Landmark-Based Anatomical Priors as Penalty Masks in Weakly Supervised Learning

Effects on Classification Performance and Heatmap Distribution in Hip Osteophyte Detection

Michel Chen[1]
Supervisor(s): Jesse Krijthe[1] **Gijs van Tulder**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Michel Chen
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Gijs van Tulder, Julia Olkhovskaia

Abstract

Weakly supervised learning can reduce the annotation burden for radiographic osteophyte detection because models can be trained with image-level labels rather than pixel-level masks. However, image-level supervision does not specify where the pathology is located, and a classifier may therefore base its decisions on irrelevant anatomical regions. This paper studies whether landmark-based anatomical priors can improve the classification performance and spatial behaviour of a weakly supervised hip osteophyte classifier. Using the CHECK and OAI datasets, we train a ResNet-18 baseline to predict four binary osteophyte targets and compare it with a prior-guided model that adds a penalty to class activation maps during training. The penalty is constructed from BoneFinder landmarks and uses a plateau Gaussian mask around four anatomical target zones. Performance is evaluated using AUC, heatmap centre-of-mass distance, peak distance, spread, paired Wilcoxon signed-rank tests, and qualitative heatmap visualizations. The prior-guided model produces more compact heatmaps that are significantly closer to the landmark-defined anatomical target zones, with mean centre-of-mass distance reductions between 23.9 and 74.0 pixels, mean peak distance reductions between 24.9 and 73.1 pixels, and spread reductions between 24.5 and 33.7 when evaluated on positive osteophyte cases only. Classification performance remains similar to the baseline, with AUC differences between -0.01 and +0.01. These findings indicate that landmark-based penalty masks can improve alignment of class-discriminative heatmaps in weakly supervised hip osteophyte detection without requiring pixel-level osteophyte annotations.

1 Introduction

Osteoarthritis (OA) is a major problem in the public health sector. A key radiographic indicator of hip OA is the development of osteophytes, or bone protrusions, which commonly develop in specific anatomical regions such as the superior and inferior acetabulum and femoral head. Accurate identification of these osteophytes is clinically important for assessing disease severity and progression.

Currently, deep learning (DL) is widely used for automated image analysis in radiology. DL models tasked with precisely localizing radiographic features are trained using strongly supervised learning methods. This requires large datasets paired with strong labels. These labels are highly accurate and manually assigned by medical experts. This makes the labeling process highly time-consuming and expensive.

Weakly supervised learning (WSL) offers a more efficient alternative by training models using image-level labels (e.g. osteophyte present vs. absent). This method can reduce the need for manual expert annotation, making it more scalable [1].

However, this efficiency comes at the cost of reduced spatial guidance. Because WSL does not provide pixel-level annotations, the model is not explicitly guided toward the anatomical region responsible for the label. This could cause the model to resort to shortcut learning, where it bases its prediction on easily detectable but clinically irrelevant cues instead of the pathological region of interest.

To assess whether weakly supervised models localize clinically relevant regions, researchers typically rely on Gradient-weighted Class Activation Mapping (Grad-CAM) [2]. This method is shown to be highly effective in providing class-discriminative visual explanations for most CNN architectures without changing the underlying model structure [2]. In this way, Grad-CAM can help reveal whether a model is focusing on the intended anatomical region or relying on shortcut cues.

To address the lack of spatial guidance in weakly supervised learning, recent work has explored the integration of anatomical constraints. Anatomical constraints can be integrated into medical image models in several ways, for example through preprocessing, region-of-interest selection, segmentation-based guidance, or attention mechanisms. Since this work uses a spatial penalty mask, we focus here on map-based approaches that guide the model's attention. Some approaches use anatomical prior maps or attention mechanisms to encourage models to focus on disease-relevant structures [3, 4], while others improve the quality of CAM-based pseudo-labels for weakly supervised medical segmentation by incorporating structural constraints [5]. These studies suggest that anatomical knowledge can improve the spatial behaviour of weakly supervised models.

However, the map-based approaches discussed above construct their anatomical priors from relatively dense spatial sources, such as bounding-box annotations, aggregated ROI maps, or model-derived pseudo-masks. Less is known about whether sparse anatomical landmark coordinates can provide similar spatial guidance. In contrast to dense prior maps that directly guide attention or segmentation, our method converts a small set of landmark coordinates into a plateau Gaussian penalty mask. This mask penalizes class activation far from the expected anatomical target zone during training, providing spatial regularization without requiring pixel-level osteophyte annotations. It remains unclear whether such a landmark-derived penalty mask can improve classification performance and Grad-CAM localization for hip osteophyte detection.

The primary research question of this study is therefore:

> How does the incorporation of landmark-based anatomical priors, implemented as a spatial penalty mask, affect the classification and localization performance of weakly supervised models for hip osteophyte detection?

This question is addressed through two sub-questions:

- What are the main localization limitations of a baseline model without spatial guidance, such as activation outside the joint region, diffuse heatmaps, or sensitivity to irrelevant and/or other anatomical structures?
- How do anatomical priors influence the spatial distribution of CAM activations, particularly in terms of concentration, spread, and proximity to expected anatomical target regions?

In addition to quantitative analyses on positive cases, negative and failure cases are inspected qualitatively.

To answer this, we present a methodology that extracts anatomical landmark coordinates using BoneFinder, an automatic method for locating proximal femur and acetabular shape landmarks in hip radiographs. In this work, BoneFinder is used to obtain the bone contour points from which the four osteophyte target zones are defined. These sparse landmark coordinates are converted into a plateau Gaussian-based spatial penalty mask. Our main contribution is the design of a custom loss function that penalizes a baseline ResNet-18 model during training if its activations are too far from the established anatomical target zones. We expect the penalty mask to significantly reduce spatial localization error. Furthermore, we aim to demonstrate this spatial improvement while ensuring classification performance matches or exceeds the baseline.

The remainder of this paper is structured as follows. Section 2 provides the methodology and problem description, the baseline model, and the prior model and its penalty mask.

Section 3 outlines the preprocessing of the data and the experimental setup. Section 4 reports the classification, localization, correlation, and heatmap results. Section 5 reflects on responsible research practices, discussing the ethical implications and reproducibility of our methods. Section 6 discusses the findings. Finally, Section 7 concludes the paper by answering the main research question and suggesting recommendations for future work.

2 Methodology and Problem Description

2.1 Problem Description

This study uses a CNN for binary classification of the presence of hip osteophytes in four locations from the same hip image: superior femoral, inferior femoral, superior acetabular, and inferior acetabular osteophytes. The model is trained using image-level labels that only indicate whether an osteophyte is present for each target. This is therefore a form of weakly supervised learning, because the labels provide target-level disease information but do not indicate the exact spatial location of the osteophyte within the image. In this training setting, the model is optimized solely for classification performance, without direct constraints on anatomical attention. Although this makes it easier to scale to larger datasets, the absence of spatial supervision often results in poor anatomical localization. Rather than focusing on clinically relevant osteophyte regions, models may focus on unrelated image structures or highly localized discriminative patterns.

The knowledge gap addressed in this paper concerns whether sparse landmark coordinates can be used to construct training-time penalty masks for a weakly supervised classifier. Anatomical priors have previously been used to guide segmentation or attention [3, 4, 5], but these approaches usually use dense anatomical information as attention guidance or as part of a segmentation pipeline. We therefore investigate whether penalty masks derived from sparse landmarks can guide class activation toward expected osteophyte regions without requiring dense, pixel-level osteophyte annotations. We hypothesize that penalizing activation outside target landmark zones will shift heatmap mass toward clinically relevant regions and reduce localization error, while improving classification performance.

2.2 Methodology

The proposed method adds a spatial prior loss to a weakly supervised classifier. The aim of this prior is to discourage class activation outside anatomically plausible regions while still training the model using only image-level osteophyte labels.

The method consists of three main parts: training a baseline weakly supervised model, using landmark points to construct a spatial penalty mask, and adding a plateau-based Gaussian penalty loss during training.

Baseline weakly supervised classifier

The baseline model is trained using binary cross-entropy. Unknown labels are excluded from the loss by masking out their corresponding loss terms. For a prediction logit z_i and binary label y_i, the baseline loss is:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log \sigma(z_i) + (1 - y_i) \log(1 - \sigma(z_i)) \right], \tag{1}$$

where σ is the sigmoid function and N is the number of valid labels in the batch.

