

Occlusion-robust Visual Markerless Bone Tracking for Computer-Assisted Orthopaedic Surgery

Xue Hu, Anh Nguyen, and Ferdinando Rodriguez y Baena, *Member, IEEE*

Abstract—Conventional computer-assisted orthopaedic navigation systems rely on the tracking of dedicated optical markers for patient poses, which makes the surgical workflow more invasive, tedious, and expensive. Visual tracking has recently been proposed to measure the target anatomy in a markerless and effortless way, but the existing methods fail under real-world occlusion caused by intraoperative interventions. Furthermore, such methods are hardware-specific and not accurate enough for surgical applications. In this paper, we propose a RGB-D sensing-based markerless tracking method that is robust against occlusion. We design a new segmentation network that features dynamic region-of-interest prediction and robust 3D point cloud segmentation. As it is expensive to collect large-scale training data with occlusion instances, we also propose a new method to create synthetic RGB-D images for network training. Experimental results show that our proposed markerless tracking method outperforms recent state-of-the-art approaches by a large margin, especially when an occlusion exists. Furthermore, our method generalises well to new cameras and new target models, including a cadaver, without the need for network retraining. In practice, by using a high-quality commercial RGB-D camera, our proposed visual tracking method achieves an accuracy of $1-2^\circ$ and $2-4$ mm on a model knee, which meets the standard for clinical applications.

Index Terms—Biomedical applications, Image processing, Neural networks, Vision-based instrumentation and measurement.

I. INTRODUCTION

RESTORING the mechanical axis of the lower limb is crucial in orthopaedic knee surgery [1]. For example, in total knee arthroplasty (TKA), the distal femur should be resected at a certain angle, and the prostheses should be congruently placed on the surrounding anatomy [2]. However, up to 20% of TKA procedures performed by experienced surgeons result in knee axis misalignment greater than 3° [3]. Implant misalignment could cause abnormal polyethylene wear, joint instability and early implant failure, all of which would have a significant impact on patients' quality of life [1].

Over the past decade, navigation systems have been recognised as a powerful tool to improve the efficacy and accuracy of knee surgery [4], [5]. By providing intraoperative measurements and pre-operative plannings in visual or numerical form, navigation systems guide the surgeon to reach the goal in various steps with greater control, precision, and consistency in time [6]. Conventional orthopaedic navigation systems usually embed an optical marker-based tracking mechanism to relate

the computer-stored information to the actual patient pose on the surgical table. These systems therefore rely on a dedicated system to track the movement of passive or active infrared (IR) markers that are rigidly pinned and registered to the target bone. The patient-specific information could be image-based (*e.g.*, pre-operative or intra-operative medical scans such as CT or MRI) [7] or image-free (*e.g.*, generic kinematic and/or morphological models parametrised onto the digitised anatomical landmarks) [8]. Surgeons need to first manually collect a set of key points such as implanted fiducials, anatomical landmarks or surface points, to which the image-based or image-free information can be registered by point-based [9] or surface-based approaches [10]. The registered initial target pose can then be updated according to the IR markers tracked by the optical tracker.

While marker-based tracking is currently regarded as the “gold standard” by many commercial navigation systems such as NAVIO (Smith & Nephew PLC) and MAKO (Stryker Corp.), three main limitations exist: first, the marker incision causes an additional scar and further surgical exposure, which may increase the risk of infection, nerve injury, and bone fracture [11], [12]. Second, surgeon involvement is required for marker preparation, fixation and marker-to-target registration. These steps have the potential to introduce additional human errors [13] and workload for surgeons [14], [15]. Finally, the bulky IR markers may interfere with surgeon's performance [16], as the immobile tracker requires constant line-of-sight to the target, which may restrict surgeon's movement in the operating room [17], [5].

Thanks to the fast development in depth sensing, commercial RGB-D cameras can be explored to replace the dedicated optical system. Once the camera sees the exposed target, the pixels associated with the target are automatically segmented from the RGB-D frames by trained neural networks, then the segmented surface is registered to a reference model in real-time to obtain the target pose. Albeit the concise workflow [18], [19], two aspects must be improved to move markerless tracking one step closer to surgical application [20]: first, as both training data collection and network design consider no target occlusion, markerless tracking drifts during intraoperative interventions (*e.g.*, bone drilling). Therefore, the target must be kept still during manipulation, which is impossible for knee surgeries. Second, the networks are trained on a dataset collected by a single consumer-level RGB-D camera. Limited by the camera's quality, the achieved accuracy is below the clinical acceptance. A more precise camera is essential to achieve higher tracking accuracy. Ideally, the network should be adaptable to new cameras without retraining.

Xue Hu and Ferdinando Rodriguez y Baena are with the Mechatronics in Medicine Lab, Imperial College London, London, SW7 2AZ UK. e-mail: xue.hu17@imperial.ac.uk.

Anh Nguyen is with the Hamlyn Centre, Imperial College London, London, SW7 2AZ UK.

This paper presents a RGB-D markerless tracking method for knee surgeries, which is robust to target occlusions and better in precision. To do so, we propose a new deep neural network for automatic RGB-D segmentation and, since collecting and labelling a large number of training data are highly tedious and time-consuming, we augment the existing real data containing no occlusion instances with synthetic RGB-D data containing various simulated target interactions. We show that, by utilising both 2D RGB images and 3D point clouds converted from depth frames, our network successfully learns to be robust to occlusion from synthetic data only, and generalises well to new RGB-D cameras and knee targets. A video is provided in supporting file to demonstrate the success of our network in the real world.

Our contributions can be summarised as follows:

- 1) We propose a robust markerless bone tracking algorithm for orthopaedic navigation, which proves the usability of RGB-D tracking for surgical navigation.
- 2) We introduce a new large-scale synthetic RGB-D dataset generated with simulated target occlusion that allows network training in an effort-efficient way.
- 3) We conduct intensive experiments to verify the effectiveness of our network design under different cameras, lighting conditions, and synthetic-to-real transfer learning scenarios.
- 4) Our method achieves clinically acceptable tracking error on a model leg with a high-quality commercial RGB-D camera. To the best of our knowledge, this is the first study that verifies the suitability of visual markerless tracking for clinical applications.

The rest of the paper is organised as follows. Starting with related work in Section II, we describe our methodology in Section III. Section IV presents an evaluation of network accuracy on real test data collected under occlusion. Section V shows the performance of markerless tracking on different targets and for different RGB-D cameras. Finally, we conclude the paper and discuss the future work in Section VI.

II. RELATED WORKS

A. Pose Measurement for Surgical Guidance

Research effort has been dedicated to improving the robustness and cost effectiveness of current navigation systems. Some studies combined the optical tracking with additional measurements to solve the line-of-sight problem: Vaccarella *et al.* fused optical and electromagnetic tracking based on an unscented Kalman filter [5]; Enayati *et al.* synchronised optical and inertial data by a quaternion-based unscented Kalman Filter [21]; Ottacher *et al.* proposed a compact 3D ultrasound system that combines conventional 2-D ultrasound with optical tracking [22]. Alternatively, for surgeries such as maxillofacial surgery [23] whose target is relatively clean and feature-rich, prominent anatomy features can be detected from RGB recording and registered for poses [24]. For surgeries with complex scenes (e.g., where the target is surrounded by blood and tissues), depth-sensing is exploited to improve the detection robustness. For example, Sta *et al.* proposed a point-pair features algorithm to estimate the pose of TKA implant

from depth captures, but such a feature-based method cannot be run in real-time.

