IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received 8 May 2025, accepted 14 May 2025, date of publication 21 May 2025, date of current version 2 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3572342

RESEARCH ARTICLE

Optimization and Benchmarking of Image Segmentation for Improved Landmark Detection in Lower Limb X-Rays and Accurate Coronal Plane Alignment of the Knee Classification

SEBASTIAN AMADOR SANCHEZ[1,2], ASHKAN ZARGHAMI[1], PHILIPPE VAN OVERSCHELDE[3], AND JEF VANDEMEULEBROUCKE[1,2,4]

[1]Department of Electronics and Informatics, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[2]Imec, 3001 Leuven, Belgium
[3]moveUP, 9000 Ghent, Belgium
[4]Department of Radiology, Universitair Ziekenhuis Brussel, 1090 Brussels, Belgium

Corresponding author: Sebastian Amador Sanchez (sebastian.amador.sanchez@vub.be)

This work was supported by the Innoviris Grant of the Brussels-Capital Region, as part of the project AugmeNTed IntelligenCe In orthopaedics TrEatments (ANTICIPATE) on Augmented Intelligence in Orthopaedics Treatments, under Grant BHF/2020-RDIR-6a.

ABSTRACT Recent studies have explored image segmentation for landmark detection in computer vision and medical imaging of the lower limb, showing promising results. However, the proposed methodologies vary significantly, and a comparison with existing methods is lacking. In the present study, we investigated image segmentation for landmark detection on full lower-limb X-rays in detail and benchmark it against conventional landmark detection approaches. We detected eight landmarks in full lower limb X-rays and investigated methodological aspects to optimize image segmentation performance: network architecture (U-Net vs. Swin-UNETR), mask size centered at the landmark position to segment, and coordinate computation technique from the segmentation map. We contrasted image segmentation against optimized heatmap, coordinate, and segmentation-guided coordinate regression methods. The evaluation assessed the landmark detection error and phenotype classification accuracy based on lower limb alignment. The optimal segmentation approach employed a U-Net to segment circular masks (radius = 15 pixels), using probability thresholding before the centroid computation. Regarding landmark detection accuracy, image segmentation (median Euclidean distance (interquartile range) = 1.16 mm (1.50 mm)) was more accurate than heatmap (1.19 mm (1.61 mm)), coordinate (3.11 mm (2.87 mm)), and segmentation-guided coordinate regression (1.47 mm (1.67 mm)). Image segmentation outperformed heatmap, coordinate, and segmentation-guided coordinate regression in phenotype classification accuracy, achieving an average F_1-score of 0.79, versus 0.72, 0.47, and 0.77, respectively. Our study led to an optimized approach for landmark detection using image segmentation, outperforming alternative detection approaches tuned and tested on the same data, highlighting image segmentation's potential for broader medical imaging research applications.

INDEX TERMS CPAK classification, deep learning, image segmentation, landmark detection, lower limb X-rays.

I. INTRODUCTION

The process of detecting landmarks, i.e., salient points, for the localization, quantification, and visualization of objects of interest in an image is termed landmarking [1]. In medical imaging, landmarks identify and characterize anatomical structures from which clinicians perform further analyses, including diagnosis, treatment planning, and monitoring of treatment and disease progression [1], [2]. For example, when measuring lower limb malalignment, the head of the

The associate editor coordinating the review of this manuscript and approving it for publication was Gina Tourassi.

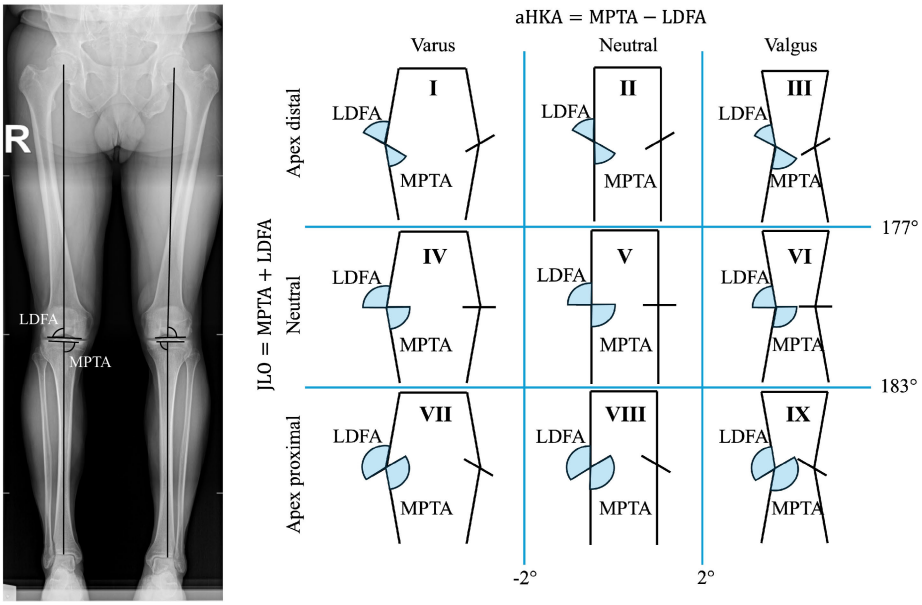

FIGURE 1. Measurement of the LDFA and MPTA in a full-leg X-ray and the nine categories distinguished by the CPAK in terms of the measured aHKA and JLO.

femur, knee notch, and ankle malleolus are typical landmark detection sites [3].

Owing to the advances in deep learning and the surge of convolutional neural networks (CNNs), approaches developed on these have replaced conventional detection methods based on either multi-scale sliding window and feature extraction or probabilistic models, such as the active appearance model [4], [5]. Currently, deep-learning-based solutions for landmark detection mainly follow two paradigms: coordinate regression and heatmap regression [5]. In the first approach, the network learns the mapping from image to pixel coordinates in an end-to-end fashion [6]. In heatmap regression, the network learns to locate the landmarks by reconstructing heatmaps, which are probability maps that follow a Gaussian distribution centered on the targeted image position [7].

Recently, researchers have explored image segmentation as an alternative to previous landmark detection schemes [3], [7], [8], [9]. Semantic segmentation is a computer vision task that aims to classify every pixel of an image, in its simplest form, as background or foreground. In landmark detection, the foreground corresponds to the pixel(s) belonging to a small region centered at the targeted landmark. Recent literature confirms the potential of image segmentation for landmark detection in traditional computer vision [7], [8] and lower limb imaging (see Section I-B). However, the available studies employ diverse methodologies, hindering the adoption of image segmentation as a landmark detection technique. Moreover, a direct comparison of image segmentation to alternative detection approaches is currently missing, limiting the insights into its relative strengths and appropriate use cases.

For this reason, our work aims to explore and benchmark image segmentation for landmark detection using full lower-limb X-rays. Our contributions are two-fold. First, we present a thorough optimization of the segmentation pipeline, evaluating network architectures, segmentation region sizes, and coordinate extraction methods. Second, we provide a comparative analysis between segmentation-based landmark detection and alternative approaches from the literature to assess their performance and limitations. This evaluation is performed at the level of landmark localization accuracy and within a clinically relevant context: alignment assessment for arthritic lower limb phenotype classification, a task of critical importance for clinical decision-making.

A. CPAK CLASSIFICATION SYSTEM

Lower limb malalignment is the loss of collinearity in the frontal plane of the hip, knee, and ankle [10]. In knee osteoarthritis, a relationship between lower limb malalignment and its progression has been observed [11]. To provide a classification system that categorizes the arthritic knee in terms of the alignment angles and assists in personalizing pre-operative planning, clinicians have proposed the Coronal Plane Alignment of the Knee (CPAK) classification system [12]. CPAK distinguishes nine classes that depend on the measured level of alignment quantified via the arithmetic hip-knee-ankle angle (aHKA) and joint line obliquity (JLO). The JLO and aHKA are expressed as follows:

$$\text{aHKA} = \text{MPTA} - \text{LDFA}, \tag{1}$$

$$\text{JLO} = \text{MPTA} + \text{LDFA}. \tag{2}$$

The MPTA corresponds to the medial proximal tibial angle, and LDFA to the lateral distal femoral angle. CPAK

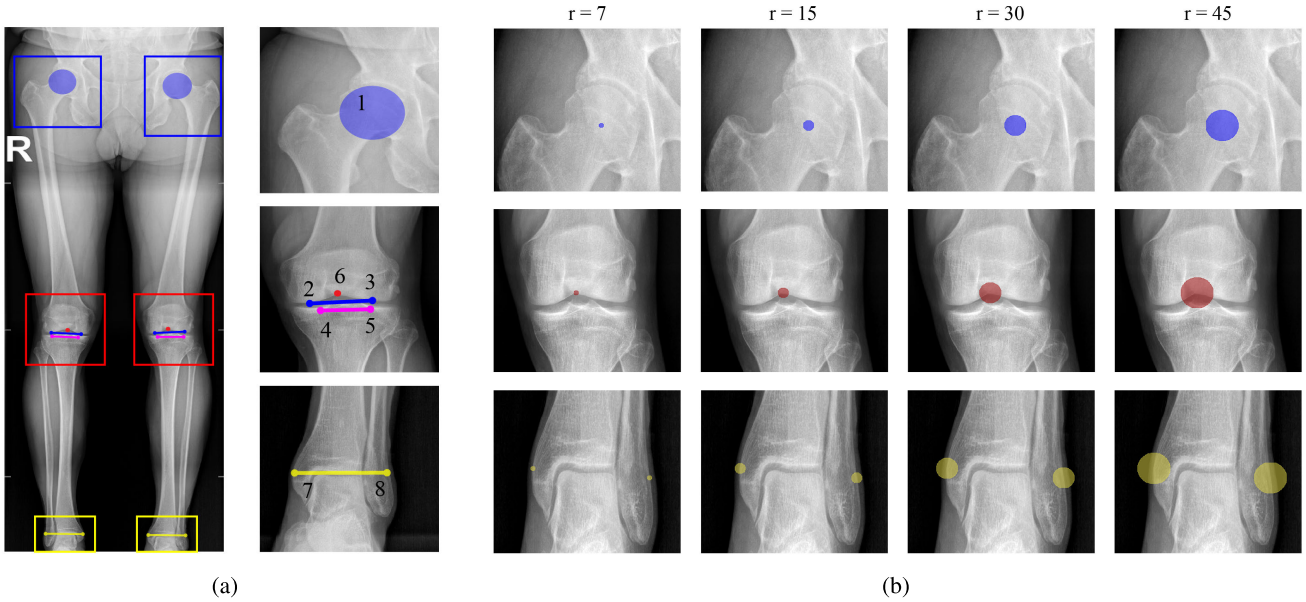

FIGURE 2. **(a) Annotations performed by the radiologist. (b) Circle masks of different sizes. For practical reasons, the single-pixel masks are not displayed.**