Anatomical prior and penalty mask

The proposed method adds a spatial prior loss that penalizes class activation outside plausible anatomical regions. Feature maps are extracted from the final convolutional layer of the model and used to compute class activation maps, following the original class activation mapping formulation by Zhou et al. [6]. Let $F_k(x, y)$ be the k-th feature map and w_{ck} the classifier weight for class c. The class activation map for class c is computed as:

$$M_c(x, y) = \mathrm{ReLU}\left(\sum_k w_{ck} F_k(x, y)\right). \tag{2}$$

Because the final convolutional feature map has lower spatial resolution than the input image, the landmark coordinates are scaled from the input resolution to the activation-map resolution. The same scale factor is applied to the Gaussian parameters.

For each activation-map location (x, y), the distance to the closest valid anatomical landmark is computed:

$$d(x, y) = \min_{p \in P_c} \|(x, y) - p\|_2 , \tag{3}$$

where P_c is the set of landmarks associated with osteophyte target c. Multiple landmarks can be associated with one target because an osteophyte may occur along a short contour segment rather than a single point. For example, the superior femoral target is represented by several landmarks along the superior femoral head contour, while acetabular targets are represented by one landmark.

A plateau Gaussian penalty is then applied. Locations within a plateau radius are not penalized. Locations outside this radius receive an increasing penalty:

$$d_{\mathrm{adj}}(x, y) = \max(0, d(x, y) - r), \tag{4}$$

$$Q_c(x, y) = 1 - \exp\left(-\frac{d_{\mathrm{adj}}(x, y)^2}{2\sigma^2}\right), \tag{5}$$

where r is the plateau radius and σ controls the smoothness of the penalty. The class activation map is normalized spatially:

$$\hat{M}_c(x, y) = \frac{M_c(x, y)}{\sum_{x,y} M_c(x, y) + \epsilon}. \tag{6}$$

The prior loss is the expected penalty under the normalized activation distribution:

$$\mathcal{L}_{\mathrm{prior}} = \frac{1}{C'} \sum_{c \in C'} \sum_{x,y} \hat{M}_c(x, y) Q_c(x, y), \tag{7}$$

where C' is the set of targets with valid landmark points. The final training objective is:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{BCE}} + \lambda \mathcal{L}_{\mathrm{prior}}, \tag{8}$$

where λ controls the strength of the anatomical prior.

It is important to note that the prior does not provide direct osteophyte supervision. The landmark points define anatomically plausible target zones rather than exact lesion boundaries. Therefore, the prior loss should be interpreted as encouraging class activation to align with expected anatomical regions, not as enforcing precise osteophyte localization. This distinction is important because true lesion localization would require manual osteophyte masks, bounding boxes, or expert point annotations, which are not available in this study.

3 Experimental Setup

3.1 Preprocessing and landmark extraction

The preprocessing script loads each DICOM image with metadata and reads anatomical landmark points extracted using BoneFinder, which is based on automatic proximal femur segmentation using random forest regression voting [7]. When a target pixel spacing is provided, the image is resampled to that spacing. The hip is then cropped around the femoral head, using the BoneFinder femoral head circle as the crop centre. Intensities are normalized using the 5th and 95th percentiles, and the processed image and labels are written to an HDF5 file.

After cropping, the BoneFinder curves are converted to crop coordinates by subtracting the crop offset. From these curves, four landmark groups are selected near the superior femoral head, inferior femoral head, superior acetabulum, and inferior acetabulum. These groups define the anatomical target zones used by the prior loss, as shown in Figure 1. Left hips are flipped horizontally so that all cropped hip images have a consistent orientation. The selected landmark coordinates are then stored in the HDF5 file for use during model training.

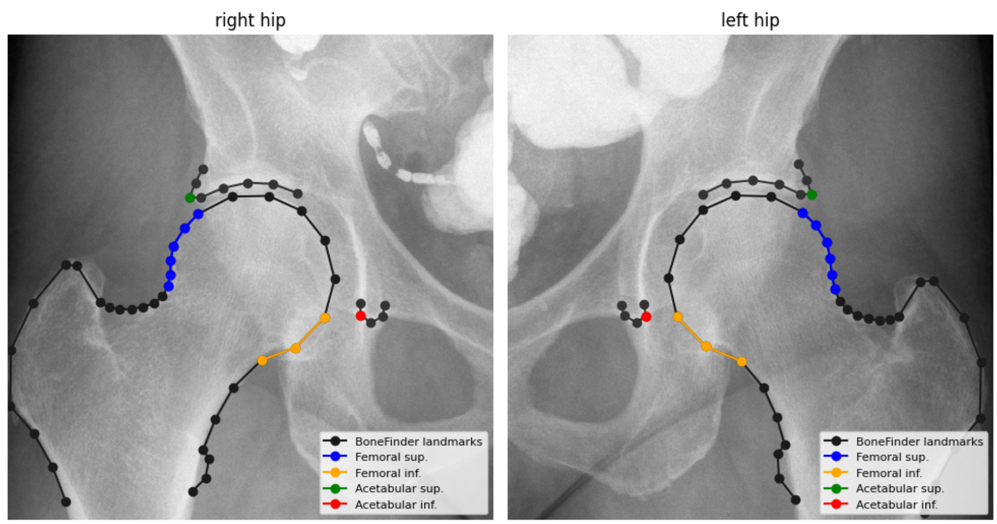


Figure 1: BoneFinder landmarks used to define anatomical target zones for the landmark-based prior. Black points indicate non-target BoneFinder landmarks along the proximal femur and acetabular contours. Colored points indicate the selected landmark groups used as anatomical priors for the four osteophyte targets: superior femoral, inferior femoral, superior acetabular, and inferior acetabular osteophytes. The right and left hip examples show that the same anatomical regions are selected on both sides.

3.2 Model training

Both baseline and prior-guided models use ResNet-18 with pretrained weights and the same subject-level 70/15/15 train-validation-test split from the CHECK and OAI datasets, which contain approximately 1,000 and 8,000 subjects, respectively. Training is done on the DelftBlue Cluster [8] using Adam with a learning rate of 10^{-5}, batch size 64, maximum 50 epochs, and random rotation augmentation. The four evaluated targets are `osteo_fem_sup`, `osteo_fem_inf`, `osteo_acet_sup`, and `osteo_acet_inf`.

For the prior-guided model, the spatial prior loss is computed on the final ResNet-18 convolutional feature map. With 224×224 input crops, this feature map has spatial resolution 7×7. Therefore, BoneFinder landmark coordinates are scaled from image coordinates to feature-map coordinates using $s = 7/224$. The same scale factor is applied to the plateau radius r and Gaussian width σ.

The prior hyperparameters are selected by lowest validation loss using a grid search. The grid search covers $\lambda \in \{0.5, 1.0, 2.0\}$, $r \in \{5, 10, 20\}$, and $\sigma \in \{10, 20, 30\}$. The best configuration in the initial grid is $\lambda = 2.0$, $r = 20.0$, and $\sigma = 30.0$. Because all three selected values are at the upper boundary of the grid, four additional configurations are evaluated: $(3, 20, 30)$, $(2, 30, 30)$, $(2, 20, 40)$, and $(3, 30, 40)$. None of these additional configurations improves the validation loss, so $\lambda = 2.0$, $r = 20.0$, and $\sigma = 30.0$ is used as the final prior-guided model configuration.

During training, both models optimize the binary classification loss. For the prior-guided model, the weighted prior loss is added during the training phase only. The baseline and prior-guided models are then evaluated on the same held-out test-set using identical classification and localization metrics.

3.3 Evaluation metrics

Classification performance is measured by the area under the curve (AUC) for each target. To visualize localization performance, we rely on Grad-CAM. Grad-CAM extracts heatmaps from the final convolutional feature maps. These heatmaps provide visual explanations of the spatial regions contributing most strongly to the model prediction. Because Grad-CAM does not require architectural modifications, it is well suited for visualizing localization behavior in CNN-based weakly supervised classification models such as ResNet-18.

Localization performance is quantified across the four anatomical target zones using two metrics. Centre-of-mass distance measures the Euclidean distance between the Grad-CAM heatmap centre of mass and the closest target landmark. Peak activation distance measures the distance between the maximum heatmap activation and the closest target landmark.

In addition to distance-to-landmark metrics, we measure heatmap spread to measure compactness. Quantitative evaluation of heatmaps is commonly recommended because visual inspection alone can be misleading, and prior work has proposed localization- or energy-based measures such as pointing-game and energy-inside-region metrics [2, 9, 10]. However, these metrics usually require ground-truth boxes, masks, or annotated relevant regions, which are not available for osteophytes in this study. Therefore, we use heatmap spread as a task-specific compactness measure.

