Latent3DU-net: Multi-level Latent Shape Space Constrained 3D U-net for Automatic Segmentation of the Proximal Femur from Radial MRI of the Hip

Guodong Zeng[1], Qian Wang[2], Till Lerch[3], Florian Schmaranzer[3], Moritz Tannast[3], Klaus Siebenrock[3], and Guoyan Zheng[1(✉)]

[1] Institute of Surgical Technology and Biomechanics, University of Bern, Bern, Switzerland
guoyan.zheng@istb.unibe.ch
[2] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[3] Department of Orthopaedic Surgery, Inselspital, University of Bern, Bern, Switzerland

Abstract. Radial 2D MRI scans of the hip are routinely used for the diagnosis of the cam-type of femoroacetabular impingement (FAI) and of avascular necrosis (AVN) of the femoral head, which are considered causes of hip joint osteoarthritis in young and active patients. However, for computer assisted planning of surgical treatment, it is highly desired to have 3D models of the proximal femur. In this paper, we propose a novel volumetric convolutional neural network (CNN) based framework to fully automatically extract 3D models of the proximal femur from sparsely hip radial slices. Our framework starts with a spatial transform to interpolate sparse 2D radial MR images to a densely sampled 3D volume data. Automated segmentation of the interpolated 3D volume data is very challenging due to the poor image quality and the interpolation artifact. To tackle these challenges, we introduce a multi-level latent shape space constrained 3D U-net, referred as *Latent3DU-net*, to incorporate prior shape knowledge into voxelwise semantic segmentation of the interpolated 3D volume. Comprehensive results obtained from 25 patient data demonstrated the effectiveness of the proposed framework.

1 Introduction

Femoroacetabular Impingement (FAI) and avascular necrosis (AVN) of the femoral head are known causes of osteoarthritis of the hip joint in young and active patients. Depending on clinical and imaging findings, two types of impingement are distinguished: pincer impingement is the acetabular cause of FAI and is characterized by focal or general over-coverage of the femoral head. Cam impingement is the femoral cause of FAI and is due to aspherical portion of the femoral-neck junction [1]. On the other hand, in AVN the blood flow to

Y. Shi et al. (Eds.): MLMI 2018, LNCS 11046, pp. 188–196, 2018.
https://doi.org/10.1007/978-3-030-00919-9_22

the femoral head is interrupted, which can progressively lead to the collapse of the hip. A lot of joint-preserving treatments have been developed in an attempt to slow or reverse its progression, as it usually affects young patients [2]. MRI has been recognized as an important assisting tool for the diagnosis and the assessment of FAI and AVN as, in addition to the non-ionizing nature, MRI can capture the vascular status of the femoral head. Moreover, MR scanners typically have the capability to directly scan planes of arbitrary orientation. A radiologist can take advantage of this in order to acquire planes perpendicular to the curvature of the acetabulum. Such a scanning protocol is often referred to as radial imaging of the hip. The appeal of using radial scans over 3D MRI for image-assisted diagnosis is its motion insensitivity and reduced scanning time, as a typical radial scan of the hip consists of much fewer slices. Radial scans around the femoral neck axis are increasingly recognized as an important tool for morphological assessment of FAI.

To enhance surgeon's ability to assess the presence, location, and severity of impingement as well as to plan hip preservation surgery, computer-assisted diagnosis and planning systems have been developed [3]. In such a system, it is highly desired to have 3D models of the proximal femur, better derived from the radial MR images of the hip to avoid extra logistic efforts and cost.

The topic of automated MR image segmentation of the hip joint has been addressed by a few studies which relied on atlas-based segmentation [4], active shape models [5] and statistical shape models [6]. Recently, with the advance of deep convolutional neural network (CNN) based techniques, deep CNN-based methods, especially those based on fully convolutional networks (FCN), are introduced for segmentation of 3D volumetric data [7–9]. Despite impressive results achieved by these methods, they all assume that densely sampled 3D volumetric data are available. To the best of our knowledge, no 3D segmentation method has been proposed for segmenting the proximal femur that relies solely on sparse radial slices, though there exists work on segmentation of other organs such as the cardiac left ventricle from radial images [10]. The method introduced in [10] depends on a matching of 3D-active shape model to sparse, arbitrarily oriented image data. The initialization of the matching is done manually. After that, the matching is driven by feature points detected using fuzzy inference.

