Adaptive Fusion of Deep Learning with Statistical Anatomical Knowledge for Robust Patella Segmentation from CT Images

Jiachen Zhao, Tianshu Jiang, Yi Lin, Lok-Chun Chan, Ping-Keung Chan, Chunyi Wen, and Hao Chen

***Abstract*— Knee osteoarthritis (KOA), as a leading joint disease, can be decided by examining shapes of patella to spot potential abnormal variations. To assist doctors in the diagnosis of KOA, a robust automatic patella segmentation method is highly demanded in the clinical practice. Deep learning methods, especially convolutional neural networks (CNNs) have been widely applied to medical image segmentation in recent years. Nevertheless, poor image quality and limited data still impose challenges to segmentation via CNNs. On the other hand, statistical shape models (SSMs) can generate shape priors which give anatomically reliable segmentation to varying instances. Thus, in this work, we propose an adaptive fusion framework, explicitly combining deep neural networks and anatomical knowledge from SSM for robust patella segmentation. Our adaptive fusion framework will accordingly adjust the weight of segmentation candidates in fusion based on their segmentation performance. We also propose a voxel-wise refinement strategy to make segmentation of CNNs more anatomically correct. Extensive experiments and thorough assessment have been conducted on various mainstream CNN backbones for patella segmentation in low-data regimes, which demonstrate that our framework can be flexibly attached to a CNN model, significantly improving its performance when labelled training data are limited and input image data are of poor quality.**

***Index Terms*— Medical Image Segmentation, Deep Learning, Statistical Shape Model, Patella Segmentation, Knee Osteoarthritis**

This work was supported by Hong Kong Innovation and Technology Fund (No. ITS/028/21FP), Project of RISA (Research Institute of Smart Ageing) (Project ID: P0043002) and HKUST UROP Program. (*Corresponding author: Hao Chen.*)

Jiachen Zhao and Yi Lin are with the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology, Hong Kong, China (e-mail: {jzhaobc, yi.lin}@connect.ust.hk).

Tianshu Jiang is with the Department of Biomedical Engineering, the Hong Kong Polytechnic University, Hong Kong, China (e-mail: tianshu.jiang@connect.polyu.hk).

Lok-Chun Chan and Chunyi Wen are with the Department of Biomedical Engineering, the Hong Kong Polytechnic University, Hong Kong, China, and also with the Research Institute of Smart Aging, the Hong Kong Polytechnic University, Hong Kong, China (e-mail: {lc-justin.chan, chunyi.wen}@connect.polyu.hk).

Ping-Keung Chan is with the Department of Orthopaedics and Traumatology, the University of Hong Kong, Hong Kong, China (e-mail: lewis@ortho.hku.hk).

Hao Chen is with the Department of Computer Science and Engineering and the Department of Chemical and Biological Engineering, the Hong Kong University of Science and Technology, Hong Kong, China (e-mail: jhc@cse.ust.hk).

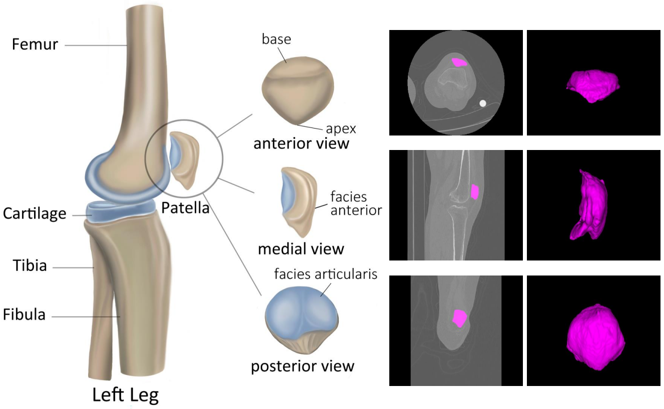


Fig. 1: Anatomy of knees. Pictures in the middle column are CT scans of knees, where regions of patella are highlighted in pink. Pictures on the right are 3D shapes of patella.

I. Introduction

KNEE osteoarthritis (KOA) affects a larger population than any other joint disease and may lead to significant annual expenditure in a country [1]. KOA occurs when degenerative changes develop in the cartilage that lines the knee joint [2]. Extensive research has been conducted to help humans better tackle OA. For instance, [3, 4] utilize neural networks to classify the severity of KOA. However, few studies have investigated how to assist the detection of KOA. Early detection is necessary because KOA is progressive and can be incurable at its later stage [5]. A crucial indicator of KOA is the degenerative change of patella [6] that is a small bone located in front of the knee joint (shown in Fig. 1). Patella protects the knee and connects the muscles in the front of the thigh to the tibia. Doctors examine the shape of patella to spot any abnormal variations that foreshadows the progression of KOA. However, conventional ways to manually observe computed tomography scans of patella can be time-consuming and demanding, which require professional anatomical knowledge and related long-term experience. An automatic segmentation method using machine is thus of great help and highly demanded in the clinical practice. Therefore, this paper focuses on proposing a robust automatic patella segmentation framework, which can help doctors more quickly capture the degenerative change of patella to diagnose KOA.

Recently, thanks to the rapid development of convolutional neural networks (CNNs), CNN-based models have become the mainstream backbones for automatic medical image segmentation. Those CNN-based models have made noticeable achievements in segmentation of many organs or tissues such as liver [7, 8, 9], lesion [10], lung [11, 12] and tumor [9, 13, 14]. However, segmentation of patellas may still impose significant challenges to mainstream CNN-based models due to ambiguous boundary and the relatively small region in the entire volume. More importantly, the amount of annotated image data is limited, which can greatly prevent CNN-based models from achieving satisfactory accuracy [15]. In addition, CNN-based models are vulnerable to poor image qualities, e.g., image artifacts and low contrast, which often appear in medical images [16].

To tackle those difficulties, researchers leverage shape priors generated through a Statistical Shape Model (SSM) to improve the segmentation of CNNs [16, 17, 18, 19]. The result of SSM can still be anatomically reliable, even when image data are limited or have poor quality. However, SSMs are sensitive to initialization [18] and need strenuous efforts to realize the location of shapes [20]. Additionally, most of past works [16, 17, 18] apply shape variation in SSM by aligning it to segmentation results of CNNs and then use that result of SSM as output, which may raise two crucial problems: (1) Lost fine details: Replacing the result of CNNs with a matched statistical shape can cause undesired decline of the final segmentation performance. Because SSMs may not accurately describe instance-wise shape variation but give general anatomy information [21]. Some detailed instance-wise information captured by CNNs may then lose its value. (2) Misleading adjustment: When CNNs fail, for example, only producing broken segmentation in the case of low-contrast input images, it may be problematic to adapt statistical shapes following results of CNN. [19] incorporates CNN into the generation of active shapes, but this method is still very vulnerable to input images of poor quality, and thus, is not robust.

To address challenges faced by CNN-based frameworks for patella segmentation and fill the gaps of utilizing SSM, this work proposes a transparent adaptive fusion framework to integrate predictions of deep learning models with anatomical knowledge represented by SSM for robust segmentation in low-data regimes. Furthermore, we propose a voxel-wise refinement strategy (VRS) between predictions of CNN and SSM when they are giving comparable segmentation results. VRS can make outputs of CNN more anatomically correct and the refined segmentation can also serve as a candidate for fusion.

Our adaptive fusion framework is motivated by a straightforward idea. For an input case, when the ground truth is available and there are multiple segmentation results (e.g., from CNN and SSM), an optimal solution is to compare their Dice-S∅rensen coefficients (DSCs) to choose the best segmentation as output. However, how can we achieve that in practice at test time when the ground truth is unknown? In order to do that, in our adaptive fusion framework, there are two modules developed. (1) We propose a module named difference score modeling (DSM) at training time. In that module, we approximate the Dice difference (ΔDSC) between segmentation results of CNN and SSM (i.e., $\text{DSC}_{\text{CNN}} - \text{DSC}_{\text{SSM}}$). We name this ΔDSC as difference score (represented by Ds) since it can reflect the performance difference of segmentations. We then conduct clustering to decide the corresponding optimal segmentation choices indicated by our estimated Ds. (2) We propose an adaptive nearest neighbor fusion (ANNF) module at test time. ANNF will decide weights of each segmentation for fusion accordingly based on its performance via referring to clusters of estimated Ds. This way of fusion enables our approach to give robust segmentation. For instance, when CNN fails in the case of low-contrast images, our adaptive fusion framework can still output a decent result by assigning more weights to segmentation with statistical anatomical knowledge. Our codes are available at https://github.com/andotalao24/PatellaSeg.

