

Knee cartilage MR images segmentation based on multi-dimensional hybrid convolutional neural network

Wenda Chen and Cao Shi*

School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China

Abstract. Accurate segmentation of knee cartilage in MR images is crucial for early diagnosis and treatment of knee conditions. Manual segmentation is time-consuming, leading researchers to explore automatic deep learning methods. However, the choice between 2D and 3D networks for organ segmentation remains debated. In this paper, we propose a hybrid 2D and 3D deep neural network approach, named UVNet, which combines the strengths of both techniques to enhance segmentation performance. Within this network structure, the 3D segmentation network serves as the backbone for feature extraction, while the 2D segmentation network functions as an information supplement network. Local and global MIP images are generated by employing various maximum intensity projection modes of knee MRI volumes as input for the information supplement network. By constructing a local and global MIP feature fusion module, the supplementary information obtained from the 2D segmentation network is fully integrated into the backbone network. We assess the quality of the proposed method using the Osteoarthritis Initiative (OAI) dataset and the 2010 Grand Challenge Knee Image Segmentation (SKI-10) dataset, comparing it to the Baseline Network and other advanced 2D and 3D segmentation methods. The experiments demonstrate that UVNet achieves competitive performance in the aforementioned two cartilage segmentation tasks.

Keywords: Convolutional neural network, maximum intensity projection, segmentation of knee cartilage

1. Introduction

The knee joint is the largest and most complex joint in the human body. Abnormal changes in the knee joint may indicate the onset of diseases, such as degenerative changes in knee cartilage. Osteoarthritis (OA) is a prevalent degenerative joint disease, affecting a substantial number of individuals [1]. OA is a major cause of morbidity and disability, resulting in considerable socioeconomic costs. In 2004, arthritis was estimated to cost the United States \$336 billion, or 3% of gross domestic product [2, 3].

OA is a complex, heterogeneous condition that commonly causes disability in the aging population [4, 5]. Since joint tissue damage is irreversible, early diagnosis is critical for patients in the early stages of osteoarthritis development. While OA damages all joint tissues, cartilage degeneration is the hallmark [6, 7]. Studies have shown that the morphometric assessment of cartilage structure, such as volume, thickness, and area, via MR images provides an accurate and precise measure of OA progression [8, 9]. Consequently, precise segmentation of knee cartilage tissue has become a critical step in knee image analysis. But a major hurdle to quantifying cartilage outcomes from MRI scans is the lack of resources necessary for tissue segmentation [10]. It may take up to 6 hours for a clinical reader to manually segment each series

*Corresponding author. Cao Shi, School of Information Science and Technology, Qingdao University of Science and Technology, 266000 Qingdao, Shandong, China. E-mail: caoshi@yeah.net.

of 3-dimensional (3D) knee MR images [11]. In the process of extensive segmentation, professionals may also make some inevitable mistakes.

Deep learning-based automatic segmentation technology has attracted a lot of attention recently since it can automatically learn features. Deep convolutional neural networks with an encoder-decoder architecture have produced outstanding results in a variety of medical picture segmentation applications since the introduction of U-Net [12]. However, the encoder requires multiple downsampling operations to capture deep semantic information, which leads to a significant loss of information and consequently affects the segmentation accuracy.

The purpose of this study is to propose a hybrid 2D and 3D CNN framework named UVNet for the automatic segmentation of knee joint cartilage from MRI data. In medical image segmentation methods, 2D approaches focus more on the details of individual slices and require fewer resources, while 3D approaches can capture global context. We aim to leverage the strengths of both methods to alleviate the issue of information loss caused by downsampling. In UVNet, 3D V-Net [13] is used for feature extraction from the 3D volume, and 2D U-Net extracts features from local and global Maximum Intensity Projection (MIP) images as supplementary information. These supplementary information sources are fused with the main network through a feature fusion module to compensate for the information loss due to downsampling. Therefore, during the training process of the 3D network, shape and location information extracted by the 2D network can assist in the segmentation of knee joint cartilage structures. In summary, we make the following three contributions:

(1) We propose a hybrid network, the UVNet, which employs V-Net as the 3D backbone network and U-Net as the information supplement subnetwork. This hybrid network architecture, which combines different dimensional segmentation methods, integrates the advantages of both 2D and 3D networks. Experimental results indicate that its segmentation performance exceeds that of independent 2D or 3D segmentation networks (refer to Table 5).

(2) Additionally, we introduced a more effective approach for the application of MIP (Maximum Intensity Projection) images. By combining local and global MIP images, U-Net can obtain additional shape and positional information to compensate for the information loss caused by downsampling in V-Net. Based on the characteristics of local and global

MIP features, we also proposed different feature fusion modules.

(3) Our method merges the segmentation features of 2D and 3D convolutional networks, thereby enhancing the performance of network segmentation. Experiments conducted on the OAI-ZBI and SKI-10 datasets reveal that our method outperforms other traditional segmentation methods across various evaluation metrics.

2. Related works

2.1. Knee joint segmentation method

Recently, deep convolutional neural networks have achieved achievements in biological image processing tasks such as classification [14] and segmentation [15–17]. Meanwhile, U-Net-like structures have been widely used in knee joint segmentation tasks. For the segmentation task of key knee joint structures, Woo et al. [18] combined anomaly information detection with the segmentation task for the first time. First, anomaly information is extracted using U-Net for knee MRI and then added to the downstream segmentation task to assist segmentation. Mao et al. [19] also proposed a similar two-stage knee segmentation network. First, significance detection was performed to obtain different bone structure location regions, and then two U-Net networks were used to segment different bone structure regions independently. For the first time, Chen et al. [20] suggested using adversarial loss to enhance the segmentation of bone structure images after resampling. In addition, a recovery network was used to restore the bone structure to its pre-sampling resolution. Meanwhile, Ambellan et al. [21] employed a CNN to locate the bone surface in the first stage, followed by the use of SSM to refine the segmentation by CNN and a smaller 3D-CNN for sub-volume segmentation. The above multi-stage segmentation methods bring segmentation performance improvements while increasing computational complexity.

There is a general problem of category imbalance in medical images, and this problem is more worthy of attention in knee images. To address this problem, Lee et al. [22] proposed a do-differential segmentation network that first segments the bone and cartilage complex (BCC) and then the bone. Cartilage is obtained by subtracting bone from BCC. This avoids segmenting the cartilage directly and alleviates the problem of category imbalance in the

segmentation task. Dai et al. [23], on the other hand, proposed a new loss function to solve the problem of category imbalance in images. In addition, the method uses a compact convolutional network to reduce the computational complexity so that the 3D knee images can be fed more completely into the network. Similarly, Raj et al. [24] proposed a multi-category loss function to obtain more accurate segmentation of different types of cartilage.

