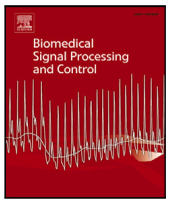
Biomedical Signal Processing and Control 106 (2025) 107754

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Patch Attention U-Net for knee cartilage segmentation in magnetic resonance images

Xiang Wang, Cao Shi *

School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, Shandong 266000, People's Republic of China

ARTICLE INFO

Keywords:
Deep learning
Encoder–decoder
Hybrid loss
Knee cartilage segmentation
Patch Attention

ABSTRACT

Knee cartilage segmentation in magnetic resonance images is a challenging task with significant clinical implications for the diagnosis and treatment of osteoarthritis. Recent advances in deep convolutional neural networks (CNNs) have shown promise in improving the accuracy of knee segmentation. However, CNNs often struggle with small samples with irregular shapes, which may result in the omission of important features such as cartilage. In this study, we propose a novel network for knee cartilage segmentation that incorporates Patch Attention (PA) block to improve the ability of network to detect small objects and a Feature Aggregation block to fuse the features from same level of the encoder and previous layer of the decoder. The PA block includes Patch-based Channel-wise Attention block and Patch-based Patch-wise Attention block, which capture intra-channel and intra-patch relationships, respectively. Our method is evaluated on publicly available datasets the 2010 Grand Challenge Knee Image Segmentation (SKI-10) dataset and Osteoarthritis Initiative (OAI) dataset, and the results demonstrate that our approach achieves impressive performance in knee cartilage segmentation.

1. Introduction

Knee cartilages are soft tissues that exist in the middle of knee bones. It is well established that knee cartilages play a crucial role in supporting and protecting the knee joint. Segmentation of knee cartilage is a necessary pre-processing step for knee osteoarthritis analysis, which can significantly improve the treatment of knee joint diseases. Compared to other imaging techniques, magnetic resonance (MR) imaging shows a higher level of specificity and sensitivity to obtain the biomedical markers of knee cartilage. Despite the availability of numerous auto-segmentation methods, the most commonly used method in clinical practice is manual annotation, which is a time-consuming and prone to intra-observer variability. Therefore, there is an urgent need for an accurate and reliable segmentation method to make the workflow in clinical scenarios more efficient.

In recent years, traditional machine learning models, such as k-nearest neighbor classifiers and support vector machines [1], have been widely employed for cartilage segmentation. However, these models often require a significant amount of computation during the learning process. After this, some methods have incorporated human prior knowledge into the initialization stage, while others [2–4] have proposed multi-stage or multi-level classification models using various techniques such as bone pre-segmentation [3,5], edge classification [6], and leveraging multiple types of images to extract rich features [5]. Although these methods have performed well, they are limited by the processed data and vary significantly across datasets, failing to generalize across cartilage images from different people and cameras, which undermines the robustness of the model. Given these limitations, there is a growing interest in exploring deep learning techniques for knee cartilage segmentation, which show great potential in overcoming some of the challenges inherent in traditional machine learning methods.

The deep learning (DL) methods, particularly CNNs [7–11], have shown great promise in improving the accuracy of knee cartilage segmentation. CNN-based approaches offer a feature learning scheme that can automatically extract informative features without relying on manual feature engineering, making them an attractive choice for medical image segmentation. Several state-of-the-art DL methods have been proposed, including 2D [8,9,11] and 3D CNNs [10,12,13], as well as semi-supervised learning frameworks [14]. For instance, 2D U-Net [9] and 2D SegNet models [7] have been used to segment various sub-compartments of the knee, including articular cartilage and meniscus. However, accurate segmentation of knee cartilage is often hindered by small sample and uncertainties in inter-cartilage boundaries and adjacent structures. To address this challenge, ongoing research efforts are focusing on developing more robust and accurate DL methods. So more and more approaches introduce attention mechanism into the network. Le et al. [15] proposed a two-branch architecture that combines region and contour information with a narrow band active contour attention model, which focuses on the informative regions

* Corresponding author.
E-mail address: caoshi@yeah.net (C. Shi).

https://doi.org/10.1016/j.bspc.2025.107754
Received 27 July 2023; Received in revised form 16 January 2024; Accepted 15 February 2025
Available online 1 March 2025

while suppressing the noisy regions. Yu et al. [16] presented a 3D segmentation method that uses a hierarchical transformer encoder with local spatial representation learning and a multi-stage framework, which improves the network's ability to detect small samples. Despite the promising results achieved by attention mechanisms in knee cartilage segmentation, there is still room for improvement in terms of the design and implementation of attention blocks. Inspired by [17–19], we redesign channel attention and spatial attention to change the poor performance of small sample segmentation caused by applying attention mechanism on the whole feature map in previous work. In this paper, we propose a novel knee cartilage segmentation method based on a patch attention block composed of Patch-based Channel-wise Attention (PCA) block and Patch-based Patch-wise Attention (PPA) block. We construct the PCA block by sequentially operating a SE block on patches to obtain the corresponding weights of each channel. This channel-level block exploits the relationships among each channel of patches. The PPA block works to find intra-patch relationships of the feature map using Multilayer Perceptron (MLP). The incorporation of the patch attention mechanism enables the network to focus better on key or small regions in images.

Our contributions in this work can be summarized as:

1. We propose a semantic segmentation network to address the issue of erroneous segmentation of small samples in knee cartilage, termed PA-UNet, that evokes channel attention and spatial attention on the patch convolutional features.
2. We propose a PCA block and a PPA block, which utilize the contextual information and the spatial information of patch feature maps rather than the whole feature maps, to segment target completely, and also propose a new Feature Aggregation (FA) block which fuses the features from the encoder and the decoder.
3. PA-UNet is one of the very first attempts to introduce both channel attention and spatial attention into knee cartilage segmentation. The comparative experiments show that ours obtains an improvement over other medical image segmentation networks.

2. Related work

2.1. Medical image segmentation

For decades, researchers have explored various automated and semi-automated techniques for segmenting the knee joint. Among these, convolutional neural networks (CNNs) have emerged as a powerful tool for medical image segmentation, achieving performance levels approaching those of human experts. One example is the work of Liu et al. [8], which proposed a method that combines CNNs with 3D simplex deformable models to achieve more accurate and efficient musculoskeletal tissue segmentation. Another promising approach is the combination of statistical shape models and CNNs, as demonstrated by Ambellan et al. [10] in their successful segmentation of knee bones and cartilages. Tan et al. [20] proposed a collaborative method that first extracts regions of interest (ROIs) for three cartilage areas and then fuses them to generate fine-grained segmentation results.

Recent studies have investigated the potential of self-attention mechanisms in capturing long-range dependencies in feature maps. One such study by Zheng et al. [21] proposed a 3D segmentation method that employs a meta-learner [22] to ensemble three 2D models and one 3D model known as base-learners. The base-learners and meta-learner are trained using labeled data and pseudo-labels, respectively, to obtain a 3D segmentation result. However, obtaining fully annotated 3D data is both costly and scarce, and hence, Zheng et al. proposed a sparse annotation strategy in a later study [23]. This strategy involves selecting the most representative 2D slices for annotation based on their low-dimensional encoding and representativeness in a set of 3D images. Moreover, Zheng et al. [24] further designed a K-head FCN to compute the pseudo-label uncertainty of each slice, which is then used to discard highly uncertain pixels from the subsequent training process. Specifically, the K-head FCN generates K different pseudo-labels for each slice, and the uncertainty of each pixel is computed as the variance of these K pseudo-labels. The pixels with high uncertainty scores are excluded from the training set to improve the accuracy and robustness of the segmentation model.

