IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 11, 2020, accepted October 26, 2020, date of publication October 28, 2020, date of current version November 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034418

Deep 3D Convolutional Networks to Segment Bones Affected by Severe Osteoarthritis in CT Scans for PSI-Based Knee Surgical Planning

DAVIDE MARZORATI[1], **MATTIA SARTI**[1], **LUCA MAINARDI**[1], **(Member, IEEE), ALFONSO MANZOTTI**[2], **AND PIETRO CERVERI**[1]

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133 Milan, Italy
[2]Department of Orthopaedics and Trauma, Luigi Sacco Hospital, ASST FBF-Sacco, 20157 Milan, Italy

Corresponding author: Pietro Cerveri (pietro.cerveri@polimi.it)

ABSTRACT Segmentation of bony structures in CT scans is a crucial step in knee arthroplasty based on personalized surgical instruments (PSI). As a matter of fact, the success of the surgery depends on the quality of the matching between the patient-specific resection jigs, manufactured exploiting the patient bony surfaces attained by segmentation, and true patient surfaces. Severe pathological conditions as chronic osteoarthritis, deteriorating the cartilages, narrowing the intra-articular spaces and leading to bone impingement, complicate the segmentation making the recognition of bony boundaries sub-optimal for traditional semi-automated methods and often extremely difficult even for expert radiologists. Deep convolutional neural networks (CNNs) have been investigated in the last years towards automatic labeling of diagnostic images, especially harnessing the encoding-decoding U-Net architecture. In this article, we implemented deep CNNs to encompass the concurrent segmentation of the distal femur and the proximal tibia in CT images and evaluate how segmentation uncertainty may impact on the surgical planning. A retrospective set of 200 knee CT scans of patients was used to train the network and test the segmentation performances. Tests on a subset of 20 scans provided median dice, sensitivity and positive predictive value indices greater than 96% for both shapes, with median 3D reconstruction error in the range of 0.5mm. Median 3D errors on both PSI femoral and tibial contact areas and surgical cut alignments were less than 2mm and 2°, respectively, which can be considered clinically acceptable. These results substantiate that deep CNN architectures can disclose the opportunity of segmenting bone shapes in CT scans for PSI-based surgical planning with promising accuracy. However, we observed that segmentation scores alone cannot be taken as representative of the 3D errors at the contact areas of the PSI. Therefore when comparing segmentation algorithms of PSI-based surgical planning the 3D errors should be explicitly analyzed.

INDEX TERMS 3D U-net, bone segmentation, CT images, deep learning, knee arthroplasty, osteoarthritis, personalized surgical instruments.

I. INTRODUCTION

Computed tomographic (CT) and magnetic resonance (MR) imaging are competing techniques to perform surgical planning in knee arthroplasty by means of personalized surgical instruments (PSI) and customized implants [1]–[5]. The three-dimensional (3D) geometry of patient bones, obtained through image segmentation and surface reconstruction, is crucial to identify clinical landmarks, establish the optimal femoral and tibial resections, decide the implant size, optimize the implant location in the different planes towards the recovery of knee joint mechanics. According to the preoperative plan, patient-specific cutting jigs are designed, manufactured and used during the surgery to accurately drive the bone resection avoiding invasive intra-medullar instruments [6]–[8]. In this clinical pipeline, image segmentation plays a fundamental role as it influences the reconstruction accuracy of digital bone surfaces, the matching of the personalized instrument to the true bone geometry and ultimately the resection alignment, leaving therefore the overall surgical

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

performance particularly susceptible to segmentation uncertainty [9]–[12]. Sub-millimetre correspondence between patient surface and jig footprint has been claimed as geometrical target for successful PSI-based knee surgery [13], [14]. Due to such demanding requirement, bone segmentation is manually performed by expert orthopaedic radiologists using clinical image management and visualization tools. Nonetheless, due to large variations in bone shape and size among individuals, the process is very time consuming, labor intensive and usually the results are subjected to significant inter-operator variability [15]–[17]. Segmentation is further complicated by the effects of pathological disorders affecting cartilages and bone surfaces, whose morphology can significantly differ from physiological conditions. For instance, osteoarthritis or post traumatic sequelae, deteriorating the cartilages and narrowing the intra-articular spaces, makes the cortical profiles jagged, thus complicating the delineation of bony boundaries and often requiring at least one additional quality cross-check performed by a different expert radiologist. As soon as the severity increases to the chronic condition, bone impingement occurs causing the interface between adjacent bones to become almost indistinguishable. In addition, the formation of osteophytes, especially in case of long-term impingement, makes the bony profiles extremely irregular and the delineation of surface boundaries difficult even for expert radiologists [18]–[20]. Not only the quality of the manual segmentation of lower limb bones is prone to such complexity but also traditional semi-automated algorithms, based on gray histograms, edge detection, region growing and statistical shape models were proven sub-optimal requiring extensive manual post-processing [21]–[24]. As a consequence, due to the limited speed and weak robustness of bone segmentation on CT scans, such techniques have had limited spread hitherto in PSI-based surgical planning for knee arthroplasty. More recently, bone segmentation of 2D images and 3D scans has been addressed as a pixel-wise classification problem by leveraging multi-layer convolutional neural networks (CNNs), trained by supervised deep learning (DL) algorithms [25]–[27]. This interest has been motivated by several successes achieved by such methodology to solve complex problems such as text translation, natural voice generation, lip reading, road sign recognition, image synthesis and game challenges [28]. With respect to traditional machine learning approaches, DL demonstrated superior ability to discover from the original data the fundamental features that determine the success of the specific task. The translation of this principle to image segmentation is that representative features (e.g. lines, boundaries, contours, 2D/3D geometries) can be learned by the DL multi-layer network directly from raw images and the corresponding labeled images, without requiring pre-processing or any a-priori assumption about the complexity of the shape to be segmented [29], [30]. Driven by the segmentation task, DL exploits, layer by layer in the network, data encoding with spatial down-sampling to synthesize significant shape characteristics, and then data decoding with spatial up-sampling to build full-resolution segmentation. The ability to learn the complexity of the image structure may be extended to the bone case in a wide spectrum of variability seen in a huge number of different pathological conditions. A variant of this encoding-decoding network is represented by the so called U-Net that has been recently proposed in the field of biomedical image segmentation [31]. Both 2D and 3D U-Net models have been investigated for segmentation of hand bones in X-ray images [32], mandibular bones in cranio-facial CT [33], femur in CT scans [30] and major skeletal bones in whole-body CT scans [34]. All such papers focused mainly on binary segmentation of bones against the image background, used low resolution data, disregarded the effects of large bone deformations and osteophytes on the segmentation quality. Overall the impact of segmentation errors the surgical planning was underestimated. In order to address such challenges, in this work we proposed to investigate CNN networks based on 2D and 3D U-Net architectures for the automatic semantic segmentation of the distal femur and proximal tibia. Segmentation quality was evaluated to verify whether such methods may be adopted with reliability to reconstruct anatomical surfaces used in surgical planning for PSI-based knee arthroplasty. Specifically, we performed the segmentation of a dataset of 200 knee CT scans, acquired on pathological patients, affected by severe osteoarthritis, who underwent total knee replacement using PSI technique. We quantified the effects of segmentation errors on contact points and resection plane directions against the reference planning based on manual segmentation performed by expert radiologists. Main contributions in this article, both from technical and clinical points of view, can be herewith outlined: 1) extensive comparison between 2D and 3D U-net architectures to demonstrate the better performance of the 3D framework; 2) specific comparison of the 3D U-net with traditional algorithms, based on semi-automatic region growing, to evaluate the ability to segment femur and tibia affected by wide morphological deformations and local spurs formed by osteophytes; 3) morphological matching quality of the PSI with reconstructed shapes, measured on the true contact areas, is accurate enough to ensure rotational alignment in agreement with the surgical planning performed by expert surgeons.

II. MATERIALS AND METHODS

A. PATIENT DATA AND PREPROCESSING

Axial CT scans of the knee, over a set of 200 patients (128 males and 72 females - 91 left against 109 right knees), acquired for planning purposes and provided in anonymized form by Medacta International SA (Castel San Pietro, Switzerland), were retrospectively available for this study [35], [36]. The patients, aged 67±10 years, reported localized knee pain, associated to osteoarthritis, and mechanical knee instability. Diagnostic imaging confirmed different degrees of cartilage defects, femoral osteophytes and shape abnormalities mainly at the condylar regions of the

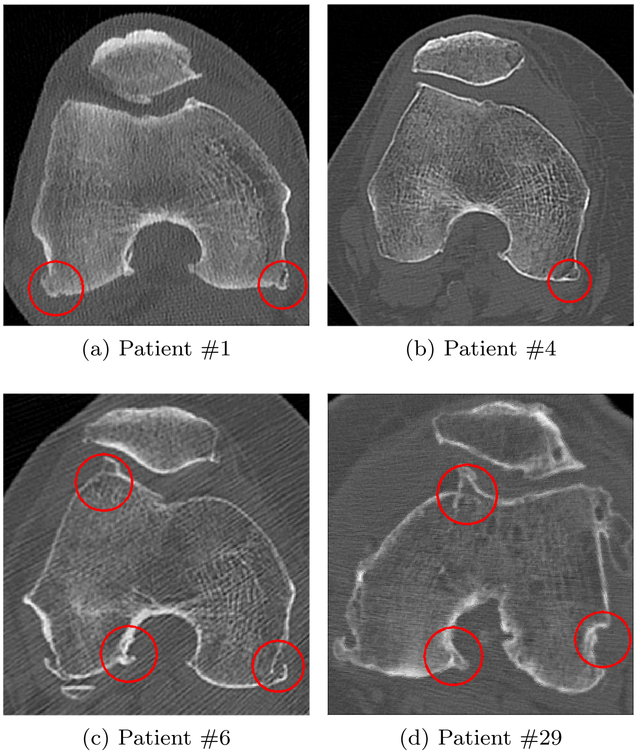


FIGURE 1. Axial slices taken from 4 patients imaging the distal femur. Deformations and osteophytes have been encircled.