Heatmap spread is defined as the weighted spatial standard deviation of the Grad-CAM activation around its centre of mass:

$$S = \sqrt{\frac{\sum_{x,y} M(x,y)\left((x-\bar{x})^2 + (y-\bar{y})^2\right)}{\sum_{x,y} M(x,y) + \epsilon}},$$

where $M(x, y)$ is the Grad-CAM activation and $(\bar{x}, \bar{y})$ is the heatmap centre of mass. Lower values indicate that the heatmap activation is more spatially concentrated. Unlike the centre-of-mass and peak-distance metrics, spread does not measure whether the heatmap is close to the anatomical target. Instead, it measures whether the explanation is diffuse or compact.

For paired statistical comparison, we compare the baseline model to the prior model using the Wilcoxon signed-rank test on the localization and compactness metrics. The samples in this test are not repeated runs of the model. Instead, each sample corresponds to one hip image from the held-out test-set. For each target and metric, we compute one value for both baseline and prior-guided models on the same image, giving a paired observation. The Wilcoxon signed-rank test is applied to these paired differences. The null hypothesis is that the median paired difference is zero. Therefore, the resulting significance tests assess whether the prior-guided model consistently improves over the baseline across images for this specific trained model pair and seed. They do not by themselves measure robustness across different random initializations, training runs, or data splits.

Only images with a positive label for the corresponding target are included, because distance to an osteophyte target zone is only meaningful when that target is present. Negative cases are evaluated qualitatively using Grad-CAM examples rather than included in the quantitative localization statistics. Images with missing labels or undefined metric values are excluded. Therefore, the reported n is the number of paired positive test-set hip images for that target and metric. Negative differences indicate lower values for the prior-guided models, corresponding to better anatomical alignment or more compact heatmaps.

4 Results

4.1 Classification Performance

The prior-guided model maintains very similar classification performance, as can be seen in Figure 2. Femoral superior AUC remains 0.92. Femoral inferior AUC decreases slightly from 0.90 to 0.89. Superior acetabular AUC increases from 0.79 to 0.80. Inferior acetabular AUC decreases slightly from 0.87 to 0.86. These differences are small compared with the localization changes reported below.

4.2 Localization Performance

The strongest effect of the prior is visible in the localization metrics. Across all four targets, centre-of-mass and peak activation distances are lower for the prior-guided model, indicating that the prior shifts Grad-CAM activation closer to the landmark-defined target zones in positive cases (Figure 3). Table 1 confirms this pattern, showing that all paired differences are negative, with shorter distances and more compact heatmaps for the prior-guided model.

Mean centre-of-mass distance reductions range from 23.9 pixels for superior femoral osteophytes to 74.0 pixels for superior acetabular osteophytes. Mean peak distance reductions range from 24.9 pixels to 73.1 pixels. Heatmap spread also decreases for all four targets, showing that the prior-guided model produces more compact heatmaps. Mean spread reductions range from 24.5 to 33.7 pixels. All paired Wilcoxon signed-rank tests have p-values $p < 0.001$, supporting the conclusion that, for the trained models evaluated in the main experiment, the localization and compactness improvements are consistent across positive paired samples.

It is particularly interesting that the baseline model has much higher centre-of-mass and peak activation distances for the superior acetabular target than for the other targets. This suggests that, before applying the prior, the baseline heatmaps for this target are especially

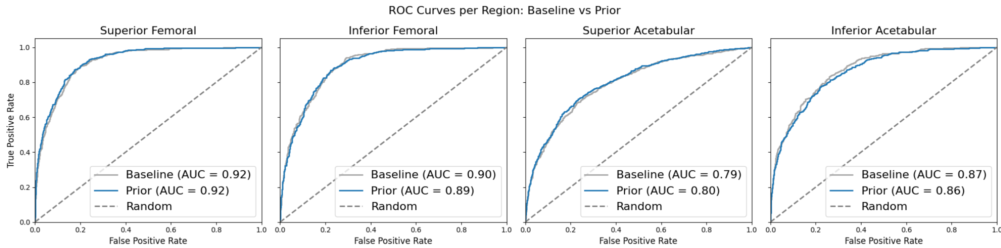


Figure 2: ROC curves for the baseline and prior-guided models. The prior-guided model preserved classification performance, with AUC differences between -0.01 and +0.01 across the four targets.

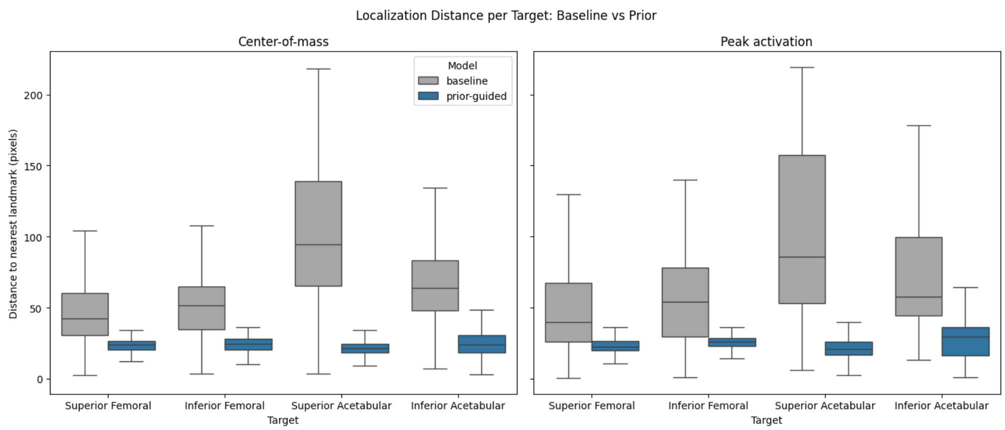


Figure 3: Localization distance distributions for centre-of-mass and peak activation on positive target labels only. Lower values indicate that Grad-CAM heatmaps are closer to the target landmarks. The prior-guided model reduces localization distance for all targets.

displaced from the landmark-defined target zone. These errors require more explanation than the distance metrics alone can provide. For this reason, we take a closer look at the aggregated heatmaps in the next section.

4.3 Aggregated Heatmap Analysis

Mean Grad-CAM heatmaps over positive samples show that the baseline model often spreads activation over a broad region of the hip, although the pattern differs between targets (Figure 4). For the superior femoral and inferior acetabular targets, the baseline heatmaps are relatively broad but remain centred near the landmark-defined target zones. For the inferior femoral target, activation is distributed across two nearby regions that both lie close to the target points. In contrast, the superior acetabular target shows two distinct activation regions, one of which is located substantially farther from the target zone. The prior-guided model concentrates activation nearer to the landmark-defined target zones across all four targets. This pattern is consistent with the quantitative results in Figure 3 and Table 1, where the larger baseline displacement for the superior acetabular target may contribute to

Table 1: Localization and compactness metrics for paired positive test-set hip images. Values are reported as mean ± standard deviation in pixels. Here, n denotes the number of positive labels. Differences are computed as prior minus baseline, so negative values indicate lower localization distance or more compact heatmaps for the prior-guided model.

Metric	Target	n	Baseline	Prior	Mean diff.
Centre-of-mass	Femoral sup.	605	47.5 ± 25.5	23.6 ± 4.3	-23.9
	Femoral inf.	364	52.1 ± 26.9	23.9 ± 5.7	-28.3
	Acetabular sup.	728	95.7 ± 48.7	21.7 ± 6.5	-74.0
	Acetabular inf.	330	70.1 ± 32.1	25.3 ± 13.8	-44.8
Peak distance	Femoral sup.	605	48.4 ± 31.6	23.5 ± 5.9	-24.9
	Femoral inf.	364	61.9 ± 48.6	25.7 ± 10.6	-36.2
	Acetabular sup.	728	95.0 ± 58.0	21.9 ± 8.3	-73.1
	Acetabular inf.	330	80.9 ± 54.9	32.1 ± 36.1	-48.8
Heatmap spread	Femoral sup.	605	57.1 ± 11.6	32.4 ± 4.6	-24.5
	Femoral inf.	364	57.0 ± 15.8	23.4 ± 4.9	-33.3
	Acetabular sup.	728	57.4 ± 17.6	23.5 ± 3.8	-33.7
	Acetabular inf.	330	52.6 ± 19.1	22.7 ± 5.4	-30.0

its higher localization distances and the larger improvement observed with the prior.

However, the aggregated heatmaps also show that the prior does not produce exact localization. In several targets, the strongest activation is near, but not exactly centered on, the landmark points. The inferior acetabular target illustrates this limitation most clearly: although the prior-guided heatmap is more constrained than the baseline, its strongest activation is close to the image boundary. These observations motivate the individual case analysis in Section 4.5.

4.4 Co-occurrence Between Osteophyte Targets

The four osteophyte targets are not independent. Table 2 shows that all off-diagonal target correlations are positive, with the strongest correlations between superior and inferior femoral osteophytes ($r = 0.445$) and between superior and inferior acetabular osteophytes ($r = 0.405$). This indicates that osteophytes in one anatomical target zone often co-occur with osteophytes in other target zones.