Therefore, our contributions are:

- We explicitly combine SSM-based anatomical knowledge and CNN-based segmentation via a transparent adaptive fusion framework, which can automatically adjust weights in fusion for optimal segmentation results.
- We propose a voxel-wise refinement strategy (VRS) for the dynamic integration with shape priors to make patella segmentation of CNN more anatomically correct. The VRS can confine the segmentation of CNN and serve as a candidate for fusion.
- We have established a benchmark computed tomography (CT) dataset for patella segmentation and the ground truth was provided by radiologists with agreement. To the best of our knowledge, we are the first to conduct segmentation of patella on CT scans in low-data regimes with deep learning models. Our proposal's effectiveness is demonstrated through extensive experiments and thorough assessment on diverse CNN backbones.

II. Related work

A. Medical Image Segmentation with SSM

Before the appearance of capable convolutional neural networks, SSMs are often employed [22] for the medical image segmentation of bones and organs which have regular anatomical features. In terms of patella segmentation, [20] adopts dual-optimization approach based on the active shape model (ASM) for lateral knee X-ray images. Because SSMs are obtained via geometric priors on the shapes of training data without referring to pixel information, artifacts or low image contrast may not exert influence. However, heuristically designed models of appearance are usually necessary for such methods to locate regions and adjust the SSM to the image data [23]. [24] located structures by modeling gray-level appearance and active shapes. In summary, SSMs can offer useful anatomical knowledge with a few labelled data, which enables robust segmentation in the case of limited data and unreliable image data.

B. Medical Image Segmentation with CNN

In recent years, CNNs have been the mainstream frameworks for various medical image segmentation tasks. CNN

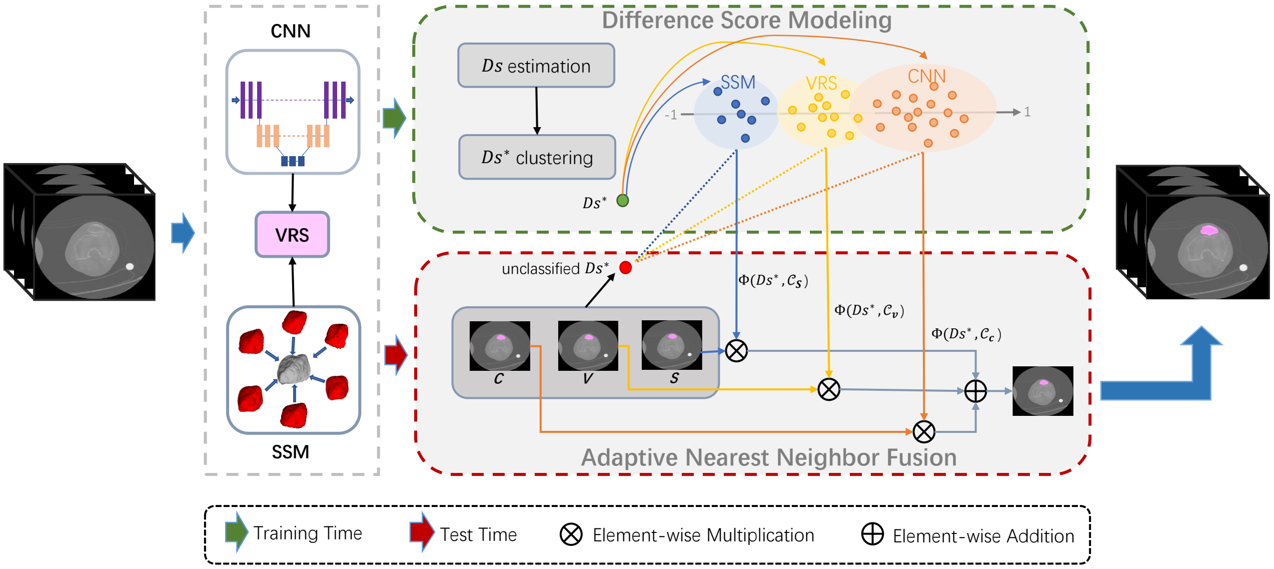


Fig. 2: An illustration of our adaptive fusion framework, which consists of three parts. Segmentations as candidates for fusion are first generated from a CNN backbone, SSM and voxel-wise refinement strategy (VRS), which are represented by C, S and V respectively. Those segmentations are then sent to Difference Score Modelling (DSM) module and Adaptive Nearest Neighbor Fusion (ANNF) module. Ds represents the difference score (i.e., $\text{DSC}_{\text{CNN}} - \text{DSC}_{\text{SSM}}$) and Ds^* stands for the estimated Ds. In ANNF, $\Phi(Ds^*, \mathcal{C}_{\text{s}}), \Phi(Ds^*, \mathcal{C}_{\text{v}})$ and $\Phi(Ds^*, \mathcal{C}_{\text{c}})$ represent memberships (i.e., possibilities) of Ds^* to be in the clusters $\mathcal{C}_{\text{s}}$, $\mathcal{C}_{\text{v}}$, and $\mathcal{C}_{\text{c}}$ formed in DSM.

based models can be broadly classified into two categories, i.e., 2D-based and 3D-based approaches. In terms of 2D-based methods, 2D convolution is conducted toward slices of input medical image data (e.g., volumetric CT or magnetic resonance imaging (MRI)) [13, 25, 26, 27]. UNet [27] especially, has been widely applied because of its reliable performance and simplicity. Based on UNet, there are many variants. UNet++ [28] reinforces the skip connections so as to aggregate features of varying semantic scales. nnUNet [29] shows that UNet [27] is capable enough for various tasks in medical image analysis through doing automatic searches on optimal hyper-parameters. In addition, ResUNet [30] is also a popular and capable model for segmentation, which replaces the encoder with a ResNet [31]. In terms of 3D-based methods, 3D convolution is conducted toward a volume, which can better utilize information across slices [7, 32, 33, 34, 35, 36]. VNet [32], for instance, incorporates residual connection into 3D UNet and is trained towards a Dice coefficient. There is also work combining both 2D and 3D methods. For example, H-DenseUNet [9] leverages both a 2D DenseUNet and a 3D counterpart for extracting intra-slice features and aggregating volumetric contexts. In terms of knee bones segmentation, [37] employed three 2D CNNs with axial, coronal and sagittal image planes as input respectively for tibial cartilage segmentation. [38] combined SegNet [26] with 3D simplex deformable modeling for the cartilage and bone segmentation of the knee joints.

C. Segmentation Combining SSM and CNN

Rather than through fusion, past works combine SSM and CNN mainly through initializing SSM with segmentation of CNN and adjusting SSM based on some schema. In other words, the final output will be given solely by SSM rather than both CNN and SSM. [16, 17] designed a complex system that consists of 2D UNet, 3D UNet and SSM. The SSM which is adapted to the segmentation of 2D UNet will serve as the input to a 3D UNet. The result of the 3D UNet may further be used as a template for the SSM to fit. This fashion of combination may neglect some instance-wise details segmented by UNet. In addition, when UNet fails to produce correct segmentation, adapting SSM to that segmentation will not be reasonable, possibly damaging the ultimate segmentation result. [18] proposed a way to utilize statistical pixel-level information by Bayesian Model to assist the adjustment of the SSM. [19] employed a CNN to adjust the shape of SSM based on input images. Those adjustment schemata utilizing either pixel-level features or CNNs may be vulnerable to poor image quality such as artifact and low-contrast. So those methods may not give robust segmentation in practice when image data are unreliable.