There are also some researchers who hope to find better segmentation methods than U-Net. Sun et al. [25] explored the effect of replacing the standard convolutional blocks in U-Net [12] and DeepLabv3plu [26] with different blocks such as residual block, residual SE block, dense block, and dense SE block on the task of knee image segmentation and found that the variant using SE block has better feature capturing capability. Sengar et al. [27] utilized a knee joint dataset to assess variants of the U-Net architecture, such as U-Net++ [28] and U-Net3+ [29], from various angles, tasks, and embedding methods. Their objective was to obtain a superior and more widely applicable alternative to U-Net. Kessler et al. [30] applied conditional generative adversarial networks (cGANs) [31] for the first time to the segmentation of knee MRI and compared them with U-Net in various aspects, and they concluded that cGANs networks have better robustness. Kessler et al. [32] used the

3D-CaSM [33] method to extract accurate regional measurements of cartilage morphology on 2D and 3D U-Net segmented knee masks. More et al. [34] proposed the “Discrete-MultiResUNet”, which combines the Discrete Wavelet Transform (DWT) and the MultiResUNet architecture and is more effective in extracting salient features for key tissue segmentation. Overall, the UNet-like method has been shown to be effective in knee segmentation tasks. It is worth noting that U-Net downsampling results in the loss of feature information, which is also present in the other UNet-like methods mentioned above. However, this problem has not been given enough attention.

3. Method

In this section, we introduce our proposed multidimensional network structure system for medical image segmentation framework, UVNet. The overall structure is shown in Fig. 1, which consists of two network branches: a backbone feature extraction network, V-Net, and an information supplementation network, U-Net. The V-Net takes the knee MRI volumes as input and the pixel-by-pixel probabilistic map as output. The U-Net, on the other hand, uses local and global MIP maps extracted from MRI volumes as input. Additional information is supplied

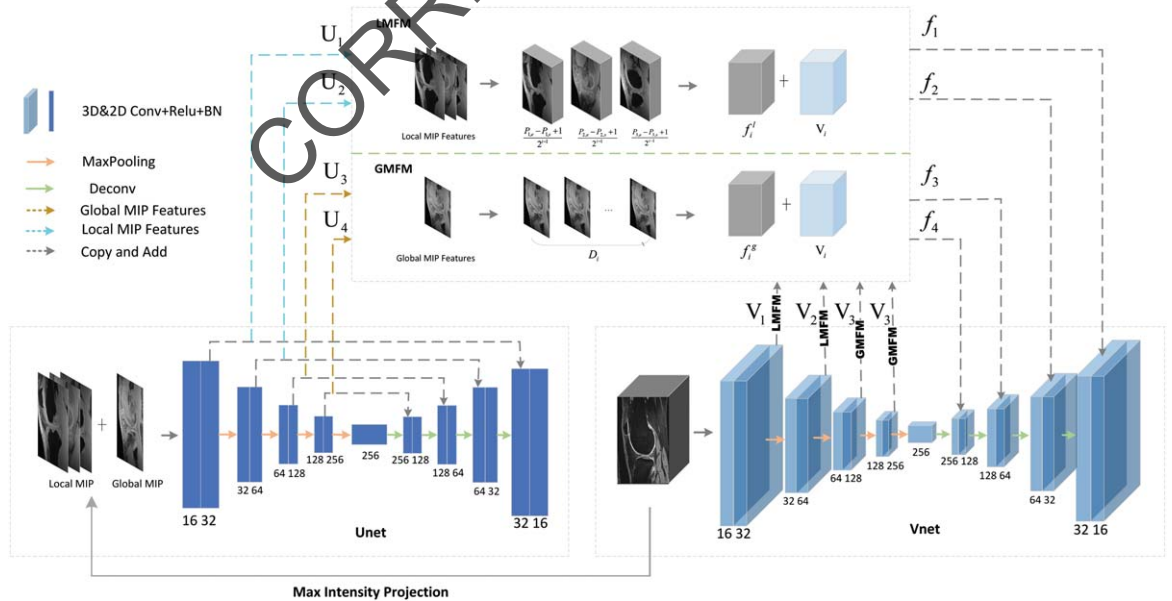


Fig. 1. The UVNet architecture consists of a main body composed of U-Net and V-Net. Additionally, there are local MIP fusion module (LMFM) and global MIP fusion module (GMFM) parts. The maximum intensity projection of various models is applied to the MRI volume to provide local and global MIP images, which are used as input for the U-Net.

mented by fusing the MIP feature maps with the hierarchical feature maps of the V-Net through the feature fusion module. This work has been inspired to some extent by [35]. In the next three sections, we explain the backbone and information supplementation network structure, the local and global MIP extraction strategy, and the feature fusion module.

3.1. 3D backbone network and 2D information supplement network

UVNet employs V-Net as the backbone feature extraction network, which has proven its effectiveness in various medical image segmentation tasks. It uses 3D convolution for feature extraction and achieves end-to-end segmentation through an efficient encoding-decoding structure. In the encoding path, each layer of features undergoes two consecutive identical convolutional layers, followed by batch normalization and the ReLU activation function, and then downsampling using a convolution operation with a stride of two. The decoding path is similar to the encoding path, but the feature map scale is recovered by upsampling, and in the last layer, a $1 \times 1 \times 1$ convolution is performed to map each component feature vector to the desired number of classes. Skip-connections are also used between the encoder and decoder to recover the information loss caused by downsampling.

Traditional medical image segmentation methods usually use a separate 2D or 3D network for feature extraction, which is normally effective. But, in order to extract the deep semantic information, it is necessary to downsample the feature maps, which inevitably results in the loss of information. Even though skip-connections are used in V-Net to supplement the lost information, the skip-connections of the next layer are still obtained by downsampling the previous layer. Knee joint cartilage has an elongated morphology and occupies a relatively small proportion of pixels in the entire medical image. After multiple downsampling steps, this information loss becomes even more critical for cartilage features, leading to issues like under-segmentation or even cartilage breakage in the segmentation mask. Therefore, in addition to using V-Net for feature extraction, we also employ U-Net to segment MIP images as additional supplementary information to alleviate information loss in skip-connections. U-Net utilizes two-dimensional convolution and has the same number of layers as V-Net, making it compatible for feature fusion with V-Net.