B. RGB-D Learning-based Markerless Tracking

Commercial RGB-D cameras can achieve fast and accurate measurement in a high resolution, making them potential new tools for surgical navigation. For real-time tracking purposes, learning-based methods were explored in the literature to extract comprehensive features automatically. Yang *et al.* designed a fully convolutional network (FCN) to automatically segment spine area from RGB-D captures so that the pre-plannings can be overlaid accordingly [19]. The RGB and depth features were encoded from input 2D maps, fused at different stages of encoder, and decoded jointly to predict the segmentation mask [19]. Liu *et al.* proposed a sequential RGB-D network for automatic femur tracking [25]. The target centre was roughly localised by a convolutional neural network (CNN) in the RGB frame first. Then the aligned depth map was cropped around the predicted centre with a fixed size of 160×160 pixels, and passed to another CNN to finely predict the femur mask. The femur area was segmented from depth maps according to the prediction, converted to 3D points, and registered to a scanned model by iterative closest point (ICP) in real-time.

Unlike these literature methods which focus on a clean target surface and train the network with collected real data [19], [25], we aim to improve the tracking robustness when the target is manipulated under occlusion, by generating a synthetic dataset to train a segmentation network with new design.

III. SYNTHETIC DATA CREATION

While the available dataset collected in [25] by a RealSense D415 camera (Intel Corp.) has a limited size (5200 independent frames on a cadaver knee and a model knee) and contains no target occlusion, a large dataset with occlusion instances is essential to train a network that works within an intraoperative scenario. To expand the current training data in a fast and efficient way, we generate synthetic data using a modular procedural pipeline, BlenderProc [26], on an open-source rendering platform, Blender [27]. The details of data generation are described below.

1) *Creation of Randomised Scenes:* A model knee is scanned by a highly precise scanner (HDI 3D scanner, LMI Technologies Inc.). The obtained frames are co-registered into a single point cloud. After hole filling [28] and Laplacian normal smoothing [29], a 3D knee model is reconstructed from the merged point cloud by application of the Screened Poisson algorithm [30]. Then, the model is manually divided into the femur and not-femur sections and imported into the Blender pipeline via the Python scripting interface.

As suggested in [31], domain randomisation is critical to overcoming the simulation-to-reality gap in RGB data. Therefore, we randomly alter the scene during image generation regarding (Fig. 1):

- The type (point or surface) and strength of lighting.

- The room background, which contains arbitrary extrusions and objects loaded from the Ikea dataset [32] as distractors. The materials of the wall, floor and loaded objects are randomly sampled from a large public material database, ambientCG [33].
- The material of skin and exposed bone, by blending a random texture with a random RGB colour.

2) *Simulation of Foreground Occlusion*: Five 3D models of human hands and surgical tools are prepared and imported as foreground distractors. These objects are randomly positioned and orientated within the camera’s line-of-sight of the exposed femur to simulate partial target occlusion. The fingertip or tooltip, defined as the origin of local object coordinates, can optionally contact and translate on the femur surface. The material of these objects is also altered following the random texture blending method mentioned above.

3) *Adding Depth Sampling Noise*: The simulation-to-reality gap was found to be more significant in depth imaging due to the complex noise profiles and diverse ecosystem of sensors [34]. A commercial structured-light depth camera mainly experiences three kinds of sampling noise:

- Pixel location offsets due to quantised disparity [35]: the final depth value at a pixel $Z(x, y)$ is interpolated from the raw sampling at adjacent pixels $Z(x + \delta_x, y + \delta_y)$.
- The IID Gaussian deviation of depth values, presumably due to sensor noise or errors in the stereo algorithm [36] (*i.e.*, $Z(x, y) = \hat{Z}(x, y) + \delta_z$).
- The depth value dropout at some pixels (*i.e.*, $Z(x_d, y_d) = 0$) due to two reasons: the interaction of the projected IR pattern with illumination and target material, or the interaction of the depth camera’s projector-receiver pair with the scene [34].

Modelling the dropout noise is extremely challenging, since the material- and illumination-dependent interaction can not be physically simulated. Besides, the dropout density is subject to specific camera properties like spatial sampling resolution and the baseline distance between projector and receiver. Therefore, we model the first two types of noise in a Gaussian distribution with $(\delta_x, \delta_y) \sim \mathcal{N}(0, 1/2)$ and $\delta_z \sim 0.08\mathcal{N}(0, 1/3)$ (in mm), according to the datasheet of D415 camera [37].

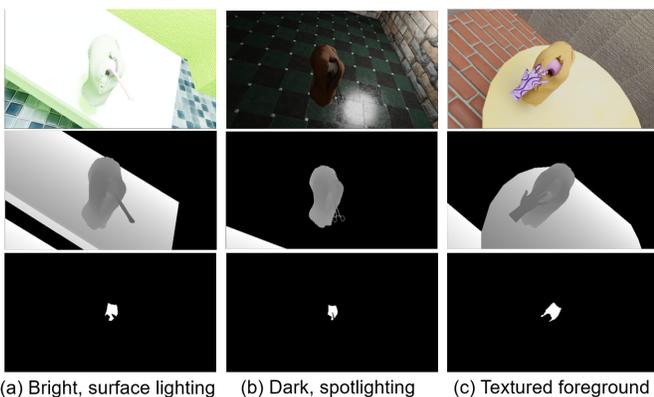


Fig. 1: Example synthetic images with variations in the strengths and types of lighting, background and foreground.

4) *Statistics*: For each image-generation session with a settled scene, 20 captures are taken with random camera poses. The viewpoint is controlled to be 0.5-1m away from the target to replicate the physical working distance. The sampling intrinsic parameters and resolution are set to the physical values of a RealSense camera calibrated by a standard routine [38]. The visibility of the exposed femur is checked for each sampling pose to ensure a meaningful capture. The simulation is repeated to produce 10,000 randomised synthetic RGB-D images together with automatically labelled binary femur masks. Fig. 1 shows some examples of generated synthetic images.

IV. MARKERLESS SEGMENTATION AND REGISTRATION

Fig. 2 shows an overview of the proposed markerless tracking workflow. The whole procedure can be divided into two steps: automatic target segmentation and real-time pose registration. In this section, we will describe how we implement each part for better tracking robustness and accuracy.

A. Automatic Segmentation Network

Similar to [25], our segmentation network contains a sequential arrangement to leverage both RGB and depth imaging (Fig. 2). The stable RGB stream ensures robust target localisation in the full scope of captures, while the depth data ensure fine segmentation, as they are less impacted by bleeding and surgical lighting [39]. The RoI box is first predicted from the global RGB frame by a RoINet, according to which the aligned depth frame is cropped and resampled into a 3D point cloud. A SegNet then predicts the femur mask from the cloud for point-wise segmentation. The details for both networks are explained below.

1) *RoINet*: Unlike [25], where the authors regress a target centre location and crop the depth frames with a fixed box size to match the input dimension of the segmentation CNN, our network directly predicts an RoI box with an adaptive size to more tightly bound the exposed femur surface. The change is required for two reasons: first, to ensure high segmentation speed for real-time tracking, only a certain number of resampled points (N) could be taken by the segmentation network. However, a sufficient number of segmentation outputs (N_f) are desired for reliable pose registration. Therefore, RoI cropping should ensure a high target occupation rate N_f/N . Second, when the camera moves towards or away from the target, or the network is deployed to a new camera with a considerably different focal length, a fixed cropping size may fail to cover the whole target dimensions. Therefore, RoI cropping should be dynamic in size to ensure a nearly constant value for N_f/N .

