty via the variance of the predictions (Gal and Ghahramani, 2016; Kendall and Gal, 2017). In medical image analysis, this idea has been applied to segmentation and landmark detection to highlight regions where the networks is less confident and to guide human review or post processing (Jungo et al., 2018; Drevickỳ and Kodym, 2020; Ye et al., 2023).

More complex uncertainty estimators are possible, but often do not fit our use criterion for downstream weighting. Ensemble methods (Rahaman et al., 2021) have been proposed as a Dropout generalization, as each Dropout iteration is often viewed as an ad hoc bootstrapped ensemble, but these require retraining and multiple network evaluations (beyond randomly sampled masks), and so they do not fit our use-case. Another family of methods is the conformal prediction (Shafer and Vovk, 2008) framework, where "conformance scores" effectively rate datapoints as out-liers or in-liers, allowing classification or regression to split its operating characteristic curves into a geometric product. Our method is not a direct classification, and conformance scores are only related to uncertainty by quantile/order; there is no guarantee of magnitude differences being related. Bayesian methods may also model and then sample parameter weights to form posterior distributions of both networks and outputs (Marinescu et al., 2020), but these sampling methods are often slow, and again require numerous network evaluations.

3. Method

3.1. Problem Setup and Groundwork

We would like to compute the pose of a patient observed under 2D fluoroscopy or radiograph relative to a standardized 3D pose, for a fixed detector (x-ray camera) position; this is equivalent to finding the pose of a detector which is observing a patient in standard position. The search space is over all 3D pose parameters $\theta = (r, t)$, composed of a rotation r and a translation t. This generally has 6 degrees of freedom, 3 dof for r and t each, even though r may be represented in overparameterized fashion, i.e., by a 3×3 matrix or a 4 dimensional quaternion.

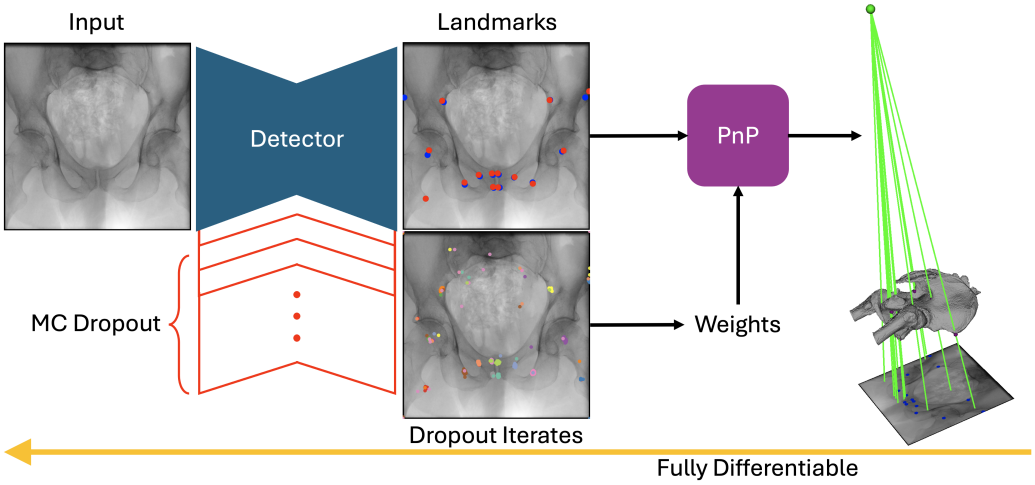

Figure 1: Overview of the uncertainty-aware pose estimation framework. Monte Carlo (MC) Dropout is used to produce uncertainty estimates (in Red, below Detector) from the primary landmarking network (in Blue, Detector), which then weight landmarks during the Perspective-n-Point (PnP) optimization (in Purple). The PnP method has no learnable parameters, and is fully differentiable, allowing registration losses to be propagated directly back to the landmarking network.

