Medical Artificial Intelligence
ISSN: 2957-5524

Zentime

Research Article https://doi.org/10.61189/242224ubhmhy

A survey on the application of deep learning in knee joint cartilage ultrasound image segmentation

Jintao Duan[1], Miao Zhou[2], Yuxiang Wang[1], Fangfang Chen[1], Liangqing Lin[3], Qinghua Wu[3], Haipo Cui[1]

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China. [2]Jiangsu Cancer Hospital, Nanjing 213164, Jiangsu Province, China. [3]The First Hospital of Putian, Putian 351100, Fujian Province, China.

Corresponding author: Haipo Cui.

Declaration of conflict of interest: None.

Received September 27, 2024; Accepted January 22, 2025; Published March 31, 2025

Highlights

- A systematic review of deep learning (DL) approaches for femoral cartilage segmentation in knee joint ultrasound images.
- Evaluation of popular DL models (U-Net, U-Net++, Siam U-Net, Mask R-CNNs) using various metrics.
- Discussion on dataset scarcity, data preprocessing methods, and future directions for DL-based ultrasound segmentation.

Abstract

Objectives: The femoral cartilage in the knee joint is prone to degenerative changes and injuries, often requiring Magnetic Resonance Imaging as the diagnostic gold standard. However, due to the high cost and limited availability of Magnetic Resonance Imaging, ultrasound is explored as a viable alternative. This paper presents a comprehensive review of current deep learning (DL) strategies for knee femoral cartilage segmentation in ultrasound images, focusing on commonly used datasets, data preprocessing techniques, and state-of-the-art DL models. **Methods:** We systematically reviewed the literature from major medical and engineering databases, summarizing key contributions to knee femoral cartilage segmentation. We focused on (1) the scarcity of large-scale public ultrasound datasets and its impact on model training, (2) popular DL architectures (e.g., U-Net variants, Mask Region-based Convolutional Neural Network), and (3) evaluation techniques, particularly the Dice Similarity Coefficient. We also examined image preprocessing and data augmentation strategies aimed at mitigating data insufficiency. **Results:** Our review shows that U-Net and its variants (e.g., Siam U-Net, U-Net++) commonly achieve competitive Dice Similarity Coefficient values (around 0.70–0.80) for knee cartilage segmentation, despite the limitations in training data. Advanced networks like Mask Region-based Convolutional Neural Network, when combined with robust image preprocessing and transfer learning (e.g., using COCO/ImageNet pretrained weights), can improve Dice Similarity Coefficient by over 20%. However, the absence of standardized public datasets limits direct comparisons between studies and affects reproducibility. **Conclusion:** DL holds significant potential for accurate and cost-effective femoral cartilage segmentation in knee joint ultrasound images, offering a feasible alternative or complement to Magnetic Resonance Imaging-based assessments. However, challenges remain due to the lack of large-scale, open-access ultrasound datasets and inconsistent evaluation protocols. Future work should focus on establishing public benchmarks, refining novel network architectures, and enhancing real-time clinical deployment to foster wider adoption and greater clinical impact.

Keywords: Deep learning, knee joint cartilage, ultrasound image segmentation

Introduction

The knee joint, one of the body's largest and most complex structures, endures substantial stress during daily activities. Femoral cartilage damage, commonly detected during arthroscopic examinations, contributes to patient discomfort, functional impairment, and decreased quality of life [1]. As populations age and lifestyles evolve, the prevalence of knee joint disor-

Address correspondence to: Haipo Cui, School of Health Science and Engineering, University of Shanghai for Science and Technology, No. 516 Jungong Road, Yangpu District, Shanghai 200093, China. E-mail: h_b_cui@163.com.

Table 1. Literature on knee femoral cartilage segmentation using deep learning and ultrasound imaging

Methods	Evaluation metrics	Authors
U-Net	DSC ROI DSCUB	Antico et al.(2020) [23]
U-Net++	DSC A/VPD HD	du Toit et al. (2022) [24]
Siam-U-Net	DSC CE	Dunnhofer et al. (2020) [25]
Mask R-CNNs	DSC, Jaccard	Kompella et al. (2019) [22]

Note: DSC, Dice Similarity Coefficient; CNNs, Convolutional neural networks; ROI, region of interest; Mask R-CNNs, Mask Region-based Convolutional Neural Networks.

ders has risen, including degenerative knee osteoarthritis, cartilage injuries, chondromalacia, and subchondral bone fractures [2]. Accurate diagnosis and timely treatment are crucial for pain relief, mitigating disease progression, and restoring knee function. Although magnetic resonance imaging (MRI) is the current diagnostic gold standard, its high cost, lengthy scan times, and limited availability reduce its practicality. Consequently, Schmitz proposed ultrasound as an effective alternative, citing its lower cost, real-time imaging capability, suitability for use in operating rooms, and value for surgical guidance [3-5]. Ultrasound-guided arthroscopic surgeries, as demonstrated by Wu, highlight its growing role in clinical applications [6].

Medical image segmentation plays a vital role in clinical tasks such as diagnosis, treatment planning, and disease monitoring, with ultrasound imaging being particularly favored for its non-invasive, real-time, and cost-effective features. However, ultrasound image segmentation presents inherent challenges due to artifacts, intensity inhomogeneities, operator-dependent probe placement, soft tissue deformation, and motion artifacts related to knee joint structures [7, 8]. These issues complicate segmentation [9]. To address these challenges, researchers have proposed various solutions, ranging from traditional digital image processing to machine learning. For example, Srensen et al. developed an automated bone segmentation method using dynamic programming, while Faisal et al. applied the Local Statistical Level Set Method for cartilage segmentation, achieving an average Dice Similarity Coefficient (DSC) of 0.91 [10-12]. Among the traditional (i.e., non-deep learning) segmentation approaches, Srensen et al. and Faisal et al. both demonstrated how explicit, model-based methods can be applied to medical image segmentation [10, 11]. Dice et al. employed a dynamic programming technique, while Faisal et al. adopted the Local Statistical Level Set Method for knee cartilage segmentation, achieving an average DSC of 0.91 [12]. Although their studies targeted different anatomical structures (bone vs. cartilage), both exemplify the effectiveness of traditional algorithms in enhancing segmentation accuracy. Despite these advances, traditional methods remain time-consuming and often require manual parameter adjustments for accurate feature delineation.