The RoINet, as shown in Fig. 3, is modified from the localisation network proposed in [25], by adding two mid-layer auxiliaries and a multi-box loss function. With a similar design to Alexnet, the first five convolutional layers extract feature maps with shrinking sizes from the input RGB image. Inspired by the Single Shot Multibox Detector (SSD) [40], M multi-scale feature maps are taken from different layers and convoluted by 3×3 kernels to produce M bounding boxes with a probability for the presence of the target in the box

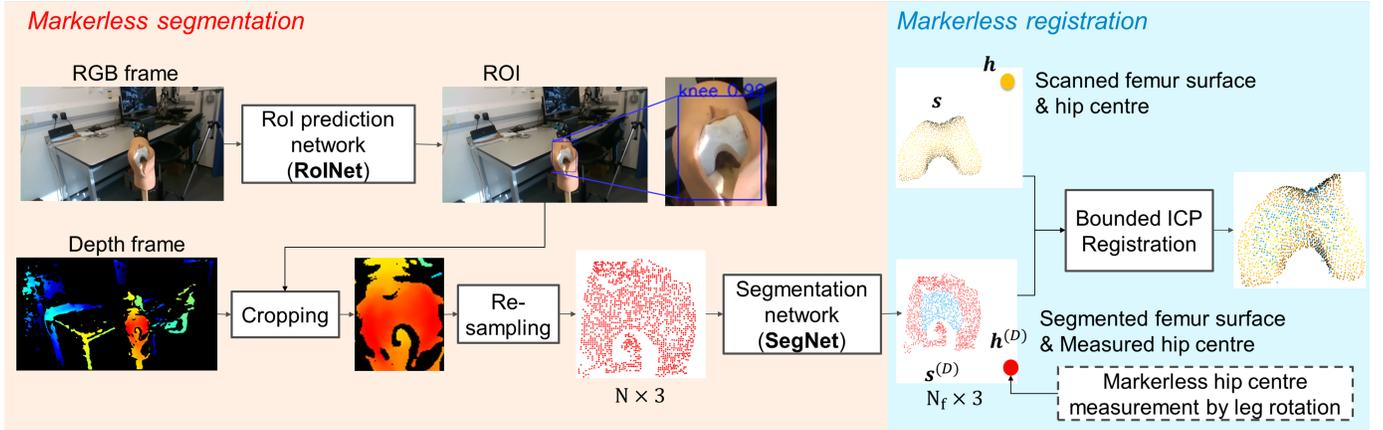


Fig. 2: Overview of markerless tracking: Starting from an aligned RGB-D frame, we initially compute the RoI for the exposed target femur using our RGB-based RoINet. After cropping the depth frame with predicted RoI, a $N \times 3$ point cloud is resampled and input to the segmentation network to predict the femur label for every point. The N_f segmented femur points are then registered to a pre-scanned reference model by a Bounded ICP algorithm implementation in real-time to obtain the target pose.

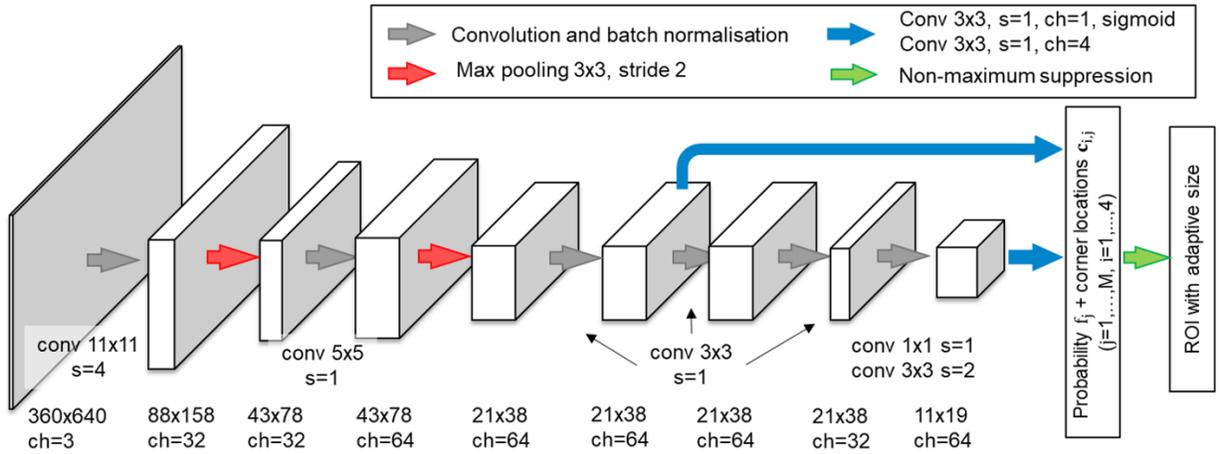


Fig. 3: Architecture of the RoI prediction network based on RGB information. We extract feature maps by an Alexnet backbone, and take multi-scale features for multi-box classification and corner regression. In our implementation, $M=21 \times 38 + 11 \times 19$.

($0 < f < 1$). Each bounding box $\mathbf{c} = [c_1, c_2, c_3, c_4]$ is uniquely decided by the x and y offset of upper-left and lower-right corners relative to the default box coordinates of $[-0.5, -0.5, 0.5, 0.5]$. The overall $M \times (4 + 1)$ predictions are processed by a non-maximum suppression to decide the best RoI box.

2) *SegNet*: The generated depth maps shown in Fig. 1 apparently lack realistic depth dropout. Fortunately, compared to 2D depth maps, the 3D point cloud representation of depth data is less vulnerable to such sampling artefacts (Section V-C). Network-learned features should be similar in both real and synthetic domains to ensure knowledge transfer; they should also be robust to camera sampling properties so that the trained network is camera-agnostic.

Consequently, as shown in Fig. 4, our SegNet is designed to learn from the 3D point cloud representation rather than the 2D depth maps (as used in [19], [25]). It takes over the PointNet architecture [41] to predict from an $N \times 3$ input point cloud $\{\mathbf{p}\}$. The input points are processed by a succession of Multi-Layer Perceptrons (MLPs) to produce $N \times 1024$ encoded

features. A symmetric maximum pooling function is applied to extract a 1024-dimensional global descriptor, which is then concatenated with a local feature vector taken from a mid-layer. Next, the combined latent features are decoded by MLPs and reshaped into an N -dimensional vector. The predicted vector is finally mapped between 0-1 by a sigmoid function. The output values $\{p\}$ represent how possible is it that each of the N points belongs to the femur surface. The point with predicted probability p_j higher than a threshold (0.8 in our implementation) can be regarded as a target femur point.