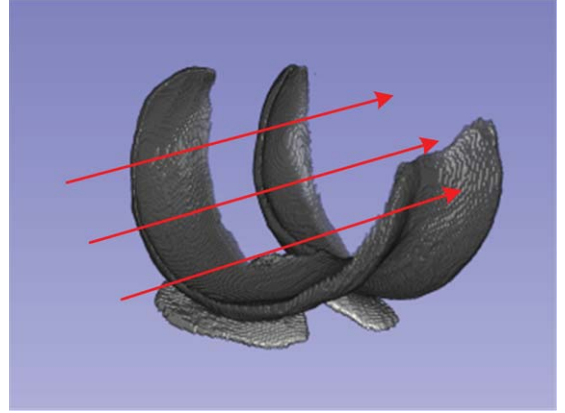


Fig. 2. Maximum Intensity Projection

3.2. Local MIP images and Global MIP image

Maximum Intensity Projection (MIP) is a projection method used for extracting structures and features from medical images. The conventional projection process involves scanning the entire MRI volume in a specific direction to find the maximum pixel intensity. Then, the maximum intensity values from all pixels along that direction are projected onto a two-dimensional image, which we refer to as the global MIP image. The projection process can be represented as illustrated in Fig. 2. By applying this process separately to the MRI volume and corresponding labels, MIP images and MIP labels are obtained. Although this projection process is relatively straightforward, it finds extensive application in medical image analysis, including tasks such as reconstruction [36, 37], detection [38, 39], and segmentation [40, 41]. During the projection process, cartilage structures are typically retained in the MIP image, providing valuable information about the shape and location of the cartilage.

It is worth noting that during the process of ray projection, lower pixel values will be overwritten by higher pixel values. Therefore, when acquiring the global MIP image, a substantial amount of valuable shape and positional information is lost, and this should be given due consideration. If the global MIP image is used exclusively as supplementary information, it may introduce interference during the training of V-Net, resulting in a decrease in accuracy. Considering the complexity of the segmentation task, relying solely on global projection as additional information may appear overly simplistic and brute-force. To mitigate the aforementioned issue of information loss and

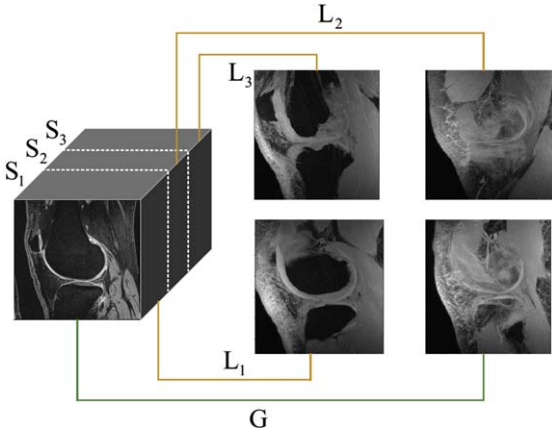


Fig. 3. For each MRI volume, three local MIP images and one global MIP image were extracted.

enrich the supplementary information structure, we have employed a method for calculating MIP based on the morphology of cartilage, which we call local MIP. We consider that the cartilage morphology in adjacent slices exhibits similarities. By performing maximum intensity projection across multiple neighboring slices, the loss of information is expected to decrease, and the morphological structure of the cartilage will be more prominently represented in the MIP image. As depicted in Fig. 3, we have partitioned the entire cartilage structure into three parts according to the morphological characteristics, denoted as S_1 , S_2 , and S_3 . The $S_j (j = 1, 2, 3)$ represents the MRI volume between slices $p_{j,s}$ and $p_{j,e}$, where $p_{j,s}$ denotes the number of the starting slice, and $p_{j,e}$ denotes the number of the last slice. We then carried out projections on these three parts separately, resulting in three local MIP images, namely L_1 , L_2 , and L_3 . These local MIP images are combined with the global MIP image, denoted as G , and together they serve as inputs to the U-Net.

3.3. Feature fusion module

To incorporate the extracted MIP image as supplementary information into the V-Net, we need to perform dimension expansion on different MIP images. We propose two feature fusion modules for this process, referred to as the Local Feature Fusion Module (LMFM) and the Global Feature Fusion Module (GMFM), as shown in Fig. 1. $U_i (i = 1, 2, 3, 4)$ represent the MIP feature maps obtained from U-Net, with dimensions $B_i^u \times C_i \times W_i \times H_i$. The hierarchical feature maps in the V-Net are

denoted as $V_i (i = 1, 2, 3, 4)$, with dimensions $B_i^v \times C_i \times W_i \times H_i \times D_i$. Here, B represents batch size, C represents the number of channels, and C, W, H, D represent the number of channels, width, height, and depth of the feature maps. We set B_i^v to 1, while B_i^u is set to 4, represented as $B_i^u = \{l_{i,1}, l_{i,2}, l_{i,3}, g_i\}$. Among these, $l_{i,1}$, $l_{i,2}$, and $l_{i,3}$ correspond to the local MIP feature maps, and g_i represents the global MIP feature map.

3.3.1. Local MIP fusion module

The U_1 and U_2 are used as inputs for LMFM to accomplish feature fusion with the V_1 and V_2 , respectively. In LMFM, only $l_{i,1}$, $l_{i,2}$, and $l_{i,3}$ undergo dimension expansion. As mentioned earlier, the $L_j (j = 1, 2, 3)$ are obtained by extracting continuous slices between $p_{j,s}$ and $p_{j,e}$. To ensure that the size of the local MIP features matches the 3D semantic feature V_i , we need to replicate $l_{i,j}$ by $(p_{j,e} - p_{j,s} + 1) / (2^{i-1})$ times and concatenate them to form f_i^l . This can be expressed as:

$$f_i^l = \left(\underbrace{l_{i,1}, \dots, l_{i,1}}_{\frac{p_{1,e}-p_{1,s}+1}{2^{i-1}}}, \underbrace{l_{i,2}, \dots, l_{i,2}}_{\frac{p_{2,e}-p_{2,s}+1}{2^{i-1}}}, \underbrace{l_{i,3}, \dots, l_{i,3}}_{\frac{p_{3,e}-p_{3,s}+1}{2^{i-1}}} \right) \quad (1)$$

The f_i^l has the same scale as V_i . The two feature maps are superimposed in the form of a weighted linear summation, followed by a group normalization operation and a ReLU activation function, calculated as:

$$f_i = \text{relu}(\text{GN}(\lambda f_i^l + (1 - \lambda)V_i)) \quad (2)$$

Where, λ is a weight parameter that adjusts the importance of the two feature maps.

