Multiplanar 3D knee MRI segmentation via UNet inspired architectures

Sengar, Sandeep Singh; Meulengracht, Christopher; Boesen, Mikael Ploug; Overgaard, Anders Føhrby; Gudbergsen, Henrik; Nybing, Janus Damm; Perslev, Mathias; Dam, Erik Bjørnager

Published in:
International Journal of Imaging Systems and Technology

DOI:
10.1002/ima.22836

Publication date:
2023

Document version
Publisher’s PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Multi-planar 3D knee MRI segmentation via UNet inspired architectures

Sandeep Singh Sengar1 | Christopher Meulengracht2 | Mikael Ploug Boesen2 | Anders Føhrby Overgaard2 | Henrik Gudbergsen2 | Janus Damm Nybing2 | Mathias Perslev1 | Erik Bjørnager Dam1

Correspondence
Sandeep Singh Sengar, Department of Computer Science, University of Copenhagen, Copenhagen, 2100, Denmark.
Email: sengar@di.ku.dk

Funding information
In this study, the work on the LIRA cohort was supported by an unrestricted grant from Novo Nordisk A/S including active and placebo medicine, and by the Cambridge Weight Plan donating dietary supplements. In addition, the work on the LIRA cohort was supported by a core grant from the Oak Foundation (OCAY-13-309) given to the Parker Institute, Copenhagen University Hospital, Bispebjerg and Frederiksberg. The funders had no influence on the design or conduct of the study and were not involved in the experimental results.

1 | INTRODUCTION

The encoder–decoder-based UNet architecture introduced in 2015 is highly popular for medical image segmentation as demonstrated by the one-citation-per-hour rate during 2020. The merits of the UNet include a good performance with limited data and impressive generalization across diverse tasks. Like other segmentation algorithms, UNet is designed to extract feature maps at different scales, which may encode different levels of specificity. The UNet has become the golden standard method for the segmentation of 2D medical images that any new method must be validated against. In recent years, a number of variations to the seminal UNet have been proposed with promising results in the papers introducing them. However, there is no clear consensus if any of these architectures generalize as well and the UNet currently remains the methodological golden standard. For the segmentation of 3D scans, UNet-inspired methods are also dominant, but there is a larger variety across applications. By evaluating the architectures in a different dimensionality, embedded in a different method, and for a different task, we aimed to evaluate if any of these UNet alternatives are promising as a new golden standard that generalizes even better than the UNet. The purpose of this study was to compare UNet inspired models for generalized 3D segmentation. To efficiently segment the 3D scans, we employed each UNet variant architecture as the central 2D segmentation core in the multi-planar UNet 3D segmentation method that previously demonstrated excellent generalization in the MICCAI Segmentation Decathlon. It would strongly support a claim of generalizability, if a promising UNet-variant consistently outperforms the UNet in this quite different setting. The experimental results show that the multi-planar-based UNet2+ (MPUNet2+) method outperforms other variants including the original multi-planar UNet (MPUNet).
semantic information. Organ boundaries or textures, for instance, can be obtained through lower-level detailed features in the early layers. Based on these, high-level semantic features, such as the position of organs, can be detected by deeper layers. The UNet attempts to map these low-dimensional features to segmentations through a learned decoder. As is considered central to its success, the UNet implements so-called skip-connections that ensure an efficient flow of information between the encoder and the decoder, which, among others, allows the decoder to retrieve positional information that would otherwise be lost in max-pooling operations of the encoder.

In recent years, several variations of UNet architectures have been proposed with promising results for 2D images. Among them, the UNet2+, UNet3+, and deeply supervised versions of UNet2+ (UNet2+ DS) and 3+ (UNet3+ DS) are the most popular UNet inspired architectures for medical imaging. UNet2+ and UNet3+ have been developed to reduce the fusion of semantically non-similar features from plain skip connections in UNet. These architectures employ nested and dense skip connections to decrease the semantic gap between the encoder and the decoder parts of the network. Deep supervision (DS) may be further employed to learn hierarchical representations from the full-scale aggregated feature maps. The performance of both standard and deeply supervised versions of UNet2+ and UNet3+ has been demonstrated to be superior to the original UNet for certain 2D medical image datasets. However, it remains unclear if these findings generalize to other tasks or 3D medical imaging modalities.

The purpose of this study was to compare the performance of the original UNet architecture to its newer UNet2+ and UNet3+ counterparts (with and without DS) when used as the central segmentation component of the multi-planar UNet (MPUNet) method. The MPUNet is a combined training- and testing data-augmentation method for 3D medical image segmentation. At its core, it uses a 2D fully convolutional neural network such as UNet to segment 2D image slices from 3D volumes. In contrast to other 2D segmentation methods, however, the MPUNet samples a large set of such 2D image slices from across multiple viewing axes spanning the 3D image volume during both training and prediction. The MPUNet ranked 5th and 6th in the first and second rounds of the 2018 Medical Segmentation Decathlon, respectively, despite the system requiring no task-specific information, no human interaction, and being based on a fixed model topology and a fixed hyper-parameter set. These properties make the MPUNet ideal for comparing the five different architectures in consideration, that is, UNet, UNet2+, UNet3+, UNet2+ DS, and UNet3+ DS.