distal femur and at the tibial plateau (Fig.1). CT scans were acquired with different imaging equipment, mostly at 512×512 pixels and 600 slices on average, with variable pixel size, ranging from 0.3mm to 0.4mm, and axial slicing, ranging from 0.3mm to 1.0mm, though. Along with the CT scans, the dataset encompassed distal femur and proximal tibia segmentation masks and the corresponding reconstructed surfaces. Expert radiological operators manually performed the image segmentation of the osseous portion of the bones using Mimics software (Materialize, Belgium). For increasing reliability, each dataset was segmented by one operator and revised later on by another one. Because of the imaging equipment and acquisition protocol variability, no common segmentation protocol was adopted and no data about segmentation uncertainty was available. As a function of the particular centering of the knee joint in the CT scan, the distal femur was segmented up to 2-4 cm away from frontal notch of the trochlear region along the femur shaft. Concurrently, the length of the tibia segmented shaft was variable across the patient set in a range of about 2-3 cm. Along with the bone morphology, the dataset included the corresponding planning surfaces. Planning data were available on planning surfaces as PSI contact points and contact areas, along with planar sections indicating the planned cuts. An example of the planning surfaces for one of the subjects included in the dataset is depicted in Fig.2. All the patients underwent knee replacement surgery between 2014 and 2016 using PSI technique exploiting the 3D surfaces reconstructed by the available segmentation. As originating from different scanning equipment, the CT volumetric datasets underwent preprocessing to make the pixel intensity distribution consistent and to arrange spatial dimensions to cope with network input. First, pixels belonging to filling background and air were automatically identified in the images, according to information gathered from Dicom header, and the corresponding intensity values put both to zero. The remaining image pixels underwent intensity normalization taking into account of different intensity scale encoding (e.g. Hounsfield units, 12-bit raw pixels). Then, each scan was cropped in the axial direction to remove the slices where segmentation was not available.

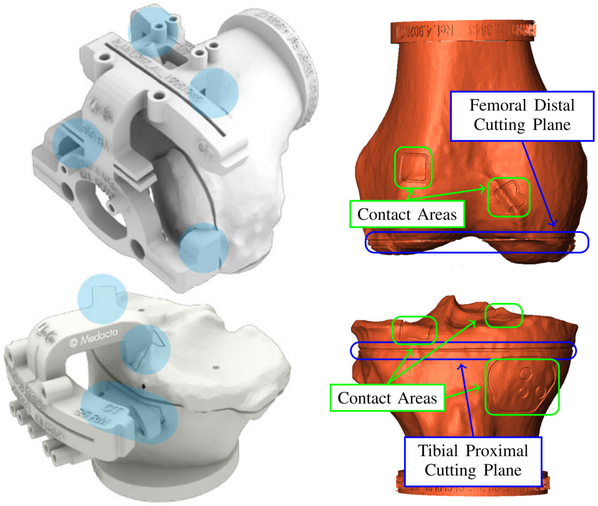


FIGURE 2. (Left panel) Femoral and tibial PSI of MyKnee system (courtesy of Medacta International Spa) with highlighted (light blue) contact points. (Right panel) Planning surfaces for patient 185 in the frontal view with contact points, areas and section planes.

B. U-NET ARCHITECTURE

The segmentation net adopted in this work took its roots from the convolutional U-Net in the 2D and 3D versions. In our implementation, 2D version of the U-Net was configured to process in input one CT axial slice and provide in output the femur and tibia segmented structures in a discrete-value image. Likewise, 3D version was configured to process in input one CT axial volume of the knee and provide in output the femur and tibia segmented structures in a 3D discrete-value volume. The U-Net mainly consists of a feed-forward architecture, which performs input encoding by means of convolutional layers, into a compressed multi-map feature representation, and then data decoding by means of deconvolutional layers. The decoding process exploits multi-scale feature fusion by concatenating the output of the encoding layer to the corresponding deconvolutional layer (cfr. Fig.3 for 3D architecture). According to the number of labels the output is configured by means of a multi-dimensional Softmax layer. In both 2D and 3D models, we envisioned each sub-module of the encoding path composed by two following convolutional layers, each one

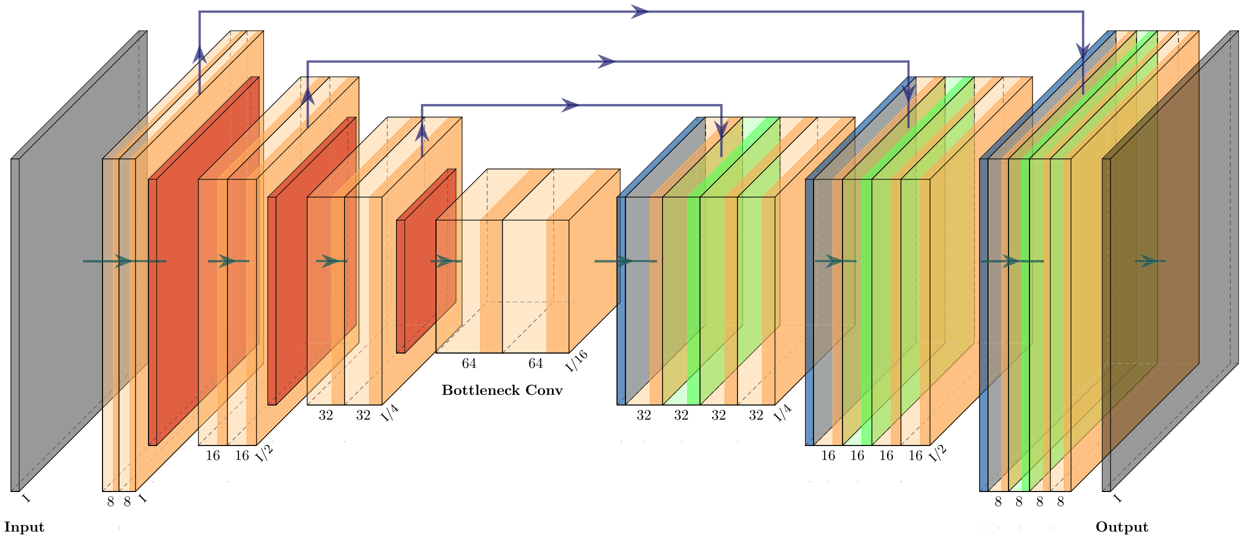


FIGURE 3. Schematic of one symmetric encoder/decoder structure, linked by a bottle-neck stage, of the U-Net. Convolutional (light red), Relu (orange), pooling (red), un-pooling (blue) and concatenating layers (light green) can be recognized in the picture. The encoding path includes three stages, with each stage embedding 2 sequential convolutions and one max pooling. In the first stage, the two convolutional layers features 8 feature maps each. The decoder stage, being symmetric to the encoder one, includes three up-sampling stages. Exactly 351435 free parameters ought to be trained for this model. By convenience this U-Net architecture was named as 8-16-32-64-32-16-8.

featuring linear activation, linked to batch normalization layer and followed by a ReLU layer (light red rectangle), and a final max pooling operator (red rectangle), ensuring a spatial compression by a factor of 2. The batch normalization layer shifted and scaled the activation distribution of the convolutional layer at each batch, by adjusting both the mean and the standard deviation of the layer activation map to optimal values during training. The bottleneck part embedded two following convolutional operators. In the decoding path, each layer featured first the up-sampling operator (blue rectangle in Fig.3), implemented by a transposed convolution with ReLU activation, whose output was then concatenated (blue arrow in Fig.3) with the corresponding output of the encoding path (light red rectangle), followed by two convolutional operators. Assuming 8 feature maps in the first stage of the encoding path and doubling the number every new stage, by convenience the resulting U-Net architecture was named as 8-16-32-64-32-16-8. The Softmax output layer was a 4D tensor featuring the volume size, the first three dimensions, and the label (background, femur, tibia) as the fourth one.

C. LOSS FUNCTIONS FOR TRAINING

The 2D (3D) network training was based on the correspondence between the dataset of CT images/volumes and the equal-sized dataset of 2D (3D) annotated masks. In our specific problem, one pixel/voxel in the annotated mask will feature alternatively background, femur or tibia labels. Considering the 3D case, the goal of training in the segmentation network is to maximize the probability of voxel membership to the corresponding label inside the volume, this attained by minimizing a proper loss function. Traditionally, for volume segmentation the voxel-wise cross-entropy loss is adopted, which aims at maximizing each estimated posterior probability that voxels belong to a specific class given the corresponding expected probability [37], [38]. As cross-entropy does not discriminate among labels, the background, which is predominant in CT images with respect to bony voxels, may easily bias the network training. Alternatively to cross-entropy, Dice similarity index and the Jaccard coefficient harness the overlap of voxels belonging to the same class between label and CT volumes [39] but the conventional formulation does not address multi-label segmentation. This issue is usually overcome by assigning a weight, proportional to the number of voxels belonging to the specific class, to each label contribution in such a way that the different frequencies of voxel for each class can be compensated and the overall loss function re-balanced. Assuming a multi-class labeling across C classes and N voxels, the loss function based on Dice can be written as:

$$\mathcal{D}(y,\hat{y}) = 1 - \sum_{i}^{C} k_i \left(\frac{2\sum^{N} y_i \cdot \hat{y}_i}{\sum^{N} y_i \cdot y_i + \sum^{N} \hat{y}_i \cdot \hat{y}_i} \right) \tag{1}$$

where y_i and $\hat{y}_i$ are respectively the true and predicted segmented volumes for the label i whose scalar product is computed over N voxels. The scalar value k_i is a coefficient weighting the contribution of each label to the loss function and it can be computed as:

$$k_i = \frac{1}{C-1}\left(1 - \frac{P_i}{N}\right) \tag{2}$$

where P_i and N are the number of voxels belonging to class i and the overall number of voxels in the volume, respectively. Similarly, the loss function based on Jaccard coefficient can be written as:

$$\mathcal{J}(y,\hat{y}) = 1 - \sum_{i}^{C} k_i\Big(\frac{\sum^N y_i \cdot \hat{y}_i}{\sum^N y_i + \sum^N \hat{y}_i + \sum^N y_i \cdot \hat{y}_i}\Big) \quad (3)$$

For weighted cross-entropy, the loss function can be written as:

$$\mathcal{CE}(y,\hat{y}) = \sum_{i}^{C} k_i(-\hat{y}_i \cdot log(y_i) + (1-\hat{y}_i)\cdot log(1-y_i)) \quad (4)$$

As far as the weights are concerned, usually they are computed on the overall training dataset but this is sub-optimal in a mini-batch paradigm, where the batch size is a fraction of the entire set. We detailed below in the next section how we addressed this issue.