This co-occurrence is relevant for weakly supervised learning because the model is trained only with image-level target labels. The baseline model may therefore learn to predict one target using evidence from a different, correlated target region. In that case, a high classification score can be obtained without target-specific localization.

Figure 5 illustrates this behaviour for one baseline example. In this sample, the superior femoral, inferior femoral, and superior acetabular labels are positive, while the inferior acetabular label is negative. However, the baseline Grad-CAMs are not clearly separated by target. The superior femoral and superior acetabular heatmaps are broad and overlap strongly around the superior joint region. More importantly, the inferior femoral heatmap peaks near the superior acetabular region rather than near the inferior femoral landmarks. The inferior acetabular heatmap also activates in the superior acetabular region, despite the inferior acetabular label being negative.

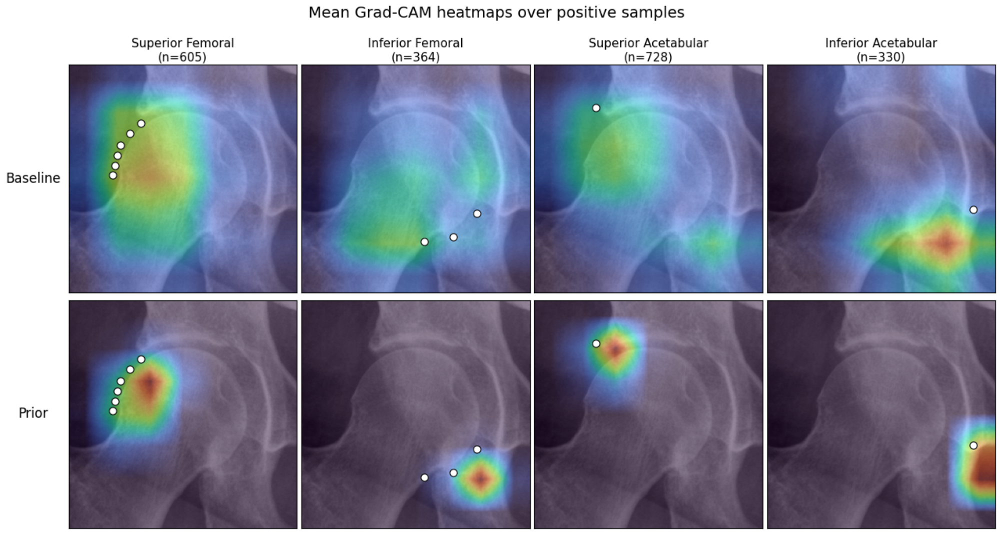


Figure 4: Mean Grad-CAM heatmaps over positive samples. Columns correspond to the four osteophyte targets. White points indicate mean target landmarks. The prior-guided model produces more compact and anatomically aligned heatmaps than the baseline.

Table 2: Pairwise target co-occurrence matrix. All off-diagonal correlations are positive, indicating that osteophytes in different target zones tend to co-occur. This co-occurrence may allow a weakly supervised baseline model to use evidence from one anatomical region as proxy information for another target.

	Sup. fem.	Inf. fem.	Sup. acet.	Inf. acet.
Sup. fem.	1.00	0.445	0.383	0.329
Inf. fem.	0.445	1.00	0.252	0.328
Sup. acet.	0.383	0.252	1.00	0.405
Inf. acet.	0.329	0.328	0.405	1.00

This example does not prove that the baseline model always uses other target zones as proxies, but it is consistent with that behaviour. Together with the positive target correlations in Table 2, it suggests that part of the baseline localization error may come from the model using correlated osteophyte evidence rather than target-specific anatomical evidence.

4.5 Individual Case Analysis

Aggregated heatmaps show the average spatial effect of the prior, but individual examples are needed to inspect how this effect appears in specific positive, negative, and failure cases. Figure 6 shows four representative Grad-CAM examples. These examples are illustrative case studies and should not be interpreted as aggregate evidence.

Figure 6a shows a positive superior acetabular case where both models predict the positive label with high confidence. The baseline score is 0.91 and the prior-guided score is 0.86.

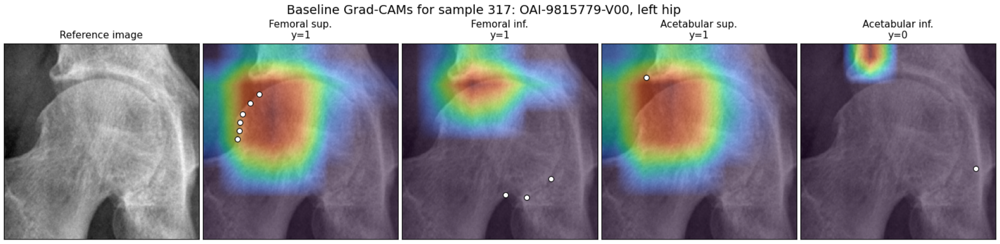


Figure 5: Baseline Grad-CAMs for all four osteophyte targets in one test sample. White points indicate the target landmarks for each output. The heatmaps show limited target specificity: the inferior femoral and inferior acetabular outputs activate in superior joint regions rather than near their corresponding landmarks, consistent with possible proxy use of correlated target zones.

The prior-guided heatmap is more compact and shifted toward the superior acetabular target region, while the baseline heatmap is broader and less target-specific.

Figure 6b shows a positive superior femoral case where the prior-guided model remains correctly positive but lowers the prediction score from 0.95 to 0.77. The prior reduces diffuse activation compared with the baseline, but the strongest activation is still slightly right of the expected femoral edge. This illustrates a partial localization improvement rather than exact osteophyte localization.

Figure 6c shows a negative superior femoral case. The baseline model gives a high false-positive-like score of 0.70, while the prior-guided model lowers the score to 0.26. The prior also produces a more compact heatmap near the superior femoral target region. However, this should be interpreted as spatial constraint toward the landmark-defined region, not proof that the model has learned the exact absence of disease.

Figure 6d shows a limitation. The true label is positive and the baseline gives a high score of 0.88, but the prior-guided score drops to 0.46. Although the prior heatmap is more spatially constrained, the reduced score shows that improved anatomical compactness can come at the cost of sensitivity.

Overall, these examples suggest that the prior mainly improves regional anatomical alignment and heatmap compactness, not exact lesion localization. This distinction is important because the prior is derived from sparse BoneFinder landmarks rather than manual osteophyte masks.

5 Responsible Research

This study uses medical imaging data, which requires careful handling of privacy, bias, and reproducibility. The radiographs and labels may contain sensitive health information. Therefore, data is stored only on the DelftBlue Cluster [8]. Even when radiographic images do not directly show names or dates, they remain medical data and should be treated as confidential.

Bias is a critical concern. The model can inherit biases from the dataset composition, disease prevalence, and annotation practices. If some demographic groups, disease severities, or imaging settings are underrepresented, the model may generalize poorly. The prior itself may also have assumptions: landmark placement errors or systematic anatomical differences