For general target volumes and their corresponding 2D images, image matching methods can be employed to efficiently search this space (Grupp et al., 2020; Gopalakrishnan et al., 2025), but these methods usually disregards any special knowledge of the target image domain (continued discussion in Section 6 and (Gao et al., 2020)). In surgical imaging we often know much more about the target region's anatomy and potential keypoints/features. Assuming we have corresponding features, we can avoid the image matching problem and instead solve a 2D/3D point-set registration. Using a fixed detector convention, we write the point correspondence problem as the following least squares problem between landmarks in 3D (p^{3D}) and their apparent 2D positions in the image (p^{2D}):

$$\min_{\theta} \sum_{i}^{L} \| \text{Proj}[T_\theta(p_i^{3D})] - p_i^{2D} \|_2^2. \quad \text{PnP Least Squares Problem} \tag{1}$$

Here, Proj is the operator that takes 3D points to their 2D positions in the camera. The least squares 2D/3D point alignment problem has been solved in the literature by the Perspective-n-Point (PnP) family of methods (Lepetit et al., 2009; Terzakis and Lourakis, 2020). We use a gradient based optimization which in practice converges to the correct solution. Notably this optimization is itself fully differentiable (Amos and Kolter, 2017), as are, generally speaking, most of the class of PnP solvers.

For landmark prediction, we employ a U-Net-based convolutional neural network to generate landmark probability maps from fluoroscopy images (Ronneberger et al., 2015), as shown in Figure 1. During inference, the coordinate for each landmark is identified as the pixel location with the maximum intensity in the corresponding predicted heatmap. For

image inputs I and neural network f parameterized by ϕ, we write this operation as:

$$p_i^{\text{2D}} = \underset{x \in \Omega}{\text{argmax}} \left([f(I; \phi)]_i \right). \tag{2}$$

Here Ω is the 2D image domain in which point x is contained. The i^{th} channel output corresponds to the i^{th} landmark.

Our specific architecture incorporates a ResNet-101 encoder (He et al., 2016) initialized with ImageNet-pretrained weights (Deng et al., 2009). To improve generalization, we utilize a dilation-erosion label augmentation scheme, which has demonstrated efficacy in orthopaedic datasets by broadening the effective training signal (Suh et al., 2023; Chan et al., 2025). The model is trained to predict these augmented heatmaps using a binary cross-entropy loss. As demonstrated in (Mo et al., 2025), because the PnP operation is differentiable, we can add a weighted PnP loss directly to the binary cross-entropy loss to improve landmark identification.

3.2. Main Contribution: Uncertainty Weighted PnP and PnP Losses

Similar to other least squares estimation problems, PnP methods are notoriously susceptible to outliers; one solution to this general problem are uncertainty weighted least squares solutions (Hastie, 2009). We introduce this same solution for our PnP solver. Modifying Eq. 1, we construct a weighted least squares PnP problem by adding weights w_i to each of the point error terms:

$$\min_{\theta} \sum_{i}^{L} w_i \| \text{Proj}[T_\theta(p_i^{\text{3D}})] - p_i^{\text{2D}} \|_2^2. \quad \text{PnP Weighted Least Squares Problem} \tag{3}$$

The w_i should be set to values that are proportional to the "trustworthiness" of each point, or equivalently inversely proportional to the uncertainty for each point.

As the points are selected by neural network, it is thus natural to use a neural network uncertainty method to estimate this quantity. We estimate w_i using Dropout uncertainty (Gal and Ghahramani, 2016), which prescribes computing Monte Carlo samples of each of the outputs, then measuring uncertainty by summary statistics on those samples; in our context, this means we should compute $p_{i,s}$ using

$$p_{i,s}^{\text{2D}} = \underset{x \in \Omega}{\text{argmax}} \left([f(I, m_s \circ \phi)]_i \right). \tag{4}$$

where m_s is the dropout mask for sample s (Kendall and Gal, 2017). We do this a total of S times for different m_s masks; for Pytorch implementations, these masks are efficiently sampled in memory, so that batched outputs see significant parallelism gains, computing all S samples at the same time, or as many as memory allows.

