Segmentation-Informed Landmark Regression for Robust Localization in Fluoroscopic Imaging

David Jozef Hresko[1] DAVID.JOZEF.HRESKO@TUKE.SK

Peter Drotar[1] PETER.DROTAR@TUKE.SK

[1] *Technical University of Kosice, Faculty of Electrical Engineering and Informatics, Department of Computers and Informatics, Kosice, Slovak republic*

Editors: Under Review for MIDL 2026

Abstract

Accurate and robust localization of anatomical landmarks in fluoroscopic images is essential for image-guided interventions, yet remains challenging due to low contrast, noise, and overlapping anatomical structures. In this work, we propose a framework that jointly performs bone structure segmentation and landmark heatmap prediction to improve landmark localization accuracy. Our method is built on an HRNet-based architecture augmented with a custom decoder, enabling high-resolution feature preservation while learning rich spatial context for both tasks. To guide the network toward anatomically consistent predictions, we introduce a composite loss function that integrates dice-sørensen coefficient combined with cross-entropy for segmentation quality, Kullback–Leibler divergence loss for heatmap regression, and a novel anatomical consistency loss that penalizes landmark predictions falling outside their corresponding segmented bone regions. Experiments demonstrate that coupling segmentation with landmark regression significantly improves localization robustness compared to other approaches, resulting in a mean localization error of just 10.25 pixels. The proposed approach provides a reliable foundation for downstream tasks in radiology.

Keywords: landmark detection, semantic segmentation, medical registration, medical imaging

1. Introduction

Accurate localization of anatomical landmarks in fluoroscopic images plays a critical role in numerous image-guided interventions, including orthopedic navigation, catheter placement, and intraoperative motion tracking. Despite its importance, reliably detecting landmarks in fluoroscopy remains a challenging problem (Croci et al., 2024). These images are often characterized by low signal-to-noise ratios, overlapping anatomical structures, motion blur, and limited contrast, all of which hinder traditional image-processing methods as well as modern deep learning–based approaches.

Recent advances in convolutional neural networks have significantly improved landmark detection performance in various medical imaging modalities (Keuth et al., 2025). However, most existing approaches treat landmark localization as an isolated prediction task and fail to incorporate anatomical context that could otherwise guide the model toward more plausible outputs (Yang et al., 2025). In fluoroscopy, where ambiguity is common and structural information is sparse, the absence of such contextual constraints frequently results in inconsistent or anatomically implausible predictions, particularly in the presence of occlusions or artifacts.

To mitigate mentioned issues, we propose an HRNet-based architecture with a custom decoder that preserves high-resolution representations while capturing the spatial context needed to perform bone segmentation and landmark regression. To ensure anatomically consistent outputs, we propose a composite loss function that combines the Dice–Sørensen coefficient with cross-entropy, Kullback–Leibler divergence term and a novel anatomical consistency loss that penalizes landmark predictions that fall outside their corresponding segmented bone regions.

The rest of the paper is organized as follows. In the methodology section we provide detailed data description and the proposed solution is explained. Finally we present the results and discuss different aspects of our solution.

2. Methods

We propose a segmentation-informed approach for anatomical landmark localization in fluoroscopic images. Our method uses an HRNet-based (Wang et al., 2020) architecture with a custom decoder that jointly predicts bone segmentation masks and landmark heatmaps from a single 2D frame. By combining structural segmentation with heatmap regression, the model gains stronger anatomical context, improving robustness in low-quality fluoroscopy. Training is guided by a composite loss that includes segmentation accuracy, heatmap similarity, and an anatomical consistency term ensuring landmarks lie within the correct bone regions. The following sections describe the dataset, model design, and training procedure.

2.1. Dataset

The fluoroscopic dataset (Grupp et al., 2020) used in this study consists of 366 manually verified X-ray images acquired from six cadaveric lower-torso specimens. All images were captured using a Siemens CIOS Fusion mobile C-arm system equipped with a 30 × 30 cm^2 flat-panel detector. The raw fluoroscopic frames have a resolution of 1536 × 1536 pixels with an in-plane pixel spacing of 0.194 mm, followed by a 50-pixel border crop to remove detector artifacts. For the experiments in this paper, we used a downsampled version of the images by a factor of 8, resulting in a final size of 180 × 180 pixels to reduce computational complexity while preserving sufficient anatomical detail for segmentation and landmark regression.