Deep learning (DL) has revolutionized computer vision, with algorithms that excel across a range of medical imaging applications. Key milestones include the development of convolutional layers and architectures such as AlexNet, U-Net, region-based convolutional neural networks (R-CNNs), and Transformer models [13-17]. DL techniques have shown impressive success in ultrasound image segmentation across various anatomical sites, such as the breast, thyroid, distal humerus cartilage, and left atrium [18-21]. However, knee ultrasound segmentation continues to present unique challenges. Recent approaches, including Mask region-based convolutional neural networks (Mask R-CNNs) introduced by Gayatri Kompella and Maria Antico, show promise for high-quality, fully automated segmentation, although research remains limited in this area [22].

This paper reviews the application of DL to femoral cartilage segmentation in knee ultrasound images. As shown in **Table 1**, we highlight four representative articles covering segmentation methods, publication years, and evaluation metrics in this field. In addition to the content in **Table 1**, Section 3 discusses additional DL architectures and various modifications applied to different knee structures, providing a more comprehensive overview. The organization of this paper is as follows: Section 2 introduces data preprocessing techniques for knee ultrasound images; Section 3 covers commonly used DL architectures, including those in **Table 1** and other relevant studies, along with evaluation metrics. Finally, Sections 4 and 5 present the discussion and conclusions, respectively.

Data and preprocessing

Dataset

In DL model training, data is a crucial compo-

Table 2. Medical image dataset

Datasets	Imaging modality	Anatomical region	Number of images
MRNet [30]	MRI	Knee	1,370
FastMRI [31]	MRI	Knee and cranial nerve	11,500
TN-SCUI 2020 [32]	Ultrasonic	Thyroid nodule	4,554
CAMUS [33]	Ultrasonic	Left ventricle and left atrial endocardium	1,000

Note: MRI, magnetic resonance imaging.

nent. It has been shown that extensive training on large datasets enhances the ability of models to learn accurate weights and parameters, thereby significantly improving performance. However, unlike other domains, medical image segmentation often faces challenges related to data scarcity and reliance on non-public datasets. These issues can lead to suboptimal model performance and problems such as overfitting, particularly when the available data is limited [26]. Creating large, publicly accessible datasets in the medical domain is particularly difficult due to patient privacy concerns and unauthorized access, leading most datasets to remain private [27]. Additionally, image annotation often requires manual labeling by experienced physicians. It is not feasible for a single doctor to annotate large datasets, so the involvement of multiple medical staff is necessary, which may result in inconsistencies in annotation standards due to individual differences among doctors. Moreover, the high financial cost of acquiring medical image data necessitates the involvement of volunteers and patients, and, due to safety concerns, some images cannot be collected in large quantities, further exacerbating data scarcity. The lack of sufficient data can lead to significant positive and negative sample biases due to the unique characteristics of medical data, ultimately hindering effective model training.

Models typically perform better when large public datasets are available. Several studies have successfully utilized public datasets, such as the Automatic Cardiac Diagnosis Challenge and the Multimodal Brain Tumor Segmentation Challenge, to achieve notable results and improve model performance [28, 29]. These datasets include a diverse range of imaging data, such as ultrasound, MRI, and computed tomography scans, and have contributed significantly to advancements in DL within the medical field. Given the current lack of a public dataset for knee ultrasound images, **Table 2** presents several public ultrasound databases, as well as datasets containing knee images, which can serve as a starting point for future research.

Data preprocessing

High-quality original medical images can be used for model training and yield reasonable results. However, implementing appropriate preprocessing steps can further enhance the utility of these images. While preprocessing cannot directly address the issue of limited data volume, it can alleviate model-fitting and performance challenges associated with data scarcity. Image enhancement techniques, such as adjusting brightness and contrast and applying blurring, combined with data augmentation methods (e.g., random flips and rotations), help expand the dataset, thus diversifying the training set and mitigating data limitations.

In medical data, data imbalance is a common challenge. Techniques like oversampling and undersampling have proven effective in reducing this issue. Ultrasound images, in particular, often suffer from speckle noise, leading to suboptimal signal-to-noise ratios. Kompella demonstrated that applying Gaussian filtering to remove high-frequency noise, combined with fine-tuning blurring parameters, reduces noise while preserving the clarity of anatomical structures, such as femoral cartilage [22]. Similarly, Antico et al. employed Matrix Laboratory (MATLAB) to segment 3D ultrasound images of the knee joint into 2D slices [23]. By adjusting pixel dimensions and filling in black pixels, they achieved favorable training results. Tang used thin-plate spline interpolation to model biological deformations and expand the dataset [34]. This interpolation method, which smoothly approximates sparse data points, supports biological variability, increasing the dataset and enhancing the model's generalization and performance, ultimately improving image model training.

Methods and Materials

Convolutional neural networks (CNNs)

CNNs are widely used in computer vision and DL due to their ability to extract features through convolution. This makes them particu-

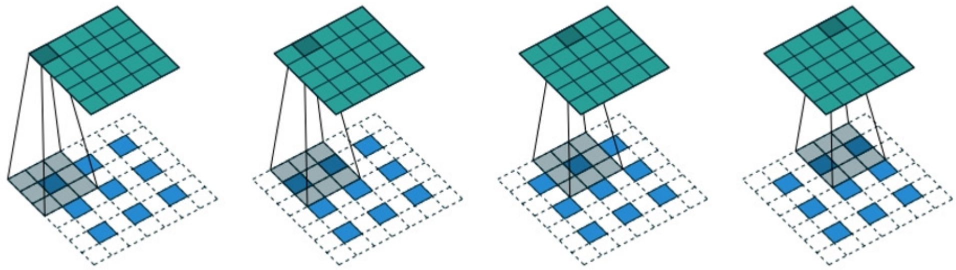

Figure 1. The fundamental architecture of CNNs. This figure is cited from [35]. CNNs, convolutional neural networks.

larly suited for tasks such as image classification, object detection, and image segmentation (**Figure 1**). CNNs are structured in a hierarchical layer configuration, with each layer processing the output from the previous one. The first layer, the input layer, directly receives the images. Subsequent layers filter the input data to identify local features, using filters (or kernels) that are adjusted through backpropagation to optimize their values. The size of these kernels is determined by the model designer.