We then compute summary statistics $\bar{p}_i$ and u_i for each of the i landmarks:

$$\bar{p}_i = \frac{1}{S} \sum_{s=1}^{S} p_{i,s}^{\text{2D}} \qquad u_i = \sqrt{\frac{1}{S} \sum_{s=1}^{S} \| p_{i,s}^{\text{2D}} - \bar{p}_i \|_2^2}. \tag{5}$$

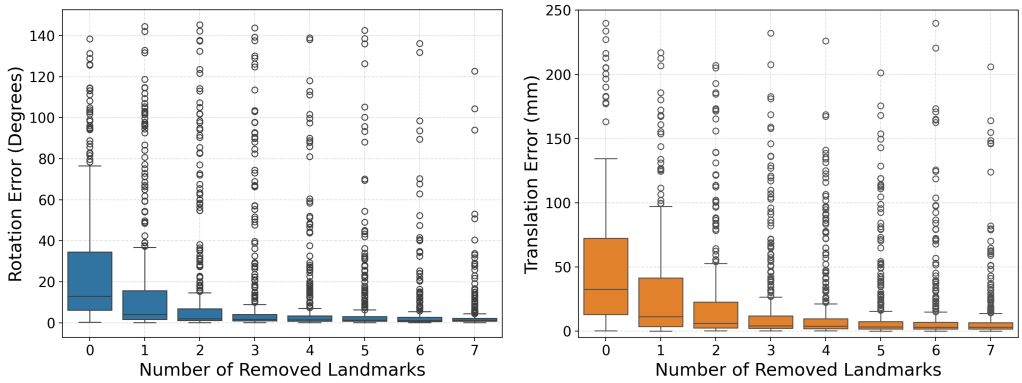

Figure 2: Oracle experiment using ground-truth 2D landmark positions. The boxplots show the rotation error and translation error as a function of number of removed landmarks. The boxplots from left to right correspond to oracle landmark filtering levels $K = 0, 1, \ldots, 7$, i.e., removing the K most erroneous landmarks.

Statistical theory (Kendall et al., 2018) suggests that our weights should be inversely proportional to their uncertainty. Whether due to inaccuracy while computing the weights, or due to other deviations, we find that it is more numerically stable to normalize and then negatively exponentiate the weights:

$$\tilde{u}_i = \frac{u_i}{\max_{i'} u_{i'} + \varepsilon} \qquad w_i = \exp\bigl(-\beta \tilde{u}_i\bigr), \tag{6}$$

with β as a hyper-parameter controlling weight "fall-off", which will correspond to outlier suppression strength in the resulting optimization of Eq. 3. For numerical stability of the optimization we also normalize w_i after this procedure. We refer to solutions of Eq. 3 as *continuous weighting*.

Implementation considerations: We can implement this scheme directly in Pytorch *both for inference and training*. The Dropout statistics themselves are composed of fully differentiable operations, as are the weight constructions and the weighted PnP optimization. However, for completely untrained networks, these weights will likely be nearly uniform (i.e., not informative). Thus, we implement using a finetuning scheme, where a network is trained to output landmarks first, before being refined by weighted PnP losses.

In backpropagating from the weighted PnP to the primary network, we need to propagate through all of our Dropout iterates. This requires a number of network activations to be held in memory that is equal to the dropout iterates; to avoid this cost, we could choose to exclude the dropout from the backpropagation. This would lead to inaccuracies in the gradient, but as we show in Table 2, this only leads to small overall performance degradation for significant memory overhead reduction.

