Published in final edited form as:

Magn Reson Med. 2020 January ; 83(1): 139–153. doi:10.1002/mrm.27920.

Fully Automated Patellofemoral MRI Segmentation using Holistically Nested Networks: Implications for Evaluating Patellofemoral Osteoarthritis, Pain, Injury, Pathology, and Adolescent Development

Ruida Cheng,

Biomedical Imaging Research Services Section (BIRSS), Office of Intramural Research, Center of Information Technology, NIH, Bethesda, MD, USA

Natalia A. Alexandridi,

Functional and Applied Biomechanics, Department of Rehabilitation Medicine, NIH, Bethesda, MD, USA

Richard M. Smith, M.D.,

Functional and Applied Biomechanics, Department of Rehabilitation Medicine, NIH, Bethesda, MD, USA

Aricia Shen,

Functional and Applied Biomechanics, Department of Rehabilitation Medicine, NIH, Bethesda, MD, USA and University of California Irvine School of Medicine, Irvine, CA 92617

William Gandler,

Biomedical Imaging Research Services Section (BIRSS), Office of Intramural Research, Center of Information Technology, NIH, Bethesda, MD, USA

Evan McCreedy,

Biomedical Imaging Research Services Section (BIRSS), Office of Intramural Research, Center of Information Technology, NIH, Bethesda, MD, USA

Matthew J. McAuliffe, Ph.D.,

Biomedical Imaging Research Services Section (BIRSS), Office of Intramural Research, Center of Information Technology, NIH, Bethesda, MD, USA

Frances T. Sheehan, Ph.D.

Functional and Applied Biomechanics, Department of Rehabilitation Medicine, NIH, Bethesda, MD, USA

Abstract

Corresponding author: Ruida Cheng (ruida@nih.gov).

Ethical Review: This study was approved by the IRB office of the National Institute of Child Health and Human Development.

Work Performed: All experimental data collection and analysis was done at the Functional and Applied Biomechanics section, Rehabilitation Medicine Department of the National Institutes of Health, Bethesda, MD, USA.

Conflict of Interest: None to report for any author.

Purpose: Our clinical understanding of the relationship between 3D bone morphology and knee osteoarthritis, as well as our ability to investigate potential causative factors of osteoarthritis, has been hampered by the time intensive nature of manually segmenting bone from MR images. Thus, we aim to develop and validate a fully automated deep learning framework for segmenting the patella and distal femur cortex, in both adults and actively growing adolescents.

Methods: Data from 93 subjects, obtained from an IRB-approved protocol, formed the study database. Three-dimensional sagittal GRE and GRE-FS images and manual models of the outer cortex were available for 86 femurs and 90 patellae. A deep learning based 2D holistically nested networks (HNN) architecture was developed to automatically segment the patella and distal femur using both single (sagittal, uni-planar) and three cardinal plane (tri-planar) methodologies. The errors in the surface-to-surface distances and the Dice coefficient were the primary measures used to quantitatively evaluate segmentation accuracy using a 9-fold cross-validation.

Results: The average absolute error for segmenting both the patella and femur was 0.33mm. The Dice coefficient was 97% and 94% for the femur and patella. The uni-planar, relative to the tri-planar, methodology produced slightly superior segmentation. Neither the presence of active growth plates nor pathology influenced the segmentation accuracy.

Conclusion: The proposed HNN with multi-feature architecture provides a fully automatic technique capable of delineating the often indistinct interfaces between the bone and other joint structures with an accuracy better than nearly all other techniques presented previously, even when active growth plates are present.

2. INTRODUCTION

The impact of knee osteoarthritis (OA) in terms of cost and quality of life losses cannot be overstated. Knee OA, combined with the less prevalent hip OA, ranks as the 11th highest contributor to global disability out of 291 conditions¹. It is associated with increased mortality, having a hazard ratio just slightly higher than smoking². Radiographic diagnosis of OA historically entailed visualizing joint space narrowing, osteophyte formation, subchondral cysts and/or sclerosis on plain-film radiographs³. With the rapid advancement of 3D imaging, particularly magnetic resonance (MR) imaging, evaluating OA has advanced to incorporate 3D cartilage and bone models. Yet, large clinical trials and longitudinal studies have been limited due to the time-intensive nature of manually creating 3D models.

In order to improve the diagnosis of and tracking the progression of knee OA, as well as investigating potential causative factors of OA, numerous algorithms that automatically or semi-automatically segment the tibiofemoral (TF)^{4–18} and patellofemoral (PF) cartilage^{4–6,9,12–14,18,19}; and, to a much lesser extent tibiofemoral^{5,7,8,20–24} and patellofemoral^{5,21} bone have been developed. The earliest segmentation techniques¹⁰ were semi-automatic, suffered from low accuracies, and typically required an expert user. The next generation focused strongly on active shape models^{6,12,16,25}, which were often touted as “automatic” techniques. Yet, these remained semi-automatic, as an intensive amount of manual pre-processing was required. Recent advances in convolutional neural networks (CNN) have achieved impressive breakthroughs in automatic knee segmentation from MR images, but the focus has remained primarily on delineating TF cartilage. Previous CNN

segmentation algorithms have been based on three architectures 2D patch¹⁵, 2D U-Net²⁰, and 3D U-Net²⁶. Prasoon et al.¹⁵ proposed 2D patch-based tri-planar CNN architecture for cartilage segmentation. The 2D patch consumes large hyper-parameters space during training, significantly downgrading the training and prediction efficiency. Liu et al.²⁰ developed an automatic segmentation pipeline utilizing an encoder-decoder (2D U-Net) architecture to generate the initial shape and then use a 3D deformation to generate the final 3D segmentation. The 2D U-Net employs a 10-stage encoding-decoding path, requiring excessive hyper-parameter space when training with a large dataset. 2D U-Net model's training and testing are limited to fixed size patches (e.g., 32×32 or 64×64 patches), as large patches (e.g., the entire image, 256×256 or 512×512) lead to overfitting issues and poor accuracies. In addition, scaling up to a large training dataset with 2D U-Net requires extensive hyper-parameter tuning. Ambellan et al.²⁶ combined statistical shape models with 2D and 3D U-Net to achieve high segmentation accuracy. The 3D U-Net enforces local spatial context within small image volumes (e.g., 32×32×32), however it lacks context in regard to the surrounding voxel intensity and texture feature. With the 3D convolution, the memory of 3D U-Net is too large to fit on a small general purpose graphics card (i.e. 5GB texture memory). This constrains the 3D CNN from scaling well to large datasets if GPU memory limitation is of concern.

An alternate to the CNN is the holistically nested networks (HNN), which simplifies the CNN architecture by eliminating the decoding path of the 2D U-Net, reducing the size by half. HNN was first proposed by Xie et al.²⁷, as a general edge boundary detection method on natural images. Roth et al.²⁸ adapted the HNN architecture for segmenting the interior of organs on medical images. Most HNN-based methodologies employ 2D convolutional layers in a slice-by-slice manner^{29,30}. The HNN architecture is simple a feed forward neural network capable of producing multi-scale outputs in a single path. This results in smaller hyper-parameters space in training and testing, alleviating the overfitting associated with the 2D U-Net and minimizing the GPU resources required in both training and testing. In addition, this multi-scale and multi-level nature of HNN overcomes the fixed image size limitation of 2D U-Net. A final key advantage of the HNN^{27,28}, relative to CNN, is that the training and prediction are based on the whole image (e.g., 256×256 and 512×512) end-to-end, holistically, which allows the segmentation to be performed using both local and global contextual information, which improves accuracy.

