What Limits Mechanical-Axis Accuracy in Full-Length Femur Reconstruction from Biplanar Radiographs?

Ashish Kumar Pokharel

ashishkumar.pokharel@gmail.com

Abstract. Deep learning can reconstruct the femur in 3D from two radiographs, but such methods are evaluated mainly with overlap metrics on the proximal femur, which do not show how well clinical measurements are recovered. We asked what limits the accuracy of a whole-bone measurement: the femoral mechanical axis, from hip centre to knee centre. We reconstructed full-length femurs from digitally reconstructed radiographs (DRRs) of 101 CT scans (5-fold cross-validation) with a small and a large network (AttentionUNet and SwinUNETR), and tested on 58 external femurs, 19 with expert landmarks. With pose augmentation, the median axis error was 0.53° for AttentionUNet and 0.55° for SwinUNETR internally, against an inter-expert disagreement of 0.30° and an error of 0.29° for landmark extraction from ground-truth masks. On the external femurs with expert landmarks, errors were 0.69° and 0.62°, with 7/19 and 6/19 exceeding 2°, the largest in femurs tilted beyond the augmented pose range. A 2D baseline that detects the landmarks directly on the radiographs was not shown to be better than 3D on either cohort (0.50–1.01° externally across training variants), and across all methods the added error came mainly from locating the hip centre (2.3–3.6 mm), not the knee. Without pose augmentation, AttentionUNet mislocated the femoral head ($>$ 20 mm) in 9/101 cases and SwinUNETR in none, at nearly equal median Dice. With idealised inputs, the main improvable target is hip-centre localisation, not a better architecture; and since the median Dice (not per-case Dice) missed these failures, reconstruction studies should report failure rates alongside overlap.

Keywords: Biplanar radiographs · 3D reconstruction · Femur · Mechanical axis · Landmark detection · Validation

1 Introduction

Lower-limb alignment is central to the planning of corrective osteotomy and knee or hip arthroplasty [24]. Its main reference is the femoral mechanical axis, the line from the centre of the femoral head to the centre of the knee. In practice it is measured on long-leg radiographs, which are two-dimensional and sensitive to limb flexion and rotation [6], or on CT, which gives a full 3D model of the bone but costs more, delivers a higher radiation dose and is not available everywhere. Recovering a 3D model of the whole femur from two ordinary radiographs, taken from the front and from the side, would combine the low cost of radiography with the measurements that a 3D model allows. Low-dose biplanar systems such as EOS already acquire simultaneous frontal and lateral images of the standing patient in clinical practice [11]. However,

such systems require dedicated and costly equipment that many hospitals, particularly in low-resource settings, do not have. We therefore study a method that would need only two radiographs and modest computing; on synthetic inputs we also check that inference can run on low-cost CPU-only hardware, though a deployment claim would require real radiographs.

Deep learning has made this kind of biplanar reconstruction practical [9, 18, 20, 27, 29, 32]. However, many studies report overlap scores such as Dice on synthetic radiographs, and several crop the femur to its proximal part. The open benchmark of Shakya and Khanal [29], for instance, uses a 128 mm proximal femur crop that cannot see the knee, and notes that better Dice does not always mean better clinical parameters. When a clinical measurement is the goal, it is therefore unclear what limits its accuracy. Three factors are often not isolated: the network architecture, whose newer variants are expected to reconstruct better; the reconstruction quality as measured by overlap; and the step that extracts landmarks from the reconstruction, which is rarely evaluated on its own.

Each factor is hard to assess. Differences in accuracy and robustness between convolutional and transformer networks can shrink when training configurations are matched [4, 21], carefully configured convolutional baselines often match newer architectures [16, 17], and the choice of metric can change the conclusion [22]. We therefore study the mechanical axis of the whole femur and separate the three factors. We compare a convolutional and a transformer model on identical data and folds across four training configurations; in the two final ones they share the optimiser, loss, augmentation and model-selection rule, with only the epoch limit differing. We relate overlap to clinical errors and failure counts. Finally, we isolate landmark extraction with a measurement floor, a learned 3D landmark extractor and a 2D landmark baseline that bypasses 3D reconstruction.

Contributions. We make three findings on a full-length femur study (101 CTs, 5-fold cross-validation, external test on 58 femurs), with public code, splits and weights.[1]

- Decomposing each landmark error along and across the axis (with a measurement floor on true masks, a learned 3D extractor and a 2D triangulation baseline) shows the across-axis error that sets the angle comes about equally from hip and knee, but what image-based methods *add* over the floor falls on the *femoral head*; the knee-rule error lies along the femur and does not tilt the axis. Direct 2D detection was not shown to be better than 3D on either cohort (bracketing it internally and externally, 0.50–1.01° externally across training variants), so the hip error is not specific to reconstruction.
- A controlled CNN (AttentionUNet) versus transformer (SwinUNETR) comparison across four training configurations: the transformer's apparent reliability advantage under weaker configurations was closed by the final configuration, so it reflected training more than architecture, though SwinUNETR was less sensitive to the configuration.
- Evidence that the median Dice was insensitive to the head-localisation failures that drove the clinical error, though per-case Dice flagged them.

[1] https://github.com/ayyp1/biplanar-xray-to-3d-femur

2 Related Work

Bone reconstruction from biplanar radiographs. X2CT-GAN [32] reconstructs CT volumes from two views, and Kasten et al. [18] train an end-to-end CNN for knee bones by repeating each radiograph along its ray direction to form a 3D input. Shakya and Khanal [29] benchmark eight encoder–decoder models with this fusion on six public datasets and four anatomies, using synthetic radiographs as input. They report Dice, distance metrics and clinical parameters such as femoral head radius and neck–shaft angle, and attention-based models perform best. Later work introduces newer architectures, such as Swin transformers [20] and Cross-RWKV blocks [27]; alternative shape representations, such as occupancy fields [9] and single-view neural implicit models [14]; and broader settings, such as multi-organ reconstruction [26] and very-low-dose imaging [19]. A recent systematic review covers 2D–3D methods for the knee [13]. Most of these works report overlap, and few measure a clinical quantity that spans the whole femur. We extend the benchmark of [29] to the full-length femur and test whether its architecture ranking holds once training configurations are comparable.

Lower-limb models and clinical measurements. The EOS line of work reconstructs the lower limb from real low-dose biplanar radiographs with statistical shape models and contour matching [8, 11]. Babazadeh Khameneh et al. [3] register a whole-femur template to real EOS radiographs with a CNN and report clinical 3D measurements, Picard et al. [28] propose a joint 2D–3D statistical shape model of the femur, and Pan et al. [25] reconstruct the tibia and fibula without 3D supervision. Arn Roth et al. [2], with networks trained on DRRs and applied to EOS radiographs, report a mean difference of 0.52° in the lower-limb (hip–knee–ankle) mechanical axis for tibial osteotomy planning (a mean on real images, not directly comparable with our medians on DRRs); their joint centres come from a 2D landmark network rather than from a reconstructed 3D bone, and their axis differs from the hip-to-knee femoral axis evaluated here. Van den Borre et al. [5] automate lower-limb measurements on weight-bearing CT. These methods rely on shape models, templates or 2D landmarks; to our knowledge, none compares end-to-end volumetric networks on a mechanical axis measured from the reconstructed whole femur.

Synthetic radiographs. Digitally reconstructed radiographs (DRRs), computed by projecting CT, are a common way to obtain paired radiographs and 3D ground truth for supervised training. Realistic simulation [30] and domain randomisation [12] help models transfer to real images. We use parallel-beam DRRs, which are simpler than real acquisition geometry; we discuss this as a limitation.

3 Methods

3.1 Data

Figure 1 gives an overview of the pipeline.

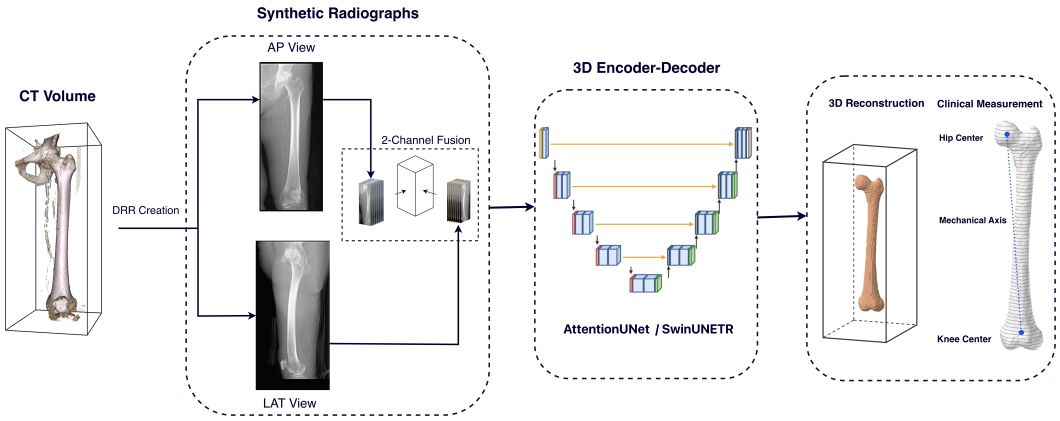


Fig. 1. Overview of our reconstruction-and-measurement pipeline, from biplanar DRRs to the mechanical axis.

HFValid. The Hip-Fracture validation collection [1] contains CT scans of one femur from each of 101 women aged 56–86 years (median 67; 64 left, 37 right), acquired on three scanner models from two vendors. All femurs were intact at the time of the scan, and subjects with implants, metal artefacts or hip deformities on the modelled side were excluded. Each femur has a surface segmentation, produced semi-automatically and manually corrected by experts, and expert landmarks. In-plane spacing is 0.59–0.98 mm (median 0.74 mm) and slice spacing 1–2 mm (median 2 mm). We use the femoral head centre (the centre of the femoral head) and the knee centre (the midpoint of the most medial and most lateral epicondyle points, a transepicondylar definition, which can differ from the notch-centre knee point used on long-leg radiographs [24]). The surfaces and landmarks are in the scanner's physical left–posterior–superior (LPS) frame, which we confirmed visually for every case (Fig. 10). We chose HFValid because it is, to our knowledge, one of the few public datasets that combine CTs of the whole femur, including the knee, with expert segmentations and expert hip- and knee-centre landmarks, which the mechanical axis requires.

VSD (external test). The lower-body collection of the Virtual Skeleton Database (VSD) [10] provides CT scans and bone surface models of 30 cadavers of both sexes, acquired on different scanners. We use 58 femurs; one subject could not be preprocessed. For 19 femurs, expert landmarks from five raters with four repetitions each are available; the hip centre is a sphere fitted through six femoral head points and the knee centre the midpoint of the medial and lateral epicondyles; the reference is the mean over all raters and repetitions. VSD is used only for testing. It serves as an external test because it differs from HFValid in population (cadavers of both sexes), scanners and landmark protocol, and its multi-rater landmarks give an independent clinical reference.