In this paper, we propose a novel volumetric FCN-based framework to fully automatically extract 3D models of the proximal femur from sparse radial MR images of the hip. More specifically, we first perform a spatial transform to interpolate the sparse radial slices to a densely sampled volumetric data. Automated segmentation of the proximal femur from such a 3D volumetric data is challenging due to the poor image quality and the interpolation artifacts. To solve these challenges, we introduce a multi-level latent shape constrained 3D U-net, referred as *Latent3DU-net*, to incorporate prior shape knowledge into a voxelwise semantic segmentation of the proximal femur from the interpolated 3D volume.

2 Methods

Figure 1 illustrate the proposed framework, which mainly consists of two steps, i.e., spatial transform and Latent3DU-net-based segmentation of the proximal femur. Below the details about each step will be presented.

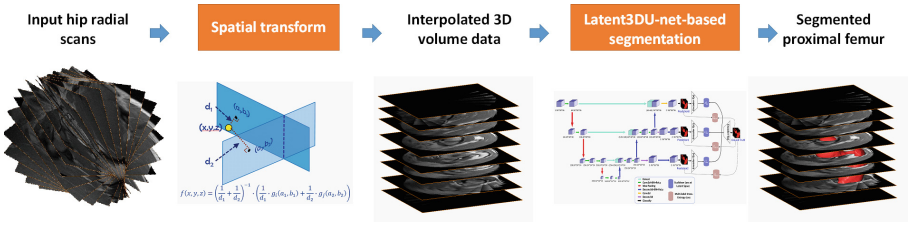

Fig. 1. A schematic illustration of the proposed framework, which mainly consists of two steps, i.e, spatial transform and Latent3DU-net-based 3D segmentation.

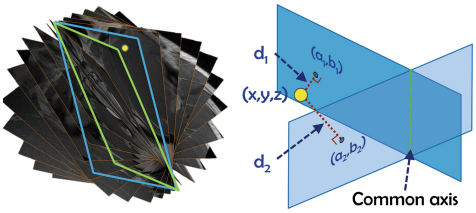

Fig. 2. A schematic illustration of how to do spatial transform.

2.1 Spatial Transform

The purpose of spatial transform is to interpolate the sparse hip radial slices to a densely sampled 3D volume data, which is done as follows. First, we compute the common axis of the radial scan by computing the intersections of all radial imaging planes. Around this axis, we then define a volume data sampling space. In order to fill the space with image data, we conduct an intensity interpolation as shown in Fig. 2. More specifically, for a point with coordinate (x, y, z) in the sampling space, we first determine the two radial planes which have the shortest distances to this point, as shown in Fig. 2-left, and denote these two plane as the ith plane and the jth plane, respectively. Assuming that the distances from this point to the two planes are d_1 and d_2, respectively, and further assuming that projections of this point onto these two planes have image coordinates of (a_1, b_1) and (a_2, b_2), respectively, we can compute the intensity value $f(x, y, z)$at

point (x, y, z) via interpolation. Although there exist many different interpolation methods, empirically we find that a distance inversely weighted interpolation as follows is enough for our purpose.

$$f(x,y,z) = (\frac{1}{d_1} + \frac{1}{d_2})^{-1} \cdot (\frac{1}{d_1} \cdot g_i(a_1, b_1) + \frac{1}{d_2} \cdot g_j(a_2, b_2)) \quad (1)$$

where $g_i(a_1, b_1)$ and $g_j(a_2, b_2)$ denotes the image intensity values of the two image point (a_1, b_1) and (a_2, b_2), respectively.

Figure 3 shows several slices extracted from an interpolated 3D volume data.

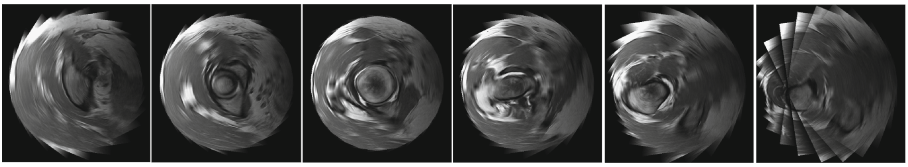

Fig. 3. Slices extracted from an interpolated 3D volume data.

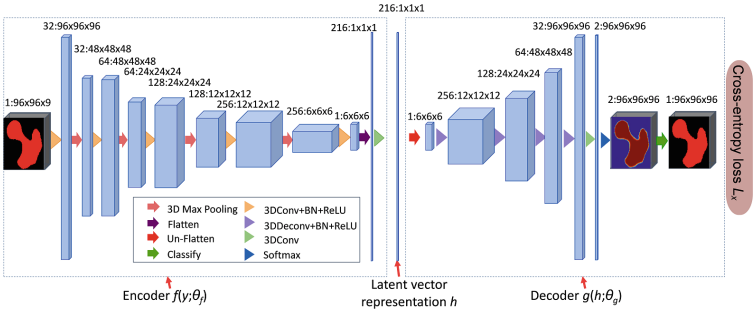

Fig. 4. Fully convolutional denoising auto-encoder for volumetric representation.