A key clinical feature that is lacking from these automatic and semi-automatic segmentation techniques, focused on the knee joint, is the ability to investigate potential precursors to OA, such as idiopathic PF pain³¹, anterior cruciate ligament injury³², patellar dislocation³³, and other injuries around/involving the knee. Unlike OA, these conditions are common in individuals from early adolescence onward and their etiology are most often associated with changes in bone morphology. Yet, only a few validated techniques^{5,7,8,20–24} are available for bone segmentation and only two studies^{5,21} report accuracies for delineating the entire patella. More importantly, no segmentation technique has been validated on immature knee joints with active growth plates. Lastly, nearly all previous techniques and their validation are based on one of two databases^{34,35}. Both databases focus on older adults with OA, lacking adolescent subjects with open femoral growth plates, and individuals with conditions that often demonstrate pathological bone morphology, with limited cartilage changes. The

open growth plates present a particular challenge because they can appear as the cortical bone-cartilage interface, creating errors for techniques that rely on pixel intensities for defining boundaries³⁶. Thus, the applicability of automatic and semi-automatic segmentation to knee joints with varying bone morphology or with open physes is unknown. Therefore, to better understand potential pre-cursors to OA, we must advance fully automated segmentation, enabling the accurate evaluation of 3D patellofemoral bone morphology in both developing and fully developed bone.

The purpose of this study was to develop and validate an automatic deep learning framework for segmenting the patellar and the distal femoral, including the distal shaft, bones using the HNN architecture. The efficiency, reliability, and holistic approach of the HNN architecture makes it ideal for our knee segmentation task. The criteria for successful implementation of the HNN architecture were that the segmentation 1) be fully automated, requiring no manual pre-processing of the data; 2) that it be validated in both mature and immature bones; 3) that it be validated for a wide range of potential bone shapes by including males and females; a range of ages; controls; and individuals with documented idiopathic knee pathology (PF pain and patellar dislocation). As part of the validation two primary questions were addressed: 1) Is segmentation accuracy the same when applied to data from subjects with active versus fused growth plates? Is the accuracy the same when applied to images from subjects with a defined knee pathology, relative to controls? 2) Can the segmentation accuracy be improved by including segmentation from all three cardinal planes (tri-planar), as opposed to a single (sagittal) cardinal plane (uni-planar)?

3. METHODS

3.1 MRI data

A database of 3D MR images was acquired from an ongoing IRB-approved NIH protocol (03-CC-0060) from April 2005-August 2017. The only inclusion criteria were that the dataset include the required images and a manually created model of the femur, patella, or both from either previous^{36,37} or ongoing studies. The final database consisted of 93 subjects (Figure 1). Consent was obtained for all adults (age ≥ 18 years) and assent was obtained for all adolescent (< 18 years) participants, with a legal guardian providing consent. Twenty-three (2 males and 21 females) datasets were from typically developing adolescents (controls, with no known serious injury, pathology, or previous surgery involving the lower limbs and 28 (2 males and 26 females) were from adolescents with a diagnosis of PFP. Only two male and four female adolescent subjects had fused femoral growth plates. Thirty-five (15 males and 20 females) datasets were from adult control participants and the remaining 7 were from adults with patellofemoral pain. One male adolescent and one female adult with patellofemoral pain had a history of dislocation. For 7 subjects (14 knees) data from both the right and left knee data were used. All researchers were blinded to the cohort associated with each dataset for the entire analysis process. Yet, it was impossible to blind the researchers to the presence of an active growth plate.

Three-dimensional sagittal GRE and GRE-FS (140 FOV, 0.27mm x 0.27mm x 1.0mm, 512 × 512 pixels) were collected for all subjects using a 3T scanner (Philips Electronics, Eindhoven, The Netherlands). In the smaller adolescents the resolution was increased by

decreasing the FOV slightly to compensate for the smaller size. To create a reference (“ground-truth”) for the evaluating auto-segmentation accuracy, the outer femoral and patellar cortices were manually segmented using MIPAV³⁸. Five different individuals were responsible for the manual segmentation. The senior author (FTS) with 15 years’ experience in creating manual musculoskeletal models reviewed all models prior to their inclusion in the segmentation atlas. The femoral segmentation was based on the GRE image set. The GRE-FS image set was consulted to ensure that the manual outline agreed with the information from both images. Similarly, the patella was segmented from the GRE-FS image set, with the GRE image set being consulted. This decision was based on the availability of models from previous studies. The GRE was the primary image used for creating femoral models, as we wanted to include the shaft and the delineation of the shaft from other structures is often indistinct in the GRE-FS images. For the patella, the GRE-FS was selected as the primary image for creating the manual model because, qualitatively, we found that there was less disagreement across researchers when segmenting the patella from the GRE-FS. The femoral and patellar atlases consisted of data from 86 and 90 subjects, based on the availability of manual models. In preparation for the training step of the project, we converted each VOI into a 3D boundary mask image set. It is important to note that the training and testing for the femur and patella were fully independent. Thus, the HNN segmentation used only a single image type. For the femur this was the GRE and for the patella this was the GRE-FS images.

3.2 Automatic Segmentation Pipeline

All images underwent an automatic pre-processing pipeline, prior to entering the HNN segmentation pipeline (Figure 2). To start, the auto-segmentation algorithm iso-tropically resampled the sagittal images. Next, it filtered the images, enhancing the edge boundary and image contrast. For training the HNN was provided with two sets of 2D slice pairs -- the isotropic MR image with the ground-truth and the isotropic Coherence Enhanced Diffusion (CED) image with the ground-truth. For testing, the ground-truth was removed. The output of the HNN model was a series of probability maps. From these probability maps, the segmentation pipeline generated a series of points delineating the outer cortical bone (volume of interest or VOI). We then converted these VOIs to 3D models for validation against the ground-truth. For our second level validation, the HNN generated VOI was converted to a binary mask for measuring the DICE score.

Image Uniformity—To create uniformity across all images, improve processing speed, and provide finer out-of-plane resolution, each 3D MR image and boundary mask set was resampled to 0.5 mm x 0.5 mm x 0.5 mm using MIPAV’s³⁸ trilinear interpolation. This resulted in a doubling of the number of images while ensuring all images consisted of 256×256 pixels.