D. NETWORK IMPLEMENTATION AND TRAINING STRATEGY

Implementation was performed leveraging on Tensorflow and Keras libraries in the Python environment. The code was run using the Colaboratory platform by Google Research (Google Colab, colab.research.google.com) equipped with 4-core CPU, 25GB RAM and with NVIDIA® Tesla® T4 GPU support, with 8GB RAM. For the training, out of the overall 200 samples, sorted alphabetically by family name, the first 160 samples, were used to train the U-Net model, namely by computing the neural weights, the next 20 samples to validate the training by checking the error for convergence. The remaining 20 samples were later used to test the segmentation performance. The training was based on Adam optimizer with a customized adaptive learning rate, starting from an initial value of 0.0003, which was updated according to the topology of the error metrics on the validation set. A heuristic threshold of 0.95 on the error metrics (see eq. 4) of femur and tibia was set to trigger the reduction of the learning rate by a factor of 2. The metrics used to evaluate the training performance on the validation set was the Intersection Over Union (IoU), computed as:

$$IoU_i = \frac{TP_i}{TP_i + FN_i + FP_i} \quad (5)$$

where TP_i, FN_i, and FP_i are the true positives, false negatives, and false positives, respectively, for class i. If the computed average error metrics remained stable (lower than 5%) for 10 consecutive epochs, the training was considered completed and the network was stored for later analysis. In order to prevent early convergence, a predefined number of epochs, set to 100, was allowed. A batch strategy in the training was selected, with a batch size in agreement with memory constraints. In order to ensure data mixing in the batches, data reshuffling was implemented in each epoch of the training. This required to compute at run-time for the specific batch the weights k_i (see eq. 1). In order to cope with the computational limits of the Colab environment that prevented to process original size data Dx, Dy, Dz, two main strategies were put in practice ensuring a reasonable trade-off among data size, network complexity and available graphics processing unit (GPU) memory. The first approach implied spatial sub-sampling while the second one required the implementation of an efficient data loading in memory. Sub-sampling was performed by generating four datasets, featuring increasing resolutions (Table 1), which also allowed to study the sensitivity of the network to spatial resolution.

TABLE 1. Datasets with different dimensions (number of pixels and slices). The corresponding voxel size (mm) are average values across the sub-sampled dataset.

Name	Dx	Dy	Dz	Δx	Δy	Δz
D1	128	128	128	1.50	1.50	1.00
D2	160	160	128	1.25	1.25	1.00
D3	184	184	160	1.10	1.10	0.80
D4	200	200	192	1.00	1.00	0.70

Efficient data loading in memory was implemented by data-generator strategy allowing to load only the batch of data required at run-time during each epoch of the optimization. The last relevant feature of the implementation was the adoption of a 5 dimensional tensor representation of data to embed the batch dimension, the volume size, and the label. As an example, considering a batch size of 4, a volume of 160 × 160×128 and the three labels, namely background, femur and tibia, the size of the 5-dimension tensor will be 4 × 160×160 × 128×3. This will be therefore the data size used in the loss function computation during the training stage.

E. SEGMENTATION ASSESSMENT

In order to test the segmentation quality and the reconstruction accuracy across different dataset spatial resolutions and network architectures, the last 20 samples, out of the 200 ones, which were not included in the training set were considered. Dice index (D), sensitivity (Se) and positive predictive value (PPV) for each segmented bone were computed and compared. According to (eq. 1), the Dice index for class bone i can be written as:

$$D_i = \frac{2TP_i}{(2TP_i + FP_i + FN_i)} \quad (6)$$

Sensitivity measures the portion of bone voxels in the labeled volume being correctly identified as bone voxels by the automatic segmentation and can be computed as:

$$Se_i = \frac{TP_i}{(TP_i + FN_i)} \quad (7)$$

Positive predictive value, also known as precision, is expressed by the proportion of correctly identified bone voxels that are true positive results and can be computed as:

$$PPV_i = \frac{TP_i}{(TP_i + FP_i)} \quad (8)$$

In addition to the above metrics, the computed labeled masks were reconstructed in 3D to obtain the corresponding surfaces $\hat{S}$. Prior to reconstruction, the masks underwent image processing to remove undue voxel spots by means of the opening morphological operator. The reconstruction accuracy, against the reference surface S, was measured by root mean squared d_R distances. Due to the relatively small sample size, non-parametric statistical significance tests were used to compare results across different conditions. Statistically significant effects were assessed at $p < 0.01$.

F. CLINICAL EVALUATION

The quality of femoral and tibial segmentation was evaluated in terms of clinical impact on the surgical planning in the total knee replacement based on MyKnee technology (Medacta International Spa, Castel San Pietro, Switzerland). Practically, the reconstructed surfaces of the test set were matched to the corresponding planning surfaces (cfr. Fig.2). For each bone, the matching was quantified in terms of distance errors at the contact areas of the PSI with the surface and angular alignment errors of the distal femoral and proximal tibial cutting planes, representing the surgical resections. On each planning surface, contact areas were sampled by picking either three or four technical landmarks each, at the vertices of the areas (see supplementary materials), using Amira software suite (Thermo Fisher Scientific Inc., Waltham, MA USA). For each landmark, the corresponding point on the reconstructed surface was determined by minimal distance. The contact area matching was computed by averaging the four distance errors. On the distal femur, two contact areas in the frontal part and two contact areas in the distal part were taken into account. On the proximal tibia, three contact areas were deemed, namely on medial and lateral tibial plateau region and one on the frontal region close to the tibial tuberosity. In order to define the femur distal resection plane, four landmarks were picked in correspondence of the planned resection sulcus, two frontally and two posteriorly. The resection plane of the tibial plateau was identified by four landmarks picked on frontal, lateral, medial and posterior aspects, in correspondence of the planned resection sulcus, respectively (cfr. Fig.2). Again, the minimal distance was used to determine the corresponding points on the femur and tibia reconstructed surface. For each bone, the normal direction of the plane fit to the four points was computed in the planning and reconstructed surfaces and the in-between angular deviation was projected on both frontal and sagittal anatomical planes, obtaining two clinically relevant measures [1], [12], [40].

III. EXPERIMENTAL TEST AND RESULTS

A. U-NET SETUP

1) 2D AND 3D COMPARISON

The first analysis carried out on the U-Net was designed to compare the performance achieved by the network, in its 3D implementation, to the results obtained with different implementations of a 2D U-Net on the same data. The used U-Net architecture included three main processing blocks for both encoding and decoding paths with the following feature map configuration ($8-16-32-64-32-16-8$) (Fig.3). All the convolutional and deconvolutional kernels had $3\times3\times3$ and $2\times2\times2$ size, respectively, leading to exactly 351435 free parameters. As far as 2D U-Net architecture was considered, three, four, five, and six processing layers in both the encoding and decoding paths were taken into account. For all the different 2D architectures, the number of feature maps in the first layer was equal to 8, while the number of features maps in the bottleneck were equal to 64, 128, 256, and 512, respectively. The number of parameters of the 4 2D architecture ranged from 121 thousands (three processing layers) up to 1,948,795 (6 processing layers). In order to compare the performance of the 2D and 3D networks using the same metrics, the images produced as outputs by the 2D U-Net were ordered and combined together to produce a single volume for each patient belonging to the test set. After completing this procedure, Dice, sensitivity, and precision values were computed on the volumes as explained in Section II-E. This analysis was carried out on dataset D1, therefore the images used in the 2D U-Net had a resolution of 128×128 pixels, and the respective three dimensional volume used in the 3D U-Net had a resolution $128\times128\times128$ pixels. Dice, sensitivity, and precision results, computed on the test set for both femur and tibia, were sensibly better for the 3D U-Net that those obtained for all the 2D U-Net versions, as shown Fig.4. This analysis reported that the 3D implementation outperformed the 2D U-Net, even with a higher number of processing layers in the encoding and decoding paths. A statistically significant difference ($p < 0.01$) was found when comparing the Dice values for both femur and tibia computed on test set using the 3D U-Net and the four different implementations of the 2D U-Net.

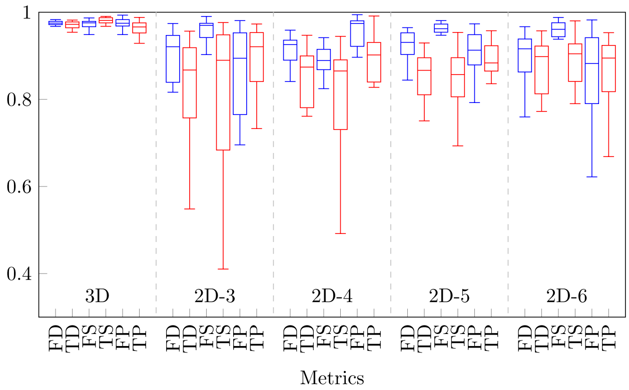


FIGURE 4. **Boxplots of dice, sensitivity, and precision values obtained on the dataset D1 using the described 3D U-Net, and four different implementations of a 2D U-Net with an increasing number of processing layers (2D-3: 3 processing layers, 2D-4: four processing layers, ...). *FD*: Femur Dice; *TD*: Tibia Dice; *FS*: Femur Sensitivity; *TS*: Tibia Sensitivity; *FP*: Femur PPV; *TP*: Tibia PPV.**

2) LOSS FUNCTION TEST

The second analysis aimed at testing the training dependency on the three different loss functions, namely Dice, Jaccard and cross-entropy indexes. The chosen U-Net architecture was identical to that one of the previous test. Dataset D1, composed of volumes with dimensions 128 × 128 × 128 was used. For all the three loss functions, the median values of Dice, Se and PPV distributions, across the 20 test patients, ranged from 0.96 to 0.98 with no statistical difference ($p > 0.1$). Dice index was therefore chosen as loss function for all subsequent experiments.

3) ABLATION TEST

The third analysis aimed at evaluating the dependency of the segmentation quality on number of convolutional layers of the U-Net. This was tested by means of an ablation test, using again dataset D1. The convolutional layers were inserted or removed symmetrically in the encoding and decoding paths maintaining the bottleneck layer. Four ($8-16-32-64-128-64-32-16-8$), three ($8-16-32-64-32-16-8$) and two ($8-16-32-16-8$) layers were taken into account. The results carried out on the network are reported in Table 2. Four, three, and two convolutional layers were taken into account for this analysis, using as input to the network dataset D1. This analysis reported that changing the number of convolutional layers produced very small changes in the Dice metric for both femur and tibia, with an overall range in the interval 0.96-0.98. No statistically significant difference ($p > 0.01$) was found when comparing the Dice values for both femur and tibia using the three different number of convolutional layers.