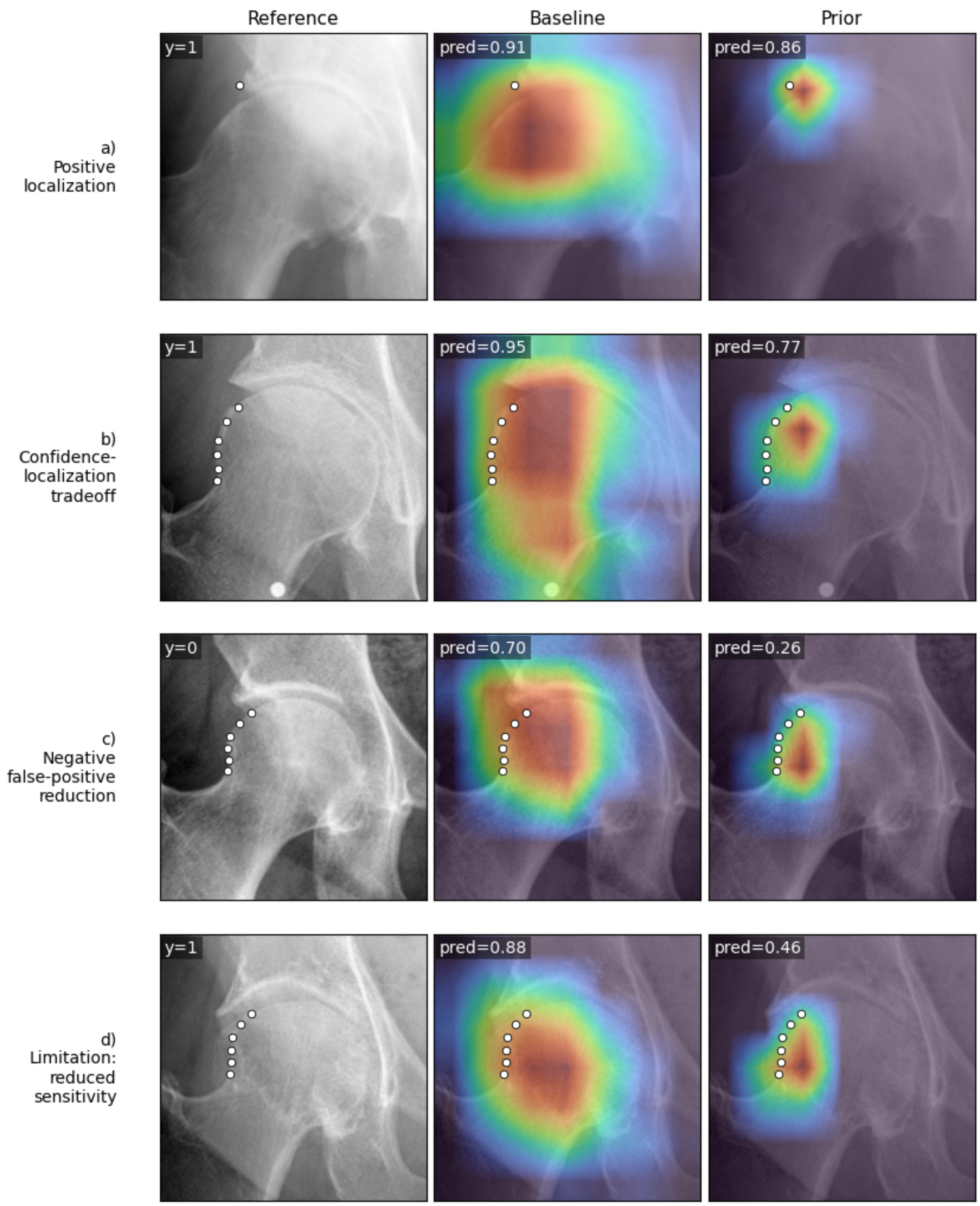


Figure 6: Grad-CAM examples for the baseline and prior-guided models. Each row shows the reference image, baseline Grad-CAM, and prior-guided Grad-CAM for one target; white points indicate the BoneFinder target landmarks. The examples show: (a) positive superior acetabular localization improvement, (b) positive superior femoral confidence-localization tradeoff, (c) negative superior femoral false-positive reduction, and (d) positive superior femoral limitation where the prior improves compactness but reduces the prediction.

could make the penalty less appropriate for some patients. Reporting results only as an average can hide such subgroup failures.

Weak supervision has an additional integrity risk. Because the model is trained only from image-level labels, good classification performance does not imply that the model has learned the intended anatomical evidence. This work explicitly addresses that issue by evaluating localization through Grad-CAM distances and heatmap visualizations. Nevertheless, Grad-CAM is still an explanation method rather than ground-truth localization. The distances are measured to BoneFinder landmarks, not to manually annotated osteophyte locations. The results should therefore be interpreted as evidence of improved anatomical alignment, not as proof of exact lesion segmentation.

Reproducibility is supported by fixed dataset splits, a documented preprocessing pipeline and model, and saved evaluation scripts. The prior hyperparameters were selected using a documented grid search, with the full model-selection results reported in Appendix A. Since the best initial configuration lay at the boundary of the grid, additional higher-value configurations were tested. This reduces the risk of overclaiming that the selected hyperparameters are globally optimal.

Generative AI was used in a limited supporting role during this project. Specifically, ChatGPT was used for language support while writing the report and for assistance with code debugging [11]. This tool was not used to generate experimental results, perform statistical analysis, define the research question, or draw scientific conclusions. All AI-generated suggestions were manually checked, adapted, and verified by the author. No radiographs, patient-level metadata, labels, or other confidential medical data were intentionally entered into these tools. Therefore, AI assistance was treated as a writing and debugging aid, while responsibility for the final text, implementation, analysis, and interpretation remains with the author.

6 Discussion

The results support the hypothesis that coordinate-based landmark priors can improve the anatomical alignment of Grad-CAM heatmaps in weakly supervised hip osteophyte detection. For positive osteophyte cases, adding the prior consistently shifts activation towards the landmark-defined target zones and reduces both centre-of-mass and peak activation distances across all four osteophyte targets. This effect is visible in the quantitative localization metrics as well as in the heatmap visualizations. However, the improvement should be interpreted as better alignment with expected anatomical regions rather than as proof of exact osteophyte localization, since no pixel-level osteophyte annotations are available.

In contrast to the localization improvements, classification performance remains largely unchanged. This is not entirely surprising, because the prior mask encourages the model to place activation closer to the expected target bone contour, but it does not add additional diagnostic information beyond the original radiograph. Therefore, the model still has to solve essentially the same classification problem from the same image. The prior can improve where the model activates without necessarily changing whether positive cases are ranked above negative cases, which is what AUC measures. This can explain why the prior-guided model achieves AUC values that are very similar to the baseline model.

The target-correlation analysis provides another possible explanation for why the baseline can still classify similarly despite poorer target-specific localization. All off-diagonal correlations between osteophyte labels are positive, indicating that osteophytes in different anatomical zones tend to co-occur. In a weakly supervised setting, this creates an opportunity for the classifier to use one osteophyte region as proxy evidence for another target. The baseline example in Figure 5 is consistent with this behaviour: several target-specific Grad-CAMs activate in overlapping superior joint regions rather than near their corresponding landmark zones.

The prior-guided model appears to reduce this ambiguity by penalizing activation outside the expected target zone for each class. This may explain why the largest localization improvement is observed for the superior acetabular where the baseline heatmaps are broad and displaced. However, the correlation analysis cannot prove that the baseline model systematically uses proxy regions. It only shows that such a shortcut is plausible. A stronger test would require counterfactual or lesion-specific annotations showing whether activation follows the true osteophyte rather than a correlated region.

The individual case studies further show that improved anatomical alignment does not always imply a higher prediction score. In successful positive cases, the prior-guided model produces more compact heatmaps near the expected landmarks while maintaining a high prediction score. In the negative example, the prior reduces a high baseline score and constrains the heatmap to the superior femoral region. However, the failure case shows that the prior can also lower the prediction score for a true positive case. This suggests a trade-off: the prior discourages diffuse or anatomically implausible activation, but it may also suppress broader classification evidence that is still useful for prediction.

These findings also motivate reflection on the methodological choices made in this study. Using BoneFinder landmarks for the penalty masks is a useful middle ground between only using image-level labels and requiring or constructing dense osteophyte annotations. The landmarks give a rough but meaningful indication of where each osteophyte type is expected to appear, so the model can be guided toward the right anatomical region. At the same time, this means the method depends on the quality of the Bonefinder landmarks. A possible alternative would be to place similar landmark points manually. Since these points only need to indicate anatomical target zones, this could potentially be done by people with sufficient anatomical knowledge rather than by medical experts.

A second methodological choice concerns the prediction setup. Predicting all four osteophyte targets with one model was chosen for computational efficiency, shared feature learning, and the possibility of analysing target interactions, such as co-occurrence. However, training four separate location-specific models could reduce ambiguity between correlated targets and potentially improve target-specific localization, although it would increase training cost. Another alternative would be to crop the input around each target region using the BoneFinder landmarks before classification, which could force the model to focus on the intended anatomy.

Several limitations remain. First, the prior depends on the quality of the BoneFinder algorithm. If landmarks are incorrect, missing, or biased, the penalty can guide the model towards the wrong region. Second, the heatmap metrics use landmarks as the reference point, while true osteophytes may extend around the landmark or appear with a different shape. Third, the final convolutional activation map has low spatial resolution. Scaling a 224-by-224 crop to a 7-by-7 activation map makes the penalty coarse, even when the original landmarks are precise. This limitation is consistent with prior work showing that weakly supervised medical localization can be strongly constrained by the spatial resolution of CAM-producing architectures [12]. Finally, the significance testing should be interpreted cautiously, since the main Wilcoxon tests are based on paired image-level differences for the evaluated models and do not fully capture uncertainty from alternative data splits and model runs.

7 Conclusions and Future Work

This paper investigated how landmark-based anatomical priors, implemented as spatial penalty masks, affect weakly supervised hip osteophyte detection. The results show that the prior substantially improves the anatomical alignment of Grad-CAM heatmaps while maintaining similar image-level classification performance. Across positive cases for all four osteophyte targets, the prior-guided model reduces both centre-of-mass distance and peak activation distance to the nearest landmark-defined target zone, and improves compactness of the heatmaps, while AUC changes remain small.

