Medical Image Analysis 57 (2019) 149–164

Contents lists available at ScienceDirect

Medical Image Analysis

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

Holistic decomposition convolution for effective semantic segmentation of medical volume images

Guodong Zeng[a,b,1], Guoyan Zheng[a,1,*]

[a] *School of Biomedical Engineering, Shanghai Jiao Tong University, No.800 Dongchuan Road, Shanghai 200240, China*
[b] *Institute for Surgical Technology and Biomechanics (ISTB), University of Bern, Stauffacherstrasse 78, Bern 3014, Switzerland*

ARTICLE INFO

Article history:
Received 24 December 2018
Revised 22 May 2019
Accepted 4 July 2019
Available online 8 July 2019

Keywords:
Semantic segmentation
Volume images
Deep learning
Convolutional neural networks,

ABSTRACT

Convolutional Neural Networks (CNNs) have achieved state-of-the-art performance in many different 2D medical image analysis tasks. In clinical practice, however, a large part of the medical imaging data available is in 3D, e.g, magnetic resonance imaging (MRI) data, computed tomography (CT) data and data generated by many other modalities. This has motivated the development of 3D CNNs for volumetric image segmentation in order to benefit from more spatial context. Due to GPU memory restrictions caused by moving to fully 3D, state-of-the-art methods depend on subvolume/patch processing and the size of the input patch is usually small, limiting the incorporation of larger context information for a better performance. In this paper, we propose a novel Holistic Decomposition Convolution (HDC), which learns a number of separate kernels within the same layer and can be regarded as an inverse operation to the previously introduced Dense Upsampling Convolution (DUC), for an effective and efficient semantic segmentation of medical volume images. HDC consists of a periodic down-shuffling operation followed by a conventional 3D convolution. HDC has the advantage of significantly reducing the size of the data for sub-sequential processing while using all the information available in the input irrespective of the down-shuffling factors. We apply HDC directly to the input data, whose output will be used as the input to sub-sequential CNNs. In order to achieve volumetric dense prediction at final output, we need to recover full resolution, which is done by using DUC. We show that both HDC and DUC are network agnostic and can be combined with different CNNs for an improved performance in both training and testing phases. Results obtained from comprehensive experiments conducted on both MRI and CT data of different anatomical regions demonstrate the efficacy of the present approach.

1. Introduction

Segmentation of important organs or structures from volumetric medical images, such as 3D magnetic resonance imaging (MRI) data and computed tomography (CT) data, is a prerequisite for many clinical applications including disease diagnosis, surgical planning and computer assisted interventions (Neubert et al., 2012; Xia et al., 2013; 2014; Castro-Mateos et al., 2014; Chandra et al., 2014; Arezoomand et al., 2015; Chen et al., 2015; Luo et al., 2016; Zeng et al., 2017; Zheng et al., 2017; Chen et al., 2017a; Karasawa et al., 2017; Dong et al., 2018; Roth et al., 2018c; Liu et al., 2018b; Li et al., 2018; Roth et al., 2018a; 2018b). For example, bony structure segmentation from hip MR images will greatly facilitate the applications of MR images for hip preservation surgical planning and simulation (Xia et al., 2013; 2014; Chandra et al., 2014; Arezoomand et al., 2015; Zeng et al., 2017); segmentation of lumbar intervertebral discs (IVDs) from 3D MR data is a step prior to the quantitative assessment of IVD abnormalities for the diagnosis of spinal disease (Neubert et al., 2012; Castro-Mateos et al., 2014; Chen et al., 2015; Zheng et al., 2017; Li et al., 2018); pancreas segmentation from 3D abdominal CT scans is crucial in computer-aided screening, diagnosis, and quantitative assessment (Karasawa et al., 2017; Roth et al., 2018a; 2018b). Manually delineating organs or structures from medical volume images is labor-intensive, tedious and subjects to intra- and inter-observer variations. This has motivated numerous research works on developing methods for automated segmentation. In this paper, we would like to develop an automatic volumetric image segmentation pipeline that can be applied off-the-shelf to segmenting both MRI and CT data of organs or structures of different volume sizes.

Automated segmentation of volumetric images is a challenging task. Firstly, as shown in Fig. 1, left, the shape and appearance of

[*] Corresponding author.
E-mail addresses: guoyan.zheng@sjtu.edu.cn, guoyan.zheng@ieee.org (G. Zheng).
[1] Authors contribute equally to the paper.

https://doi.org/10.1016/j.media.2019.07.003
1361-8415/© 2019 Elsevier B.V. All rights reserved.