3.2 Preprocessing and Synthetic Radiographs

Each expert surface is rasterised to a mask on the CT grid. Right femurs are mirrored so that every femur looks like a left one. CT and mask are then resampled to a common box of $96 \times 128 \times 256$ voxels (x, y, z) at 2 mm, centred on the bounding box of the femur mask, with linear interpolation for CT. The mask is interpolated linearly and thresholded at 0.5, which avoids the one-voxel shift of mirrored femurs that nearest-neighbour interpolation causes. The box covers $192 \times 256 \times 512$ mm, so the whole femur fits in every case; 2 mm matches the median slice spacing, keeps the volume small enough for full-resolution training, and makes every dimension divisible by 32 as SwinUNETR requires.

We use synthetic radiographs because no public dataset pairs real biplanar radiographs of the whole femur with CT-based 3D ground truth and landmarks [20, 29]; computing them from the same CT gives an exact 2D–3D correspondence and is the common approach in prior work [2, 18, 20, 29]. Synthetic anteroposterior (AP) and lateral (LAT) radiographs are computed by parallel projection along the volume axes:

$$\begin{gathered} \mu = \max\!\big(0,\ 0.02\,\big(1 + \tfrac{\mathrm{HU}}{1000}\big)\big)\,, \\ I_{\mathrm{AP}}(z,x) = \Delta \sum_{y} \mu(z,y,x), \qquad I_{\mathrm{LAT}}(z,y) = \Delta \sum_{x} \mu(z,y,x), \end{gathered} \tag{1}$$

with $\Delta = 2$ mm, and divided by a fixed constant of 6.0, set just above the largest HFValid value (5.6) to avoid clipping and reused unchanged for VSD. Parallel projection keeps an exact pixel-to-voxel correspondence, which the fusion input relies on.

For VSD, the coordinate frames of the meshes and CTs are not consistent across subjects, so each femur is oriented automatically from the anatomy: the left femur lies at $+x$ of the right one, the patella is anterior to the femur, and the femoral head is proximal. The orientation of every femur was then checked visually (Fig. 9). We crop the CT around the femur with the median margins of the HFValid scans, which removes the other leg, and fill the hollow marrow canal of the raw meshes before rasterising them.

3.3 Input Fusion and Models

Each radiograph is repeated along its own ray direction to form a 3D volume, and the two volumes are stacked as a 2-channel input of size $2 \times 256 \times 128 \times 96$ (channels $\times\, z \times y \times x$), as in [18, 29]. We compare AttentionUNet [23] (channels 16–128, three downsampling steps, 1.5M parameters), the best femur model in [29], with SwinUNETR [15], a widely used transformer-based 3D segmentation network built on Swin transformers, as is the recent biplanar method Swin-X2S [20] (feature size 48, gradient checkpointing, 62M parameters), both from MONAI [7]. The networks output one logit per voxel; a sigmoid gives a probability, which is thresholded at 0.5. Connected components smaller than 5% of the volume of the largest component are removed, which deletes stray voxels but keeps a femoral head that is separated from the shaft by a thin gap. The same post-processing is used for every model. On HFValid, each patient is predicted by the model of the fold that held it out; on VSD, the probabilities of the five fold models are averaged before thresholding.

3.4 Training Configurations

All models are trained with 5-fold cross-validation on HFValid using one fixed split (seed 42), saved to disk so that every model sees identical folds. Table 1 lists the four configurations. Configuration A gives AttentionUNet the optimiser and learning rate of SwinUNETR v1. Configuration B adds longer training, cosine decay and mild augmentation. Configuration C adds CT-based pose augmentation with re-projection, stronger intensity augmentation and per-image normalisation.

Table 1. Training configurations. All use the same 5 folds, Dice + cross-entropy loss, batch size 1 and mixed precision, and keep, for each fold, the epoch with the best validation Dice.

Configuration	Setting
v1	Both models. AttentionUNet: Adam, lr 10^{-3}; SwinUNETR: AdamW, lr 10^{-4}. 60 epochs, patience 12, no augmentation, fixed intensity scaling.
A	AttentionUNet with the SwinUNETR v1 optimiser and learning rate (AdamW, lr 10^{-4}).
B	Both models. A + cosine learning rate, up to 150 epochs (SwinUNETR 80), patience 20, mild intensity augmentation and a random 3D shift of up to ± 4 voxels applied consistently to both views and the mask.
C	Both models. B + CT pose augmentation with re-projection ($p = 0.8$), wider intensity augmentation and per-image z-score normalisation (details below).

Configuration C. With probability 0.8 per training sample, the CT and mask are rotated together about the box centre (flexion $\pm 15°$, abduction $\pm 10°$, axial rotation $\pm 15°$), shifted by up to ± 8 mm, and the DRRs are recomputed from the rotated CT. The model thus sees radiographs of femurs in new poses rather than rotated images. Each view then receives gamma (0.6–1.6), contrast (0.7–1.3) and brightness (± 0.1) changes, Gaussian blur ($p = 0.3$, σ 0.3–1.5 pixels) and Gaussian noise ($p = 0.5$, s.d. up to 0.04), and is z-score normalised. AttentionUNet is trained for up to 150 epochs and SwinUNETR for 80 because of its higher cost (about 4.3 versus 1.9 GPU-hours per fold on an NVIDIA T4). Both use AdamW (lr 10^{-4}, weight decay 10^{-5}) with cosine decay and early stopping after 20 epochs without improvement. SwinUNETR ran to its 80-epoch limit in every fold, with its best epoch between 72 and 79 and a gain of only about 0.003 in validation Dice over the last 20 epochs; longer training might have improved it slightly. Appendix A lists all settings.

3.5 Landmark Extraction and Measurement Floor

Hip centre. The surface voxels of the mask (mask minus its erosion) are restricted to the top 35% of the femur's height and the medial 60% of its width, which keeps the femoral head and excludes the greater trochanter. A sphere is fitted with RANSAC (500 random 4-point samples, radius 18–28 mm, inlier tolerance 4 mm), and the best sphere is refined by linear least squares on its inliers with tolerances of 4, 3 and 2 mm. Its centre is the hip centre.

Knee centre. In the distal 30% of the mask, we find the axial level where the femur is widest from medial to lateral and take the midpoint between the two edges, averaged over ± 2 slices, with the antero-posterior position given by the mean of that level. Because the expert knee centre lies slightly above and behind the widest level, a constant offset (about 8 mm: 6 mm posterior and 5 mm proximal) is added. It is learned on the true masks of the training folds only, and the value learned on all of HFValid is used for VSD. Figure 2 shows both steps on one case, and Appendix B gives further details.

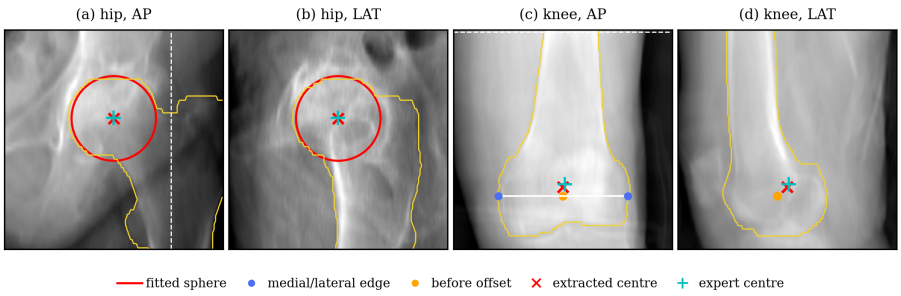


Fig. 2. Landmark extraction on one HFValid case, shown on DRRs rendered at 0.5 mm with background suppression for display, with the true femur outline (yellow). (a, b) Hip: sphere fitted to the proximal–medial surface (dashed line: medial cut). (c, d) Knee: widest medial–lateral level within the distal 30% (dashed line: its upper limit), its midpoint before (orange) and after the offset (red).

Axis error and floor. The mechanical axis is the vector from the knee centre to the hip centre, and the axis error is the 3D angle between the predicted and the expert axis. We also report its coronal (frontal-plane, varus–valgus) and sagittal components: the signed differences between the angles of the two axes projected onto these planes, in the canonical left-femur frame, reported as absolute values. Hip and knee errors are Euclidean distances in mm. Running the same extractor on the *true* masks gives the *measurement floor*: the extractor's own error with a perfect reconstruction. It serves as a reference for this extractor rather than a strict lower bound, and comparing it with a model's error indicates, descriptively, how much error the reconstruction adds.

Learned 3D extractor. To test whether the hand-crafted rules limit the axis, a learned extractor predicts both centres from a 3D mask. For each fold, a 3D U-Net (MONAI, channels 8–128, two output heatmaps) is trained on the true masks of the training folds with the spatial-softmax cross-entropy of the 2D baseline (Sect. 3.6, Gaussian targets with $\sigma = 2$ voxels), and the landmark is the probability-weighted centroid around the heatmap maximum. So that it transfers to reconstructions, training masks receive the pose augmentation of configuration C and random corruptions that imitate reconstruction errors: smoothing and re-thresholding, erosion or dilation, and small spherical bumps and bites on the surface. Training uses AdamW (lr 10^{-3}) with cosine decay for

at most 80 epochs, with early stopping (patience 15) on the mean landmark error on the true masks of an inner validation split: 12 patients drawn from the training folds (seed 42+fold) and excluded from training, so the outer fold is used only for evaluation. Predicted masks are used neither for training nor for selection. Each HFValid reconstruction is evaluated with the extractor of its own fold; on VSD, the heatmaps of the five fold extractors are averaged. Appendix B gives the settings of both learned extractors.

Error geometry. To see why a landmark error does or does not tilt the axis, each hip and knee error vector is split into a component along the expert axis and a component across it. To first order, the axis error equals the across-axis component of the difference between the hip and knee errors divided by the femur length (a small-angle approximation, with the femur length taken as each case's expert hip-to-knee distance), so errors along the axis barely change the angle. We also test whether the hip error is a consistent bias or a scatter: the mean signed error is subtracted, with the mean estimated on the other folds only, and the errors are recomputed. Finally, the radius of the fitted head sphere on each reconstruction is compared with that on the true mask, to test whether the reconstructed head has the right size. Observer variability is estimated on the 19 VSD femurs with expert landmarks. Each of the 20 observations (five raters, four repetitions) gives a hip centre, the sphere through the six head points, and a knee centre. Because the HFValid reference is a single annotation per femur [1], the methods are compared with the disagreement between two observations by different raters, split in the same way and converted to an axis difference; the 3040 pairs come from only 19 femurs and are not independent. For the VSD reference, which is the consensus, each observation is instead compared with the mean of the other 19 observations. Finally, as a post-hoc check of why external cases fail, each femur's tilt is measured on the ground-truth mask as the angle between the mask's principal axis and the box long axis, split into abduction (coronal) and flexion (sagittal) components; its association with Dice is summarised per configuration-C model by Spearman's ρ ($n = 58$).