III. Method

In this section, we develop an adaptive fusion framework whose work flow is shown in Fig. 2. There are three stages in our framework. At the first stage, we will respectively get segmentations from a CNN backbone and SSM and obtain the result of voxel-wise refinement strategy (VRS). Those segmentation results serve as candidates for fusion. In Sec. III-A, we first introduce the background of SSM and how we generate it in a label-free way. Then in Sec. III-B, we explain the working mechanism of our proposed VRS. At the second stage, the difference score modelling (DSM) will estimate the difference score (Ds) and conduct clustering on the approximated Ds (represented as Ds^*). Then we are able to predict the best segmentation based on Ds^* without ground truth. The DSM is explained in Sec. III-C. At the third stage, for each input test case, our proposed adaptive nearest neighbor fusion (ANNF) module will decide the possibility of each segmentation result to be the best via a weighted k-nearest neighbor algorithm based on clusters of Ds^*. Those possibilities are deemed as weights in fusion to produce the ultimate segmentation result. The ANNF is explained in Sec. III-D.

A. Revisiting SSM

SSM gives the average and general shapes and has a number of parameters for controlling the variation of shapes [39]. For the construction of an SSM, there are generally two stages, which are shape alignment and dimension reduction. The generalized Procrustes alignment (GPA) is a popular method for shape alignment, as described by [40]. The procedure to minimize the distance between two shapes is implemented iteratively to align a group of shapes to their unknown mean. After alignment, we can simply average over all samples to obtain the mean shape as $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, where $\bar{x}$ is the mean shape, x_i indicates the i^{th} shape and N is the total number of all shapes. Instead of using variance on all points of a shape, we can reduce the dimension to a small set of modes for the description of variation since there may be inter-point correlation. This process is usually accomplished by using principal component analysis (PCA) [41]. Finally, the segmented shape in the training set can be represented by using the mean shape and a weighted sum of deviations: $x = \bar{x} + P\theta$, where $P = (p_1 p_2 ... p_t)$ is the matrix of the first t eigenvectors, and $\theta = (\theta_1 \theta_2 ... \theta_t)^T$ is a vector of weights. Weights can be learnt through regression models to adapt the SSM to certain image data. Further details for calculation and construction can be found in [23].

The conventional procedures of generating an SSM include arduous landmarks labeling. Landmarks are characteristic points that are distinguishable and represent features of shapes. In our work, to avoid laborious and challenging annotation of landmarks, we use all points sampled from segmented training data by an edge detector [42]. Additionally, although SSM can represent average shapes and variation as well [43], we found that when combining SSM with neural models, applying further variations [16] in SSM leads to degraded performance compared with using average shapes alone (see our experiments in Sec. IV-D). That may be attributed to lost fine details and misleading adjustment when applying variations by some schema. Thus, in this work, segmentation of SSM refers to results by employing average shapes if without specification.

B. Voxel-wise Refinement Strategy

VRS works under the observation that CNN is good at capturing instance-wise shape variation, while it lacks anatomical knowledge [15]. So, we utilize a weighted average to integrate shape priors into segmentation of CNN when CNN and SSM are giving comparable segmentation results. Let C, S and V stand for the segmentation results of CNN, SSM and VRS respectively. The VRS is represented as follows:

$$V = \mu \otimes C + (1 - \mu) \otimes S, \tag{1}$$

where μ is a weight matrix and $\otimes$ stands for element-wise multiplication. μ is decided by the following equation:

$$\mu = \begin{cases} 1 - \Delta \otimes \frac{1}{r\sigma} & \Delta < r\sigma \\ \quad 0.5 & \Delta \geqslant r\sigma \end{cases}, \tag{2}$$

where σ is the standard deviation of shapes to the mean that can be derived along with the generation of SSM at training time; Δ is the absolute difference value between C and S; r is a scalar parameter (see Sec. IV-C for its value decision during implementation). Weight μ is adjustable following a schema that when the deviation to the mean shape at a position is large, more weights will be given to the C because shapes are more variable at that position; when Δ is too large, more weights will be assigned to S in order to constrain the shape. To complete the definition of Eq. (2), we need to reconsider μ in the boundary case where C and S are not comparable ($\Delta \geqslant r \times \sigma$). We conduct a simple arithmetic mean between S and C. This naive setting is empirically found effective and reduces the number of hyper-parameters in our framework. Overall, the VRS can confine the segmentation of CNN and remove outliers, which complements results of CNN with shape priors.

C. Difference Score Modeling

In this section, we propose to estimate the difference score Ds (i.e., $\text{DSC}_{\text{CNN}} - \text{DSC}_{\text{SSM}}$) and implement clustering on the approximated results Ds^*. However, instead of precise quantitative approximation, the goal of DSM is to give qualitative prediction by forming general range-wise mappings from Ds^* to true Ds. More specifically, for instance, when Ds^* is *small*, it is expected to indicate a fairly *small* Ds (e.g., a negative value). Thus we can know SSM outperforms CNN and in this case, SSM should be the optimal segmentation choice.

We start our derivation of Ds^* with the original formula of the Ds:

$$Ds = \frac{2\,|C \cap M|}{|C| + |M|} - \frac{2\,|S \cap M|}{|S| + |M|}, \tag{3}$$

where M represents the ground truth; S and C are segmentation results of SSM and CNN; $|\cdot|$ denotes a volume; $|C \cap M|$ and $|S \cap M|$ stand for the intersection of S and C with M. In order to focus on the difference between segmentation results, we then split S and C respectively into two parts:

$$C = I \cup C_e, \; S = I \cup S_e, \tag{4}$$

where I is the common area of S and C (i.e., $|C \cap S|$); S_e and C_e are portions of S and C that exclude the intersection I (i.e., $S_e = S \setminus I$, $C_e = C \setminus I$). S_e and C_e actually showcase the disagreement of SSM and CNN on the segmentation of the same instance.

We next accordingly define γ, α and β as the percentage of I, C_e and S_e to be correctly segmented (i.e., lying inside M).

$\gamma = \frac{|I \cap M|}{|I|}$, $\alpha = \frac{|C_e \cap M|}{|C_e|}$, $\beta = \frac{|S_e \cap M|}{|S_e|}$. We can then have the following representations:

$$|C \cap M| = \gamma\,|I| + \alpha\,|C_e|\,, \tag{5}$$

$$|S \cap M| = \gamma\,|I| + \beta\,|S_e|\,. \tag{6}$$

Then replace $|C \cap M|$ and $|S \cap M|$ in Eq. 3 with Eq. 5 and Eq. 6. We can obtain:

$$Ds^* = 2\left(\frac{\gamma\,|I| + \alpha\,|\mathrm{C}_e|}{|C| + |M|} - \frac{\gamma\,|I| + \beta\,|S_e|}{|S| + |M|}\right). \tag{7}$$

Because I is agreed on by both C and S, we estimate γ as 1, indicating that I has high confidence to be fully correct. We then estimate the $|M|$ as the average of $|S|$ and $|C|$ because S and C are expected to be generally close to M in terms of the size alone. Experiments in Sec. IV-F support the effectiveness of our approximation. Therefore, we come to the following approximation.

$$\gamma \approx 1\,, \tag{8}$$

$$|M| \approx \frac{|S| + |C|}{2}. \tag{9}$$

After plugging Eq. 8 and Eq. 9 into Eq. 7, we can finally obtain our Ds^*:

$$Ds^* = 2\left(\frac{|I| + \alpha\,|C_e|}{|C| + \frac{|S|+|C|}{2}} - \frac{|I| + \beta\,|S_e|}{|S| + \frac{|S|+|C|}{2}}\right). \tag{10}$$

As a result, only α and β are left undecided in Eq. 10. We conjecture that there exist monotonic relations between α and C_e, β and S_e. The analysis is as follows: Let m be the ground truth for C_e and S_e, i.e., the area that C and S disagree. We speculate the disagreement is mainly because m is a changeable area specific to instances (e.g., contour) that tends to be captured by CNN rather than SSM [15]. Therefore, when C_e and S_e are increasing, meaning that the m is more diverse from the shape priors given by SSM, then α may increase and β may decrease.