In CNNs, the area of an input image influenced by a neuron is called its receptive field. Larger kernels expand this field, enabling the model to capture more contextual information. This capability enhances the model's ability to distinguish between targets and backgrounds, improving its reasoning. However, enlarging the receptive field can increase computational demands and reduce local detail resolution.

Activation layers are typically added to introduce non-linearity, and pooling layers are used to reduce the dimensionality of the output, thus lowering the computational load. Pooling can be categorized into max pooling, which highlights features and preserves finer details, and average pooling, which may reduce overfitting. Connected layers are then used to identify higher-level features. During training, kernel weights and neural connections are iteratively optimized through backpropagation until an optimal setup is achieved.

This canonical CNNs framework has paved the way for architectures such as LeNet and GoogLeNet [36, 37]. Depending on the input image type, CNNs can be 2D, 2.5D, or 3D. The 2.5D configuration, which captures spatial information more comprehensively than 2D but requires fewer computational resources than 3D, has been shown to be effective. Prasoon et al. demonstrated its utility by creating a 2.5D representation of a knee joint by aggregating three orthogonal 2D MRI image slices (XY, YZ, and XZ) [38]. They successfully segmented knee cartilage using this 2.5D approach.

U-Net

Following the advent of CNNs, U-Net was introduced as a groundbreaking neural network architecture, as shown in **Figure 2** [15]. Prior to the innovative proposal of U-Net by Ronneberger et al., efforts to segment medical images had been made but largely produced suboptimal results [15]. The introduction of U-Net, however, significantly advanced medical image segmentation, culminating in a victory at the 2015 ISBI Cell Tracking Challenge. U-Net represents a pivotal development in this field, with numerous algorithms subsequently built upon this model, achieving notable successes.

As shown in **Figure 2**, the U-Net architecture consists of two main components: an encoder (contracting pathway) and a decoder (expanding pathway). The encoder, similar to a conventional CNNs, uses layers of down-sampling and convolution to extract features from images. The decoder, on the other hand, employs up-sampling and skip connections to the contracting pathway, which enhances the resolution of features extracted by the encoder and transmits them to the final convolution layer, resulting in a fully segmented image. The name "U-Net" comes from the distinctive U-shaped architecture formed by the encoder-decoder interaction.

Antico et al. utilized a five-layer U-Net model to process 3D knee ultrasound images by segmenting them into 2D slices along the sagittal plane [23]. These slices were then used for model training, enabling successful segmentation of the distal femoral cartilage. They applied the Dice coefficient as the loss function, ReLU for the activation function, and Adam as the optimizer, which contributed to promising results [39, 40].

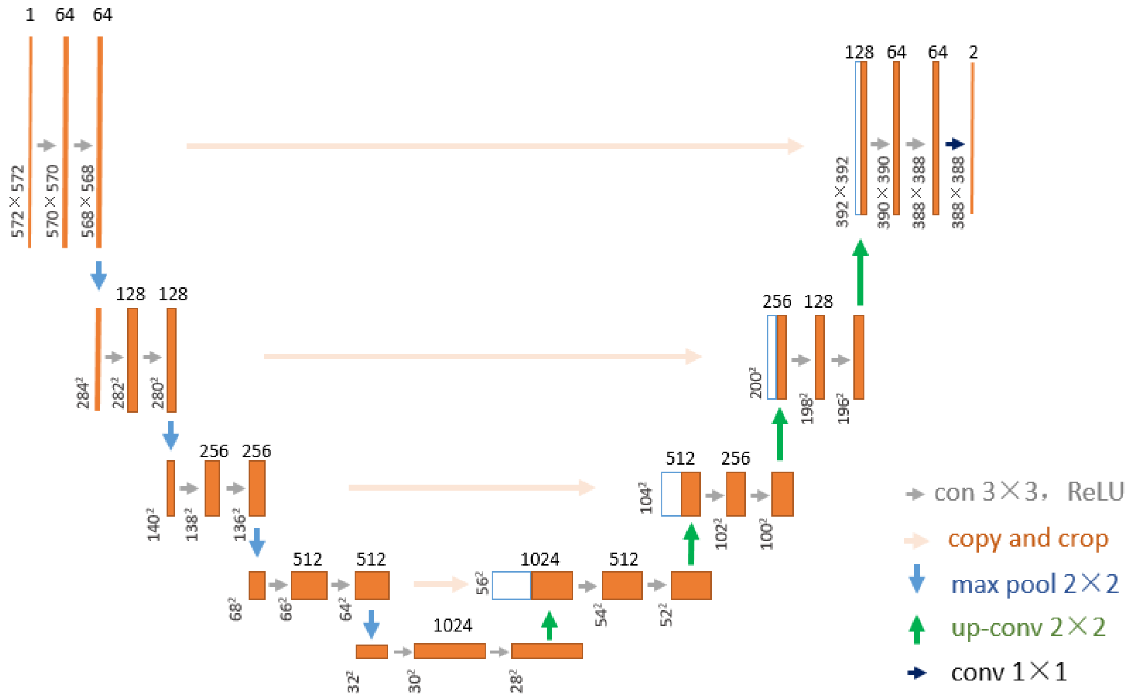

Figure 2. The fundamental architecture of U-Net. This figure is cited from [15].

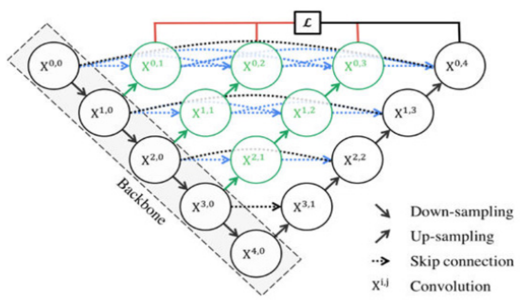

Figure 3. The fundamental architecture of U-Net++. This figure is cited from [41].