MR images have a range of intensities across images, which can even vary across images for the same subject, using the same scan parameters on the same scanner. To curb this intensity non-uniformity, a histogram normalization scheme -- median and IQR (interquartile range) normalization was applied. The scheme subdivided each image intensity range with five points of four quartiles: min, first quartile (25%), median, third quartile (75%) and max. The

normalization computed the difference (Eq. 1) between the first and third quantiles, removing the low and high signals, which were typically associated with noise. Then, the image intensity was scaled to a range between 0 and 1000, ensuring similar intensity levels across all given datasets. To further reduce unwanted noise, a CED filter³⁹ was utilized to obtain the boundary enhanced feature image (CED image).

$$result = 1000 * \frac{I_{xyz}}{median(I) + 2 * [75\%(I) - 25\%(I)]}, \quad 1$$

where I is the intensity set of the whole 3D image after sorting and I_{xyz} is the voxel intensity.

Holistically Nested Networks (HNN) Pipeline—After the automatic pre-processing for image uniformity, we used the holistically nested networks (HNN) as the basic building block for the automatic patellofemoral bone segmentation pipeline. For our segmentation methodology, the HNN network (Figure 3) was comprised of a single stream fully convolutional neural network (FCN) with multi-layered architecture. To extract feature maps, the HNN network was divided into 5 staged convolution layers, with each layer convolving the input image using a different size and number of filters. The generated feature maps were then used as input to the next layer. Each output layer (side-output) was inserted after each convolutional layer to produce a probability map by merging the extracted feature maps at each stage. The side-outputs produced fine-to-coarse probability maps in the feed forward path. The shallow layer (side1) generated the probability map with finer details, such as messy edge boundary. The deep layer (side5) generated the probability map at a coarse level, containing more global contextual information such as the shapes of femur and patella bones. One weighted fusion layer was added to compute the average from side-output probability maps and produces the final prediction. Each side-output probability map was bilinearly interpolated into the original image size before fusion. The entire network was trained with multiple error propagation paths (Figure 3, color dashed lines) between fusion layer and each side-output layer. The error propagation paths passed the error loss computed from the objective function (see explanation in the Appendix formulation section) between fusion layer and side-output layers. The input ground truth binary mask was used to guide each side-output layer in compensating for the weak edges loss and played an important role as deep supervision. The error propagation paths in the multi-layered architecture informed each layer about the final objective of prediction, rather than relying on the final layer (Conv5) to back propagate the loss to all the shallower layers (Conv1 to 4). The single stream and feedforward fashion of the HNN architecture enhanced data processing efficiency, especially in the testing phase, by eliminating the back propagation path.

Experimentally, we found that training and testing the HNN with the MR image data alone occasionally produced unexpected irregular shapes and smaller scattered regions, primarily due to low edge signals in the MR images. To minimize these problems and increase stability, we created an enhanced HNN model using multi-feature image pairs in both the training and testing phases. Each MRI and CED image set was paired with the

corresponding ground-truth binary mask. Each image pair was a standalone basic unit and together constitute the training set. The HNN architecture holistically learned the deep representation from both MRI and the CED images, building the internal convolution neural networks between the two image features and corresponding ground truth image maps. The HNN produced per-pixel prediction map pairs (Figure 3) from each side-output layer. This allowed the HNN architecture to naturally learn the appropriate weights and biases across different image features in the training phase.

Implementation—We implemented the HNN network using the Caffe library⁴⁰. The pre-processing, post-processing, 3D surface reconstruction and mesh optimization are all implemented with the MIPAV. The proposed HNN network is fine-tuned from the initialization of the pre-trained 5 stage VGG Net⁴¹ model, which composed the 5 basic stages of the HNN architecture. Each stage block in HNN network composited from multiple convolution, ReLU, and pooling layers. The sigmoid cross-entropy loss is used to optimize the network. The overall objective loss of HNN is minimized using standard stochastic gradient descent (see appendix for further details on the loss function). The parameters of HNN network include: learning rate (10^{-7}), gamma (0.1), weight decay (0.0002), batch size (1), momentum (0.9), iteration size (10), and step size (10,000). Iteration size sets the number of epochs of the training dataset. The step size means to drop the learning rate by a factor of 10 (i.e. multiply learning rate by a factor of gamma = 0.1) every 10000 iterations. One HNN model (~100,000 training iterations) can be trained in ~2 days. During the testing phase, it takes 20 msecs to generate a single prediction map from a 2D slice image and ~ 3 min (including pre-processing, prediction and post-processing) to generate the segmentation for one test image ($512 \times 512 \times 124$), using the uni-planar method. The majority of this time is used in applying the coherence enhanced diffusion (CED) filter and converting the predicted probability maps to VOI contours. The training and testing experiments were conducted on 64 bit Ubuntu 16.04 operating system, with the following hardware configurations: Intel i7, 3.40 GHz, 64 GB RAM, one Nvidia Titan X Pascal card with 12 GB texture memory. The N-fold cross validation experiments were executed on NIH Biowulf GPU clusters with a single Nvidia k80 card. For replication purpose, the code, the model, and the results are available at https://github.com/ruida/MRI_knees_seg.

Automatic Tri-planar Segmentation Pipeline—To potentially provide improved accuracy, we expanded the HNN methodology by incorporating a tri-planar volumetric approach. Immediately after completing of the automatic pre-processing pipeline (Figure 2), we transformed the sagittal images and corresponding binary masks into axial and coronal view images. The image sets in each cardinal plane were paired with binary masks to train a standalone HNN deep learning model. During testing phase, the test image was pre-processed into the three cardinal (axial, sagittal, and coronal) views. Then, each cardinal image set was run against the training model to generate the VOI contours. The three automatically segmented VOI contours were merged together to formulate the final tri-planar auto-segmented VOI.

3.3 Validation of the uni- and tri-planar segmentation

To maximize the information provided to the training pipeline, we validated the automatic segmentation using a 9-fold cross-validation. We randomly split the dataset into 9 groups (folds). For each unique validation, one group (10 datasets) was assigned as the testing set, and the remaining groups as the training set (80 datasets). For the final femoral validation, only 6 datasets were allocated to the testing set. We tested the quality of these uni- and tri-planar automatic segmentations using both direct comparison (the point-to-surface error between the automatically and manually generated bone models) and DICE similarity.

Direct comparison—We defined the accuracy of the uni-planar and tri-planar automatic segmentations as the distance between the points defining the bone surface, derived through auto-segmentation, to surface of the 3D manually segmented model. To start the analysis, we imported each subject's automatically and manually generated VOIs into Geomagic Wrap (3D systems, Research Triangle Park, NC). The tri-planar segmentation created VOIs with approximately 3 times the number of points as the uni-planar, thus we reduced the number of points of the tri-planar VOIs by approximately one-third using Geomagic's uniform sampling algorithm, insuring that the number of points in the tri-planar model was within 10% of the uni-planar model. Next, as done in previous studies^{37,42} models were created from the VOIs by wrapping a surface mesh to the points. For the femur, we cut the model using an axial plane (as defined by the MR coordinate system) at the most superior location that contained a complete wrap for the shaft. We then converted the automatically generated (uni- and tri-planar) models to point clouds, allowing us to calculate the distances from all points representing the automatically generated model to the manually generated model. Comparing surface-to-surface generated identical results, but the point-model improved visualization. To facilitate comparisons to previous and future studies, the errors between the automatically and manually generated bone models were reported as an average error, an average absolute error, and an RMS error.