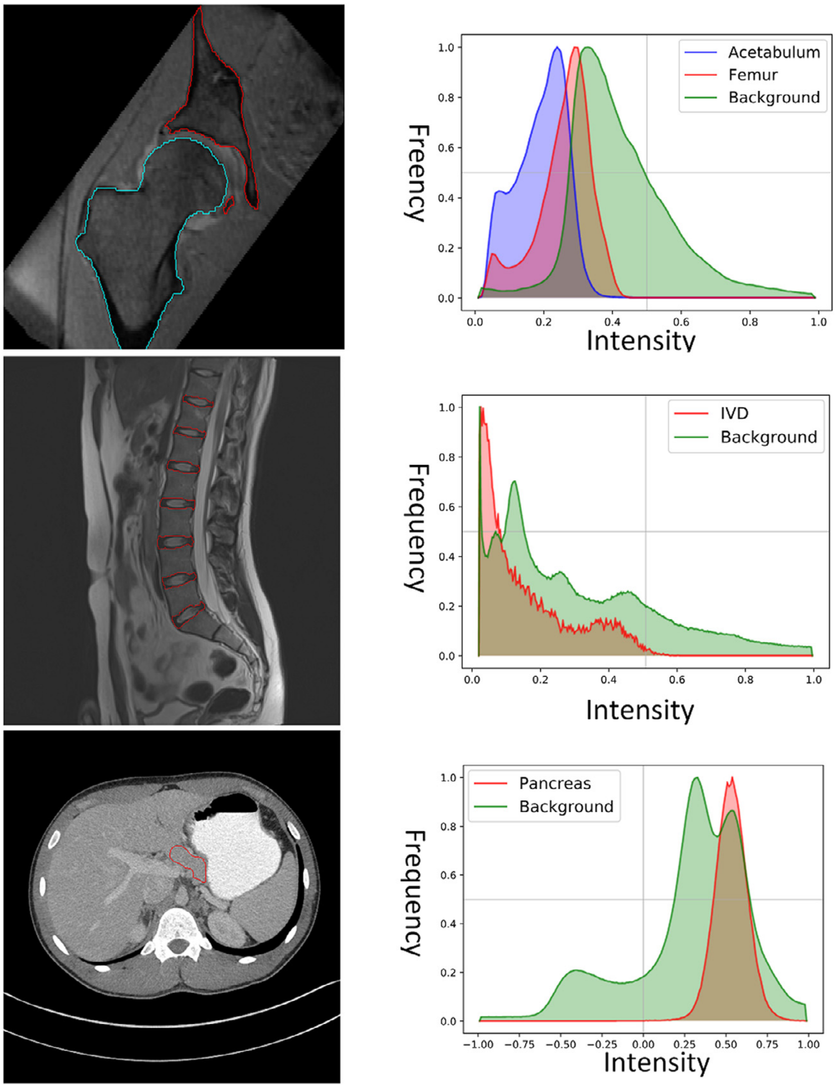


Fig. 1. Slices overlapped with ground truth segmentations of 3 different organs (left) and the corresponding volumetric intensity distributions (right). From top to down: hip segmentation from T1 MR images (the 1st row), IVD segmentation from T2 MR images (the 2nd row), and pancreas segmentation from CT data (the 3rd row).

the target organs or structures vary greatly from task to task. For example, in comparison with the hip joint, both the IVDs and the pancreas occupy a very small fraction of the complete volume. Secondly, subjected to the low contrast and analogous intensity distributions between the targeted organs or structures and surrounding tissues (Fig. 1, right), it is difficult to estimate the ambiguous boundaries from the complex background. Thirdly, these anatomical structures often have highly varying shapes, scales and appearance among patients, which are hard to capture. Lastly, we have to deal with not only mono-modality problems but also multi-modality problems. In the past decades, various algorithms have been introduced in order to meet the challenge. Large number of previous efforts focus on incorporating prior knowledge in the form of statistical shape models, deformable models, level sets, multi-atlas or graph models in order to deal with large variations of appearance and shape. Despite significant progress, many are not used in clinical practice due to the requirement of physician input and/or the requirement of technical assistance (Kronman and Joskowicz, 2016). Later on, machine learning-based methods have gained more and more interest. A crucial step in the design of such systems is the extraction of discriminant features from the images, which is usually done by human researchers. The limited representation capability of these hand-crafted features makes it difficult to handle large variations of appearance and shape.

The more recent development of deep neural networks, and in particular convolutional neural networks (CNN) (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015; Long et al., 2015), suggests another course of methods to solve the challenging image segmentation tasks (Litjens et al., 2017). Contrary to conventional shallow learning methods, where feature design is crucial, deep learning methods automatically learn hierarchies of relevant features directly from the training data. Early works (Prasoon et al., 2013; Roth et al., 2014) treat the image segmentation as a classification problem with sliding window where CNNs are applied to input patches to classify the central pixel/voxel of each patch. As classifying each pixel/voxel in a sliding window fashion results in orders of magnitude of redundant calculation, most of recent works (Ronneberger et al., 2015; Havaei et al., 2017; Kamnitsas et al., 2017) are based on Fully Convolutional Networks (FCN) (Long et al., 2015), which can process the input image in an end-to-end way and can provide a full resolution segmentation map (Litjens et al., 2017). In several biomedical image segmentation benchmarking competitions, methods built on CNNs (Ronneberger et al., 2015; Chen et al., 2016b; Havaei et al., 2017; Zheng et al., 2017; Yu et al., 2017b) are on the top list of the associated leaderboard. Despite the fact that CNN-based methods have achieved state-of-the-art performance in many different 2D medical image analysis tasks, in clinical practice, however, a large part of the medical imaging data available is in 3D. This has motivated the development of 3D CNNs for volumetric image segmentation of both soft and hard tissues in order to benefit from more spatial context. Due to GPU memory restrictions caused by moving to fully 3D, state-of-the-art methods (Çiçek et al., 2016; Milletari et al., 2016; Kamnitsas et al., 2017; Li et al., 2017; Dou et al., 2017; Chen et al., 2017a; Liu et al., 2018b; Roth et al., 2018b; Shi et al., 2018) depend on subvolume/patch processing. The size of the input patch is usually small if no specialized hardware with large GPU memory is used, limiting the incorporation of larger context information for a better performance. To tackle these challenges, we present a novel and efficient approach which allows for using large size of patches for an effective and efficient semantic segmentation of volumetric images. Our contributions can be summarized as follows:

- First, we propose a novel Holistic Decomposition Convolution (HDC), which learns a number of separate kernels within the same layer and can be regarded as an inverse operation to the previously introduced Dense Upsampling Convolution (DUC) (Shi et al., 2016; Wang et al., 2018). HDC consists of a periodic down-shuffling operation followed by a conventional 3D convolution. HDC has the advantage of significantly reducing the size of the data for sub-sequential processing while using all the information available in the input irrespective of the down-shuffling factors. We apply HDC directly to the input data, whose output will be used as the input for sub-sequential CNNs. In order to achieve volumetric dense prediction at final output, we need to recover full resolution, which is done by using DUC.
- Second, we conduct extensive ablation studies to validate the proposed approach on the task of segmentation of both large and small objects using mono-modality and multi-modality MR images respectively. We show that HDC and DUC are network agnostic and can be combined with different FCNs for an improved performance on both training and testing phases. More specifically, we demonstrate that the improved performance can be obtained when HDC and DUC are used with 3D U-net (Çiçek et al., 2016), 3D V-net (Milletari et al., 2016), and HighRes3DNet (Li et al., 2017), respectively. We investigate the influence of the down-shuffling factors on the segmentation results. We further demonstrate that the best results are obtained when HDC and DUC are used with 3D U-net (Çiçek et al., 2016). We refer such a net as 3D Large Patch U-net (3D LP-U-Net) as it allows for using large patches than the previously introduced 3D U-net. Finally, we investigate the advantages of HDC and DUC when compared to strided convolution and transposed convolution, respectively.
- Third, in addition to two tasks in the ablation studies, we apply the 3D LP-U-net off-the-shelf to two other typical yet highly challenging segmentation tasks, i.e., IVD segmentation from T2 MR images and pancreas segmentation from CT data. We conduct comprehensive cross-validation experiments on open datasets. We compare the performance of 3D LP-U-net with that of state-of-the-art methods. We have achieved comparable or better results than many state-of-the-art methods in all four segmentation tasks.

2. Related work

In this section, we first provide a review of CNN-based volumetric image segmentation methods, followed by a review of other relevant segmentation methods for different image modalities and organs.

2.1. CNN-based volumetric image segmentation

Recent literature witnessed the successful applications of CNN-based methods for volumetric medical image segmentation tasks. CNN-based volumetric medical image segmentation methods can be divided into two categories: 2D CNN-based and 3D CNN-based.