defines neutral alignment when $-2° \leq$ aHKA $\leq 2°$, varus deformity when aHKA $< -2°$, and valgus malalignment when aHKA $> 2°$. In parallel, a neutral line obliquity occurs for $177° \leq$ JLO $\leq 183°$, apex distal occurs if JLO $< 177°$, and apex proximal corresponds to JLO $> 183°$. Fig. 1 outlines the required angles for the computation of CPAK and shows the nine CPAK classes.

From Fig. 1, it is clear that the CPAK classification depends on accurately measuring the MPTA and LDFA angles, as even minor measurement errors can lead to significant misclassification and result in an incorrect surgical approach. For instance, misclassifying a subject from class II to V results in a recommendation for mechanically aligned surgery instead of anatomically aligned surgery. Manual measurement of alignment has been shown to be prone to errors [13], time-consuming [14], and greatly influenced by clinician expertise [15]. Moreover, given the expected rise in the number of surgical interventions [16], there is a clear need for novel and automated solutions that aid clinicians in assessing lower limb X-rays.

B. LANDMARK DETECTION ON LOWER LIMB X-RAY IMAGING

Researchers have proposed multiple approaches for landmark identification, aiming to automate the measurement of leg length [17], limb malalignment [3], [9], [18], [19], [20], [21], [22], [23], [24], [25], and CPAK classification [26], [27]. However, few report their findings at the landmark detection level, making the comparison and identification of the best approach challenging.

Tsai [17] employed heatmap regression to detect six lower limb landmarks, achieving an average localization error of 2.6 mm. Kim et al. [23] used heatmap regression to localize 19 landmarks and measure malalignment. However, no landmark detection error values were reported. Meanwhile, Tack et al. [20] utilized a cascade of coordinate regression networks to detect three landmarks with an L_2 error of up to 1.94 mm. Conversely, Bonin et al. [22] used multiple coordinate regression networks to identify 13 landmarks (with an average error of 24.8 px) in frontal knee X-rays.

Meng et al. [9] detected 20 landmarks via image segmentation using circular masks with an area of 5 mm^2, achieving an overall successful detection rate at 3 mm over 90%. Gai et al. [21] proposed a multi-task approach based on image segmentation to segment lower limb bones and detect nine landmarks, yielding a total RMSE of 2.14 mm. Sanchez et al. [3] investigated the identification of eight anatomical landmarks through the segmentation of circular masks, reaching a mean Euclidean distance error of 1.96 mm. Mika et al. [24] combined a dilation-erosion technique with a U-Net to detect five landmarks via image segmentation and achieved an overall RMSE of 2.38 px.

Simon et al. [25] introduced LAMA, a deep learning software for measuring malalignment (accuracy = 89.2%) via landmark segmentation. However, landmark detection errors were not mentioned. LAMA was subsequently used to analyze X-rays and explore the relationship between CPAK and gender [28]. Jo et al. [27] utilized image segmentation to detect 15 anatomical landmarks, achieving normalized distance errors of 0.88 to 2.92. Their work enabled the automatic measurement and analysis of CPAK, linking it to osteoarthritis severity [29]. Steele et al. [26] segmented regions from which landmarks were derived. Next, CPAK was computed from the MPTA and LDFA, which were

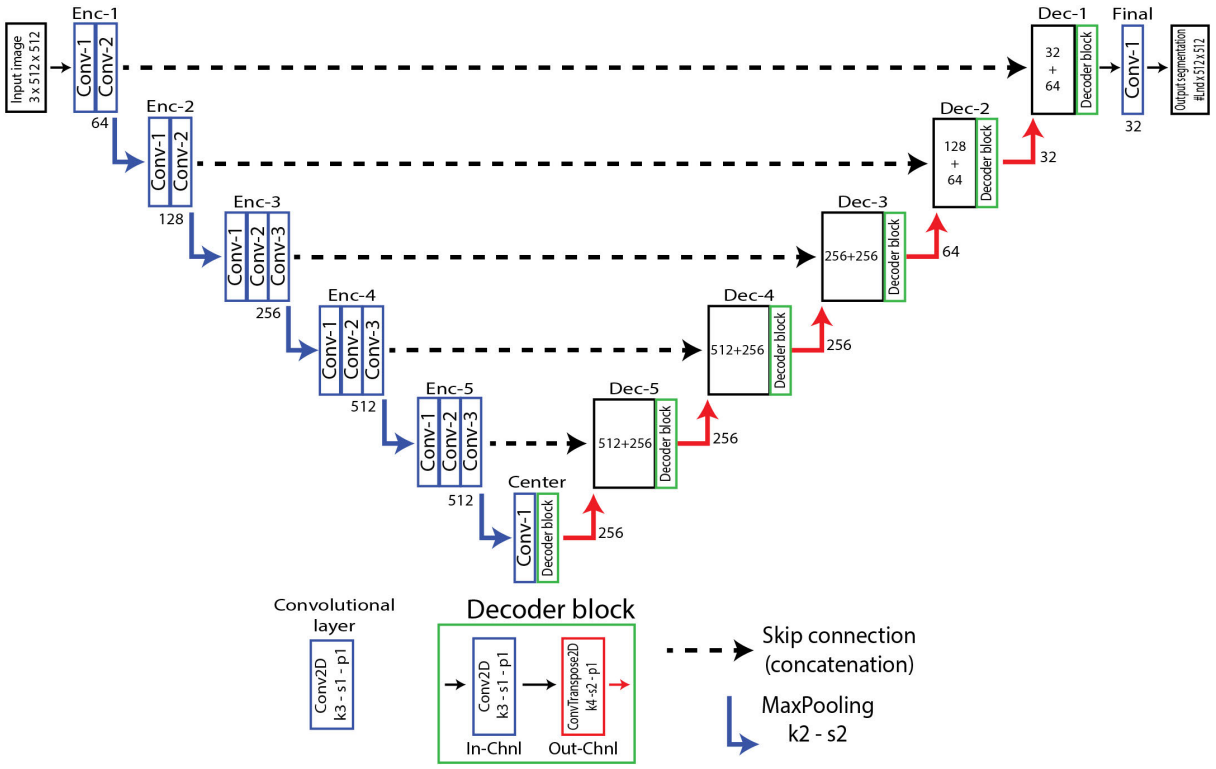

FIGURE 3. Architecture of the implemented U-Net: each (blue) *Conv* box is a convolutional layer. In the encoder, output channels are shown below the final box. In the decoder, output channels are noted next to each red arrow. A (black) box in the decoder indicates concatenation between the encoder and corresponding decoder blocks, with input channels specified inside. Each convolution and transposed convolution is followed by a ReLU operation. **Notes:** ***k:*** **kernel size,** ***s:*** **stride,** ***p:*** **padding.**

measured from the landmarks, resulting in angular errors between 0.8° and 1.2°.

II. METHODOLOGY

A. DATA

A private anonymized dataset of 919 full lower limb X-rays was employed. We retrospectively collected pre- and post-operative X-rays, and one radiologist manually annotated the X-rays. First, he placed bounding boxes surrounding the target joint. Then, he annotated the anatomical landmarks, as shown in Fig. 2a. To delineate the alignment angles (see Fig. 2), we defined the LDFA as the lateral angle between the mechanical axis of the femur ($1 \rightarrow 6$) and the tangent to the femoral condyles ($2 \rightarrow 3$). We determined the MPTA by measuring the medial angle between the mechanical axis of the tibia (center between 7 and $8 \rightarrow 6$) and the tangent to the tibial plateaus ($4 \rightarrow 5$). Lastly, aHKA and JLO were measured using (1) and (2), respectively. For the CPAK calculation, we measured the angles using the annotated landmarks and compared them with those derived from the estimated landmark positions obtained through the investigated landmark detection approaches.

We placed a circular binary mask (foreground $= 1$) centered at the desired landmark position to train the image segmentation networks. These segmentation ground truths consisted of the same number of channels as the number of targeted landmarks. To train each network, we cropped regions of interest from the full lower limb X-rays using the annotated bounding boxes. Subsequently, we resampled the extracted images to a fixed size of 512×512 pixels because of the varying image sizes among the regions of interest (ROIs). Next, we rescaled the intensities to the range [0, 1] by dividing them by the maximum intensity level of the image. During experimentation, we employed 80% of the X-rays for training and hyperparameter tuning and 20% for testing.