The considered architectures were evaluated for cartilage segmentation in knee MRI. Cartilage segmentation is an important step in clinical research on the progression of osteoarthritis and is likely in the future clinical practice. We considered three knee MRI cohorts from the Osteoarthritis Initiative (OAI), the Center for Clinical and Basic Research (CCBR), and the Liraglutide trial (LIRA). Using three cohorts with highly distinct MRI sequences provided a challenging evaluation setup.

The rest of this paper is organized as follows. First, a brief introduction of related works is given in Section 2. The investigated UNet methods and the multi-planar UNet framework are elaborated in Section 3. Results are provided in Section 4. Discussion of the paper’s results is given in Section 5. Finally, the conclusion of the work is provided in Section 6.

2 | RELATED WORKS

Knee osteoarthritis is a whole joint disease originated by a multifactorial combination of biochemical, biomechanical, systemic, and intrinsic risk factors. The segmentation of cartilages in knee MRI has been employed in a broad range of OA-related studies, for example, (1) classification and detection of knee OA progression (Ashinsky et al., Tiulpin et al., Ashinsky et al., Chang et al., Almajalid et al.), imaging biomarker analysis (Hafezi-Nejad et al., Schaefer et al., Shah et al., Williams et al.), biomechanical modeling (Luukkonen et al.), and stimulation of cartilage degeneration (Peuna et al., Luukkonen et al., Mononen et al.). Automated cartilage segmentation has been extensively researched, but remains an active research problem, in part due to the thin to extremely thin (few millimeters to submillimeter) structures of knee cartilage. Furthermore, patella, femoral, and tibial cartilages have distinctively different and changing shapes across image slices, which may pose challenges in particular to classical segmentation methods (e.g., 2D statistical shape-based models). Today, most methods are based on deep learning, see for a review. Notable deep-learning-based work includes Liu et al., who proposed a 2D convolutional neural network and slice-wise knee cartilage segmentation approach. Ambellan et al. suggested a similar approach while also adding a Statistical Shape Model (SSM) to refine the output of their UNet-based bone segmentation technique. The SSM was employed to remove false positive voxels from the tibia and femur and fill holes in segmentation masks due to poor intensity contrast. However, the good performance of this model is achieved using a multi-step approach which increases complexity and computational requirements.
Perslev\textsuperscript{11} proposed a simple and thoroughly evaluated multi-planar UNet (MPUNet) for the segmentation of arbitrary medical image volumes. This model requires no human interaction, no task-specific information, and is based on a fixed hyperparameter set and fixed model topology.\textsuperscript{11} Thus, it aims to eliminate the process of model selection in order to reduce the risk of overfitting on a method level. Xu\textsuperscript{33} employed the MPUNet to segment 3D jaw images. Also aiming to leverage the use of multiple views, Zhou et al.\textsuperscript{34} suggested the deep multi-planar co-training-based semi-supervised approach for 3D abdominal multi-organ segmentation. The relevance of multi-planar MRI acquisitions has also been assessed for prostate segmentation using deep learning techniques.\textsuperscript{35} In contrast to the MPUNet, this method uses an ensemble of convolutional neural networks each independently trained on a single imaging view. Similarly, QuickNAT\textsuperscript{40} has been proposed for quickly segmenting MRI Neuroanatomy. QuickNAT also explored multi-view aggregation by training the model for each of the three principal views and aggregating their scores.

Oktay et al.\textsuperscript{36} proposed a novel 3D UNet model with attention gates for automatic pancreas segmentation. In this work, attention gates provided increased performance, but at the cost of additional model parameters which comes at the cost of higher computational cost. The use of additional residual connections within the UNet was also investigated but did not provide any significant performance improvement, while Huang et al.\textsuperscript{7} later found that their UNet3+ method outperforms the attention UNet on a liver and spleen segmentation task.

Qin et al.\textsuperscript{37} proposed the $U^2$-Net for salient object detection. This model uses residual U-block at each level of encoder and decoder, which provides high performance but may increase computational cost. Additionally, the performance of this model based on 2D datasets is lower than our proposed work. Heinrich et al.\textsuperscript{38} proposed the TernaryNet method—a UNet variant model for segmenting the pancreas, which uses ternary convolutions (where both activations and weights are restricted to values in $\{-1, 0, 1\}$) to significantly reduce the computational requirements of the UNet. Isensee et al.\textsuperscript{39} proposed the nnUNet approach that relies on selecting an appropriate model topology and/or cascade from an ensemble of candidate models through cross-validation. In contrast to this, we have the interest to develop a task-agnostic segmentation method that depends on a single architecture and learning procedure. This is to make the model lightweight and easily transferable to clinical settings with limited computing resources. QuickNAT\textsuperscript{40} has been proposed for quickly segmenting the MRI Neuroanatomy. This model has been applied as a backbone for multi-view aggregation (similar to multi-planar UNet).

However, the accuracy of our proposed work is far better than this.

3 | Method

A total of five UNet variants were compared on the task of cartilage segmentation in knee MRI across three cohorts. In the following, we first introduce the considered models. Then, we briefly describe the multi-planar framework under which each model was trained and evaluated. Finally, we introduce the evaluation cohorts and the statistical analysis used to compare each model.

3.1 | Models

The performance of the following UNet variants were compared:

- **UNet**: The original UNet introduced by Ronneberger et al., 2015,\textsuperscript{1} but as implemented by Perslev et al., 2019.\textsuperscript{11} Notably, the Perslev et al., 2019 implementation scales the number of filters in each convolutional layer up by a factor $\sqrt{2}$, uses nearest-neighbor upsampling and batch normalization layers.