TABLE 2. Results of the ablation test performed on the network using dataset D1 and a different number of convolutional layers.

Layers	Dice	
	Femur	Tibia
2	0.97 (0.97-0.98)	0.97 (0.96-0.98)
3	0.97 (0.97-0.98)	0.97 (0.96-0.98)
4	0.97 (0.97-0.98)	0.97 (0.97-0.98)

4) DEPENDENCY ON FEATURE MAP SIZE AND NUMBER

In the fourth test, dataset D2, composed of volumes with dimensions 160 × 160 × 128, was considered to analyze the variability of the segmentation performance as a function of the feature map size and number in the convolutional layers. The architecture $8-16-32-64-32-16-8$ for the network was taken into account again, whereas the 3D filter size of the convolutional kernels was made variable across three different values: 3×3×3, 4×4×4 and 5×5×5. Boxplot charts of dice, sensitivity and positive predictive values obtained with the different filter sizes were summarized in Fig.5. The analysis showed that the change in the filter size elicited very small effects with an overall range of sensitivity in the interval 0.94-0.99 for both femur and tibia, and of positive predictive value in the interval 0.93-0.99. After testing the effects of filter size, two networks with two different feature map configurations, namely 8-16-32-64-32-16-8 and 12-24-48-96-48-24-12, were compared. Results attained with the two different feature maps configurations were similar, with little to none difference between them. No statistically significant difference ($p > 0.01$) was found when comparing dice, sensitivity and positive predictive values computed using the two different feature map configurations.

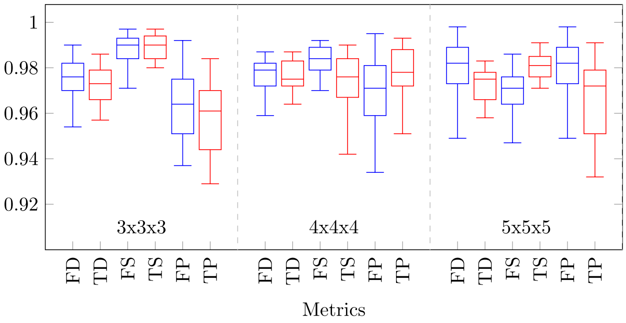


FIGURE 5. Boxplots of dice, sensitivity and positive predictive values obtained on the test set using dataset D2 with three different filter sizes. Boxplots in blue refer to the femur class, while boxplots in red refer to the tibial class. *FD*: Femur Dice; *TD*: Tibia Dice; *FS*: Femur Sensitivity; *TS*: Tibia Sensitivity; *FP*: Femur PPV; *TP*: Tibia PPV.

B. QUALITATIVE ANALYSIS FOR THE 3D U-NET

A qualitative analysis of the segmented scans was performed using the 3D U-Net (loss function: dice, configuration: 8-16-32-64-32-16-8, convolutional filter size: 3 × 3 × 3, deconvolutional filter size: 2 × 2 × 2 size) trained on dataset D2 (Table 1). The segmentation was effective in accurately labeling the distal femur and proximal tibia, excluding at the same time the other bones such as the patella and the fibula, as shown in Fig.6. Interestingly, the trained network was able to properly separate adjacent tibial and femoral surfaces also in presence of very narrow spaces in between. Bony spurs on the ridge of the trochlear region of the femur and on the tibial plateau boundary were also correctly segmented. Interestingly, condylar osteophytes, both in medial and lateral side, were again correctly segmented excluding contiguous tissues (Fig.7). In order to verify in which of the two paths of the U-Net, and at what level, the removal of non-target tissues occurred, we visually analyzed the activation maps layer by layer in the encoder and in the decoder, corresponding to the volumes input. As expected, the training specialized the encoding path as edge and boundary detector at decreasing spatial scale when increasing the path depth, while the decoding path was tailored to remove background and discriminate between target and non-target bones during progressive image up-sampling. In Fig.8, the visualization of the activation maps in the net trained on dataset D2, corresponding to the 78th slice of the volume scan #185 can be appreciated. One image in the first column represents the image output of one feature map of the 8 available (for sake of clarity only 5 of them were depicted) in the first convolutional layer in the encoding path. One image in the second column represents the image output of one feature map of the 16 available (for sake of clarity only 6 of them were depicted) in the second convolutional layer in the encoding path. One image in the

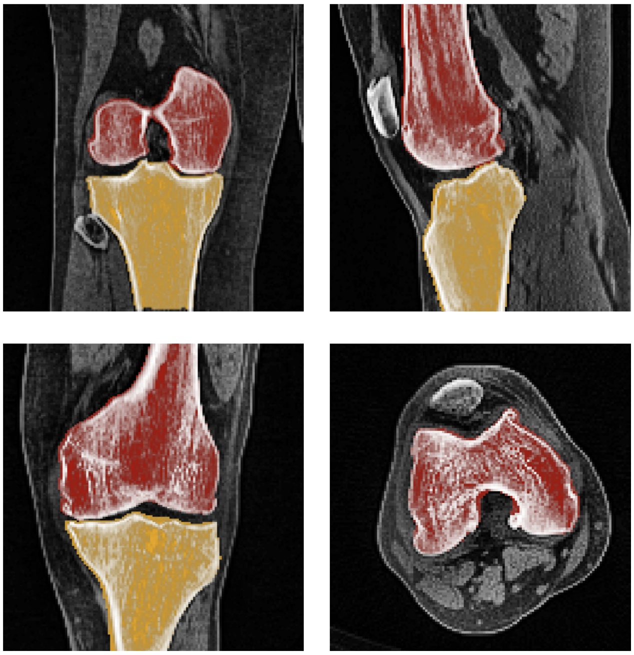


FIGURE 6. (Upper panels) Frontal and sagittal views of the segmented CT scan of patient #185. (lower panels) Frontal and axial views of the segmented CT scan of patient #196. In both frontal views, in correspondence of the narrow space between tibia (yellow) and femur (dark red), correct segmentation can be appreciated. As it can be noticed as well in the axial view of patient #196, one spur on the medial ridge of the trochlear region and one osteophyte on internal surface of the medial condyle are both properly segmented.

third column represents the image output of one feature map of the 32 available (for sake of clarity only 9 of them were depicted) in the third convolutional layer in the encoding path. One image in the fourth column represents the image output of one feature map of the 64 available (for sake of clarity only 17 of them were depicted) in the bottleneck convolutional layer. One image in the fifth layer represents the image output of one feature map of the 32 available in the first convolutional layer in the decoding path. One image in the sixth layer represents the image output of one feature map of the 16 available in the second convolutional layer in the decoding path. One image in the seventh layer represents the image output of one feature map of the 8 available in the third convolutional layer in the decoding path. The output of the last column will be the input of the Softmax classifier (see supplementary multimedia).

C. QUANTITATIVE ANALYSIS FOR THE 3D U-NET

1) SEGMENTATION ERRORS

The four datasets (Table 1) with different resolution setups were used to train the earlier U-Net architecture. Fig.9 shows the boxplot charts of Dice index distributions obtained on femur and tibia across the test set. In all datasets, D, Se and PPV median values were all greater than 0.96 for both femur and tibia segmentation (Table 3). First and third quartile of sensitivity values across the four datasets laid within the range 0.97-0.99. PPV median values for the femur ranged from a minimum of 0.96, on dataset D2, to a maximum of 0.98, on both datasets D3 and D4. PPV median for the tibia, instead, were of 0.96 on dataset D2 and 0.97 on the remaining datasets. Kruskal-Wallis test was used to assess the statistical difference between femur and tibia classes in all the three measures for each dataset. Sensitivity values between femur and tibia were statistically different, at 1% of significance, for dataset D1 and D3. Positive predictive values of femur and tibia class were statistically different only for dataset D3 ($p < 0.01$).

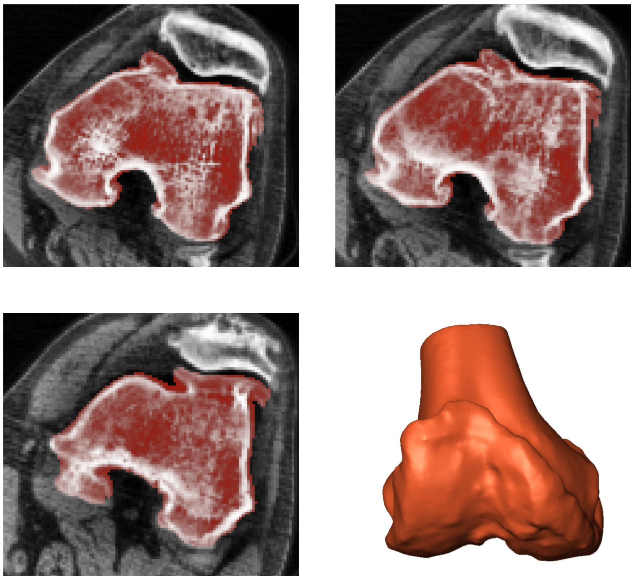


FIGURE 7. Three axial views of the segmented femur (transparent dark red) of the patient #197. From the corresponding reference surface, it can be noticed the relevant deformation of the trochlear ridge anteriorly, with the corresponding segmentation that follows correctly the osseous profile.

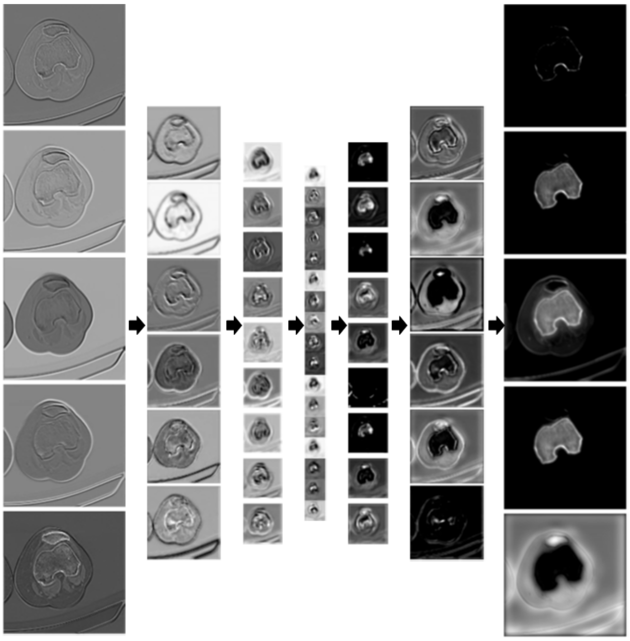


FIGURE 8. Simplified graphical representation of the processing of the 78^{th} slice, out of 128, in the U-Net (8-16-32-64-32-16-8) for the patient #185 trained on dataset D2.