2D CNN-based methods (Ronneberger et al., 2015; Roth et al., 2015; Moeskops et al., 2016; Havaei et al., 2017) often perform volumetric image segmentation slice by slice, and then fuse the 2D segmentation results to obtain a 3D segmentation. Before the introduction of FCNs, segmentation is mostly done in a sliding window fashion, leading to redundant calculation. For example, Roth et al. (2015) introduced multi-level deep convolutional networks for automated pancreas segmentation where they used a sliding window approach to extract 2.5D image patches composed of axial, coronal and sagittal planes at each candidate location. After the introduction of FCNs, almost all the 2D segmentation methods are built upon on 2D FCNs. Havaei et al. (2017) introduced a two-pathway FCN for brain tumor segmentation. The U-net (Ronneberger et al., 2015) is one of the most popular 2D deep learning networks and has been applied to many different medical image segmentation tasks (Liu et al., 2018a; Norman et al., 2018). It is based on encoder-decoder type architecture, where the encoder part focuses on analysis and feature representation from the input data while the decoder part generates segmentation map, relying on the learned features from the encoder part. Shortcut connections are established between layers of equal resolution in the encoder and decoder paths to facilitate forward and backward information flow.

Though 2D or 2.5D CNN-based methods achieved greatly improved results over non-CNN-based approaches, they cannot make full use of the spatial context information encoded in the volumetric data. This has motivated the development of various 3D CNN-based methods (Çiçek et al., 2016; Milletari et al., 2016; Kamnitsas et al., 2017; Dou et al., 2017; Li et al., 2017; Liu et al., 2018b; Roth et al., 2018c; Shi et al., 2018), aiming at taking full advantage of powerful volumetric representation across all three spatial dimension. For example, Kamnitsas et al. (2017) proposed a dual pathway, 11 layers deep 3D multi-scale CNN with fully connected Conditional Random Field (CRF) for brain lesion segmentation and achieved state-of-the-art performance. Çiçek et al. (2016) proposed the 3D U-net as an extension to the 2D U-net by replacing all 2D operations with their 3D counterparts. By incorporation of residual blocks and using a similar architecture as the 3D U-net, Milletari et al. (2016) proposed the 3D V-net for volumetric medical image segmentation. Another 3D CNN-based approach that benefits from residual learning is VoxResNet (Chen et al., 2017b), aiming at increasing the network depth to 25 layers for generating more representative features to deal with the large variations of 3D brain MR images. Inspired by Xie and Tu (2015), Dou et al. (2017) proposed 3D Deeply Supervised Networks (DSN) for automated segmentation of volumetric medical images by including auxiliary supervision via side outputs. Recently, Chen et al. (2017a) proposed a 3D feature-enhance network for automatic femur segmentation. Liu et al. (2018b) introduced a cascaded deep convolutional neural network for joint segmentation and genotype prediction of brainstem gliomas. Roth et al. (2018c) introduced cascaded 3D full convolutional networks for medical image segmentation.

One thing common to all these 3D CNN-based approaches is that they all follow a fully convolutional downsample-upsample pathway. More specifically, the downsampling path tries to achieve higher-level feature abstraction by gradually downsampling low-level features with high spatial resolutions while the upsampling path aims to upsample the learned high-level features to achieve a full-resolution segmentation. Deviating from the fully convolutional downsample-upsample pathway, Li et al. (2017) proposed a high-resolution network architecture which they referred as “HighRes3DNet” for the segmentation of fine structures in volumetric images. HighRes3DNet preserves the spatial resolution throughout the layers and the enlargement of the receptive field is then achieved by incorporating dilated convolution. Shi et al. (2018) introduced an approach called “Bayesian VoxDRN” which was built upon dilated convolution and residual connection by extending 2D dilated residual network (Yu et al., 2017a) to 3D. In this work, in order to achieve volumetric dense prediction, they proposed to use DUC (Shi et al., 2016; Wang et al., 2018) to get voxel-level predictions at the output. Additional advantage of Bayesian VoxDRN includes the output of a measure of model uncertainty, which is achieved by a dropout-based Monte Carlo sampling during testing to generate a posterior distribution of the voxel class labels.

2.2. Segmentation of hip MR images

Automated segmentation of bony structures in hip MR images will greatly facilitate the applications of MR images for planning and simulation of hip preservation surgery such as femoroacetabular impingement (FAI) treatment and periacetabular osteotomy (PAO) surgery. The topic of automated MR image segmentation of the hip joint has been addressed by a few studies which relied on atlas-based segmentation (Xia et al., 2013), graph-cut (Xia et al., 2014), deformable model (Gilles and Magnenat-Thalmann, 2010; Arezoomand et al., 2015), or statistical shape model (Chandra et al., 2014). Recent works on using CNNs to segment 3D musculoskeletal images have been focused on developing learning-based segmentation models in 2D and then using post-processing such as 3D deformable models (Liu et al., 2018a; Norman et al., 2018) to provide final segmentation mask. These networks are usually based on encoder-decoder type architecture such as the U-net. The additional incorporation of deformable modelling steps in the segmentation pipeline impedes with the benefits of end-to-end learning-based segmentation approaches. This has motivated the development of fully 3D CNN-based methods. For example, Zeng et al. (2017) introduced a method based on 3D U-net with multi-level deep supervision. In another study, Deniz et al. (2018) compared the performances of 3D U-net with that of 2D U-net on segmenting the proximal femur from MR images and they found that 3D U-net provided better segmentation accuracy.

2.3. IVD segmentation from MR images

Although almost every medical imaging modality has been used to evaluate lumbar degenerative disc disease, MRI is widely recognized as the imaging technique of choice for the assessment of lumbar IVD abnormalities due to its excellent soft tissue contrast and no ionizing radiation. This, in turn, has sparked specific interest in developing methods for automated image analysis and quantification for the diagnosis of spinal diseases using MR images. For IVD segmentation, there exist methods based on Hough Transform (Shi et al., 2007), watershed algorithm (Chevrefils et al., 2009), atlas registration (Michopoulou et al., 2009), Adaboost and normalized-cut (Huang et al., 2009), graph cuts with geometric priors from neighboring discs (Ayed et al., 2011), statistic shape model (Neubert et al., 2012) and machine learning (Chen et al., 2015). With the recent advance of deep learning techniques, many researchers have proposed deep learning based methods to segment IVDs from MR images (Chen et al., 2016a; Zeng and Zheng, 2017; Li et al., 2018). For example, Chen et al. (2016a) introduced a 3D deeply supervised FCN to localize and segment IVDs, which achieved the state-of-the-art localization performance in MICCAI 2015 IVD localization and segmentation challenge (Zheng et al., 2017). Zeng and Zheng (2017) developed a deeply supervised multi-scale FCN for automatic segmentation of IVDs in 3D MR images and achieved better results than those reported in MICCAI 2015 IVD localization and segmentation challenge (Zheng et al., 2017). Recently, Li et al. (2018) introduced a 3D multi-scale FCN with random modality voxel dropout learning for IVD localization and segmentation from multi-modality MR images.