Test-time filtering and discrete selection: We find that instead of performing weighted squares, wholly excluding low weight landmarks from the optimization empirically produces strong performance. Ranking landmarks by uncertainty, we define an uncertainty-filtered subset by discarding the K most uncertain visible landmarks. This method we

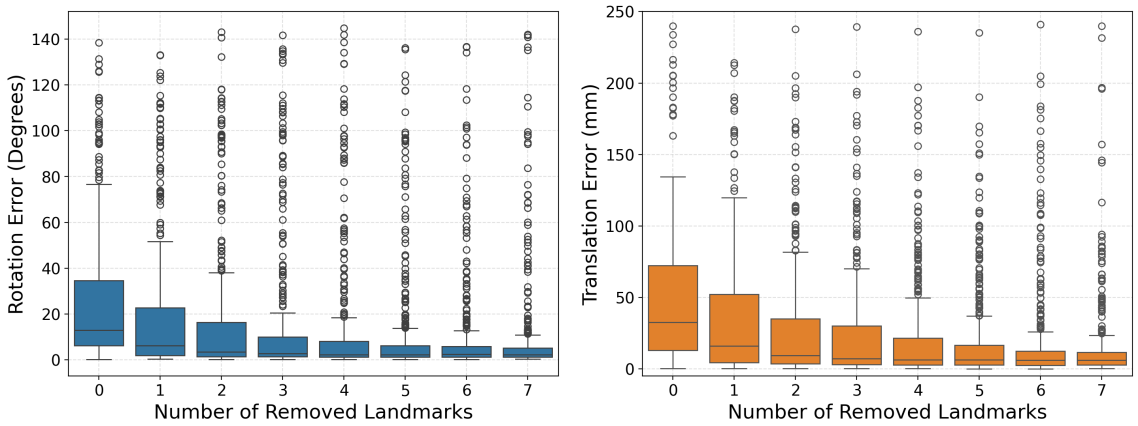

Figure 3: Rotation and translation errors across landmark dropout iterations ($K = 0, \ldots, 7$). Each boxplot summarizes the per-image error for a given dropout number K after aggregating results across all patients. The translation axis is truncated at 250 mm for outliers exceeding this threshold in $K = 0, 1, 2, 3, 5$ (1.39%), $K = 4$ (1.11%), $K = 6$ (0.83%), and $K = 7$ (0.28%).

call *discrete selection*; while it is not easily optimized over, at test time it provides good performance as shown in Section 5.

4. Experiments

We validated our approach using the public CT dataset (Grupp et al., 2020), which provides paired 3D anatomical landmarks and CT volumes (see Appendix A). Using these volumes, we synthesized DRRs via DiffDRR (Gopalakrishnan and Golland, 2022). The imaging geometry was standardized with a source-to-detector distance of 1020 mm and a volume-to-detector distance of 400 mm. To simulate diverse patient positioning, we randomized camera poses (pelvic poses) with rotations drawn from $[-45°, 45°]$ along the x- and y-axes and $[-15°, 15°]$ along the z-axis. Translations were sampled independently from $[-50, 50]$ mm along each axis. For every rendered DRR (512×512), the 14 3D landmarks were projected to generate ground-truth 2D labels, forming the basis for detection and registration tasks.

We employed a leave-one-subject-out cross-validation strategy, holding out one volumetric image and all associated 2D images as a test set while training and fine-tuning on the remaining subjects. To quantify uncertainty, we incorporated MC dropout ($p = 0.1$) within the decoder. To estimate the uncertainty, the model was evaluated S times per image ($S = 40$ for fine-tuning, $S = 100$ for testing) to generate a distribution of sample predictions $\{p_{i,s}\}_{s=1}^{S}$ for each landmark i.

To assess the benefit of removing erroneous landmarks, we conducted an oracle experiment (Figure 2). For each landmark i, we compute the oracle detection error as $d_i = \|p_i^{\text{2D}} - p_i^*\|_2$ and filtered out the top-K highest-error landmarks to form an oracle filter w_i^{gt}. Excluding these landmarks consistently improved rotation and translation accuracy, motivating our use of uncertainty u_i as a proxy for the unknown error d_i.