3) *Network Training*: The whole dataset is randomly divided into training and validation sets by a ratio of 8:2. The two networks are separately trained using the Tensorflow library [42]. For the RoINet, the batch size is set to 4 and the training loss C_{rol} is defined as the total difference of predicted box corner locations and probabilities ($c_{i,j}, f_j$) of the labelled

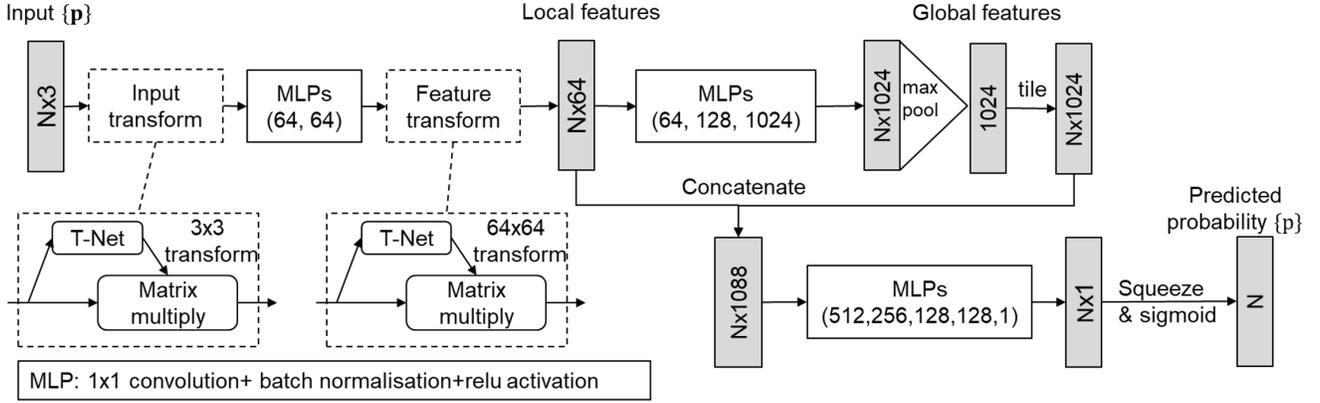


Fig. 4: Architecture of the 3D SegNet. $N \times 3$ input points are resampled from the cropped depth frames according to the predicted RoI. The encoder-decoder structure follows the design of PointNet. The decoded 1-channel output is additionally processed by a sigmoid function to predict the probability. $N=2000$ in our implementation.

values $(\hat{c}_{i,j}, \hat{f}_j)$:

$$C_{\text{RoI}} = \sum_{j=1}^M \left(\sum_{i=1}^4 |c_i - \hat{c}_{i,j}| + |f_j - \hat{f}_j| \right) \quad (1)$$

For the SegNet, the batch size is set to 32 and the training loss C_{Seg} is defined as the sum of absolute differences between the predicted probabilities p_j and binary femur labels \hat{p}_j .

$$C_{\text{Seg}} = \sum_{j=1}^N |p_j - \hat{p}_j| \quad (2)$$

The Adam optimiser with an exponentially decaying learning rate starting from 0.001 is used for both training to ensure a steady rate of learning.

B. Markerless Registration

As the segmented points have a limited spatial spread over the partially exposed femur area, classical ICP-based registration is vulnerable to rotational misalignment [43]. For higher registration accuracy and better robustness against wrongly segmented points, we thus adopt a previously validated and published Bounded ICP (BICP) method [18]. BICP uses a remote pair of corresponding features (*e.g.*, the model hip \mathbf{h} and measured hip centre $\mathbf{h}^{(D)}$) to bound the registration error between the scanned model surface \mathbf{s} and the automatically segmented femur surface $\mathbf{s}^{(D)}$:

$$\mathbf{P}^{(D)}(t) = \text{BICP}(\mathbf{s}^{(D)}(t), \mathbf{s}, \mathbf{h}^{(D)}(t), \mathbf{h}) \quad (3)$$

\mathbf{s} , \mathbf{h} and $\mathbf{h}^{(D)}(t)$ are obtained with a setup similar to that proposed in [18]:

1) *Model surface and hip location*: In our laboratory setup, before online tracking, the model surface \mathbf{s} is digitised from the femur under maximum skin exposure. The hip centre \mathbf{h} is sphere-fitted from the probed surface points of the ball joint (Fig. 5). In a clinical setup, the model \mathbf{s} and \mathbf{h} would instead be reconstructed from pre-operative images such as CT or MRI.

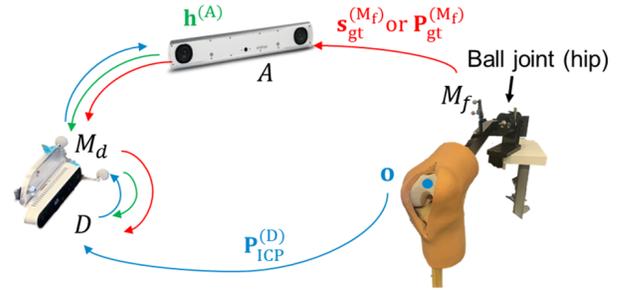


Fig. 5: The system setup for BICP registration and evaluation. Blue lines: transformation from the tracked landmark \mathbf{o} into $\mathbf{o}^{(A)}$ for hip centre calculation; Green lines: online transformation of the fitted hip centre $\mathbf{h}^{(A)}$ into $\mathbf{h}^{(D)}$. Red lines: transformation of ground truth femur surface or pose for labelling or evaluation.

2) *Measured hip location*: As shown in Fig. 5, an optical marker M_d is rigidly anchored to the depth camera D , which is tracked by a global optical tracker A (FusionTrack 500, Atracsys LLC.) to obtain a hip centre estimate. The hip centre can be modelled as the pivot point around which the leg is rotated. To track a femur landmark during rotation in a markerless way, we combine the aforementioned automatic segmentation with ICP registration to track a rough femur pose $\mathbf{P}_{\text{ICP}}^{(D)}(t)$. The local model origin $\mathbf{o} = [0, 0, 0, 1]^T$ is chosen as the landmark to avoid any projection bias due to the rotational registration error. The landmark positions tracked by ICP-based markerless tracking are transformed into global coordinates by the hand-eye calibrated transformation $M_d^A \mathbf{T}$ and optically tracked $M_d^A \mathbf{T}$ as follows:

$$\mathbf{o}^{(A)}(t) = M_d^A \mathbf{T}(t) \times M_d^D \mathbf{T} \times \mathbf{P}_{\text{ICP}}^{(D)}(t) \times \mathbf{o} \quad (4)$$

During rotation, more than 40 frames of $\mathbf{o}^{(A)}$ are recorded, from which the still hip centre $\mathbf{h}^{(A)}$ is computed by a sphere-fitting algorithm [18]. The estimated global hip location $\mathbf{h}^{(A)}$ is

finally transformed back to the depth camera frame as $\mathbf{h}^{(D)}(t)$ by ${}^M_d\mathbf{T}(t)$ for online BICP registration (green path in Fig. 5).