U-Net++

Building on the U-Net framework, U-Net++ represents an evolution in neural network architectures for image segmentation, as illustrated in **Figure 3** [41]. This model features an interconnected network structure, enhancing its ability to capture features and understand contextual information. U-Net++ incorporates a branching system that combines features from both the encoder and decoder levels, improving feature transmission and information integration. These branching connections enhance the model's ability to capture context and preserve fine details, leading to more accurate semantic segmentation. Compared to the original U-Net, these architectural improvements make U-Net++ more effective at extracting contextual knowledge, optimizing parameter usage, and enhancing robustness.

When applied to imaging tasks, the 3D U-Net++ model benefits from the unique data characteristics of 3D images compared to 2D images. Through convolution and pooling operations across dimensions, the 3D U-Net++ model excels at capturing and utilizing the inherent information in 3D image data. In fields requiring the analysis of volumetric shapes and structures, such as medical image segmentation, the 3D U-Net++ model is particularly effective at delineating contours and configurations. Its ability to extract features in three dimensions allows it to retain more information during convolution and pooling than its 2D counterpart. However, it comes with the drawback of reduced efficiency compared to the 2D version, due to the larger amount of data in 3D medical imaging. On the other hand, acquiring image data for 2D U-Net models is simpler and has a broader range of applications.

Addressing the limitations of 2D ultrasound imagery, which heavily relies on probe placement and operator skill, and exploring bedside alternatives that compensate for MRI's constraints, du Toit et al. employed a U-Net++ model with a ResNet-50 backbone instead of conventional U-Net models [24, 42]. They partitioned 3D ultrasound images into 2D slices and applied data augmentation techniques, including random assortments, translations, rotations, and scalings, to train the model. After reconstructing the 2D ultrasound slices, the predictions were reinserted into their respective slots within the 3D volume, linking the boundary points

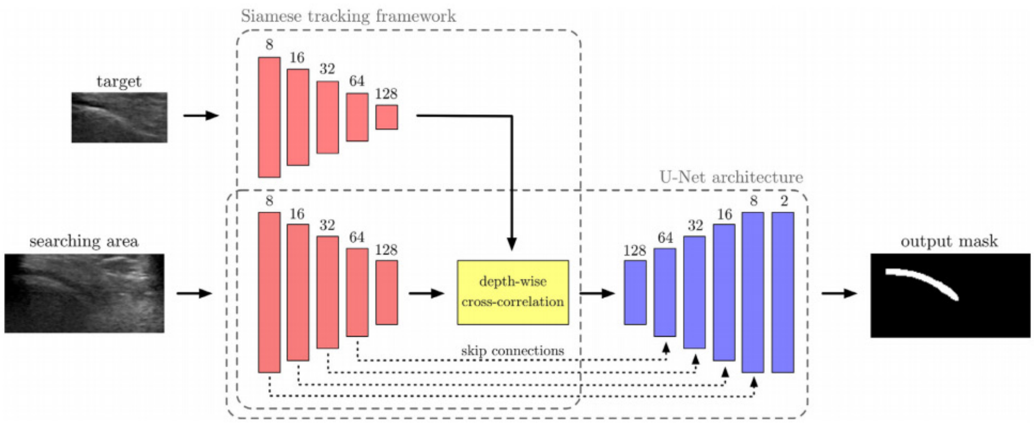

Figure 4. The fundamental architecture of Siam U-Net. This figure is cited from [25].

of adjacent slices. A smoothing filter was then applied, and the volume's extremities were sealed, resulting in a complete 3D surface. Despite these advancements, U-Net demonstrated a slight edge over U-Net++ in all evaluated metrics.

Siam U-Net

The Siam U-Net model is an enhanced version of the original U-Net architecture, specifically designed to improve performance in knee cartilage image segmentation. Dunnhofer et al. successfully implemented this model for segmenting ultrasound images of knee cartilage, achieving significant results [25]. The model's innovation lies in integrating the U-Net with the Siamese framework, as shown in **Figure 4** [43, 44]. It accepts two cropped images as inputs: one for preliminary segmentation and the other for finer segmentation. This dual-level segmentation approach allows the model to process larger images while preserving small details.

The images are processed through an encoder consisting of five computational units, each containing a 3x3 convolutional layer and a 2x2 max pooling operation. The encoder further refines each convolutional layer using batch normalization, ReLU activation, and a dropout layer. After processing, a depth-level cross-correlation operation is performed on the feature maps. Zero-padding is then applied to these cross-correlated feature maps to match the dimensions of the search area embedding. This enables the comparison of the target cartilage image with a specific slice region, resulting in a richer similarity map than standard cross-correlation operations.

The output binary mask is generated through a decoder branch sequence, consisting of four blocks: bilinear upsampling of the previous feature map, followed by a 2x2 convolution; connection with the feature representations from each encoder block in the search area (referred to as skip connections); and two 3x3 convolutional layers. Finally, a 1x1 convolutional layer is used to generate the output segmentation, consisting of two channels: one for predicting the foreground object (cartilage) and the other for predicting the background pixels.

The incorporation of depth-level cross-correlation and skip connections enables more accurate cartilage segmentation within the search area. Compared to the original U-Net, the Siam U-Net features smaller computational block sizes, reducing computational load and increasing processing speed. Additionally, the inclusion of a dropout layer improves the model's generalization capabilities.

Mask R-CNNs

Kompella and colleagues utilized a Mask R-CNNs for cartilage image segmentation [22]. In their study, 3D images of the cartilage were sliced sagittally into 256 2D slices, each with dimensions of 272x510 pixels, which were then used for training, as shown in **Figure 5**.