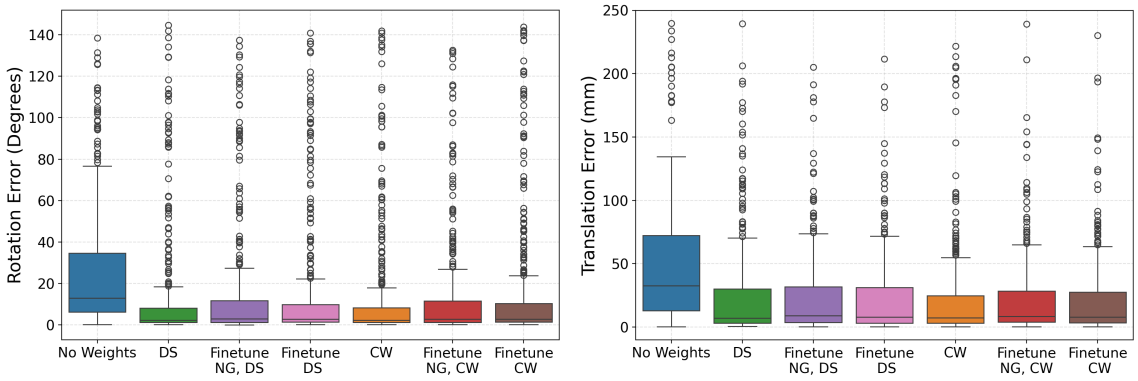

Figure 4: Rotation and translation error distributions comparing No Weights, Discrete Selection (DS, top-3 landmark filtering) and Continuous Weighting (CW). Variants include finetuning without gradient updates on MC dropout model (Finetune NG) and fully finetuned model (Finetune). The translation axis is truncated at 250 mm for outliers. In the translation boxplot, from left to right, 1.11%, 0.83%, 0.28%, 0.28%, 0.83%, 0.28%, 0.28% of the datapoints were truncated.

We evaluated pose estimation performance using the Euler angle difference (degrees) for rotation and root mean squared error (RMSE, mm) for translation. To assess our framework, we designed four experimental stages. We begin by quantifying registration stability by applying uncertainty based top-K filtering ($K = 0, \ldots, 7$). Next, we compare our uncertainty weight based fine-tuned model against a baseline that treats all landmarks equally as well as test time only weighting methods. We then analyze per-landmark uncertainty to show which anatomical regions benefit from uncertainty-aware handling. Finally, we perform an error retention analysis by progressively excluding high-uncertainty images to validate the effectiveness of the metric as an outlier detector.

5. Results

We evaluated the efficacy of the estimated uncertainty as a criterion for outlier rejection, independent of ground-truth labels. Figure 3 illustrates the distributions of rotation and translation errors as we progressively exclude out the top-K most uncertain landmarks ($K = 0, \ldots, 7$). We observed a sharp, monotonic decay in both error and interquartile range (IQR) as K increases. This suggests that the uncertainty metric u_i effectively isolates the long-tail outliers that disproportionately destabilize the registration solver. By removing these high-uncertainty points, the system recovers a geometrically consistent subset of landmarks, resulting in precise pose estimation even in the presence of detection noise.

Figure 4 and Table 2 present a 3D pose estimation performance across seven experimental configurations. Experiment that uses all landmarks for the pose estimation with no weights shows high variance with frequent outliers, resulting in a mean rotation error of 26.14 degrees and translation error of 51.18 mm. Introducing uncertainty-based discrete selection (DS) in inference nearly halves both errors (14.22 degrees, 24.79 mm), and fine-tuning

Experiment	Rotation Error (degrees)		Translation Error (mm)	
	Mean	Median	Mean	Median
No Weights (Grupp et al., 2020)	26.14	12.96	51.18	32.70
DS	14.22	2.31	24.79	6.38
Finetune + NG + DS	16.52	2.87	21.98	7.46
Finetune + DS	16.33	2.84	21.92	7.21
CW	13.94	2.27	24.18	6.35
Finetune + NG + CW	15.65	2.73	22.14	7.44
Finetune + CW	15.84	2.73	20.63	6.97