- **UNet2+**: UNet2+\textsuperscript{6} combines multi-scale features by employing a dense network of skip connections as an intermediary grid between the encoder and decoder. This full-scale skip connection architecture aims to make better use of multi-scale features and to minimize the loss of semantic information between the encoder and decoder.

- **UNet3+**: UNet3+\textsuperscript{7} is another variant of UNet also with increased skip-connectivity between the encoder and decoder as compared to the original UNet, but lesser so than UNet2+. The aim of introducing additional skip connections in UNet3+ is similar to that of UNet2+ to make full use of the multi-scale features, but with less number of parameters.

- **UNet2+ DS & UNet 3+ DS**: UNet2+ DS\textsuperscript{6} and UNet3+ DS\textsuperscript{7} adds deep supervision of the aggregated feature maps of UNet2+ and UNet3+, respectively, with the aim of learning better hierarchical representations. We have adopted deep supervision of UNet2+ and UNet3+ as implemented by Huang et al.\textsuperscript{7}

All models are fully convolutional neural networks\textsuperscript{49} that maps a (potentially multi-channel) 2D image slices $x$ of size $w \times h$ and $C$ channels to a softmax segmentation map of similar size $w \times h$ and $K$ channels for predicting $K$ classes. Batch normalization\textsuperscript{46} was implemented between all successive convolutional layers for all models.
We term each model as integrated into the MPUNet training method MPUNet2+, MPUNet3+, MPUNet2 + DS, and MPUNET3 + DS, respectively. A schematic overview of the distinctive skip-connection architecture of the default UNet, UNet2+ and UNet3+, respectively, is shown in Figure 1.

### 3.2 Multi-planar UNets

Each model was evaluated as the core segmentation component of the MPUnet. At its core, the MPUnet method works by traversing the 3D image volume in 2D slices from different orientations. During training, the central UNet-typed segmentation model is trained on randomly sampled 2D images from across all views. This scheme enforces the model to learn to segment the target as seen through multiple views. The sampling mechanism allows to sample of near-arbitrary numbers of distinct 2D images, that is, the technique resembles data augmentation. However, each sampled image displays biologically relevant structures, because they are sampled from the true training data without distortion (ignoring interpolation artifacts).

During inference, the learned ability to segment the target 3D volumes from multiple directions is further utilized by traversing the volume along all views. This results in a prediction volume generated from each traversal orientation. These prediction volumes are then linearly fused into a single segmentation volume using a learned fusion model. The steps are explained in more detail below.

### 3.3 Pre- and postprocessing

The MPUnet uses a minimum of image preprocessing applied outside of the segmentation model itself. An image and channel-wise outlier robust preprocessing is applied to scale the intensity values. The scaling is performed based on inter-quartile range and median statistics calculated only over foreground voxels in each channel. Here, foreground voxels are separated by having intensities greater than the first percentile of the intensity

---

**FIGURE 1** Architecture of (A) UNet; (B) UNet2+; (C) UNet3+. The depth of each node is displayed under the circle.
distribution. The 1 percentile threshold was a heuristic designed to encounter situations where most voxels in a scan had values of approximately 0 intensity. In some extreme cases, the IQR might also be 0 (or very close), and the normalization blows up or is undefined. Ignoring voxels with values less than or equal to the 1st percentile ignores all those 0-valued voxels and ensures that all voxels with actual information are normalized to 0 median and 1 IQR.

### 3.4 Multi-planar input data generation

During both training and testing, 2D image slices are generated by sampling from multiple planes of random orientation spanning the 3D image. The procedure is visualized in Figure 2. To generate these 2D inputs, we first define a set $V = \{v_1, v_2, \ldots, v_k\}$ of $k$ unit vectors in $\mathbb{R}^3$. Each vector may be manually specified or randomly sampled. Given this set $V$ of sampling orientations, we can get $k$ sets of 2D slices $L_v = \{(x_n^v, y_n^v)\}_{n=1}^{N_v}$, where $N_v$ is the number of 2D slices obtained from plane $v \in V$; $x$ and $y$ are the 2D slice and its corresponding label map, respectively; $y_i^v \in \{0, 1, \ldots, K\}$ is the cartilage label ($K$ segmentation targets plus background) of the $i$th pixel in $x$. During training, image slices sampled randomly across the views in $V$ are fed to the segmentation model. The model is not made aware of which view generated each received image. In effect, the multi-planar sampling acts to significantly augment the input data. Non-linear random elastic deformation transformations\cite{41} is further applied to each sampled 2D image slice with a probability 1/3 and with similar transformation parameter sampling strategy to that of Perslev et al., 2019\cite{11} (i.e., deformation intensity multipliers, $\sigma$, and elastic constants, $\alpha$, sampled uniformly from $[0, 450]$ and $[19, 29]$ respectively).

In our experiments, we considered $k \in \{1, 3, 6\}$, that is, we experiment with sampling from 1, 3, and 6 different random orientations when training and evaluating each of our five candidate models on each dataset. For $k = 3$ and $k = 6$, each view vector was sampled randomly under the restriction of each pair of random vectors having at least a 60° angle between them. The same set $V$ of view vectors was used for all models for a given $k$ and a given dataset to ensure maximum comparability between the results obtained using each model type.

In the single-plane ($k = 1$) setting, we considered the sagittal axis (i.e., the view was not randomly sampled). Because using a single view only usually leads to worse performance (see Perslev et al., 2019\cite{11}), this setting was only explored for a subset of datasets and methods as detailed in the Section 4 in order to limit computational costs.