B. NETWORK ARCHITECTURES

We assessed two network architectures commonly employed for image segmentation, U-Net [30] and Swin-UNETR [31]. Both architectures consist of an encoder that extracts features at different resolution levels, a decoder that up-samples the features to the targeted resolution, and a set of skip connections that transfer the information from the encoder to the decoder for better localization information. In both architectures, the encoders were initialized with pretrained (ImageNet) weights [32], as it has been shown advantageous for performance [3]. For the U-Net model, this involved transferring the weights of the convolutional layers from a pretrained VGG-16 network into the encoder path of the U-Net. For the Swin-UNETR network, the pretrained weights from a Swin-S transformer were transferred to the corresponding encoder layers of the architecture. We modified the

last convolutional layer to yield masks with the same number of channels as detected landmarks.

The U-Net was based entirely on convolutional operations and followed the topology proposed by Shvets et al. [33], where the encoder relied on a VGG-16 network.[1] The VGG-16 encoder comprises 13 convolutional layers, ReLU operations, and five max pooling layers, which are utilized to extract features at various resolution levels. Fig. 3 shows the arrangement of the employed U-Net.

The Swin-UNETR proposed by Hatamizadeh et al. [31] incorporates Swin transformer operations on the encoder and convolutions on the decoder, where both branches are combined using skip connections. The shifted window (Swin) transformer [34] extracts feature representation in different resolutions by retrieving image patches of several sizes. The Swin-UNETR model employs this approach to generate non-overlapping patches from the input images and perform self-attention across various resolutions during the encoder phase. We used the available MONAI[2] implementation of the Swin-UNETR, and employed a patch size of 4×4 and a feature size of 96. The number of layers in each stage was 2, 2, 18, and 2, and the number of attention heads was set as 3, 6, 12, and 24, respectively. We set the remaining network configuration parameters to the default values defined by MONAI. These included the dropout rate and feature normalization approach.

C. MASK SIZE ANALYSIS

Previous studies have approached landmark segmentation in several ways: using a single pixel as a mask [7], [8] or employing masks with circular shapes of various sizes [3], [9], [24], [27]. To investigate how mask size affects image segmentation performance, we created binary masks for training, which consisted of a single pixel or circles with different radii (r) values: 7, 15, 30, and 45 pixels (see Fig. 2b).

D. COMPUTING LANDMARK COORDINATES FROM SEGMENTATION MAPS

The output of the image segmentation model is a probability map that assigns a floating-point value between zero and one to each pixel, indicating the likelihood that it belongs to the foreground. Since the x-y coordinates of the targeted landmark can be derived from the probability maps in multiple ways, we investigated and compared five different methods.

The first approach, *Method 1*, computes the centroid of the output probability map of the segmentation network as follows:

$$C_x = \frac{\sum_{x,y} xP(x,y)}{\sum_{x,y} P(x,y)}, \quad C_y = \frac{\sum_{x,y} yP(x,y)}{\sum_{x,y} P(x,y)}. \tag{3}$$

where C_x and C_y are the coordinates of the centroid, x and y are the corresponding x and y coordinates, and $P(x, y)$ is the probability at the pixel coordinates (x, y) of belonging to the landmark class. The second approach, *Method 2*, replicates what was proposed by Meng et al. [9] and Gai et al. [21], where the largest component in the probability map is determined before applying (3).

Method 3, our proposed methodology, processes the probability map of each landmark channel. First, we identify the landmark channel's minimum and maximum probability values. Next, the channel is thresholded at the half value between these two probabilities, retaining only the higher ones. The largest connected component is then extracted, and the centroid from the processed image is computed using (3).

Method 4 replicates the methodology implemented by Sanchez et al. [3] and computes the centroids over a binary image created using a threshold of 0.5. After thresholding, a circular ($r = 2$) binary morphological open filter is applied to remove small components. Because of the nature of single-pixel segmentation, we did not use the open filter after thresholding in this case. Lastly, *Method 5* locates the positions of the landmarks as proposed by Hsu et al. [7], where the pixel with the maximum probability is regarded as the landmark position. If a cluster of pixels that shares the maximum probability value is present, as is the case for masks with $r \geq 7$, the cluster centroid is computed.

E. COMPARISON TO ALTERNATIVE LANDMARK DETECTION METHODS

We compared image segmentation with other common landmarking approaches: (1) heatmap regression, (2) coordinate regression, and (3) segmentation-guided coordinate regression (SGR). Heatmap regression, which is widely used in medical imaging [5], followed the same U-Net topology employed for image segmentation and was trained with Gaussian heatmaps ($\sigma = 15$) generated from labeled coordinates with channels equal to the number of landmarks. Coordinate regression, also used for landmark localization [18], [20], [22], was based on a VGG-16 encoder [35] with a fully connected layer with a total of nodes equal to the landmark pixel coordinates. Finally, we implemented SGR as described by Sanchez et al. [3], a methodology that combines image segmentation with coordinate regression, achieving a balance between accuracy and robustness.

F. TRAINING AND EVALUATION

We trained an independent network for each joint and landmark detection approach using the training dataset and images of size $3 \times 512 \times 512$ (channels, height, and width). We performed 5-fold cross-validation to identify the optimal learning rate, batch size, and number of epochs for each network. Table 1 lists the values we explored. We ensured that the same fold partitions were utilized during the training for all the evaluated networks. After cross-validation, we trained the networks using the optimal hyperparameters and the complete training set. We selected the final model based on the lowest loss achieved with the optimal number of epochs.

[1] https://github.com/ternaus/robot-surgery-segmentation
[2] https://docs.monai.io/en/stable/networks.html

TABLE 1. Explored hyperparameters during the 5-fold cross-validation.

Hyperparameter	Values
Batch size	2, 4, 8
Learning rate	10^{-4}, 10^{-5}, 10^{-6}
Epochs	25, 50, 100

Lastly, we ran inference on the hold-out test set once and computed the final performance metrics.

We employed the mean squared error (MSE) as a loss function for the regression-based approaches, and a uniform weighted combination of the Dice and cross-entropy (Dice-CE) losses for the segmentation approaches. All the networks had pretrained (ImageNet) weight initialization for their encoder subparts and utilized the Adam optimizer combined with layer-wise learning rate decay. The training of the networks included a data augmentation scheme that randomly ($p = 0.5$) applied affine transformations (rotation ($\pm 10°$), scaling (± 0.1), and translation (± 0.1)) to the images and masks.

During inference, we remapped the estimated coordinates from the 512×512 coordinate space to the original coordinate space and converted them to physical points. We computed the estimated x-y coordinates for the image segmentation-based models using the methods described in Section II-D. Meanwhile, we took the maximum pixel coordinates as the estimated x-y coordinates for the heatmap networks. The coordinate regression and SGR approaches directly produce landmark coordinates in their final fully connected layer.

1) METRICS

We used the Euclidean distance to evaluate the accuracy of the developed landmark detection networks. The calculation of the Euclidean distance is:

$$d(p, q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}. \tag{4}$$

where (p_x, p_y) and (q_x, q_y) correspond to the x-y physical points of the ground truth and estimation, respectively. After measuring the Euclidean distance error, we computed the successful detection rate (SDR), which consists of determining the percentage of detections below a specific distance threshold:

$$\text{SDR}_z = \frac{d_z}{D}. \tag{5}$$

In (5), d_z represents the detections with an Euclidean distance error below the z distance threshold, and D is the total number of detections. For the assessment of CPAK classification, we employed the median absolute error (MedAE) to evaluate the error between the angles registered using the ground truth position of the landmarks and the estimated detections, and it can be expressed as:

$$\text{MedAE}(v, \hat{v}) = \text{median}(|v_1 - \hat{v_1}|, \ldots, |v_n - \hat{v_n}|). \tag{6}$$

In (6), n is the number of samples, v are the values calculated using the ground truth landmark positions, and $\hat{v}$ are the values obtained via the detected landmark positions. To assess the classification performance, we used the precision, recall, and the $\text{F}_1 - \text{score}$:

$$\text{Precision}_I = \frac{\text{TP}_I}{\text{TP}_I + \text{FP}_I}, \tag{7}$$

$$\text{Recall}_I = \frac{\text{TP}_I}{\text{TP}_I + \text{FN}_I}, \tag{8}$$

$$\text{F}_1 - \text{score}_I = 2 \times \frac{\text{Precision}_I \times \text{Recall}_I}{\text{Precision}_I + \text{Recall}_I}. \tag{9}$$

where TP, FP, and FN are the true positives, false positives, and false negatives for Class I. The macro-average metrics for the N number of classes (nine in total) are:

$$\text{Precision}_{\text{Macro-Avg.}} = \frac{\sum_I^N \text{Precision}_N}{N}, \tag{10}$$

$$\text{Recall}_{\text{Macro-Avg.}} = \frac{\sum_I^N \text{Recall}_N}{N}, \tag{11}$$

$$\text{F}_1 - \text{score}_{\text{Macro-Avg.}} = \frac{\sum_I^N F_1 - \text{score}_N}{N}. \tag{12}$$

2) INTERPRETABILITY ANALYSIS

We used Seg-Grad-CAM [36], [37] to investigate the image regions relevant to the networks during the inference process of landmark segmentation. Seg-Grad-CAM is an extension of Grad-CAM [38] that employs pixel indexes (typically those corresponding to the output of the segmentation network) to highlight the areas that contributed to the segmentation decision. To fully explore the decision-making of the networks, we analyzed the outputs of the five decoder blocks (Dec-5 to Dec-1 in Fig. 3).

3) IMPLEMENTATION

We developed the experiments using Python 3.10.4, PyTorch 1.12.0 [39], Torchvision 0.13.1 [32], and MONAI 1.0.1 [40]. PyTorch and Torchvision are two deep learning frameworks for developing computer vision models. Meanwhile, MONAI is a Python package based on PyTorch that specializes in medical imaging. It includes novel networks, loss functions, and data augmentation techniques. We used an NVIDIA A100 GPU to train the networks.