DICE Similarity Comparison—For the Dice comparison, the tri-planar segmentation surface was converted to 2D sagittal plane binary masks. The first step in this process was to create high density smoothed 3D surfaces of the femur or patella, utilizing Ball-pivoting⁴³ and Poisson surface reconstruction⁴⁴, and VTK Laplacian smoothing algorithm⁴⁵. This model was then resliced to create binary masks corresponding to the sagittal planes of the high-resolution MR dataset. Both uni-planar and tri-planar segmentation binary masks were compared with the ground truth binary mask for the dice score (EvaluateSegmentation tool⁴⁶). Five metrics are used, 1) Dice similarity coefficient (DSC), 2) Jaccard index (IoU), 3) Hausdorff distance (HD, mm), 4) Average minimum distance (AD, mm), and 5) Volumetric similarity coefficient (VSC).

3.4 Secondary Validations

We investigated the impact of preprocessing on segmentation by re-running the segmentations on the complete in-house dataset with just the HNN network. The preprocessing pipeline (isotropic sampling, CED filter, and normalization) was omitted. Performance was defined using the Dice similarity comparison (DSC, IoU, HD, AD, VSC).

To further validate the effectiveness of the HNN model, we validated the femoral segmentation using an external public database (SKI10)⁴⁷. Due to issues associated with the online access to the database, only the 100 training datasets were used in the 10-fold validation inactive issues. Performance was defined using the Dice similarity comparison (DSC, IoU, HD, AD, VSC).

3.5 Statistics

After assuring data homogeneity, a Student's t-test was used to compare the accuracies and Dice scores between the cohort with open growth plates and the cohort with fused growth plates, as well as the control cohort and the cohort with knee pathologies. A paired Student's t-test was used to compare accuracy of the uni- and tri-planar approaches, as well as, to compare the accuracy of segmentation with and without use of the pre-processing step. Finally, a Student's t-test was used to determine if the accuracy of segmenting was different between the in-house dataset and the SKI10 dataset. A $p < 0.05$ was considered a significant difference.

4. RESULTS

We successfully created a fully automatic deep learning framework for segmenting the patella and distal femur using holistically nested networks (HNN) architecture. The average absolute error for segmenting the patella and femur was 0.33mm for both bones (Table 1). The Dice similarity coefficient was 97% and 94% for the femur and the patella (Table 2). The errors were homogeneously distributed (Figure 4) and the error patterns were similar across all subjects (Figures 5 and 6). For the patella, the highest errors tended to be on the outer rim of the patella, with the lowest errors being at the anterior and posterior (bone-cartilage interface) surfaces (Figure 5). The auto-segmentation diverged from the manual on the medial and lateral edges of the patella due to the rapid changes in contours between images. They diverged at the most posterior-superior aspect of the patella due to the sharp contour changes in this region. Finally, the image contrast between the central inferior patella and the joint space was highly variable across subjects, creating focal errors in this region. Similar to the patella, the femur had higher errors on the medial and lateral sides of the bone with the lowest errors at the bone-cartilage interface (Figure 6). The other region of error was the edges of the posterior notch due to rapid changes in the bone contours in this region.

There were no significant differences in the accuracy of automatically segmenting immature bone, relative to mature bone (Table 3, Figure 6). The error was evenly distributed across segmentations for immature and mature bone (Figure 4). There were no differences in the accuracy of segmenting data from the control cohort, relative to the cohort with diagnosed knee pathologies (Table 3). Similarly, the error was evenly distributed across both cohorts.

In general, the uni-planar segmentation out-performed the tri-planar. The tri-planar approach did significantly improve measures of point-to-surface accuracy in terms of average error (both bones) and the average absolute error for the patella (Table 2). Ultimately, these differences were too small to be clinically relevant.

Removing the pre-processing step significantly reduced the accuracy, with the exception that the DSC, which slightly improved for the patella (Table 2). Although the segmentation accuracy was significantly better for the in-house dataset, relative to the SKI10 (Table 2), the differences were too small to be clinically relevant.

5. Discussion

The successful implementation of the proposed HNN segmentation pipeline provides a crucial step forward in our ability to evaluate patellofemoral bone morphology. This technique supports future advances in both clinical and engineering based research studies by providing a fully automated technique that is able to decipher the often indistinct interfaces between the bone and other joint structures (e.g., tendon, ligament, muscle, etc) with an accuracy better than nearly all other techniques presented previously^{5,7,8,20–24} (Table 4). Our segmentation methodology is currently the only technique that can automatically, without any manual pre-processing steps, segment the complete patellofemoral bones in both mature and actively growing bone. The ability to properly interpret the growth plates as an internal bone structure will enable studies beyond those focused solely on adults. This in turn will expand our knowledge of bone development and pathology in the pediatric/adolescent population⁴⁸, allowing us to better understand the etiology of pathology and the impact of injury in this population, as well as to investigate potential precursors to OA.

The HNN model we developed addresses many of the challenging issues in segmenting bone, which often has indistinct demarcation from other tissues within the joint in MR images. Our implementation was supported by a relatively large training and testing database (Table 4). Beside the robust, fully-automatic segmentation performance, the HNN model has the distinct advantages over other traditional methods^{5,7,8,21–24}. The HNN model is an end-to-end training and prediction system that integrates the edge enhanced multi-feature into the HNN network to naturally learn the enhanced edge boundaries during training phase, improving the robustness and accuracy of segmentation. It aggregates the multi-scale and multi-level deep convolutional neural networks to capture the complex variation of femoral and patellar bone shape and texture appearance in MR images. The model is image size agnostic, due to multi-scale and isotropic sampling step in the pre-processing. Thus, there is no need to recreate a training set when the resolution, size, and (potentially) contrast levels change in the testing set. This pre-processing improved accuracy by stabilizing the segmentation, while only minimally increasing processing time. Specifically, when the HNN pipeline was used without the pre-processing step, the generated segmentations contained scattered small regions well outside the bone, resulting in the extremely high HD numbers observed. However, even when the pre-processing was removed, the accuracy of the current segmentation model remained superior to previous studies (Table 4). As both the uni- and tri-planar methods are 2D segmentations, they are computationally efficient and scale well to large datasets and fit onto a general purpose graphics card.

To the best of our knowledge, only two studies^{5,21} presented accuracies for segmenting the full patellar bone, as the focus has been primarily on segmenting the tibiofemoral bone and cartilage surfaces^{4,9–11,13–15,17–19}. One study on patellar bone segmentation reported⁵ more

accurate segmentation (Table 4) than the current. Yet, this method is only semi-automatic, requiring a time-intensive manual “pre-segmentation” of the images. The active shape modeling methodology presented by Fripp and colleagues²¹ is fully automatic. They avoided the manual pre-processing step by using an affine transformation between the test image and the reference image. Unfortunately, Fripp and colleagues limited their validation to a database of 20 knees and the methodology was unsuccessful in completing the patellar segmentation for 9/20 datasets. The current study did not have such limitations. Further, previous studies evaluating the semi- or automatic segmentation of PF cartilage^{1,5,9–11,13,14,17,18} reported accuracies (DICE similarity coefficients ranged from 63% to 84%) that were inferior to the current results for PF bone segmentation. The stability of the HNN architecture, its ability to capture texture variation across the full image context, and the large training set likely promoted the superior performance of the HNN model.