2.2. Attention mechanism

Channel attention mechanisms are a class of techniques that recalibrate channel-wise features by exploiting inter-channel dependencies in a data-driven manner. One of the seminal works in this field is the Squeeze-and-Excitation (SE) block, which enhances the representational power of convolutional neural networks (CNNs) through dimensionality reduction [25,26]. However, the SE block comes with a trade-off between performance improvement and computational cost, as it increases the model complexity. To address this issue, a recent study proposed the Efficient Channel Attention (ECA) mechanism, which achieves similar or better performance compared to SE blocks with significantly fewer parameters [27]. Channel attention mechanisms have been widely applied to various medical image segmentation tasks, demonstrating their effectiveness and robustness [28,29].

Spatial attention is a mechanism that explores pixel relationships in the spatial domain by calculating the similarity between each pair of pixels in feature maps. It was initially introduced in Non-local networks [30] and Transformer [31] to enhance the feature representation of each pixel based on its similarity with others. While the spatial self-attention mechanism overcomes the limitations of fixed-range approaches [32,33] that fail to capture long-range dependencies, it also incurs high computational costs due to the need to process the entire feature map for each pixel. Specifically, the attention similarity for each pixel affects all other pixels, leading to high computational costs. Recently, several methods [34,35] have been proposed to optimize the GPU memory costs of spatial self-attention.

Both spatial and channel attention mechanisms have demonstrated promising results in various medical image segmentation tasks [34–36], as they effectively capture long-range dependencies that are typically missed by multiple fixed-range approaches [32,33]. However, combining these two mechanisms presents several challenges, as they may produce conflicting results when applied separately or in sequence. For example, a pixel belonging to a semantic class within a small region in the feature maps may be highlighted by spatial attention but ignored by channel attention, as its channel representation may not be critical from the entire feature maps' perspective. Conversely, a pixel belonging to a semantic class within a large region in the feature maps may be suppressed by spatial attention but amplified by channel attention, as its channel representation may dominate the feature maps. Hence, integrating spatial and channel attention is not a straightforward task and requires careful design and evaluation [37].

3. Method

In this section, we present the proposed Patch-based Attention U-Net (PA-UNet) in detail. We first provide an overview of PA-UNet in Section 3.1. followed by an introduction to each component of PA-UNet from Section 3.2 to Section 3.5. Next, we give the joint loss in Section 3.6. Finally, we discuss the experimental details for reproduction in Section 3.7.

3.1. Overview

This section provides a detailed introduction to our method PA-UNet. PA-UNet is based on two main blocks: Patch-based Attention

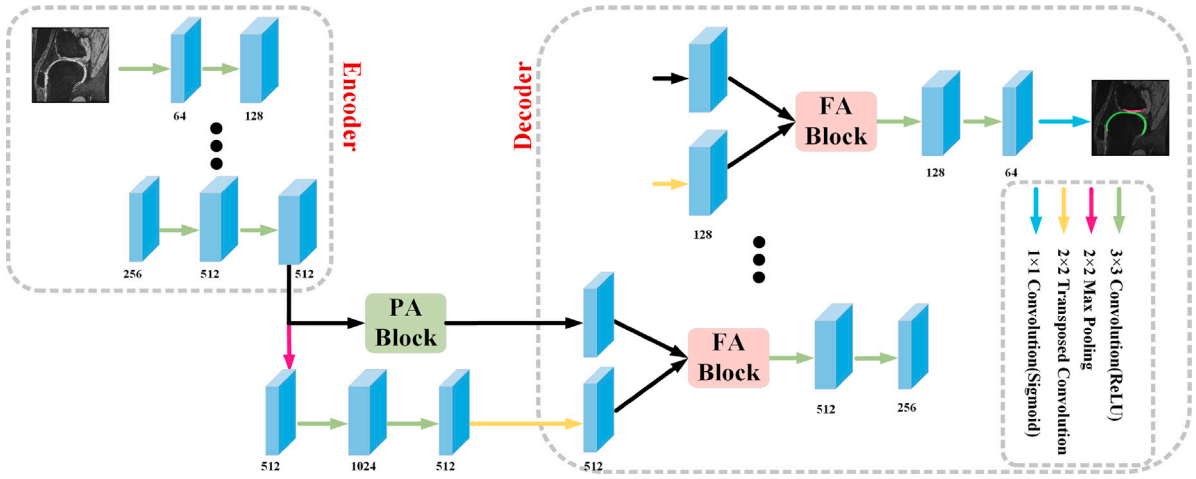


Fig. 1. The Structure of PA-UNet. It combines PA block for enhancing feature representation in the encoder stage and FA block for feature fusion in the decoder stage.

(PA) block and Feature Aggregation (FA) block. In particular, Patch-based Channel-wise Attention (PCA) block and Patch-based Patch-wise Attention (PPA) block are developed in PA block.

The overall framework of PA-UNet is illustrated in Fig. 1. An image $I \in \mathbb{R}^{H\times W\times C}$ is taken as the input of U-Net, where H, W and C represent the height, the width and the number of channels of the image I, respectively. First, the encoder of the U-Net efficiently downsamples the input image by several consecutive convolutional blocks to generate different levels of feature maps. And between every two stages of the encoder from layer 2 to layer 5, the PA blocks take the features from low-level and enhance features representation of the encoder for same-level branch of the decoder. Then, as the core of the PA-UNet, the FA block of decoder takes the feature maps from encoder and decoder as inputs for feature aggregation. In the encoder, the PCA block considers the relationships among the neighboring channels within the local area and the PPA block further considers the relationships among patches as well. In the decoder, the FA block completes the effective fusion of features from the encoder and decoder.

3.2. Patch-based attention block

For semantic segmentation of knee cartilage, it is important to encode both the spatial and contextual information of the knee joint for predicting score maps. A common issue in knee cartilage segmentation is that the cartilage regions of interest can be of different sizes and shapes, and if the entire knee cartilage is treated as a whole, some local information may be lost. To address this, we propose a PA block to consider the relationships among neighboring channels within each local patch and the relationships among patches in the feature maps. Unlike most attention methods that focus on the entire knee joint image, our method focuses on each local patch extracted from the input features. This block mainly consists of a PCA block and a PPA block to enhance the contextual information and spatial details, respectively. Fig. 2 shows the proposed PA block. We apply PCA block and then another step of PPA block sequentially. The PA block takes feature $F \in \mathbb{R}^{C\times H\times W}$ and outputs feature map of the same size as the input.

3.3. Patch-based channel-wise attention

In knee cartilage segmentation, many channels are generated by convolution operations, but some may contain less relevant feature information. To address this issue, the channel attention mechanism is used to distinguish the importance of different channels. However, traditional channel attention methods, such as SE Net, rely on spatial pooling operations like max pooling or average pooling, which can lead to a loss of contextual information. To address this problem, we propose a PCA block for knee cartilage segmentation. The PCA block divides the input knee joint image into patches and deploys a SE block on each patch to obtain intra-channel relationships. Each patch then reassigns weights to its channels to obtain a channel-weighted feature map that preserves contextual information. This approach improves the accuracy of knee cartilage segmentation while minimizing the computational cost.