Anatomically, six bone-structure classes are segmented: left hemipelvis, right hemipelvis, left femur, right femur, vertebrae, and upper sacrum. The dataset includes 14 bilateral anatomical landmarks: anterior superior iliac spine (ASIS), femoral head center (FH), greater sciatic notch (GSN), interior obturator foramen (IOF), medial obturator foramen (MOF), superior pubic symphysis (SPS), and inferior pubic symphysis (IPS), each with left and right variants. These 2D landmarks are derived by projecting expert-annotated 3D CT landmarks into the fluoroscopic images.

2.2. Network architecture

For joint bone segmentation and landmark localization, we employ a custom HRNet-based architecture with a multi-head decoder. The network consists of a high-resolution encoder that extracts multi-scale features from the fluoroscopic images. We use the final encoder feature map as input to two parallel decoding branches: one for landmark heatmap regression and one for bone segmentation.

Both decoders share a similar structure, consisting of successive upsampling convolutional layers followed by residual blocks, which progressively recover spatial resolution while refining feature representations. Each residual block contains two convolutional layers with batch normalization and ReLU activation, and includes identity skip connections to improve gradient flow and feature reuse. The final layer of the landmark decoder outputs 14 heatmaps, one per anatomical landmark, whereas the segmentation decoder outputs seven segmentation class logits for all bone structures.

To facilitate landmark training, the landmark heatmaps are passed through a log-softmax activation, enabling the use of KL-divergence loss. Additionally, a visibility vector is computed to enforce anatomical consistency by checking whether each predicted landmark lies within its corresponding segmented bone region. Each element of the vector takes a value of 0 if the landmark is not present, 0.5 if the landmark is predicted but lies outside the correct anatomical segment, and 1 if the landmark is predicted within the correct segmented region.

To evaluate the correctness of the predicted visibility, we employ a weighted mean absolute error loss between the predicted visibility vector $\hat{\mathbf{v}} \in \{0, 0.5, 1\}^{14}$ and the ground-truth visibility vector $\mathbf{v} \in \{0, 1\}^{14}$. False positives (i.e., cases where $\hat{v}_i > 0$ while $v_i = 0$) are penalized more strongly through a weight of 2, while all other elements receive a weight of 1.

Formally, the visibility loss is defined as

$$\mathcal{L}_{\text{vis}} = \frac{1}{B} \sum_{b=1}^{B} \left(\frac{1}{14} \sum_{i=1}^{14} w_{b,i} \, |\hat{v}_{b,i} - v_{b,i}| \right),$$

with element-wise weights

$$w_{b,i} = \begin{cases} 2, & \text{if } v_{b,i} = 0, \\ 1, & \text{if } v_{b,i} = 1. \end{cases}$$

This formulation imposes a higher penalty on incorrect visibility predictions for landmarks that should be absent, promoting greater anatomical consistency in the predicted landmark set. By incorporating this dual-branch design, the network can effectively leverage structural segmentation cues, enhancing both the accuracy and robustness of landmark localization in challenging fluoroscopic images.

2.3. Landmark regression

Landmark localization is performed using a heatmap-based regression approach. For each landmark, the network predicts a heatmap where higher intensity values indicate a higher probability of the landmark being present at that spatial location. To extract the landmark coordinates (x, y) in the image, the heatmap is first passed through a sigmoid activation to normalize its values between 0 and 1, and then the position of the maximum value is taken as the predicted coordinate:

$$(x, y) = \text{argmax } \sigma(H),$$

where H denotes the predicted heatmap for a given landmark and $\sigma(\cdot)$ represents the sigmoid function. This heatmap-based representation allows the network to model uncertainty in landmark positions and provides sub-pixel localization when combined with appropriate post-processing.