### 3.5 Multi-planar prediction

During prediction, the model is applied along each view generating a set of $k$ segmentation volumes $S = \{S_v \in \mathbb{R}^{w \times h \times d \times K} | v \in V\}$. Every $S_v$ is mapped to the...
input space by allocating to each voxel in the input data, the value of its closest predicted point in \( S \). Distances are calculated in physical coordinates. The set of predictions, \( S \), are then fused to a single prediction using a linear model \( f_{\text{fusion}} : \mathbb{R}^{|V| \times K} \rightarrow \mathbb{R}^K \) computes \( z(x)_c = \sum_{c=1}^{K} W_{k,c} \cdot p_{k,x,c} + \beta_c \), where, \( p_{k,x,c} \) represents the probability of class \( c \) at voxel \( x \) as predicted by segmentation \( S_k \). The \( \beta \in \mathbb{R}^K \) is a bias parameter, which affects the general tendency to predict a given class, and \( W \in \mathbb{R}^{|V| \times K} \) weighs the probabilities of each individual class as predicted in each view. The validation data are used to learn the \( f_{\text{fusion}} \) parameters.

### 3.6 Evaluation datasets

For evaluation purposes, knee MRI images with independent manual segmentations from three distinct cohorts were considered. These were acquired from the OAI, CCBR, and LIRA. For all three cohorts, only a single knee from a given patient was included. It is the study population representative of the subjects that would be included in a clinical trial (investigating a potential OA treatment). For OAI, this subset was selected as representative for a clinical trial (i.e., KL 2 and 3). For CCBR and LIRA, the subsets were selected as representative of the population in these studies. The CCBR study is a community study with an overrepresentation of OA cases whereas LIRA is a clinical trial. Overall, the subjects used in this paper are representative of clinical trial cohorts. Each dataset is briefly described in the following:

- **OAI**: The OAI is a multi-center, ten-year observational study of men and women of age between 45 and 79 years and baseline body mass index (BMI) of \( 26.9 \pm 4.44 \) kg/m\(^2\). The study was funded by the National Institutes of Health (NIH). The OAI subcohort mainly includes knees of Kellgren and Lawrence (KL) scores 2 and 3, making the cohort representative of an OA clinical trial. Around one-third of the knees have KL scores of 0 or 1. In this study, 60 unique knee MRI volumes from the OAI were considered.

- **CCBR**: A community study of men and women of age 54 ± 15 and BMI 26 ± 4 kg/m\(^2\) from the Greater Copenhagen region, Denmark. The cohort included healthy individuals and subjects with mild to advanced radiographic OA. In this study, 140 knee MRI volumes from the CCBR were considered. Half the subjects of the CCBR cohort had no baseline radiographic signs of OA, defined as KL score 0, and the rest were roughly evenly distributed from KL 1 to KL 3.

- **LIRA**: A sub-study to the randomized LOSEIT trial investigating the effect of liraglutide on body weight and pain in overweight or obese patients (men and women) with knee osteoarthritis (ClinicalTrials.gov NCT02905864). Eligibility criteria for the patient trial were as follows: age 18–74 years, BMI \( \geq 27 \) kg/m\(^2\), clinical diagnosis of knee OA as confirmed by radiology, but restricted to definite radiographic OA at early to moderate-stages (Kellgren-Lawrence grades 1, 2, or 3), and motivation for weight loss. All patients were subjected to an 8-week diet intervention with a minimum 5% initial weight loss, after which the patients were randomized to receive either 3 mg liraglutide or 3 mg placebo for an additional 52 weeks. MRI was performed at eight weeks and at randomization week 0 and again after 52 weeks. In this study, 39 knee MRI scans from LIRA were considered.

Overall, the subjects considered for evaluation in this paper are representative of OA clinical trial cohorts. Technical details on the MRI acquisition protocols are given for each dataset in Table 1.

### 3.7 Optimization and hyperparameters

All three cohorts were randomly divided with the ratio of 60:20:20 for training, validation, and testing, respectively, in five-fold cross-validation experiments (i.e., in each fold, the training data was split again to create a validation set for the fold used for model selection via early stopping). Each multi-planar UNet variant was trained in each fold using 1, 3, or 6 input planes. All models were implemented in TensorFlow 2.3.0\(^{10}\) and trained on an Nvidia-Titan X graphics processing unit with 8GB memory. The cross-entropy loss function was minimized using the Adam optimizer\(^{45}\) with a fixed learning rate of \( \eta = 5 \times 10^{-5} \) and otherwise default Adam optimizer parameters \( (\beta_1 = 0.9, \beta_2 = 0.99, \text{ and } \epsilon = 10^{-8}) \). A variable batch size of 4–16 depending on model size was used (16 by default, reduced by 2 until batches fit in GPU memory). The effect of using a variable batch size is discussed below. Each model was trained for a maximum of 500 epochs and early stopping after 15 epochs of no observed improvement based on the validation Dice score. The observed model which scored to highest validation Dice score was then selected for testing in each fold. Because any number of images may be sampled under the MPUNet framework in each epoch, we defined one training and one validation epoch as spanning 2500 and 3500 image slices, respectively. All other non-constant hyperparameters were automatically deduced from the training data by the MPUNet software. See Perslev et al., 2019\(^{11}\) and the MPUNet GitHub repository details are available at https://github.com/sandeepsinghsengar/MPUNet2Plus. [Correction added on 29 December, 2022, after first online publication: GitHub repository link provided].
3.8 | Quantitative and qualitative evaluation

Each model was used to segment the full 3D MRI volume of all subjects in each test-fold split. Quantitative performance was evaluated by computing the Dice score overlap between the predicted and manual segmentations for each target compartment. Performance differences between pairs of models were tested using the two-sided, Wilcoxon signed-rank test at significance threshold $\alpha = 0.05$. In addition to performance, each model was also evaluated for its computational cost. Specifically, the number of trainable parameters and the total training time (mean across folds) were computed for each model and experiment.