3.6 2D Landmark Baseline

To separate landmark extraction from reconstruction, a 2D baseline predicts the hip and knee centres directly on each radiograph, without any 3D reconstruction. For each fold, one 2D U-Net per view (MONAI, channels 16–256, two output heatmaps) is trained with a spatial-softmax cross-entropy against Gaussian targets ($\sigma = 2$ pixels) centred on the expert landmarks projected onto that view; the predicted landmark is the probability-weighted centroid around the heatmap maximum. Because the views are orthogonal parallel projections, the 3D point follows directly: x from the AP view, y from the lateral view and z as the mean of both. Training uses the same folds, pose and intensity augmentation and normalisation as configuration C, AdamW (lr 10^{-3}) with cosine decay (schedule length 150 epochs). The primary 2D checkpoint is selected per fold on the held-out outer fold (lowest mean landmark error), as the 3D checkpoints are; because this selects on nearly the evaluated quantity, we also report three less optimistic variants as a selection bracket (Sect. 4.5): a nested variant that selects the epoch on an inner validation split the outer fold never enters, a refit variant that takes the same epoch

count and retrains on all training patients, and a fixed-schedule run with no checkpoint selection. On VSD, the heatmaps of the five fold models are averaged. Unlike the rule-based 3D extractor, the baseline learns the expert knee definition directly and needs no offset. As a post-hoc resolution check, the fixed-schedule 2D baseline was trained and evaluated at 1 mm pixel size and compared with a matched 2 mm run (box size in voxels and the Gaussian target width scaled accordingly), to test whether 2 mm pixels limit hip localisation.

3.7 Metrics and Statistics

Overlap is measured with Dice, the 95th-percentile Hausdorff distance (HD95), the average symmetric surface distance (ASD) and the normalised surface distance at 2 mm (NSD) [22]. Clinical metrics are the axis error (primary endpoint), the hip and knee errors, the proportion of cases below 1° and 2°, and the maximum error. A *head failure* (head-localisation failure) occurs when the predicted hip centre is more than 20 mm from its reference. On HFValid, the reference is the expert hip centre; on VSD, it is the hip centre extracted from the true mask, because expert landmarks exist for only 19 femurs. This operational threshold, comparable to the femoral head radius (our sphere fit allows 18–28 mm), flags substantial localisation errors and is not a validated clinical cut-off; it was chosen after inspecting the first results, and Sect. 4.2 reports its sensitivity. All values are per patient and summarised as median [interquartile range]. Each pairwise comparison of two models on the same cases uses two-sided Wilcoxon signed-rank tests on the per-patient differences of the seven continuous metrics (Dice, HD95, ASD, NSD, hip, knee and axis error), Holm-corrected within that comparison (not across comparisons), and 95% confidence intervals of the median paired difference from 10 000 bootstrap resamples of patients with replacement, resampling paired observations together. For comparisons without overlap metrics (2D baseline), the Holm family is the hip, knee and axis errors. Failure counts are compared with an exact McNemar test. For the final models, equivalence of the axis error is tested with two one-sided Wilcoxon signed-rank tests (TOST, margin $\pm 0.5^\circ$, a pragmatic choice rather than a validated clinical threshold, $\alpha = 0.05$) on the paired differences shifted by $+0.5^\circ$ and -0.5°. Zero differences are discarded and ties given average ranks, so the tested location is the pseudo-median. Equivalence is concluded if both tests are significant. The association between Dice and axis error is measured with Spearman's ρ for each model separately. Proportions of cases below a threshold and the maximum error are descriptive and not tested. On VSD, the two femurs of one cadaver are treated as independent, which may understate the uncertainty of the 58-femur comparisons; the 19 femurs with landmarks come from different cadavers.

4 Results

4.1 Cross-validation on HFValid

Table 2 summarises the final models. Because checkpoints were selected on the held-out fold, the HFValid values are validation-selected estimates (see Limitations); VSD is the

independent test. Both models reach a median Dice of 0.94 and an axis error of about half a degree, with no head failures. The paired axis difference (AttentionUNet minus SwinUNETR) is -0.01° (95% CI -0.13 to 0.03), not significant after correction. This interval excludes differences larger than about 0.13°, and TOST also concludes equivalence within the pragmatic $\pm 0.5^\circ$ margin ($p < 0.001$). The only significant difference is the hip error (median paired difference -0.29 mm, 95% CI -0.55 to -0.04 mm, $p_{\mathrm{Holm}} = 0.041$), in favour of AttentionUNet. AttentionUNet's worst axis error is 1.99° and SwinUNETR's 2.43°. The floor's worst case (2.26°) exceeds AttentionUNet's, so the floor is not a per-case lower bound.

Table 2. HFValid 5-fold cross-validation with configuration C, median [IQR] over 101 patients. The floor is the landmark extractor applied to the true masks. Paired p: Wilcoxon, Holm-corrected. Checkpoints were selected on the held-out fold (see Limitations). On the 2 mm grid, HD95 takes few distinct values (e.g. 2, 2.83, 3.46 mm).

	AttentionUNet	SwinUNETR	p	Floor
Dice	0.939 [0.932, 0.945]	0.937 [0.930, 0.945]	0.47	–
HD95 (mm)	2.83 [2.83, 3.46]	2.83 [2.83, 3.46]	1.00	–
ASD (mm)	1.01 [0.93, 1.15]	1.02 [0.91, 1.12]	1.00	–
NSD @ 2 mm	0.922 [0.898, 0.943]	0.914 [0.891, 0.942]	0.23	–
Hip error (mm)	2.47 [1.75, 3.26]	2.69 [1.97, 3.90]	0.041	0.39 [0.25, 0.55]
Knee error (mm)	3.88 [2.59, 5.86]	4.05 [3.11, 5.58]	1.00	3.11 [2.12, 4.48]
Axis error ($^\circ$)	0.53 [0.32, 0.72]	0.55 [0.33, 0.85]	0.23	0.29 [0.14, 0.39]
$< 1^\circ$ / $< 2^\circ$ (%)	90 / 100	82 / 98	–	–
Worst axis ($^\circ$)	1.99	2.43	–	2.26
Head failures	0/101	0/101	–	–

The floor indicates how the error splits. On the true masks the extractor has a median axis error of 0.29°, against 0.53° and 0.55° on the reconstructions, so roughly half of the remaining error is already present with a perfect reconstruction (a descriptive comparison of medians). The median hip error rises from 0.39 mm on the true masks to 2.47–2.69 mm, whereas the knee error rises only from 3.11 mm to 3.88–4.05 mm; the relatively high knee floor reflects the difference between our knee rule and the expert definition. The coronal error was 0.26° [0.12, 0.47] for AttentionUNet and 0.30° [0.16, 0.50] for SwinUNETR (below 1° in 99% and 97% of cases), against a sagittal error of 0.36° [0.20, 0.56] and 0.36° [0.19, 0.67], so the sagittal component was slightly larger. The floor is almost entirely sagittal (coronal 0.06°, sagittal 0.27°), which suggests that the extractor's own error lies mainly in the antero-posterior position of the landmarks.

4.2 Configuration Ablation

Table 3 shows how the comparison changes with the configuration. With the v1 configurations, AttentionUNet had 12 head failures and SwinUNETR none. Giving AttentionUNet the SwinUNETR optimiser and learning rate (A) did not help (15 failures). Longer training with cosine decay and mild augmentation (B) brought its typical

accuracy level with the transformer (median axis 0.53° versus 0.59°) but left 9 failures, against none for SwinUNETR (McNemar $p = 0.004$). Under configuration B one would conclude that the transformer is more reliable.

Table 3. Configuration ablation on HFValid (101 patients): median Dice, hip error (mm), axis error, head failures and worst axis error. SwinUNETR was not trained with configuration A, which is the SwinUNETR v1 configuration applied to AttentionUNet.

	AttentionUNet					SwinUNETR				
Configuration	Dice	Hip	Axis	Fail	Worst	Dice	Hip	Axis	Fail	Worst
v1	0.909	4.70	0.75°	12	6.8°	0.923	4.01	0.71°	0	2.8°
A	0.914	4.25	0.68°	15	11.8°	–	–	–	–	–
B	0.930	2.96	0.53°	9	11.4°	0.934	3.09	0.59°	0	4.6°
C	0.939	2.47	0.53°	0	2.0°	0.937	2.69	0.55°	0	2.4°

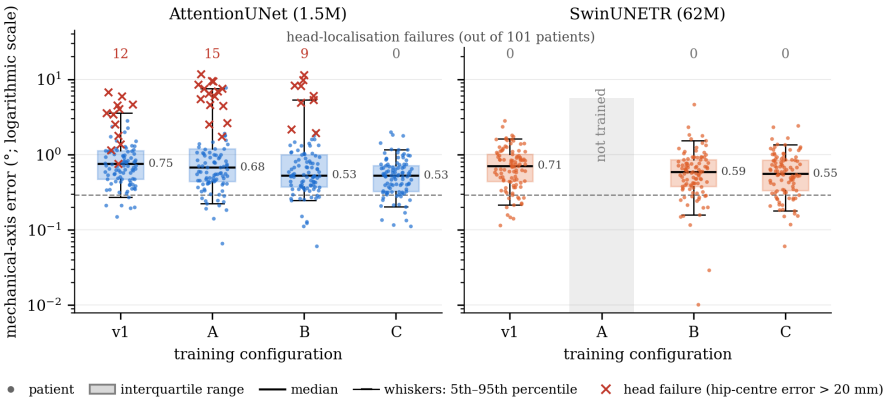


Fig. 3. Per-patient mechanical-axis error on HFValid ($n = 101$) for each training configuration. The y-axis is logarithmic because errors span more than two orders of magnitude. The dashed horizontal line indicates the landmark-extraction floor (0.29°), obtained by applying the landmark extractor to the reference segmentation masks; numbers above the boxes are head-localisation failures out of 101 patients. Configuration C leaves the median of AttentionUNet unchanged but removes its tail of head failures, after which the two architectures look alike.