The main contribution is a reproducible method for converting sparse BoneFinder landmarks into a spatial penalty for weakly supervised classification. This approach does not require dense, pixel-level osteophyte annotations and can therefore be useful in settings where image-level labels are available but detailed lesion annotations are expensive to obtain.

The results should be interpreted as improved alignment with expected anatomical target regions, not as proof of exact osteophyte localization. The evaluation relies on BoneFinder landmarks rather than expert-drawn osteophyte masks, and the final ResNet-18 feature map has low spatial resolution. In addition, the target-correlation analysis suggests that weakly supervised baseline models may use co-occuring osteophyte regions as proxy evidence, but this behaviour cannot be confirmed without stronger lesion-level annotations.

Future work should evaluate calibration, test more prior hyperparameters beyond the current grid boundary, and compare alternative penalty mask shapes. In addition, the current shared four-output model should be compared with separate location-specific models to assess whether target-specific training reduces ambiguity between correlated osteophyte regions. Another useful comparison would be landmark-guided cropping, where each input is cropped around the relevant BoneFinder target zone before classification. This could test whether restricting the image to the expected anatomical region improves localization or classification performance. Finally, a stronger localization reference, such as expert-drawn osteophyte masks, bounding boxes, or point annotations for a subset of images, would make it possible to measure true osteophyte localization rather than distance to target-zone landmarks.

References

[1] L. Misera, G. Müller-Franzes, D. Truhn, and J. N. Kather, "Weakly supervised deep learning in radiology," vol. 312, no. 1, p. e232085.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," vol. 128, no. 2, pp. 336–359.

[3] M. I. Hossain, M. Zunaed, M. K. Ahmed, S. M. J. Hossain, A. Hasan, and T. Hasan, "ThoraX-PriorNet: A novel attention-based architecture using anatomical prior probability maps for thoracic disease classification," vol. 12, pp. 3256–3273.

[4] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," 2018.

[5] Z. Chen, Z. Tian, J. Zhu, C. Li, and S. Du, "C-CAM: Causal CAM for weakly supervised semantic segmentation on medical image," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11666–11675, IEEE.

[6] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *CoRR*, vol. abs/1512.04150, 2015.

[7] C. Lindner, S. Thiagarajah, J. M. Wilkinson, T. arcOGEN Consortium, G. A. Wallis, and T. F. Cootes, "Fully automatic segmentation of the proximal femur using random forest regression voting," *IEEE Transactions on Medical Imaging*, vol. 32, no. 8, pp. 1462–1472, 2013.

[8] Delft High Performance Computing Centre (DHPC), "DelftBlue Supercomputer (Phase 2)." `https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2`, 2024.

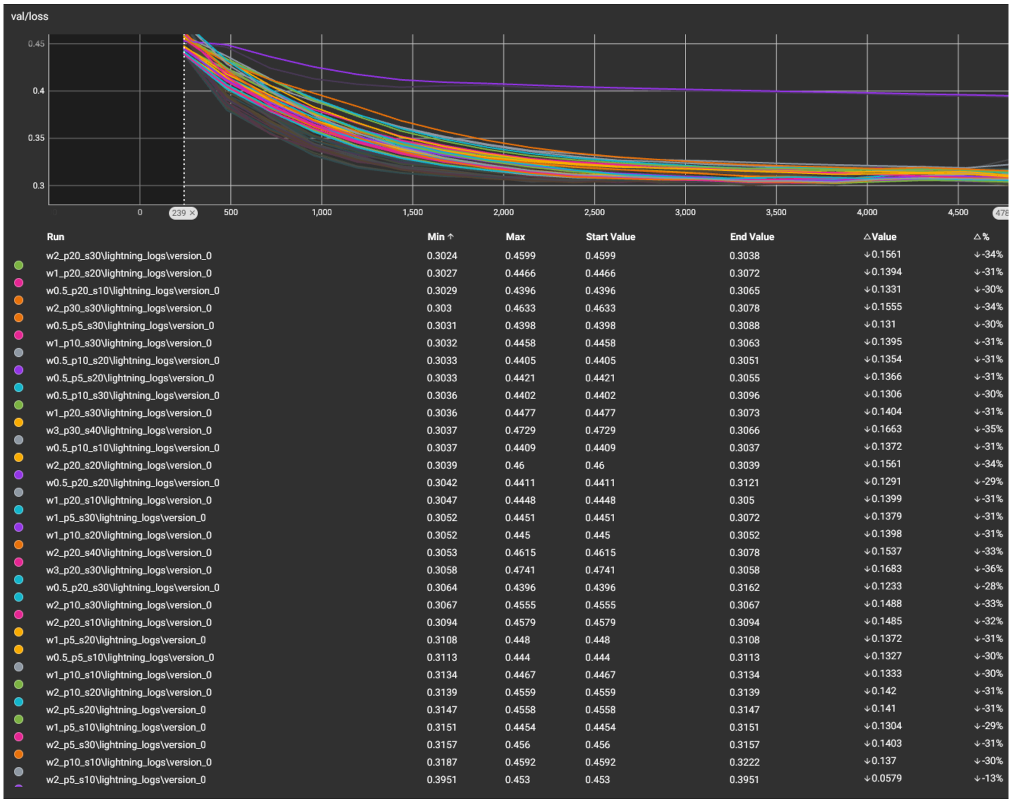
[9] S. Poppi, M. Cornia, L. Baraldi, and R. Cucchiara, "Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis," 2021.

[10] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S. Q. H. Truong, C. D. T. Nguyen, V.-D. Ngo, J. Seekins, F. G. Blankenberg, A. Y. Ng, M. P. Lungren, and P. Rajpurkar, "Benchmarking saliency methods for chest x-ray interpretation," *Nature Machine Intelligence*, vol. 4, no. 10, pp. 867–878, 2022.

[11] OpenAI, "ChatGPT." `https://chatgpt.com/`, 2026. Accessed: Jun. 17, 2026.

[12] K. Preechakul, S. Sriswasdi, B. Kijsirikul, and E. Chuangsuwanich, "High resolution weakly supervised localization architectures for medical images," 2020.

A Appendix



Run	Min ↑	Max	Start Value	End Value	△Value	△%
w2_p20_s30\lightning_logs\version_0	0.3024	0.4599	0.4599	0.3038	↓0.1561	↓-34%
w1_p20_s20\lightning_logs\version_0	0.3027	0.4466	0.4466	0.3072	↓0.1394	↓-31%
w0.5_p20_s10\lightning_logs\version_0	0.3029	0.4396	0.4396	0.3065	↓0.1331	↓-30%
w2_p30_s30\lightning_logs\version_0	0.303	0.4633	0.4633	0.3078	↓0.1555	↓-34%
w0.5_p5_s30\lightning_logs\version_0	0.3031	0.4398	0.4398	0.3088	↓0.131	↓-30%
w1_p10_s30\lightning_logs\version_0	0.3032	0.4458	0.4458	0.3063	↓0.1395	↓-31%
w0.5_p10_s20\lightning_logs\version_0	0.3033	0.4405	0.4405	0.3051	↓0.1354	↓-31%
w0.5_p5_s20\lightning_logs\version_0	0.3033	0.4421	0.4421	0.3055	↓0.1366	↓-31%
w0.5_p10_s30\lightning_logs\version_0	0.3036	0.4402	0.4402	0.3096	↓0.1306	↓-30%
w1_p20_s30\lightning_logs\version_0	0.3036	0.4477	0.4477	0.3073	↓0.1404	↓-31%
w3_p30_s40\lightning_logs\version_0	0.3037	0.4729	0.4729	0.3066	↓0.1663	↓-35%
w0.5_p10_s10\lightning_logs\version_0	0.3037	0.4409	0.4409	0.3037	↓0.1372	↓-31%
w2_p20_s20\lightning_logs\version_0	0.3039	0.46	0.46	0.3039	↓0.1561	↓-34%
w0.5_p20_s20\lightning_logs\version_0	0.3042	0.4411	0.4411	0.3121	↓0.1291	↓-29%
w1_p20_s10\lightning_logs\version_0	0.3047	0.4448	0.4448	0.305	↓0.1399	↓-31%
w1_p5_s30\lightning_logs\version_0	0.3052	0.4451	0.4451	0.3072	↓0.1379	↓-31%
w1_p10_s20\lightning_logs\version_0	0.3052	0.445	0.445	0.3052	↓0.1398	↓-31%
w2_p20_s40\lightning_logs\version_0	0.3053	0.4615	0.4615	0.3078	↓0.1537	↓-33%
w3_p20_s30\lightning_logs\version_0	0.3058	0.4741	0.4741	0.3058	↓0.1683	↓-36%
w0.5_p20_s30\lightning_logs\version_0	0.3064	0.4396	0.4396	0.3162	↓0.1233	↓-28%
w2_p10_s30\lightning_logs\version_0	0.3067	0.4555	0.4555	0.3067	↓0.1488	↓-33%
w2_p20_s10\lightning_logs\version_0	0.3094	0.4579	0.4579	0.3094	↓0.1485	↓-32%
w1_p5_s20\lightning_logs\version_0	0.3108	0.448	0.448	0.3108	↓0.1372	↓-31%
w0.5_p5_s10\lightning_logs\version_0	0.3113	0.444	0.444	0.3113	↓0.1327	↓-30%
w1_p10_s10\lightning_logs\version_0	0.3134	0.4467	0.4467	0.3134	↓0.1333	↓-30%
w2_p10_s20\lightning_logs\version_0	0.3139	0.4559	0.4559	0.3139	↓0.142	↓-31%
w2_p5_s20\lightning_logs\version_0	0.3147	0.4558	0.4558	0.3147	↓0.141	↓-31%
w1_p5_s10\lightning_logs\version_0	0.3151	0.4454	0.4454	0.3151	↓0.1304	↓-29%
w2_p5_s30\lightning_logs\version_0	0.3157	0.456	0.456	0.3157	↓0.1403	↓-31%
w2_p10_s10\lightning_logs\version_0	0.3187	0.4592	0.4592	0.3222	↓0.137	↓-30%
w2_p5_s10\lightning_logs\version_0	0.3951	0.453	0.453	0.3951	↓0.0579	↓-13%