V. NETWORK EVALUATION

A. Test Data Collection

To evaluate the performance of the trained segmentation network in the real world, we collected 800 RGB-D captures by a RealSense D415 camera, during which the target femur was partially occluded by hands or tools. To automatically label the femur pixels, an optical marker M_f was inserted into the metal leg so that the ground truth (gt) femur surface could be optically tracked (red path in Fig. 5). After a standard exposure, the femur surface was manually digitised as $\mathbf{s}_{gt}^{(A)}$. The probed surface was then calibrated to M_f as $\mathbf{s}_{gt}^{(M_f)}$, and further transformed into D according to:

$$\mathbf{s}_{gt}^{(D)}(t) = {}^D_M\mathbf{T} \times {}^M_A\mathbf{T}(t) \times {}^A_{M_f}\mathbf{T}(t) \times \mathbf{s}_{gt}^{(M_f)} \quad (5)$$

As suggested by [25], the transformed surface points $\mathbf{s}_{gt}^{(D)}$ were finally registered to the raw depth capture by a standard ICP algorithm to identify the matching pixels that should be labelled as femur points. However, when hands or tools occluded the target surface, the registration between digitised surfaces and unsegmented captures became highly unreliable. To ensure correct annotation under target occlusion, we utilised pairwise captures. As shown in Fig. 6, the target was first captured with no surface occlusion or contact, then labelled by ICP-based point matching as described above (frame 1). Subsequently, without moving the camera or target, another capture was carried out for the femur surface while being partially occluded by a hand in a purple glove or a tool wrapped in purple tape to simplify the segmentation process, as follows. The femur mask labelled in frame 1 was applied to frame 2's RGB frame to segment an RoI, which was then converted to hue saturation and value (HSV) format, and filtered by a band-pass hue filter in the purple colour range to identify the pixels that belong to the foreground. The gt femur pixels for frame 2 were finally computed by subtracting the femur pixels in frame 1 by the detected foreground pixels in frame 2. The gt RoI box was computed as the smallest rectangle that covers all gt femur pixels.

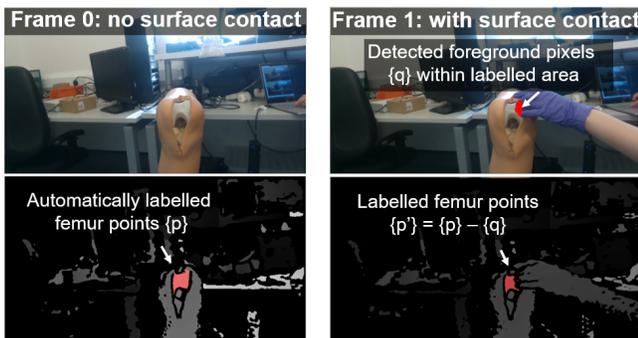


Fig. 6: Generation of the ground truth label mask for a target femur under surface contact based on a pairwise capture.

B. Results

Fig. 7 shows some example images with overlaid Grad-CAM heat maps obtained by the proposed RoI prediction network. Regardless of hand occlusion, tool manipulation, capturing perspective and human presentation, the network properly pays attention to the exposed femur. If the intersection over union (IoU) between the predicted RoI and gt RoI is higher than 0.5, the prediction is regarded as successful. The overall accuracy is presented by the success rate of predictions over the entire test dataset. The localisation network trained in [25] is also tested as a reference for comparison. The predicted RoI is regarded as the box drawn around the inferred target location, with the same size as the ground truth RoI box.

Depending on the gt label (positive: is femur; otherwise negative) and the correctness of prediction (true: prediction matches gt; otherwise false), the N points can be classified as true positive (TP), true-negative (TN), false positive (FP) and false-negative (FN). To avoid the bias arising from a large number of TN predictions for background points, the segmentation accuracy is defined as the IoU score in each frame:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (6)$$

The overall accuracy is presented by the mean and standard deviation of IoU values over the full dataset. Table I lists the evaluated accuracy of our networks and the reference networks proposed in [25]. Our networks are almost twice more accurate than the reference networks.

TABLE I: Accuracy comparison of RoI prediction and point/pixel segmentation, between our networks and the reference networks proposed in [25].

	RoI	Seg	RoI+Seg
Liu <i>et al.</i> [25]	67.54%	42.03±32.96%	39.45±30.18%
Ours	94.78%	85.42±12.43%	84.20±14.43%

C. Ablation Study

The higher accuracy of our trained networks may be due to the new network structures, or the synthetic data included for training. We run ablation tests to study the effect of each component. Specifically, we want to answer three questions:

- 1) Are the synthetic data helpful in improving the robustness to occlusion for our networks?
- 2) Can other (*e.g.*, Liu *et al.* [25]) networks be improved by learning on synthetic data?
- 3) What is the critical factor that causes a difference in transferring ability?

To answer the first two questions, we additionally trained the proposed networks on the real part of the data only, and the reference network [25] on our synthetic-included dataset. By comparing the “Real” with “Real+sim” group shown in Table II, the simulated images significantly improve the robustness of our RGB and depth networks against real-world occlusion, while it harms the reference networks [25].

For the last question, we investigated the segmentation network first. The proposed structure learns 3D geometric



Fig. 7: The femur class GRAD-CAM activation heat map with the predicted ROI box and confidence.

TABLE II: Ablation study for the effect of synthetic images and network structure.

Network	Training data	RGB	D
Liu <i>et al.</i> [25]	Real	67.54%	42.03±32.96%
	Real+sim	60%	0
	Real+sim with dropout	-	57.81±31.22%
Ours	Real	52.87%	76.81±17.83%
	Real+sim	94.78%	85.42±12.43%
	Real+sim with dropout	-	85.37±11.70%
Ours without RGB auxiliary	Real	32.95%	-
	Real+sim	74.80%	-

features from an unorganised point cloud, whereas the reference structure learns 2D features from a cropped depth map. As shown in Table III, the depth dropout artefact makes the simulated 2D depth maps clearly different from real captures, but has less effect on the converted 3D point cloud since both data are sampled from the same 3D geometry. To prove the correlation between depth dropout and transferring ability, we generated 5,000 synthetic depth frames with simulated partial dropout noise caused by the interaction between scene and projector-receiver: an extra viewpoint was set up in Blender as the pattern projector in addition to the main viewpoint as the signal receiver. The ray cast from the projector to each sampled pixel was computed to find the pixels that cannot receive projected patterns. The depth values of those pixels were then overridden by zeros, resulting in a more realistic 2D depth map (Table III). Our proposed network and reference segmentation network [25] were then trained with dropout-included synthetic data. By comparing the “Real+sim” and “Real+sim with dropout” group shown in Table II, as expected, the partially modelled dropout artefact improves the knowledge transfer for the reference network, but makes no difference to the proposed network.

We then turned our attention to the RoI prediction network. Compared to the reference network, our network is different by having two mid-layer auxiliaries and a multi-box loss for training. We removed the mid-layer auxiliaries from the proposed architecture and trained the modified network on the proposed dataset. As shown in Table II, the network can still learn from simulated images (*i.e.*, with an improvement from 32.95% to 74.80%), but the prediction accuracy was reduced by around 20%. The degradation implies the importance of mid-layers for prediction accuracy, but not for synthetic-to-real transfer. Therefore, we speculate that the learning ability on synthetic data mainly comes from the training on multi-

	Real capture	Simulated, no dropout noise	Simulated, with dropout noise
2D depth map			
3D point cloud			

TABLE III: Examples of simulated depth data represented in 2D depth map and 3D point cloud.

box loss. In fact, our RoINet localises the target in RGB frames by multi-box classification rather than direct regression. The classification is more tolerant of inconsistent features on different domains.

VI. EXPERIMENTS ON MARKERLESS TRACKING

A. Implementation

While the network inference was scripted in Python, for faster speed, the BICP registration was coded in C++ and compiled into a dynamic linked library (DLL) that could be called in Python. Executed on a computer (Intel®Core™i5-8250U processor) with no dedicated graphics processing unit, each RoI prediction took approximately 0.01s, the point segmentation took approximately 0.04s, and the BICP registration took approximately 0.05s. Two threads were executed in parallel for the frame acquisition and inference, and the BICP registration, respectively. Given the RealSense camera’s 30 Hz frame rate, the overall markerless tracking update frequency was found to be around 12 Hz.