Configuration C removed all nine failures (Fig. 3). Its median axis error is almost unchanged from B (0.53° for both), although the median of the per-patient paired differences (C minus B) is $-0.07°$ ($p_{\mathrm{Holm}} = 0.014$), and the tail shrank: the proportion of cases below 1° rose from 75% to 90% and the worst case fell from 11.4° to 2.0° (failures 9 to 0, McNemar $p = 0.004$). Overlap also improved (median paired Dice difference

$+0.010$, HD95 -0.8 mm, both $p < 10^{-9}$). SwinUNETR gained less from C (Dice $+0.003$, a smaller hip improvement than AttentionUNet, axis essentially unchanged). The reliability difference seen under configuration B was absent under configuration C, which suggests that the training configuration strongly influenced the comparison. Because configuration C changes pose augmentation, intensity range and normalisation together, we cannot attribute the gain to one of them. It was also developed after inspecting the configuration B failures on these folds, so the HFValid improvement may be optimistic; the VSD results (Sect. 4.4), which played no part in its design, are the independent check. These conclusions do not depend on the failure threshold. At 15, 20 and 25 mm, AttentionUNet B has 10, 9 and 9 failures, SwinUNETR B 1, 0 and 0, and both C models none. At every threshold, McNemar $p \leq 0.012$ for AttentionUNet B versus both SwinUNETR B and AttentionUNet C.

4.3 Robustness

Table 4 shows the models under test-time perturbations. AttentionUNet trained with configuration B was highly sensitive to intensity changes, with head failures in more than half of the cases under gamma changes, whereas SwinUNETR B was much more robust and had no head failures under any intensity change. With configuration C, neither model produced head failures under any perturbation except two AttentionUNet cases under strong blur. Both C models showed a similar increase in axis error when the lateral view was rotated, from about 0.55° to about 0.8° at 10°. For the C models, the gamma and noise perturbations lie within or near the training augmentation range, so they test in-distribution robustness; strong blur and lateral rotation lie outside it. To test the ground-truth-centred crop, we also shifted the test radiographs as if the femur lay 10, 20 or 30 mm off-centre in a random direction (training used shifts of at most 8 mm). Median Dice stayed at 0.94 and the median axis error at 0.51–0.61° for both C models; there were no head failures up to 20 mm, and 2/101 for AttentionUNet at 30 mm. These perturbation results are descriptive.

Table 4. Robustness on HFValid (test-time perturbations of the held-out radiographs): median axis error (°) / head failures out of 101. Blur and noise are applied to both views; gamma is applied to the intensities of both views before normalisation; for lateral rotation, the CT is rotated about the vertical axis and only the lateral DRR is re-projected, so the two views are no longer orthogonal.

Perturbation	AttentionUNet B	SwinUNETR B	AttentionUNet C	SwinUNETR C
None	0.53 / 9	0.59 / 0	0.53 / 0	0.55 / 0
Noise (s.d. 0.05)	0.81 / 16	0.63 / 0	0.66 / 0	0.68 / 0
Blur ($\sigma = 3$ px)	1.01 / 20	0.63 / 0	0.74 / 2	0.66 / 0
Gamma 0.7	2.99 / 54	0.64 / 0	0.58 / 0	0.61 / 0
Gamma 1.4	3.56 / 60	0.59 / 0	0.54 / 0	0.57 / 0
Lateral rot. 5°	0.69 / 9	0.66 / 0	0.62 / 0	0.65 / 0
Lateral rot. 10°	0.93 / 9	0.97 / 2	0.77 / 0	0.82 / 0

4.4 External Test on VSD

On the external cohort (Table 5), overlap drops across all 58 femurs (median Dice 0.86) and head failures appear (8 and 6), but on the 19 femurs with expert landmarks the median axis error stays below 0.7° for both models. No metric differs significantly between the two models (all $p_{\mathrm{Holm}} \geq 0.46$; failures McNemar $p = 0.63$). With 19 femurs the equivalence test is inconclusive (TOST $p = 0.057$): the errors are similar, but equivalence within $\pm 0.5°$ is not established. Configuration C improved both architectures over configuration B on VSD. On the 19-femur subset, AttentionUNet's median axis error fell from 5.2° to 0.7° ($p_{\mathrm{Holm}} < 0.001$) and SwinUNETR's from 1.1° to 0.6° ($p_{\mathrm{Holm}} = 0.009$); across all 58 femurs, AttentionUNet's failures fell from 26 to 8 (McNemar $p < 10^{-5}$) and SwinUNETR's from 8 to 6 ($p = 0.63$). At failure thresholds of 15, 20 and 25 mm, the C models had 9, 8 and 8 (AttentionUNet) and 7, 6 and 6 (SwinUNETR) failures. Of the C-model failures, 4 (AttentionUNet) and 2 (SwinUNETR) fall within the 19 femurs with landmarks. Coronal errors on the 19 femurs were 0.39° and 0.28°, and sagittal errors 0.45° and 0.40°. The extractor's knee error on the true VSD masks (4.51 mm) was already close to that on the reconstructions (4.57 and 5.89 mm), so much of the external knee error is present without any reconstruction error, consistent with the different landmark protocol (Limitation iv).

The head failures relate to femur pose (Fig. 4). Post hoc, each femur's tilt was measured on the ground-truth mask as the angle between the shaft (mask principal axis) and the box long axis. Tilt beyond the $\approx 18°$ corner spanned by the pose augmentation (the combination of $\pm 15°$ flexion and $\pm 10°$ abduction, $\sqrt{15^2 + 10^2} = 18°$, fixed by the training design rather than read off the plot) was a strong risk factor: of the 5 femurs tilted beyond 18°, 4 failed under AttentionUNet and 4 under SwinUNETR, against 4/53 and 2/53 below it, and failures had a median tilt of about 19° against 5° (Spearman $\rho = -0.29$ between tilt and Dice for each model, $n = 58$). The extreme tilts were in abduction (up to 29°), augmented only to $\pm 10°$: of the 6 femurs with abduction beyond $\pm 10°$, 5 (AttentionUNet) and 4 (SwinUNETR) failed, against 3/52 and 2/52 within the augmented range, so wider abduction augmentation is an obvious next step. Some failures still occurred at low tilt, so pose explains the gross failures but not all of them, and this measure does not capture axial rotation, which remains untested. Figure 6 shows the best, median and worst cases on both cohorts.

4.5 2D Landmark Baseline

The 2D baseline brackets the 3D models depending on how its checkpoint is selected (Table 6). With held-out selection on landmark error, which is more optimistic than the Dice-based selection of the 3D models, its median axis error on HFValid was 0.46° (88% below 1°, worst 2.46°, no head failures); with nested selection (no peeking) it was 0.67° (three head failures), and a single fixed-schedule run of 70 epochs with no selection gave 0.56°. The 3D models (0.53–0.55°) lie within this bracket, so the internal comparison is undetermined (see Limitations). At the optimistic (held-out) end, the paired differences favour the 2D baseline—against SwinUNETR C on all three errors (axis $-0.10°$, 95% CI -0.19 to -0.02, $p_{\mathrm{Holm}} = 0.0007$; hip -0.56 mm,

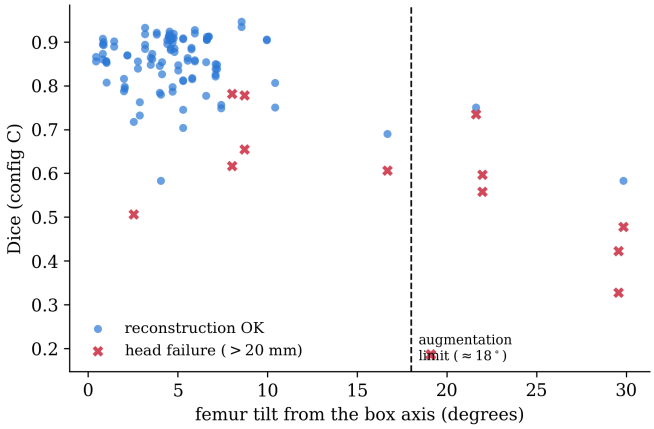


Fig. 4. External head failures against femur tilt, one panel per configuration-C model (one point per VSD femur in each panel); tilt is the angle of the shaft (ground-truth mask principal axis) from the box axis. Of the 5 femurs beyond the ≈18° corner spanned by the pose augmentation (dashed), 4 failed under each model; a few failures also occur at low tilt, and one femur above the corner does not fail under each model.

Table 5. External test on VSD with configuration C (5-fold ensemble), median [IQR]. Overlap and failures over 58 femurs; errors against expert raters over the 19 femurs with landmarks. Floor values from the extractor on the true masks.

	AttentionUNet	SwinUNETR	Floor
Dice ($n = 58$)	0.856 [0.779, 0.905]	0.854 [0.800, 0.895]	–
HD95 (mm, $n = 58$)	9.90 [8.28, 14.70]	10.00 [8.31, 12.60]	–
ASD (mm, $n = 58$)	2.19 [1.70, 3.29]	2.26 [1.79, 2.94]	–
NSD @ 2 mm ($n = 58$)	0.742 [0.648, 0.831]	0.726 [0.661, 0.820]	–
Head failures ($n = 58$)	8	6	–
Axis vs experts (°, $n = 19$)	0.69 [0.42, 2.24]	0.62 [0.46, 2.12]	0.38
Hip vs experts (mm, $n = 19$)	4.35 [2.42, 11.60]	4.57 [2.91, 8.72]	0.57
Knee vs experts (mm, $n = 19$)	4.57 [3.09, 20.11]	5.89 [4.54, 10.39]	4.5
Axis $< 2°$ ($n = 19$)	12/19	13/19	–

$p_{\mathrm{Holm}} = 0.0006$; knee -1.10 mm, $p_{\mathrm{Holm}} < 10^{-4}$), and on the knee against AttentionUNet C (-0.71 mm, $p_{\mathrm{Holm}} = 0.0002$; axis and hip not significant)—but this reflects its more favourable selection. Its coronal and sagittal errors (held-out) were 0.24° and 0.29°. On VSD none of these variants is selected (each is simply applied), and the three vary widely: the median axis error was 0.50° (held-out, 5/58 head failures), 0.68° (nested, 6/58) and 1.01° (fixed-schedule, 9/58), against 0.62–0.69° and 6–8/58 for the 3D models. This spread across training variants (0.50–1.01°) is wider than any 2D-versus-3D difference, so the external comparison is also undetermined; the 2D baseline was not shown to be better than 3D on either cohort. No variant differed significantly from either 3D model in axis error or failure count: the held-out and nested variants

gave all $p_{\mathrm{Holm}} \geq 0.5$ (failures McNemar $p \geq 0.25$), and the fixed-schedule variant, despite its higher median, gave axis $p_{\mathrm{Holm}} \geq 0.12$ and McNemar $p \geq 0.51$ (its knee error alone was higher than SwinUNETR's, $p_{\mathrm{Holm}} = 0.025$). The nested variant is pessimistic for two reasons: it does not peek at the outer fold, and it also trains on 12 fewer patients, so part of its gap is data rather than selection (the refit variant, Sect. B, trains on all patients and gives 0.62°). The AP and lateral views agreed on landmark height to a median of 1.4 mm at the hip and 1.4 mm at the knee.