III. RESULTS

Table 2 shows the optimal hyperparameters after performing 5-fold cross-validation using the U-Net and Swin-UNETR models. Table 3 presents the results obtained by comparing the two investigated segmentation architectures using masks of varying sizes and the methods explored to calculate the coordinates of the landmarks. In Table 3, the landmarks of each joint are grouped, and we report the median Euclidean distance error among the folds of the cross-validation process and the interquartile range (IQR). Table 3 also indicates the 95^{th} percentile of the Euclidean distance error and the specific metrics (number of false positives and missed detections) for *Method 4*.

TABLE 2. Optimal hyperparameters for each type of joint using a U-Net and a Swin-UNETR in terms of Euclidean distance error. We report the median and IQR (in parentheses) from the 5-fold cross-validation scheme. In this case, the x-y landmarks' coordinates were estimated by computing the centroids of the masks without post-processing and masks with a radius of $r = 15$.

Joint	Epochs		Batch size		Learning rate		Euclidean distance [mm]	
	U-Net	Swin-UNETR	U-Net	Swin-UNETR	U-Net	Swin-UNETR	U-Net	Swin-UNETR
Femur	25	100	4	4	1.00E-05	1.00E-04	1.15 (1.10)	1.21 (1.07)
Knee	25	100	4	8	1.00E-05	1.00E-04	1.06 (1.53)	1.07 (1.53)
Ankle	25	50	2	4	1.00E-05	1.00E-04	1.70 (1.92)	1.94 (2.30)

TABLE 3. Comparison between network architectures using different landmark centroid computation techniques and masks of different sizes. The radii values are expressed in pixels. We report the median, IQR (in parentheses), and the 95^{th} percentile of the Euclidean distances of the 5-fold cross-validation scheme. The bottom shows the average number of false positives and missed detections registered only for *Method 4* over the implemented cross-validation process. Other methods have neither, and always return one set of landmark coordinates.

Median (interquartile range) of the Euclidean distance [*mm*]						
Network	Methodology	Single-pixel	$r = 7$	$r = 15$	$r = 30$	$r = 45$
U-Net	Method 1	12.90 (29.55)	1.29 (2.27)	**1.21 (1.63)**	**1.28 (1.60)**	1.44 (1.81)
	Method 2	9.39 (37.44)	1.35 (2.70)	**1.21 (1.61)**	**1.28 (1.57)**	1.43 (1.79)
	Method 3	16.98 (47.31)$^{\wedge}$	1.33 (2.11)$^{+\wedge}$	**1.21 (1.60)**$^{*\wedge}$	**1.25 (1.58)**$^{*+\wedge}$	1.41 (1.78)$^{*+\wedge}$
	Method 4	49.27 (24.54)	1.13 (1.59)	1.21 (1.61)	1.28 (1.60)	1.39 (1.65)
	Method 5	21.85 (63.72)	1.33 (2.39)	1.59 (1.77)	**3.57 (3.49)**	**5.91 (5.59)**
Swin-UNETR	Method 1	**2.36 (4.86)**	**1.21 (1.77)**	1.28 (1.73)	1.29 (1.70)	1.45 (1.74)
	Method 2	**2.13 (4.33)**	**1.20 (1.75)**	1.27 (1.71)	1.28 (1.67)	1.43 (1.74)
	Method 3	**2.16 (4.56)**	**1.20 (1.75)**$^{\wedge}$	1.27 (1.71)$^{\wedge}$	1.28 (1.67)$^{*+\wedge}$	1.41 (1.75)$^{*+\wedge}$
	Method 4	3.55 (6.22)	1.23 (1.90)	1.30 (1.76)	1.30 (1.73)	1.44 (1.77)
	Method 5	**2.07 (4.19)**	**1.19 (1.78)**	1.52 (1.76)	4.48 (4.21)	6.33 (6.25)
95^{th} percentile of the Euclidean distance [*mm*]						
Network	Methodology	Single-pixel	$r = 7$	$r = 15$	$r = 30$	$r = 45$
U-Net	Method 1	50.33	56.01	5.11	4.98	6.53
	Method 2	67.23	101.96	4.92	4.88	6.37
	Method 3	102.99	70.67	4.91	4.88	6.43
	Method 4	132.34	5.69	5.11	5.14	5.28
	Method 5	117.67	72.3	5.39	7.02	11.87
Swin-UNETR	Method 1	18.89	6.22	5.48	5.21	5.39
	Method 2	23.14	6.23	5.46	5.19	5.34
	Method 3	22.91	6.23	5.48	5.19	5.33
	Method 4	21.18	6.99	5.97	5.58	5.67
	Method 5	18.27	6.43	5.72	8.30	11.48
Method 4: Supplementary detection error analysis						
Network	Metric	Single-pixel	$r = 7$	$r = 15$	$r = 30$	$r = 45$
U-Net	False positives	121836.6	34	19.4	20.8	20.4
	Missed detections	729.6	325.6	11.6	4.6	60.2
Swin-UNETR	False positives	5688.2	100.4	34.4	22.6	22.4
	Missed detections	438.2	5.4	0.4	0.0	0.0

Bold numbers represent a significant difference (lower error) between U-Net and Swin-UNETR (Wilcoxon test with $\alpha = 0.05$).
We compared Method 3 to Methods 1, 2, and 5 for statistical (lower error) differences (Wilcoxon test with $\alpha = 0.05$):
* Statistical difference between Method 3 and Method 1 (Wilcoxon test with $\alpha = 0.016$ after Bonferroni correction)
$^{+}$ Statistical difference between Method 3 and Method 2 (Wilcoxon test with $\alpha = 0.016$ after Bonferroni correction)
$^{\wedge}$ Statistical difference between Method 3 and Method 5 (Wilcoxon test with $\alpha = 0.016$ after Bonferroni correction)

Table 2 indicates that the U-Net architecture slightly outperformed the Swin-UNETR in two of the three analyzed joints. U-Net achieved lower Euclidean errors for the femur and ankle joints. Conversely, no apparent difference was noticed between the models for the knee joint. In Table 2, divergences in the training of the architectures are recognized. We found that the Swin-based models required more epochs and larger learning rate values to achieve a performance equivalent to that obtained using fully CNN-based networks.

When comparing the network architectures (Table 3), we observe that at smaller mask sizes (single pixel and $r = 7$), Swin-UNETR leads to a more accurate performance in all methods except for *Method 4*. In *Method 4*, we distinguish that both architectures fall short in the single-pixel scenario, and a trade-off between false positives and missed detections occurs when $r \geq 7$. Compared with the Swin-UNETR model, U-Net yielded fewer false positives but more missed detections. Meanwhile, Swin-UNETR is more robust in handling missed detections but produces a higher number of false positives. When $r = 15$ or $r = 30$, the U-Net model tends to outperform the Swin-UNETR in most centroid calculation methodologies. U-Net achieved better metrics with these radii values than Swin-UNETR at $r = 7$. Conversely, there was no marked difference between the architectures when $r = 45$, except for *Method 5*, where U-Net performed better than Swin-UNETR. However, the previous

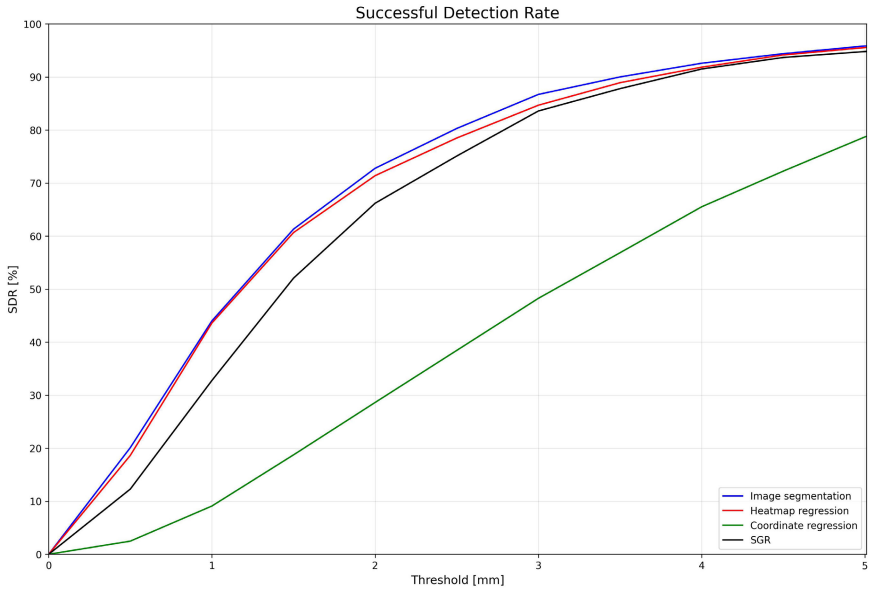

FIGURE 4. Successful detection rate at different Euclidean distance threshold levels.

TABLE 4. Comparison between the four different landmark detection techniques in terms of the median Euclidean distance error [mm] and IQR (in parentheses).

Landmark	Image segmentation	Heatmap regression	Coordinate regression	SGR
HF	1.07 (0.97)	**0.90 (0.87)***	1.55 (1.26)	1.69 (1.26)
R-FC	**1.36 (1.85)**	1.40 (1.90)	3.48 (2.27)	1.63 (1.81)
L-FC	**1.18 (1.47)***	1.26 (1.72)	4.22 (2.67)	1.38 (1.85)
R-TP	**1.10 (1.49)***	1.27 (1.74)	3.99 (2.63)	1.51 (1.81)
L-TP	**1.07 (1.40)***	1.32 (1.86)	4.97 (2.67)	1.51 (1.80)
CK	**0.61 (0.66)**	0.61 (0.67)	3.59 (1.93)	0.89 (0.72)
R-A	1.80 (1.96)	**1.55 (1.96)**	1.93 (1.65)	1.69 (1.88)
L-A	1.69 (1.84)	1.63 (1.93)	1.99 (1.78)	**1.56 (1.74)**
All	**1.16 (1.50)***	1.19 (1.61)	3.11 (2.87)	1.47 (1.67)

Bold numbers represent the method with the lowest metric.
* Lower error between Image segmentation and Heatmap regression with statistical significance difference (Wilcoxon test with $\alpha = 0.05$).
HF: head of the femur, *R-FC:* right femur condyle, *L-FC:* left femur condyle, *R-TP:* right tibial plateau, *L-TP:* left tibial plateau, *CK:* center of the knee, *R-A:* right ankle, and *L-A:* left ankle.

difference is minor compared with the metrics obtained for smaller radii values.