A qualitative evaluation was also performed to compare the segmentation quality of each model on a single MRI. Surface models were fitted to all (manually and automatically) segmented compartments and visually compared. The single MRI on which the baseline MPUNet model performed closest to its median performance across all subjects was selected for the qualitative evaluation.

4 | RESULTS

4.1 | Quantitative evaluation

The 1-plane versions of the MPUNet, MPUNet2+, and MPUNet3+ provided segmentations across the test folds with Dice scores (mean across subjects and compartments) of 0.847, 0.872, and 0.852, respectively, on the OAI dataset, and 0.795, 0.815, and 0.817, respectively, on the CCBR dataset. To limit computational costs, the deep supervision versions of the MPUNet2+ and MPUNet3+ were not evaluated in the single-plane setup, and no models were evaluated on the LIRA dataset in the single-plane setup. As only the single sagittal plane was considered in these experiments, they test the performance of each of the three base segmentation models when applied to each dataset in a standard 2D setting without multi-planar data augmentation and prediction.

The evaluation results of the 3-plane and 6-plane experiments are shown in Table 2. The table also shows the average computation time and number of trainable parameters across all three datasets. The highest scores are highlighted in bold together with those scores for other models within the same experimental setting (same dataset and number of planes) that are not significantly different from the best score. Supporting $p$-values computed from Wilcoxon rank-sum tests comparing different pairs of models are shown in Table 3. The MPUNet2+ model performs significantly better or statistically non-different from all other models on all three datasets in the 6-plane setting, and only significantly inferior to the MPUNet3+ model on the CCBR dataset in the 3-plane setting. The MPUNet3+ model performs similarly to the MPUNet2+ model on both OAI and CCBR (6-planes), while significantly inferior on the LIRA dataset (6-planes). The baseline MPUNet performs inferior to the best of the MPUNet2+ and MPUNet 3+ models in all
experiments except OAI (6-planes). Across all but two experiments, the DS versions of the MPUNet2+ and MPUNet3+ models perform worse than their non-DS counterparts.

For most models and datasets, the average performance increases when going from 3 planes to 6 planes with a few exceptions that show relatively similar performance in both settings. These findings that more views usually improve performance, and never decrease performance, align with those of Perslev et al., 2019.11

As also evident from Table 2, the total number of trainable parameters is reduced when using the MPUNet2+ and MPUNet3+ implementations instead of the original MPUNet. The average computation time is also significantly shorter using those models on the OAI dataset in both the 3- and 6-plane setups, while computation time is similar on the CCBR dataset and LIRA dataset (except MPUNet3+ on LIRA, 6-planes, which computes faster, but also performed significantly worse). Interestingly, the average computation time for a given model both sometimes increases and sometimes decreases when going from the 3- to 6-plane setup. This finding is discussed below.

Box and whisker plots are shown in Figure 3 for a cartilage-wise comparison between the MPUNet and MPUNet2+ models. Combined Dice scores across all three datasets (OAI, CCBR, and LIRA) are shown for the medial tibial and medial femoral cartilages only, as only these two cartilages are common in all three datasets. In Figure 3, green triangles represent the mean Dice score corresponding to particular cartilage. The associated statistical description of Figure 3 is shown in Table 4. All of the minimum, maximum, mean, median, 25th, and 75th percentile statistics are higher for MPUNet2+ as compared to MPUNet3+ (OAI, CCBR, and LIRA datasets, 6-planes). The results have been computed with 3 and 6 number of planes. The performance is reported as mean Dice score and training time (in hour) averaged over five-folds. Here, the displayed mean Dice score is calculated over non-background classes only. The number of parameters (#param.) for all methods is also reported. The highest Dice scores are highlighted in bold together with those that are not statistically significantly different.

### TABLE 2 Performance comparison on OAI, CCBR, and LIRA datasets. The results have been computed with 3 and 6 number of planes. The performance is reported as mean Dice score and training time (in hour) averaged over five-folds. Here, the displayed mean Dice score is calculated over non-background classes only. The number of parameters (#param.) for all methods is also reported. The highest Dice scores are highlighted in bold together with those that are not statistically significantly different.