2.4. Pancreas segmentation from CT data

Accurate pancreas segmentation from 3D abdominal CT scans is crucial for many clinical applications. Because of the large shape and size variations across different patients and because of the low contrast to surrounding tissues, automated pancreas segmentation from CT data remains a very challenging task. Early attempts are mainly based on the multi-atlas registration and label fusion (MALF) framework (Shimizu et al., 2010; Wolz et al., 2013; Karasawa et al., 2017). With the advent of deep learning based methods, both 2D CNN-based methods as well as 3D CNN-based methods have been proposed for automatic pancreas segmentation. For example, Roth et al. (2015) proposed to use a set of multi-scale and multi-level deep CNNs applied in a sliding window fashion on local image patches for automated pancreas segmentation. They achieved an average Dice Overlap Coefficient (DOC) of 71.8 ± 10.7%. Later on, they proposed a two-stage approach combining random forest regression based pancreas localization with holistically-nested CNNs on the three orthogonal axial, sagittal and coronal views (Roth et al., 2018a). An average DOC of 81.27 ± 6.27% was reported. Cai et al. (2017) proposed a new convolutional/recurrent neural network architecture to address the contextual learning and segmentation consistency problem caused by processing sequences of 2D image slices independently. Zhou et al. (2017) proposed a coarse-to-fine framework where the rough pancreas region was found in the coarse level and then a 2D FCN-based fixed-point model was used to refine the pancreas region iteratively. The same group later proposed a recurrent saliency transformation network for segmenting pancreas where a saliency transformation module was used to repeatedly convert segmentation probability map from the previous iteration as spatial weights to the current iteration (Yu et al., 2018).

3. Methods

In this section, we will first briefly present the usage of DUC for semantic segmentation, followed by a detailed description of HDC. We will then show how to combine HDC and DUC with FCNs for effective segmentation of volumetric images.

3.1. Dense upsampling convolution for semantic segmentation

For a typical FCN-based approach that follows the downsample-upsample pathway, in order to achieve volumetric dense prediction, we need to recover full resolution at output. Conventional methods such as bilinear upsampling (Yu et al., 2017a) is not attractive as the upsampling parameters are not learnable. Deconvolution could be an alternative but, unfortunately, it can easily lead to “uneven overlap”, resulting in checkboard artifacts (Aitken et al., 2017). In Shi et al. (2016), DUC, which consists of low-resolution convolutions with a periodic up-shuffling operation (see Fig. 2

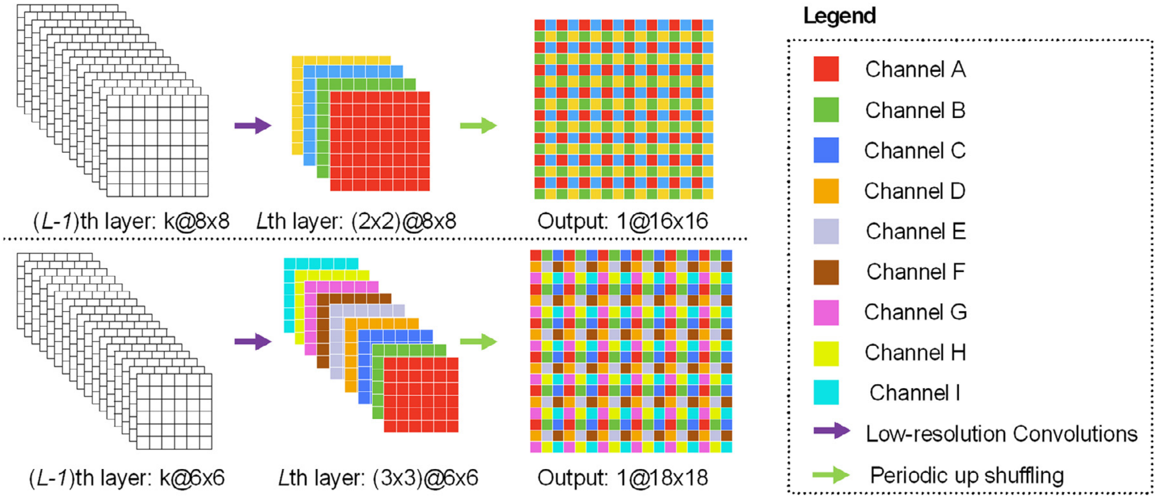


Fig. 2. A schematic view of the DUC in 2D which consists of low-resolution convolutions and a periodic up-shuffling operation. Top row: a 2D case when the upsampling factor along each dimension is 2; bottom row: a 2D case when the upsampling factor along each dimension is 3. The numbers at the bottom of each figure follow a format "number of channels@feature dimensions".

for a schematic view in 2D), was proposed to jointly learn the feature extraction and upsampling weights for super-resolution reconstruction. DUC was later used as the last layer for semantic segmentation in Wang et al. (2018) and Shi et al. (2018) where mathematically DUC was defined as below:

$$DUC(I^{HR}; W_L, b_L) = PUS\big(W_L * f^{L-1}(I^{LR}) + b_L\big) \tag{1}$$

where L is the number of layers in our neural network; $*$ is convolution operation; $f^{L-1}(I^{LR})$ is the low-resolution (LR) feature maps at the $(L-1)$th layer; W_L, b_L are trainable weights and bias for the last layer, respectively; the final layer (the Lth layer as shown in Fig. 2) has $(C \times (n_x \times n_y \times n_z))$ channels with C being the number of classes; n_x, n_y, and n_z being the upsampling factors along three spatial axes, respectively; PUS is a periodic up-shuffling operator which aims to rearrange the elements of a $(d \times h \times w) \times (C \times (n_x \times n_y \times n_z))$ tensor T_{LR} to a high-resolution (HR) tensor T_{HR} of shape $(n_x \times d) \times (n_y \times h) \times (n_z \times w) \times C$ (see Fig. 2 for a schematic illustration), where d, h, and w denote depth, height, and width of the LR feature maps, respectively; $(n_x \times d)$, $(n_y \times h)$, and $(n_z \times w)$ are the depth, height, and width of the HR feature maps, respectively. The mapping is done in 3D as follows:

$$\begin{aligned} T_{HR}(x, y, z, c) = {} & T_{LR}(\lfloor x/n_x \rfloor, \lfloor y/n_y \rfloor, \lfloor z/n_z \rfloor, c + mod(x, n_x) \cdot C \\ & + mod(y, n_y) \cdot n_x \cdot C + mod(z, n_z) \cdot n_x \cdot n_y \cdot C) \end{aligned} \tag{2}$$

where x, y, z, c are the coordinates of the voxels in the HR space with $x \in [0, (n_x \times d) - 1], y \in [0, (n_y \times h) - 1], z \in [0, (n_z \times w) - 1], c \in [0, C - 1]$.