Table 2: Quantitative analysis of pelvic pose estimation comparing No Weights, Discrete Selection (DS, top-3 landmark filtering) and Continuous Weighting (CW). Variants include finetuning without gradient updates on MC dropout model (Finetune NG) and fully finetuned model (Finetune). We report the mean and median Euler angle difference for rotation (degrees) and RMSE for translation (mm).

with continuous weighting (CW) maintains similar performance with lower translation error (21.92 mm). CW applied directly in inference achieves the lowest mean rotation error (13.94 degrees), while the combined strategy with fine-tuning yields the best overall translation accuracy (20.63 mm). Notably, median rotation errors drop from 12.96 degrees in the *No Weights* to 2.27–2.87 degrees with uncertainty-based methods, and median translation errors from 32.70 mm to below 8 mm across all proposed configurations.

Figure 5 presents error retention curves showing a monotonic reduction in residual error as samples with high uncertainty, defined as the mean spatial deviation across all the landmarks in the image, are progressively excluded. In intra-operative guidance, this facilitates graceful failure allowing the system to withhold prediction on ambiguous frames rather than outputting misleading guidance. Therefore, clinical workflows can strategically trade off temporal for reliability to ensure surgical decision making to be informed exclusively by high confidence pose estimates.

6. Discussion

Our results demonstrate that integrating epistemic uncertainty, through both continuous weighting during training and inference and discrete selection at inference, significantly improves registration accuracy. However, this study highlights two primary limitations regarding the definition of ground truth and the parameterization of the uncertainty.

Specifically, because anatomical landmarks often represent abstract geometric constructs rather than visually distinct features, human annotation inherently contains subjective variability. Consequently, high model uncertainty frequently captures this ill-defined nature rather than a failure to learn, justifying our strategy of down-weighting ambiguous points to prevent overfitting. However, the current mechanisms for this weighting and filtering rely on fixed hyperparameters (β and K), which do not account for the significant variance in clinical image quality. Because static thresholds may inadvertently suppress useful gradients or retain noise depending on the acquisition, an optimal deployment requires dynamic

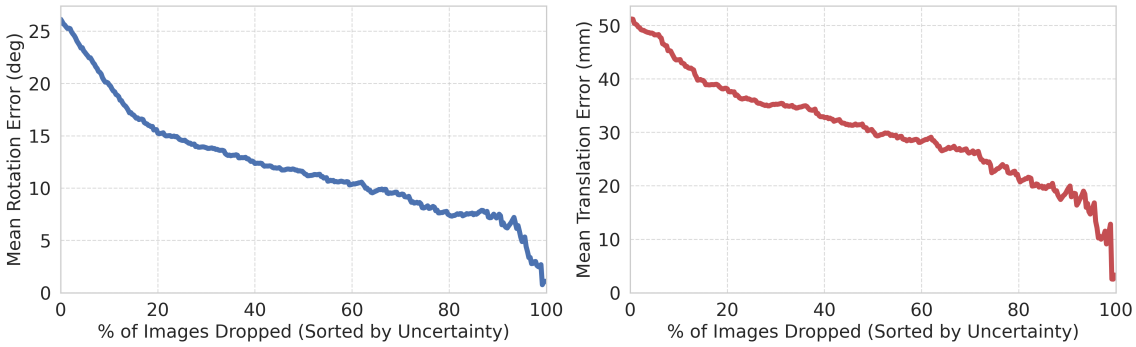

Figure 5: Mean error as samples are removed in order of their estimated uncertainty. The curves show the reduction in mean rotation (left) and translation (right) errors as the most uncertain samples are progressively filtered out (0% to 100%).

parameterization, where these values adaptively shift according to the specific noise profile of the intra-operative image.

The proposed uncertainty-aware framework opens an opportunity for fully automated, label-free uncertainty estimation. Current methods are bottlenecked by the need for expert annotation of semantically defined anatomical points. Future work could investigate a patient-specific approach where random, geometrically distinct points are sampled automatically from the patient's CT segmentation, replacing manual landmarks. By combining this with our uncertainty estimation, the system could autonomously discover and prioritize the most reliable geometric features for registration without human supervision.