<table>
<thead>
<tr>
<th>Model</th>
<th>3 planes Dice score</th>
<th>3 planes Time (h)</th>
<th>6 planes Dice score</th>
<th>6 planes Time (h)</th>
<th>#Param.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset: OAI, Number of classes: 7</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPUNet</td>
<td>0.850</td>
<td>93</td>
<td>0.868</td>
<td>67</td>
<td>≈ 62M</td>
</tr>
<tr>
<td>MPUNet2+</td>
<td>0.873</td>
<td>26</td>
<td>0.867</td>
<td>32</td>
<td>≈ 36M</td>
</tr>
<tr>
<td>MPUNet3+</td>
<td>0.861</td>
<td>47</td>
<td>0.869</td>
<td>35</td>
<td>≈ 27M</td>
</tr>
<tr>
<td>MPUNet2+ DS</td>
<td>0.854</td>
<td>24</td>
<td>0.846</td>
<td>24</td>
<td>≈ 36M</td>
</tr>
<tr>
<td>MPUNet3+ DS</td>
<td>0.854</td>
<td>57</td>
<td>0.856</td>
<td>55</td>
<td>≈ 27M</td>
</tr>
<tr>
<td><strong>Dataset: CCBR, Number of classes: 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPUNet</td>
<td>0.817</td>
<td>15</td>
<td>0.816</td>
<td>16</td>
<td>≈ 62M</td>
</tr>
<tr>
<td>MPUNet2+</td>
<td>0.830</td>
<td>18</td>
<td>0.842</td>
<td>16</td>
<td>≈ 36M</td>
</tr>
<tr>
<td>MPUNet3+</td>
<td>0.836</td>
<td>18</td>
<td>0.841</td>
<td>20</td>
<td>≈ 27M</td>
</tr>
<tr>
<td>MPUNet2+ DS</td>
<td>0.827</td>
<td>18</td>
<td>0.836</td>
<td>14</td>
<td>≈ 36M</td>
</tr>
<tr>
<td>MPUNet3+ DS</td>
<td>0.824</td>
<td>17</td>
<td>0.832</td>
<td>19</td>
<td>≈ 27M</td>
</tr>
<tr>
<td><strong>Dataset: LIRA, Number of classes: 5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPUNet</td>
<td>0.832</td>
<td>24</td>
<td>0.864</td>
<td>34</td>
<td>≈ 62M</td>
</tr>
<tr>
<td>MPUNet2+</td>
<td>0.861</td>
<td>26</td>
<td>0.858</td>
<td>37</td>
<td>≈ 36M</td>
</tr>
<tr>
<td>MPUNet3+</td>
<td>0.845</td>
<td>29</td>
<td>0.850</td>
<td>26</td>
<td>≈ 27M</td>
</tr>
<tr>
<td>MPUNet2+ DS</td>
<td>0.835</td>
<td>23</td>
<td>0.844</td>
<td>20</td>
<td>≈ 36M</td>
</tr>
<tr>
<td>MPUNet3+ DS</td>
<td>0.853</td>
<td>34</td>
<td>0.856</td>
<td>43</td>
<td>≈ 27M</td>
</tr>
</tbody>
</table>

### TABLE 3 p-values for the Wilcoxon rank-sum test performed on the grand mean Dice score (mean across classes and subjects) obtained by the MPUNet2+ versus comparison methods (i.e., MPUNet, MPUNet3+, MPUNet2+ DS, and MPUNet3+ DS all at 6-planes) on the OAI, CCBR, and LIRA datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MPUNet2+ vs. MPUNet</th>
<th>MPUNet2+ vs. MPUNet3+</th>
<th>MPUNet2+ vs. MPUNet2+ DS</th>
<th>MPUNet2+ vs. MPUNet3+ DS</th>
</tr>
</thead>
<tbody>
<tr>
<td>OAI</td>
<td>0.028</td>
<td>0.005</td>
<td>4 e-30</td>
<td>4 e-34</td>
</tr>
<tr>
<td>CCBR</td>
<td>2 e-25</td>
<td>0.923</td>
<td>4 e-12</td>
<td>1 e-25</td>
</tr>
<tr>
<td>LIRA</td>
<td>4 e-4</td>
<td>2 e-29</td>
<td>4 e-29</td>
<td>1 e-17</td>
</tr>
</tbody>
</table>
compared to MPUNet for both cartilages. Furthermore, there is a higher number of (negative) outliers for MPU-Net than MPUNet2+.

In all, the MPUNet2+ was found to outperform other models and notably the baseline MPUNet in terms of both segmentation accuracy and computational cost across most experiments. A detailed cartilage-wise Dice scores of the MPUNet2+ model is provided for each individual dataset in Table 5, showing high segmentation accuracy on all individual compartments across datasets, with a minimum observed mean Dice of 0.78 on the Medical Femoral compartment of the LIRA dataset, and > 0.80 mean Dice on all other compartments.

4.2 Qualitative evaluation

The qualitative evaluation for the performance comparison is shown in Figure 4. For fair visualization, we have selected a single 2D slice from each dataset corresponding to the median Dice score. The full segmentation result of all the available cartilages is displayed for the selected 2D slice. The columns of Figure 4 show the ground truth and the results of MPUNet2+, MPUNet, MPUNet3+, MPUNet2+ DS, and MPUNet3+ DS, respectively. Each row corresponds to a knee from the OAI, CCBR, and LIRA datasets, respectively. Overall, while all methods provide highly reasonable segmentations, the MPUNet2+ provides a visually preferable segmentation on difficult and thin cartilages best exemplified by the medial femoral cartilage segmentations on the CCBR dataset (middle row).

Finally, a cartilage-wise visual comparison of the total count of false positives (FPs) and false negatives (FNs) in the segmented results on each cohort for the MPUNet and MPUNet2+ models was performed. For display purposes, the FN and FP voxel predictions were projected down from three to two dimensions. These results are shown in Figures S1–S12 for the following compartments, respectively: (1) Medial Tibial, Lateral Tibial, Medial Femoral, Lateral Femoral, and Medial Meniscus for the OAI dataset, (2) Medial Femoral and Medial Femoral Meniscus.