3.2. Holistic decomposition convolution

HDC can be regarded as the inverse operation to DUC. As shown in Fig. 3, HDC consists of a periodic down-shuffling operation and low-resolution convolutions. HDC is designed to be directly applicable to the input data with the aim to reduce the size of the data for sub-sequential processing while using all the information available in the input irrespective of the down-shuffling factors. This is also the reason why we call this novel operation as "Holistic Decomposition Convolution". More specifically, let's assume that the size of the input data (I^{HR}) is $(n_x \times d) \times (n_y \times h) \times (n_z \times w) \times C$ and the size of the output from HDC is $d \times h \times w \times k$, where $(n_x \times d)$, $(n_y \times h)$, $(n_z \times w)$ and C denote depth, height, width, and number of channels of the input data, respectively; n_x, n_y, n_z are the down-shuffling factors along the three spatial axes, respectively; k is the number of feature maps in the output of HDC. Instead of applying convolutions to the high-resolution input data, we first apply a periodic down-shuffling operator to the input data to get $(C \times (n_x \times n_y \times n_z))$ channels of low-resolution data and then apply convolutions with a kernel size of $3 \times 3 \times 3$ to get the k feature maps of size $(d \times h \times w)$. Mathematically, this can be described as:

$$HDC(I^{LR}; W_1, b_1) = \phi\big(W_1 * PDS(I^{HR}) + b_1\big) \tag{3}$$

where ϕ is an non-linear activation function that is applied element-wise; W_1, b_1 are trainable weights and bias, respectively; PDS is a periodic down-shuffling operation which aims to rearrange a HR tensor (T_{HR}) in the shape of $(n_x \times d) \times (n_y \times h) \times (n_z \times w) \times C$ to a LR tensor (T_{LR}) in the shape of $(d \times h \times w) \times (C \times (n_x \times n_y \times n_z))$. And the operation $T_{LR} = PDS(T_{HR})$ can be mathematically described as below:

$$\begin{aligned} T_{LR}(x', y', z', c') = T_{HR}\big(& x' \cdot n_x + \lfloor mod(c', n_x \cdot C)/C \rfloor, \\ & y' \cdot n_y + \lfloor mod(c', n_x n_y \cdot C)/(n_x \cdot C) \rfloor, \\ & z' \cdot n_z + \lfloor c'/(n_x n_y \cdot C) \rfloor, \\ & mod(c', C) \big) \end{aligned} \tag{4}$$

where x', y', z', c' are the coordinates of the voxels in the LR space, and $x' \in [0, d-1], y' \in [0, h-1], z' \in [0, w-1], c' \in [0, C \cdot n_x \cdot n_y \cdot n_z - 1]$.

3.3. HDC and DUC augmented FCNs for volumetric image segmentation

Both DUC and HDC are network agnostic and can be combined with existing FCNs such as the 3D U-net (Çiçek et al., 2016), the 3D V-net (Milletari et al., 2016), and the HighRes3dNet (Li et al., 2017) for semantic segmentation as shown in Fig. 4, as long as the dimensions of the output from HDC satisfy the input requirement of the deep neural networks. Fig. 5 shows an example of combining HDC and DUC with the 3D U-net for segmenting 3D hip MR images of limited field of view. The advantage of such a pipeline is apparent. When a HDC with down-shuffling factors of (n_x, n_y, n_z) is applied to the input data, both the computational and the storage cost for the underlying 3D U-net will be reduced by a factor of $(n_x \times n_y \times n_z)$, allowing one to use large patch as the input. The full resolution segmentation map is then obtained at the final output by applying a DUC with up-shuffling factors of (n_x, n_y, n_z). To

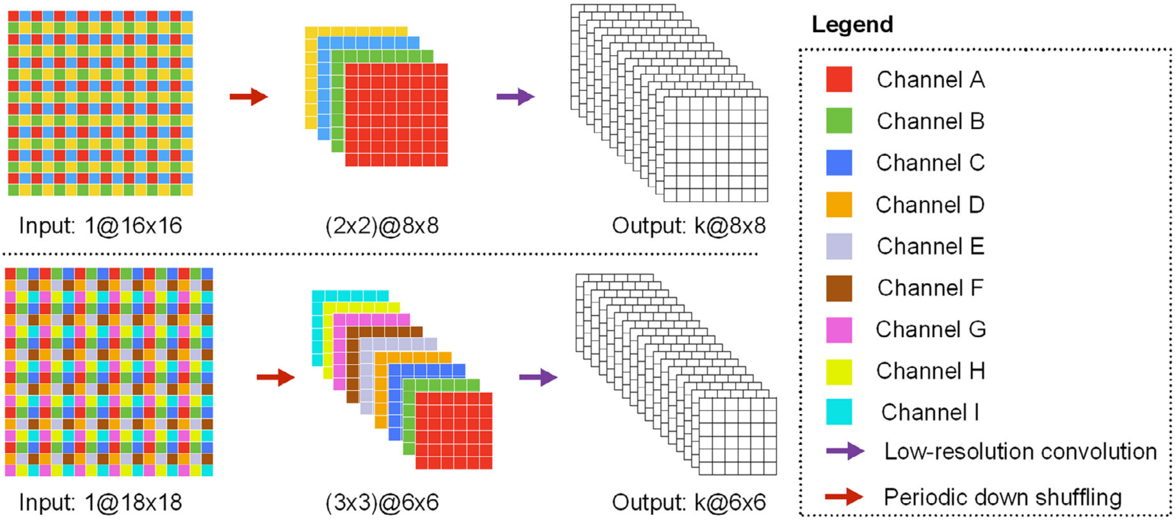


Fig. 3. A schematic view of the HDC in 2D which consists of a periodic down-shuffling operation and low-resolution convolutions. Top row: a 2D case when the downsampling factor along each dimension is 2; bottom row: a 2D case when the downsampling factor along each dimension is 3.

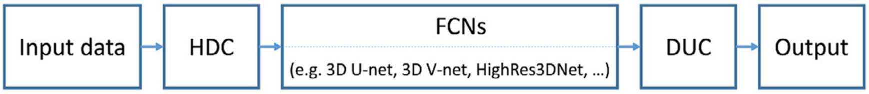


Fig. 4. A schematic view of how to augment existing FCNs with HDC and DUC for semantic segmentation.