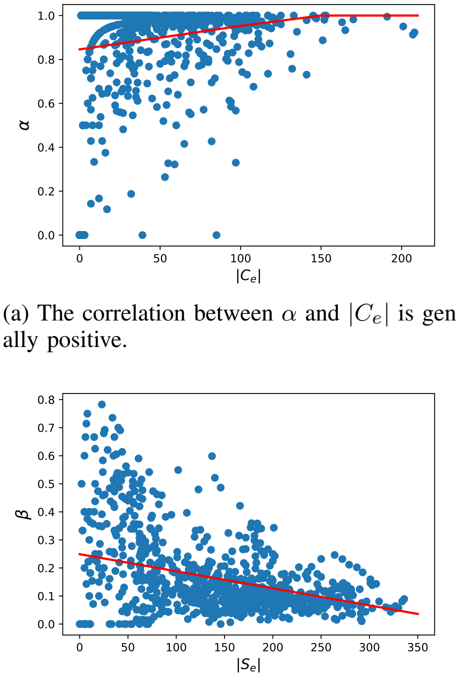
We thus implement linear regression to learn speculated monotonic relations between α and $|C_e|$, β and $|S_e|$. The expression for our regression is as follows.

$$\alpha = \frac{w_\alpha}{\lambda}\,|C_e| + b_\alpha, \tag{11}$$

$$\beta = \frac{w_\beta}{\lambda}\,|S_e| + b_\beta. \tag{12}$$

Here, w_α and w_β are scalars rather than matrices. λ is a constant. We divide the w_α and w_β by λ so as to accelerate and stablize the learning process, because $|S_e|$ or $|C_e|$ is usually much larger than α or β. We may need to clip the predicted ratios into the range between 0 and 1. We visualize the relation between α and $|C_e|$, β and $|S_e|$ and our regression result as well in Fig. 3a and Fig. 3b. The ratio α tends to be larger with $|C_e|$ becoming bigger; the relation between $|S_e|$ and β is the opposite. Such monotonic correlations align with our conjecture above. Linear regression can efficiently describe such major relations. In Sec. IV-F, we further empirically show linear regression is simple but effective enough for our approach.

Then, we conduct Ds^* clustering at training time in order to decide what the optimal segmentation is indicated by certain range of Ds^*. Clustering is implemented on a separate set where data are unseen by segmentation models so that their true performance can be reflected by DSC. For each input instance, we compare the DSCs of SSM, CNN and VRS to decide which outperforms. Then, the unclassified Ds^* will be assigned to the corresponding cluster for that best segmentation. The step-by-step implementation is shown in Algorithm 1.



(a) The correlation between α and $|C_e|$ is generally positive.

(b) The correlation between β and $|S_e|$ is generally negative.

Fig. 3: The blue dots show the correlation between α and $|C_e|$, β and $|S_e|$. The red line is the result of our regression. The majority of dots cluster around the regression line.

Algorithm 1 Ds^* Clustering

```
Require: The segmentation $C$ of CNN; The segmentation $S$ of SSM; The segmentation $V$ of VRS; The cluster $\mathcal{C}_c$ for CNN; The cluster $\mathcal{C}_s$ for SSM; The cluster $\mathcal{C}_v$ for VRS; The ground truth $M$.
Ensure: The $Ds^*$ is assigned to one of $\mathcal{C}_c$,$\mathcal{C}_s$ and $\mathcal{C}_v$.
  Calculate the $Ds^*$ based on Eq. 10.
  Let $f$ be the function to calculate DSC.
  Define $\mathcal{X}$ as the segmentation with the highest DSC.
  Calculate Max$(f(C, M), f(S, M), f(V, M))$,
    to decide $\mathcal{X}$.
  Add the $Ds^*$ to $\mathcal{C}_\mathcal{X}$, $\mathcal{C}_\mathcal{X} \in \{\mathcal{C}_c, \mathcal{C}_s, \mathcal{C}_v\}$.
```

D. Adaptive Nearest Neighbour Fusion

Based on clusters formed in DSM, we can predict the optimal segmentation choice indicated by Ds^* computed at test time. We adopt a weighted k-nearest neighbor algorithm rather than a deterministic one, which outputs the possibility of each segmentation result to be the best. Such a weighted algorithm will take every segmentation choice into consideration, thus reducing the influence of errors introduced by our estimation of Ds. The predicted possibilities are deemed as weights for fusion. The adaptiveness of our ANNF is that during fusion, among all segmentation results, the better one segmentation is, the larger weight it will have.

For the implementation of our algorithm, we are inspired by the concept of membership from "fuzzy set" [44]. In fuzzy set theory, each element in a set has a grade of membership indicating its possibility of belonging to that set. In our case, membership can be considered as the possibility of each segmentation result to be the best. We follow steps in [45] to decide the memberships of unclassified Ds^* at test time. First, we need to find the set $\mathbb{S} = \{Ds_i^* \mid i = 1, 2, ..., k\}$ that contains the k nearest Ds^* in the clusters to the current unclassified Ds^*. Then, let $\Phi(Ds^*, \mathcal{C}_{\mathcal{X}})$ be the membership of Ds^* to be in the cluster $\mathcal{C}_{\mathcal{X}}$; Let $\Phi(Ds_i^*, \mathcal{C}_{\mathcal{X}})$ be the membership of elements of $\mathbb{S}$ in the $\mathcal{C}_{\mathcal{X}}$, where $\mathcal{X} = C, S, V$. The elements of $\mathbb{S}$ are assigned memberships in a way that each Ds_i^* has complete membership (i.e., equals to 1) for the cluster which it belongs to and zero membership for the other clusters. Following [45], we define $\Phi(Ds^*, \mathcal{C}_{\mathcal{X}})$ as follows:

$$\Phi(Ds^*, \mathcal{C}_{\mathcal{X}}) = \frac{\sum_{i=1}^{k} \Phi(Ds_i^*, \mathcal{C}_{\mathcal{X}})/d + \epsilon}{\sum_{i=1}^{k} 1/d + \epsilon}, \ d = |Ds^* - Ds_i^*|^2, \tag{13}$$

where ϵ is a very small constant for smoothing and d is the distance. The assigned membership of Ds^* is affected by the inverse distance to the nearest neighbors and their memberships. Therefore, the membership for $\mathcal{C}_{\mathcal{X}}$ will be larger if Ds^* is closer to some Ds_i^* in $\mathcal{C}_{\mathcal{X}}$. Finally, the result after fusion will then be:

$$\text{Output} = \sum^{\mathcal{X}} \Phi(Ds^*, \mathcal{C}_{\mathcal{X}}) \otimes \mathcal{X}, \ \mathcal{X} = C, S, V. \tag{14}$$

Memberships $\Phi(Ds^*, \mathcal{C}_{\text{c}}), \Phi(Ds^*, \mathcal{C}_{\text{s}})$ and $\Phi(Ds^*, \mathcal{C}_{\text{v}})$ are namely the weights for segmentations C, S and V in the fusion. Therefore, we can give an output that combines all segmentation results, where the optimal segmentation contributes the most.

IV. Experiments

A. Dataset

We collected CT series of knee bones from 85 patients from January 2019 to October 2020 at Queen Mary Hospital in Hong Kong. Ethical approval for this data collection was obtained from the Hospital Authority of Hong Kong (approval number UW 22-090). To ensure the utmost privacy protection, all patients' personal information has been meticulously removed from the dataset. CT images were reconstructed with slice thicknesses of 0.625 mm and spacing of 0.625 mm. The patella area of each CT slice was annotated by two professional radiologists (through consensus) who have five-year-long clinical experience. The manual segmentation is conducted by using ITK-SNAP [1]. In this work, we randomly split the dataset into 50 CT scans for training, 15 CT scans for validation and 20 CT scans for test. Note that no patient will be in the different sets for a fair comparison.

B. Evaluation Metrics

Five metrics are used to measure the accuracy of segmentation results. For volumetric measures, we calculate volumetric overlap error (VOE), relative volume difference (RVD) and DSC. For surface distance measures, average symmetric surface distance (ASD) and maximum surface distance (MSD) are employed. Surface distance measures can better represent small features, such as osteophytes that are crucial to diagnostic purposes. For VOE, RVD, ASD and MSD, the smaller the value is, the better the segmentation result is. The value of DSC refers to Dice per case score. The detailed formulas for those metrics can be found in [46].