TABLE 3. Summary of dice (D), sensitivity (Se) and precision (PPV) indexes obtained with the four datasets. Data shown are median (first quartile-third quartile) values.

Dataset	D		Se		PPV	
	Femur	Tibia	Femur	Tibia	Femur	Tibia
D1	0.97 (0.97-0.98)	0.97 (0.96-0.98)	0.98 (0.97-0.98)	0.98 (0.98-0.99)	0.97 (0.97-0.98)	0.97 (0.95-0.98)
D2	0.98 (0.97-0.98)	0.97 (0.96-0.98)	0.99 (0.98-0.99)	0.99 (0.98-0.99)	0.96 (0.95-0.97)	0.96 (0.94-0.97)
D3	0.98 (0.97-0.98)	0.98 (0.97-0.98)	0.98 (0.97-0.98)	0.98 (0.98-0.99)	0.98 (0.97-0.99)	0.97 (0.95-0.98)
D4	0.99 (0.98-0.99)	0.98 (0.98-0.99)	0.99 (0.99-0.99)	0.99 (0.98-0.99)	0.98 (0.97-0.99)	0.98 (0.97-0.99)

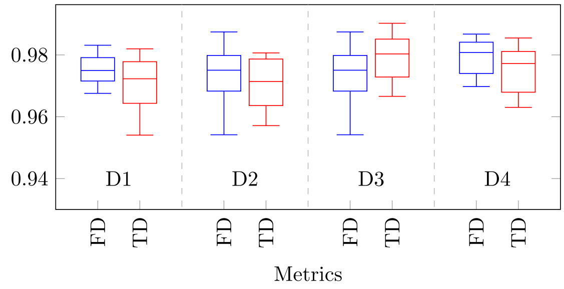


FIGURE 9. Boxplots of Dice values obtained on the test set with the four different datasets. Boxplots in blue refer to the femur class, while boxplots in red refer to the tibial class. *FD*: Femur Dice; *TD*: Tibia Dice.

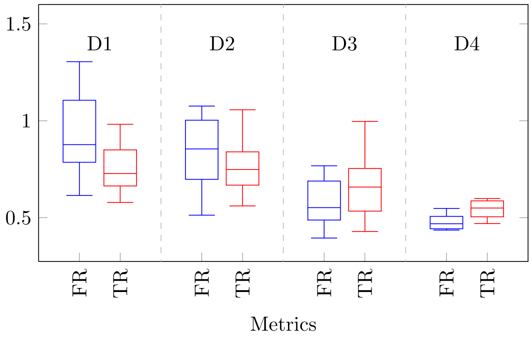


FIGURE 10. Boxplots of RMS errors (mm) obtained on the test set with the three different datasets. Boxplots in blue refer to the femur class, while boxplots in red refer to the tibial class. *FR*: Femur RMSE; *TR*: Tibia RMSE.

2) THREE-DIMENSIONAL RECONSTRUCTION ERRORS

For the models trained with the four datasets, the root mean square distance (RMS) were computed. The median values of distance distributions ranged between 0.5 and 1mm, with a maximum error lower than 1.5mm (Fig.10). It is noteworthy that in dataset D4 the values computed for the femur and tibia classes were very akin and close to 0.5mm. Furthermore, for dataset D4 the maximum IQR was lower than 0.15mm. As it was expected, a general decreasing trend for the 3D reconstruction errors was obtained as the spatial resolution increases, thus suggesting that augmenting the spatial resolution by approaching the original voxel size could still improve the reconstruction quality even up to the CT planning scans. Kruskal-Wallis test was used to assess the statistical difference in RMS distances between femur and tibia classes, and Conover post-hoc analysis with Holm correction was carried out to compare the results across the datasets. Differences statistically significant were found ($p < 0.005$) when comparing dataset D1 with datasets D3 and D4, while no differences were found ($p > 0.1$) between dataset D3 and dataset D4.

3) COMPARISON WITH SURGICAL PLANNING OBTAINED BY MANUAL SEGMENTATION

The computed errors at contact areas located on the frontal femur and tibial plateau locations were quite similar (range: 1-2mm) across the four training datasets, as it can be appreciated in Fig.11. The contact area error at condylar locations distally was conversely lower than 1mm, in all the four setups. These findings were not surprising as the distal condylar surfaces feature less local complexity than frontal and medio-lateral areas. As a matter of fact, pathological deformations tend to flatten the distal areas while condylar ridges both frontally and laterally usually undergo wider deformations. These deformation effects make the segmentation more challenging, here confirmed by greater errors. The median angular deviation of the direction of the femoral distal resection plane, projected on both frontal and sagittal anatomical planes, was close to 0° with IQR in the range of 1° in all the four setups (Fig.12). While no statistical difference was found between the angular values of frontal and sagittal planes ($p > 0.33$), a slightly greater IQR was attained for the sagittal plane. The direction of tibial plateau resection plane deviated from the corresponding planning direction less than 1.5° and no statistical difference was detected ($p > 0.07$)

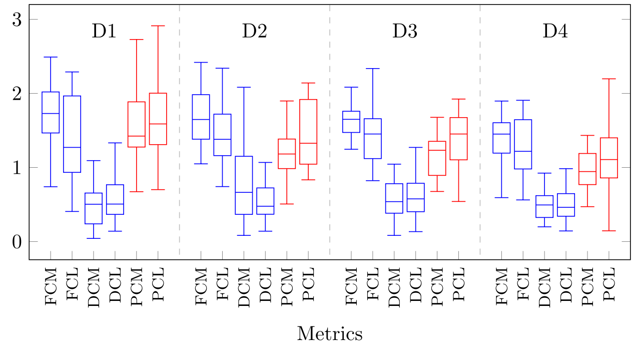


FIGURE 11. Distance error distributions (mm) computed at the contact areas for the four training setups. Blue and red boxplot charts refer to the femur and tibia, respectively. *FCM*: Femur Frontal Contact Medial; *FCL*: Femur Frontal Contact Lateral; *DCM*: Femur Distal Contact Medial; *DCL*: Femur Distal Contact Lateral; *PCM* Tibia Plataeu Contact Medial; *PCL* Tibia Plateau Contact Lateral.

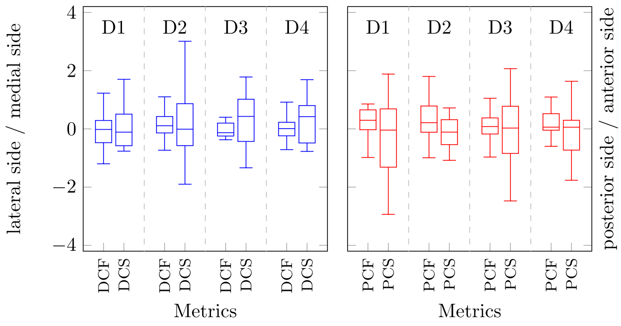


FIGURE 12. Boxplot charts of angular error (°) values obtained on the test set using the different datasets. Boxplots in blue refer to the femur class, while boxplots in red refer to the tibial class. ***DCF***: Femur Distal Cut Frontal; ***DCS***: Femur Distal Cut Sagittal; ***PCF***: Tibia Plateau Cut Frontal; ***PCS***: Tibia Plateau Cut Sagittal. For the frontal plane, values greater and less than 0° indicate medial and lateral side, respectively. For the sagittal plane, values greater and less than 0° indicate anterior and posterior side, respectively.

between the angular values of frontal and sagittal planes. Again, the alignment in the sagittal plane was more uncertain than that in the frontal plane.

4) 3D U-NET AGAINST REGION GROWING ALGORITHM

The segmentation performance of the 3D U-Net, trained using dataset D4, was compared with the "Fast GrowCut" implementation of the region growing algorithm [41], available into 3D Slicer (www.slicer.org). The test set of the dataset D4 was imported into 3D Slicer to semi-automatically perform the segmentation. Segmented images and reconstructed surfaces (femur and tibia) were considered to compute DICE and 3D RMSE, with respect to the original reference segmentation and 3D surfaces. The DICE results, reported in Table 4, demonstrated a better performance of the U-Net, for both tibia and femur, supported by a significant statistical difference ($p < 0.01$). Similarly, 3D RMSE results were in favour of the U-Net with a significant statistical difference ($p < 0.01$).

TABLE 4. Comparison between 3D U-Net, trained on dataset D4, and region growing algorithm in terms of median (first quartile-third quartile) Dice values and 3D RMS errors (mm), computed on the test set.

Metric	Class	Algorithm	
		U-Net	Region Growing
Dice	Femur	0.99 (0.98 - 0.99)	0.95 (0.95 - 0.95)
	Tibia	0.98 (0.98 - 0.99)	0.94 (0.93 - 0.94)
3D RMSE (mm)	Femur	0.44 (0.38 - 0.48)	0.96 (0.92 - 1.08)
	Tibia	0.57 (0.48 - 0.90)	1.08 (1.04 - 1.19)

IV. DISCUSSION

A. MAIN FINDINGS

Many recent research articles extensively described the use of U-Net for medical image segmentation and the majority confirmed the superiority of the 3D architecture, with respect to the traditional 2D, for the processing of 3D scans. It was indeed shown that 3D convolution allows for the directly modeling of the spatial connectivity of the target anatomical regions during training [25], [42]–[44]. The results of the first analysis performed in this work supported such previous findings (cfr. Fig.4). Using 3D U-Net, the binary segmentation of bones in CT and MRI images was demonstrated feasible with high accuracy [29], [30] while the feasibility of multi-region labeling in a semantic segmentation approach and the role of pathological deformations of bones were not systematically addressed in the literature. Effects of segmentation quality on image-based surgical planning had not been tested so far. In order to deal with such challenges, in the present paper we adopted the 3D U-Net paradigm to address the semantic segmentation of CT images of the knee to extract concurrently femur and tibia regions, affected by severe pathological conditions, and evaluate how segmentation errors might have impacted on PSI-based surgical planning. While retrospective, the extensive dataset of 200 samples allowed for evaluating the extrapolation properties of the training. The obtained results on the test set confirmed that both femur and tibia regions were successfully segmented with high accuracy featuring both PPV and sensitivity greater than 96%. Such quality was corroborated by analyzing the 3D reconstruction errors with the median value in the range of 1mm. The high quality of the 3D matching between the reconstructed surfaces and the contact areas of the virtual resection guides, represented by the planning surfaces, was proved by a maximum median error lower than 2mm. Likewise, very low angular deviations (2°) of the resection planes further supported the achieved segmentation quality and the overall implemented methodology.