The same setup shown by the red path in Fig. 5 was used to obtain the gt femur pose for accuracy evaluation. The pre-scanned model \mathbf{s} was first registered to the manually digitised bone surface for the initial pose $\mathbf{P}_{gt}^{(A)}$, then transformed into M_f as a time-invariant local pose $\mathbf{P}_{gt}^{(M_f)}$. The registered initial gt pose can be updated continuously based on optical tracking:

$$\mathbf{P}_{gt}^{(D)}(t) = {}^D\mathbf{T} \times {}^{M_d}\mathbf{T}(t) \times {}^A\mathbf{T}(t) \times \mathbf{P}_{gt}^{(M_f)} \quad (7)$$

The real-time tracking error was defined as the relative transformation between the markerless-tracked femur pose and the gt pose in D:

$$\mathbf{P}_{err} = \mathbf{P}_{gt}^{(D)}(t)^{-1} \times \mathbf{P}^{(D)}(t) \quad (8)$$

\mathbf{P}_{err} was decomposed into the 3D rotational and translational misalignment. During each experiment, the RGB-D camera was held by a tripod and randomly placed at 10 different locations around the target knee. More than 50 frames of evaluated \mathbf{P}_{err} were collected from each camera position to quantify the overall tracking error.

B. Comparison with the Literature

The RealSense D415 camera was first tested on the same model knee used for synthetic data creation. The markerless tracking proposed in [25] was implemented and tested under the same setup, as a reference for performance comparison. As shown in Fig. 8, without target interaction, both reference and proposed methods track properly. When the femur is partially occluded by the hand, the ROI centre predicted by the reference RGB network drifts slightly from the actual femur centre. Fortunately, as the fixed cropping size (*i.e.*, 160) is large enough at the working distance, the cropped depth frames may still contain the target femur. However, the reference segmentation network fails to identify the femur pixels, resulting in an unreliable pose. In contrast, both the proposed ROI prediction and segmentation networks work well under hand occlusion.

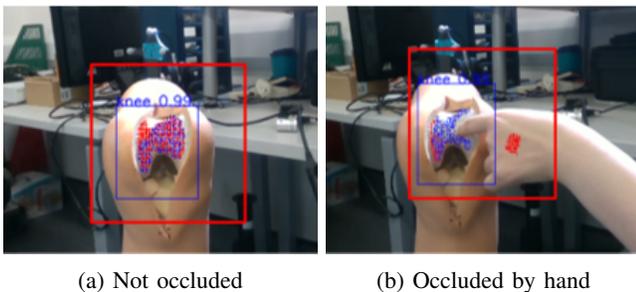


Fig. 8: Overlaid markerless segmentation (predicted ROI and segmented points) by Liu *et al.* (red) and our networks (blue).

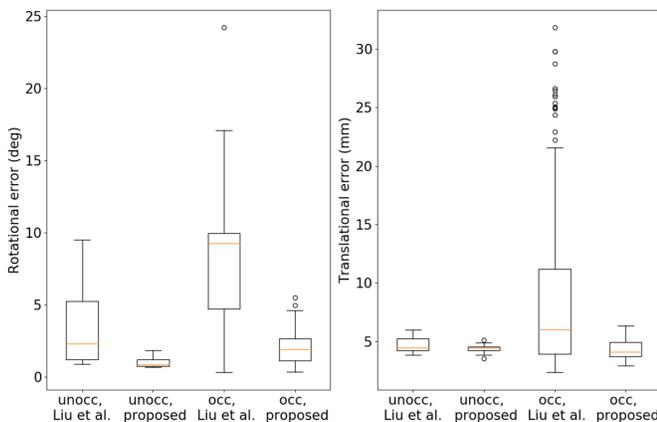


Fig. 9: Accuracy of our proposed method and the reference method by Liu *et al.* [25] with/without target occlusion.

Fig. 9 compares the BICP-based markerless tracking accuracy obtained by the proposed and reference segmentation

networks. The Kruskal-Wallis test was used to check whether the difference between obtained results is statistically significant. No matter whether the target occlusion exists, the proposed tracking can achieve better accuracy than the reference tracking (p-values < 0.001 in both rotation and translation). The proposed markerless tracking achieves $1.02^\circ \pm 0.33^\circ$, $4.39 \text{ mm} \pm 0.33 \text{ mm}$ error with no occlusion (unocc), and $2.05^\circ \pm 1.10^\circ$, $4.33 \text{ mm} \pm 0.78 \text{ mm}$ error under occlusion (occ). There is no significant difference in translation (p-value = 0.21) but in rotation (p-value < 0.001).

C. Camera Agnostic Performance

Despite the promising results, the RealSense D415 camera is not designed for highly precise tasks. Therefore, a more accurate depth camera should be adopted for future clinical applications. To test the generalisability of the proposed network on new cameras, and to show the potential of markerless tracking in achieving higher accuracy, we deployed the trained network with an Acusense RGB-D camera (Revopoint 3D Technologies Inc.) that claims sub-millimetre accuracy within the 1-meter working distance. The Acusense camera, based on the coded IR structured light technology, has a higher RGB resolution (600×800) and much longer focal length (*e.g.*, $f_x = 2061$ compared to $f_x = 460$ for RealSense camera). We subsampled the raw RGB frames into 300×400 and padded the margin by white pixels into the designed input size of 360×640 for RoINet. The predicted box corners were then mapped to the depth frames for cropping. Given the camera's much lower frame rate of around 6-7 Hz, the overall markerless tracking reached a 5-6 Hz refresh rate with multi-threading computation.

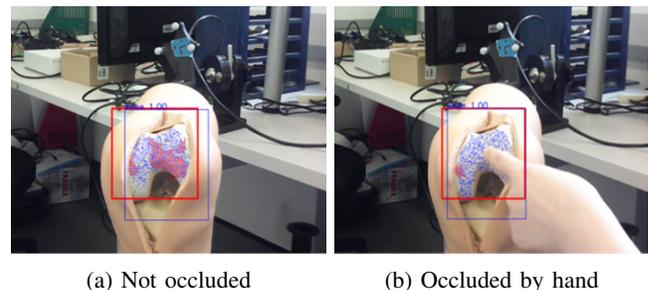


Fig. 10: Overlaid markerless segmentation by Liu *et al.* (red) and proposed networks (blue) with a new Acusense camera.