Table 6. 2D landmark baseline against the 3D reconstructions (configuration C), median [IQR]. Three training variants bracket the comparison: *held-out* selects the epoch on the outer fold (on landmark error, so more optimistic than the Dice-based 3D selection), *nested* selects it on an inner split the outer fold never enters (no peeking), and *fixed-sched.* is a single run of 70 fixed epochs with no selection (the other variants use a 150-epoch schedule); its augmentation is unseeded, as for all runs here, so its values are one draw and vary run to run (Limitation iii). VSD axis error and $< 2°$ over the 19 femurs with expert landmarks; head failures over all 58.

	HFValid ($n = 101$)			VSD		
	Axis (°)	Hip (mm)	Knee (mm)	Axis (°)	$< 2°$	Failures
2D (held-out)	0.46 [0.27, 0.68]	2.33 [1.49, 3.26]	2.84 [2.00, 4.02]	0.50 [0.37, 1.04]	15/19	5/58
2D (fixed-sched.)	0.56 [0.33, 0.82]	2.88 [2.00, 4.04]	3.24 [2.39, 4.96]	1.01 [0.46, 2.61]	14/19	9/58
2D (nested)	0.67 [0.36, 1.05]	3.61 [2.46, 5.35]	3.33 [2.39, 5.52]	0.68 [0.54, 1.43]	15/19	6/58
AttentionUNet C	0.53 [0.32, 0.72]	2.47 [1.75, 3.26]	3.88 [2.59, 5.86]	0.69 [0.42, 2.24]	12/19	8/58
SwinUNETR C	0.55 [0.33, 0.85]	2.69 [1.97, 3.90]	4.05 [3.11, 5.58]	0.62 [0.46, 2.12]	13/19	6/58

4.6 Learned 3D Extractor

The learned extractor reduced the knee error on all three HFValid inputs (Table 7): from 3.11 to 2.39 mm on the true masks (median paired difference, learned minus rule, -0.54 mm, 95% CI -0.99 to -0.12), from 3.88 to 2.52 mm on AttentionUNet C reconstructions (-0.77 mm, -1.55 to -0.40) and from 4.05 to 2.87 mm on SwinUNETR C reconstructions (-0.99 mm, -1.50 to -0.66; all $p < 0.001$). Its hip error was slightly higher than that of the sphere fit ($+0.12$ to $+0.46$ mm, $p \leq 0.003$). The axis error did not change: 0.53 to 0.53° for AttentionUNet C (paired $-0.02°$, -0.04 to 0.02, $p = 0.73$), 0.55 to 0.57° for SwinUNETR C ($-0.02°$, -0.05 to 0.01, $p = 0.35$) and 0.29 to 0.28° on the true masks ($p = 0.76$). On VSD, the learned extractor was accurate on the true masks (knee 4.23 against 4.51 mm) but less robust on reconstructions: on badly reconstructed femurs it sometimes misplaced the knee entirely (upper-quartile knee error 217 mm for AttentionUNet C, against 20 mm with the rule; paired test on all 19 knee errors, $p = 0.03$). These p-values are from unadjusted Wilcoxon signed-rank tests.

4.7 Where the Axis Error Comes From

The first-order approximation tracks the measured axis error almost exactly (Spearman $\rho = 0.999$, maximum absolute difference 0.08°; median femur length 375 mm), so

Table 7. Rule-based (sphere fit and widest-level knee rule) against learned 3D landmark extraction on the true masks and on the configuration C reconstructions, median [IQR]; the learned extractor's epoch was selected on an inner validation split. VSD over the 19 femurs with expert landmarks.

		HFValid ($n = 101$)			VSD ($n = 19$)		
Mask	Extractor	Axis (°)	Hip (mm)	Knee (mm)	Axis (°)	Knee (mm)	$< 2°$
True	rule	0.29 [0.14, 0.39]	0.39 [0.25, 0.55]	3.11 [2.12, 4.48]	0.38 [0.28, 0.50]	4.51 [3.12, 7.15]	19/19
	learned	0.28 [0.16, 0.41]	0.93 [0.61, 1.26]	2.39 [1.71, 3.48]	0.45 [0.27, 0.60]	4.23 [2.85, 6.54]	18/19
AttentionUNet C	rule	0.53 [0.32, 0.72]	2.47 [1.75, 3.26]	3.88 [2.59, 5.86]	0.69 [0.41, 2.24]	4.57 [3.09, 20.1]	12/19
	learned	0.53 [0.31, 0.70]	2.60 [1.88, 3.27]	2.52 [1.66, 3.77]	0.74 [0.34, 33.9]	5.59 [2.84, 217]	13/19
SwinUNETR C	rule	0.55 [0.33, 0.85]	2.69 [1.97, 3.90]	4.05 [3.11, 5.58]	0.62 [0.46, 2.12]	5.89 [4.54, 10.4]	13/19
	learned	0.57 [0.34, 0.86]	2.98 [2.17, 4.11]	2.87 [1.89, 3.85]	0.68 [0.38, 2.97]	4.94 [3.13, 12.1]	12/19

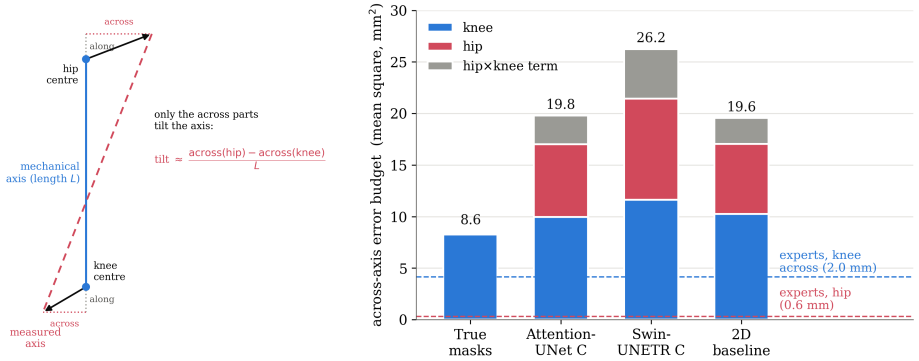


Fig. 5. (a) Each landmark error splits into a component along the mechanical axis and one across it; only the across components tilt the axis, by about $[\mathrm{across(hip)} - \mathrm{across(knee)}]/L$, so along-axis error (e.g. most of the knee-rule error) barely changes the angle. (b) Across-axis error budget (mean squares over 101 patients, $\mathrm{E}|\mathbf{h} - \mathbf{k}|^2 = \mathrm{E}|\mathbf{h}|^2 + \mathrm{E}|\mathbf{k}|^2 - 2\,\mathrm{E}[\mathbf{h} \cdot \mathbf{k}]$; the 2D baseline is the held-out run): on the true masks it is almost all knee, whereas the reconstructions and the 2D baseline add their extra error mainly at the hip, bringing the hip and knee contributions to roughly equal. Dashed lines: inter-expert disagreement.

the axis error is set by the across-axis part of the hip−knee error difference (Fig. 5a, Table 8).

On the true masks the across-axis budget is almost entirely the knee (96%); on the reconstructions and the 2D baseline it grows and moves toward the hip (Fig. 5b). An error equally likely in every direction would put two-thirds of its squared size across the axis; the reconstructed hip error was close to this (73–79% across), whereas the knee-rule error lay mainly along the femur (33–36% across), which is why correcting the knee with a learned extractor left the axis unchanged. A small negative hip−knee correlation adds a rigid-tilt component. This hip error is not specific to volumetric reconstruction: the 2D baseline, which reconstructs nothing, had a comparable hip error (2.33 mm for the held-out variant) and the same pattern, a large hip increment over the floor and a small knee one, so the hip and knee contributions end up roughly equal.

Nor is the hip error mainly a resolution effect: in a separate, earlier matched pair (a fixed-schedule 2D baseline rendered at 2 mm and at 1 mm), the hip error fell only from about 3.0 to 2.5 mm. The per-pixel error rose and the CT slice spacing (1–2 mm) was unchanged, so pixel size accounts for only part of the hip error, not the bulk.

At the hip the error was a scatter, not a bias: the mean signed error (0.3–0.6 mm) was far below the root-mean-square error (3.1–3.7 mm), and removing the per-fold mean did not reduce the hip or axis error. Nor was the head distorted: the fitted head radius showed no detectable difference from the true mask. The hip error is therefore a scatter in the position of the head, not a change in its size.

Observer variability on VSD puts these in context. Two experts differed by 0.57 mm at the hip and, across the axis, 2.04 mm at the knee, with axes differing by 0.30°; a single observation deviated from the consensus by 0.21°. The across-axis knee error of all methods (1.5–2.7 mm) was near the expert level, whereas every image-based hip error (2.3–3.6 mm across the 3D models and the 2D selection variants) was 4.1–6.3 times the expert hip disagreement. Against the single HFValid annotation, the floor (0.29°) already matched expert disagreement (0.30°), the reconstructions were 1.8 times it and the 2D baseline 1.5 to 2.2 times; in the tail, 18–30% of image-based cases exceeded the expert 95th percentile (0.81°) against 3% for the floor. Externally the 3D axis errors (0.62–0.69°) were about 3.0–3.3 times the single-observer deviation. These are a reference scale, not an exact bound (see Limitations).

Table 8. Error geometry on HFValid ($n = 101$), medians in mm (axis in degrees). Across and along refer to the expert mechanical axis; "across (hip−knee)" is the across-axis part of the difference between the hip and knee error vectors, which divided by the femur length gives the axis error to first order. The 2D baseline row uses the held-out run.

Mask	Extractor	Hip across	Hip along	Knee across	Knee along	Across (hip−knee)	Axis (°)
True	rule	0.25	0.21	1.82	2.08	1.82	0.29
	learned	0.62	0.43	1.45	1.16	1.84	0.28
AttentionUNet C	rule	2.07	1.05	2.04	2.49	3.50	0.53
	learned	2.29	1.14	1.62	1.27	3.52	0.53
SwinUNETR C	rule	2.28	1.13	2.12	3.16	3.83	0.55
	learned	2.36	1.10	1.84	1.62	3.72	0.57
2D baseline	–	1.90	0.84	1.84	1.63	3.05	0.46

4.8 Dice versus Clinical Error

Dice correlates only moderately with the axis error (Spearman $\rho = -0.33$ and -0.44 for the C models, -0.45 and -0.40 for the B models, all $p < 0.001$). Because the femoral shaft accounts for most of the segmented volume, an incorrectly reconstructed femoral head has only a limited effect on Dice but can move the hip centre by centimetres (Fig. 8). Under configuration B, AttentionUNet's median Dice was 0.930 against SwinUNETR's 0.934, a small though significant difference, while it had nine head failures and the transformer none. Per case, low Dice did flag most failures: the nine had

Dice 0.40–0.89, and only 5 of the other 92 cases scored at or below the highest failure Dice (0.89). The failures were thus hidden in the median comparison rather than invisible to Dice.