B. COMPARISON WITH THE LITERATURE

Binary segmentation of bones in CT and MRI images exploiting neural networks and DL was already demonstrated as feasible. Segmentation of spine in CT scans using a deep CNN provided sensitivity of 97% and 3D surface distance error of 7.4mm [27]. However, the study was performed on a small dataset of 32 patients that reduced the span of the results. Accuracy of about 97% for femur segmentation in CT images was recently reported [30]. While including a dataset of 150 patients, the work focused only of the one single bone and performed the study using only a low resolution version of the original dataset with an inter-slice resolution (3mm) smaller than twice as many the resolution used in this work. This can raise some doubts about the quality of the 3D reconstruction attainable on the basis of such a segmentation. Bone segmentation in dual energy CT scans by means of U-Net architecture, applied to 15 patients, featured Dice index of about 96%. [45]. However, the study performed binary segmentation only and used a very low volume resolution. The U-Net architecture was used in bone segmentation in 53 low-quality low-dose whole-body CT scans leading dice score of 95% [34]. However, the generalization of the results was reduced as the dataset was acquired with a unique scanner. In our work, we used conversely volumes acquired with four different scanners, namely Philips, Canon Medical Systems, GE Medical Systems and Toshiba. CNNs were applied to segment the skull in 20 CT scans for treatment planning applications achieving 92% of sensitivity and surface reconstruction errors in the range of 1.5mm [46]. A specific study exploiting a lightweight U-Net for hand bone segmentation in X-ray images reported 94% of sensitivity [32]. The results of the present work basically agreed also with such results, even obtained on different bones. 2D U-Net was applied to multi-label segmentation of 12 different structures in knee joint by processing 20 MRI scans achieving a mean Dice index for femur and tibia of about 90% [26], again in agreement with our results. Overall, we can assert that the achieved results are basically in line with the present literature, arguing also that their scope can be regarded to a wider clinically extent because of the heterogeneity of pathological severity into the dataset and the proof of the results quality into the PSI-based surgical planning application.

C. TECHNICAL CHALLENGES

It is well known that 3D CNNs are computationally more demanding than 2D CNNs and can lead to higher overfitting due to the increased number of trainable weights. In this article, we tackled the first issue by down-sampling the original data into four different dataset sizes that allowed us to study the potential sensitivity of the segmentation to pixel size and inter-slice distance using non-isotropic voxels. From Dice, sensitivity, and positive predictive values, we concluded that the segmentation performance of the corresponding four trained networks was very similar. Conversely but expected, the analysis of the 3D accuracy demonstrated that the errors on the reconstructed surfaces were decreasing with the increase of voxel resolution. In order to address the second issue, we leveraged the validation error to drive the training stop to successfully overcome potential overfitting. This led to get very similar IoU metrics in the training and validation datasets. As far as dataset extent is concerned, no data augmentation was necessary. Usually such an approach, implemented by rotation and translation of the original samples, is suggested to increase the sample size and enhance the spatial variability of the target regions in the images. However, in this work we deemed the available samples sufficient to constrain the training, endorsed by both labeling and reconstruction results. The spatial invariance achieved by the network was further proved by the ability of the network to generalize the true learnt semantic segmentation to the labeling of both left and right regions into a single 3D scan (Fig.13). Finally, regarding computational demands, each epoch during the training of the greater resolution dataset (D4), took on average about 4 minutes and the overall training took approximately 12h. Completing one fully automatic segmentation with the trained model took about 3s. Conversely, the region growing algorithm took approximately 45 minutes to complete a segmentation, requiring accurate initialization of seeds performed manually.

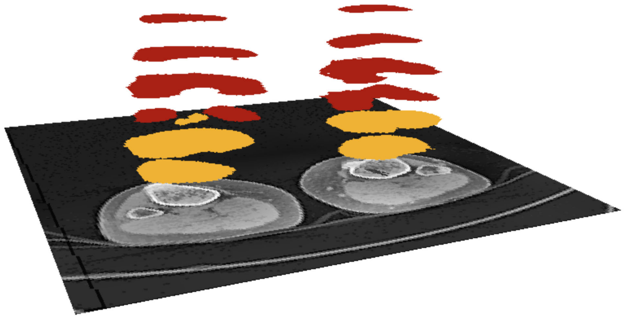


FIGURE 13. Segmentation using the trained U-Net (dataset D4) on a 3D scan imagining the two knees of patient #198. As expected, the network is performing a true semantic segmentation labeling at the same time the left and right shapes of tibia and femur.

D. CLINICAL CHALLENGES

As described, the matching of the femoral resection component of MyKnee PSI mainly depends on the reconstruction precision of both frontal and distal aspects of the distal femur. For the tibia, the frontal aspect of the tibial plateau mainly affects the matching with the tibial resection component. Nominally, the final coupling tolerance should reflect the combination of uncertainty in scanning resolution with accuracy of image segmentation, surface reconstruction, surface smoothing, digital representation of the surface in the production system and finally precision of manufacture. While it is not unequivocal to exactly quantify the error chain in each step, we give a feasible hypothesis of the final matching error between the PSI and the true patient anatomy. The scanning resolution, namely the voxel size, which is typically 0.5 × 0.5×0.5mm for both planning CT and MRI, is the first source of uncertainty. When carefully performed, the segmentation quality, which however depends on the image modality (in the MRI the bone-soft tissue boundary is less contrasted than that in the CT) and on the operator performing the task, introduces additional errors that, in the best conditions, are at sub-voxel size scale. It is well known however that the high quality segmentation process is very time consuming and prone to errors, especially with severe clinical conditions, worsening the manual segmentation. The surface reconstruction quality, which strongly depends on slicing thickness, and the surface smoothing further increase the difference between the true and the digital patient bones. Assuming that the virtual matching in the PSI design is perfect, no additional errors should be considered at this stage. Contrary, the prototyping process can decrease further the matching accuracy due to the internal representation of the surface data and the manufacturing precision. For instance, modern 3D printers ensure a precision of at least 0.1 mm. In conclusion, it is realistic to assume that, in the best condition, the uncertainty of the coupling is in the range of 1 mm. It was reported that such an uncertainty can induce rotational differences in the coronal and sagittal planes, between the planned and intra-operative alignment, in the range of about 2° [5], [47]. We showed that it is possible to attain a reconstruction accuracy in terms of RMSE less than 2.0 mm and 1.5 mm, for both tibia and femur, results that could be further improved by processing datasets at higher resolution. Such 3D errors led to deviations of the femoral distal cut direction in the frontal plane less than ±1°, being akin to alignment errors (±3°) reported in the recent literature [13], [17], [48], [49] and recognized in the range of tolerable surgical errors, demonstrating the clinical potential of the proposed segmentation approach. However, the slightly greater deviations in the sagittal plane should be carefully taken into consideration, as the corresponding mal-alignment of the mechanical axis of the femur modifies the tibio-femoral extension gap, which in turn affects the patellofemoral joint kinematics and ligament balancing [49]. Similarly, the tibial angular error in the sagittal plane, while being lower than ±2°, must be carefully evaluated as it may lead to prosthesis impingement issues in the knee flexion [8], [50].

In general, Dice, sensitivity and PPV are assumed predictive indices of the 3D reconstruction error. However, the segmentation quality on the overall CT scan should be carefully examined and not immediately considered representative of local areas of the bony regions, especially those ones corresponding to the PSI contacts (Fig.14). As a result, small variations of the segmentation quality might correspond to sensible error variation even greater than 1 mm as shown for the tibial contact areas. In addition, morphological deformations are heterogeneously distributed across the overall bony shape (see the difference between distal and frontal contacts in the femur) making in both cases the segmentation quality alone not enough to ensure accurate matching between the PSI and the reconstructed surface. In synthesis, improvements in the segmentation in terms of Dice, Se or PPV does not necessarily mean an improvement of the clinical value of segmentation. As a final remark, we highlight that the role of the PSI surgical technique has been topic of many debates with no general consensus about the accuracy and reliability for a large-scale surgery [12], [50], [51]. Nonetheless, two recent meta-analysis studies, comparing PSI-based interventions to traditional surgery using invasive instruments on approximately 5000 patients, reported significant differences with regards to operative time and blood loss in favor of PSI [52], [53]. In addition, knee surgery based on PSI has been very recently reported to improve functional kinematics with respect to traditional surgery [8]. In general, while PSI cannot be regarded as the gold standard in total knee replacement, advanced osteoarthritis conditions can be surgically addressed through such a technique, especially in conditions of bone deformity, which can prevent the use of intra-medullar bars [54].

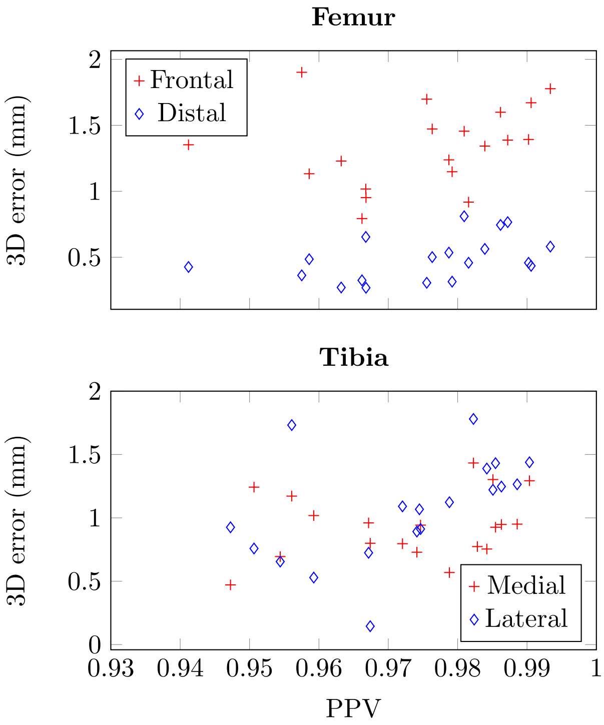


FIGURE 14. Relation between PPV and contact errors (mm) of the PSI for both femur and tibia (D4 dataset).