Mask R-CNNs is an extension of the R-CNNs model, designed for object detection and segmentation tasks. It excels at both detecting object locations and generating corresponding segmentation masks. The Mask R-CNNs architecture consists of four components: a backbone network, a Region Proposal Network (RPN), an object detection network, and a pixel-level mask generation network. The backbone network, such as ResNet or VGG, is responsible for extracting image features [45]. By utilizing this backbone, Mask R-CNNs extracts high-level features from the input images to support subsequent object detection and seg-

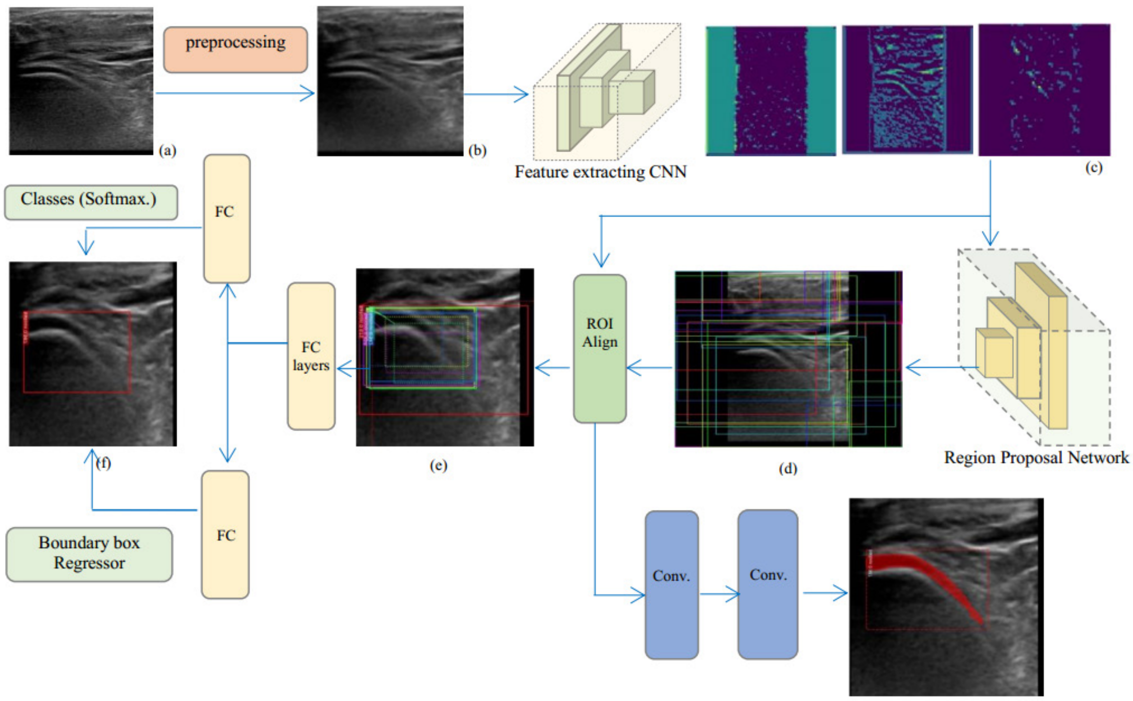

Figure 5. The fundamental architecture of Mask R-CNNs. (a) The original US image is fed into a feature-extracting CNNs; (b) The RPN generates candidate ROIs; (c) Each ROI is precisely aligned via ROI Align, avoiding misalignments from standard pooling; (d) FC layers predict object classes (Softmax) and refine bounding boxes (regressor); (e) A separate branch outputs instance-level segmentation masks; (f) Final predicted bounding box obtained from the bounding box regressor, along with the object probability obtained from the softmax classifier, after passing through the fully convolutional network. Through this multi-branch approach, Mask R-CNNs concurrently performs object detection and segmentation, enabling automatic knee cartilage analysis in ultrasound images. This figure is cited from [22]. US, ultrasound; CNNs, convolutional neural networks; RPN, Region Proposal Network; ROIs, regions of interest; FC, Fully connected; Mask R-CNNs, Mask Region-based Convolutional Neural Networks.

mentation tasks.

The RPN component generates object regions within the image. It is a convolutional network that captures features at various positions and scales using a sliding window approach, determining whether an object is present at each location. The RPN also offers boundary box regression adjustments for each region. The object recognition network then takes these regions, classifies them, and refines the bounding boxes for precise object delineation. This network typically consists of fully connected layers or a combination of connected layers.

The unique feature of Mask R-CNNs is its pixel-level mask generation network, which receives candidate regions from the RPN, extracts corresponding features, and generates pixel-level masks for the object using a Fully Convolutional Network [46]. This allows the network to represent each object instance with a detailed mask. Mask R-CNNs thus performs simultaneous object detection and pixel-level segmentation, providing both accurate object localization and detailed segmentation results. The backbone network extracts high-level semantic features, the RPN generates candidate regions, the object detection network classifies and refines the bounding boxes, and the mask generation network produces pixel-level masks for each object.

Evaluation metrics

Model performance in segmentation is typically evaluated using a variety of metrics. Relying on a single metric to assess a model's performance is unreliable; hence, using a combination of metrics provides a more comprehensive evaluation. Among these, the DSC is commonly regarded as the most reliable metric for evaluating segmentation models, measuring the overlap between two binary masks [12]. Segmentation accuracy is quantitatively assessed by comparing the model's output with manual annotations, considered the gold standard. Precision is typically evaluated through overlap measurements or surface distance. The DSC, which measures the overlap between two regions, is employed in this study to compare the proposed method's segmentation results with the ground truth. As shown in Equation 1, $R1$ represents the ground truth area, while $R2$ is

Table 3. Comparison of Mean and Min/Max Dice Similarity Coefficients (DSC) Achieved by U-Net and Intra-Observer Expert

Results	Mean DSC		Min/Max DSC	
	U-Net	Intra-observer	U-Net	Intra-observer
1	0.65	0	0/0.95	
2	0.68	0.77	0/0.94	0/0.95
3	0.69	0.73	0/0.95	0/0.92
4	0.68	0.40	0/0.94	0/0.92
5	0.71	0.70	0/0.94	0/0.92
Overall	0.68	0.64	-	-

Note: DSC, Dice Similarity Coefficient.