The overall structure of the PCA block is shown in the left part of Fig. 2. The input to the block is the feature map $F \in \mathbb{R}^{C\times H\times W}$ obtained from the encoding convolutional block, and it outputs a channel-weighted feature map with the same size as the input. Firstly, the $C\times W\times H$ feature map F is divided into a collection of $C\times ps\times ps$ patches P, where C, W, H and ps represent the channel, width, height and patch size of the patch feature map respectively.

$$F = [P_1, P_2, \dots, P_N] \tag{1}$$

where $N = \frac{H}{ps}\times\frac{W}{ps}$. Each patch passes through a SE block separately to obtain a reweighted patch feature. In this work, we set the patch size to be a hyperparameter and found that 32 works best. Since the SE block is common used, we simplified its structure with its name in Fig. 2. In this process, the SE block transforms the input $P \in \mathbb{R}^{C\times ps\times ps}$ to feature map $U \in \mathbb{R}^{C\times ps\times ps}$. In the notation that follows we take transformation F_{tr} to be a convolutional operator and use $V = [v_1, v_2, \dots, v_C]$ to denote the learned set of filter kernels, where v_c refers to the parameters of the cth filter. We can then write the outputs as $U = [u_1, u_2, \dots, u_C]$, where

$$u_c = v_c * P = \sum_{s=1}^{C} v_v^s * p^s \tag{2}$$

Here denotes convolution, $v_c = [v_c^1, v_c^2, \dots, v_c^C]$, $P = [p^1, p^2, \dots, p^C]$ and $u_c \in \mathbb{R}^{H\times W}$. Finally, all Ps are stacked to their original position, resulting in the weighted feature map F' where

$$F' = [U_1, U_2, \dots, U_N] \tag{3}$$

3.4. Patch-based patch-wise attention

Semantic segmentation of knee cartilage in medical images suffers greatly from the problem of incomplete division, since the classification of cartilage regions is a comprehensive task affected by both the size and the number of the cartilage samples. If we treat the knee image as a whole, the small cartilage samples in the image may have a very small or even negligible weight contribution in the entire training process. Meanwhile, besides focusing on the relation inside the cartilage patch, the link between cartilage patches is also very tight. Intuitively, when a cartilage sample shows up at a certain position, its neighboring position

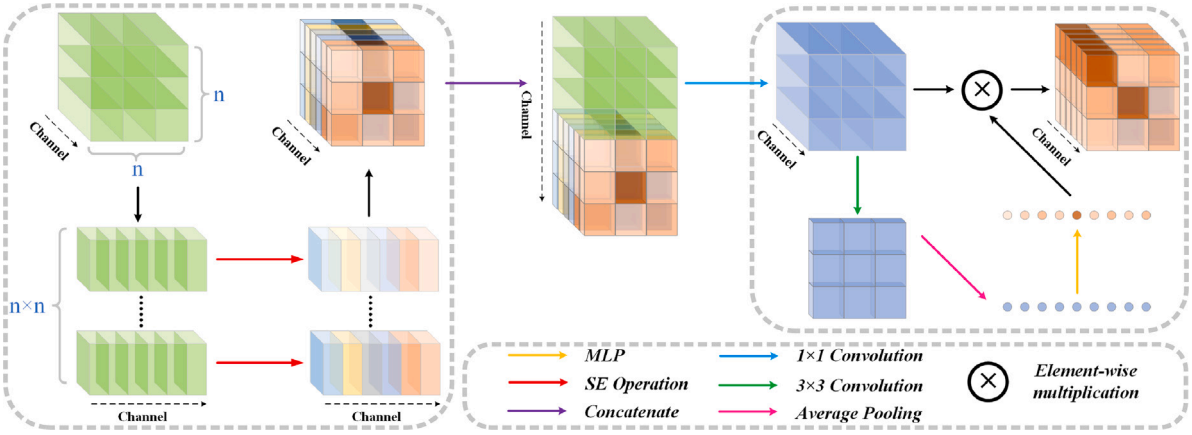


Fig. 2. The whole structure of PA block consists of PCA block showed in the left part and PPA block described in the right part. The two blocks are connected by concatenating the output of PCA and the original feature map along the channel axis, and applying a 1 × 1 convolution layer to decrease the channel size.

in the knee probably has a cartilage sample too. Thus, establishing the connection between cartilage patches is especially important. Based on the above analysis, we propose the PPA block which considers the relation among the cartilage patches. It assigns different weights to each patch, and through this way, we obtain the intra-patch relationship feature map.

Next, we discuss the design of a PPA block, which is similar to the structure of the PCA block. To generate the PPA feature map, we first concatenate the channel refined feature $F' \in \mathbb{R}^{C\times H\times W}$ and original image from PCA. Next, we pass the feature map through 1×1 convolutional layer $Conv_s$ to obtain the input feature map $F_{in} \in \mathbb{R}^{C\times H\times W}$ of the PPA.

$$F_{in} = Conv_s([F'; F]) \tag{4}$$

Next, we extract patch $F_j^{in} \in \mathbb{R}^{1\times H\times W}$ from the feature F_{in} through convolutional layer $Conv_j$ with a kernel size of 1×1 following a ReLU activation, where j denotes the patch of jth position in F_{in}. Notably, F_j^{in} and P_i represent the same area of the scene if $j = i$. We feed this feature map F_j^{in} to a corresponding average-pool layer with kernel size of $ps \times ps$. This operation aggregates the spatial information to generate a spatial attention $E_{in} \in \mathbb{R}^{1\times \frac{H}{ps}\times \frac{W}{ps}}$.

$$E_{in} = AvgPool(\sigma(Conv_j(F_{in}))) \tag{5}$$

Then we reshape $E_{in} \in \mathbb{R}^{1\times \frac{H}{ps}\times \frac{W}{ps}}$ into $M_{in} \in \mathbb{R}^{1\times \frac{HW}{ps^2}}$. The M_{in} is then passed through a MLP to reassign the attention weight of each patch. Finally, we reshape the M_{in} to obtain $M_{out} \in \mathbb{R}^{1\times \frac{H}{ps}\times \frac{W}{ps}}$. After expanding M_{out} over both channel and spatial dimension, we multiply it with F_{in} to obtain a patch-wise refined feature F''.

3.5. Feature aggregation

Considering striking a balance between segmentation accuracy and training consumption, we devise a Feature Aggregation block, termed FA, which effectively combines features from encoder and decoder. Notably, the features from the encoder undergo processing by the PA block, thereby incorporating additional contextual and spatial information during information aggregation. The FA architecture is depicted in Fig. 3, where the feature $T \in \mathbb{R}^{C\times H\times W}$ from encoder and feature $D \in \mathbb{R}^{C\times H\times W}$ from decoder are first enhanced by Feature Enhancement (FE) block to yield $T' \in \mathbb{R}^{C\times H\times W}$ and $D' \in \mathbb{R}^{C\times H\times W}$, respectively. Following the acquisition of the enhanced features, our primary objective was to devise an efficient approach to fuse the two features. Thus we introduce a Feature Selection (FS) block to facilitate feature fusion, which is designed to compute a set of weights that determine the relative contributions of the enhanced features, rather than selecting a single feature. Specifically, the enhanced features $T' \in \mathbb{R}^{C\times H\times W}$ and $D' \in \mathbb{R}^{C\times H\times W}$ are concatenated and input into FS block, yielding two attention feature maps $P \in \mathbb{R}^{1\times H\times W}$ and $Q \in \mathbb{R}^{1\times H\times W}$ that guide the feature fusion process. The fused feature map $O \in \mathbb{R}^{C\times H\times W}$ is ultimately computed through element-wise addition of the weighted features.

It is worth noting that the proposed FA block is able to better integrate information from the encoder and decoder on single scale at a time. In previous works on natural scene tasks, an important challenge is to accurately classify all objects appearing in an image. Multi-scale information provide richer spatial context information, which helps to better understand the relationships between objects and backgrounds in an image. However, in medical image segmentation tasks, the object categories are limited, and the positions of objects and backgrounds are basically fixed. In addition, due to the differences in object shapes and sizes in natural scene, multi-scale features may better capture all the details of all objects. However, in the case of cartilage segmentation, cartilage varies between different slices, the sizes and shapes of cartilage can be divided into very few categories when considering the entire dataset. Finally, in practical applications of natural scene images, the size and resolution of images may differ due to factors such as the shooting equipment and distance, whereas medical images are captured using fixed-position devices with the same parameters, and there is few drastic variation owing to differences in the subjects being imaged. Therefore, considering lower computational complexity and memory usage of single-scale makes it even more suitable for our task.