Figure A1: Full grid search results for the prior-guided model, showing validation loss over training for different combinations of prior weight λ, plateau radius r, and Gaussian smoothness σ. The best configuration was selected based on the lowest validation loss.

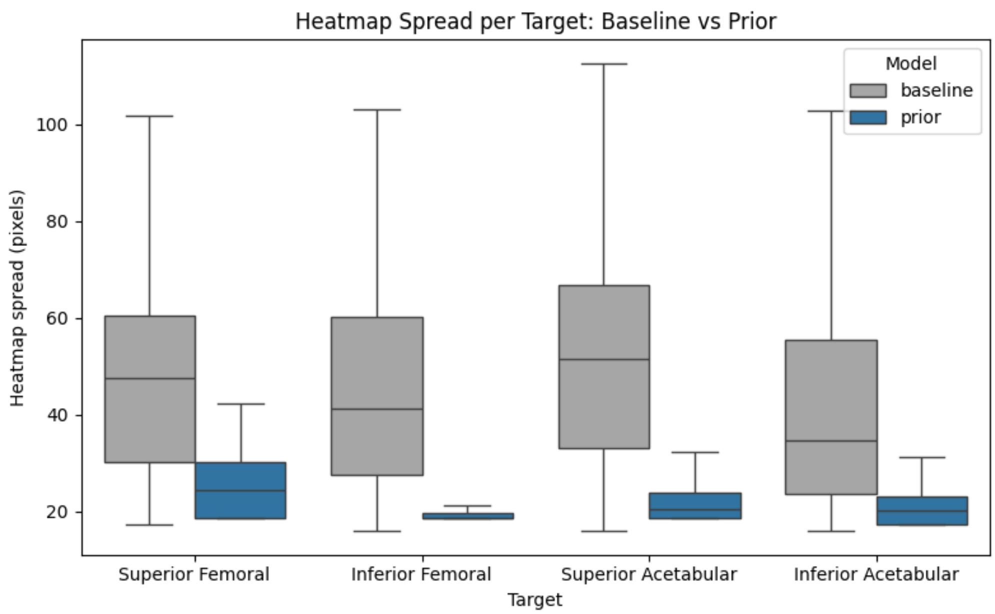


Figure A2: Spread distributions for the baseline and prior-guided models on positive target labels only. Lower values indicate more spatially concentrated Grad-CAM heatmaps. The prior-guided model shows consistently lower spread across all targets, indicating that the anatomical prior produces more compact heatmaps.

Table A1: Wilcoxon signed-rank analysis of localization and compactness metrics over 13 repeated runs, using different seed for the model dataset split. Values are reported as the mean metric value across runs for the baseline and prior-guided models. Differences are computed as prior minus baseline, so negative values indicate lower localization distance or more compact heatmaps for the prior-guided model. For all metrics, the prior-guided model improved over the baseline in all 13 runs, resulting in identical Wilcoxon signed-rank test results of $W = 0.0$ and $p = 0.000244$.

Metric	Target	n_{runs}	Baseline	Prior	Mean diff.	Improved	p-value
CoM	Femoral sup.	13	51.6	23.4	−28.2	13/13	0.000244
	Femoral inf.	13	61.6	22.4	−39.2	13/13	0.000244
	Acetabular sup.	13	88.4	21.8	−66.6	13/13	0.000244
	Acetabular inf.	13	90.0	22.8	−67.1	13/13	0.000244
Peak	Femoral sup.	13	53.0	23.3	−29.7	13/13	0.000244
	Femoral inf.	13	64.5	24.8	−39.7	13/13	0.000244
	Acetabular sup.	13	90.0	22.1	−67.9	13/13	0.000244
	Acetabular inf.	13	90.7	26.4	−64.3	13/13	0.000244
Spread	Femoral sup.	13	53.9	31.8	−22.1	13/13	0.000244
	Femoral inf.	13	55.2	26.4	−28.8	13/13	0.000244
	Acetabular sup.	13	53.6	23.8	−29.8	13/13	0.000244
	Acetabular inf.	13	50.4	23.2	−27.2	13/13	0.000244

Table A2: Overview of all repeated evaluation runs. Each run used the same seed for both the model training and the data split. The table reports model-level localization and compactness metrics for the baseline and prior-guided models. Differences are computed as prior minus baseline, so negative values indicate lower localization distance or more compact heatmaps for the prior-guided model.

Seed	Metric	Target	Baseline mean	Prior mean	Diff.
0	CoM	Femoral sup.	58.2	24.3	−33.9
0	CoM	Femoral inf.	54.3	22.8	−31.5
0	CoM	Acetabular sup.	89.0	21.7	−67.3
0	CoM	Acetabular inf.	76.8	18.2	−58.6
0	Peak	Femoral sup.	58.2	22.8	−35.4
0	Peak	Femoral inf.	54.1	24.7	−29.5
0	Peak	Acetabular sup.	91.3	21.6	−69.7
0	Peak	Acetabular inf.	78.5	18.9	−59.6
0	Spread	Femoral sup.	55.3	30.9	−24.5
0	Spread	Femoral inf.	53.8	26.5	−27.2
0	Spread	Acetabular sup.	57.7	23.6	−34.1
0	Spread	Acetabular inf.	45.4	23.0	−22.4
1	CoM	Femoral sup.	55.3	24.1	−31.2
1	CoM	Femoral inf.	53.7	21.5	−32.1

Continued on next page

Table A2: Overview of all repeated evaluation runs continued.