3.3.2. Global MIP fusion module

The U_3 and V_3 , as well as the U_4 and V_4 , respectively, serve as inputs for GMFM to complete fusion. The difference from LMFM is that in GMFM, we only perform dimension expansion on the global MIP feature g_i to match its scale with V_i . The specific expansion process involves duplicating g_i by D_i times, yielding f_i^g , which can be expressed as:

$$f_i^g = \underbrace{(g_i, g_i, \dots, g_i)}_{D_i} \quad (3)$$

Similar to the LMFM, a weighted linear sum is performed between the obtained 3D global MIP map f_i^g and the output 3D feature map V_i of the V-Net

encoder. After the group normalization operation and ReLU activation function are applied, the final feature map is calculated as:

$$f_i = \text{relu}(GN(\lambda f_i^g + (1 - \lambda)V_i)) \quad (4)$$

The final feature map f_i is then passed through a skip-connection to the i layer of the V-Net network decoder.

4. Experiments

4.1. loss function

In the training process, DiceLoss was selected as the loss function for both 2D and 3D models and defined as,

$$L_{2D}(P_{mip}, G_{mip}) = 1 - \frac{2|P_{mip} \cap G_{mip}|}{|P_{mip}| \cup |G_{mip}|} \quad (5)$$

Where, P_{mip} is the segmentation result of 2D networks, and G_{mip} is the MIP ground truth.

$$L_{3D}(P, G) = 1 - \frac{2|P \cap G|}{|P| \cup |G|} \quad (6)$$

Where, P is the segmentation result of 3D networks, and G is the ground truth of femoral cartilage and tibial cartilage.

The overall loss function of network is defined as:

$$L = \omega L_{2D} + (1 - \omega) L_{3D} \quad (7)$$

Our final loss consists of losses from two paths, where L_{2D} and L_{3D} are loss functions of the 2D network and the 3D network, respectively, and ω represents the weight between the two losses. In our experiment, ω was set at 0.35 for optimal experimental performance.

4.2. Datasets

4.2.1. SKI-10

Our model utilized the SKI-10 database from the SKI-10 MICCAI challenge [42], which comprises 100 knee MR images along with expert-annotated femur and tibia segmentation images. The SKI-10 database is based on MR images provided by Biomet, Inc. These images were collected at 80 different centers across the United States using equipment from all major manufacturers, such as General Electric, Siemens, Philips, Toshiba, and Hitachi. All images were captured in the sagittal plane with a rectangular

pixel spacing of 0.4×0.4 mm and a slice distance of 1 mm, without the use of contrast agents. Experts at Biomet, Inc. interactively segmented all MR images into four categories: femur, femoral cartilage, tibia, and tibial cartilage [42]. In this study, we only used femoral and tibial cartilage.

4.2.2. OAI-ZIB

The Osteoarthritis Initiative dataset (OAI-ZIB) consists of 507 MRI volumes with manual segmentations by experts at Zuse Institute Berlin. Images were acquired using a 3T Siemens MRI scanner and quadrature transmit-receive knee coil (USA Instruments, Aurora, OH) at one of four sites using a 3D sagittal water-excited T1-weighted (TE 5 ms, TR 16 ms, flip angle 25°). Dual Echo in the Steady State (DESS) sequence with in-plane resolution of 0.365×0.365 mm, Flip Angle 25° dual echo in the steady state (DESS) sequence with in-plane resolution of 0.365×0.365 mm, matrix size 384×384 , and slice thickness of 0.7mm [43].

4.3. Experimental detail

In our study, we divided the two datasets into 80% for network training and 20% for testing. Since the diagnosis of osteoarthritis in clinical practice mainly refers to the morphological manifestations of cartilage, only cartilage tissue labels were retained in the data set. After removing the labels of bone, the dataset contained three categories: 0-background, 1-femoral cartilage, and 2-tibial cartilage. Then, the images are randomly cropped into patches with a size of $256 \times 256 \times 96$, which are sent to the V-Net for training.

During the training period, the learning rate is set to 0.002 in the initial stage, the training batch size of V-Net is set to 1, and the training batch size of U-Net is set to 4, so as to ensure that the training of each MIP is non-interference. The total training epochs are set at 500. The Adam [44] optimizer is used for optimization, and the attenuation coefficient is 0.0001.

4.4. Evaluation metrics

We used the Dice Similarity Coefficient (DSC), Recall, and Precision [45–47] as evaluation metrics, all of which were commonly used to evaluate similarity in the field of medical image segmentation.

$$DSC = \frac{2TP}{2TP + FP + FN} \quad (8)$$

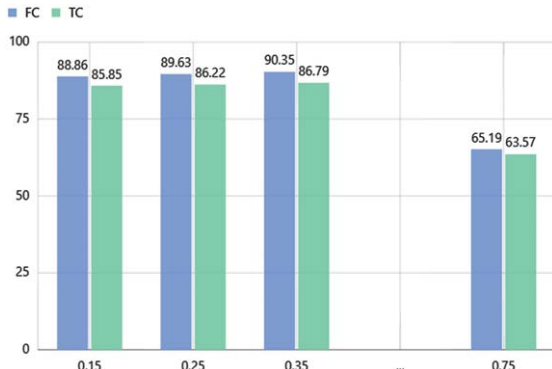


Fig. 4. The influence of hyper-parameter ω on network performance.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

where TP, FP, and FN are the number of true positive, false positive, and false negative predictions, respectively.

5. Results and discussion

5.1. Comparison of loss hyper-parameters

Firstly, we study the influence of hyper-parameters on the partial performance of the loss function. We experimented with the effect of different values of ω from 0.15 to 0.75 on the performance of the hybrid network. The experimental results are shown in Fig. 4. When $\omega = 0.35$, the experimental performance is the best.

5.2. Comparison on OAI-ZIB dataset

In order to evaluate the performance of our proposed network framework, we conducted a comparative analysis with several popular deep learning-based medical image segmentation methods, such as U-Net [12], U-Net++ [28], V-Net [13], and 3D U-Net [48], on the OAI dataset. Table 1 presents the performance comparison of the different methods on the OAI dataset, with the best results emphasized in bold.

Our proposed hybrid network model, UVNet, which incorporates local and global MIPs, achieved impressive DSC scores of 90.35 for femoral cartilage

(FC) and 86.79 for tibial cartilage (TC). These scores outperformed those of other commonly used segmentation methods. The U-Net model obtained DSC scores of 86.91 and 83.40, while the V-Net model achieved DSC scores of 87.45 and 85.24. By utilizing a hybrid network model, we successfully enhanced the segmentation performance for both femoral and tibial cartilage.