In terms of mask size, in Table 3, we notice that the overall accuracy improves as the dimensions of the masks increase. However, as the mask size increases, the accuracy stagnates, and the performance declines. For example, in the U-Net case, we observe an improvement from the single-pixel approach to the $r = 7$ and $r = 15$ instances. However, at $r \geq 30$, the Euclidean distance error increases. In the Swin-UNETR model, this behavior occurs from the single-pixel approach to the $r = 7$ case, and subsequently, the accuracy levels decline for larger mask sizes.

Regarding landmark coordinate identification methodologies, computing the centroid of the resultant probability map is overall more beneficial than binarizing or identifying the pixel(s) with the highest probability value(s). Regardless of the network architecture and when $r \geq 15$, centroid computation results in more accurate detections, effectively reducing erroneous detections. Moreover, adding a post-processing phase before computing the centroid of the landmarks is advantageous when $r \geq 15$. In Table 3, compared to the other methodologies, the proposed *Method 3* accurately detected landmarks in the different evaluated scenarios. Overall, *Method 3* leads to more accurate detections in terms of Euclidean distance error compared to directly calculating the centroid.

We selected the image segmentation model (U-Net with mask size $r = 15$) that performed best in the cross-validation process and trained it on the entire training dataset. We then evaluated it, using the proposed *Method 3* and the testing dataset, and compared it to other landmarking approaches commonly used in the literature. Fig. 4 and Table 4 summarize the results achieved for each landmark detection methodology using the test set. Table 4 indicates the median Euclidean distance errors and IQR on the targeted landmarks. Meanwhile, Fig. 4 groups all the detections and shows

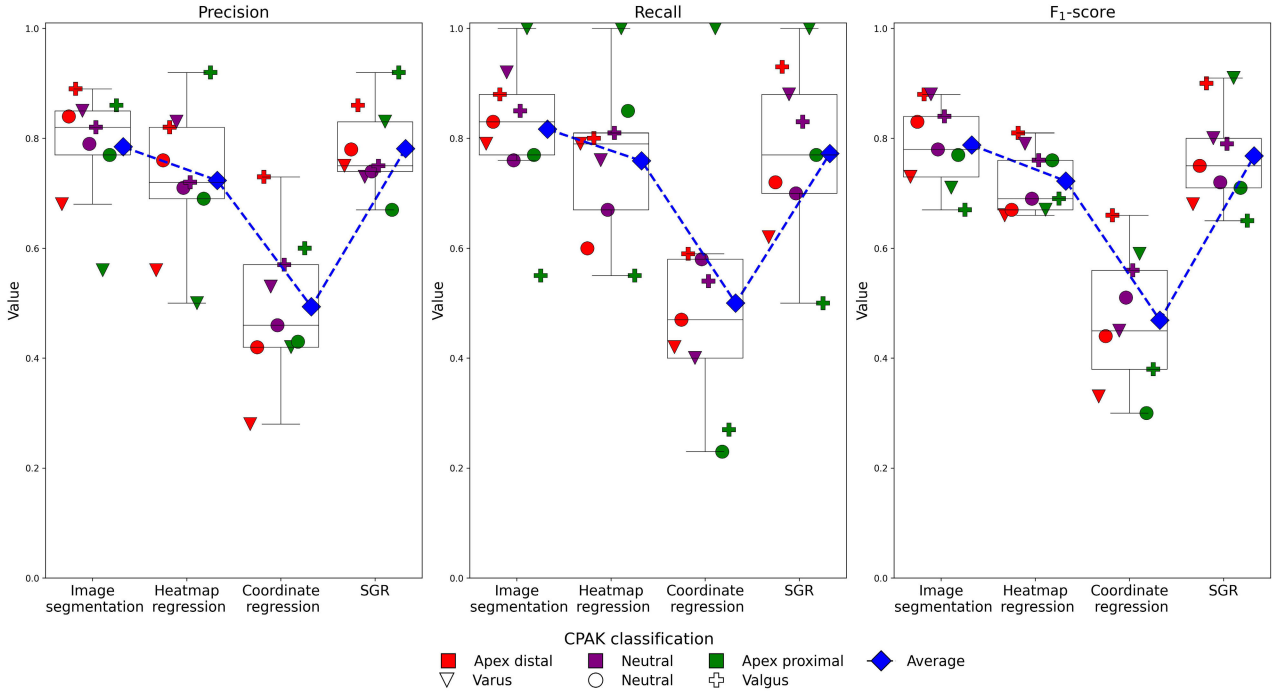

FIGURE 5. Classification metrics achieved using each landmark detection method. We computed the metrics for each class and then calculated the macro-average across all classes.

TABLE 5. Comparison between the four different landmark detection techniques to measure the angles involved in the CPAK calculation in terms of the median absolute error [°] and IQR (in parentheses).

Metric	Image segmentation	Heatmap regression	Coordinate regression	SGR
mLDFA	**0.26 (0.42)**	0.43 (0.50)	0.98 (1.38)	0.36 (0.46)
mMPTA	**0.41 (0.57)**	0.45 (0.68)	0.97 (1.30)	0.52 (0.71)
aHKA	**0.58 (0.74)**	0.71 (0.98)	1.36 (1.79)	0.61 (0.90)
JLO	**0.47 (0.80)**	0.63 (0.88)	1.48 (1.90)	0.70 (0.94)

Bold numbers indicate the best result (lower error) between Image segmentation and Heatmap regression (Wilcoxon test with $\alpha = 0.05$).
mLDFA: Mechanic lateral distal femoral angle, *mMPTA:* Mechanic medial proximal tibial angle, *aHKA:* Arithmetic hip-knee-ankle angle, *JLO:* Joint line obliquity.

the SDR of each evaluated approach as a function of the Euclidean distance cut-off value.

Table 4 demonstrates the improved performance of the image segmentation and heatmap regression approaches compared with the other methods. Heatmap regression most accurately identifies the femoral head landmark. Meanwhile, image segmentation excels in detecting knee landmarks compared with other approaches. Heatmap regression and the SGR approach are the most accurate techniques for identifying ankle landmarks. We found image segmentation to be the most accurate, with an overall median distance error of 1.16 mm (1.50), mainly due to substantially lower errors for landmarks in the knee joint. Heatmap regression closely followed the image segmentation methodology. SGR and coordinate regression were sub-optimal compared with image segmentation and heatmap regression, with coordinate regression being the inferior landmarking technique.

From Fig. 4, we can see that image segmentation performs better than the alternative methods, closely followed by heatmap regression. At 1 mm, the image segmentation achieved an $\text{SDR}_{1mm} = 44.02\%$. Meanwhile, heatmap regression yielded an $\text{SDR}_{1mm} = 43.61\%$. This difference was more pronounced at 3 mm, as image segmentation yielded $\text{SDR}_{3mm} = 86.72\%$ and heatmap regression achieved $\text{SDR}_{3mm} = 84.68\%$. Compared to SGR, the heatmap regression approach at lower thresholds had higher SDR values. However, as the threshold value increased, SGR ($\text{SDR}_{5mm} = 94.80\%$) achieved detection rates comparable to those of the heatmap ($\text{SDR}_{5mm} = 95.55\%$). Fig. 4 confirms the inferior performance of the coordinate regression approach.

We evaluated the landmark detection approaches in the CPAK classification task to provide deeper insight into their differences. Table 5 and Fig. 5 compare the detection methodologies in terms of their alignment measurement errors and classification metrics, respectively. Table 5 shows