Fig. 10 demonstrates the strength of our tracking over the proposed method by Liu *et al.* [25] regarding device dependency or the lack thereof. While the fixed-size cropping by [25] fails to cover the full target, our dynamic RoINet efficiently adapts to a larger cropping size. The segmentation network in [25] is also less robust than our SegNet, which could be caused by the different features in 2D depth maps, since the Acusense camera has a higher spatial resolution and less dropout effect around the edges. Fig. 11 shows how the tracking accuracy changes after using a more precise RGB-D camera. There is a significant accuracy improvement in translation (p-values < 0.001 for both occ and unocc) but not in rotation (unocc: p-value = 0.25; occ: p-value = 0.24). The

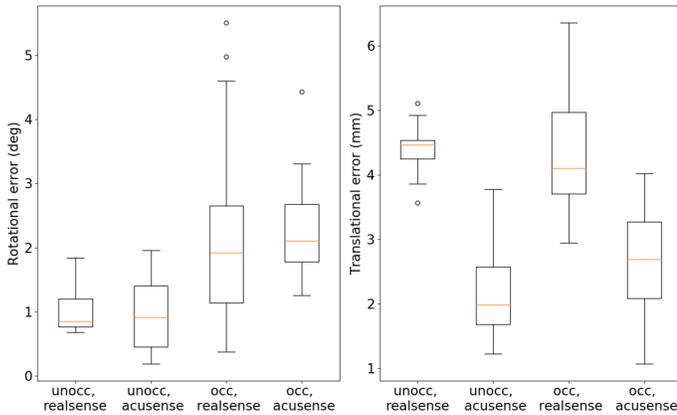
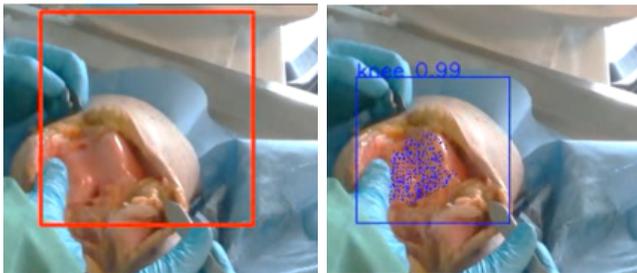


Fig. 11: Accuracy of the proposed markerless tracking with/without occlusion, tested on different cameras.

markerless tracking error is $0.95^\circ \pm 0.55^\circ$, $2.17 \text{ mm} \pm 0.62 \text{ mm}$ with no occlusion, and $2.24^\circ \pm 0.73^\circ$, $2.62 \text{ mm} \pm 0.85 \text{ mm}$ under occlusion. According to the quantitative score table for guide concepts proposed by Audenaert *et al.* [44], the accuracy obtained here is in the clinically “acceptable” range (*i.e.*, error less than 4° and 4 mm).

D. Generalisation Ability

A general question is whether the network will still work if the target anatomy is different from the model/s used for training. We tested the qualitative segmentation performance on a cadaveric knee during a partial joint replacement dissection study (approved by Imperial College Healthcare NHS Trust Tissue Bank with the number R15022 for the use of human cadavers), where we had no ground truth to compare to, and quantitative tracking accuracy on a new (and different) model knee. Both of the targets had never been seen by the network during training.



(a) By Liu *et al.*

(b) By ours

Fig. 12: Markerless segmentation by Liu *et al.* (red) and proposed networks (blue) on a new cadaver knee under occlusion. Results are shown in pairwise recording.

As shown in Fig. 12, while the method proposed by Liu *et al.* fails under occlusion, our network gives reliable predictions. Fig. 13 shows the quantified tracking accuracy by the proposed method on a new model knee. Although never seeing the target, the tracking accuracy remains high (*i.e.*, $1.07^\circ \pm 0.25^\circ$, $4.94 \text{ mm} \pm 0.23 \text{ mm}$ with no occlusion, and

$2.82^\circ \pm 1.22^\circ$, $5.21 \text{ mm} \pm 0.83 \text{ mm}$ with occlusion), indicating good generalisability to new geometry. Compared to the old knee, the new target experiences slightly higher tracking error in unoccluded translation, occluded rotation and occluded translation (p -values < 0.05), suggesting that including more instances of target geometry for training may further improve the network performance.

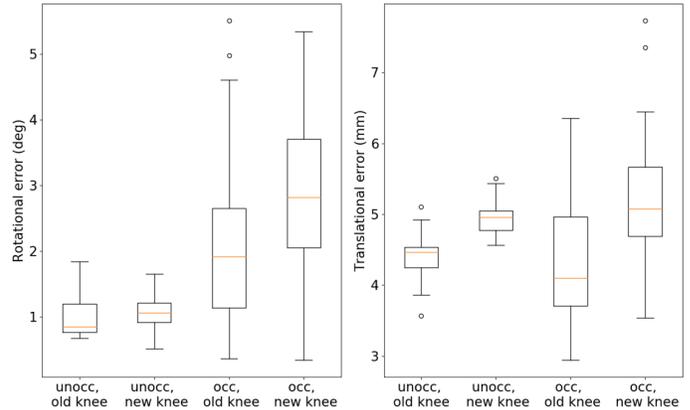


Fig. 13: Accuracy of the proposed markerless tracking with/without occlusion, tested on different model knees.

VII. CONCLUSION AND FUTURE WORK

In this work, we proposed a new RGB-D based occlusion-robust markerless femur tracking method for computer-assisted knee surgeries. By training the network on a padded dataset with synthetic images, the robustness to target occlusion is learned in a cost and effort-efficient way. To ensure effective synthetic-to-real transfer, we show that the multi-box loss is critical for RoI prediction and learning on the 3D point cloud is vital for robust segmentation. While the state-of-the-art markerless tracking fails under target occlusion, our method can achieve a stable accuracy of around 2° and 4 mm no matter whether the target is fully visible to the camera or not. The proposed tracking can be deployed on new target geometries (including a cadaver knee) and with new RGB-D cameras without the need for network retraining. Consequently, we demonstrated here that, by simply using a more precise camera, we could achieve a tracking error of around 1 - 2° and 2 - 4 mm , a marked performance improvement that now meets the requirements for clinical deployment. The results indicate the possible use of RGB-D imaging as a new modality for surgical applications.

Our synthetic training data can be further improved for better network performance. The imported knee model is currently considered as a rigid body with a fixed femur exposure. By modelling the skin part as a non-rigid body controlled by some critical nodes, various extents of skin exposure could be included in the synthetic images. Besides, as suggested by results on a new knee, including more target geometries in simulation could enrich the generated data. Finally, we will fully demonstrate the overall markerless navigation workflow in a cadaveric study in the future.