The proposed HNN with multi-feature architecture also outperformed current state-of-the-art semi- or fully-automatic MR-based femoral and tibial bone segmentation methods. It did so with a unique dataset that expands the type of bone that can be studied in the future. The majority of the studies that were comparable to the current (Table 4) used the OAI database. This was not feasible for the current study, as the OAI database did not contain immature bone and does not provide ground truth models. However, the imaging used for the OAI dataset was similar to the current in that it was acquired using a 3T scanner with a resolution ($0.3625 \times 0.3625 \times 0.7\text{mm}$) comparable to the current ($0.5 \times 0.5 \times 1.0\text{mm}$, resampled from a captured resolution of $0.27 \times 0.27 \times 1\text{mm}$). There was a significant, but extremely small degradation in the accuracy when the HNN model was applied to the SKI10 dataset. Thus, it is unlikely that variations in image quality were the source for the differences in accuracies between the current and previous studies. The similarity in accuracy for segmenting images with active growth plates versus mature bone not only opens up avenues for the study of adolescent bone, it indicates that this technique will likely be easily adaptable to a variety of image types, as the internal bone signal varied greatly across immature and mature bone. This capacity to apply the current methodology to other image types is supported by the automatic pre-processing step that fosters a uniformity across images.

Deriving accurate results based on two distinct image types is a strength of the current study, as this demonstrates algorithm is capable of segmenting multiple image types. Three-dimensional sagittal GRE and GRE-FS (140 FOV, $0.27\text{mm} \times 0.27\text{mm} \times 1.0\text{mm}$, acquisition matrix 232×220 pixels, image size: 512×512 pixels) were collected for all subjects using a 3T scanner (Phillips Electronics, Eindhoven, The Netherlands). Using the GRE images for the femoral segmentation and the GRE-FS images for the patellar segmentation required approximately ~10 minutes of scan time to acquire both data sets. As with other techniques, the accuracy of delineating the patellar and femoral bones on image types other than the ones we used remains unknown. For the femoral segmentation, the contrast between the bone and musculotendon structures on the GRE image allowed for accurate delineation of the bone from other knee structures (e.g. ACL, PCL, muscle, etc.) and for an accurate delineation of the shaft. The use of the fat-suppressed image to create the manual patellar models demonstrated that current algorithm could accurately segment the superior and inferior patellar bone from the quadriceps and patellar tendons (Figure 5), even when minimal to no contrast existed between these structures. Due to the contrast of the fat-

suppressed images, the manual models for the current, as with previous patellar bone segmentation studies^{5,21}, incorporated the tendon crossing over the anterior patella into the patellar bone. Biologically, a distinct anatomical separation between these two structures does not exist, as the tendon interdigitates with the bone. Thus, the accuracy of automatically separating the anterior ligament from the patella was not tested.

We focused our study on bone modeling, based on our research interests. As the errors were consistently lowest at the bone-cartilage interface, the current technique shows the potential to expand into cartilage segmentation. As the current work does not directly address the accuracy of segmenting cartilage, future work is needed to determine the accuracy of segmenting cartilage using the current HNN architecture. Although the multi-scale and isotropic sampling step provided benefits to the segmentation, it did falsely increase errors in some subjects. This sampling step doubles the number of slices for auto-segmentation. Yet, the manual segmentation was completed on the original dataset. Thus, the manual segmentation only created outlines for the odd images in the re-sampled images, whereas the auto-segmentation produced outlines for every image. This often resulted in an extra outline 0.5mm beyond the manual segmentation, creating a false error (Figure 6, bottom row, image 166). The tri-planar segmentation did not provide uniform improvements in segmentation accuracy. The lower DICE scores for the femoral tri-planar segmentation were primarily due to the over-smoothing involved in transitioning the tri-planar point cloud back to binary masks for evaluation. Regardless, at the current time, the faster uni-lateral approach remains the superior technique. However, the tri-planar approach may prove useful for segmentation where the uni-planar approach cannot produce the needed precision in all cardinal planes.

In conclusion, this work advances the field of auto-segmentation of MR images by developing a fully automated technique capable of deciphering the often indistinct interfaces between the bone and other joint structures. This methodology is the only one to demonstrate fully automatic segmentation (without any manual pre-processing steps) that can segment both mature bone and bone with active growth plates. The accuracy is better than nearly all previous techniques. This opens future research paths in the crucial topic of the relationship of bone morphology, pathology, and the long term bone-cartilage health. Moving forward, we are interested in the generalization of applying deep transfer learning methods across different MR scanning protocols. We hope that training a single unified deep learning model can be generalized to vast knee segmentation tasks to aid in the diagnosis and understanding of knee pathologies.

Acknowledgements:

We would like to thank the Radiology Department at the Clinical Center of the National Institutes of Health for their support of this work. This work was funded by the Intramural Research Program of the National Institutes of Health Clinical Center and the National Institute of Biomedical Imaging and Bioengineering Intramural Program, Bethesda, MD, USA. This research was also made possible through the NIH Medical Research Scholars Program, a public-private partnership supported jointly by the NIH and generous contributions to the Foundation for the NIH from Pfizer, the Leona M. and Harry B. Helmsley Charitable Trust, and the Howard Hughes Medical Institute, as well as other private donors. For a complete list, visit the website at <http://fnih.org>. This work was supported in part by Intramural Research Program of National Institutes of Health. We thank NVIDIA for Titan X GPU donation. We also appreciate NIH Biowulf staffs to support us to run experiments on Biowulf GPU clusters, which employed Nvidia K20x, K80 graphics cards.

8. APPENDIX

HNN Formulation

To formulate the HNN architecture, we followed the notation of Xie et al.³². We denoted the training set as $S = \{(X_n, Y_n), n = 1, \dots, N\}$. X_n is the raw input images (MRI or CED pairs), Y_n is the corresponding ground truth label map. The whole objective is to learn the network hyper-parameters (W) which enabled binary mask detection at multi-scale. The HNN networks had $M(5)$ side-output layers. Each output had an associated classifier with corresponding weight $W = (w^{(1)}, w^{(2)}, \dots, w^{(M)})$. We defined the objective loss function of the side-output layers as:

$$\mathcal{L}_{side}(W, w) = \sum_{m=1}^M a_m l_{side}^{(m)}(W, w^{(m)}), \quad 1$$

where $l_{side}^{(m)}$ is the cross-entropy loss for the m^{th} side-output. We computed the loss function over all pixels in the image (X) and label map (Y). $\mathcal{L}_{side}$ denoted an image-level loss function for side-outputs. We calculated the femur and patella interior map prediction $\hat{Y}_{side} = \alpha(\hat{A}(m)_{side})$ as each side-output layer, where $\hat{A}(m)_{side}$ are the prediction map from side-output layer m . We defined the loss function at the fusion layer $\mathcal{L}_{fuse}$ as:

$$\mathcal{L}_{side}(W, w, h) = Dist(Y, \hat{Y}_{fuse}) \quad 2$$

Where $\hat{Y}_{fuse} \equiv \alpha\left(\sum_{m=1}^M h_m \hat{A}_{side}^{(m)}\right)$ with $h = (h_1, \dots, h_m)$ being fusion weight. $Dist(\cdot, \cdot)$ was the distance between fused predictions and ground truth label map, which is the cross-entropy loss. The overall objective loss function of HNN is minimized using standard stochastic gradient descent.