The FA block in our network architecture comprises two main blocks, the FE block and the FS block, which aim to enhance feature representation from the encoder or decoder and fuse the feature information, respectively. The FS block consists of a 1×1 convolutional layer $Conv_{1\times1}$ followed by a sigmoid activation function σ. This generates an attention map with two channels (P and Q), each corresponding to the enhanced feature map (T' or D') from the encoder or decoder, respectively. Therefore, this process is formulated as:

$$[P, Q] = \delta(Conv_{1\times1}([T', D'])) \tag{6}$$

where the $P \in \mathbb{R}^{1\times H\times W}, Q \in \mathbb{R}^{1\times H\times W}$ are the output of the FS block. Then the feature map T from encoder or D from decoder is fed into the FE block which consists of three branches. The first branch passes through a base block to enrich the feature representation, while the second branch does not undergo any processing. The base block contains two 3×3 convolutional layers $Conv_{3\times3}$ followed by a BN layer. The third branch is processed by the FS block to adaptively learn two weight tensors for first branch and second branch. We denote η as the activate function $ReLU$. So this process can be formulated as:

$$T' = \eta(P \otimes T) + \eta(Q \otimes Conv_{3\times3}(Conv_{3\times3}(T))) \tag{7}$$

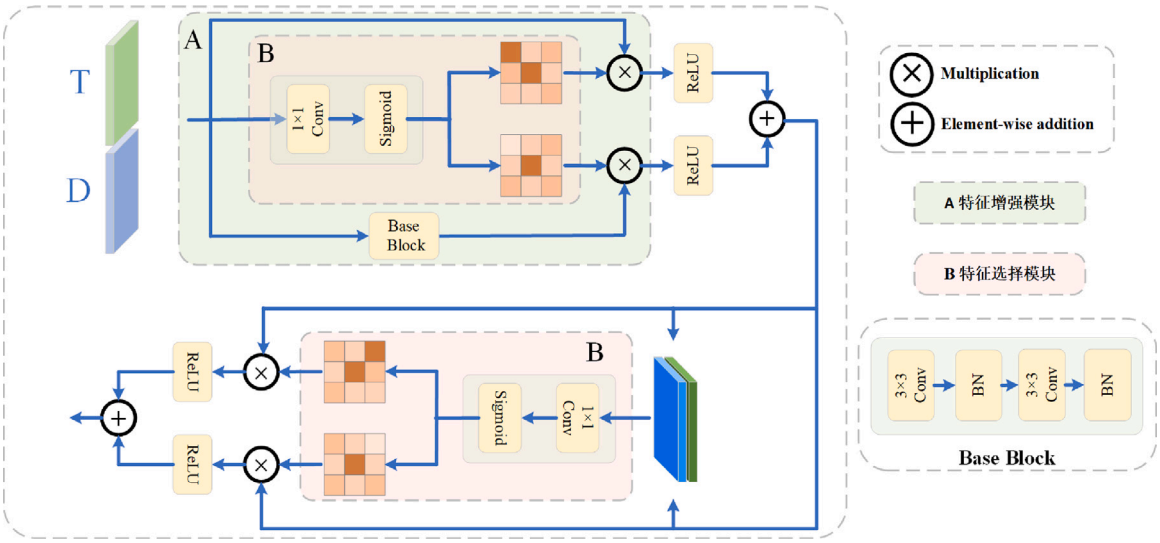


Fig. 3. The Structure of FA Block. The feature map T represents the output of the attention block of the encoder, and the D is the feature map from previous layer of the encoder.

where the $\otimes$ means the element-wise multiplication operation and the D' is obtained in the same way. So the final output O can be obtained by:

$$O = \eta(P \otimes T') + \eta(Q \otimes D') \tag{8}$$

3.6. Loss function

Small sample segmentation is a challenge in semantic segmentation especially in medical images [38]. The difficulty is mainly caused by the extremely imbalanced sample distribution. In our experimental datasets, the object, small-volumed femoral and tibial cartilages, account for nearly 1/100000 of the whole MR images. Since the semantic segmentation that is a pixel classification task essentially, such small sample contributes minimally to loss function. In our design, we employ a hybrid loss function consisting of contributions from focal loss [39], dice loss [40] and boundary loss [41] to address this issue. The focal loss introduces a factor $(1 - p_t)^{\gamma}$ based on cross entropy loss for class imbalance encountered during training process. The dice loss addresses this problem by turning pixel-wise classification problem into class-wise distance problem. Both dice loss and focal loss are region-based, but we argue that a boundary loss can mitigate the issues related to regional losses in highly unbalanced segmentation problems. Therefore, we introduce a boundary loss that takes the form of a distance metric on the space of contours (or shapes), not regions.

We feed source images(x^s) into the encoder(E) and pixel-level classifier (C) to generate prediction maps. Then we optimize the whole segmentation network by the total loss and the total loss can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{seg} &= \mathcal{L}_{focal}(C_i(E(x^s)), y^s) \\ &+ \mathcal{L}_{dice}(C_i(E(x^s)), y^s) \quad \text{i=1, 2, 3} \\ &+ \mathcal{L}_{boundary}(C_i(E(x^s)), y^s)\end{aligned} \tag{9}$$

3.7. Other information

We conduct all experiments on workstation with Ubuntu 18.04 operating system, RTX3080ti GPU, 12G memory and Pytorch 1.8 deep learning framework. In the training phase, the initial value of the learning rate $initial_{lr}$ is set to 0.02, and is decayed according to the formula: $lr = initial_{lr} \times \gamma^{\frac{epoch}{step\_size}}$, where the initial value γ is set to 0.7, and the $step\_size$ is set to 4. We used standard Adam to optimize the objective function. The batch size is empirically set to 2, which is the maximum batch size that the proposed method can run on our GPU with 12 GB memory. Since most training procedures converged around epoch of 60, we empirically set epoch to 60 to ensure effective training.

4. Experiment

In this section, we evaluate the effectiveness of our proposed method from both qualitative and quantitative perspectives on challenging knee cartilage benchmarks (SKI10 and OAI-ZIB datasets). Specifically, we first introduce the datasets and evaluation metrics in Sections 4.1 and 4.2. We then compare PA-UNet with several semantic segmentation methods on OAI-ZIB and SKI10 datasets in Sections 4.3 and 4.4, respectively. Finally, we conduct ablation studies to analyze the effectiveness of each key component of the PA block in Section 4.5.

4.1. Dataset

The experiments were conducted on the publicly available datasets OAI-ZIB and SKI10.

The OAI-ZIB dataset is comprised of the femoral and tibial cartilages and bones of 507 MR volumes from the publicly available Osteoarthritis Initiative dataset [10]. The MR images were acquired on Siemens 3T Trio systems using a 3D double echo steady state (DESS) sequence with water excitation. Outlines of femoral and tibial cartilages and bones were generated using a statistical shape model [42] with manual adjustments performed by experts at Zuse Institute Berlin. We only segment the cartilage as our target. For 3D volumes, each of its 160 slices has a resolution of 384 × 384. Therefore, the OAI dataset can extract a total of 81120 2D image data from it. Along the slice dimension, we pick 40th, 80th and 120th slices for comparative experiments of 2D models and all slices for 3D model in this work. The reason is that there are no images of missing one or both cartilage in these slices.