V. CONCLUSION

The developed deep convolutional network was trained and validated to concurrently segment distal femur and proximal tibia, exhibiting severe pathological deformations, in knee CT volumes acquired throughout different scanners. Interestingly, within the U-Net paradigm, encoder and decoder were proved suitable for filtering and semantic reconstruction, respectively, resembling processing of brain neural pathways, and the network exhibited reasonable generalization capabilities when exposed to both knees within the same CT volume, as expected from the convolutional spatial invariance. Specific results were shown valuable in terms of segmentation, surface reconstruction and surgical planning performances, being comparable to results obtainable by means of expert segmentation. We can argue therefore that fostering deep CNN in clinical tools may offer the opportunity of removing the current prohibitive barriers of time and effort during CT image segmentation, assuring high accuracy and making PSI-based knee arthroplasty closer at hand.

ACKNOWLEDGMENT

The authors would like to thank Medacta International SA for providing patient data.

REFERENCES

[1] M. Pietsch, O. Djahani, M. Hochegger, F. Plattner, and S. Hofmann, "Patient-specific total knee arthroplasty: The importance of planning by the surgeon," *Knee Surg., Sports Traumatology, Arthroscopy*, vol. 21, no. 10, pp. 2220–2226, Oct. 2013.

[2] O. Cartiaux, L. Paul, B. G. Francq, X. Banse, and P.-L. Docquier, "Improved accuracy with 3D planning and patient-specific instruments during simulated pelvic bone tumor surgery," *Ann. Biomed. Eng.*, vol. 42, no. 1, pp. 205–213, Jan. 2014.

[3] L. Mattei, P. Pellegrino, M. Calò, A. Bistolfi, and F. Castoldi, "Patient specific instrumentation in total knee arthroplasty: A state of the art," *Ann. Transl. Med.*, vol. 4, p. 126, Apr. 2016.

[4] J. Vide, T. P. Freitas, A. Ramos, H. Cruz, and J. P. Sousa, "Patient-specific instrumentation in total knee arthroplasty: Simpler, faster and more accurate than standard instrumentation—A randomized controlled trial," *Knee Surg., Sports Traumatology, Arthroscopy*, vol. 25, no. 8, pp. 2616–2621, Aug. 2017.

[5] A. Alvand, T. Khan, C. Jenkins, J. L. Rees, W. F. Jackson, C. A. F. Dodd, D. W. Murray, and A. J. Price, "The impact of patient-specific instrumentation on unicompartmental knee arthroplasty: A prospective randomised controlled study," *Knee Surg., Sports Traumatology, Arthroscopy*, vol. 26, no. 6, pp. 1662–1670, Jun. 2018.

[6] I. Nizam and A. V. Batra, "Accuracy of bone resection in total knee arthroplasty using CT assisted-3D printed patient specific cutting guides," *SICOT-J*, vol. 4, p. 29, Apr. 2018.

[7] E. Rex, C. Gaudelli, E. Illical, J. Person, K. Arlt, B. Wylant, and C. Anglin, "Guiding device for the patellar cut in total knee arthroplasty: Design and validation," *Bioengineering*, vol. 5, no. 2, p. 38, May 2018.

[8] K.-S. Shih, C.-C. Lin, H.-L. Lu, Y.-C. Fu, C.-K. Lin, S.-Y. Li, and T.-W. Lu, "Patient-specific instrumentation improves functional kinematics of minimally-invasive total knee replacements as revealed by computerized 3D fluoroscopy," *Comput. Methods Programs Biomed.*, vol. 188, May 2020, Art. no. 105250.

[9] P. Cerveri, A. Manzotti, N. Confalonieri, and G. Baroni, "Automating the design of resection guides specific to patient anatomy in knee replacement surgery by enhanced 3D curvature and surface modeling of distal femur shape models," *Computerized Med. Imag. Graph.*, vol. 38, no. 8, pp. 664–674, Dec. 2014.

[10] Y. D. Levy, V. V. G. An, C. J. W. Shean, F. R. Groen, P. M. Walker, and W. J. M. Bruce, "The accuracy of bony resection from patient-specific guides during total knee arthroplasty," *Knee Surg., Sports Traumatology, Arthroscopy*, vol. 25, no. 6, pp. 1678–1685, Jun. 2017.

[11] G. C. Wernecke, S. Taylor, P. Wernecke, S. J. MacDessi, and D. B. Chen, "Resection accuracy of patient-specific cutting guides in total knee replacement," *ANZ J. Surg.*, vol. 87, no. 11, pp. 921–924, Nov. 2017.

[12] S. Gong, W. Xu, R. Wang, Z. Wang, B. Wang, L. Han, and G. Chen, "Patient-specific instrumentation improved axial alignment of the femoral component, operative time and perioperative blood loss after total knee arthroplasty," *Knee Surg., Sports Traumatology, Arthroscopy*, vol. 27, no. 4, pp. 1083–1095, Apr. 2019.

[13] W. Anderl, L. Pauzenberger, R. Kölblinger, G. Kiesselbach, G. Brandl, B. Laky, B. Kriegleder, P. Heuberer, and E. Schwameis, "Patient-specific instrumentation improved mechanical alignment, while early clinical outcome was comparable to conventional instrumentation in TKA," *Knee Surg., Sports Traumatology, Arthroscopy*, vol. 24, no. 1, pp. 102–111, Jan. 2016.

[14] T. Ogura, K. Le, G. Merkely, T. Bryant, and T. Minas, "A high level of satisfaction after bicompartmental individualized knee arthroplasty with patient-specific implants and instruments," *Knee Surg., Sports Traumatology, Arthroscopy*, vol. 27, no. 5, pp. 1487–1496, May 2019.

[15] S. Lee, J. Y. Kim, J. Hong, S. H. Baek, and S. Y. Kim, "Ct-based navigation system using a patient-specific instrument for femoral component positioning: An experimental *in vitro* study with a sawbone model," *Yonsei Med. J.*, vol. 59, pp. 769–780, Aug. 2018.

[16] T. Ogawa, M. Takao, T. Sakai, and N. Sugano, "Factors related to disagreement in implant size between preoperative CT-based planning and the actual implants used intraoperatively for total hip arthroplasty," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 4, pp. 551–562, Apr. 2018.

[17] M. Miura, S. Hagiwara, J. Nakamura, Y. Wako, Y. Kawarai, and S. Ohtori, "Interobserver and intraobserver reliability of computed tomography-based three-dimensional preoperative planning for primary total knee arthroplasty," *The J. Arthroplasty*, vol. 33, pp. 1572–1578, May 2018.

[18] L. I. Wang, M. Greenspan, and R. Ellis, "Validation of bone segmentation and improved 3-D registration using contour coherency in CT data," *IEEE Trans. Med. Imag.*, vol. 25, no. 3, pp. 324–334, Mar. 2006.

[19] J. G. Tamez-Pena, J. Farber, P. C. Gonzalez, E. Schreyer, E. Schneider, and S. Totterman, "Unsupervised segmentation and quantification of anatomical knee features: Data from the osteoarthritis initiative," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 4, pp. 1177–1186, Apr. 2012.

[20] D. Wu, M. Sofka, N. Birkbeck, and S. K. Zhou, "Segmentation of multiple knee bones from CT for orthopedic knee surgery planning," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 8673, P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds. Cham, Switzerland: Springer, 2014, doi: 10.1007/978-3-319-10404-1_47.

[21] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," *Med. Image Anal.*, vol. 13, no. 4, pp. 543–563, Aug. 2009.

[22] J. Zhang, C.-H. Yan, C.-K. Chui, and S.-H. Ong, "Fast segmentation of bone in CT images using 3D adaptive thresholding," *Comput. Biol. Med.*, vol. 40, no. 2, pp. 231–236, Feb. 2010.

[23] C. Lindner, S. Thiagarajah, J. Wilkinson, T. Consortium, G. Wallis, and T. Cootes, "Fully automatic segmentation of the proximal femur using random forest regression voting," *IEEE Trans. Med. Imag.*, vol. 32, no. 8, pp. 1462–1472, Aug. 2013.

[24] J. T. Lynch, M. T. Y. Schneider, D. M. Perriman, J. M. Scarvell, M. R. Pickering, M. Asikuzzaman, C. R. Galvin, T. F. Besier, and P. N. Smith, "Statistical shape modelling reveals large and distinct subchondral bony differences in osteoarthritic knees," *J. Biomechanics*, vol. 93, pp. 177–184, Aug. 2019.

[25] G. Zeng, X. Yang, J. Li, L. Yu, P. A. Heng, and G. Zheng, "3D U-net with multi-level deep supervision: Fully automatic segmentation of proximal femur in 3D MR images," in *Machine Learning in Medical Imaging. MLMI* (Lecture Notes in Computer Science), vol. 10541, Q. Wang, Y. Shi, H. I. Suk, and K. Suzuki, Eds. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-67389-9_32.

[26] Z. Zhou, G. Zhao, R. Kijowski, and F. Liu, "Deep convolutional neural network for segmentation of knee joint anatomy," *Magn. Reson. Med.*, vol. 80, no. 6, pp. 2759–2770, Dec. 2018.

[27] M. Vania, D. Mureja, and D. Lee, "Automatic spine segmentation from CT images using convolutional neural network via redundant generation of class labels," *J. Comput. Des. Eng.*, vol. 6, no. 2, pp. 224–232, Apr. 2019.

[28] P. P. Brahma, D. Wu, and Y. She, "Why deep learning works: A manifold disentanglement perspective," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 1997–2008, Oct. 2016.

[29] C. M. Deniz, S. Xiang, R. S. Hallyburton, A. Welbeck, J. S. Babb, S. Honig, K. Cho, and G. Chang, "Segmentation of the proximal femur from MR images using deep convolutional neural networks," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 16485.