7. Conclusion

We present an uncertainty-aware framework for landmark-based 2D/3D pelvis registration that integrates epistemic uncertainty into both continuous weighting and discrete landmark selection. By estimating per-landmark reliability with MC dropout and incorporating these estimates into the PnP formulation, our method reduces sensitivity to erroneous detections and substantially improves rotational and translational accuracy compared to conventional landmark-based registration. The results demonstrate that uncertainty provides a principled mechanism for identifying unreliable keypoints, stabilizing pose estimation, and enabling graceful failure on ambiguous frames. While remaining limited by the subjective nature of anatomical landmark definitions and fixed hyperparameters, this framework establishes a foundation for adaptive, uncertainty-guided registration strategies and suggests future opportunities for fully automated, label-free geometric feature selection in image-guided surgical workflows.

Acknowledgments

This work was supported in part by NSF 2321684 and a VISE Seed Grant.

References

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International conference on machine learning*, pages 136–145. PMLR, 2017.

Bastian Bier, Mathias Unberath, Jan-Nico Zaech, Javad Fotouhi, Mehran Armand, Greg Osgood, Nassir Navab, and Andreas Maier. X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 55–63. Springer, 2018.

Peter YW Chan, Courtney E Baker, Yehyun Suh, Daniel Moyer, and J Ryan Martin. Development of a deep learning model for automating implant position in total hip arthroplasty. *The Journal of Arthroplasty*, 2025.

Sue Min Cho, Robert B Grupp, Catalina Gomez, Iris Gupta, Mehran Armand, Greg Osgood, Russell H Taylor, and Mathias Unberath. Visualization in 2d/3d registration matters for assuring technology-assisted image-guided surgery. *International journal of computer assisted radiology and surgery*, 18(6):1017–1024, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Dusan Drevickỳ and Oldrich Kodym. Evaluating deep learning uncertainty measures in cephalometric landmark localization. In *BIOIMAGING*, pages 213–220, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Cong Gao, Xingtong Liu, Wenhao Gu, Benjamin Killeen, Mehran Armand, Russell Taylor, and Mathias Unberath. Generalizing spatial transformers to projective geometry with applications to 2d/3d registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 329–339. Springer, 2020.

Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng. Complete solution classification for the perspective-three-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 25(8):930–943, 2003.

Vivek Gopalakrishnan and Polina Golland. Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In *Workshop on Clinical Image-Based Procedures*, pages 1–11. Springer, 2022.

Vivek Gopalakrishnan, Neel Dey, and Polina Golland. Intraoperative 2d/3d image registration via differentiable x-ray rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11662–11672, 2024.

Vivek Gopalakrishnan, Neel Dey, David-Dimitris Chlorogiannis, Andrew Abumoussa, Anna M Larson, Darren B Orbach, Sarah Frisken, and Polina Golland. Rapid patient-specific neural networks for intraoperative x-ray to volume registration. *ArXiv*, pages arXiv–2503, 2025.

Robert B Grupp, Mehran Armand, and Russell H Taylor. Patch-based image similarity for intraoperative 2d/3d pelvis registration during periacetabular osteotomy. In *International Workshop on Computer-Assisted and Robotic Endoscopy*, pages 153–163. Springer, 2018.

Robert B Grupp, Mathias Unberath, Cong Gao, Rachel A Hegeman, Ryan J Murphy, Clayton P Alexander, Yoshito Otake, Benjamin A McArthur, Mehran Armand, and Russell H Taylor. 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.

Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Alain Jungo, Raphael Meier, Ekin Ermis, Marcela Blatti-Moreno, Evelyn Herrmann, Roland Wiest, and Mauricio Reyes. On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 682–690. Springer, 2018.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81(2):155–166, 2009.

Shiqi Li, Chi Xu, and Ming Xie. A robust o (n) solution to the perspective-n-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1444–1450, 2012.