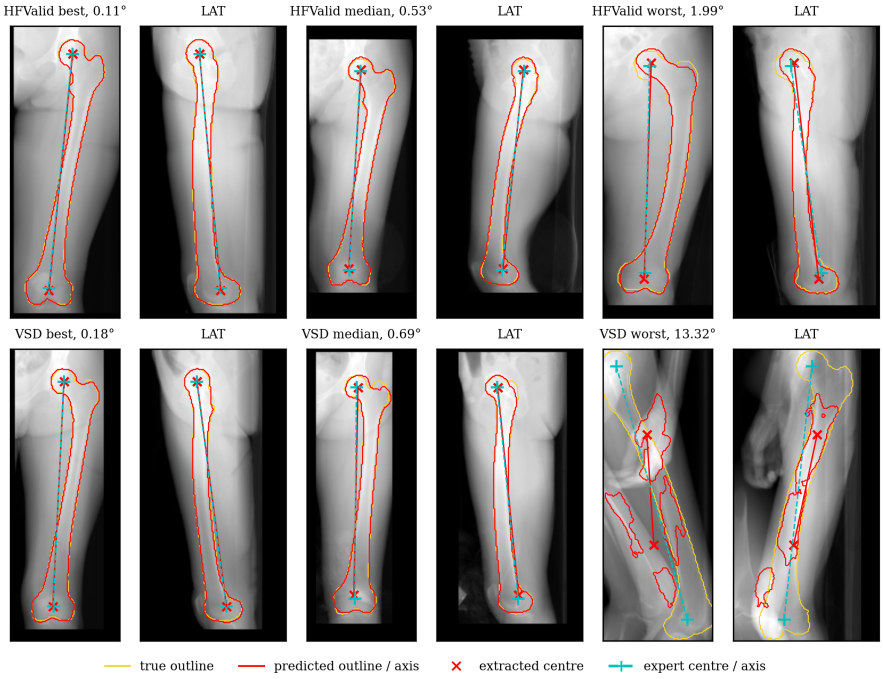


Fig. 6. Best, median and worst AttentionUNet C cases by axis error on HFValid (top) and VSD (bottom), chosen by rank rather than by appearance. Each case shows the AP and lateral DRR with the true (yellow) and predicted (red) outlines, the extracted (red $\times$) and expert (cyan $+$) hip and knee centres, and the predicted and expert axes. The worst VSD case (Dice 0.19) is a gross reconstruction failure in a femur that is strongly tilted in the image.

5 Discussion

What limits axis accuracy? No single landmark limits the axis: the across-axis error that sets the angle comes about equally from hip and knee (Sect. 4.7). What image-based methods *add* over the measurement floor falls on the femoral head, and the 2D baseline, which reconstructs nothing, shows a comparable or larger hip error (2.3–3.6 mm across variants against 2.5–2.7 mm for the 3D models), so this is a problem of locating the head in the radiographs rather than of volumetric reconstruction. The knee-rule error, by contrast, lies along the femur, so a learned extractor cut it without moving the axis, and the across-axis knee error is already near expert disagreement. More precise placement

of the femoral head is therefore the main target for improvement, while median overlap was a poor guide to any of this.

2D versus 3D. Whether direct 2D detection or 3D reconstruction is preferable for this measurement is not settled here. Internally the 2D baseline brackets the 3D models by checkpoint selection (0.46–0.67° against 0.53–0.55°); externally its axis error spans 0.50–1.01° across training variants, wider than any 2D-versus-3D difference. The comparison is therefore undetermined on both cohorts, and 2D was not shown to be better than 3D on either. For a point-defined parameter such as the mechanical axis it is a simpler alternative, though not shown here to be more accurate. Orthogonal parallel-beam views also make 2D triangulation exact, so any point-defined parameter could be triangulated the same way; the case for 3D reconstruction rests instead on parameters that depend on surface geometry (a fitted neck axis, condylar tangents, joint-line orientation) and on settings where triangulation degrades.

Training configuration and architecture. Under configuration C, the 1.5M-parameter AttentionUNet performed comparably to the 62M-parameter SwinUNETR in overlap and axis accuracy. The two also behaved similarly under the tested perturbations and on the external cohort, although the external clinical comparison rests on 19 femurs. Under the earlier configurations, SwinUNETR showed fewer head-localisation failures, an apparent reliability advantage that a study stopping at configuration B would have reported as an architecture effect; configuration C closed this gap. This agrees with reports that CNN–transformer gaps can shrink when training is matched [4, 21] and that well-configured CNN baselines remain competitive [17]. The gains came from the combined configuration C, which adds augmentation that mimics variation in limb pose and image appearance. SwinUNETR, however, had no head failures on the unperturbed cross-validation images under any configuration, and its VSD failures changed little from B to C (8 to 6, against 26 to 8 for AttentionUNet). It was therefore less sensitive to the training configuration, although with one seed per fold we cannot attribute this to the architecture alone.

Overlap metrics. Median overlap can conceal errors that matter for clinical planning. Under configuration B, the median Dice scores differed by only 0.004, yet AttentionUNet had nine head-localisation failures compared with none for SwinUNETR. In this experiment, failure counts, worst-case errors and clinical measurements therefore revealed a difference that median Dice alone did not capture, which supports reporting clinically relevant measurements alongside overlap metrics [22, 29].

Clinical meaning. Under idealised projection, the axis error of about half a degree is smaller than the changes that limb flexion and rotation cause in conventional radiographic measurements [6], although real radiographs would add such effects; in the coronal plane, which determines varus–valgus alignment, the median error was about 0.3°. The 3D axis error (0.53–0.55°) was, however, about 1.8 times the median disagreement between two experts (0.30°), and 19–30% of the reconstructions exceeded the VSD-derived 95th percentile of that disagreement against about 5% expected at observer level, so the method does not yet reach observer-level reliability; the extractor on

perfect masks (0.29°, 3%) already does. On the external cohort, moreover, the median accuracy hides a heavy tail: 7/19 and 6/19 femurs exceeded 2°, the upper-quartile hip error reached 11.6 mm, and median Dice fell to 0.86, so accuracy outside the training population is not yet sufficient for planning. All of these are geometric accuracies under synthetic conditions, not evidence of clinical acceptability.

Deployment. Inference needs no GPU: the AttentionUNet five-fold ensemble, exported to ONNX, runs on a Raspberry Pi 4 in about 110 s per case with 1.7 GB of memory (Appendix C). This only shows the compute is modest; a clinical deployment would still require automatic femur localisation and validation on real radiographs.

Limitations. (i) *Synthetic, idealised inputs.* The DRRs are parallel-beam, without the scatter, magnification or soft tissue of real radiographs (no public dataset pairs real whole-femur biplanar images with CT ground truth), and the crop is centred on the true femur mask, so a deployed system would still need to localise the femur. The results should be read as a controlled benchmark. (ii) *Model selection.* The 3D checkpoints were selected on the held-out fold by Dice, so the cross-validation figures are validation-selected and optimistic; VSD is the unbiased test. The 2D baseline brackets the 3D models by selection: choosing its epoch on landmark error (nearly the evaluated quantity) is more optimistic still and gives 0.46°; a single fixed-schedule run with no checkpoint selection gives 0.56°; and a nested variant, selected on an inner split the outer fold never enters, gives 0.67°. The 3D models (0.53–0.55°) lie within this bracket, so the internal comparison is undetermined. SwinUNETR reached its best epoch at 72–79 of 80 for a 0.003 Dice gain, so its own selection effect is small. Nested selection of the 3D models was beyond our compute budget. (iii) *Training variance and post-hoc configuration.* Each fold used one seed, so small differences between models may be training variance [31], and only two architectures were compared. Configuration C was tuned after inspecting configuration-B failures and changes several factors at once, and SwinUNETR ran fewer epochs, so the comparison may favour AttentionUNet. The 2D training variants show how large this variance can be: two nested-style variants (selecting the epoch on an inner split, or refitting on all training patients for the selected epoch count) differed by about 0.05° and two head failures (0.67°/three versus 0.62°/one), and externally the 2D variants, which differ in schedule length, selection and seed, spanned 0.50–1.01° – more than the gap between 2D and 3D. More directly, re-running the fixed-schedule protocol unchanged – the only same-protocol repeat here – moved the median axis error by about 0.03° internally (0.55–0.58° over three runs) and 0.12° externally (1.13° and 1.01°, with 10 and 9 head failures), so the single-seed 3D models may vary similarly. (iv) *Cohort and landmark protocol.* Training is a single cohort (older women, supine), and the knee definition and observer variability differ in protocol and population between HFValid and VSD, so the external knee error and the observer scale are approximate rather than exact bounds. The pose augmentation covered abduction only to $\pm 10°$, and most external femurs tilted beyond that range failed (Sect. 4.4).

Future work. Methodologically, the next steps are validation on real biplanar radiographs, including standing acquisitions from hospitals in low-resource settings, more

realistic radiographic simulation [12, 30], wider pose augmentation (especially in abduction, which the external failures exceeded), experiments with multiple random seeds and comparison against an nnU-Net baseline [16]. Clinically, the reconstruction should be evaluated against 2D landmarks on parameters that depend on surface geometry, such as femoral anteversion from a fitted neck axis and the mechanical lateral distal femoral angle (mLDFA) from the condylar joint line, which are relevant to arthroplasty and distal femoral osteotomy planning, and the femoral head should be placed more precisely, for example by refining it in the images. Extending the reconstruction to the tibia would be a subsequent step towards estimating the full hip–knee–ankle angle.

6 Conclusion

We asked what limits the femoral mechanical axis recovered from two synthetic radiographs, and separated three factors. The knee-rule error lay along the femur, so correcting it left the axis unchanged; the extra across-axis error of every image-based method, 2D or 3D, came mainly from scattering the femoral head, several times more than experts disagree. Direct 2D detection was not shown to be better than 3D reconstruction on either cohort (its axis error bracketed the 3D models both internally and externally, depending on the training variant), so the hip problem is not specific to volumetric reconstruction. Under the final training configuration a CNN 40 times smaller than the transformer matched it in cross-validation accuracy and reliability, so the transformer's earlier reliability edge reflected the configuration more than the architecture; and the median Dice did not reflect the head-localisation failures that drove the clinical error (per-case Dice still flagged them). The main improvable target is therefore more precise placement of the femoral head, and for 3D reconstruction, parameters that depend on surface geometry rather than on point landmarks.

Acknowledgements. Training used free GPU time on Kaggle.

Disclosure of interests. The author has no competing interests to declare.