C. Implementation Details

We implement our models using Keras with TensorFlow 2.1.0 as backend. For CNN backbones, we use the ADAM optimizer [47] with a learning rate 10^{-4}. The batch size is 32 and the training epoch is 60 based on validation results. All experiments were conducted using an Intel CPU, and a NVIDIA RTX2080s GPU. DSM is implemented on the validation set (869 CT slices) since DSM requires segmentation on unseen data. For weighted k-nearest neighbor algorithm in ANNF, we set k as three, which we found gives a balance between computation time and performance of fusion.

For parameters in Eq. 2 of VRS, we set r as the maximum among the averages of maximums of Δ/σ of training data at each position. As a result, $r \times \sigma$ can cover the majority of the displacement of CNN from SSM at each position and outliers can thus be detected and filtered. For parameters in linear regression (i.e., Eq. 11 and Eq. 12) of DSM, λ is set as the average of two standard deviations from the mean size of $|S_e|$ and two standard deviations from the mean size of $|C_e|$. As a result, $|C_e|/\lambda$ and $|S_e|/\lambda$ can be scaled within 0 and 1, becoming comparable to α and β. The ϵ in Eq. 13 of ANNF is set as 0.001.

D. Results and Analysis

Since we are the first to conduct patella segmentation with CT images, we first test four widely adopted CNN backbones in medical image analysis, namely UNet [27], ResUNet [30], DenseUNet [9] and VNet [32] to set up baselines. The UNet follows the structure in [16] that is constructed for knee bones segmentation and gives a leading performance on the public dataset of the MICCAI grand challenge for knee image segmentation. For ResUNet, we use ResNet-34 as the encoder. The specific structure is illustrated in [30]. For DenseUNet, we adopt the structure in [9], which demonstrates a competitive performance on three organ segmentation tasks in Liver Tumor Segmentation (LiTS) Challenge. We adopt VNet from [32] that was proven capable on a public challenge dataset for prostate

[1] `http://www.itksnap.org`; v. 3.8.0; open-source software

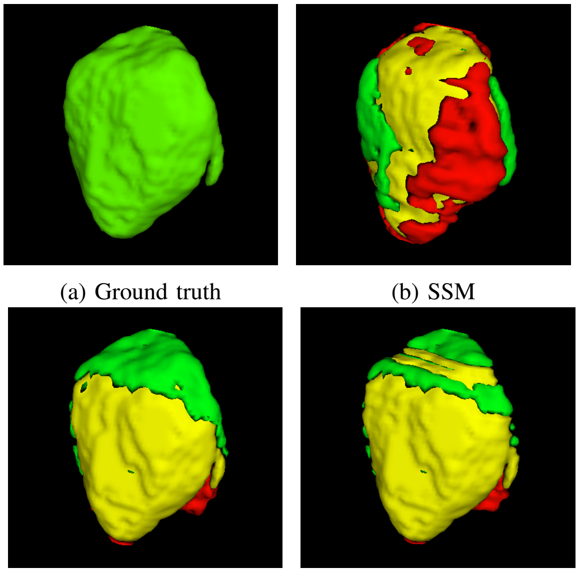
Method	DSC (%)	VOE (%)	RVD (%)	ASD (mm)	MSD (mm)
UNet	84.16 ± 0.30	27.34 ± 10.76	2.71 ± 7.09	2.57 ± 5.05	4.92 ± 6.41
+ Adjusting SSM [16]	80.37 ± 0.33	32.81 ± 10.96	4.93 ± 7.13	4.22 ± 4.21	7.32 ± 7.28
+ NN-Fusion	84.32 ± 0.26	26.78 ± 11.05	2.69 ± 7.34	2.23 ± 4.49	4.34 ± 6.88
+ Our Approach w/o VRS	84.73 ± 0.38	26.49 ± 9.84	2.20 ± 7.41	1.94 ± 4.82	4.16 ± 7.55
+ Our Approach w VRS	**85.38 ± 0.43**	**25.79 ± 9.32**	**1.93 ± 7.23**	**1.78 ± 4.29**	**3.89 ± 7.21**
ResUNet	84.32 ± 0.13	27.11 ± 10.64	2.61 ± 7.02	2.50 ± 5.97	4.46 ± 6.92
+ Adjusting SSM [16]	80.74 ± 0.28	32.29 ± 10.46	4.44 ± 8.58	4.05 ± 5.37	7.24 ± 6.10
+ NN-Fusion	84.55 ± 0.31	26.25 ± 11.35	2.47 ± 7.55	2.33 ± 5.10	4.34 ± 7.21
+ Our Approach w/o VRS	85.07 ± 0.41	25.98 ± 10.61	2.10 ± 7.36	1.74 ± 6.16	3.96 ± 6.14
+ Our Approach w VRS	**85.53 ± 0.34**	**25.18 ± 11.89**	**1.64 ± 8.13**	**1.70 ± 4.04**	**3.58 ± 7.38**
DenseUNet	84.38 ± 0.28	26.89 ± 11.05	2.57 ± 7.53	2.12 ± 4.27	4.32 ± 6.13
+ Adjusting SSM [16]	80.70 ± 0.26	32.59 ± 11.43	4.57 ± 8.95	4.13 ± 4.37	7.29 ± 7.12
+ NN-Fusion	84.54 ± 0.33	26.70 ± 12.15	2.44 ± 7.83	2.01 ± 4.99	4.28 ± 7.37
+ Our Approach w/o VRS	84.98 ± 0.23	26.13 ± 11.27	2.16 ± 7.26	1.86 ± 3.84	4.10 ± 6.95
+ Our Approach w VRS	**85.50 ± 0.28**	**25.47 ± 10.60**	**1.84 ± 7.91**	**1.75 ± 3.32**	**3.69 ± 6.73**
VNet	83.79 ± 0.26	27.02 ± 10.11	2.73 ± 9.68	2.69 ± 5.61	5.12 ± 6.05
+ Adjusting SSM [16]	79.81 ± 0.31	34.35 ± 11.73	5.11 ± 9.28	4.32 ± 5.44	8.23 ± 6.45
+ NN-Fusion	84.01 ± 0.21	26.89 ± 10.64	2.56 ± 7.74	2.47 ± 5.26	4.79 ± 6.69
+ Our Approach w/o VRS	84.70 ± 0.33	26.72 ± 10.53	2.22 ± 7.95	2.09 ± 4.35	4.21 ± 6.10
+ Our Approach w VRS	**85.34 ± 0.37**	**26.13 ± 10.67**	**2.10 ± 7.51**	**1.98 ± 4.54**	**3.91 ± 6.76**

TABLE I: Evaluation results of different post-processing methods on four mainstream CNN baselines for segmentation.

segmentation [48]. Except VNet, all the models are 2D-based. The final results are shown in Table I. Four CNN baselines show similar segmentation performances on patella CT scans according to the diverse metrics. The DSCs are approximately 84%.

Then we apply our adaptive fusion framework to all four CNN backbones to evaluate its effectiveness on improving their segmentation results. We implement our approach with and without the VRS as a candidate for fusion to examine the effect of VRS. The results of Table I demonstrate that our fusion framework can bring noticeable improvement on all those CNN baselines. DSCs of all four CNN backbones can be increased by around 0.6% when VRS is not employed. When VRS is applied, the performance is further boosted. In terms of DSC, the average increase of all CNN backbones can then achieve around 1.2%. Additionally, we observed a more dramatic increment in the performance of CNN baselines when the data size is further decreased (see Sec. IV-E).