5.3. Comparison on SKI-10 dataset

Table 2 presents the quantitative results for the SKI-10 MICCAI challenge knee cartilage segmentation dataset. According to the experimental evaluation, the FC and TC DSC scores for the UVNet model are 83.57 and 80.12, respectively. These values are 6.04 and 6.44 points higher than the U-Net and are 4.3 and 2.7 points higher than the baseline V-Net. Furthermore, we compared our work with U-Net++ and 3D U-Net, which have served as robust baselines for various image segmentation tasks. Our DSC scores for FC and TC surpass those of U-Net++ by 5.66 and 5.96 points, and outperform 3D U-Net by 5.4 and 3.47 points, respectively. Upon careful visual analysis of the qualitative results, it is evident that the UVNet model produces superior segmentation outcomes. In comparison to other network models, UVNet demonstrates better performance in terms of TC integrity segmentation results, which is often a crucial factor in clinical examinations.

5.4. Ablation experiment

We conducted two experimental directions on the OAI-ZIB dataset to validate the structural rationality of our model. The first direction explores the impact of varying the number of local MIP images on model performance, while the second direction examines the influence of different feature embedding methods on model performance.

(1) Different Number of Local MIP Images: To address the issue of information loss in global MIP images, we introduced local MIP images. A combination of local and global MIP images was used as input to the information augmentation subnet, U-Net. Table 3 illustrates the effect of the extracted number of local MIP images on the model's segmentation performance. Using too many local MIP images can interfere with the feature extraction of the backbone network, while using too few can result in significant information loss, essentially reverting to global MIP

Table 1
Quantitative valuation of different methods on the OAI-ZIB images

Method	DSC		Recall		Precision	
	FC	TC	FC	TC	FC	TC
U-Net	86.91	83.40	86.72	83.56	90.31	85.23
U-Net++	86.67	83.23	85.65	82.77	90.45	84.93
3D-UNet	88.76	85.79	87.46	84.71	90.58	87.64
V-Net	87.45	85.24	84.95	83.77	90.53	87.57
Our	90.35	86.79	90.39	86.38	90.43	86.95

Table 2
Quantitative valuation of different methods on the SKI-10 images

Method	DSC		Recall		Precision	
	FC	TC	FC	TC	FC	TC
U-Net	77.51	73.68	74.91	73.12	77.64	73.89
U-Net++	77.91	74.16	75.42	73.69	79.28	74.91
3D U-Net	78.17	76.65	78.81	77.81	82.32	78.75
V-Net	79.27	77.42	78.39	79.22	84.62	79.15
Our	83.57	80.12	83.92	81.52	83.48	81.39

Table 3
The influence of different number of local MIP image on the experimental results of UVNet

Method	DSC		Recall		Precision	
	FC	TC	FC	TC	FC	TC
One slices	88.57	86.12	87.27	85.62	87.12	85.89
Two slices	89.92	85.93	89.56	84.79	88.87	85.53
Three slices	90.35	86.79	90.39	86.38	90.43	86.95
Four slices	89.12	85.23	89.12	85.54	87.69	83.54

Table 4
The influence of different feature embedding methods on the experimental results of UVNet

Method	DSC		Recall		Precision	
	FC	TC	FC	TC	FC	TC
L: $V_1 V_2$ G: $V_3 V_4$	90.35	86.79	90.39	86.38	90.43	86.95
L: $V_3 V_4$ G: $V_1 V_2$	90.12	85.43	88.21	83.92	87.71	86.27
L: $V_1 V_2 V_3 V_4$	90.95	84.67	90.42	83.79	89.52	84.37
G: $V_1 V_2 V_3 V_4$	88.57	86.12	87.27	85.62	87.12	85.89

Table 5
Relative improvement of U-Net, V-Net, and UV-Net in DSC

Dataset	FC				TC			
	U-Net	V-Net	UVNet	(U-Net/V-Net)	U-Net	V-Net	UVNet	(U-Net/V-Net)
OAI-ZIB	86.91	87.45	90.35	3.44/2.9	83.40	85.24	86.79	4.7/1.55
SKI-10	77.51	79.27	83.57	6.04/4.3	73.68	77.42	80.12	6.44/2.7

images. It can be observed that the model achieves optimal performance when three slices of local MIP images are used. This approach effectively reduces information loss while minimizing interference during fusion with the backbone network. This method ensures a balanced presentation of cartilage structures and surrounding tissues, resulting in more precise segmentation results.

(2) Different Feature Embedding Methods: Another key challenge is how to embed local and global MIP feature maps into the V-Net. We compared four different embedding methods:

(a) L: $V_1 V_2$ G: $V_3 V_4$ represent embedding local MIP features into $V_1 V_2$ and global MIP features into $V_3 V_4$;

(b) L: $V_3 V_4$ G: $V_1 V_2$ represent embedding local MIP features into $V_3 V_4$ and global MIP features into $V_1 V_2$;

(c) L: $V_1 V_2 V_3 V_4$ represent embedding local MIP features into $V_1 V_2 V_3 V_4$;

(d) G: $V_1 V_2 V_3 V_4$ represent embedding global MIP features into $V_1 V_2 V_3 V_4$;

From Table 4, we can see that the model performs best when adopting the (a). This is attributed