Basis for the SKI10 challenge are 100 3D MR volumes acquired in a multi-site, multi-vendor and multi field-strength setting. Cases of left and right knees are distributed approximately equally [43]. Reference labels of the femoral and tibial cartilage along with respective bones were also provided. Only femoral and tibial cartilage labels are used in this work. Additionally, labels for load bearing regions of tibial and femoral cartilages were provided for evaluation. 80 MR volumes were used for training and 20 for testing. The slicing selection strategy is consistent with the OAI-ZIB dataset.

All MR images are saved from original format to Portable Network Graphics (PNG) format. In cartilage labels, we label 0 as background, 1 as femoral cartilage, and 2 as tibial cartilage. Both training image and testing images are padding to 512 × 512 due to our design.

4.2. Evaluation metrics

In our testing phase, we choose 2D and 3D models to contrast with our model, so different metrics applied in different space are selected in our experiments. For 2D space, there are three evaluation metrics in our experiment, including the dice similarity coefficient (DSC), the intersection over Union (IoU) and Hausdorff distance (HD). We compare the segmented results of knee cartilage (A) to the corresponding gold standard (B) using a slice-wise comparison per patient.

The DSC is the most used measure in medical image segmentation to compare the overlap of two regions and higher dice value indicates better segmentation performance. And IoU is also known as the Jaccard index (or the Jaccard similarity coefficient) which has been widely used to measure the similarity between finite sample sets.

Hausdorff Distance (HD) measures how far two subsets of a metric space are from each other. HD is one of the most informative and useful criteria because it is an indicator of the largest segmentation error and can be formulated as:

$$HD = max(h(|A|, |B|), h(|B|, |A|)) \tag{10}$$

$$h(|A|, |B|) = \max_{a \in |A|} \{ \min_{b \in B} \|a - b\| \} \tag{11}$$

$$h(|B|, |A|) = \max_{a \in |B|} \{ \min_{b \in A} \|b - a\| \} \tag{12}$$

For testing performance of 3D models, we choose DSC, Volume Overlap Error (VOE) and Average Symmetric Surface Distance (ASSD). Given that our model operates in 2D space, when comparing it with 3D models, we need to put the segmented slices together to corresponding volume and use them as a basis for comparison. Assuming that, A is the segmentation result of the knee cartilage, and B is the ground truth, then the definitions of the two metrics are as follows:

VOE is calculated using the ratio between intersection and union between two sets of segmentations (A and B):

$$VOE(A, B) = 1 - \frac{|A \cup B|}{|A \cap B|} \tag{13}$$

Average Symmetric Surface Distance (ASSD) is the average distance between the surfaces of segmentation results A and the ground truth B, where d(p, S(X)) represents the shortest Euler distance from voxel p to the surface voxel of the segmentation result:

$$ASSD(A, B) = \frac{1}{|S(A)| + |S(B)|} \Big(\sum_{p \in S(A)} d(p, S(B)) + \sum_{p \in S(B)} d(p, S(A)) \Big) \tag{14}$$

4.3. Comparison with the SOTA methods on OAI-ZIB dataset

Quantitative analysis of segmentation accuracy

We performed a comprehensive comparison between our model and state-of-the-art 2D, 3D medical image segmentation models. With regard to 2D methods, there are only a few models designed for knee cartilage segmentation in recent years, so we compared ours against U-Net, DeepLabV3+ and SegNet which are not designed specially for knee cartilage segmentation whereas perform well on medical image segmentation.

From Table 1, it can be seen that our proposed model showed higher performance in the comparisons on the three evaluation metrics. Especially for the DSC, the proposed method obtains the highest mean value of 88.8% on femoral cartilage and 82.7% tibial cartilage. The best performance of our model further confirmed that attention mechanism plays a vital role in medical image segmentation tasks.

For verifying whether our design in 2D space is superior to methods working in 3D space, we performed a series of comparative tests on all OAI-ZIB dataset samples. We choose U-Net++ (3D), Ensemble method [44], TransUNet (3D), Attention U-Net (3D) nnFormer [45] and CTC-Net [46] as our comparing factors. All these methods are specially designed for medical imaging segmentation which bring more fair condition. As showed in Table 2, we also notice that our model outperforms other methods in almost all evaluation metrics, but falls slightly behind U-Net++ (3D) and Ensemble methods in terms of ASSD. Compared with Attention U-Net (3D), we further confirm that fully exploiting the information buried in 2D space also can achieve high performance than methods in 3D space.

Table 1
Comparison of 2D models with Ours on OAI-ZIB dataset.

Method	DSC(↑)		IoU(↑)		HD(↓)	
	femoral	tibial	femoral	tibial	femoral	tibial
U-Net	86.7	81.3	77.8	70.8	10.1	12.6
SegNet	86.4	80.2	76.7	69.3	12.3	15.3
DeepLabV3+	87.3	81.6	77.4	71.3	10.5	13.5
Ours	**88.8**	**82.7**	**77.9**	**72.4**	**9.1**	**11.7**

Table 2
Comparison of 3D models with Ours on OAI-ZIB dataset.

Method	DSC(↑)		VOE(↓)		ASSD(↓)	
	femoral	tibial	femoral	tibial	femoral	tibial
Ensemble Method	88.13	84.64	22.42	26.52	0.193	0.215
U-Net++	88.22	84.31	21.83	25.63	0.192	0.242
TransUNet	88.66	83.86	21.36	26.01	0.238	0.245
Attention U-Net	88.90	84.99	21.77	25.14	0.247	0.252
nnFormer	87.50	83.59	22.37	25.83	0.224	0.235
CTC-Net	87.86	83.66	22.58	25.32	0.233	0.239
Ours	**90.18**	**85.74**	**20.12**	**24.25**	**0.183**	**0.201**

Visual analysis on segmentation result

Fig. 4 provides comparisons of the visual segmentation of representative sample between the proposed method and other classical models in 2D space. The columns from $(a) - (b)$ represent the source image, ground true and predictions of SegNet, U-Net, DeepLabV3+ and ours. We organize samples $(1) - (6)$ into three pairs, with each pair representing a distinct level of segmentation difficulty: moderate, easy, and hard. In the first group, the femoral cartilage occupies a relatively smaller area than the tibial cartilage, but both together occupy a moderate area. Therefore, when segmenting such targets, the problem often arises at the ends rather than the center of the cartilage. Compared to columns c, d, and e, our method performs better at the endpoints. In the second group, both femoral and tibial cartilages have relatively large areas, and the segmentation performance of different models is not significantly different. However, in the most challenging group, both femoral and tibial cartilages have very small areas, and in some cases only the tibial cartilage is present, which can greatly affect the segmentation accuracy. Compared to columns c, d, and e, our model not only detect the target but also achieve complete segmentation of the target.