2.4. Training and Validation

The model was trained for 1000 epochs using the AdamW optimizer, which decouples weight decay from gradient updates to improve generalization. The learning rate was set to 1×10^{-4} to ensure stable and reliable convergence. In total, six separate models were trained following this procedure. A batch size of 32 was used throughout all experiments. The full dataset was randomly divided into 90% training data and 10% validation data, with validation performed every 100 epochs. For final performance assessment, a leave-one-out specimen protocol was employed: one entire cadaver specimen was excluded from training and validation and used exclusively as a hold-out test set. Landmark localization accuracy was quantified using the Euclidean distance between the predicted and ground-truth coordinates for each landmark. During training, the model achieving the lowest mean Euclidean distance on the validation set was saved as the best checkpoint, ensuring optimal performance for final evaluation on the hold-out specimen.

The optimization objective consisted of three components: a Dice-based segmentation loss $\mathcal{L}_{\text{dice}}$, a KL-divergence loss $\mathcal{L}_{\text{KL}}$ for heatmap regression, and a visibility loss $\mathcal{L}_{\text{vis}}$ enforcing anatomical consistency between predicted landmarks and segmented bone structures. The total loss was defined as:

$$\mathcal{L} = w_{\text{dice}}\,\mathcal{L}_{\text{dice}} + w_{\text{KL}}\,\mathcal{L}_{\text{KL}} + w_{\text{vis}}\,\mathcal{L}_{\text{vis}},$$

Here, $w_{\text{dice}} = 1.0$ determines the contribution of the segmentation loss, $w_{\text{KL}} = 0.1$ controls the influence of the KL-divergence term used for landmark heatmap regression, and $w_{\text{vis}} = 10.0$ specifies the weight of the visibility loss, which penalizes landmarks predicted outside their anatomically valid region. This weighted composition ensures that landmark prediction and segmentation are jointly optimized while strongly enforcing anatomical plausibility.

To improve robustness and reduce overfitting, several data augmentations were applied during training, each with a probability of 0.5. Table 1 summarizes the augmentations and their parameter ranges. In addition to these augmentations, all images were intensity-scaled to a normalized range of $[0, 1]$, and both images and labels were padded to a fixed spatial resolution of 192×192 pixels to ensure consistent input dimensions for the network.

All augmentations were implemented using the MONAI framework (Cardoso et al., 2022), ensuring consistent transformation of images, segmentation masks, and landmark heatmaps. These augmentations simulate realistic variations encountered in fluoroscopic acquisition, improving the model's generalization capability.

Table 1: Data augmentations and their parameter ranges.

Augmentation	**Parameter Range**
RandZoom	1.0 – 1.3
RandRotate	$(-0.17, 0.17)$ rad
RandAdjustContrast	$\gamma \in (0.7, 1.5)$

3. Results

Model performance was evaluated using the Euclidean distance between predicted and ground-truth landmark coordinates. The HRNet model was trained separately using three different loss functions—MSE (mean squared error), NCC (normalized cross-correlation), and our proposed loss. Six leave-one-out experiments were conducted, where one specimen was held out for testing while the remaining five were used for training. Table 2 presents the localization errors for each loss function across all held-out specimens, including per-specimen results as well as the overall mean and standard deviation for each loss.

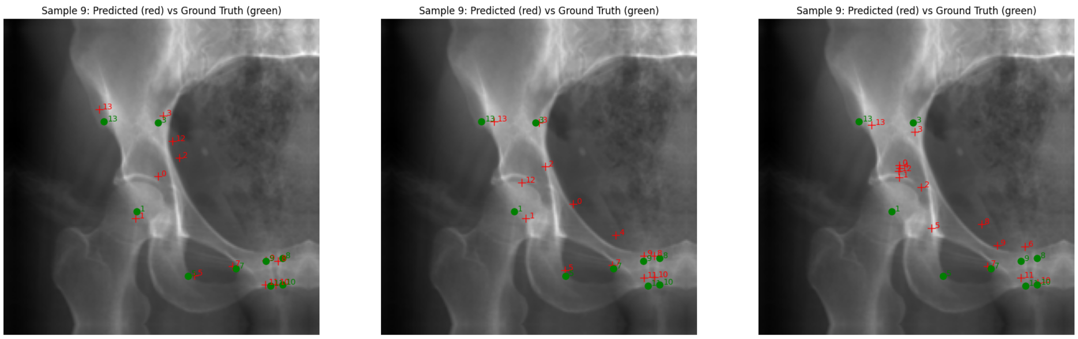
Table 2: Euclidean distance error (in pixels) for six leave-one-out specimen test splits. Lower values indicate better performance.