Table 4.Comparison of mean-DSC acheived by 2D Transverse and 3D Reconstruction Segmentation Performance Using U-Net and U-Net++

Results	U-Net Mean-DSC	U-Net++ Mean-DSC
2D Transverse	87.10±4.60	83.80±2.90
3D Reconstruction	73.10±3.90	74.30±1.70
P-value	0.063	0.064

Note: DSC, Dice Similarity Coefficient.

the segmented area. A higher DSC value indicates better segmentation accuracy. Given the diverse evaluation metrics used in the reviewed studies, the DSC was selected as the primary evaluation criterion for this study.

$$\mathrm{DSC} = 2 * \frac{|R1 \cap R2|}{|R1| + |R2|} \quad (1)$$

Results

Performance of U-Net on 2D ultrasound data

The U-Net architecture has become a pivotal model in medical image segmentation, serving as the foundation for numerous derivatives, while maintaining strong performance on its own. Antico et al. applied the U-Net model to segment femoral cartilage images from knee ultrasound scans, obtaining promising results [23]. A comparative analysis was conducted between the DSC obtained from the U-Net model and those from expert manual annotations. **Table 3** presents the calculated average DSC, along with the maximum and minimum DSC values. As shown in the table, the U-Net model achieved an average DSC between 0.65 and 0.71, with an overall mean of 0.68. In comparison, expert annotations across test sets (Test Sets 1 to 5) ranged from 0 to 0.77, with an average of 0.64. Zero DSC values from experts, as seen in Test 1 of **Table 3**, indicate an anomalous case. The data suggests that the U-Net model performs well in image segmentation, with its DSC values approximately 1-2% higher than those of the experts for both complete images and selected regions. This implies that the U-Net's segmentation performance is either comparable to or slightly better than that of human experts. However, both the U-Net and expert segmentations are less accurate in images with blurred boundaries or poorly defined cartilage. To improve performance, specialized strategies may be needed to address such challenging cases.

Performance of U-Net++ on 2D/3D ultrasound data

The U-Net++ model is an enhanced version of the original U-Net, featuring a more complex network architecture with additional layer hierarchies and branching connections. This modification enhances the model's ability to capture more detailed features and integrate contextual information. However, its segmentation performance does not always surpass that of the original U-Net model, and under certain conditions, it may produce inferior results.

du Toit et al. applied both U-Net and U-Net++ models to 3D knee ultrasound images, which were sliced into 2D images for training [24]. The 2D segmented images were then reassembled into their original 3D structure, connecting the boundaries of adjacent slices. A smoothing filter was applied, and the volume extremes were sealed to reconstruct a continuous 3D surface. In their study, du Toit et al. assessed the mean DSC between the 2D slices and the

Table 5. Comparison of mean DSC among Siam-U-Net, Temporal and Spatio-Temporal Tracking, U-Net, and Léger's U-Net

Results	Mean DSC		Mean DSC		
	Temporal tracking	Spatio-temporal	Siam-U-Net	U-Net	Léger's U-Net [47]
1	0.74±0.15	0.73±0.16	0.74±0.15	0.61±0.24	0.60±0.23
2	0.69±0.20	0.71±0.16	0.69±0.20	0.62±0.22	0.65±0.23
3	0.69±0.16	0.70±0.15	0.69±0.16	0.68±0.18	0.68±0.21
4	0.69±0.17	0.68±0.18	0.69±0.17	0.62±0.23	0.64±0.22
5	0.73±0.14	0.73±0.14	0.73±0.14	0.66±0.21	0.67±0.20
6	0.69±0.15	0.68±0.16	0.69±0.15	0.63±0.23	0.62±0.28
Overall	0.68	0.71±0.16	0.70±0.16	0.64±0.22	0.64±0.23

Note: DSC, Dice Similarity Coefficient.

Table 6. Comparison of segmentation results with and without preprocessing, using ImageNet, COCO, and the Level Set Method

Results	Pre-processing	Mean-DSC
Pretrain with ImageNet [48]	No	0.22
	Yes	0.64
Pretrain with COCO [49]	No	0.67
	Yes	0.80
Level Set [50]	No	0.52
	Yes	0.54

Note: DSC, Dice Similarity Coefficient.

reconstructed 3D images, and evaluated the P-value to analyze the variance in segmentation performance, as shown in **Table 4**. The results showed that the segmentation quality of the 2D images was significantly better than that of the 3D reconstructions. This discrepancy was due to the reconstruction method, which produced rectangular surfaces that intersected three distinct regions, as manually segmented. Additionally, the smoothing filter simplified the 2D predictions, lowering the DSC. However, it also removed outliers, making the segmentation results for both 2D and 3D images more comparable for both U-Net models.

Further performance tests with knee osteoarthritis images revealed no significant differences in segmentation outcomes between U-Net and U-Net++, aligning with results from healthy knee images.

Performance of Siam-U-Net on temporal and spatio-temporal sequences of 2D ultrasound data

Siam-U-Net is an innovative variant of the U-Net neural network, integrating a Siamese tracking framework. Dunnhofer et al. collected a dataset comprising 35 3D+time series knee ultrasound images [25]. Manual segmentation of 2D ultrasound images was performed by an orthopedic surgeon, generating temporal and spatio-temporal 2D+time series from the 3D+-time series data. A total of 18,278 slices within the sagittal planes of the ultrasound volume were annotated. The dataset was split into six subsets, with one used for testing and five for training. The model was applied to segment femoral cartilage in 2D ultrasound images. The average DSC for both temporal and spatio-temporal tracking were compared, alongside the performance of the original U-Net and the U-Net variant proposed by Léger et al., as shown in **Table 5** [47]. The results indicate that Siam-U-Net outperforms these comparative models in terms of performance metrics.