4.4. Comparison with the SOTA methods on SKI10 dataset

Quantitative analysis on segmentation accuracy

To further verify the generalization ability of the proposed method, same as experiments on OAI-ZIB dataset, we also set two types of models, working in 2D and 3D space, to achieve qualitative and visual result presentation. Under 2D space, we choose DeepLabV3+, Norman et al. [9] and U-Net as our comparative models. For 3D space, we compare our method with U-Net++ (3D), transUNet (3D), Attention U-Net (3D) and Ensemble method. Table 3 shows experiment results on different models. We observe similar the same excellent performance on this dataset despite that SKI10 is not good as OAI-ZIB in image quality. In the 3D space, as showed in Table 4, we also see that our method also

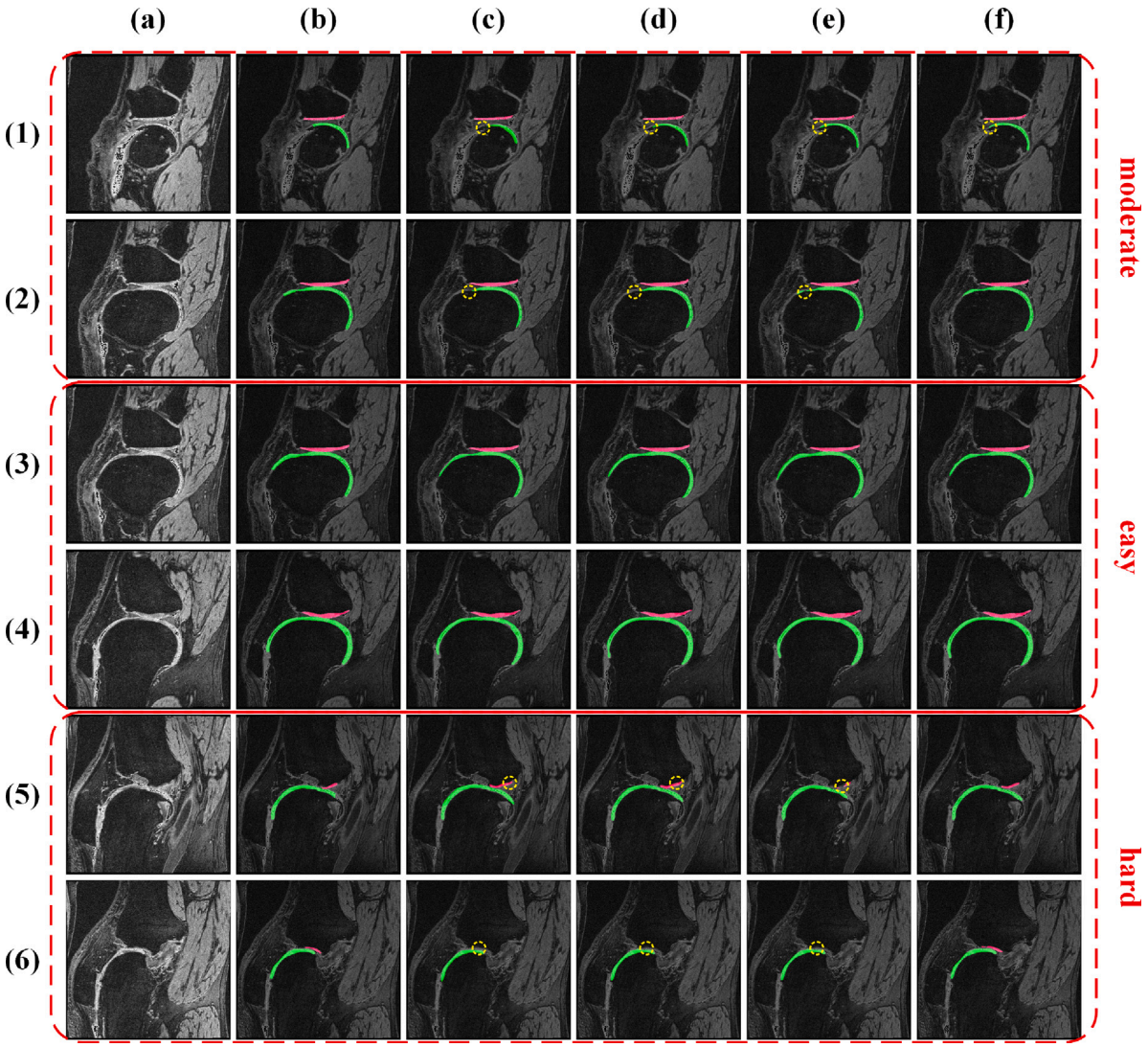


Fig. 4. The visual results of 2D models and Ours on OAI-ZIB dataset. The red part represents the predictions of femoral cartilages and the green represents the predictions of tibial cartilages. yellow dot circles represent the difference between reference model and the label.

Table 3
Comparison of 2D models with Ours on SKI10 dataset.

Method	DSC(↑)		IoU(↑)		HD(↓)	
	femoral	tibial	femoral	tibial	femoral	tibial
U-Net	77.4	71.2	67.6	64.1	17.5	20.1
DeepLabV3+	76.4	70.5	66.5	63.7	20.3	25.7
Norman et.al	77.8	70.6	67.3	64.2	18.6	20.4
Ours	**78.2**	**71.6**	**68.5**	**64.6**	**16.4**	**18.7**

Table 4
Comparison of 3D models with Ours on SKI10 dataset.

Method	DSC(↑)		VOE(↓)		ASSD(↓)	
	femoral	tibial	femoral	tibial	femoral	tibial
U-Net++	77.20	71.40	20.86	20.07	0.392	0.409
TransUNet	77.89	74.29	21.36	20.01	0.343	0.345
Attention U-Net	79.93	76.01	18.77	18.14	0.316	0.295
Ensemble method	78.89	75.67	18.47	18.19	0.333	0.315
nnFormer	78.33	75.68	20.55	19.32	0.358	0.355
CTC-Net	39.65	76.42	20.51	10.64	0.379	0.340
Ours	**81.32**	**77.65**	**18.31**	**17.25**	**0.309**	**0.287**

exhibits excellent performance across almost all evaluation metrics. It is only slightly inferior to the Attention U-Net model in the evaluation of the ASSD metric. Our model achieves best performance in both 2D and 3D space in metric dice score at 78.2%/71.6% and 80.45%/77.23% respectively which further proves the robustness of our method.

Visual analysis on segmentation result

Here, we present our visual segmentation result with other 3D models in Fig. 5 in 2D slice form. Different from OAI-ZIB dataset, the shapes of data in the Ski10 dataset are more diverse. From column (a) to column (g), each column represents the source image, ground truth, and the predictions of the Ensemble method (3D), U-Net++ (3D), TransUNet (3D), Attention U-Net (3D), and ours, respectively. From the visualization results, we observed that U-Net++(3D) had the worst experimental performance, followed by TransUNet. Both Attention U-Net and the Ensemble method show similar performance to our method, but Attention U-Net had a slight edge in terms of accuracy. However, our method still outperformed Attention U-Net in many aspects by examining samples (1) to (5). Because we notice that other models have varying degrees of under-segmentation issues, especially in the