$$\mathcal{L}_{HNN}(W, w, h) = \operatorname{argmin}(\mathcal{L}_{side}(W, w) + \mathcal{L}_{fuse}(W, w, h)) \quad 3$$

6. REFERENCES

1. Cross M, Smith E, Hoy D, et al. The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. *Ann Rheum Dis*. 2014;73(7):1323–1330. [PubMed: 24553908]
2. Kasai T, Hasegawa Y, Imagama S, et al. The impact of musculoskeletal diseases on mortality-comparison with internal diseases: A 15-year longitudinal study. *J Orthop Sci*. 2017;22(6):1126–1131. [PubMed: 28754502]
3. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthritis. *Ann Rheum Dis*. 1957;16(4):494–502. [PubMed: 13498604]

4. Ahn C, Bui TD, Lee YW, Shin J, Park H. Fully automated, level set-based segmentation for knee MRIs using an adaptive force function and template: data from the osteoarthritis initiative. *Biomedical engineering online*. 2016;15(1):99. [PubMed: 27558127]
5. Yin Y, Zhang X, Williams R, Wu X, Anderson DD, Sonka M. LOGISMOS--layered optimal graph image segmentation of multiple objects and surfaces: cartilage segmentation in the knee joint. *IEEE transactions on medical imaging*. 2010;29(12):2023–2037. [PubMed: 20643602]
6. Baldwin MA, Langenderfer JE, Rullkoetter PJ, Laz PJ. Development of subject-specific and statistical shape models of the knee using an efficient segmentation and mesh-morphing approach. *Computer methods and programs in biomedicine*. 2010;97(3):232–240. [PubMed: 19695732]
7. Dodin P, Martel-Pelletier J, Pelletier JP, Abram F. A fully automated human knee 3D MRI bone segmentation using the ray casting technique. *Medical & biological engineering & computing*. 2011;49(12):1413–1424. [PubMed: 22038239]
8. Lee JG, Gumus S, Moon CH, Kwok CK, Bae KT. Fully automated segmentation of cartilage from the MR images of knee using a multi-atlas and local structural analysis method. *Medical physics*. 2014;41(9):092303.
9. Dam EB, Lillholm M, Marques J, Nielsen M. Automatic segmentation of high- and low-field knee MRIs using knee image quantification with data from the osteoarthritis initiative. *J Med Imaging (Bellingham)*. 2015;2(2):024001.
10. Folkesson J, Dam EB, Olsen OF, Pettersen PC, Christiansen C. Segmenting articular cartilage automatically using a voxel classification approach. *IEEE transactions on medical imaging*. 2007;26(1):106–115. [PubMed: 17243589]
11. Liukkonen MK, Mononen ME, Tanska P, Saarakkala S, Nieminen MT, Korhonen RK. Application of a semi-automatic cartilage segmentation method for biomechanical modeling of the knee joint. *Computer methods in biomechanics and biomedical engineering*. 2017;20(13):1453–1463. [PubMed: 28895760]
12. Neogi T, Bowes MA, Niu J, et al. Magnetic resonance imaging-based three-dimensional bone shape of the knee predicts onset of knee osteoarthritis: data from the osteoarthritis initiative. *Arthritis and rheumatism*. 2013;65(8):2048–2058. [PubMed: 23650083]
13. Norman B, Pedoia V, Majumdar S. 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*. 2018:172322.
14. Ozturk CN, Albayrak S. Automatic segmentation of cartilage in high-field magnetic resonance images of the knee joint with an improved voxel-classification-driven region-growing algorithm using vicinity-correlated subsampling. *Computers in biology and medicine*. 2016;72:90–107. [PubMed: 27017069]
15. Prasoon A, Petersen K, Igel C, Lauze F, Dam E, Nielsen M. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. *Medical image computing and computer-assisted intervention : MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention* 2013;16(Pt 2):246–253.
16. Solloway S, Hutchinson CE, Waterton JC, Taylor CJ. The use of active shape models for making thickness measurements of articular cartilage from MR images. *Magnetic resonance in medicine*. 1997;37(6):943–952. [PubMed: 9178247]
17. Tamez-Pena JG, Farber J, Gonzalez PC, Schreyer E, Schneider E, Totterman S. Unsupervised segmentation and quantification of anatomical knee features: data from the Osteoarthritis Initiative. *IEEE transactions on bio-medical engineering*. 2012;59(4):1177–1186. [PubMed: 22318477]
18. Zhang K, Lu W, Marziliano P. Automatic knee cartilage segmentation from multi-contrast MR images using support vector machine classification with spatial dependencies. *Magnetic resonance imaging*. 2013;31(10):1731–1743. [PubMed: 23867282]
19. Carballido-Gamio J, Majumdar S. Atlas-based knee cartilage assessment. *Magnetic resonance in medicine*. 2011;66(2):574–583. [PubMed: 21773988]
20. Liu F, Zhou Z, Jang H, Samsonov A, Zhao G, Kijowski R. Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. *Magnetic resonance in medicine*. 2018;79(4):2379–2391. [PubMed: 28733975]