Razvan V Marinescu, Daniel Moyer, and Polina Golland. Bayesian image reconstruction using deep generative models. *arXiv preprint arXiv:2012.04567*, 2020.

Chou Mo, Yehyun Suh, J Ryan Martin, and Daniel Moyer. Enhanced landmark detection model in pelvic fluoroscopy using 2d/3d registration loss. *arXiv preprint arXiv:2511.21575*, 2025.

Rahul Rahaman et al. Uncertainty quantification and deep ensembles. *Advances in neural information processing systems*, 34:20063–20075, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Yehyun Suh, Aleksander Mika, J Ryan Martin, and Daniel Moyer. Dilation-erosion methods for radiograph annotation in total knee replacement. In *Medical Imaging with Deep Learning, short paper track*, 2023.

George Terzakis and Manolis Lourakis. A consistently fast and globally optimal solution to the perspective-n-point problem. In *European Conference on Computer Vision*, pages 478–494. Springer, 2020.

Mathias Unberath, Jan-Nico Zaech, Sing Chun Lee, Bastian Bier, Javad Fotouhi, Mehran Armand, and Nassir Navab. Deepdrr–a catalyst for machine learning in fluoroscopy-guided procedures. In *International conference on medical image computing and computer-assisted intervention*, pages 98–106. Springer, 2018.

Mathias Unberath, Cong Gao, Yicheng Hu, Max Judish, Russell H Taylor, Mehran Armand, and Robert Grupp. The impact of machine learning on 2d/3d registration for image-guided interventions: A systematic review and perspective. *Frontiers in Robotics and AI*, 8:716007, 2021.

Ziyang Ye, Haiyang Yu, and Bin Li. Uncertainty-aware u-net for medical landmark detection. *arXiv preprint arXiv:2303.10349*, 2023.

Appendix A. Data Collection and Preprocessing

A.1. CT Volume and Segmentation Reorientation

The DeepFluoro dataset (Grupp et al., 2020) provides each CT volume with voxel array V, voxel spacing $s = (s_x, s_y, s_z)$, direction matrix $R \in \mathbb{R}^{3\times3}$, and physical origin $o \in \mathbb{R}^3$. The scanner index–to–world affine matrix is

$$A = \begin{bmatrix} R\,\mathrm{diag}(s_x, s_y, s_z) & o \\ 0 & 1 \end{bmatrix}. \tag{7}$$

To match the NIfTI (x, y, z) axis convention, we apply a permutation of voxel indices followed by a reflection. Let $V(i, j, k)$ denote the original voxel array with axes (i, j, k).

First, we permute axes 0 and 2 via

$$\tilde{V}(i,j,k) = V(k,j,i). \tag{8}$$

Next, we flip the new x–axis (dimension i) by reflecting it:

$$V'(i,j,k) = \tilde{V}(N'_x - 1 - i,\ j,\ k), \tag{9}$$

where N'_x is the size of the permuted first dimension. Together, this yields the reoriented volume V' used for all NIfTI exports. This flip corresponds to the matrix

$$F = \begin{bmatrix} -1 & 0 & 0 & (N_z - 1)s_z \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{10}$$

so the final NIfTI affine is

$$A' = AF. \tag{11}$$

This guarantees that (V', A') preserves the original scanner physical coordinates.

A.2. 3D Landmark Conversion

Each anatomical landmark is provided in physical scanner coordinates $p = (x, y, z)^\top$. To map it into the voxel space of the reoriented CT, we apply

$$v = (A')^{-1} \begin{bmatrix} p \\ 1 \end{bmatrix}, \qquad v = (i, j, k). \tag{12}$$

The resulting voxel coordinates are rounded to the nearest integer index.

A.3. 2D Projection Extraction

Projection images are intensity-normalized to $[0, 255]$ and saved as PNG files. The provided 2D landmark coordinates (u, v) are written directly unless they lie outside the image bounds, in which case the corresponding landmark is marked as invisible for the ground truth landmarks.