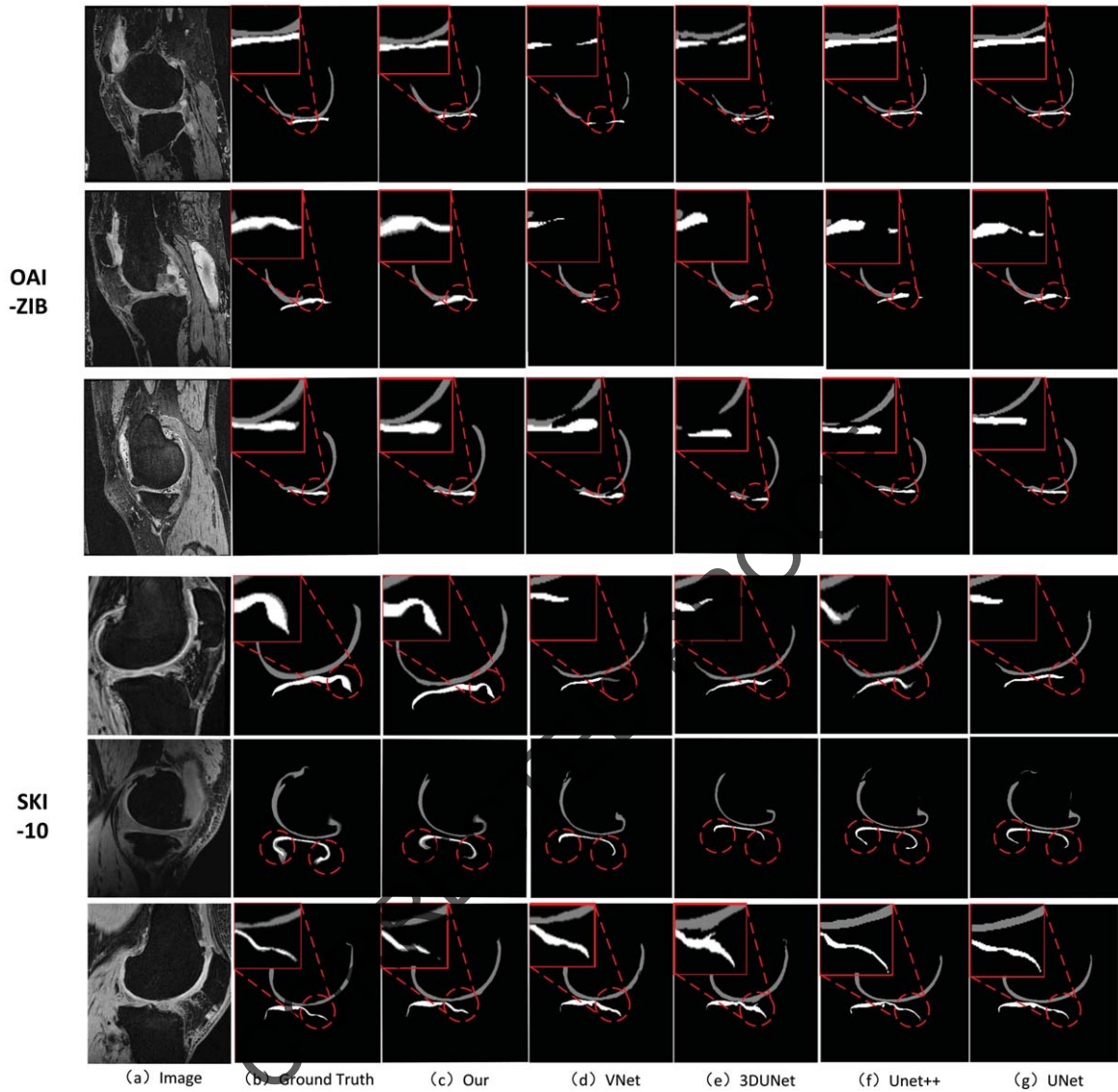


Fig. 5. The illustration of two sampled segmentation results of proposed method and the compared methods 3D U-Net, 3D V-Net, U-Net and U-Net++ on OAI-ZIB and SKI-10 datasets. The cartilage structure of the knee joint is described as: femoral cartilage (dark gray), tibial cartilage (light gray).

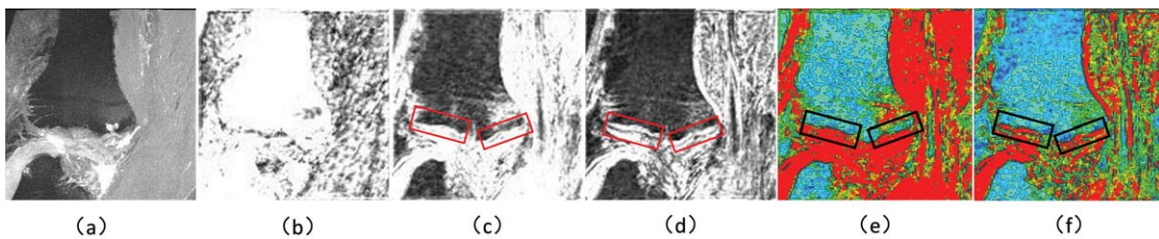


Fig. 6. Visualization of the feature map fusion process. (a) the MIP image, (b) the MIP feature map, (c) the feature map of V-Net after one downsampling, (d) the hierarchical features fused with the MIP feature, (e) Color fill for (c), (f) Color fill for (d).

to the fact that local MIP feature maps focus more on shape and boundary information, while global MIP feature maps emphasize positional information and the overall morphology of the cartilage. Embedding local MIP feature maps into the shallow features $V_1 V_2$ effectively supplements the boundary information lost in the middle cartilage, ensuring cartilage continuity. Embedding global MIP feature maps into the deep semantic features $V_3 V_4$ enhances cartilage localization and reduces morphological deficiencies in the cartilage structure.

5.5. Discussion

In knee MRI, cartilage is wrapped by other tissues, which are very similar in their morphology performance, and the cartilage occupies a small proportion in the whole image, which can easily cause the loss of boundary information during convolution and downsampling. The above phenomenon may lead to the problem of cartilage under-segmentation or even cartilage breakage in the final mask. To address the above problems, UVNet can compensate for missing boundaries by obtaining additional shape and position information from the MIP feature maps.

Figure 5 shows examples of cartilage segmentation on the OAI-ZIB and SKI-10 datasets by the proposed and comparative methods. It can be observed that our method shows smoother and more continuous cartilage with an overall morphology closer to the ground truth. As shown in the second and fourth rows of Fig. 5, TC tends to be under-segmented in the results of the baseline network and the other compared methods, while the segmentation loss of TC is compensated in our proposed method. Similarly, in the first, and third rows of Fig. 5, the cartilage mask obtained from the baseline network performs poorly, and some of the cartilage is even broken, whereas the UVNet corrects this defect and ensures the integrity of the cartilage morphology. It can be concluded that this is mainly attributed to the fact that the local and global MIP images are able to replenish the boundary information lost by downsampling, enhance the cartilage's morphological representation in the whole image, and make the cartilage's boundary more continuous. Our method can well guarantee the integrity of cartilage segmentation and reduce the possibility of cartilage breakage in the final segmentation mask.

In Section 3.2, we explain how to combine the morphological representation of knee cartilage to extract MIP images as an information supplement. In Section 3.3, it is explained how they can be fused with the hier-

archical features of V-Net and added to the training. Figure 6 demonstrates the advantages of our proposed method by visualizing the MIP feature map and the feature fusion process. The specific fusion process is done by the feature fusion module, and these representative feature maps are obtained by averaging feature maps of different channels. It can be observed that after one downsampling, the cartilage morphology of the feature map (c) becomes blurred and the boundary distinguishability is relatively poor, which will affect the accuracy of the subsequent segmentation. The fusion of the MIP feature map (b) with (c), which contains additional information, can supplement the lost cartilage boundary information in the feature map (c) and give the cartilage a more complete morphology. We can see that the boundary of cartilage in the fused feature map (d) is more recognizable, and the demarcation line between cartilage and bone as well as cartilage and other tissues is clearer. After color filling, we can observe this more clearly in (e) and (f). The richer expression of the information will be very helpful for subsequent segmentation.