2.2 Segmentation of the Proximal Femur

It is useful to incorporate prior shape knowledge into image segmentation algorithms to obtain more accurate and plausible results. As summarized in a recent survey paper on deep learning in medical image analysis [11], most of the classification and regression models utilize a pixel-level loss function such as cross-entropy or Dice loss. Prior knowledge is usually incorporated in a post-processing step. Recently, based on the TL networks of [12], Oktay et al. [13] proposed anatomically constrained neural networks to incorporate anatomical prior knowledge into CNNs. In this paper, inspired by the fully convolutional denoising auto-encoder of [14], we propose a fully convolutional volumetric auto-encoder

that learns volumetric representation from noisy data. The learned volumetric representation can then be treated as a denoised generative vector representation of anatomical knowledge in a latent space. We further propose a multi-level latent shape constrained 3D U-net, referred as *Latent3DU-net*, for accurate segmentation of the proximal femur from the interpolated volume data.

Fully convolutional denoising auto-encoder. Figure 4 shows the architecture of the fully convolutional denoising auto-encoder to learn an end-to-end, voxel-to-voxel mapping. The left half of our network can be seen as an encoder stage that results in a condensed representation (indicated by "latent vector representation"). In the second stage (right half), the network reconstructs back the input from the latent vector representation by deconvolutional (3DDeconv) layers. The network is trained using cross-entropy loss. After training, the encoder $f(y;\theta_f)$ can be used to map a noisy volumetric label to a vector representation h in the latent space.

Latent3DU-net. Figure 5 illustrates the architecture of the Latent3DU-net. It is an extension of 3D U-net [7] with multi-level deep supervision. We further leverage multi-level Euclidean losses calculated at the latent space to enforce the prediction to follow the learnt shape/label distributions. More specifically, let W be the weights of main network and $\{w^c\}$ be the weights of classifiers. Then the cross-entropy loss function of a classier is: $L^c_{ce}(\chi; W, w^c) = \sum_{x_i \in \chi} -\log p(y_i = t(x_i)|x_i; W, w^c)$, where χ represents the training samples; y_i is the ground truth label; $p(y_i = t(x_i)|x_i; W, w^c)$ is the probability of target class label $t(x_i)$ corresponding to sample $x_i \in \chi$. Additionally, as shown in Fig. 5, the Euclidean loss at latent space of a classifier is: $L^c_{he} = \|f(\phi(x)^c; \theta_f) - f(y; \theta_f)\|_2^2$, where $\phi(x)^c$ is the prediction of the cth classifier and y is the ground truth segmentation. Then the total loss function of the Latent3DU-net is:

$$L(\chi, W, \{w^c\}) = \sum_c (\alpha^c L^c_{ce}(\chi, W, w^c) + \lambda^c L^c_{he}) + \gamma(\psi(W) + \sum_c \psi(w^c)) \quad (2)$$

where $\psi()$ is the regularization term (L_2 norm in our experiment) with hyper parameter γ, $\{\alpha^c\}$ and $\{\lambda^c\}$.

For both fully convolutional denoising auto-encoder as shown in Fig. 4 and Latent3DU-net as shown in Fig. 5. All convolutional layers use kernel size of $3 \times 3 \times 3$ and strides of 1 and all max pooling layers use kernel size of $2 \times 2 \times 2$ and strides of 2. In the convolutional and deconvolutional layers of our networks, batch normalization (BN) [15] and rectified linear units (ReLU) [16] are adopted to speed up the training and to enhance the gradient back-propagation.

3 Experiments and Results

3.1 Dataset and Preprocessing

We evaluated the proposed framework on a dataset consisting of MR gadolinium-enhanced radial scans of 25 patients with symptomatic FAI or AVN. No 3D MR data was available for these patients. The intra-slice spacing of these radial scans

is 0.28 mm and the size of the images is either 448×448 or 512×512. There are 14 radial slices in each radial scan. A reference, manual segmentation of every slice of the radial scans was also provided. From the 2D manual segmentation of each radial scan, we used the method introduced by Carr et al. [17] to reconstruct a smooth 3D surface model of the proximal femur. We then conducted spatial transform for all radial scans. After that, we also converted the reconstructed 3D surface models of the proximal femur into dense binary volumetric labels. As there was no 3D MR scan available for these patients, we took the corresponding dense binary volumetric labels as the ground truth segmentation.