21. Frupp J, Crozier S, Warfield SK, Ourselin S. Automatic segmentation of the bone and extraction of the bone-cartilage interface from magnetic resonance images of the knee. *Physics in medicine and biology*. 2007;52(6):1617–1631. [PubMed: 17327652]
22. Ababneh SY, Prescott JW, Gurcan MN. Automatic graph-cut based segmentation of bones from knee magnetic resonance images for osteoarthritis research. *Medical image analysis*. 2011;15(4):438–448. [PubMed: 21474362]
23. Gandhamal A, Talbar S, Gajre S, Razak R, Hani AFM, Kumar D. Fully automated subchondral bone segmentation from knee MR images: Data from the Osteoarthritis Initiative. *Computers in biology and medicine*. 2017;88:110–125. [PubMed: 28711767]
24. Kashyap S, Oguz I, Zhang H, Sonka M. Automated Segmentation of Knee MRI Using Hierarchical Classifiers and Just Enough Interaction Based Learning: Data from Osteoarthritis Initiative. *Medical image computing and computer-assisted intervention : MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention* 2016;9901:344–351.
25. Fitzpatrick CK, Baldwin MA, Laz PJ, FitzPatrick DP, Lerner AL, Rullkoetter PJ. Development of a statistical shape model of the patellofemoral joint for investigating relationships between shape and function. *Journal of biomechanics*. 2011;44(13):2446–2452. [PubMed: 21803359]
26. Ambellan F, Tack A, Ehlke M, Zachow S. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative. *Medical image analysis*. 2019;52:109–118. [PubMed: 30529224]
27. Xie SN, Tu ZW. Holistically-Nested Edge Detection. *International Journal of Computer Vision*. 2017;125(1–3):3–18.
28. Roth HR, Lu L, Lay N, et al. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. *Medical image analysis*. 2018;45:94–107. [PubMed: 29427897]
29. Cheng RD, Lay N, Mertan F, et al. Deep Learning with Orthogonal Volumetric Hed Segmentation and 3d Surface Reconstruction Model of Prostate Mri. 2017 Ieee 14th International Symposium on Biomedical Imaging (Isbi 2017) 2017:749–753.
30. Cheng RD, Roth HR, Lay N, et al. Automatic magnetic resonance prostate segmentation by deep learning with holistically nested networks. *J Med Imaging*. 2017;4(4).
31. Utting MR, Davies G, Newman JH. Is anterior knee pain a predisposing factor to patellofemoral osteoarthritis? *The Knee*. 2005;12(5):362–365. [PubMed: 16146626]
32. Lecoq FA, Parienti JJ, Murison J, et al. Graft Choice and the Incidence of Osteoarthritis After Anterior Cruciate Ligament Reconstruction: A Causal Analysis From a Cohort of 541 Patients. *The American journal of sports medicine*. 2018;46(12):2842–2850. [PubMed: 30199646]
33. Salonen EE, Magga T, Sillanpaa PJ, Kiekara T, Maenpaa H, Mattila VM. Traumatic Patellar Dislocation and Cartilage Injury: A Follow-up Study of Long-Term Cartilage Deterioration. *The American journal of sports medicine*. 2017;45(6):1376–1382. [PubMed: 28298062]
34. Nevitt M, Felson DT, Lester G. The osteoarthritis initiative protocol for the cohort study. In. <https://oai.epi-ucsf.org/datarelease/2006>.
35. Heimann T, Morrison B, Warfield SK, Styner M, Neithammer M. Segmentation of Knee Images 2010. In:2010.
36. Smith RM, Boden BP, Sheehan FT. Increased Patellar Volume/Width and Decreased Femoral Trochlear Width Are Associated With Adolescent Patellofemoral Pain. *Clinical orthopaedics and related research*. 2018;476(12):2334–2343. [PubMed: 30422967]
37. Borotikar BS, Sipprell WH 3rd, Wible EE, Sheehan FT. A methodology to accurately quantify patellofemoral cartilage contact kinematics by combining 3D image shape registration and cine-PC MRI velocity data. *Journal of biomechanics*. 2012;45(6):1117–1122. [PubMed: 22284428]
38. McAuliffe MJ, Lalonde FM, McGarry D, Gandler W, Csaky K, Trus BL. Medical image processing, analysis and visualization in clinical research. *14th IEEE Symposium on Computer-Based Medical Systems*; 2001.
39. Weickert J Coherence-Enhancing Diffusion Filtering. *International Journal of Computer Vision*. 1999;31(2):111–127.
40. Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. 2014.

41. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014.
42. Smith R, Boden B, Sheehan F. Increased Patellar Volume/Width and Decreased Femoral Trochlear Width are Associated with Female Adolescent Patellofemoral Pain. *Clinical Orthopaedics and Related Research*. in press.
43. Bernardini F, Mittleman J, Rushmeier H, Silva C, Taubin G. The ball-pivoting algorithm for surface reconstruction. *Ieee T Vis Comput Gr*. 1999;5(4):349–359.
44. Kazhdan M, Hoppe H. Screened Poisson Surface Reconstruction. *Acm T Graphic*. 2013;32(3).
45. Schroeder w, Martin k, Lorensen b. *Visualization Toolkit: An Object-Oriented Approach to 3D Graphics*, 4th Edition Pearson Education, Inc; 2006.
46. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15:29. [PubMed: 26263899]
47. Heimann T J Morrison B, Styner M, Niethammer M, Warfield S. Segmentation of Knee Images: A Grand Challenge. 2010.
48. Sheehan FT, Brainerd EL, Troy KL, Shefelbine SJ, Ronsky JL. Advancing quantitative techniques to improve understanding of the skeletal structure-function relationship. *J Neuroeng Rehabil*. 2018;15(1):25. [PubMed: 29558970]
49. Raynauld JP, Martel-Pelletier J, Bias P, et al. Protective effects of licofelone, a 5-lipoxygenase and cyclo-oxygenase inhibitor, versus naproxen on cartilage loss in knee osteoarthritis: a first multicentre clinical trial using quantitative MRI. *Ann Rheum Dis*. 2009;68(6):938–947. [PubMed: 18653484]

Figure 1. Demographic constitution of database.

In total data for 93 knees were included in the current study. In the final column the average demographics for each subgroup is provided with the standard deviation in parentheses.

Abbreviations: years old (yo);PFP (patellofemoral pain); and NA – not applicable.

Figure 2. Proposed automatic segmentation pipeline for the uni-planar approach.

The same pathway is followed for the tri-planar approach, by applying the same pathway to axial and coronal images.

Figure 3.
HNN architecture for bone segmentation (femoral example provided)

Distribution of the Average Absolute Error for the Patella

Distribution of the Average Absolute Error for the Femur

Figure 4: Distribution of error across study cohort.

The x-axis denotes the subject's placement in the error hierarchy (1 being the lowest error). Thus, the identical number in the top (patellar error) and the bottom (femoral error) does not denote the same subject. The bars indicate the value of the average absolute error for each subject. The circles indicate errors associated with adolescents with active growth plates and the squares indicate errors associated with control subjects. The stars denote subject's for which the visual representation of the manual and automatic segmentations is provided (Figure 5 and 6)

Figure 5: Visual Interpretation of Auto-Segmentation Errors for the Patella:

The subject with the lowest error (6196L) and the 2nd highest error (3002R) are shown. The highest error subject was not used, as their error stood apart from the group due to an odd signal in the patella, which caused the auto-segmentation to miss the bone-cartilage interface, which did not happen in any other subject. All other subjects had similar error patterns. The images begin with the left most manual segmentation and then are spaced every 7mm. The last image is the final segmentation. The systematic selection of images prevented us selecting the images demonstrating only the most favorable results. The manual outline (ground truth) is in red, with the auto-segmentation in yellow. The 3D model on the right demonstrates the point errors, as displayed in Geomagic. The viewpoint is from posterior to anterior. Pure red indicates errors of +2.00 mm and pure blue indicate errors of

–1.3 mm. The red arrows denote the image planes for each image (left to right). The red and blue spheres denote the location of the maximum and minimum errors. The yellow arrows indicate areas of highest errors.