[30] F. Chen, J. Liu, Z. Zhao, M. Zhu, and H. Liao, "Three-dimensional feature-enhanced network for automatic femur segmentation," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 1, pp. 243–252, Jan. 2019.

[31] J. Leng, Y. Liu, T. Zhang, P. Quan, and Z. Cui, "Context-aware U-Net for biomedical image segmentation," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 2535–2538.

[32] L. Ding, K. Zhao, X. Zhang, X. Wang, and J. Zhang, "A lightweight U-Net architecture multi-scale convolutional network for pediatric hand bone segmentation in X-Ray image," *IEEE Access*, vol. 7, pp. 68436–68445, 2019.

[33] B. Qiu, J. Guo, J. Kraeima, H. H. Glas, R. J. H. Borra, M. J. H. Witjes, and P. M. A. van Ooijen, "Automatic segmentation of the mandible from computed tomography scans for 3D virtual surgical planning using the convolutional neural network," *Phys. Med. Biol.*, vol. 64, no. 17, Sep. 2019, Art. no. 175020.

[34] A. Klein, J. Warszawski, J. Hillengaß, and K. H. Maier-Hein, "Automatic bone segmentation in whole-body CT images," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 1, pp. 21–29, Jan. 2019.

[35] P. Cerveri, C. Sacco, G. Olgiati, A. Manzotti, and G. Baroni, "2D/3D reconstruction of the distal femur using statistical shape models addressing personalized surgical instruments in knee arthroplasty: A feasibility analysis," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 13, no. 4, Dec. 2017, Art. no. e1823.

[36] P. Cerveri, A. Belfatto, G. Baroni, and A. Manzotti, "Stacked sparse autoencoder networks and statistical shape models for automatic staging of distal femur trochlear dysplasia," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 14, no. 6, p. e1947, Dec. 2018.

[37] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, "AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Med. Phys.*, vol. 46, no. 2, pp. 576–589, Feb. 2019.

[38] G. Zeng, Q. Wang, T. Lerch, F. Schmaranzer, M. Tannast, K. Siebenrock, and G. Zheng, "Latent3DU-net: Multi-level latent shape space constrained 3D U-net for automatic segmentation of the proximal femur from radial MRI of the hip," in *Machine Learning in Medical Imaging*, vol. 11046, Y. Shi, H.-I. Suk, and M. Liu, Eds. Cham, Switzerland: Springer, 2018, pp. 188–196. Accessed: Feb. 1, 2019. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00919-9_22

[39] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells III, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index1: Scientific reports," *Academic Radiol.*, vol. 11, no. 2, pp. 178–189, 2004.

[40] P. Cerveri, M. Marchente, W. Bartels, K. Corten, J.-P. Simon, and A. Manzotti, "Towards automatic computer-aided knee surgery by innovative methods for processing the femur surface model," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 6, no. 3, pp. 350–361, Sep. 2010.

[41] L. Zhu, I. Kolesov, Y. Gao, R. Kikinis, and A. Tannenbaum, "An effective interactive medical image segmentation method using fast growcut," in *Proc. MICCAI Workshop Interact. Med. Image Comput.*, vol. 17, 2014, pp. 1–9.

[42] B. Norman, V. Pedoia, and S. Majumdar, "Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry," *Radiology*, vol. 288, no. 1, pp. 177–185, Jul. 2018.

[43] G. Zeng and G. Zheng, "Deep learning-based automatic segmentation of the proximal femur from MR images," *Adv. Exp. Med. Biol.*, vol. 1093, pp. 73–79, 2018, doi: 10.1007/978-981-13-1396-7_6.

[44] Y.-J. Huang, Q. Dou, Z.-X. Wang, L.-Z. Liu, Y. Jin, C.-F. Li, L. Wang, H. Chen, and R.-H. Xu, "3D RoI-aware U-Net for accurate and efficient colorectal tumor segmentation," 2018, *arXiv:1806.10342*. [Online]. Available: http://arxiv.org/abs/1806.10342

[45] J. C. González Sánchez, M. Magnusson, M. Sandborg, Å. C. Tedgren, and A. Malusek, "Segmentation of bones in medical dual-energy computed tomography volumes using the 3D U-Net," *Phys. Medica*, vol. 69, pp. 241–247, Jan. 2020.

[46] J. Minnema, M. van Eijnatten, W. Kouw, F. Diblen, A. Mendrik, and J. Wolff, "CT image segmentation of bone for medical additive manufacturing using a convolutional neural network," *Comput. Biol. Med.*, vol. 103, pp. 130–139, Dec. 2018.

[47] V. J. León-Muñoz, F. Martínez-Martínez, M. López-López, and F. Santonja-Medina, "Patient-specific instrumentation in total knee arthroplasty," *Expert Rev. Med. Devices*, vol. 16, pp. 555–567, Jul. 2019.

[48] T. J. Heyse and C. O. Tibesku, "Improved femoral component rotation in TKA using patient-specific instrumentation," *Knee*, vol. 21, no. 1, pp. 268–271, Jan. 2014.

[49] A. Mannan and T. O. Smith, "Favourable rotational alignment outcomes in PSI knee arthroplasty: A level 1 systematic review and meta-analysis," *Knee*, vol. 23, no. 2, pp. 186–190, Mar. 2016.

[50] H. C. Gemalmaz, K. Sarıyılmaz, O. Ozkunt, M. Sungur, I. Kaya, and F. Dikici, "Postoperative mechanical alignment analysis of total knee replacement patients operated with 3D printed patient specific instruments: A prospective cohort study," *Acta Orthopaedica et Traumatologica Turcica*, vol. 53, no. 5, pp. 323–328, Sep. 2019.

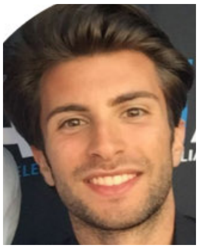
[51] N. M. Kosse, P. J. C. Heesterbeek, J. J. P. Schimmel, G. G. van Hellemondt, A. B. Wymenga, and K. C. Defoort, "Stability and alignment do not improve by using patient-specific instrumentation in total knee arthroplasty: A randomized controlled trial," *Knee Surg., Sports Traumatology, Arthroscopy*, vol. 26, no. 6, pp. 1792–1799, Jun. 2018.

[52] E. Thienpont, P.-E. Schwab, and P. Fennema, "Efficacy of patient-specific instruments in total knee arthroplasty: A systematic review and meta-analysis," *J. Bone Joint Surg., Amer.*, vol. 99, pp. 521–530, Mar. 2017.

[53] K. Kizaki, A. Shanmugaraj, F. Yamashita, N. Simunovic, A. Duong, V. Khanna, and O. R. Ayeni, "Total knee arthroplasty using patient-specific instrumentation for osteoarthritis of the knee: A meta-analysis," *BMC Musculoskeletal Disorders*, vol. 20, no. 1, p. 561, Nov. 2019.

[54] J.-K. Seon, H.-W. Park, S.-H. Yoo, and E.-K. Song, "Assessing the accuracy of patient-specific guides for total knee arthroplasty," *Knee Surg., Sports Traumatology, Arthroscopy*, vol. 24, no. 11, pp. 3678–3683, Nov. 2016.

DAVIDE MARZORATI received the M.Sc. degree in biomedical engineering, in 2017, and the M.Sc. degree in bioengineering from the Politecnico di Milano and the University of Illinois at Chicago. He is currently pursuing the Ph.D. degree in biomedical engineering with the Politecnico di Milano. He also works in the field of deep learning for biological signal processing and image segmentation. His main research interest includes the development of sensors and devices for exhaled breath analysis with the aim of early disease diagnosis.



MATTIA SARTI received the M.Sc. degree in biomedical engineering, in 2020. He worked as a Teaching Assistant in Biomedical Image Processing and completed his studies as a Researcher in the area of neural network applications for microscopy image analysis. His main research interest includes deep learning, with a previous experience in several projects involving convolutional neural networks for signal classification and image segmentation.

LUCA MAINARDI (Member, IEEE) received the M.Sc. degree in electronics engineering, in 1990, and the Ph.D. degree in bioengineering, in 1996. He is currently a Professor with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, Italy, and the Co-Chair of the SPiNLab. He is the author of more than 120 peer reviewed articles on international journals and more than 140 conference papers. He authored 12 book chapters. He is an Editor of the book *Understanding Atrial Fibrillation: The Signal Processing Contribution* (USA: Morgan & Claypool). His research interests include biomedical signal and image processing, and biomedical system modeling with applications to the cardiovascular systems. He is a member of the Board of Computing in Cardiology (CinC) Annual Conference and the Chair of the IEEE EMBS Technical Committee on Biomedical Signal Processing. He is the Coordinator of the EU Marie-Curie Project MY-ATRIA through the H2020 Program Framework.

ALFONSO MANZOTTI received the M.D. degree from the Università degli Studi di Milano, Italy, in 1992. He has been the Chief of the Department of Orthopaedic and Trauma, Luigi Sacco Hospital, ASST FBF-Sacco, Milan, since 2015. Since 2015, he has been a Lecturer of the Orthopaedic and Traumatology Residency Program at the Università degli Studi di Milano. His research interests include customized joint replacement with mini-invasive approaches and small implants, computer assisted reconstructive surgery, and arthroscopic surgery. He is a member of the Italian Society of Orthopedics and Traumatology (SIOT), the Italian Society of Knee Surgery, Arthroscopy, Cartilage and Orthopedics Technology (SIGASCOT), the European Society of Sports Traumatology, Knee Surgery and Arthroscopy (ESSKA), and the European Knee Associates (EKA).

PIETRO CERVERI received the M.Sc. degree in electronics engineering, in 1994, and the Ph.D. degree in bioengineering, in 2001. He has been an Associate Professor in Bioengineering with the Politecnico di Milano, since 2015. He is the author or a coauthor of more than 100 scientific articles published on ISI journals. His research interest includes technologies for biomedical applications, with a special focus on diagnosis and therapy. He received the Best Innovation Award at K-Idea—Scientific Technological Park, Kilometro Rosso—Bergamo, Italy, for robotic technologies for the vision in mini-invasive transluminal endoscopic surgery, in 2008. Since 2012, he has been collaborating with the National Center for Oncological Hadron Therapy, Pave, Italy.