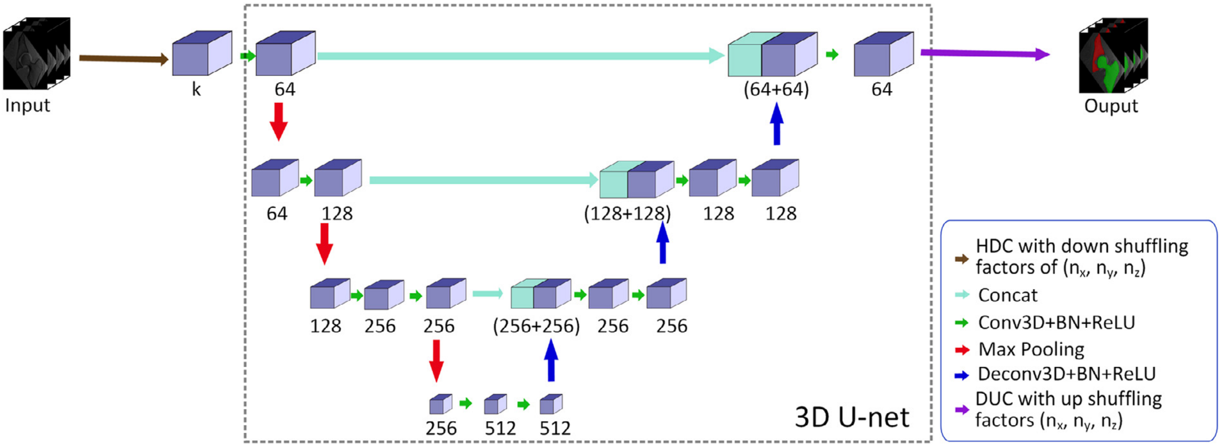


Fig. 5. A schematic view of how to augment the 3D U-net with HDC and DUC for segmenting 3D hip MR images of limited field of view. The numbers below each block represent the number of feature maps.

differentiate from the original 3D U-net, we call the 3D U-net augmented with HDC and DUC as 3D large patch U-net (3D LP-U-net). Similarly we can derive 3D LP-V-net and LP-HighRes3DNet respectively by augmenting the original 3D V-net (Milletari et al., 2016) and the HighRes3DNet (Li et al., 2017) with HDC and DUC. In this study, we take the original 3D U-net, 3D V-net and HighRes3DNet as the baseline networks to evaluate the performance of the associated networks augmented with HDC and DUC. For all the studies, a combination of cross entropy loss with Dice loss as introduced in Milletari et al. (2016) is used.

3.4. Implementation details

All methods reported in this study were implemented in Python using TensorFlow framework and were trained and tested on a desktop with a 3.6 GHz Intel(R) i7 CPU and a NVIDIA GTX 1080 Ti graphics card with 11 GB GPU memory. We empirically fixed the number of output feature maps from the HDC as $k = 64$. All the weights were initialized from Gaussian distribution ($\mu = 0$, $\sigma = 0.01$) and were then updated by the stochastic gradient descent (SGD) algorithm (momentum = 0.9, weight decay = 0.005). For the baseline networks, the initial learning rate was chosen to be 0.001 and was halved every 3000 iterations. For the networks augmented with HDC and DUC, empirically we found that when the shuffling factors become large, high initial learning rates could be used for a fast convergence. During a training stage, we randomly cropped sub-volume patches of a fixed size from training samples. Each sampled patch was normalized as zero mean and unit variance before fed into network. During a testing stage, given a test volumetric image, we extracted overlapped sub-volume patches with the same size as we used in the associate training stage and fed them to the trained network to get prediction probability maps. For the overlapped voxels, the final probability maps would be the average of the overlapped patches, which were then used to derive the final segmentation results.

4. Experiments and results

In this section, we present experimental results of the proposed pipeline for volumetric medical image segmentation. Four datasets, i.e., an in-house dataset consisting of 25 T1 hip MR images with limited field of view, an open dataset consisting of 24 multi-modality MR images obtained from the MICCAI 2018 challenge on IVD localization and segmentation (Zeng et al., 2018), another open dataset consisting of 25 T2 MR spine images obtained from the MICCAI 2015 IVD Localization and Segmentation challenge (Zheng et al., 2017), and a publicly available dataset from NIH containing 82 abdominal contrast enhanced 3D CT scans (Roth et al., 2015), were used in our study. More specifically, first, we conducted ablation studies on the in-house hip dataset and on the 3D multi-modality MR images to evaluate the influence of different factors on the performance of the proposed pipeline. Based on the findings from the ablation study, we chose the 3D LP-U-net for our remaining studies. Second, we conducted comprehensive studies on all four datasets to compare the performance of the 3D LP-U-Net with other state-of-the-art methods. Below we will start with the description of evaluation metrics, followed by a detailed presentation of the experimental setup and results for each study.

4.1. Evaluation metrics

Assuming the automatically segmented set of voxels as *AS* and the manually defined ground truth as *GT*, we compute following metrics.

4.1.1. Dice overlap coefficients (DOC)

DOC quantifies the match of two sets by normalizing the size of their intersection over the average of their sizes and is defined as follows:

$$DOC = \frac{2|AS \cap GT|}{|AS| + |GT|} \tag{5}$$

where the operator $|\cdot|$ returns the number of voxels contained in a region.

4.1.2. Distance-based metrics

Before we present the definitions of different distance-based metrics, we first define a distance measure for a voxel x from a set of voxels A as:

$$d(x, A) = \min_{y \in A} d(x, y) \tag{6}$$

where $d(x, y)$ is the Euclidean distance of the voxels incorporating the real spatial resolution of the volume data.