As comparison, we experiment with the method of [16] which also proposes to integrate CNN with SSM. That method outputs an SSM which is adjusted to fit the segmentation results of CNN. According to results in Table I, adjusting SSM turns out to decrease segmentation performance in our case. This may be attributed to the shortage of SSM that it is not good at describing instance-wise variation but general shape information. Outputting an adjusted SSM as segmentation result may neglect some details captured by CNN. Besides, adapting SSM to the segmentation of CNN is based on the assumption that CNN gives a reliable output which may not be true, especially in the case of noisy or unclear input images. However, because our adaptive fusion framework can automatically adjust the contribution of CNN and SSM to fusion based on their respective segmentation performance, our approach is less vulnerable to that concern on the quality of input images.


(a) Ground truth

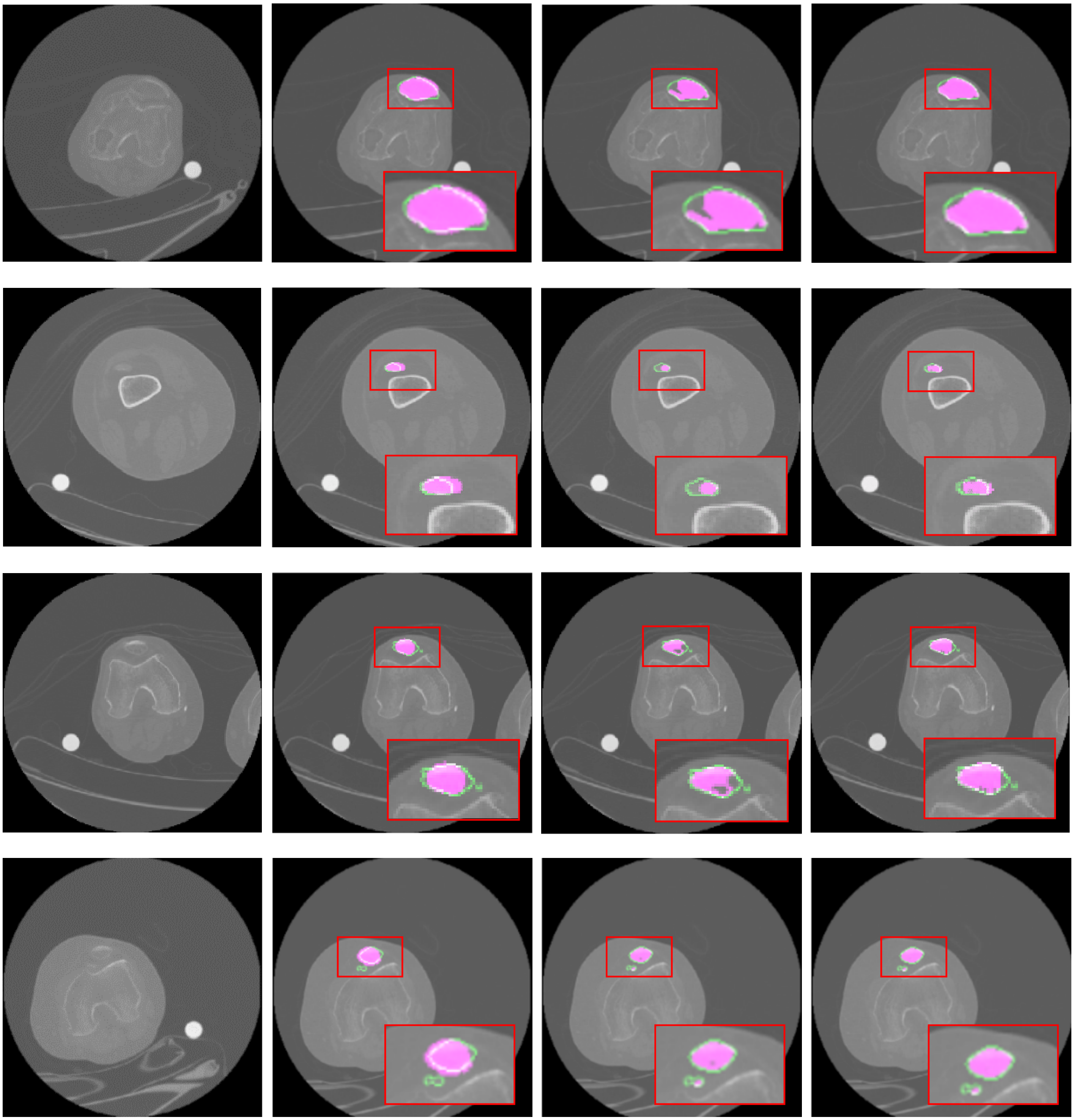
(b) SSM

(c) CNN

(d) Ours

Fig. 4: Segmentation results comparison in 3D view. Green parts stand for the ground truth. Red parts represent the segmentation given by different approaches, whose intersection with the ground truth is shown as yellow portions. The more yellow portions there are in one segmentation, the better that segmentation is.

In addition, to compare our framework with an implicit fusion approach, we construct a neural network for fusion (NN-Fusion). In terms of the structure of NN-Fusion, we input the concatenation of segmentation results of CNN and SSM into a series of convolutional layers. The output layer consists of a $7 \times 7 \times 1$ convolutional filter followed by a Sigmoid function. Each hidden layer is made up of a $7 \times 7 \times 4$ convolutional filter followed by ReLU. NN-Fusion is trained with the same data where DSM is implemented. We experimented with different number of layers in NN-Fusion. We select the NN-Fusion with two layers as comparison since it turns out to give the best performance. Results in Table I demonstrate that NN-Fusion can give slight improvement on four CNN backbones. For instance, the average DSC increase on UNet is 0.16%

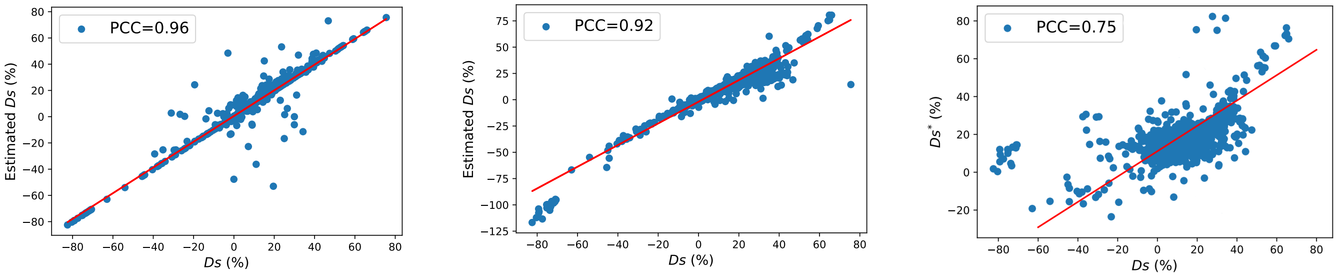


(a) Input image (b) SSM (c) CNN (d) Ours

Fig. 5: We display a series of segmentation results of CNN, SSM and our approach for comparison. Green contours represent ground-truth, and pink areas represent segmentation results. Results of CNN are given by UNet.

Methods	25 training cases		50 training cases	
	DSC (%)	Increment (%)	DSC (%)	Increment (%)
UNet	81.53 ± 0.24	-	84.16 ± 0.30	-
+Adjusting SSM [16]	78.86 ± 0.27	-2.67 ± 0.38	80.37 ± 0.33	-3.79 ± 0.47
+NN-Fusion	82.04 ± 0.17	0.51 ± 0.20	84.32 ± 0.33	0.16 ± 0.31
+Our Approach w/o VRS	82.52 ± 0.20	0.99 ± 0.10	84.73 ± 0.38	0.57 ± 0.26
+Our Approach w VRS	**83.33 ± 0.38**	**1.76 ± 0.28**	**85.38 ± 0.43**	**1.23 ± 0.22**
ResUNet	81.40 ± 0.14	-	84.32 ± 0.13	-
+Adjusting SSM [16]	78.88 ± 0.37	-2.52 ± 0.44	80.74 ± 0.28	-3.57 ± 0.28
+NN-Fusion	81.88 ± 0.13	0.48 ± 0.19	84.55 ± 0.31	0.23 ± 0.29
+Our Approach w/o VRS	82.48 ± 0.24	1.08 ± 0.14	85.07 ± 0.41	0.74 ± 0.35
+Our Approach w VRS	**83.05 ± 0.42**	**1.64 ± 0.35**	**85.53 ± 0.34**	**1.21 ± 0.32**

TABLE II: Evaluation of our approach on CNN frameworks with different data sizes.