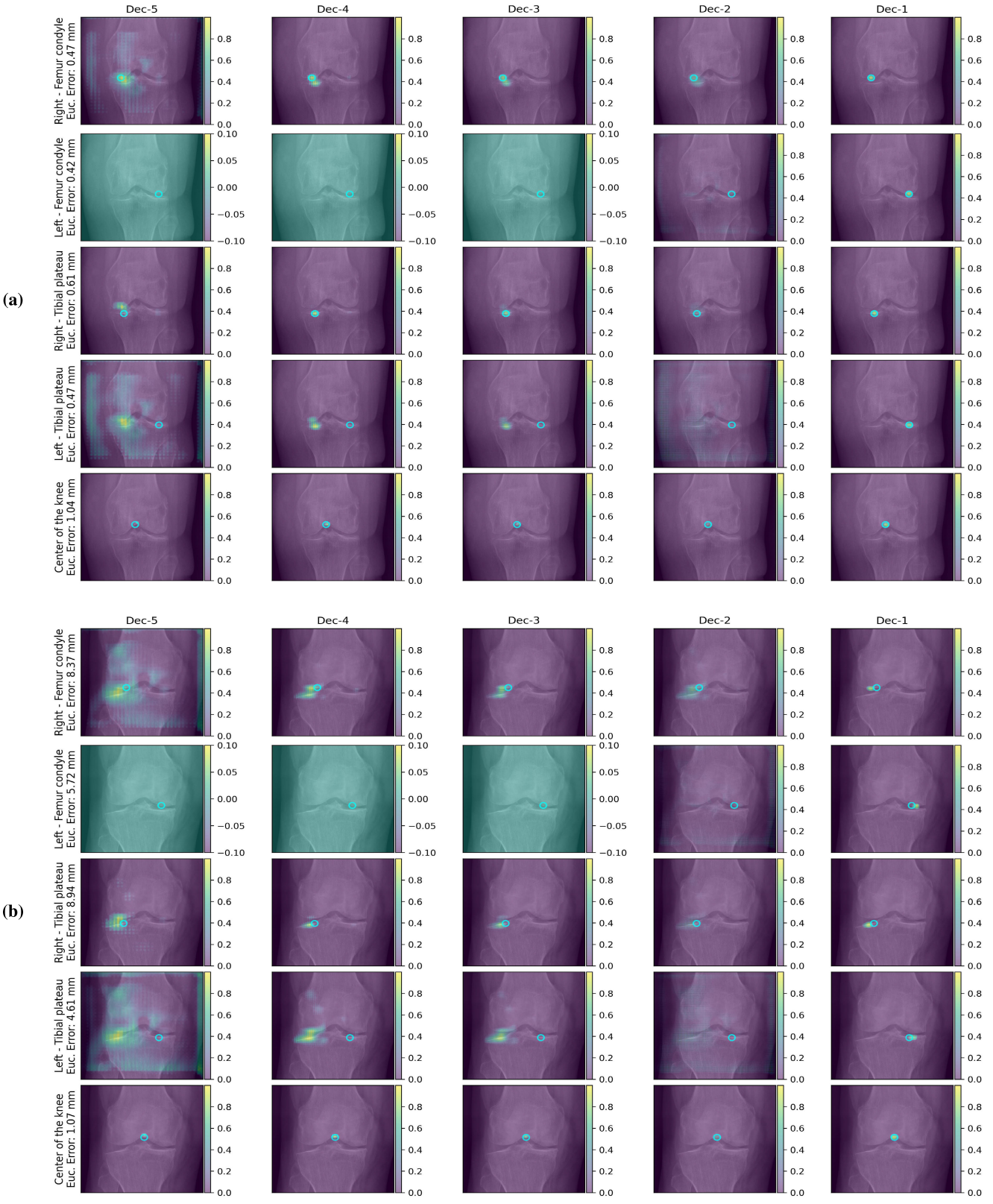

FIGURE 6. Interpretability analysis applying Seg-Grad-CAM to knee images and the decoder sections. Image with (a) landmark detection errors under 1 mm and (b) over 1 mm. Ground-truth positions are highlighted using a circle. Please refer to the online version for a better visualization.

how accurately detecting landmarks translates to lower errors when measuring alignment, since image segmentation performs better in the four addressed angles than the other detection strategies. Image segmentation is followed by heatmap regression and SGR, where the latter achieves better metrics in measuring mLDFA and aHKA than heatmap regression. Coordinate regression remains the suboptimal approach.

Fig. 5 shows that lower alignment errors lead to determining the correct CPAK class. Since image segmentation achieved the lowest errors in the alignment metrics, it registered better classification metrics. In Fig. 5, image segmentation consistently outperformed the other methods regarding precision, recall, and F_1-score in four of the nine classes (classes II, IV, V, and VI). In the remaining classes, no method is superior. What we observe is a trade-off between metrics. For instance, SGR leads in recall and F_1-score for class III, as well as precision and F_1-score for class VII. Image segmentation yields higher precision and F_1-score values for class VIII, whereas heatmap regression is the most effective methodology for class IX. For class I, no consistent method is observed. Overall, image segmentation and SGR had the highest macro-average precision (0.78), followed by the heatmap (0.72) and coordinate regression (0.49). Image segmentation yielded the highest recall (0.82) and F_1-score (0.79), followed by SGR (0.77 for both), heatmap regression (0.76 recall and 0.72 F_1-score), and coordinate regression (0.5 recall and 0.47 F_1-score).

For each leg side and type of joint image in our test set, we applied Seg-Grad-CAM and analyzed the multiple decoder blocks of the proposed U-Net. We computed the global average for each ROI and observed the overall performance across the analyzed joint and decoder blocks. Subsequently, we randomly selected 50 joints from each leg side, visually inspected the interpretability maps for recurring patterns, and compared them with the global average performance.

Fig. 6 shows two representative knee joint cases where Seg-Grad-CAM was applied on multiple decoder sections: one where the landmark detection errors were approximately 1 mm and a second where the errors exceeded 1 mm. First, we observe that for landmarks with errors of around 1 mm, the network identifies correct locations in earlier decoder stages, as the heatmap is more noticeable around the desired position (see Dec-5 for the right femur condyle and tibial plateau). This is not the case for landmarks with higher detection errors, where, in early decoder blocks, the heatmaps pinpoint broader areas and their positioning is refined in later decoder layers.

Additionally, from Fig. 6, we observe how the landmark detection refinement process occurs at various decoder levels. For instance, the right femur condyle and tibial plateau landmarks display highlighted areas surrounding the target locations on Dec-5, shifting towards expected positions closer to Dec-1. We can also observe that the center of the knee is easily detected, as the network already identifies it in Dec-5. Meanwhile, landmarks on the left side of the knee (from the patient's perspective) exhibit a pattern where, in Dec-5, the network fails to focus on relevant features or highlights features from the opposite side. It is not until Dec-1 that the correct left landmark positioning is shown.

In the Supplementary Material, Figures 1 and 2 display two representative cases of the hip and ankle joints: one case with correct (approximately 1 mm error) landmark detection and a second with detections exceeding 1 mm. For the femoral head landmark, accurate detections are paired with outputs from Seg-Grad-CAM that indicate an early focus (Dec-5) in the desired region. In contrast, detections with higher landmark detection errors are associated with outputs where the correct location appears only in the later sections of the decoder. A performance comparable to that of the left-placed knee landmarks is observed for the ankle landmarks. At the earlier stages of the decoder (Dec-5), Seg-Grad-CAM highlights features from the opposite side, with subsequent correction and refinement occurring.

IV. DISCUSSION

A key finding from our work arises from comparing the fully CNN-based U-Net model with the Swin-UNETR, which incorporates Swin transformers on the encoder path. The current literature on medical imaging has revealed a growing interest in transformer-based deep learning architectures [41]. However, evidence indicates that CNN-based networks can achieve superior results compared to transformers [42]. In this work, our findings reveal that Swin-UNETR outperforms U-Net at lower mask sizes (single pixel and $r = 7$), whereas U-Net is superior for $r = 15$ or $r = 30$. At $r = 45$, we observed an equivalent performance.

In *Method 4* and $r \geq 7$, we observed that Swin-UNETR was more robust to missed detections, but false positives hampered its performance. U-Net exhibited the opposite behavior: fewer false positives at the cost of more missed detections. We theorize that the multi-head attention layer, which enables transformers to model dependencies and contextual relationships between pixels [43], contributes to the Swin-UNETR being more robust at lower mask sizes and achieving better metrics. As the mask size increases, it becomes easier for the CNN to learn the features required to segment circular masks from an image. These results encourage further exploration of the best deep learning architecture. A CNN could be a first option if resources are limited, since transformers require more training time and data [44]. In our case, the best performance was found by CNNs, so we selected this architecture for further experiments.

A second finding stems from analyzing the impact of the mask size. We observed that smaller masks, particularly the single-pixel ones, yielded less accurate detections. This finding contradicts those reported by He et al. [8] and Hsu et al. [7]. Nonetheless, He et al. employed an output resolution of 56×56, and Hsu et al. of 224×224, significantly reducing the search space. In our case, the output resolution

was 512 × 512, making it challenging to segment smaller masks. Additionally, in the case of Hsu et al., their objective was to segment 64 facial landmarks. This condition facilitates segmentation, as the network learns to segment multiple single-pixel regions instead of just one, as observed in the case of the femoral head landmark.

Furthermore, we observed that as the size of the masks increased, the detection accuracy improved. However, beyond a certain point, the accuracy plateaus and then begins to decrease. We theorize that when the masks are too small, the networks struggle to learn the mapping from the X-ray to the mask, primarily because of the 512 × 512 output resolution. Conversely, when the masks are oversized, the network fails to correctly learn to differentiate between anatomical landmarks, resulting in segmentations that are not centered in the correct location. Therefore, using a medium-size mask ($r = 15$) generally leads to accurate detections.

As a third finding, we observed that determining the centroid from the probability map resulted in accurate detections. Employing the maximum probability (*Method 5*) to determine the landmark's positions was sub-optimal, likely due to the fact that it bases its estimate on a single pixel, and is more sensitive to noise. Computing the centroid avoids such behavior because it captures the average position, resulting in a less sensitive and more accurate estimation. Due to the binarization of the probability maps using *Method 4*, we encountered missed detections and false positives. Methods 1 to 3 avoid erroneous detections because they compute the centroid directly from the probability map, rather than from a binary mask.

Overall, adding a post-processing scheme, either *Method 2* or *Method 3*, translates to more accurate detections compared with directly calculating the centroid from the probability map. Only in the U-Net with $r = 7$ did *Method 1* result in a more accurate approach. We theorize that in this particular case, the application of a post-processing step leads to the selection of an erroneous component, resulting in inaccurate detections. For values of $r \geq 30$, the proposed centroid computation method achieved better accuracy metrics.

Equivalent to what was done by Hsu et al. [7] in the facial landmark recognition domain, we considered it necessary to compare landmarking approaches to help authors in future lower-limb X-ray imaging research. In the literature on whole lower limb X-ray imaging, solutions based on either image segmentation, heatmap, or coordinate regression can be found; see Section I-B. In our study, image segmentation, on average, and specifically for landmarks located in the knee, proved to be the most accurate approach. Regression-based approaches achieved better metrics for landmarks placed on the head of the femur and ankle.