Figure 6: Visual Interpretation of Auto-Segmentation errors for the Femur:

All subjects had similar error patterns. Thus, the images for subject 5708L were used, instead of the lowest error subject, to demonstrate a segmentation in the presence of active growth plates. Subject 6205L was selected, instead of the highest error, to demonstrate the algorithm's ability to correctly segment, even in the presence of pathological shape (e.g., image 85). The images begin with the left most manual segmentation and then are spaced every 10mm. The last image is the final segmentation. The systematic selection of images prevented us selecting the images demonstrating only the most favorable results. The manual outline (ground truth) is in red, with the auto-segmentation in yellow. The largest errors are denoted with yellow arrows. The 3D model on the right demonstrates the point errors, as displayed in Geomagic. The viewpoint is from anterior to posterior. Pure red indicates errors of +2.00 mm and pure blue indicates errors of -1.3 mm. The red arrows denote the image planes for each image (right to left). The red and blue spheres denote the maximum and minimum errors. The yellow arrows indicate areas of highest errors.

Table 1:
Point-to-Surface Accuracy:

The top half of the table compares the uni- with tri-planar methods for the femur, whereas the bottom half does this for the patella. Where appropriate the standard deviation is provided in parenthesis. The p-value associated with the differences between the uni- and tri-planar techniques are provided. The result represents the best result for each bone in each column when comparing uni- versus tri-planar is bolded. If both techniques are equivalent, nothing is bolded for that column.

			# Pts	Average Error	RMS error	Average Absolute Error
Femur	uni	mean (SD)	6765(652)	0.24 (0.07)	0.44 (0.06)	0.33(0.04)
	tri	mean (SD)	6544(715)	0.22 (0.08)	0.45(0.05)	0.34(0.04)
	p			<0.001	0.013	0.026
Patella	uni	mean (SD)	2040(234)	0.17(0.15)	0.41(0.09)	0.33(0.06)
	tri	mean (SD)	2345(253)	0.08(0.16)	0.41(0.08)	0.32(0.06)
	p			0.007	0.448	<0.001

Table 2.
Comparison of Segmentations.

The mean is provided with 1 standard deviation in parenthesis. The 2 rows of p-values provided in the top portion of the table, “*with pre-processing (NIH data)*”, indicated the results of a paired Student’s t-test comparing the uni- and tri-planar approaches. The blue color in this section indicates the best results when comparing the uni- to the tri planar. The p-values provided in the middle portion of the table, “*without pre-processing (NIH data)*”, indicated the results of a paired Student’s t-test comparing the combine pre-processing and HNN network approach (top of table) to the isolated HNN approach. The one blue result indicates the one result that improved when the pre-processing was removed. The p-values provided in the bottom portion of the table, “*with pre-processing (SKI10 data)*”, indicate the results of a Student’s t-test comparing the segmentation of the SKI10 dataset⁴⁷ and the NIH dataset (top of the table) using the combined pre-processing and HNN pipeline approach. NS – not significant; Uni: uni-planar segmentation; Tri: tri-planar segmentation; Fem: femur, Pat: patella. NIH: National Institutes of Health (indicates in house database); DSC: Dice similarity coefficient; IoU: Jaccard index; HD: Hausdorff distance; AD: Average minimum distance; VSC: Volumetric similarity coefficient, and NS: not significant.

			DSC (%)	IoU (%)	HD(mm)	AD(mm)	VSC(%)
<i>with pre-processing (NIH data)</i>							
Fem	uni	Mean	97.19(0.54)	94.54(1.02)	6.85(2.49)	0.03(0.01)	94.54(1.02)
	tri	Mean	95.57(1.08)	91.54(1.95)	9.26(2.20)	0.08(0.04)	98.08(0.88)
	p		<0.001	<0.001	<0.001	<0.001	<0.001
Pat	uni	Mean	94.39(1.33)	89.41(2.36)	5.37(1.84)	0.07(0.02)	95.63(1.99)
	tri	Mean	94.68(1.13)	89.91(2.01)	6.87(3.42)	0.07(0.03)	97.19(1.70)
	p		<0.001	<0.001	<0.001	NS	<0.001
<i>without pre-processing (NIH data)</i>							
Fem	uni	Mean	95.92(1.95)	92.24(3.45)	110(51)	0.29(0.63)	97.10(1.77)
	p		<0.001	<0.001	<0.001	<0.001	<0.001
Pat	uni	Mean	95.18(1.27)	90.83(2.27)	9.79(15.59)	0.06(0.02)	95.77(1.59)
	p		<0.001	<0.001	<0.001	<0.001	<0.001
<i>with pre-processing (SKI10 data)</i>							
Fem	uni	Mean	97.39(0.77)	94.93(1.44)	9.47(6.04)	0.04 (0.03)	98.98(0.84)
	p		0.041	0.037	<0.001	NS	<0.001

Table 3:
Point-to-Surface Accuracy Across Cohorts:

All results are for the uni-planar segmentation, using the complete pipeline. The top half of the table lists femoral values and the bottom patellar. The accuracy, expressed as an average absolute error (AAE), was compared between all subjects with fused and unfused femoral growth plates, along with a comparison between controls and patients, with the p-value provided. Where appropriate the standard deviation is provided in parenthesis.

			Fused	Unfused	p-value	Patient	Control	p-value
Femur	AAE	mean (SD)	0.33 (0.04)	0.34 (0.04)	0.280	0.34 (0.04)	0.33 (0.04)	0.161
	Dice	mean (SD)	0.97 (0.01)	0.97 (0.00)	0.595	0.97 (0.01)	0.97 (0.00)	0.053
Patella	AAE	mean (SD)	0.32 (0.07)	0.34 (0.05)	0.068	0.32(0.06)	0.34 (0.06)	0.211
	Dice	mean (SD)	0.95 (0.01)	0.94 (0.01)	0.065	0.95 (0.01)	0.94 (0.01)	0.070

Table 4:
Previous Patellar and Femoral Bone Segmentation Accuracies.

Comparison of existing segmentation methods capable of femoral and/or patellar bone segmentations: The Dice similar coefficient (DSC); the root mean square errors (RMS) and average absolute errors (AAE) comparing the automatically and manually segmented models are listed, as provided by each study. ^ Used a leave-one-out cross-validation²¹ or a 9-fold (current, NIH) or 10-fold (current, SKI10) cross-validation (current). The final two rows (current) used the full uni-planar algorithm (pre-processing and HNN network)

<i>Authors</i>	<i>Algorithm</i>	DATA			DSC (%)		RMS (MM)		AAE (MM)	
		Train	Test	source	Femur	Patella	Femur	Patella	Femur	Patella
<i>Ababneh</i> ²²	graph-cut	50	190	OAI	93					
<i>Dodin</i> ⁷	ray-casting	0	161	prev. study ⁴⁹	94				0.5	
<i>Fripp</i> ²¹	3D ASM	20	20^	in house	95	86				
<i>Gandham</i> ²³	level-set	8	8	OAI	90		2.1		1.43	
<i>Kashyap</i> ²⁴	LOGISMOS	28	53	OAI					0.55	
<i>Lee</i> ⁸	multi-atlas	100	50	SKI10			1.05		0.63	
<i>Liu</i> ²⁰	2D U-Net	60	40	SKI10			1.08		0.56	
<i>Yin</i> ⁵	LOGISMOS	9	20	OAI					0.22	0.23
<i>current</i>	HNN	80	90^	NIH	97	96	0.44	0.41	0.33	0.33
<i>current</i>	HNN	90	100^	SKI10	97					