Test Specimen	**Proposed loss**	**MSE loss**	**NCC loss**
6	18.28	32.26	42.01
5	18.41	34.92	45.48
4	6.54	22.46	45.61
3	6.39	27.07	42.81
2	5.13	23.85	28.52
1	6.76	17.22	26.17
Mean ± Std	**10.25 ± 5.79**	**26.30 ± 6.04**	**38.43 ± 7.98**

Overall, the proposed loss consistently achieved the lowest Euclidean localization error across all test specimens, demonstrating superior accuracy and robustness compared with both the MSE and NCC losses. Predictions obtained with our loss were consistently well-aligned with the ground-truth landmarks, showing minimal variability across different specimens and anatomical variations.

In contrast, the MSE loss, while generally reliable, produced larger deviations in more challenging cases and struggled to fully capture subtle anatomical structures, leading to less precise landmark placement. The NCC loss exhibited even greater variability, with errors more pronounced in specimens that presented complex bone geometries or less distinct fluoroscopic features.

These results highlight the advantage of integrating segmentation-informed consistency and landmark-specific constraints in the training process. By leveraging structural priors and anatomical guidance, our proposed loss not only improves average accuracy but also ensures more stable and reliable predictions across diverse test conditions. To further highlight these differences, Figure 1 provides a qualitative comparison of the landmark locations predicted by each of the evaluated loss functions.


Sample 9: Predicted (red) vs Ground Truth (green)
Sample 9: Predicted (red) vs Ground Truth (green)
Sample 9: Predicted (red) vs Ground Truth (green)

(*a*) Proposed loss (*b*) MSE loss (*c*) NCC loss

Figure 1: Qualitative comparison of landmarks locations predicted by evaluated loss functions. Green dots represent ground truth locations and red crosses represent predicted locations.

4. Discussion

In this work, we demonstrated that incorporating a visibility-guided consistency loss can meaningfully improve anatomical landmark localization by penalizing predictions outside segmented bone regions. However, we still observed cases where the model predicts a non-zero visibility even when the true landmark is absent, which suggests that segmentation errors or coarse mask boundaries can mislead the visibility branch. Improving the precision of the segmentation mask, for example by providing more detailed subregion delineation or estimating segmentation uncertainty in an anatomically aware manner (Adiga et al., 2024), could further enhance the model's ability to correctly suppress implausible landmark predictions.

A second limitation relates to the weighting scheme and granularity of the visibility signal. In our loss formulation, all "outside-bone" predictions are uniformly penalized, regardless of how far they lie from the correct anatomical region. This binary weighting does not reflect the actual anatomical cost of different errors: a landmark predicted just outside the bone boundary may be less problematic than one predicted far away. A more refined approach might modulate the penalty by distance to the bone mask or by the local geometry of the bone surface. Similarly, our discrete visibility values may limit the expressiveness of uncertainty; adopting a continuous visibility confidence could yield smoother gradients and more informative feedback during training. These ideas are in line with work (Chen et al., 2022) on semi-supervised anatomical landmark detection, where a global shape prior is used to regularize pseudo-labels and enforce more coherent landmark predictions.

While our dual-branch design leverages structural cues from segmentation, it does not explicitly model spatial dependencies between landmarks. Previous work has shown that incorporating global shape priors or anatomical constraints can regularize landmark configurations and improve robustness. For example (Pang et al., 2024) proposed a prior-guided coarse-to-fine framework for 3D landmark localization, which exploits annotation priors and the correlation between landmarks to better maintain anatomical relationships.