Seed	Metric	Target	Baseline mean	Prior mean	Diff.
1	CoM	Acetabular sup.	92.2	22.1	−70.1
1	CoM	Acetabular inf.	89.1	24.0	−65.1
1	Peak	Femoral sup.	58.3	23.4	−34.9
1	Peak	Femoral inf.	57.6	24.1	−33.5
1	Peak	Acetabular sup.	90.9	22.4	−68.5
1	Peak	Acetabular inf.	88.6	27.2	−61.4
1	Spread	Femoral sup.	52.8	31.5	−21.3
1	Spread	Femoral inf.	53.3	26.9	−26.4
1	Spread	Acetabular sup.	50.6	23.8	−26.9
1	Spread	Acetabular inf.	48.2	22.8	−25.4
2	CoM	Femoral sup.	53.9	23.3	−30.5
2	CoM	Femoral inf.	58.6	22.0	−36.6
2	CoM	Acetabular sup.	90.9	22.4	−68.5
2	CoM	Acetabular inf.	95.0	24.2	−70.9
2	Peak	Femoral sup.	55.2	23.6	−31.6
2	Peak	Femoral inf.	62.1	24.8	−37.3
2	Peak	Acetabular sup.	94.5	23.3	−71.2
2	Peak	Acetabular inf.	94.0	28.9	−65.1
2	Spread	Femoral sup.	50.0	33.1	−16.8
2	Spread	Femoral inf.	53.4	28.2	−25.2
2	Spread	Acetabular sup.	52.1	24.8	−27.3
2	Spread	Acetabular inf.	44.6	22.9	−21.6
3	CoM	Femoral sup.	48.8	22.3	−26.4
3	CoM	Femoral inf.	71.4	21.3	−50.1
3	CoM	Acetabular sup.	75.6	21.4	−54.2
3	CoM	Acetabular inf.	100.4	24.6	−75.8
3	Peak	Femoral sup.	50.3	23.5	−26.8
3	Peak	Femoral inf.	76.7	25.1	−51.6
3	Peak	Acetabular sup.	76.7	22.1	−54.6
3	Peak	Acetabular inf.	100.9	31.6	−69.3
3	Spread	Femoral sup.	53.4	32.2	−21.1
3	Spread	Femoral inf.	55.4	27.7	−27.7
3	Spread	Acetabular sup.	52.4	24.2	−28.1
3	Spread	Acetabular inf.	51.3	24.0	−27.3
4	CoM	Femoral sup.	59.3	23.0	−36.2
4	CoM	Femoral inf.	55.2	22.3	−32.9
4	CoM	Acetabular sup.	93.3	21.9	−71.5
4	CoM	Acetabular inf.	75.2	22.6	−52.6
4	Peak	Femoral sup.	60.8	23.5	−37.2
4	Peak	Femoral inf.	57.3	24.9	−32.4
4	Peak	Acetabular sup.	94.9	22.2	−72.8
4	Peak	Acetabular inf.	79.4	26.1	−53.3
4	Spread	Femoral sup.	59.0	33.0	−26.0
4	Spread	Femoral inf.	55.1	25.2	−29.9

Continued on next page

Table A2: Overview of all repeated evaluation runs continued.

Seed	Metric	Target	Baseline mean	Prior mean	Diff.
4	Spread	Acetabular sup.	53.4	23.8	−29.7
4	Spread	Acetabular inf.	56.8	23.4	−33.4
5	CoM	Femoral sup.	54.6	23.2	−31.4
5	CoM	Femoral inf.	46.9	21.4	−25.5
5	CoM	Acetabular sup.	94.7	21.5	−73.1
5	CoM	Acetabular inf.	84.7	22.4	−62.3
5	Peak	Femoral sup.	55.8	23.3	−32.5
5	Peak	Femoral inf.	46.0	24.6	−21.4
5	Peak	Acetabular sup.	97.5	22.3	−75.2
5	Peak	Acetabular inf.	82.8	24.8	−58.0
5	Spread	Femoral sup.	54.8	31.5	−23.3
5	Spread	Femoral inf.	55.7	27.2	−28.5
5	Spread	Acetabular sup.	61.7	23.6	−38.1
5	Spread	Acetabular inf.	45.8	24.4	−21.4
6	CoM	Femoral sup.	52.4	23.3	−29.0
6	CoM	Femoral inf.	52.3	22.1	−30.2
6	CoM	Acetabular sup.	98.6	22.6	−76.0
6	CoM	Acetabular inf.	75.3	20.7	−54.6
6	Peak	Femoral sup.	54.3	23.3	−31.0
6	Peak	Femoral inf.	55.3	24.8	−30.5
6	Peak	Acetabular sup.	99.0	22.4	−76.6
6	Peak	Acetabular inf.	75.0	22.0	−53.0
6	Spread	Femoral sup.	52.8	32.7	−20.0
6	Spread	Femoral inf.	53.9	27.3	−26.6
6	Spread	Acetabular sup.	52.3	23.8	−28.5
6	Spread	Acetabular inf.	44.5	23.0	−21.6
7	CoM	Femoral sup.	49.9	23.2	−26.7
7	CoM	Femoral inf.	60.5	23.1	−37.5
7	CoM	Acetabular sup.	77.5	22.1	−55.4
7	CoM	Acetabular inf.	95.6	24.3	−71.3
7	Peak	Femoral sup.	48.4	23.2	−25.3
7	Peak	Femoral inf.	61.6	25.0	−36.5
7	Peak	Acetabular sup.	80.5	22.4	−58.1
7	Peak	Acetabular inf.	96.1	30.3	−65.8
7	Spread	Femoral sup.	52.1	31.6	−20.6
7	Spread	Femoral inf.	54.6	24.8	−29.8
7	Spread	Acetabular sup.	56.5	24.2	−32.3
7	Spread	Acetabular inf.	53.5	23.0	−30.4
8	CoM	Femoral sup.	58.9	23.5	−35.4
8	CoM	Femoral inf.	51.8	22.5	−29.3
8	CoM	Acetabular sup.	93.0	21.4	−71.5
8	CoM	Acetabular inf.	99.1	23.0	−76.1
8	Peak	Femoral sup.	62.1	22.8	−39.3
8	Peak	Femoral inf.	53.1	25.4	−27.7

Continued on next page

Table A2: Overview of all repeated evaluation runs continued.

Seed	Metric	Target	Baseline mean	Prior mean	Diff.
8	Peak	Acetabular sup.	96.4	21.4	−74.9
8	Peak	Acetabular inf.	98.5	27.1	−71.3
8	Spread	Femoral sup.	49.7	31.6	−18.1
8	Spread	Femoral inf.	46.2	27.5	−18.7
8	Spread	Acetabular sup.	48.5	23.4	−25.2
8	Spread	Acetabular inf.	44.5	23.7	−20.9
9	CoM	Femoral sup.	45.5	23.3	−22.2
9	CoM	Femoral inf.	83.8	22.7	−61.1
9	CoM	Acetabular sup.	87.7	21.7	−66.0
9	CoM	Acetabular inf.	94.2	25.6	−68.6
9	Peak	Femoral sup.	46.3	23.2	−23.1
9	Peak	Femoral inf.	93.5	24.6	−68.9
9	Peak	Acetabular sup.	87.5	22.0	−65.5
9	Peak	Acetabular inf.	93.0	30.7	−62.3
9	Spread	Femoral sup.	51.6	31.4	−20.3
9	Spread	Femoral inf.	59.6	26.4	−33.2
9	Spread	Acetabular sup.	52.7	23.7	−29.1
9	Spread	Acetabular inf.	52.5	23.6	−28.8
10	CoM	Femoral sup.	46.9	23.6	−23.3
10	CoM	Femoral inf.	77.1	22.8	−54.3
10	CoM	Acetabular sup.	83.2	21.5	−61.7
10	CoM	Acetabular inf.	96.0	21.1	−74.9
10	Peak	Femoral sup.	45.6	23.6	−22.0
10	Peak	Femoral inf.	84.6	24.5	−60.1
10	Peak	Acetabular sup.	82.6	21.8	−60.8
10	Peak	Acetabular inf.	95.4	22.1	−73.3
10	Spread	Femoral sup.	52.8	30.9	−21.9
10	Spread	Femoral inf.	54.7	25.5	−29.2
10	Spread	Acetabular sup.	50.8	23.5	−27.3
10	Spread	Acetabular inf.	53.5	21.7	−31.8
11	CoM	Femoral sup.	40.4	24.0	−16.4
11	CoM	Femoral inf.	83.4	22.2	−61.2
11	CoM	Acetabular sup.	76.9	21.6	−55.3
11	CoM	Acetabular inf.	118.7	22.2	−96.5
11	Peak	Femoral sup.	46.7	23.6	−23.1
11	Peak	Femoral inf.	82.9	24.4	−58.5
11	Peak	Acetabular sup.	82.6	21.8	−60.8
11	Peak	Acetabular inf.	123.7	26.2	−97.5
11	Spread	Femoral sup.	59.8	31.4	−28.4
11	Spread	Femoral inf.	65.1	26.2	−38.9
11	Spread	Acetabular sup.	50.7	23.6	−27.1
11	Spread	Acetabular inf.	61.8	23.5	−38.3
123	CoM	Femoral sup.	47.0	23.5	−23.5
123	CoM	Femoral inf.	51.6	23.8	−27.9

Continued on next page

Table A2: Overview of all repeated evaluation runs continued.

Seed	Metric	Target	Baseline mean	Prior mean	Diff.
123	CoM	Acetabular sup.	96.3	21.5	−74.9
123	CoM	Acetabular inf.	69.3	23.7	−45.6
123	Peak	Femoral sup.	47.4	23.4	−24.0
123	Peak	Femoral inf.	53.7	25.2	−28.6
123	Peak	Acetabular sup.	96.1	21.7	−74.4
123	Peak	Acetabular inf.	73.2	26.8	−46.5
123	Spread	Femoral sup.	57.1	32.4	−24.7
123	Spread	Femoral inf.	57.0	23.4	−33.6
123	Spread	Acetabular sup.	57.4	23.5	−33.9
123	Spread	Acetabular inf.	52.6	22.7	−29.9