TABLE 4 Statistical description of box and whisker plots in Figure 3.

<table>
<thead>
<tr>
<th>Cartilage/Method</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>25th Percentile</th>
<th>Median</th>
<th>75th Percentile</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medial Tibial</td>
<td>0.622</td>
<td>0.918</td>
<td>0.857</td>
<td>0.837</td>
<td>0.866</td>
<td>0.885</td>
<td>0.048</td>
</tr>
<tr>
<td>MPUNet2+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medial Tibial</td>
<td>0.543</td>
<td>0.916</td>
<td>0.843</td>
<td>0.824</td>
<td>0.852</td>
<td>0.878</td>
<td>0.054</td>
</tr>
<tr>
<td>MPUNet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medial femoral</td>
<td>0.589</td>
<td>0.922</td>
<td>0.836</td>
<td>0.810</td>
<td>0.848</td>
<td>0.875</td>
<td>0.065</td>
</tr>
<tr>
<td>MPUNet2+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medial femoral</td>
<td>0.524</td>
<td>0.918</td>
<td>0.819</td>
<td>0.787</td>
<td>0.827</td>
<td>0.866</td>
<td>0.079</td>
</tr>
<tr>
<td>MPUNet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE 5 Compartment-wise performance (mean Dice score across subjects) of MPUNet2+ 6-planes on the OAI, CCBR, and LIRA datasets. “—” represents the absence of manual segmentations for this compartment in the dataset.

<table>
<thead>
<tr>
<th>Cartilage</th>
<th>OAI</th>
<th>CCBR</th>
<th>LIRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medial Tibial</td>
<td>0.862</td>
<td>0.849</td>
<td>0.876</td>
</tr>
<tr>
<td>Lateral Tibial</td>
<td>0.905</td>
<td>—</td>
<td>0.896</td>
</tr>
<tr>
<td>Medial Femoral</td>
<td>0.875</td>
<td>0.835</td>
<td>0.782</td>
</tr>
<tr>
<td>Lateral Femoral</td>
<td>0.893</td>
<td>—</td>
<td>0.801</td>
</tr>
<tr>
<td>Patellar</td>
<td>0.815</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lateral Meniscus</td>
<td>0.888</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Medial Meniscus</td>
<td>0.836</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Tibia</td>
<td>—</td>
<td>—</td>
<td>0.983</td>
</tr>
</tbody>
</table>
Tibial for the CCBR dataset, and (3) Medial Tibial, Lateral Tibial, Medial Femoral, Lateral Femoral, and Tibia for the LIRA dataset are shown. In the aforementioned figures, FNs and FPs are shown in the first and second rows, respectively, while results for MPUNet and MPUNet2+ are displayed in the first and second columns, respectively. It is evident from these figures that the MPUNet2+ has fewer FNs and FPs than the MPUNet across most cartilages and cohorts.

5 DISCUSSION

We investigated the performance of multiple multi-planar-based UNet architectures for the segmentation of cartilages in 3D knee MRI images. The target is to produce an efficient method that can work with the smaller size of data because one of the disadvantages of medical imaging is the lack of a high amount of labeled dataset. Each method was evaluated with respect to its ability to perform accurately with fixed hyperparameters across cohorts of variable data sizes, demographics, and scanner sequences. In addition, the computational requirements and the total number of parameters of each model were considered, favoring models which require fewer computational resources. All models were evaluated in a fixed compute-resource environment with a single GPU and corresponding fixed available memory pool. This choice of setting had the important implication that not all models could be trained with identical batch sizes. As the batch size is an important hyperparameter, which may affect performance, the model comparisons of this paper should be interpreted as investigating which model performs the best on the specific task of knee MRI segmentation given a restricted compute budget. The results do not provide information on which model types are generally superior given an unlimited computing environment. Similarly, these results alone do not show why one model works better than another. For instance, the default MPUNet and the MPUNet2+ differ significantly in their number of trainable parameters, which may affect performance, as they may be over- or under-parameterized, respectively, given the task. Thus, any observed differences in performance may be driven by any combination of the variable batch sizes used during training, different numbers of trainable parameters, and the actual model architecture differences (e.g., skip connectivity).

Each architecture was evaluated using 1, 3, and 6 numbers of input views over three different datasets (OAI, CCBR, and LIRA). The quantitative evaluations may be summarized as: (a) Table 2 shows that the Dice scores of the MPUNet2+ are significantly better than or indifferent to other methods across datasets. Only on the dataset, CCBR does MPUNet3+ perform indifferent from MPUNet2+ with a p-value of 0.92, while p < 0.05 for all other pairwise comparisons to the MPUNet2+, see Table 3. The DS version of the MPUNet2+ model performs worse than the MPUNet2+ model in all cases. It is unclear if DS adds unnecessary complications given the task of knee cartilage segmentation, or if the DS models should be trained with other hyperparameters to reach better performance. Interestingly, the performance of a single model tends to vary more across two different datasets than the performance varies between different models on the same dataset. This phenomena is likely driven by both the difference in the number of target compartments being segmented for each cohort (i.e., differences in the fundamental difficulty of the segmentation problem), as well as differences in scanner sequences and patient cohort demographics. (b) On average, the six planes setting of the MPUNet2+ performs better than the corresponding 1 and 3 plane settings. (c) The computation time of MPUNet2+ is typically less than the methods, most noticeably significantly shorter than the baseline MPUNet. The MPUNet2+ also has significantly fewer trainable parameters compared to MPU-Net (36 M vs. 62 M) and a lower GPU memory requirement. Interestingly, the time required to train each model does not scale linearly in the number of input views. This may indicate that the target function being approximated does not scale trivially in complexity with the number of views, because some information learned in one view is also useful in others (e.g., filters that respond to edges, textures, simple shapes, and other such view-independent features). (d) Box and whisker plots were displayed in Figure 3, showingDice scores for 6 planes versions of the MPUNet and MPUNet2+ on Medial tibial and Medial femoral cartilages across all three datasets pooled together. From these plots and the corresponding statistical description in Table 4, it is evident that all of the minimum, maximum, mean, 25th percentile, median, and 75th percentile are higher for the MPUNet2+ model. (e) Finally, the MPUNet2+ model performs well on all individual cartilages as seen in Table 5.