Performance of Mask R-CNNs on 2D ultrasound data

Mask R-CNNs, an advanced extension of R-CNNs, simultaneously detects object locations and generates their masks. Using a 4D ultrasound probe (5-13 MHz frequency range), Kompella et al. acquired static 3D ultrasound images of knee joints from healthy volunteers [22]. These images were manually segmented by an experienced physician and split into 256

2D slices in the coronal plane, with 85% allocated to the training set and 15% to the test set. Preprocessing was done according to COCO and ImageNet datasets before training. The efficacy of Mask R-CNNs was evaluated using the Mean DSC, comparing segmentation accuracy among networks pretrained on COCO and ImageNet datasets to those with no pretraining. Additionally, Mask R-CNNs' performance was compared with the conventional level-set segmentation method, which requires manual initialization. As shown in **Table 6**, networks pretrained on the COCO2016 dataset exhibited a 20% increase in DSC over those without pretraining. Pretraining with the ImageNet dataset resulted in a 190% improvement in performance, surpassing the level-set technique's accuracy.

Performance comparison with different models

Comparing DSC scores (U-Net: 0.68, Intra-observer: 0.64), U-Net slightly outperforms intra-observer segmentation. Furthermore, in recognizing 2D (87.1, 73.1) transverse and 3D reconstructed images of femoral cartilage, U-Net outperforms U-Net++ (83.8, 73.1). This is due to U-Net++'s increased structural complexity, stemming from additional pathways and skip connections that enable it to incorporate more image information. Despite these improvements, optimal results are not achieved in femoral cartilage image recognition. The integration of the U-Net model with a Siamese network structure in Siam-U-Net allows for the use of two cropped images as inputs, one for preliminary segmentation and the other for precise segmentation. This multi-level segmentation architecture improves the model's ability to preserve fine details while handling larger images, resulting in a slight increase in DSC from U-Net (0.64) to Siam-U-Net (0.70). Preprocessing also plays a significant role in performance, as demonstrated by DSC scores in Mask R-CNNs. Preprocessing the COCO dataset increased the DSC score from 0.67 to 0.80, with similar enhancements observed in other datasets. These improvements were due to the extraction of relevant features, noise reduction, and the use of data augmentation techniques that enriched the data and improved the model's generalization capability.

Discussion

Summary

The literature reviewed in this article demonstrates the partial application of DL for the segmentation of femoral cartilage in knee ultrasound images. The article begins with an introduction to common data types used in DL, followed by an overview of publicly available datasets and preprocessing methods for data management. Various methods suitable for femoral cartilage segmentation in knee ultrasound images are discussed, with a focus on the U-Net architecture and its variants, which have been predominantly used. Other models, such as Mask R-CNNs, have also shown promising results. During the review, it was noted that some studies employed less common evaluation metrics, and many relied on non-public datasets. The performance of various models is highlighted, with the DSC being the most frequently used evaluation metric.

Challenges

Despite these advancements, challenges remain in the DL segmentation of femoral cartilage in knee ultrasound images. First, there is a lack of large-scale, publicly accessible datasets specifically for knee femoral cartilage ultrasound imaging. While data augmentation and the fusion of supervised and unsupervised learning can help alleviate this issue, the effectiveness of the models is still constrained by their inherent limitations. Thus, the creation of a standardized, publicly available dataset is crucial. Second, early research on knee femoral cartilage ultrasound segmentation often failed to incorporate the latest neural network models and architectures, leading to performance gaps compared to newer approaches. Ongoing research is expected to address these gaps. Additionally, there is growing interest in intelligent medical devices for medical image processing. However, the significant computational power required for processing medical images presents challenges in ensuring both accuracy and real-time performance when deploying models on medical devices. The model size is also a critical factor. These challenges may be gradually overcome with advancements in the miniaturization of DL models for medical image segmentation and the development of intelligent medical devices.

Future research directions

Future research in automated femoral cartilage detection, powered by DL, aims to enhance real-time operational efficiency, accuracy, and user-friendliness of the model. By integrating the model with ultrasound images of femoral cartilage obtained via the probe, immediate identification could be achieved, thus improving the efficiency of healthcare providers and strengthening the reliability of clinical decision-making. Another key focus is the development of a comprehensive public dataset for femoral cartilage. As the dataset grows, model detection accuracy is expected to improve. Simultaneously, given the current data limitations, there is a need for progress in identifying various distinct morphological forms of femoral cartilage. Therefore, a major research direction is to enhance the model's ability to generalize and identify these unique cartilage morphologies.

Conclusion

This research reviews the latest advancements in DL for knee ultrasound image segmentation and compiles publicly accessible medical datasets and methods employed in femoral cartilage segmentation. The study also includes an analysis of several evaluation metrics. DL has the potential to significantly enhance the diagnostic efficiency of physicians. For patients in medically underserved areas, online availability of trained models could provide substantial support, improving medical services. This survey aims to offer valuable insights and encourage further research on femoral cartilage segmentation in knee ultrasound images.

Author contributions: Jintao Duan: Conceptualization of the review; Literature search and screening; Writing—original draft preparation; Data interpretation; Final manuscript revision. Miao Zhou: Assisted in literature collection; Contributed to data organization; Reviewed and edited the manuscript for language and structure. Yuxiang Wang: Data curation; Figure/table preparation; Provided feedback on manuscript sections related to deep learning architectures. Fangfang Chen: Literature analysis; Drafted parts of the methodology and discussion; Verified references for accuracy. Liangqing Lin: Provided medical guidance on knee joint clinical aspects; Ensured the clinical relevance of the review; Critically reviewed the manuscript for important intellectual content. Qinghua Wu: Offered medical guidance regarding ultrasound imaging protocols; Gave expert opinions on clinical feasibility; Revised the manuscript sections involving ultrasound data. Haipo Cui (Corresponding Author): Overall supervision; Funding acquisition; Methodology design in deep learning perspective; Final approval and submission of the manuscript.

References

[1] Curl WW, Krome J, Gordon ES, et al. Cartilage injuries: a review of 31,516 knee arthroscopies. Arthroscopy 1997;13(4):456-460.

[2] Cui A, Li H, Wang D, et al. Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. EClinicalMedicine 2020;29-30:100587.

[3] Schmitz RJ, Wang HM, Polprasert DR, et al. Evaluation of knee cartilage thickness: a comparison between ultrasound and magnetic resonance imaging methods. Knee 2017;24(2):217-223.