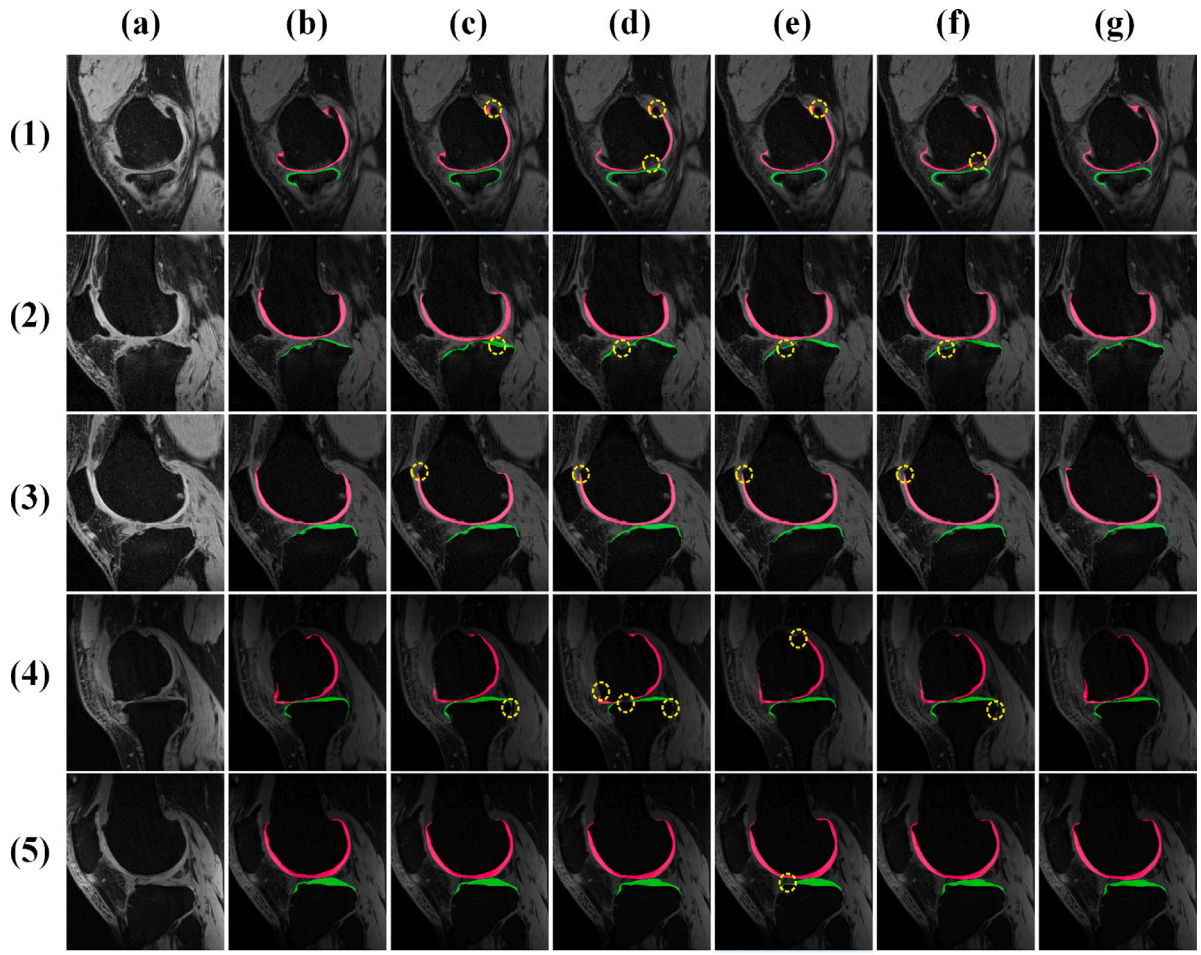


Fig. 5. The visual results of 3D models and Ours on SKI10 dataset. The red part represents the predictions of femoral cartilages and the green represents the predictions of tibial cartilages. yellow dot circles represent the difference between reference model and the label.

endpoints or thin areas of the cartilage, which do not occur in ours.

4.5. Ablation test

In this section, we design multiple experiments on OAI-ZIB dataset to investigate the effectiveness of each component of PA-UNet (including PA block and FA block). Specifically, to verify the improvement of the PA block on segmentation performance, we conduct two sets comparative experiments. The first set compares the performance of approach U-Net with approaches +U-Net, ++U-Net and +++U-Net, while the second set compares the performance of approach U-Net with SE block (in layer from 2 to 4) with approaches +U-Net, ++U-Net and +++U-Net. Then to verify the effectiveness of the FA block, we designed three sets of comparative experiments: +U-Net and U-Net+, ++U-Net and ++U-Net++, and +++U-Net and +++U-Net+++. In each set, the FA model was incrementally added after the attention model. This experimental design allows us to systematically investigate the contribution of the FA model in enhancing the feature learning capability of our network and to assess the usefulness of the information learned by the attention module. The symbol + preceding U-Net indicates the introduction of the PA block in U-Net, while symbol + that comes after U-Net indicates the introduction of the FA block in U-Net. And the number of + represents the quantity of PA or FA blocks used. For example, the meaning of "+U-Net+" is the U-Net with a FA block and a PA block in the second layer.

From Fig. 6, by comparing U-Net with +U-Net, ++U-Net and +++U-Net, we can see that the PA block brings performance improvement from 86.8%/81.6% to 88.5%/82.3%, with further improvements as the number of FA blocks increases. We also compare U-Net with three SE blocks to models +U-Net, ++U-Net and +++U-Net. We find that when only one PA block was used, our model lags behind U-Net with three SE blocks. However, when two or more FA blocks were used, our model outperform it by a large margin. This demonstrates the superiority of PA block that integrates both channel and spatial attention information. Next, to demonstrate the effectiveness of the FA block, we compare models +U-Net, ++U-Net and +++U-Net with models +U-Net+, ++U-Net++ and ours, and find that the performance is further improved by 0.4%/0.3%, 0.5%/0.5% and 0.4%/0.4% after adding the FA block. Finally, through internal comparison, we draw a conclusion that the performance improves from +U-Net to +++U-Net with the metrics DSC increasing from 86.8%/81.4% to 88.5%/82.3%, and from +U-Net+ to our model with the metrics DSC increasing from 87.2%/81.7% to 88.9%/82.7%, demonstrating that the information is indeed being effectively utilized.

5. Conclusion

In this study, we propose a new network called PA-UNet that enhances knee cartilage segmentation using an attention mechanism on different datasets. Our novel PA block addresses the challenges of small sample segmentation. Compared with other classic methods, our segmentation performance consistently outperforms them across both femoral and tibial cartilage parts. Additionally, we conduct an ablation study to show how our PA block and FA block surpass others. Finally, our method has the potential for further improvement to better generalize in the target domain, and we plan to explore this direction in our future work.

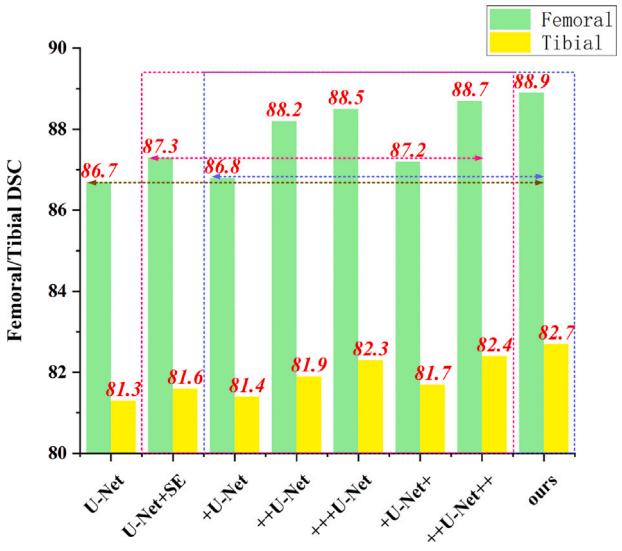


Fig. 6. The result of ablation test, green represents the accuracy of femoral cartilage segmentation and the yellow represents the accuracy of tibial cartilage segmentation. In each dashed box, bidirectional arrows represent the comparison between the base model and the optimal model.

CRediT authorship contribution statement

Xiang Wang: Methodology, Writing – original draft, Writing – review & editing. **Cao Shi:** Funding acquisition, Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This study was supported by the National Natural Science Foundation of China (grant numbers 61806107, 61702135 and 62201314).

Data availability

The authors do not have permission to share data.

References

[1] J. Folkesson, E.B. Dam, O.F. Olsen, P.C. Pettersen, C. Christiansen, Segmenting articular cartilage automatically using a voxel classification approach, IEEE Trans. Med. Imaging 26 (1) (2006) 106–115.

[2] P. Dodin, J.-P. Pelletier, J. Martel-Pelletier, F. Abram, Automatic human knee cartilage segmentation from 3-D magnetic resonance images, IEEE Trans. Biomed. Eng. 57 (11) (2010) 2699–2711.