Data and code availability. HFValid (CC BY 4.0) [1] and VSD (CC BY-NC-SA 4.0) [10] are public. Code, the fold split, configuration files and trained weights are available at https://github.com/ayyp1/biplanar-xray-to-3d-femur.

References

1. Aldieri, A., La Mattina, A.A., Szyszko, J.A., Baruffaldi, F., Viceconti, M.: HFValid collection: Hip-Fracture validation collection (2025). https://doi.org/10.6092/UNIBO/AMSACTA/8431, https://amsacta.unibo.it/id/eprint/8431
2. Arn Roth, T., Jokeit, M., Sutter, R., Vlachopoulos, L., Fucentese, S.F., Carrillo, F., Snedeker, J.G., Esfandiari, H., Fürnstahl, P.: Deep-learning based 3D reconstruction of lower limb bones from biplanar radiographs for preoperative osteotomy planning. International Journal of Computer Assisted Radiology and Surgery **19**(9), 1843–1853 (Apr 2024). https://doi.org/10.1007/s11548-024-03110-5, http://dx.doi.org/10.1007/s11548-024-03110-5

3. Babazadeh Khameneh, N., Cresson, T., Lavoie, F., de Guise, J., Vázquez, C.: Automatic CNN-based 3D/2D non-rigid registration platform for fast 3D femur reconstruction and clinical 3D measurements from Bi-planar radiographs. Computers in Biology and Medicine **196**, 110676 (2025). https://doi.org/10.1016/j.compbiomed.2025.110676, http://dx.doi.org/10.1016/j.compbiomed.2025.110676
4. Bai, Y., Mei, J., Yuille, A., Xie, C.: Are Transformers More Robust Than CNNs? In: Advances in Neural Information Processing Systems 34 (NeurIPS) (2021)
5. Van den Borre, I., Audenaert, E., Vermue, H., Huysentruyt, R., Van Vynckt, L., Vanhauwe, R., Pas, V., Pizurica, A., Van Oevelen, A.: Combining deep learning with statistical shape modelling enables automated lower limb measurements with observer-level reliability using weight-bearing computed tomography. Journal of Experimental Orthopaedics **13**(1), e70669 (Jan 2026). https://doi.org/10.1002/jeo2.70669, http://dx.doi.org/10.1002/jeo2.70669
6. Brunner, J., Jörgens, M., Weigert, M., Kümpel, H., Degen, N., Fuermetz, J.: Significant changes in lower limb alignment due to flexion and rotation—a systematic 3D simulation of radiographic measurements. Knee Surgery, Sports Traumatology, Arthroscopy **31**(4), 1483–1490 (Jan 2023). https://doi.org/10.1007/s00167-022-07302-x, http://dx.doi.org/10.1007/s00167-022-07302-x
7. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A.: MONAI: An open-source framework for deep learning in healthcare (2022). https://doi.org/10.48550/ARXIV.2211.02701, https://arxiv.org/abs/2211.02701
8. Chaibi, Y., Cresson, T., Aubert, B., Hausselle, J., Neyret, P., Hauger, O., de Guise, J.A., Skalli, W.: Fast 3D reconstruction of the lower limb using a parametric model and statistical inferences and clinical measurements calculation from biplanar X-rays. Computer Methods in Biomechanics and Biomedical Engineering **15**(5), 457–466 (2012). https://doi.org/10.1080/10255842.2010.540758
9. Chen, J., Lin, Y., Sun, H., Li, X.: Spatial-Division Augmented Occupancy Field for Bone Shape Reconstruction from Biplanar X-Rays. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. LNCS, vol. 15007, p. 668–678. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72104-5_64, http://dx.doi.org/10.1007/978-3-031-72104-5_64
10. Fischer, M.C.M.: VSDFullBodyBoneReconstruction: Segmentations and surface models of bones of the entire lower body created from cadaver CT scans from the VSDFullBody collection (2023). https://doi.org/10.5281/ZENODO.8302448, https://zenodo.org/doi/10.5281/zenodo.8302448
11. Gajny, L., Girinon, F., Bayoud, W., Lahkar, B., Bonnet-Lebrun, A., Rouch, P., Lazennec, J.Y., Skalli, W.: Fast quasi-automated 3D reconstruction of lower limbs from low dose biplanar radiographs using statistical shape models and contour matching. Medical Engineering & Physics **101**(1), 103769 (Feb 2022). https://doi.org/10.1016/j.medengphy.2022.103769, http://dx.doi.org/10.1016/j.medengphy.2022.103769
12. Gao, C., Killeen, B.D., Hu, Y., Grupp, R.B., Taylor, R.H., Armand, M., Unberath, M.: Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. Nature Machine Intelligence **5**(3), 294–308 (Mar 2023). https://

doi.org/10.1038/s42256-023-00629-1, http://dx.doi.org/10.1038/s42256-023-00629-1
13. Gurbanli, Y., Mohammed, J., Travascio, F.: Two-dimensional to three-dimensional knee reconstruction from radiographs and fluoroscopy: A systematic review of methods and accuracy (1995–2025). PLOS One **21**(8), e0356987 (Aug 2026). https://doi.org/10.1371/journal.pone.0356987, http://dx.doi.org/10.1371/journal.pone.0356987
14. Guven, G., Ugurdag, F., Ates, H.F.: X2BR: High-fidelity 3D bone reconstruction from a planar X-ray image with hybrid neural implicit methods. Medical & Biological Engineering & Computing **64**(5), 1931–1944 (Mar 2026). https://doi.org/10.1007/s11517-026-03549-5, http://dx.doi.org/10.1007/s11517-026-03549-5
15. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. LNCS, vol. 12962, p. 272–284. Springer International Publishing (2022). https://doi.org/10.1007/978-3-031-08999-2_22, http://dx.doi.org/10.1007/978-3-031-08999-2_22
16. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
17. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jäger, P.F.: nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. LNCS, vol. 15009, p. 488–498. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72114-4_47, http://dx.doi.org/10.1007/978-3-031-72114-4_47
18. Kasten, Y., Doktofsky, D., Kovler, I.: End-To-End Convolutional Neural Network for 3D Reconstruction of Knee Bones from Bi-planar X-Ray Images. In: Machine Learning for Medical Image Reconstruction. p. 123–133. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-61598-7_12, http://dx.doi.org/10.1007/978-3-030-61598-7_12
19. Lin, Y., Sun, H., Li, Y., Aslam, R., Tse, L.F., Cheng, T., Chui, C.S., Yau, W.F., Le Meur, V.R., Amangeldy, M., Cho, K., Ye, Y., Zou, J., Zhao, W., Li, X.: Real-time reconstruction of 3D bone models via very-low-dose protocols. npj Digital Medicine **9**(1), 353 (Mar 2026). https://doi.org/10.1038/s41746-026-02389-9, http://dx.doi.org/10.1038/s41746-026-02389-9
20. Liu, K., Ying, Z., Jin, J., Li, D., Huang, P., Wu, W., Chen, Z., Qi, J., Lu, Y., Deng, L., Chen, B.: Swin-X2S: Reconstructing 3D Shape from 2D Biplanar X-ray with Swin Transformers (2025). https://doi.org/10.48550/ARXIV.2501.05961, https://arxiv.org/abs/2501.05961
21. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11966–11976 (2022). https://doi.org/10.1109/CVPR52688.2022.01167
22. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M.A., Wiesenfarth, M., Kavur, A.E., Sudre, C.H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M.B., Cardoso, M.J., Cheplygina, V., Cimini, B.A., Collins, G.S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Haase, R., Hashimoto, D.A., Hoffman, M.M., Huisman, M., Jannin, P., Kahn,

C.E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B.A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A.L., Mattson, P., Meijering, E., Menze, B., Moons, K.G.M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C.I., Shetty, S., van Smeden, M., Summers, R.M., Taha, A.A., Tiulpin, A., Tsaftaris, S.A., Van Calster, B., Varoquaux, G., Jäger, P.F.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21**(2), 195–212 (Feb 2024). https://doi.org/10.1038/s41592-023-02151-z, http://dx.doi.org/10.1038/s41592-023-02151-z

23. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention U-Net: Learning Where to Look for the Pancreas. In: Medical Imaging with Deep Learning (MIDL) (2018)
24. Paley, D., Herzenberg, J.E., Tetsworth, K., McKie, J., Bhave, A.: Deformity Planning for Frontal and Sagittal Plane Corrective Osteotomies. Orthopedic Clinics of North America **25**(3), 425–465 (1994). https://doi.org/10.1016/s0030-5898(20)31927-1, http://dx.doi.org/10.1016/S0030-5898(20)31927-1
25. Pan, K., Zhang, Y., Zhao, L., Huang, S.: Self-supervised 3D Reconstruction of Tibia and Fibula from Biplanar X-rays. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). p. 21692–21699. IEEE (Oct 2025). https://doi.org/10.1109/iros60139.2025.11247754, http://dx.doi.org/10.1109/iros60139.2025.11247754
26. Pan, Z., Zhou, H., Ren, Q., Hou, X., Dai, J., Liu, X., Che, Y., Zhao, X., Xie, Y., Li, Z., Liang, D., Liang, X.: X2Shape: CT-free 3D multi-organ reconstruction with biplanar X-rays. Medical Image Analysis **112**, 104085 (2026). https://doi.org/10.1016/j.media.2026.104085, http://dx.doi.org/10.1016/j.media.2026.104085
27. Pan, Z., Zhu, J., Zhou, H., Yang, G., Dai, J., Xie, Y., Liang, X.: X2Bone: Reconstructing 3D Bone Structures from 2D Biplanar X-Rays via Cross-RWKV. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2026. p. 399–408. Springer Nature Switzerland (2026). https://doi.org/10.1007/978-3-032-38186-6_38, http://dx.doi.org/10.1007/978-3-032-38186-6_38
28. Picard, F.D., Poudel, P., Shehata, N., Lavoie, F., Lombaert, H.: A Joint 2D-3D Statistical Shape Model for Orthopedic Reconstruction (2026). https://doi.org/10.48550/ARXIV.2609.09010, https://arxiv.org/abs/2609.09010
29. Shakya, M., Khanal, B.: Benchmarking Encoder-Decoder Architectures for Biplanar X-ray to 3D Bone Shape Reconstruction. In: Advances in Neural Information Processing Systems 36. p. 20469–20481. NeurIPS 2023, Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2023). https://doi.org/10.52202/075280-0899, http://dx.doi.org/10.52202/075280-0899
30. Unberath, M., Zaech, J.N., Lee, S.C., Bier, B., Fotouhi, J., Armand, M., Navab, N.: DeepDRR – A Catalyst for Machine Learning in Fluoroscopy-Guided Procedures. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. LNCS, vol. 11073, p. 98–106. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-00937-3_12, http://dx.doi.org/10.1007/978-3-030-00937-3_12
31. Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digital Medicine **5**, 48 (2022). https://doi.org/10.1038/s41746-022-00592-y
32. Ying, X., Guo, H., Ma, K., Wu, J., Weng, Z., Zheng, Y.: X2CT-GAN: Reconstructing CT From Biplanar X-Rays With Generative Adversarial Networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 10611–10620. IEEE (2019).

https://doi.org/10.1109/cvpr.2019.01087, http://dx.doi.org/10.1109/cvpr.2019.01087

A Hyperparameters

Table 9. Final (configuration C) settings. Configuration files: `configs/attnC.yaml`, `configs/swinC.yaml`.