(a) Comparison between Ds and estimated Ds with γ as 1 in Eq. 7.

(b) Comparison between Ds and estimated Ds with $|M|$ as $\frac{|S|+|C|}{2}$ in Eq. 7.

(c) Comparison between Ds and Ds^* after all steps of approximation.

Fig. 6: We conduct experiments to examine the effects of our proposed method for approximating Ds (i.e. $\text{DSC}_{CNN} - \text{DSC}_{SSM}$). The Pearson correlation coefficient (PCC) is calculated to reflect the linear relatedness of our estimation. In Fig. 6a and Fig. 6b, since only partial approximation is conducted, the label of y axis is named as “estimated Ds” to distinguish from Ds^*.

which is much lower than the 1.23% increment given by our approach. We suspect that the inferior performance of NN-Fusion is caused by the imbalanced distribution of cases where the fusion between CNN and SSM is needed. The majority of cases may be that CNN significantly outperforms SSM, leading to the unnecessity of fusion. Therefore, NN-Fusion may be trained to become biased toward the segmentation of CNN, neglecting fusion with SSM. However, our framework which conducts fusion explicitly in a statistical way may be less vulnerable to that concern.

We display a group of 3D models to visualize segmentation results in Fig. 4. The visualization indicates that our approach can further complete and refine the segmentation of CNN models by integrating with the results of SSM. Our approach can then produce a more trustworthy segmentation. For further illustration, we also display 2D input images and segmentation results in Fig. 5: (1) The first three rows show situations where the CNN suffers from severe low contrast of the input image and give broken segmentation. However, in those cases, the SSM still gives relatively plausible segmentation with some areas outside the true boundary. Our fusion framework then removes wrong portions of segmentation of the SSM and fills holes inside the segmentation of CNN, producing a more complete and reliable result. (2) On the other hand, the last row shows a case where the SSM fails. Because some KOA patients may develop osteophytes on the patella, there can be two separate regions of interest in one slice. Such occasional unexpected shape variation is difficult for SSM to predict, which has been trained mainly on connected shapes. But CNN succeeds to identify both regions of interest. In this case, our fusion framework chooses to output a segmentation that is almost the same as that of CNN. In summary, our adaptive fusion framework can automatically adjust weights of results of CNN and SSM for fusion based on their segmentation performances.

E. Effect of Training Data Size

In this section, we further decrease the total training data by half for CNN models and SSM generation to investigate the effect of our fusion framework in the case of more limited data. UNet and ResUNet are chosen as CNN backbones. As is shown in Table II, when CNN models are trained with 25 cases, we obtain more dramatic performance improvement. The DSCs of UNet and ResUNet are respectively increased by 1.76% and 1.64% on average when VRS is applied. In addition, it is worthwhile to notice that the performance of CNN backbones trained with 25 cases can be raised to a level comparable to that of CNN backbones trained with double of the data. This implies that our approach can be especially useful to increasing the performance of CNNs when there are limited annotated data.

F. Effect of Approximation in DSM

To verify the influence of our estimation in the Eq. 8 and Eq. 9, we replace one value with our estimation at a time in Eq. 7 as comparison with the true Ds. We also calculate the Pearson correlation coefficient (PCC) to measure the linear correlation of our estimation with the ground truth. The results are shown in Fig. 6a and Fig. 6b. It is noticeable that after replacing γ with 1 in Eq. 7, the relation between our estimation and ground truth is quite linear and the PCC can reach 0.96, meaning that there is little influence of such estimation. When replacing $|M|$ with $\frac{|S|+|C|}{2}$, the estimation starts to diverge when Ds becomes large with the increased difference between segmentation results of CNN and SSM. But the overall relation between estimation and ground truth is still close to linearity and the PCC can achieve 0.92.

After conducting regression to get values of α and β in Eq. 10, we evaluate our Ds^* by comparing it with Ds. As is shown in Fig. 6c, our Ds^* and Ds are highly correlated with the PCC as 0.75. More specifically, our Ds^* is getting larger with Ds increasing. A range of our Ds^* can thus be mapped to a certain range of Ds following a monotonic relation, which suffices to predict the magnitude of actual performance difference between CNN and SSM.

V. Discussion

Automatic patella segmentation plays an important role in clinical diagnosis. It provides the precise contour of patella, which can assist the detection of early symptoms of osteoarthritis for doctors. In this paper, we present an adaptive fusion framework combining convolutional neural networks and statistical shape models for robust segmentation. The adaptive fusion framework can make instance-wise decisions on the contribution of CNN and SSM to the output of fusion. This is crucial in the clinical practice where CT scans may have fairly low contrast. In such cases where a CNN-based model may fail, our proposed algorithm will assign more weights to SSM in fusion, outputting a decent segmentation result. Moreover, when CNN and SSM are giving comparable results, we design a voxel-wise refinement strategy (VRS) to integrate segmentation of CNN with shape priors to make it more anatomically correct. The VRS can confine the segmentation of CNN and remove outliers.

To show the capability of our approach, we test it on a range of mainstream CNN backbones, where our fusion framework demonstrates to give great improvement on results of all those CNN models. We further test our approach with half of the training data for CNN backbones. Experiments show that our adaptive fusion framework works better when data are more limited. Through the post-processing of our fusion framework, the performance of CNN baselines can be raised to the level comparable to that of CNN models trained with double of the data. This can be very useful in practice, where annotated medical image data are usually insufficient and difficult to be acquired.

We also investigate a vanilla neural network for fusion in our case, which demonstrates minimum performance improvement on CNN backbones. A plausible explanation is that for the majority of cases, CNN outperforms SSM, leading to the nonnecessity of the integration with SSM. Thus, in training data, there are relatively fewer cases where fusion between CNN and SSM is needed, which may cause a neural network biased toward the sole segmentation of CNN. However, our framework which conducts explicit fusion in a statistical way may avoid that issue. On the other hand, our approach is more transparent and interpretable compared with a neural network which is criticized to be a black box [49]. For clinical application, transparency is important because doctors may need to understand how and why the model produces its output for credible diagnosis [50]. In the future, we will further investigate constructing an end-to-end interpretable deep neural network to combine CNN with SSM in our future work.

At test time, our framework adopts an adaptive nearest neighbor fusion (ANNF) based on clusters formed in the difference score modelling (DSM) module during training. In our framework, Ds^* is manually crafted and separate to the segmentation performance of fusion. In the future, we will investigate incorporating those two processes to make the derivation of Ds^* learnable based on the feedback from the ultimate segmentation performance of fusion. We will also look into how to include credible shape variation in SSM for better segmentation in our approach. Past methods of adjusting SSM can be very vulnerable to noisy or corrupted image data, since they rely on pixel information or segmentation results of CNN models. Thus, we exclusively employ the mean shapes of SSM in this work, as they comprise the most general anatomical information [21]. This work primarily concentrates on the fusion of statistical shape information and neural networks. In future research, we aim to investigate advanced methods for exploiting SSM.

VI. Conclusion

In this paper, we propose an adaptive fusion framework to integrate the segmentation of CNN with shape priors provided by SSM. The experiment results demonstrate that our adaptive fusion framework can help CNN models to yield more accurate and more robust patella segmentation. We propose a difference score modelling module (DSM) that approximates the ΔDSC between CNN and SSM in order to benchmark their performance difference without knowing the ground truth. Then at test time, an adaptive nearest neighbor fusion module (ANNF) is proposed to utilize results of DSM. ANNF can automatically decide the contribution of CNN and SSM to the ultimate result based on their segmentation performance. Furthermore, we propose a voxel-wise refinement strategy (VRS) that utilizes an adjustable weighted average to amend results of CNN with shape priors. When the VRS is taken into consideration in fusion, the segmentation of CNN can be more anatomically correct. Extensive empirical experiments on diverse CNN backbones demonstrate the effectiveness of our approach on improving segmentation performance over different metrics. In the future, we plan to validate our approach on segmentation of other organs to develop it into an adaptable framework to improve deep learning methods with statistical shape modeling.