In addition, generalization remains a key challenge. Fluoroscopic images often contain complex artifacts or suffer from low contrast, which may hinder the translation of models trained on cadaver specimens to live clinical settings. To mitigate this, future work could explore alternative learning strategies, similar to those proposed in (Lu et al., 2023), which take advantage of pseudo-labeled data and consistency-based regularization. Incorporating such approaches could improve the model's adaptability to diverse imaging conditions and enhance robustness in real-world applications.

In summary, our results demonstrate that visibility-based regularization effectively promotes anatomical consistency in landmark localization. At the same time, they underscore several avenues for improvement, including enhanced segmentation accuracy, more expressive modeling of landmark visibility, and the incorporation of explicit spatial priors. Tackling these challenges has the potential to further improve the accuracy, interpretability, and generalizability of landmark detection in fluoroscopic and other medical imaging modalities.

5. Conclusion

In this work, we presented a robust framework for anatomical landmark localization in fluoroscopic images, combining heatmap-based regression with a visibility-guided consistency loss. The visibility loss enforces anatomical plausibility by penalizing landmarks predicted outside their segmented bone regions, while the heatmap representation allows precise and probabilistic coordinate regression. Experimental results using a leave-one-out specimen protocol demonstrated that our approach accurately predicts landmark positions across challenging fluoroscopic images, achieving lower mean Euclidean distances on unseen specimens compared to standard loss functions such as MSE or NCC.

By integrating segmentation information through the visibility loss, the proposed dual-branch design effectively leverages structural cues, improving both the accuracy and robustness of landmark localization. Overall, our method provides a strong foundation for further applications in image-guided interventions, automated anatomical analysis, and scenarios where precise landmarking is critical.

Acknowledgments

This work was supported by the Ministry of Education, Science, Research, and Sport of the Slovak Republic under contract no. VEGA 1/0174/24, by the Slovak Research and Development Agency under contract no. APVV-23-0411

References

Sukesh Adiga, Jose Dolz, and Herve Lombaert. Anatomically-aware uncertainty for semi-supervised image segmentation. *Medical Image Analysis*, 91:103011, 2024.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Runnan Chen, Yuexin Ma, Lingjie Liu, Nenglun Chen, Zhiming Cui, Guodong Wei, and Wenping Wang. Semi-supervised anatomical landmark detection via shape-regulated self-training. *Neurocomputing*, 471:335–345, 2022.

Eleonora Croci, Hanspeter Hess, Fabian Warmuth, Marina Künzler, Sean Börlin, Daniel Baumgartner, Andreas Marc Müller, Kate Gerber, and Annegret Mündermann. Fully automatic algorithm for detecting and tracking anatomical shoulder landmarks on fluoroscopy images with artificial intelligence. *European Radiology*, 34(1):270–278, 2024.

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(5):759–769, 2020.

Ron Keuth, Lasse Hansen, Maren Balks, Ronja Jäger, Anne-Nele Schröder, Ludger Tüshaus, and Mattias Heinrich. Denseseg: joint learning for semantic segmentation and landmark detection using dense image-to-shape representation. *International Journal of Computer Assisted Radiology and Surgery*, 20(3):441–451, 2025.

Liyun Lu, Mengxiao Yin, Liyao Fu, and Feng Yang. Uncertainty-aware pseudo-label and consistency for semi-supervised medical image segmentation. *Biomedical Signal Processing and Control*, 79:104203, 2023.

Yijie Pang, Pujin Cheng, Junyan Lyu, Fan Lin, and Xiaoying Tang. Prior guided 3d medical image landmark localization. In *Medical Imaging with Deep Learning*, pages 1163–1175. PMLR, 2024.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

Yukang Yang, Yu Wang, Tianyu Liu, Miao Wang, Ming Sun, Shiji Song, Wenhui Fan, and Gao Huang. Anatomical prior-based vertebral landmark detection for spinal disorder diagnosis. *Artificial Intelligence in Medicine*, 159:103011, 2025.