6. Conclusion

In this study, we combine 2D U-Net with 3D V-Net to propose a new knee cartilage segmentation method, which we name UVNet. We propose for the first time to extract local MIP and global MIP based on the morphology of knee cartilage and apply them as complementary information in the segmentation to minimize information loss due to downsampling. Specifically, the local and global MIP images are feature extracted using U-Net, and the features are fused into the decoding subnetwork of V-Net through the local and global MIP feature fusion module, which effectively enhances the morphological representation of knee cartilage tissues in the images and ensures the completeness of cartilage segmentation. We evaluated our proposed method using two knee MRI datasets, OAI-ZIB (507 samples) and SKI-10 (100 samples). Experimental results show that UVNet performs significantly better than other segmentation methods.

In our future work, we will focus more on simplifying the model architecture while retaining its segmentation capability. The limitation of UVNet is that it has more parameters and longer training and inference times. We plan to use a 2D network with smaller parameters to segment MIP images to reduce the overall number of parameters in the model

without compromising segmentation performance. In addition, we will collect more clinical patient data, such as knee data from the Johnston County Osteoarthritis Program (JoCoOA) [49], to validate our proposed segmentation method and extend the application area of our method.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61806107, 61702135 and 62201314, the Opening Project of State Key Laboratory of Digital Publishing Technology, and the Shandong Province “Double-Hundred Talent Plan” on 100 Foreign Experts and 100 Foreign Expert Teams Introduction (WST2021020).

References

- [1] M.T. Nieminen, V. Casula, M.T. Nevalainen and S. Saarakkala, Osteoarthritis year in review 2018: imaging, *Osteoarthritis and Cartilage* **27**(3) (2019), 401–411.
- [2] C.R. Chu, A.A. Williams, C.H. Coyle and M.E. Bowers, Early diagnosis to enable early treatment of pre-osteoarthritis, *Arthritis Research & Therapy* **14**(3) (2012), 212.
- [3] E. Yelin, S. Weinstein and T. King, The burden of musculoskeletal diseases in the United States, in *Seminars in Arthritis and Rheumatism* **46**(3) (2016), 259.
- [4] D.T. Felson, An update on the pathogenesis and epidemiology of osteoarthritis, *Radiol Clin North Am* **42**(1) (2004), 1–9.
- [5] P. Sarzi-Puttini, et al., Osteoarthritis: An overview of the disease and its treatment strategies, *Semin Arthritis Rheum* **35**(1 Suppl. 1) (2005), 1–10.
- [6] P. Creamer and M.C. Hochberg, Osteoarthritis, *Lancet* **350** (1997), 503–508.
- [7] V.B. Kraus, F.J. Blanco, M. Englund, M.A. Karsdal and L.S. Lohmander, Call for standardized definitions of osteoarthritis and risk stratification for clinical trials and clinical use, *Osteoarthritis Cartilage* **23** (2015), 1233–1241.
- [8] F. Eckstein, F. Cicuttini, J.-P. Raynaud, J.C. Waterton and C. Peterfy, Magnetic resonance imaging (MRI) of articular cartilage in knee osteoarthritis (OA): Morphological assessment, *Osteoarthr Cartil* **14** (2006), 46–75.
- [9] J. Lee, S. Gumus, C. Moon, C.K. Kwok and K.T. Bae, Fully automated segmentation of cartilage from the MR images of knee using a multi-atlas and local structural analysis method, *Medical Physics* **41**(9) (2014), 092303.
- [10] A.A. Gatti, & M.R. Maly, Automatic knee cartilage and bone segmentation using multi-stage convolutional neural networks: data from the osteoarthritis initiative, *Magnetic Resonance Materials in Physics, Biology and Medicine* **34** (2021), 859–875.
- [11] Y. Du, R. Almajalid, J. Shan and M. Zhang, A novel method to predict knee osteoarthritis progression on MRI using machine learning methods, *IEEE Transactions on Nanobioscience* **17**(3) (2018), 228–236.
- [12] O. Ronneberger, P. Fischer and T. Brox, U-net: Convolutional networks for biomedical image segmentation[C]. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, 2015:234–241.
- [13] F. Milletari, N. Navab and S. Ahmadi, V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, *2016 Fourth International Conference on 3D Vision (3DV)*, (2016), 565–571.
- [14] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, et al., Convolutional neural networks for medical image analysis: full training or finetuning, *IEEE Trans Med Imaging* **35** (2016), 1299–312. doi: 10.1109/TMI.2016.2535302
- [15] K. Kamnitsas, C. Ledig, V.F.J. Newcombe, J.P. Simpson, A.D. Kane, D.K. Menon, et al., Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, *Med Image Anal* **36** (2017), 61–78. doi: 10.1016/j.media.2016.10.004
- [16] D.M. Vigneault, W. Xie, C.Y. Ho, D.A. Bluemke and J.A. Noble, Ω -Net (Omega-Net): fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks, *Med Image Anal* **48** (2018), 95–106.
- [17] Cao Shi, Canhui Xu, Jianfei He, Yinong Chen, Yuanzhi Cheng, Qi Yang and Haitao Qiu, Graph-based convolution feature aggregation for retinal vessel segmentation, *Simulation Modelling Practice and Theory* **121** (2022), 102653.
- [18] B. Woo, C.B. Engstrom, W. Baresic, J. Fripp, S. Crozier and S.S. Chandra, Automated anomaly-aware 3D segmentation of bones and cartilages in knee MR images from the Osteoarthritis Initiative, *ArXiv, abs/2211.16696*, (2022).
- [19] J. Mao, P. Men, H. Guo and J. An, Region-based two-stage MRI bone tissue segmentation of the knee joint, *IET Image Process* **16** (2022), 3458–3470.
- [20] H. Chen, N. Zhao, T. Tan, Y. Kang, C. Sun, G. Xie, N. Verdonchot and A.M. Sprengers, Knee Bone and Cartilage Segmentation Based on a 3D Deep Neural Network Using Adversarial Loss for Prior Shape Constraint, *Frontiers in Medicine* **9** (2022).
- [21] F. Ambellan, A. Tack, M. Ehlke and S. Zachow, Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative, *Medical Image Analysis* **52** (2019), 109–118.
- [22] H.S. Lee, H. Hong and J. Kim, BCD-NET: A novel method for cartilage segmentation of knee MRI via deep segmentation networks with bone-cartilage-complex modeling, *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (2018), 1538–1541.
- [23] W. Dai, B. Woo, S. Liu, M. Marques, F. Tang, S. Crozier, C. Engstrom and S.S. Chandra, Can3d: Fast 3d Knee MRI Segmentation Via Compact Context Aggregation, *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, (2021), 1505–1508.
- [24] A. Raj, S. Vishwanathan, B. Ajani, K. Krishnan and H. Agarwal, Automatic knee cartilage segmentation using fully volumetric convolutional neural networks for evaluation of osteoarthritis, *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, (2018), 851–854.
- [25] M. Sun, L. Lu, I.A. Hameed, C.P. Kulseng and K. Gjesdal, Detecting Small Anatomical Structures in 3D Knee MRI Segmentation by Fully Convolutional Networks, *Applied Sciences* (2021).