	AttentionUNet	SwinUNETR
Parameters	1.5M	62M
Channels / feature size	16–128, 3 downsamplings	48, checkpointing
Input / output	$2 \times 256 \times 128 \times 96$ / $1 \times 256 \times 128 \times 96$	
Loss	Dice + cross-entropy (sigmoid)	
Optimiser	AdamW, lr 10^{-4}, weight decay 10^{-5}, cosine decay	
Epochs (max) / patience	150 / 20	80 / 20
Batch size, precision	1, mixed precision	
Pose augmentation	$p = 0.8$; $\pm 15^\circ$ / $\pm 10^\circ$ / $\pm 15^\circ$; shift ± 8 mm	
Intensity augmentation	gamma 0.6–1.6, contrast 0.7–1.3, brightness ± 0.1	
Blur / noise	$p = 0.3$, σ 0.3–1.5 px / $p = 0.5$, s.d. ≤ 0.04	
Normalisation	per-image z-score	
Training time per fold (T4)	$\approx$1.9 h	$\approx$4.3 h

B Landmark Extraction Details

Rule-based 3D extractor. The sphere is fitted with the linear form $\|\mathbf{p}\|^2 = 2\mathbf{c} \cdot \mathbf{p} + d$, with radius $r = \sqrt{d + \|\mathbf{c}\|^2}$, which turns the fit into a linear least-squares problem. RANSAC keeps the sphere with the most surface points within the tolerance; the refit with progressively tighter tolerances removes points on the femoral neck. On the true HFValid masks the extractor gives a hip error of 0.39 mm, a knee error of 3.11 mm (8.00 mm without the offset) and an axis error of 0.29° (median over 101 patients, offset learned on the training folds). Figure 7 illustrates the sphere selection and the distribution of the floor.

Learned 3D extractor. The network is a MONAI 3D U-Net (channels 8, 16, 32, 64 and 128, strides 2, one residual unit) with the full $256 \times 128 \times 96$ binary mask as input and one heatmap per landmark as output. Targets are separable 3D Gaussians ($\sigma = 2$ voxels) centred on the expert landmarks. The loss is the cross-entropy between the softmax over all voxels of each heatmap and the normalised target. A landmark is decoded as the probability-weighted centroid in a $7 \times 7 \times 7$ window around the heatmap maximum. Training masks are augmented in this order: with probability 0.8,

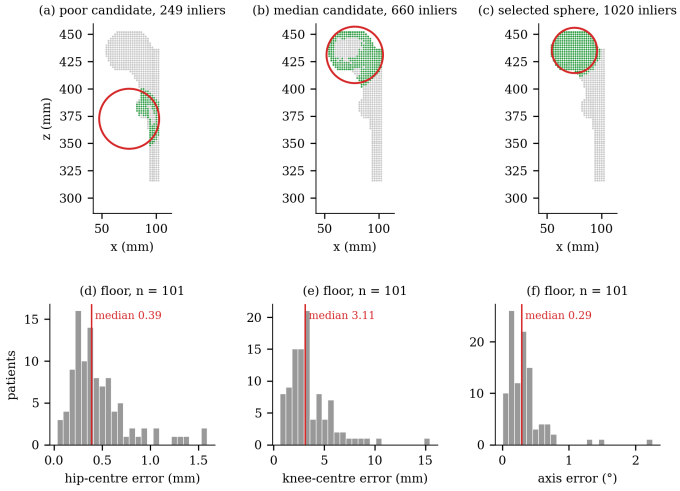


Fig. 7. Landmark extractor. (a–c) Three of the 500 RANSAC candidates for one HFValid femur, projected on the coronal plane: the proximal–medial surface points (grey), the inliers within 4 mm of each candidate sphere (green) and the sphere (red). The candidate with the most inliers (c) is kept and then refined. (d–f) Hip-centre, knee-centre and axis errors of the extractor on all 101 true masks (the measurement floor), with medians in red.

the mask and its landmarks are rotated (flexion $\pm 15°$, abduction $\pm 10°$, axial rotation $\pm 15°$) and shifted (up to ± 4 voxels), and the pose is kept only if both landmarks stay inside the box; with probability 0.5, the mask is smoothed (Gaussian, σ 0.5–1.5 voxels) and re-thresholded at 0.35–0.65; with probability 0.25 each, it is eroded or dilated by one voxel; and with probability 0.4, one to four spheres of radius 2–5 voxels, centred on random surface voxels, are added to or removed from the mask. Training uses batch size 2, mixed precision, AdamW (lr 10^{-3}, weight decay 10^{-5}) with cosine decay over at most 80 epochs, and early stopping after 15 epochs without improvement in the mean hip and knee error on the true masks of the inner validation split (12 of the training patients, drawn with seed $42+$fold and excluded from training). A fold took 12–26 minutes on an NVIDIA T4.

2D landmark baseline. Each view has its own MONAI 2D U-Net (channels 16, 32, 64, 128 and 256, strides 2, two residual units), with one DRR as input and one heatmap per landmark as output; the two networks of a fold are trained jointly with the summed loss of both views. Targets ($\sigma = 2$ pixels), loss and decoding (7×7 window) follow the 3D extractor. The 3D landmark is $(x_{\mathrm{AP}}, y_{\mathrm{LAT}}, (z_{\mathrm{AP}} + z_{\mathrm{LAT}})/2)$, where the AP view gives x and z and the lateral view y and z; the two height estimates differed by a median of 1.4 mm (hip) and 1.4 mm (knee). Training samples receive the pose augmentation of configuration C (with re-projection of the DRRs and landmarks), its intensity augmentation and its per-image normalisation. Training uses batch size 4 and AdamW (lr 10^{-3}, weight decay 10^{-5}) with cosine decay over a 150-epoch schedule. The primary checkpoint is selected per fold on the held-out outer fold, as for the 3D

models. The selection bracket adds three variants that do not use the outer fold: a *nested* variant that selects the epoch by early stopping (patience 20) on an inner split of 12 patients drawn from the training folds and keeps that model; a *refit* variant that takes the same epoch count and retrains on all training patients of the fold; and a *fixed-schedule* run of 70 epochs with no selection.

C CPU-only Deployment

Table 10. Inference of AttentionUNet C on a Raspberry Pi 4 (4 GB), 64-bit Raspberry Pi OS, 4 CPU threads, float32, no cooling.

PyTorch, one model pass	334 s
ONNX Runtime, one model pass	21–23 s
ONNX Runtime, 5-fold ensemble	109 s
Landmark extraction	0.3–0.5 s
Peak memory	1.7 GB
Maximum temperature	74.5 °C, no throttling

Times are per case on the stored DRR images and exclude DRR generation; the ensemble time covers all five fold models and the landmark extraction is listed separately. The ONNX model reproduced PyTorch on a held-out case (maximum difference in voxel probability 5×10^{-4}; identical Dice of 0.936 and identical hip, knee and axis errors), and gave the same result for the stored radiograph images and for radiographs projected directly from CT. This check covers one case.

D Additional Figures

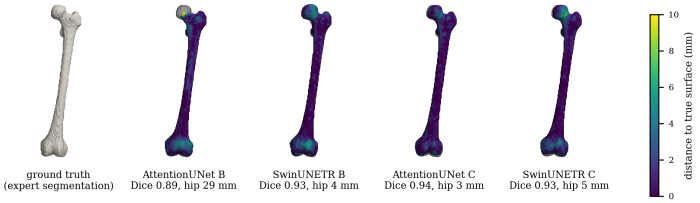


Fig. 8. 3D reconstructions of the same HFValid case, the configuration B head failure of AttentionUNet with the highest Dice, by both architectures under configurations B and C (front view). Each prediction is coloured by its distance to the true surface, with the true femur shown faintly in grey behind it. AttentionUNet B misses the femoral head (Dice 0.89, hip error 29 mm), whereas SwinUNETR B and both configuration C models reconstruct it (hip error 3–5 mm).

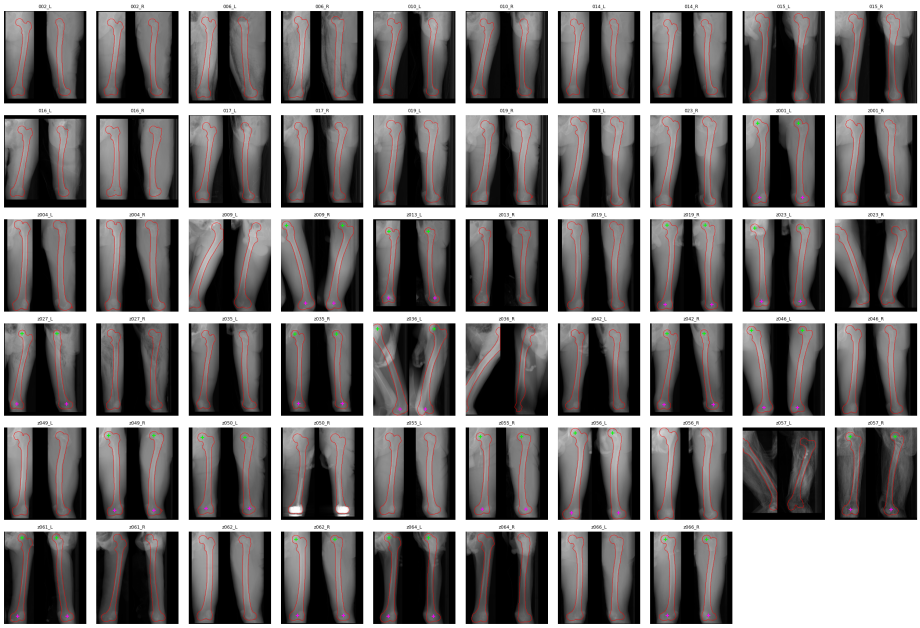


Fig. 9. All 58 VSD femurs prepared in the same way (AP and lateral DRR side by side, true outline in red). Expert hip (green) and knee (magenta) centres are shown for the 19 femurs with landmarks. Several femurs are strongly flexed or tilted, unlike the HFValid scans.



Fig. 10. All 101 HFValid femurs after preprocessing: AP and lateral DRRs with the true femur outline (red) and the expert hip (green) and knee (magenta) centres. Right femurs are mirrored to look like left ones.