All the interpolated volume data and the corresponding binary volumetric labels were rescaled to a size of $96 \times 96 \times 96$ due to memory restrictions. To enlarge the training samples and to mitigate possible over-fitting problem, random noise was injected: random value between $(-3, 3)$ was added to each voxel. Finally, each training sample was normalized as zero mean and unit variance before fed into the network. A standard 5-fold cross-validation study was performed to evaluate the performance of the proposed framework.

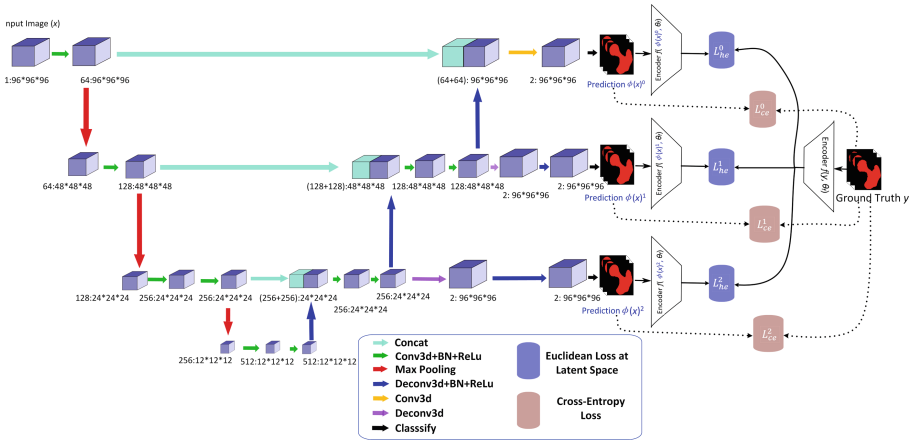

Fig. 5. Illustration of the architecture of Latent3DU-net.

3.2 Training

We trained our networks from scratch. The training was done in two stages. In the first stage, the fully convolutional denoising auto-encoder was trained for $5,000$ iterations. After that, we trained the Latent3DU-net for another $5,000$ iterations. All weights were initialized from a Gaussian distribution ($\mu = 0, \sigma = 0.01$) and were updated by the stochastic gradient descent(SGD) algorithm (momentum $= 0.9$, weight decay $= 0.005$). For each stage of the training, the initial learning rate was initialized as 1×10^{-3} and halved by every $1,500$ times.

3.3 Testing and Evaluation

In the inference phase, only the prediction $\phi(x)^0$ of the 0th classifier was used to generate the segmentation result. After that, the segmentation was rescaled back to the original size. Implemented with Python using TensorFlow framework on a workstation with a 3.6GHz Intel(R) i7 CPU and a GTX 1080 Ti graphics card with 11GB GPU memory, on average it took Latent3DU-net about 10 s to finish one test case while the spatial transform took another 30 s.

The segmented results were compared with the associated ground truth segmentation. For each test case, we evaluated the distance between the surface models extracted from different segmentation as well as the volume overlap measurements including Dice overlap coefficient, precision and recall.

For further comparison, we implemented the 3D U-net with multi-level deep supervision (we referred it as "3DU-net-MLDS") as introduced in [9], which reported state-of-the-art results when applied to segmentation of the proximal femur from 3D MR images, and a 3D U-net [7].

Table 1. Comparison of the results achieved by different methods. HD: Hausdorff distance; ASD: average surface distance; DC: Dice Coefficient.

Methods	DC	HD (mm)	ASD (mm)	Precision	Recall
Latent3DU-net	0.954	6.18	0.74	0.958	0.950
3DU-net-MLDS	0.943	12.07	0.83	0.959	0.929
3DU-net	0.941	10.36	0.92	0.943	0.940

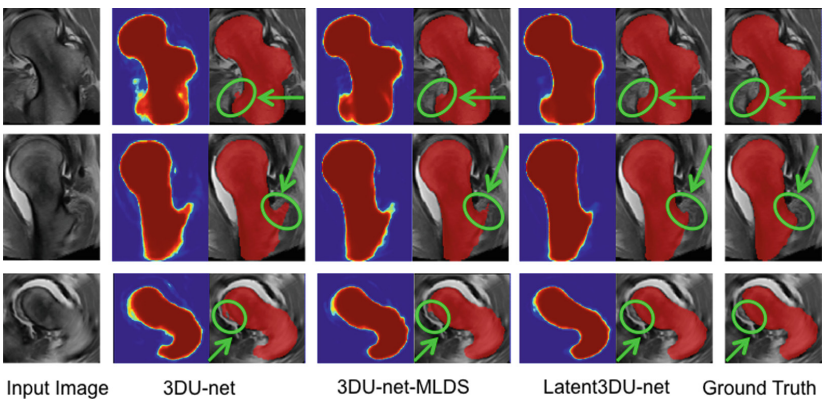

Fig. 6. Qualitative comparison of different methods. Data cropped for visualization purpose. For each method, the probability maps and the segmentation results are displayed. Green circles highlight the differences of different methods.