We further define the directed Hausdorff measure from a point set A to a point set B as the maximum distance, for all points in A, to the closest point in B. Mathematically this is given as:

$$\vec{d}_H(A, B) = \max_{x \in A}(\min_{y \in B}(d(x, y))) \tag{7}$$

With these definitions, we can define two distance-based metrics used in our studies to quantify the dissimilarity of the automatic segmentation from the ground truth:

- *Average Surface Distance (ASD)* - It is defined as the average of all the distances from points on the boundary of AS (we denote them as B_{AS}) to the boundary of GT (B_{GT}):

$$ASD = \frac{1}{|B_{AS}|} \sum_{x \in B_{AS}} d(x, B_{GT}) \tag{8}$$

- *Hausdorff Distance (HD)* - it is defined as the maximum distance between two objects:

$$HD = \max\left\{\vec{d}_H(A, B), \vec{d}_H(B, A)\right\} \tag{9}$$

4.2. Ablation study on hip MR images with limited field of view

4.2.1. Data and augmentation

In this study, we used 25 3D T1 MR images, acquired from patients with hip pain due to FAI or hip dysplasia. Those images were acquired by using a dual-flip angle 3D gradient-echo technique (TR/TE = 15/3.3 ms; flip angles: 4° and 24°; slice thickness: 1.0 mm; field of view: 160 × 160 mm^2). All images were resampled to have a uniform size of 480 × 480 × 160 voxels with an average voxel spacing of 0.374 mm × 0.363 mm × 1.078 mm. Slice by slice manual segmentation was used to create the reference ground truth segmentation. We randomly distributed the 25 datasets into two groups with one group containing 20 datasets as the training date and the remaining 5 datasets as the testing data. During training, data augmentation was used to enlarge the training samples. Specifically, we applied a smooth deformation field on both image data and ground truth labels. For this, we sampled random vectors from a normal distribution with a standard deviation of 15 voxels in a 2 × 2 × 2 grid of control points and then applied a B-spline interpolation. For each training sample, we generated four additional augmented samples. All the networks used in this study were trained on the augmented training data for 10,000 iterations.

4.2.2. Ablation study

We first investigated the influence of patch sizes on the performance of the original 3D U-net. The results are presented in Table 1. It was observed that better performance was obtained when larger patch size was used. Due to the GPU memory constraint, (200 × 200 × 40) was the maximum size that we could use.

We then examined the effect of different shuffling factors on the performance of the 3D LP-U-net when a fixed patch size of (400 × 400 × 80) was used. The results are reported in Table 2. From this table, we can see that (1) the higher the shuffling factors, empirically the bigger the initial learning rate that we could use; (2) the higher the shuffling factor, in general the less accurate the results but the best results were achieved when the shuffling factor was (4, 4, 2); (3) even with a shuffling factor as high as (25, 25, 2), we still get sub-millimeter average surface distance for both the acetabulum and the proximal femur; and (4) in comparison with the results reported in Table 1, 3D LP-U-net with a shuffling factor smaller than (16, 16, 2) achieved better results than the original 3D U-net with the largest patch size.

Fig. 6 visually compares the segmentation results obtained by the 3D LP-U-net with a fixed patch size of (400 × 400 × 80) but different shuffling factors and the 3D U-net with different patch sizes. In this figure, we show both the overall segmentation and the probability of each structure as well as the results around the hip joint. From this figure, we observe that (1) less false positive segmentation was observed when comparing the results obtained by the 3D LP-U-net with those by the 3D U-net; and (2) for the 3D LP-U-net, the larger the shuffling factors, the higher the uncertainty around the boundary of each structure.

Table 3 shows the required training time when all the models are trained for 10,000 iterations. When the down-shuffling factor was chosen to be (2, 2, 2), the training time of the 3D LP-U-net was slightly longer than the baseline model due to the additional computations required by HDC and DUC. Further increasing the down-shuffling factor led to a significant reduction of the required training time even when the size of the input patches was chosen to be (400 × 400 × 80).

Finally, we checked the influence of different architectures of the underlying FCNs on the performance of the proposed pipeline. Table 4 shows the results when the original 3D V-net and the original HighRes3DNet were used with different patch sizes. Please note that caused by high spatial resolution, HighRes3DNet (Li et al., 2017) requires the largest GPU memory to

Table 1
Results of investigation of different patch sizes on the performance of the original 3D U-net. Ace: the acetabulum; Femur: the proximal femur.

Patch size	(50 × 50 × 40)		(96 × 96 × 96)		(200 × 200 × 40)	
Anatomy	Ace	Femur	Ace	Femur	Ace	Femur
DOC (%)	37.45 ± 5.73	30.62 ± 3.55	91.30 ± 5.84	95.89 ± 1.21	92.06 ± 5.37	96.84 ± 0.90
ASD (mm)	28.15 ± 5.04	29.27 ± 4.90	5.11 ± 7.57	1.41 ± 1.11	0.88 ± 0.76	0.63 ± 0.31
HD (mm)	111.10 ± 10.41	95.1 ± 8.98	33.92 ± 29.0	29.78 ± 20.35	13.71 ± 5.07	10.85 ± 6.09

Table 2
Results when different shuffling factors were used for the 3D LP-U-net. The size of the input patch is fixed to (400 × 400 × 80).

Shuffling factors	(2, 2, 2)		(4, 4, 2)		(8, 8, 2)		(16, 16, 2)		(25, 25, 2)	
Initial learning rate	1.0E−03		2.0E−03		3.0E−03		5.0E−03		2.0E−02	
Anatomy	Ace	Femur	Ace	Femur	Ace	Femur	Ace	Femur	Ace	Femur
DOC (%)	96.77 ± 1.27	97.41 ± 1.34	96.77 ± 1.26	97.95 ± 0.63	96.30 ± 0.97	97.25 ± 0.59	94.24 ± 1.73	95.75 ± 1.02	91.57 ± 2.03	93.82 ± 1.52
ASD (mm)	0.39 ± 0.28	0.43 ± 0.28	0.39 ± 0.28	0.33 ± 0.16	0.37 ± 0.11	0.41 ± 0.09	0.62 ± 0.23	0.64 ± 0.16	0.86 ± 0.25	0.96 ± 0.25
HD (mm)	7.73 ± 3.81	6.23 ± 2.32	8.57 ± 5.68	3.59 ± 3.95	6.97 ± 3.15	5.15 ± 1.43	10.69 ± 7.44	6.39 ± 1.50	12.64 ± 2.87	8.18 ± 0.66

Table 3
Training time for different models, where "3D_U-net_P_X_Y_Z" means the results obtained from the 3D U-net with a patch size of ($X \times Y \times Z$) and "3D-LP-U-net_S_x_y_z" means the results obtained from the 3D LP-U-net with a shuffling factor of (x, y, z) and a fixed patch size of (400 × 400 × 80).

Model	3D_U-net_P_200_200_40	3D-LP-U-net_S_2_2_2	3D-LP-U-net_S_4_4_2	3D-LP-U-net_S_8_8_2	3D-LP-U-net_S_16_16_2	3D-LP-U-net_S_25_25_2
Time (min)	294	443.5	186.3	131	121	117

Table 4
Results when the original 3D V-net and the original HighRes3DNet were used with different patch sizes.