REFERENCES

- [1] A. F. Mavrogenis, O. D. Savvidou, G. Mimidis, J. Papanastasiou, D. Koulalis, N. Demertzis, and P. J. Papagelopoulos, "Computer-assisted navigation in orthopedic surgery," *Orthopedics*, vol. 36, no. 8, pp. 631–642, 2013.
- [2] N. Sugano, "Computer-assisted orthopedic surgery," *Journal of Orthopaedic Science*, vol. 8, no. 3, pp. 442–448, 2003.
- [3] J. Mahalaxmivala, M. Bankes, P. Nicolai, C. Aldam, and P. Allen, "The effect of surgeon experience on component positioning in 673 press fit condylar posterior cruciate-sacrificing total knee arthroplasties," *The Journal of arthroplasty*, vol. 16, no. 5, pp. 635–640, 2001.
- [4] W. Siebert, S. Mai, R. Kober, and P. F. Heeckt, "Technique and first clinical results of robot-assisted total knee replacement," *The Knee*, vol. 9, no. 3, pp. 173–180, 2002.
- [5] A. Vaccarella, E. De Momi, A. Enquobahrie, and G. Ferrigno, "Unscented kalman filter based sensor fusion for robust optical and electromagnetic tracking in surgical navigation," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 7, pp. 2067–2081, 2013.
- [6] G. Figueras-Benítez, L. Urbano, A. Acero, M. Huerta, and M. Castro, "Surgical navigation systems: A technological overview," in *VII International Conference on Electrical Engineering*, 2014.
- [7] J. Victor and D. Hoste, "Image-based computer-assisted total knee arthroplasty leads to lower variability in coronal alignment," *Clinical Orthopaedics and Related Research*, vol. 428, pp. 131–139, 2004.
- [8] B. A. Rebal, O. M. Babatunde, J. H. Lee, J. A. Geller, D. A. Patrick Jr, and W. Macaulay, "Imageless computer navigation in total knee arthroplasty provides superior short term functional outcomes: a meta-analysis," *The Journal of arthroplasty*, vol. 29, no. 5, pp. 938–944, 2014.
- [9] J. Hong and M. Hashizume, "An effective point-based registration tool for surgical navigation," *Surgical endoscopy*, vol. 24, no. 4, pp. 944–948, 2010.
- [10] X. Chen, Z. Song, and M. Wang, "Automated global optimization surface-matching registration method for image-to-patient spatial registration in an image-guided neurosurgery system," *Journal of Medical Imaging and Health Informatics*, vol. 4, no. 6, pp. 942–947, 2014.
- [11] R. W. Wysocki, M. B. Sheinkop, W. W. Virkus, and C. J. Della Valle, "Femoral fracture through a previous pin site after computer-assisted total knee arthroplasty," *The Journal of arthroplasty*, vol. 23, no. 3, pp. 462–465, 2008.
- [12] A. P. Schulz, K. Seide, C. Queitsch, A. Von Haugwitz, J. Meiners, B. Kienast, M. Tarabolsi, M. Kammal, and C. Jürgens, "Results of total hip replacement using the robodoc surgical assistant system: clinical outcome and evaluation of complications for 97 procedures," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 3, no. 4, pp. 301–306, 2007.
- [13] D. K. Bae and S. J. Song, "Computer assisted navigation in knee arthroplasty," *Clinics in orthopedic surgery*, vol. 3, no. 4, pp. 259–267, 2011.
- [14] D. C. Beringer, J. J. Patel, and K. J. Bozic, "An overview of economic issues in computer-assisted total joint arthroplasty," *Clinical Orthopaedics and Related Research*, vol. 463, pp. 26–30, 2007.
- [15] A. D. Pearle, P. F. O'Loughlin, and D. O. Kendoff, "Robot-assisted unicompartamental knee arthroplasty," *The Journal of arthroplasty*, vol. 25, no. 2, pp. 230–237, 2010.
- [16] P. Rodrigues, M. Antunes, C. Raposo, P. Marques, F. Fonseca, and J. P. Barreto, "Deep segmentation leverages geometric pose estimation in computer-aided total knee arthroplasty," *Healthcare Technology Letters*, vol. 6, no. 6, pp. 226–230, 2019.
- [17] A. Chan, J. Aguilon, D. Hill, and E. Lou, "Precision and accuracy of consumer-grade motion tracking system for pedicle screw placement in pediatric spinal fusion surgery," *Medical engineering & physics*, vol. 46, pp. 33–43, 2017.
- [18] X. Hu, H. Liu, and F. R. y Baena, "Markerless navigation system for orthopaedic knee surgery: A proof of concept study," *IEEE Access*, 2021.
- [19] C. Yang, M. Jiang, M. Chen, M. Fu, J. Li, and Q. Huang, "Automatic 3d imaging and measurement of human spines with a robotic ultrasound system," *IEEE Transactions on Instrumentation and Measurement*, 2021.
- [20] X. Hu, F. R. y Baena, and F. Cutolo, "Head-mounted augmented reality platform for markerless orthopaedic navigation," *IEEE Journal of Biomedical and Health Informatics*, 2021.
- [21] N. Enayati, E. De Momi, and G. Ferrigno, "A quaternion-based unscented kalman filter for robust optical/inertial motion tracking in computer-assisted surgery," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 8, pp. 2291–2301, 2015.
- [22] D. Ottacher, A. Chan, E. Parent, and E. Lou, "Positional and orientational accuracy of 3-d ultrasound navigation system on vertebral phantom study," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6412–6419, 2020.
- [23] H. Suenaga, H. H. Tran, H. Liao, K. Masamune, T. Dohi, K. Hoshi, and T. Takato, "Vision-based markerless registration using stereo vision and an augmented reality surgical navigation system: a pilot study," *BMC medical imaging*, vol. 15, no. 1, pp. 1–11, 2015.
- [24] Y. Liu, Z. Song, and M. Wang, "A new robust markerless method for automatic image-to-patient registration in image-guided neurosurgery system," *Computer Assisted Surgery*, vol. 22, no. suppl, pp. 319–325, 2017.
- [25] H. Liu and F. R. Y. Baena, "Automatic markerless registration and tracking of the bone for computer-assisted orthopaedic surgery," *IEEE Access*, vol. 8, pp. 42010–42020, 2020.
- [26] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam, "Blenderproc," *arXiv preprint arXiv:1911.01911*, 2019.
- [27] B. O. Community, *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.
- [28] P. Liepa, "Filling holes in meshes," in *Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pp. 200–205, 2003.
- [29] L. R. Herrmann, "Laplacian-isoparametric grid generation scheme," *Journal of the Engineering Mechanics Division*, vol. 102, no. 5, pp. 749–907, 1976.
- [30] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 3, pp. 1–13, 2013.
- [31] S. James, A. J. Davison, and E. Johns, "Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task," in *Conference on Robot Learning*, pp. 334–343, PMLR, 2017.
- [32] J. J. Lim, H. Pirsivash, and A. Torralba, "Parsing IKEA Objects: Fine Pose Estimation," *ICCV*, 2013.
- [33] "ambientcg: Free public domain materials for physically based rendering." <https://ambientcg.com/>. Accessed: 2021-06-25.
- [34] C. Sweeney, G. Izatt, and R. Tedrake, "A supervised approach to predicting noise in depth images," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 796–802, IEEE, 2019.
- [35] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for rgb-d visual odometry, 3d reconstruction and slam," in *2014 IEEE international conference on Robotics and automation (ICRA)*, pp. 1524–1531, IEEE, 2014.
- [36] J. T. Barron and J. Malik, "Intrinsic scene properties from a single rgb-d image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 17–24, 2013.
- [37] A. Grunnet-Jepsen, J. N. Sweetser, and J. Woodfill, "Best-known-methods for tuning intel® realsense™ d400 depth cameras for best performance," *Intel Corporation: Satan Clara, CA, USA*, vol. 1, 2018.
- [38] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.
- [39] S. Sta, J. Ogor, H. Letissier, E. Stindel, C. Hamitouche, and G. Dardenne, "Towards markerless computer assisted surgery: Application to tka," *The International Journal of Medical Robotics and Computer Assisted Surgery*, p. e2296, 2021.
- [40] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.
- [41] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.
- [42] A. Martin and et al., "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283, 2016.
- [43] F. Rodriguez y Baena, T. Hawke, and M. Jakopec, "A bounded iterative closest point method for minimally invasive registration of the femur," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 227, no. 10, pp. 1135–1144, 2013.
- [44] E. Audenaert, K. De Smedt, F. Gelaude, T. Clijmans, C. Pattyn, and B. Gebelen, "A custom-made guide for femoral component positioning in hip resurfacing arthroplasty: development and validation study," *Computer Aided Surgery*, vol. 16, no. 6, pp. 304–309, 2011.