Since image segmentation accurately detected the knee landmarks, the mLDFA and mMPTA angles were precisely measured, ensuring the proper calculation of the aHKA and JLO. This situation led to the correct classification of each lower limb compared to other landmark detection approaches on average, as well as for four of the CPAK classes. Conversely, coordinate regression performed inferiorly in detecting landmarks of the lower limb, resulting in erroneous alignment measurements and incorrect classifications. Hence, there is a relationship between accurate landmark detection and the correct classification of the lower limbs using the CPAK system.

Such a relationship is worth further analysis, as a trade-off between aHKA and JLO was observed when comparing heatmap regression to SGR, despite heatmap regression yielding more accurate landmark detections than SGR. The positioning of the detected landmarks could explain this situation. Although closer to the ground-truth positions, the detections from the heatmap regression approach could be ill-positioned, resulting in deviations in the delineation of the axes and, consequently, in the measurement of alignment. Such minor deviations, combined with the nature of the hard thresholds defined on CPAK, could explain the better performance of SGR over heatmap regression in the CPAK classification.

Although the original implementation of Seg-Grad-CAM [36] proposed examining layers closer to the network's bottleneck, Hasany et al. [45] pointed out that analyzing layers in the decoder section is more informative since the refinement of the final segmentation can be observed. In the cases we examined, this process was observed for most landmarks, and we also noted the impact of the detection error. For landmarks with higher distance errors, we found that the model was unable to identify the targeted regions in the earlier decoder layers, and landmark identification was refined in subsequent layers.

Our interpretability analysis also revealed that specific landmarks seemed easier to detect. For example, the knee center and femoral head appeared early in the decoder, whereas landmarks lateral or medial to the joint center appeared later. For some of the latter landmarks, the network appears uncertain during the earlier layers about whether to focus on the right or left side of the image. We theorize that the specifics of our landmark definition and network training played a role. Landmarks are typically defined anatomically, considering the medial and lateral sides. Meanwhile, our processing of the regions corresponding to both limbs considered landmarks placed on the right or left side of the image. We did not perform image mirroring [20] or use images from a single leg side [27] to align the anatomical and image points of view during network training, which we theorized has hindered the early identification of landmark locations and should be investigated further.

To shed some light on the quality of the achieved results, we compared image segmentation, where possible, with the results presented in the literature for equivalent landmarks in lower limb X-rays. Interpretation should be

TABLE 6. Landmark localization accuracy comparison between our proposed methods against what is reported in the literature in terms of mean Euclidean distance error [mm] and ± standard deviation. A direct comparison is impractical due to the differences in the datasets employed between methods. Hence, we only show the localization error of equivalent landmarks to observe possible trends.

Landmark	Ours	Sanchez S.A. *et al.* [3]	Tack A. *et al.* [20]	Tsai A. [17]
HF	1.58 ± 4.94	1.43 ± 1.04	1.72 ± 1.00	3.7 ± 5.3
R-FC	1.87 ± 1.77	2.29 ± 2.25	-	-
L-FC	1.65 ± 1.60	2.11 ± 1.79	-	-
R-TP	1.64 ± 2.00	2.45 ± 2.53	-	-
L-TP	1.63 ± 2.14	2.49 ± 2.37	-	-
CK	0.74 ± 0.57	1.10 ± 1.14	1.94 ± 1.33	2.9 ± 6.3
R-A	2.91 ± 6.64	2.30 ± 1.91	1.54 ± 1.33[1]	4.2 ± 1.8[1]
L-A	3.18 ± 10.74	2.09 ± 1.48		

HF: head of the femur, *R-FC*: right femur condyle, *L-FC*: left femur condyle, *R-TP*: right tibial plateau, *L-TP*: left tibial plateau, *CK*: center of the knee, *R-A*: right ankle, and *L-A*: left ankle.
[1]They computed the center of the ankle.

conducted carefully, as a direct and objective comparison is not possible due to the different data employed in each work. Consequently, the following findings focus on the observed trends. Table 6 shows that image segmentation outperformed heatmap regression [17] in all the landmarks. Compared to us, no detection framework was employed to identify the regions of interest to detect the landmarks, affecting the estimation of the coordinates.

Compared with SGR [3] and cascade coordinate regression [20], image segmentation is more accurate for knee landmarks. SGR is more accurate than image segmentation for the femoral head landmark, but image segmentation is superior to cascade regression. Meanwhile, the SGR and cascade coordinate regression better localize the ankle landmarks. This analysis indicates a relationship between the nature of the detected landmarks and the performance of image segmentation.

We believe that the previous interpretation is worth exploring, as we observed a decreased performance in the ankle joint compared to landmarks in other joints. After visual inspection, the results revealed that image segmentation performed suboptimally when the ankle was poorly imaged. Under this condition, image segmentation produces elongated outputs that affect the computation of the centroid. Since SGR and cascade regression learn to output pixel coordinates directly, these calculations are not hampered as in the case of image segmentation. A solution to such a condition could be to redefine the ankle landmarks, for example, as done by Tack et al. [20] and Tsai [17], where a single landmark at the center of the tibio-talar joint was employed.

In line with the comparison with regression-based methodologies, we contrasted our findings with those of Meng et al. [9], who also employed image segmentation as a detection technique. As mentioned above, we report general findings. Compared with ours, Meng et al. achieved better detection rates for equivalent landmarks in terms of SDR_{3mm}. This condition could be due to the network architecture employed (VB-Net) or the use of independent networks for the right- and left-leg images. Similar to our results, the results achieved by Meng et al. highlight the advantageous use of image segmentation as a landmark detection technique.

A. LIMITATIONS AND FUTURE WORK

Several limitations should be mentioned. In this study, we used a dataset from a single institution. Moreover, we evaluated image segmentation for landmark detection using lower-limb X-rays. Whether our findings hold when using other imaging modalities, such as CT or MRI, or images from other anatomic regions, such as the chest or brain, remains to be examined. Finally, we performed a qualitative interpretability analysis through manual inspection of a subset of the test set. Future work should include a more formal and quantitative interpretability analysis, and independently validate the performance of the developed methodology.

V. CONCLUSION

Image segmentation leads to superior results in terms of landmark detection (Mdn. 1.16 mm (1.50)) and CPAK classification (F_1-$score_{Macro\text{-}Avg.} = 0.79$) accuracy compared to alternative detection approaches tuned and tested on the same data, and comparable results with respect to methods recently reported in the literature, highlighting its potential as a landmark detection technique.

REFERENCES

[1] B. Ibragimov and T. Vrtovec, "Landmark-based statistical shape representations," in *Statistical Shape and Deformation Analysis*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 89–113.
[2] S. K. Zhou and Z. Xu, "Landmark detection and multiorgan segmentation: Representations and supervised approaches," in *Handbook of Medical Image Computing and Computer Assisted Intervention*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 205–229.
[3] S. Amador Sanchez, P. van Overschelde, and J. Vandemeulebroucke, "Segmentation-guided coordinate regression for robust landmark detection on X-rays: Application to automated assessment of lower limb alignment," *IEEE Access*, vol. 12, pp. 61484–61497, 2024.
[4] R. Kaur and S. Singh, "A comprehensive review of object detection with deep learning," *Digit. Signal Process.*, vol. 132, Jan. 2023, Art. no. 103812.
[5] J. Kang, K. Oh, and I.-S. Oh, "Accurate landmark localization for medical images using perturbations," *Appl. Sci.*, vol. 11, no. 21, p. 10277, Nov. 2021.

[6] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Comput. Surv.*, vol. 56, no. 1, pp. 1–37, 2023.

[7] C.-F. Hsu, C.-C. Lin, T.-Y. Hung, C.-L. Lei, and K.-T. Chen, "A detailed look at CNN-based approaches in facial landmark detection," 2020, *arXiv:2005.08649*.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[9] X. Meng, Z. Wang, X. Ma, X. Liu, H. Ji, J.-Z. Cheng, and P. Dong, "Fully automated measurement on coronal alignment of lower limbs using deep convolutional neural networks on radiographic images," *BMC Musculoskeletal Disorders*, vol. 23, no. 1, p. 869, Sep. 2022.

[10] D. Paley, *Principles of Deformity Correction*. Cham, Switzerland: Springer, 2014.

[11] D. J. Hunter, L. Sharma, and T. Skaife, "Alignment and osteoarthritis of the knee," *JBJS*, vol. 91, pp. 85–89, Jan. 2009.

[12] S. J. MacDessi, W. Griffiths-Jones, I. A. Harris, J. Bellemans, and D. B. Chen, "Coronal plane alignment of the knee (CPAK) classification: A new system for describing knee phenotypes," *Bone Joint J.*, vol. 103, no. 2, pp. 329–337, Feb. 2021.

[13] I. E. W. G. Laven, F. F. Schröder, F. de Graaff, J. C. Rompen, R. A. G. Hoogeslag, and A. H. van Houten, "Accuracy, inter- and intrarater reliability, and user-experience of high tibial osteotomy angle measurements for preoperative planning: Manual planning PACS versus semi-automatic software programs," *J. Experim. Orthopaedics*, vol. 9, no. 1, p. 44, Dec. 2022.

[14] J. Schock, D. Truhn, D. B. Abrar, D. Merhof, S. Conrad, M. Post, F. Mittelstrass, C. Kuhl, and S. Nebelung, "Automated analysis of alignment in long-leg radiographs by using a fully automated support system based on artificial intelligence," *Radiol., Artif. Intell.*, vol. 3, no. 2, Mar. 2021, Art. no. e200198.

[15] R. Vaishya, V. Vijay, V. P. Birla, and A. K. Agarwal, "Inter-observer variability and its correlation to experience in measurement of lower limb mechanical axis on long leg radiographs," *J. Clin. Orthopaedics Trauma*, vol. 7, no. 4, pp. 260–264, Oct. 2016.