3.4 Results

Table 1 shows the segmentation results achieved by different methods. On average, our method achieved a mean ASD of 0.74 mm, a mean HD of 6.18 mm, a mean DC of 0.954, a mean precision of 0.958 and a mean recall of 0.95. When evaluated on the same dataset, the method introduced in [9] achieved a mean ASD of 0.83 mm, a mean HD of 12.07 mm, and a mean DC of 0.943 while 3D U-net achieved a mean ASD of 0.92 mm, a mean HD of 10.36 mm, and a mean DC of 0.941. Pairwise T-test on the DC measurements demonstrated that the difference between our method and the method introduced in [9] is statistically significant (p-value < 0.01). Figure 6 shows a qualitative comparison of the results achieved by these three methods.

4 Conclusions

In this paper, we presented a deep CNN-based framework to fully automatically extract a 3D model of the proximal femur from sparse hip radial slices. To the best of our knowledge, this is probably the first study addressing such a problem using deep learning. We compared the results achieved by our method to those achieved by a state-of-the-art methods. The experimental results clearly demonstrated the effectiveness of incorporating the latent space constraint for accurate segmentation of the proximal femur.

Acknowledgments. This study was partially supported by the Swiss National Science Foundation via project 205321_163224/1.

References

1. Leunig, M., Beaule, P., Ganz, R.: The concept of femoroacetabular impingement: current status and future perspectives. Clin. Orthop. Relat. Res. **467**, 616–622 (2009)
2. Chughtai, M., Piuzzi, N.: An evidence-based guide to the treatment of osteonecrosis of the femoral head. Bone Joint J. **99**(10), 1267–1279 (2017)
3. Tannast, M., Kubiak-Langer, M.: Noninvasive three-dimensional assessment of femoroacetabular impingement. J. Orthop. Res. **25**(1), 122–131 (2007)
4. Xia, Y., Fripp, J.: Automated bone segmentation from large field of view 3d MR images of the hip joint. Phys. Med. Biol. **58**(20), 7375–7390 (2013)
5. Arezoomand, S., Lee, W.: A 3d active model framework for segmentation of proximal femur in MR images. Int. J. CARS **10**(1), 55–66 (2015)
6. Chandra, S.S., Xia, Y., et al.: Focused shape models for hip joint segmentation in 3d magnetic resonance images. Med. Image Anal. **18**(3), 567–578 (2014)
7. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49

8. Dou, Q., Yu, L.: 3d deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. **41**, 40–54 (2017)
9. Zeng, G., Yang, X., Li, J., Yu, L., Heng, P.-A., Zheng, G.: 3D U-net with multi-level deep supervision: fully automatic segmentation of proximal femur in 3D MR images. In: Wang, Q., Shi, Y., Suk, H.-I., Suzuki, K. (eds.) MLMI 2017. LNCS, vol. 10541, pp. 274–282. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_32
10. Van Assen, H., Danilouchkine, M.: Spasm: a 3d-ASM for segmentation of sparse and arbitrarily oriented cardiac MRI data. Med. Image Anal. **10**(2), 286–303 (2006)
11. Litjens, G., Kooi, T.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
12. Girdhar, R., Fouhey, D.F., Rodriguez, M., Gupta, A.: Learning a predictable and generative vector representation for objects. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 484–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_29
13. Oktay, O., Kamnisas, K.: Anotomically constrained neural networks (ACNN): application to cardiac image enhancement and segmentation. IEEE Trans. Med. Imaging **37**(2), 384–395 (2018)
14. Sharma, A., Grau, O., Fritz, M.: VConv-DAE: deep volumetric shape learning without object labels. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 236–250. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_20
15. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. Proc. ICML 448–456 (2015)
16. Krizhevsky, A., Ilya, S., Hinton, G.: Imagenet classification with deep convolutional neural networks. Proc. NIPS 1097–1105 (2012)
17. Carr, J., Beatson, R., et al.: Reconstruction and representation of 3d objects with radial basis functions. Computer Graphics (2001) 67–76