[3] S. Lee, S.H. Park, H. Shim, I.D. Yun, S.U. Lee, Optimization of local shape and appearance probabilities for segmentation of knee cartilage in 3-D MR images, Comput. Vis. Image Underst. 115 (12) (2011) 1710–1720.

[4] C.N. Öztürk, S. Albayrak, Automatic segmentation of cartilage in high-field magnetic resonance images of the knee joint with an improved voxel-classification-driven region-growing algorithm using vicinity-correlated subsampling, Comput. Biol. Med. 72 (2016) 90–107.

[5] J. Pang, P. Li, M. Qiu, W. Chen, L. Qiao, Automatic articular cartilage segmentation based on pattern recognition from knee MRI images, J. Digit. Imaging 28 (6) (2015) 695–703.

[6] Q. Wang, D. Wu, L. Lu, M. Liu, K.L. Boyer, S.K. Zhou, Semantic context forests for learning-based knee cartilage segmentation in 3D MR images, in: International MICCAI Workshop on Medical Computer Vision, Springer, 2013, pp. 105–115.

[7] Z. Zhou, G. Zhao, R. Kijowski, F. Liu, Deep convolutional neural network for segmentation of knee joint anatomy, Magn. Reson. Med. 80 (6) (2018) 2759–2770.

[8] F. Liu, Z. Zhou, H. Jang, A. Samsonov, G. Zhao, R. Kijowski, Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging, Magn. Reson. Med. 79 (4) (2018) 2379–2391.

[9] B. Norman, V. Pedoia, S. Majumdar, Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry, Radiology 288 (1) (2018) 177–185.

[10] F. Ambellan, A. Tack, M. Ehlke, S. Zachow, Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative, Med. Image Anal. 52 (2019) 109–118.

[11] J. Schock, M. Kopaczka, B. Agthe, J. Huang, P. Kruse, D. Truhn, S. Conrad, G. Antoch, C. Kuhl, S. Nebelung, et al., A method for semantic knee bone and cartilage segmentation with deep 3D shape fitting using data from the osteoarthritis initiative, in: Shape in Medical Imaging: International Workshop, ShapeMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings, Springer, 2020, pp. 85–94.

[12] A. Raj, S. Vishwanathan, B. Ajani, K. Krishnan, H. Agarwal, Automatic knee cartilage segmentation using fully volumetric convolutional neural networks for evaluation of osteoarthritis, 2018.

[13] A. Tack, S. Zachow, Accurate automated volumetry of cartilage of the knee using convolutional neural networks: data from the osteoarthritis initiative, in: 2019 IEEE 16th International Symposium on Biomedical Imaging, ISBI 2019, IEEE, 2019, pp. 40–43.

[14] W. Burton II, C. Myers, P. Rullkoetter, Semi-supervised learning for automatic segmentation of the knee from MRI with convolutional neural networks, Comput. Methods Programs Biomed. 189 (2020) 105328.

[15] N. Le, T. Bui, V.K. Vo-Ho, K. Yamazaki, K. Luu, Narrow band active contour attention model for medical segmentation, Diagnostics 11 (8) (2021) 1393.

[16] X. Yu, Q. Yang, Y. Zhou, L.Y. Cai, R. Gao, H.H. Lee, T. Li, S. Bao, Z. Xu, T.A. Lasko, et al., UNesT: Local spatial representation learning with hierarchical transformer for efficient medical segmentation, 2022, arXiv preprint arXiv:2209.14378.

[17] T. Shan, J. Yan, SCA-Net: A spatial and channel attention network for medical image segmentation, IEEE Access 9 (2021) 160926–160937.

[18] D. Chen, F. Zhang, P. Hao, F. Wu, T. Dong, 2D medical image segmentation combining multi-scale channel attention and boundary enhancement, J. Comput.-Aided Des. Comput. Graph. (2022).

[19] A. Amer, T. Lambrou, X. Ye, MDA-unet: a multi-scale dilated attention U-net for medical image segmentation, Appl. Sci. 12 (7) (2022) 3676.

[20] C. Tan, Z. Yan, S. Zhang, K. Li, D.N. Metaxas, Collaborative multi-agent learning for MR knee articular cartilage segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22, Springer, 2019, pp. 282–290.

[21] H. Zheng, Y. Zhang, L. Yang, P. Liang, Z. Zhao, C. Wang, D.Z. Chen, A new ensemble learning framework for 3D biomedical image segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, 2019, pp. 5909–5916.

[22] L. Yu, J.Z. Cheng, Q. Dou, X. Yang, H. Chen, J. Qin, P.A. Heng, Automatic 3D cardiovascular MR segmentation with densely-connected volumetric convnets, in: Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20, Springer, 2017, pp. 287–295.

[23] H. Zheng, Y. Zhang, L. Yang, C. Wang, D.Z. Chen, An annotation sparsification strategy for 3D medical image segmentation via representative selection and self-training, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 04, 2020, pp. 6925–6932.

[24] H. Zheng, S.M. Motch Perrine, M.K. Pitirri, K. Kawasaki, C. Wang, J.T. Richtsmeier, D.Z. Chen, Cartilage segmentation in high-resolution 3d micro-ct images via uncertainty-guided self-training with very sparse annotation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23, Springer, 2020, pp. 802–812.

[25] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[26] Z. Zhao, K. Chopra, Z. Zeng, X. Li, Sea-net: Squeeze-and-excitation attention net for diabetic retinopathy grading, in: 2020 IEEE International Conference on Image Processing, ICIP, IEEE, 2020, pp. 2496–2500.

[27] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.

[28] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19, Springer, 2016, pp. 424–432.

[29] C. Guo, M. Szemenyei, Y. Hu, W. Wang, W. Zhou, Y. Yi, Channel attention residual u-net for retinal vessel segmentation, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 1185–1189.

[30] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[32] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 801–818.

[33] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[34] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 603–612.

[35] Z. Zhu, M. Xu, S. Bai, T. Huang, X. Bai, Asymmetric non-local neural networks for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 593–602.

[36] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.

[37] Y. Huang, D. Kang, W. Jia, L. Liu, X. He, Channelized axial attention–considering channel relation within spatial attention for semantic segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 1, 2022, pp. 1016–1025.

[38] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[39] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[40] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision, 3DV, IEEE, 2016, pp. 565–571.

[41] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, I.B. Ayed, Boundary loss for highly unbalanced segmentation, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2019, pp. 285–296.

[42] H. Seim, D. Kainmueller, H. Lamecker, M. Bindernagel, J. Malinowski, S. Zachow, Model-based auto-segmentation of knee bones and cartilage in MRI data, in: Proc. MICCAI Workshop Medical Image Analysis for the Clinic, 2010, pp. 215–223.

[43] T. Heimann, B.J. Morrison, M.A. Styner, M. Niethammer, S. Warfield, Segmentation of knee images: a grand challenge, in: Proc. MICCAI Workshop on Medical Image Analysis for the Clinic, Vol. 1, Beijing, China, 2010.

[44] H. Zheng, Y. Zhang, L. Yang, C. Wang, D.Z. Chen, An annotation sparsification strategy for 3D medical image segmentation via representative selection and self-training, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 04, 2020, pp. 6925–6932.

[45] H.Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, Y. Yu, nnFormer: volumetric medical image segmentation via a 3D transformer, IEEE Trans. Image Process. (2023).

[46] F. Yuan, Z. Zhang, Z. Fang, An effective CNN and transformer complementary network for medical image segmentation, Pattern Recognit. 136 (2023) 109228.