[4] Antico M, Sasazawa F, Wu L, et al. Ultrasound guidance in minimally invasive robotic procedures. Med Image Anal 2019;54:149-167.

[5] Lueders DR, Smith J, Sellon JL. Ultrasound-guided knee procedures. Phys Med Rehabil Clin N Am 2016;27(3):631-648.

[6] Wu L, Jaiprakash A, Pandey AK, et al. 29 - Robotic and Image-Guided Knee Arthroscopy. In: Abedin-Nasab MH, editors. Handbook of Robotic and Image-Guided Surgery. Elsevier, 2020. p. 493-514.

[7] Vlad VM. Pitfalls in Musculoskeletal Ultrasound. In: El Miedany Y, editors. Musculoskeletal Ultrasonography in Rheumatic Diseases. Cham. Springer International Publishing, 2015. p. 21-55.

[8] Shrimali V, Anand RS, Kumar V. Current trends in segmentation of medical ultrasound B-mode images: a review. IETE Tech Rev 2009;26:8-17.

[9] Noble JA, Boukerroui D. Ultrasound image segmentation: a survey. IEEE Trans Med Imaging 2006;25(8):987-1010.

[10] Srensen TA, Srensen TJ. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skr Dan Vid Sel 1948;5:1-34

[11] Faisal A, Ng SC, Goh SL, et al. Knee cartilage segmentation and thickness computation from ultrasound images. Med Biol Eng Comput 2018;56(4):657-669.

[12] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297-302.

[13] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 1980;36:193-202.

[14] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60:84–90.

[15] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 2015;234-241.

[16] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition 2014;580-587.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems 2017;6000–6010.

[18] Chen G, Liu Y, Qian J, et al. DSEU-net: A novel deep supervision SEU-net for medical ultrasound image segmentation. Expert Syst Appl 2023;223:119939.

[19] Chen Y, Zhang X, Li D, et al. Automatic segmentation of thyroid with the assistance of the devised boundary improvement based on multicomponent small dataset. Appl Intell (Dordr) 2023:1-16.

[20] Zhao W, Su X, Guo Y, et al. Deep learning based ultrasonic visualization of distal humeral cartilage for image-guided therapy: a pilot validation study. Quant Imaging Med Surg 2023;13(8):5306-5320.

[21] Degel MA, Navab N, Albarqouni S. Domain and Geometry Agnostic CNNs for Left Atrium Segmentation in 3D Ultrasound. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018 2018;630-637.

[22] Kompella G, Antico M, Sasazawa F, et al. Segmentation of Femoral Cartilage from Knee Ultrasound Images Using Mask R-CNN. Annu Int Conf IEEE Eng Med Biol Soc 2019;2019:966-969.

[23] Antico M, Sasazawa F, Dunnhofer M, et al. Deep learning-based femoral cartilage automatic segmentation in ultrasound imaging for guidance in robotic knee arthroscopy. Ultrasound Med Biol 2020;46(2):422-435.

[24] du Toit C, Orlando N, Papernick S, et al. Automatic femoral articular cartilage segmentation using deep learning in three-dimensional ultrasound images of the knee. Osteoarthr Cartil Open 2022;4(3):100290.

[25] Dunnhofer M, Antico M, Sasazawa F, et al. Siam-U-Net: encoder-decoder siamese network for knee cartilage tracking in ultrasound images. Med Image Anal 2020;60:101631.

[26] Hesamian MH, Jia W, He X, et al. Deep learning techniques for medical image segmentation: achievements and challenges. J Digit Imaging 2019;32(4):582-596.

[27] Razzak MI, Naz S, Zaib A. Deep Learning for Medical Image Processing: Overview, Challenges and the Future. In: Dey N, Ashour AS, Borra S, editors. Classification in BioApps: Automation of Decision Making. Cham Springer International Publishing, 2018. p. 323-350.

[28] Bernard O, Lalande A, Zotti C, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging 2018;37(11):2514-2525.

[29] Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 2015;34(10):1993-2024.

[30] Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med 2018;15(11):e1002699.

[31] Zbontar J, Knoll F, Sriram A, et al. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. ArXiv 2018;abs/1811.08839.

[32] Shusharina N, Heinrich MP,Huang R. Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data: MICCAI 2020 Challenges, ABCs 2020, L2R 2020, TN-SCUI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings. 2021;XIX,156.

[33] Leclerc S, Smistad E, Pedrosa J, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans Med Imaging 2019;38(9):2198-2210.

[34] Tang Z, Duan J, Sun Y, et al. A combined deformable model and medical transformer algorithm for medical image segmentation. Med Biol Eng Comput 2023;61(1):129-137.

[35] Dumoulin V, Visin F. A guide to convolution arithmetic for deep learning. ArXiv 2016;abs/1603.07285.

[36] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998;86:2278-2324.

[37] Szegedy C, Wei L, Yangqing J, et al. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015;1-9.

[38] Prasoon A, Petersen K, Igel C, et al. Deep Feature Learning for Knee Cartilage Segmentation Using a Triplanar Convolutional Neural Network. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013 2013;246-253.

[39] Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. International Conference on Artificial Intelligence and Statistics 2011.

[40] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014.

[41] Zhou Z, Rahman Siddiquee MM, Tajbakhsh N,

et al. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support 2018;3-11.

[42] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016;770-778.

[43] Bertinetto L, Valmadre J, Henriques JF, et al. Fully-Convolutional Siamese Networks for Object Tracking. Computer Vision – ECCV 2016 Workshops 2016;850-865.

[44] Bertinetto L, Henriques JF, Valmadre J, et al. Learning feed-forward one-shot learners. Proceedings of the 30th International Conference on Neural Information Processing Systems 2016;523–531.

[45] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 2014.

[46] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015;3431-3440.

[47] Léger J, Brion E, Javaid U, et al. Contour Propagation in CT Scans with Convolutional Neural Networks. Advanced Concepts for Intelligent Vision Systems 2018;380-391.

[48] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition 2009;248-255.

[49] Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context. Computer Vision – ECCV 2014 2014;740-755.

[50] Osher S, Sethian JA. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. J Comput Phys 1988;79:12-49.