- [26] L. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, *European Conference on Computer Vision* (2018).
- [27] S.S. Sengar, C. Meulengracht, M.P. Boesen, A.F. Overgaard, H. Gudberg, J.D. Nybing and E.B. Dam, UNet Architectures in Multiplanar Volumetric Segmentation – Validated on Three Knee MRI Cohorts, *ArXiv, abs/2203.08194*, (2022).
- [28] Z. Zhou, M.M. Siddiquee, N. Tajbakhsh and J. Liang, UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation, *IEEE Transactions on Medical Imaging* **39** (2019), 1856–1867.
- [29] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. Chen and J. Wu, UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation, *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2020), 1055–1059.
- [30] D.A. Kessler, J.W. MacKay, V.A. Crowe, F.M. Henson, M.J. Graves, F.J. Gilbert and J.D. Kaggie, The optimisation of deep neural networks for segmenting multiple knee joint tissues from MRIs, *Computerized Medical Imaging and Graphics* **86** (2020).
- [31] P. Isola, J. Zhu, T. Zhou and A.A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 5967–5976.
- [32] D.A. Kessler, J.W. MacKay, S.M. McDonnell, R.L. Janiczek, M.J. Graves, J.D. Kaggie and F.J. Gilbert, Segmentation of Knee MRI Data with Convolutional Neural Networks for Semi-Automated Three-Dimensional Surface-Based Analysis of Cartilage Morphology and Composition, *Osteoarthritis Imaging*, (2022).
- [33] J.W. MacKay, J.D. Kaggie, G.M. Treece, S.M. McDonnell, W.S. Khan, A.R. Roberts, R.L. Janiczek, M.J. Graves, T.D. Turmezei, A.W. McCaskie and F.J. Gilbert, Three-Dimensional Surface-Based Analysis of Cartilage MRI Data in Knee Osteoarthritis: Validation and Initial Clinical Application, *Journal of Magnetic Resonance Imaging* **52** (2020).
- [34] S. More and J. Singla, Discrete-MultiResUNet: Segmentation and feature extraction model for knee MR images, *J Intell Fuzzy Syst* **41** (2021), 3774–3781.
- [35] Y. Liu, H.-S. Kwak and I.-S. Oh, Cerebrovascular Segmentation Model Based on Spatial Attention-Guided 3D Inception U-Net with Multi-Directional MIPs, *Appl Sci* **12** (2022), 2288.
- [36] H. Marquis, D. Deidda, A. Gillman, K. Willowson, Y. Gholami, T. Hioki, E. Eslick, K. Thielemans and D. Bailey, Theranostic SPECT Reconstruction for Improved Lesion Dosimetry in Radionuclide Therapy, *J Nucl Med* **62**(Suppl. 1) (2021), 1533.
- [37] S. Li, Y. Zhao and Y. Ye, Improved minimum intensity projection in holographic reconstruction via SNR-enhanced holography, *J Mod Opt* **68** (2021), 322–326.
- [38] N. Kawel, B. Seifert, M. Luetolf and T. Boehm, Effect of Slab Thickness on the CT Detection of Pulmonary Nodules: Use of Sliding Thin-Slab Maximum Intensity Projection and Volume Rendering, *Am J Roentgenol* **192** (2009), 1324–1329.
- [39] S. Fujii, E. Matsusue, Y. Kanasaki, Y. Kanamori, J. Nakanishi, S. Sugihara, J. Kigawa, N. Terakawa and T. Ogawa, Detection of peritoneal dissemination in gynecological malignancy: Evaluation by diffusion-weighted MR imaging, *Eur Radiol* **18** (2008), 18–23.
- [40] S. Jadhav, G. Deng, M. Zawin and A.E. Kaufman, COVID-view: Diagnosis of COVID-19 using Chest CT, *IEEE Trans Vis Comput Graph* 2021.
- [41] F. Yousefirizi, P. Martineau, C. Uribe and A. Rahmim, Enhancement of conventional segmentation techniques to achieve deep framework performance for lymphoma lesion segmentation in PET images, *J Nucl Med* **62** (Suppl. 1) (2021), 1427.
- [42] T. Heimann, B. Morrison, M.A. Styner, M. Niethammer and S. Warfield, Segmentation of Knee Images: A Grand Challenge, (2010).
- [43] F. Ambellan, A. Tack, M. Ehlke and S. Zachow, Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative, *Medical Image Analysis* **109**(52) (2019), 118.
- [44] D.P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, *CoRR, abs/1412.6980*, 2014.
- [45] L.R. Dice, Measures of the amount of ecologic association between species, *Ecology* **26**(3) (1945), 297–302.
- [46] J.A. Swets, Measuring the accuracy of diagnostic systems, *Science* **240**(4857) (1988), 1285–1293.
- [47] D.M.W. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation, *Journal of Machine Learning Technologies* **2**(1) (2011), 37–63.
- [48] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox and O. Ronneberger, 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2016).
- [49] J.M. Jordan, C.G. Helmick, J.B. Renner, G. Luta, A.D. Dragomir, J. Woodard, F. Fang, T.A. Schwartz, L.M. Abbate, L.F. Callahan, W.D. Kalsbeek and M.C. Hochberg, Prevalence of knee symptoms and radiographic and symptomatic knee osteoarthritis in African Americans and Caucasians: the Johnston County Osteoarthritis Project, *The Journal of Rheumatology* **34**(1) (2007), 172–80.