[16] I. Shichman, M. Roof, N. Askew, L. Nherera, J. C. Rozell, T. M. Seyler, and R. Schwarzkopf, "Projections and epidemiology of primary hip and knee arthroplasty in medicare patients to 2040–2060," *JBJS Open Access*, vol. 8, no. 1, p. 22, Jan. 2023.

[17] A. Tsai, "Anatomical landmark localization via convolutional neural networks for limb-length discrepancy measurements," *Pediatric Radiol.*, vol. 51, no. 8, pp. 1431–1447, Jul. 2021.

[18] T. P. Nguyen, D.-S. Chae, S.-J. Park, K.-Y. Kang, W.-S. Lee, and J. Yoon, "Intelligent analysis of coronal alignment in lower limbs based on radiographic image with convolutional neural network," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103732.

[19] F. Erne, P. Grover, M. Dreischarf, M. K. Reumann, D. Saul, T. Histing, A. K. Nüssler, F. Springer, and C. Scholl, "Automated artificial intelligence-based assessment of lower limb alignment validated on weight-bearing pre- and postoperative full-leg radiographs," *Diagnostics*, vol. 12, no. 11, p. 2679, Nov. 2022.

[20] A. Tack, B. Preim, and S. Zachow, "Fully automated assessment of knee alignment from full-leg X-rays employing a 'YOLOv4 and resnet landmark regression algorithm,' (YARLA): Data from the osteoarthritis initiative," *Comput. Methods Programs Biomed.*, vol. 205, Jun. 2021, Art. no. 106080.

[21] L. Gai, Z. Qiao, L. Fan, X. Meng, S. Fang, P. Dong, and Z. Qian, "MMAN: Multi-task and multi-scale attention network for concurrently lower limbs segmentation and landmark detection," in *Proc. IEEE 20th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2023, pp. 1–5.

[22] M. Bonnin, F. Müller-Fouarge, T. Estienne, S. Bekadar, C. Pouchy, and T. A. S. Selmi, "Artificial intelligence radiographic analysis tool for total knee arthroplasty," *J. Arthroplasty*, vol. 38, no. 7, pp. S199–S207.e2, Jul. 2023.

[23] S. E. Kim, J. W. Nam, J. I. Kim, J.-K. Kim, and D. H. Ro, "Enhanced deep learning model enables accurate alignment measurement across diverse institutional imaging protocols," *Knee Surg. Rel. Res.*, vol. 36, no. 1, p. 4, Jan. 2024.

[24] A. P. Mika, Y. Suh, R. W. Elrod, M. Faschingbauer, D. C. Moyer, and J. R. Martin, "Novel dilation-erosion labeling technique allows for rapid, accurate and adjustable alignment measurements in primary TKA," *Comput. Biol. Med.*, vol. 185, Feb. 2025, Art. no. 109571.

[25] S. Simon, G. M. Schwarz, A. Aichmair, B. J. H. Frank, A. Hummer, M. D. DiFranco, M. Dominkus, and J. G. Hofstaetter, "Fully automated deep learning for knee alignment assessment in lower extremity radiographs: A cross-sectional diagnostic study," *Skeletal Radiol.*, vol. 51, no. 6, pp. 1249–1259, Jun. 2022.

[26] J. R. Steele, S. J. Jang, Z. R. Brilliant, D. J. Mayman, P. K. Sculco, S. A. Jerabek, and J. M. Vigdorchik, "Deep learning phenotype automation and cohort analyses of 1,946 knees using the coronal plane alignment of the knee classification," *J. Arthroplasty*, vol. 38, no. 6, pp. S215–S221.e1, Jun. 2023.

[27] C. Jo, D. Hwang, S. Ko, M. H. Yang, M. C. Lee, H.-S. Han, and D. H. Ro, "Deep learning-based landmark recognition and angle measurement of full-leg plain radiographs can be adopted to assess lower extremity alignment," *Knee Surg., Sports Traumatol., Arthroscopy*, vol. 31, no. 4, pp. 1388–1397, Apr. 2023.

[28] S. Huber, J. A. Mitterer, S. M. Vallant, S. Simon, F. Hanak-Hammerl, G. M. Schwarz, A. Klasan, and J. G. Hofstaetter, "Gender-specific distribution of knee morphology according to CPAK and functional phenotype classification: Analysis of 8739 osteoarthritic knees prior to total knee arthroplasty using artificial intelligence," *Knee Surg., Sports Traumatol., Arthroscopy*, vol. 31, no. 10, pp. 4220–4230, Oct. 2023.

[29] S. E. Kim, S. MacDessi, D. Song, J. I. Kim, B. S. Choi, H.-S. Han, and D. H. Ro, "Coronal plane alignment of the knee (CPAK) type shifts toward constitutional varus with increasing Kellgren and Lawrence grade: A radiographic analysis of 17,365 knees," *JBJS*, vol. 107, no. 3, pp. 229–236, Feb. 2021.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[31] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Proc. Int. MICCAI Brainlesion Workshop*, Jan. 2022, pp. 272–284.

[32] T. Maintainers Contributors. (2016). *Torchvision: PyTorch's Computer Vision Library*. [Online]. Available: https://github.com/pytorch/vision

[33] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2018, pp. 624–628.

[34] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[36] K. Vinogradova, A. Dibrov, and G. Myers, "Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 13943–13944.

[37] J. Gildenblat contributors. (2021). *Pytorch Library for CAM Methods*. [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam

[38] 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," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[39] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.

[40] M. J. Cardoso et al., "MONAI: An open-source framework for deep learning in healthcare," 2022, *arXiv:2211.02701*.

[41] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—A contemplative retrospection," *Eng. Appl. Artif. Intell.*, vol. 122, Jun. 2023, Art. no. 106126.

[42] D. Ma, M. R. H. Taher, J. Pang, N. U. Islam, F. Haghighi, M. B. Gotway, and J. Liang, "Benchmarking and boosting transformers for medical image classification," in *Proc. MICCAI Workshop Domain Adaptation Represent. Transf.*, Jan. 2022, pp. 12–22.

[43] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Appl. Sci.*, vol. 13, no. 9, p. 5521, Apr. 2023.

[44] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 5, pp. 7478–7498, Jun. 2023.

[45] S. N. Hasany, F. Mériaudeau, and C. Petitjean, "The do's and don'ts of grad-CAM in image segmentation as demonstrated on the synapse multi-organ CT dataset," in *Medical Imaging with Deep Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=rnQUJLbODk

SEBASTIAN AMADOR SANCHEZ received the B.Eng. degree in biomedical engineering from the National Polytechnic Institute, Mexico City, Mexico, in 2015, and the M.S. degree in biomedical engineering from the Vrije Universiteit Brussel, Brussels, Belgium, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Electronics and Informatics.

As a Ph.D. candidate, he has collaborated closely on projects to develop computer-aided diagnosis systems to analyze COVID-19 CT images, detect landmarks in lower limb X-rays, and suppress bone structure in chest X-rays. His research interests include deep learning, computer vision, image segmentation, image registration, and object detection.

ASHKAN ZARGHAMI received the B.Sc. and M.Sc. degrees in electrical engineering from Azad University, Iran, in 2014 and 2016, respectively, and the M.Sc. degree in biomedical engineering from Vrije Universiteit Brussel, in 2023.

He is currently a member of the Department of Electronics and Informatics, Vrije Universiteit Brussel. During the second M.Sc. degree, he worked on detecting landmarks in lower limb X-rays. His research interests include deep learning, computer vision, image segmentation, transient radar methods, microwave engineering, and wall scanning technologies.

PHILIPPE VAN OVERSCHELDE received the medical degree from the University of Gent, in 1999, the Master of Science degree in biomedical and clinical engineering, in 2004, and an Executive M.B.A. degree from the Vlerick Business School, in 2016.

He completed the Orthopaedic Surgery Training at the University of Gent. He spent a year as a fellow of Hôpital Sainte Marguerite, France, working with Prof. Jean-Noël Argenson, and also briefly at the Melbourne Orthopaedic Group with Dr. David Young. In 2006, he joined the Hip and Knee Department, AZ Maria-Middelares, Ghent. His knee research centers on soft tissue balancing in total knee arthroplasty (TKA) and alignment strategies. He has over 20 years of experience with the medial pivot concept in total knee arthroplasty (TKA) and was the first in Europe to use augmented reality in TKA surgery, incorporating robotics for personalized alignment. His hip research focuses on anatomic restoration through the direct anterior approach, with expertise in complex acetabular defects and hip resurfacing via the posterolateral approach. He co-founded a Digital Therapeutics startup, in 2015, to address patient dissatisfaction after joint replacements using data-driven solutions. He is actively involved in numerous national and international societies and serves on the board of the Belgian Knee Society.

JEF VANDEMEULEBROUCKE received the master's degree in electronic engineering from the University of Ghent, Belgium.

In Granada, Spain, he specialized in artificial intelligence for one year. He performed postgraduate training in numerical optimization techniques at the Federal University of Santa Catarina (UFSC), Florianópolis, Brazil. His doctoral research focused on lung motion estimation and modeling for image-guided radiation therapy, conducted in collaboration with the Creatis Laboratory, University of Lyon 1, France, and the Center for Machine Perception, Czech Technical University in Prague, Czech Republic. He is currently an Associate Professor of medical image analysis with the Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Belgium. He is also an Affiliated Researcher with Imec, an international research and innovation hub in nanoelectronics and digital technologies. His current research interests include medical image analysis for applications in computer-aided diagnosis and image-guided interventions, with a focus on thoracic, whole-body, and dynamic imaging for oncology and musculoskeletal pathologies.