The qualitative evaluations and visualizations may be summarized as follows: (a) Figure 4 displays the segmented results of one selected 2D-slice from OAI, CCBR, and LIRA datasets, respectively, and for all tested methods. Qualitatively, the results of MPUNet2+ again seem more similar to the ground truth across datasets as compared to other methods. (b) The individual cartilage-wise total number of false negatives (FNs) and false positives (FPs) were projected from a complete 3D segmented cohort to the 2D plane as shown in the supplementary material. In line with the above findings, these visual
representations of segmentation accuracy also show that MPUNet2+ has both fewer FNs and FPs than the baseline MPUNet.

In all, the MPUNet2+ model appears to be both more accurate and less compute-intensive than the original MPUNet model for the task of knee cartilage segmentation. As for the original MPUNet, it was found robust across cohorts without hyperparameter tuning. Noticeably, the MPUNet2+ also seems to perform better than the MPUNet3+ model under our experimental setup, and DS seems to negatively affect both the MPUNet2+ and MPUNet3+ models. As a drop-in replacement to the original MPUNet, the MPUNet2+ inherits the benefits of the multi-planar framework, for example, good performance even with small amounts of training data and the ability to maintain a fixed hyperparameter set across tasks, that is, the model generally does not require task-specific information or tuning. The software does not require deep learning expertise to use making it applicable also for non-technical users.

As the baseline, MPUNet has already been found to perform as accurately or better than other deep-learning-based methods (see, for example,48,51) as well as classical methods such as KIQ,47 the MPUNet2+ may be considered a strong candidate models for general knee cartilage segmentation tasks. Finally, our results indicate that the UNet2+ model should be considered as an alternative to the current UNet gold standard also for 3D medical image segmentation tasks, as it transferred without modification and with excellent performance to both a new domain (3D knee MRI cartilage segmentation) and the multi-planar training and evaluation setup.

6 CONCLUSIONS

Five UNet variants were evaluated as the core segmentation component of the multi-planar UNet method for the task of knee cartilage segmentation on three distinct cohorts and MRI sequences. Both quantitative and qualitative evaluations demonstrated that the MPUNet2+ model had higher accuracy with less training time as compared to other methods, including the baseline MPUNet, on all considered datasets.

The ability of the UNet2+ model to transfer with high performance to an experimental setting quite different from what it was originally proposed for suggests that UNet2+ should be considered as an alternative to the golden standard UNet also for 3D medical image segmentation tasks.

FIGURE 4 Visual comparison of the performance in the form of full segmentation of all the methods obtained on a 2D slice corresponding to the median Dice score of all the datasets: where (A) Ground truth, (B) MPUNet2+, (C) MPUNet, (D) MPUNet3+, (E) MPUNet2+ DS, (F) MPUNet3+ DS; top to bottom row: OAI, CCBR, LIRA datasets.
ACKNOWLEDGMENTS

Work on the LIRA cohort was supported by an unrestricted grant from Novo Nordisk A/S including active and placebo medicine, and by the Cambridge Weight Plan donating dietary supplements. In addition, the work on the LIRA cohort was supported by a core grant from the Oak Foundation (OCAY-13-309) given to the Parker Institute, Copenhagen University Hospital, Bispebjerg and Frederiksberg. We also would like to acknowledge the staff of the Department of Radiology Bispebjerg-Frederiksberg and the patients in support of collecting the LIRA cohort. The funders had no influence on the design or conduct of the study and were not involved in data collection or analysis, in the writing of the manuscript, or in the decision to submit it for publication.

Work on the LIRA cohort was supported by an unrestricted grant from Novo Nordisk A/S including active and placebo medicine, and by the Cambridge Weight Plan donating dietary supplements. In addition, the work on the LIRA cohort was supported by a core grant from the Oak Foundation (OCAY-13-309) given to the Parker Institute, Copenhagen University Hospital, Bispebjerg and Frederiksberg. We also would like to acknowledge the staff of the Department of Radiology Bispebjerg-Frederiksberg and the patients in support of collecting the LIRA cohort. The funders had no influence on the design or conduct of the study and were not involved in data collection or analysis, in the writing of the manuscript, or in the decision to submit it for publication.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Sandeepr Singh Sengar https://orcid.org/0000-0003-2171-9332

Erik Bjørnager Dam https://orcid.org/0000-0002-8888-2524

REFERENCES


51. Desal AD, Caliva F, Iriondo C, et al. The international workshop on osteoarthritis imaging knee MRI segmentation

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.