References

[1] J. A. Buckwalter, C. Saltzman, and T. Brown, “The impact of osteoarthritis: implications for research,” *Clinical Orthopaedics and Related Research®*, vol. 427, pp. S6–S15, 2004.

[2] R. C. Lawrence, D. T. Felson, C. G. Helmick, L. M. Arnold, H. Choi, R. A. Deyo, S. Gabriel, R. Hirsch, M. C. Hochberg, G. G. Hunder *et al.*, “Estimates of the prevalence of arthritis and other rheumatic conditions in the united states: Part II,” *Arthritis & Rheumatism*, vol. 58, no. 1, pp. 26–35, 2008.

[3] E. Christodoulou, S. Moustakidis, N. Papandrianos, D. Tsaopoulos, and E. Papageorgiou, “Exploring deep learning capabilities in knee osteoarthritis case study for classification,” in *International Conference on Information, Intelligence, Systems and Applications*. IEEE, 2019, pp. 1–6.

[4] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, “Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach,” *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[5] H. A. Wieland, M. Michaelis, B. J. Kirschbaum, and K. A. Rudolphi, “Osteoarthritis—an untreatable disease?” *Nature Reviews Drug Discovery*, vol. 4, no. 4, pp. 331–344, 2005.

[6] J. F. A. Eijkenboom, J. H. Waarsing, E. H. G. Oei, S. M. A. Bierma-Zeinstra, and M. van Middelkoop, “Is patellofemoral pain a precursor to osteoarthritis?” *Bone & Joint Research*, vol. 7, no. 9, pp. 541–547, 2018.

[7] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, “3d deeply supervised network for automatic liver segmentation from ct volumes,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 149–157.

[8] T. Heimann, B. van Ginneken, M. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. K. Binnig, H. Bischof, A. Bornik, P. Cashman, Y. Chi, A. Cordova, B. M. Dawant, M. Fidrich, J. D. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmüller, R. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H. Meinzer, G. Németh, D. S. Raicu, A. Rau, E. M. van Rikxoort, M. Rousson, L. Ruskó, K. A. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. M. Waite, A. Wimmer, and I. Wolf, “Comparison and evaluation of methods for liver segmentation from CT datasets,” *IEEE Trans. Medical Imaging*, vol. 28, no. 8, pp. 1251–1265, 2009.

[9] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, “H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes,” *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[10] Y. Li, L. Luo, H. Lin, H. Chen, and P.-A. Heng, “Dual-consistency semi-supervised learning with uncertainty quantification for covid-19 lesion segmentation from ct images,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 199–209.

[11] L. Liu, Q. Dou, H. Chen, J. Qin, and P. Heng, “Multitask deep model with margin ranking loss for lung nodule analysis,” *IEEE Trans. Medical Imaging*, vol. 39, no. 3, pp. 718–728, 2020.

[12] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, Q. Huang, M. Cai, and P. Heng, “Weakly supervised deep learning for whole slide lung cancer image analysis,” *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3950–3962, 2020.

[13] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, “Brain tumor segmentation with deep neural networks,” *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

[14] L. Lin, Q. Dou, Y.-M. Jin, G.-Q. Zhou, Y.-Q. Tang, W.-L. Chen, B.-A. Su, F. Liu, C.-J. Tao, N. Jiang *et al.*, “Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma,” *Radiology*, vol. 291, no. 3, pp. 677–686, 2019.

[15] Z. Tang, K. Chen, M. Pan, M. Wang, and Z. Song, “An augmentation strategy for medical image processing based on statistical shape model and 3D thin plate spline for deep learning,” *IEEE Access*, vol. 7, pp. 133 111–133 121, 2019.

[16] 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*, vol. 52, pp. 109–118, 2019.

[17] A. Tack, A. Mukhopadhyay, and S. Zachow, “Knee menisci segmentation using convolutional neural networks: data from the osteoarthritis initiative,” *Osteoarthritis and Cartilage*, vol. 26, no. 5, pp. 680–688, 2018.

[18] J. Ma, F. Lin, S. Wesarg, and M. Erdt, “A novel bayesian model incorporating deep neural network and statistical shape model for pancreas segmentation,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 480–487.

[19] F. Milletari, A. Rothberg, J. Jia, and M. Sofka, “Integrating statistical prior knowledge into convolutional neural networks,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 161–168.

[20] H. Chen, C. Wu, C. Lin, Y. Liu, and Y. Sun, “Automated segmentation for patella from lateral knee x-ray images,” in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2009, pp. 3553–3556.

[21] H. Seim, D. Kainmueller, M. Heller, H. Lamecker, S. Zachow, and H. Hege, “Automatic segmentation of the pelvic bones from CT data based on a statistical shape model,” in *Proceedings of the Eurographics Workshop on Visual Computing for Biomedicine*, C. P. Botha, G. L. Kindlmann, W. J. Niessen, and B. Preim, Eds. Eurographics Association, 2008, pp. 93–100.

[22] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3d medical image segmentation: a review," *Medical Image Analysis*, vol. 13, no. 4, pp. 543–563, 2009.

[23] D. Kainmueller, *Deformable meshes for medical image segmentation: accurate automatic segmentation of anatomical structures*. Springer, 2014.

[24] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam, "The use of active shape models for locating structures in medical images," in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 1993, pp. 33–47.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2015, pp. 3431–3440.

[26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[28] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3–11.

[29] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[30] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote. Sens. Lett.*, vol. 15, no. 5, pp. 749–753, 2018.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] F. Milletari, N. Navab, and S. 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.

[33] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[34] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images," *NeuroImage*, vol. 170, pp. 446–455, 2018.

[35] H. Chen, Q. Dou, X. Wang, J. Qin, J. C. Cheng, and P.-A. Heng, "3d fully convolutional networks for intervertebral disc localization and segmentation," in *International Conference on Medical Imaging and Augmented Reality*. Springer, 2016, pp. 375–382.

[36] L. Yu, J.-Z. Cheng, Q. Dou, X. Yang, H. Chen, J. Qin, and P.-A. Heng, "Automatic 3d cardiovascular mr segmentation with densely-connected volumetric convnets," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 287–295.

[37] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2013, pp. 246–253.

[38] Y. Liu, P. Zhao, X. Liu, M. Wu, L. Duan, and X. Li, "Learning user dependencies for recommendation," in *International Joint Conference on Artificial Intelligence*, C. Sierra, Ed. ijcai.org, 2017, pp. 2379–2385.

[39] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Underst.*, vol. 61, no. 1, pp. 38–59, 1995.

[40] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[41] I. T. Jolliffe, "Principal components in regression analysis," *Principal Component Analysis*, pp. 167–198, 2002.

[42] S. Belongie, J. Malik, and J. Puzicha, "Matching shapes," vol. 1, 02 2001, pp. 454–461 vol.1.

[43] F. Ambellan, H. Lamecker, C. v. Tycowicz, and S. Zachow, "Statistical shape models: understanding and mastering variation in anatomy," in *Biomedical Visualisation*. Springer, 2019, pp. 67–84.

[44] L. A. Zadeh, "Fuzzy sets," in *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh*. World Scientific, 1996, pp. 394–432.

[45] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Trans. Syst. Man Cybern.*, vol. 15, no. 4, pp. 580–585, 1985.

[46] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC medical imaging*, vol. 15, no. 1, pp. 1–28, 2015.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[48] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang *et al.*, "Evaluation of prostate segmentation algorithms for mri: the promise12 challenge," *Medical image analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[49] V. Buhrmester, D. Münch, and M. Arens, "Analysis of explainers of black box deep neural networks for computer vision: A survey," *Machine Learning and Knowledge Extraction*, vol. 3, no. 4, pp. 966–989, 2021.

[50] W. Liao, B. Zou, R. Zhao, Y. Chen, Z. He, and M. Zhou, "Clinical interpretable deep learning model for glaucoma diagnosis," *IEEE J. Biomed. Health Informatics*, vol. 24, no. 5, pp. 1405–1412, 2020.