Architectures (Used patch size)	3D V-net (96 × 96 × 96)		3D V-net (200 × 200 × 40)		HighRes3DNet (100 × 100 × 80)		HighRes3DNet (200 × 200 × 20)	
Anatomy	Ace	Femur	Ace	Femur	Ace	Femur	Ace	Femur
DOC (%)	88.71 ± 6.21	92.27 ± 3.68	92.78 ± 0.50	96.67 ± 0.85	90.66 ± 6.68	86.18 ± 5.08	93.04 ± 4.31	93.58 ± 2.46
ASD (mm)	1.77 ± 1.29	1.75 ± 0.77	0.97 ± 0.97	0.59 ± 0.23	1.80 ± 2.36	2.37 ± 0.59	1.77 ± 2.34	1.50 ± 0.82
HD (mm)	15.77 ± 6.16	14.0 ± 3.70	12.15 ± 6.81	9.92 ± 4.26	15.94 ± 11.70	17.27 ± 4.93	22.79 ± 13.92	16.67 ± 6.53

Table 5
Results when different shuffling factors were used for the 3D LP-V-net and the LP-HighRes3DNet. The size of the input patch was chosen to be (400 × 400 × 80).

Results of the 3D LP-V-Net with different shuffling factors										
Shuffling factors	(2, 2, 2)		(4, 4, 2)		(8, 8, 2)		(16, 16, 2)		(25, 25, 2)	
Anatomy	Ace	Femur	Ace	Femur	Ace	Femur	Ace	Femur	Ace	Femur
DOC (%)	95.58 ± 1.43	97.11 ± 0.63	94.98 ± 1.81	96.62 ± 0.38	93.21 ± 1.74	94.55 ± 0.88	91.66 ± 2.06	93.45 ± 1.40	90.05 ± 2.83	92.69 ± 1.47
ASD (mm)	0.63 ± 0.58	0.49 ± 0.15	0.56 ± 0.34	0.51 ± 0.06	0.69 ± 0.25	0.82 ± 0.14	0.85 ± 0.27	1.0 ± 0.22	1.05 ± 0.36	1.10 ± 0.21
HD (mm)	11.21 ± 9.97	7.24 ± 2.01	8.51 ± 4.33	5.97 ± 1.74	10.77 ± 6.88	6.76 ± 0.97	11.22 ± 4.76	7.85 ± 1.26	11.73 ± 6.62	7.48 ± 1.29
Results of the LP-HighRes3DNet with different shuffling factors										
Shuffling factors	(4, 4, 1)		(4, 4, 2)		(8, 8, 2)		(16, 16, 2)		(25, 25, 2)	
Anatomy	Ace	Femur	Ace	Femur	Ace	Femur	Ace	Femur	Ace	Femur
DOC (%)	95.99 ± 1.18	97.38 ± 0.52	95.35 ± 1.30	96.62 ± 1.08	93.72 ± 1.69	95.52 ± 0.94	91.15 ± 2.09	92.41 ± 1.69	88.21 ± 2.48	90.0 ± 2.24
ASD (mm)	0.43 ± 0.18	0.44 ± 0.12	0.53 ± 0.29	0.60 ± 0.33	0.66 ± 0.21	0.75 ± 0.19	0.90 ± 0.25	1.22 ± 0.32	1.27 ± 0.51	1.55 ± 0.36
HD (mm)	8.21 ± 4.05	6.76 ± 2.68-	8.48 ± 4.33	7.85 ± 4.11	11.31 ± 4.47	8.38 ± 3.75	12.95 ± 7.14	9.53 ± 2.25	16.23 ± 5.12	9.62 ± 2.30

store intermediate results among all three architectures, though it has the smallest number of training parameters. Thus, the maximally allowed size of the input patch for the HighRes3DNet was (200 × 200 × 20). In comparison, the results of the 3D LP-V-net and the LP-HighRes3DNet with different shuffling factors are reported in Table 5. From the results reported in Tables 2, 4, and 5, we can see that (1) results achieved by the 3D LP-V-net and the LP-HighRes3DNet are better than those achieved by the associate baseline networks when the chosen shuffling factor is not too big. For example, even with a shuffling factor of (8, 8, 2), the performance of the LP-HighRes3DNet is much better than that achieved by the original HighRes3DNet with the largest patch size allowed; (2) the bigger the shuffling factor, the less accurate the results; and (3) when the same shuffling factor was used, the 3D LP-U-net achieved the best results.

4.3. Ablation study on multi-modality IVD MR images

4.3.1. Data description

We conducted experiments on the dataset obtained from the MICCAI 2018 challenge on automatic IVD localization and segmentation from 3D multi-modality MR (M3) images (Zeng et al., 2018). There are 24 3D multi-modality MRI datasets of at least 7 IVDs of the lower spine, collected from 12 subjects in two different stages in a study investigating the effect of prolonged bed rest (spaceflight simulation) on the lumbar IVDs. Each subject at each stage was scanned with a 1.5-Tesla MRI scanner of Siemens using Dixon protocol: slice thickness = 2.0 mm, pixel spacing = 1.25 mm, repetition time (TR) = 10.6 ms, echo time (TE) = 4.76 ms. Thus, each 3D multi-modality MRI dataset contains four aligned high-resolution 3D volumes: in-phase, opposed-phase, fat and water images. In total we have 96 high resolution 3D MRI volume data. All images were resampled to 2 mm × 1.25 mm × 1.25 mm and their dimension size were about $36 \times 256 \times 256$ voxels. For each IVD, reference manual segmentation was provided in the form of binary mask. The multi-modality challenge data were divided into two subsets: 64 volume images from 8 subjects were used as training data and the remaining volume images from 4 other subjects were used as testing data. During training, the input to our network was a concatenation of multi-modality sub-volumes as shown in Fig. 8. Each sub-volume was in the size of $32 \times 240 \times 240$ voxels. The shuffling factor was chosen to be (1,3,3). Each sub-volume was normalized as zero mean and unit variance before fed into the network. All parameters were trained from scratch and initialized as from Gaussian distribution ($\mu = 0, \sigma = 0.01$). All parameters were updated by the stochastic gradient descent (SGD) algorithm (momentum = 0.9, weight decay = 0.005). The neural network was trained in total 10,000 iterations. The initial learning rate was 1×10^{-3} and halved by every 3,000 iterations.

4.3.2. Ablation study

In this study we investigated the effect of different components on the performance of the proposed approach. More specifically, we compared the performance of the proposed approach to that of two different variants of the present approach. The first variant is to replace HDC in Fig. 8 by strided convolutions with a stride of size (1,3,3) and a kernel of size (3,3,3) while the second variant is to not only replace HDC by strided convolutions but also replace DUC by transposed convolutions (also called fractionally strided convolutions or deconvolutions). The results of the ablation study are shown in Table 6 where three different metrics as introduced in Zeng et al. (2018) are used, i.e., Mean Dice Similarity Coefficient (MDSC), Mean Average Surface Distance (MASD) and Mean Localization Distance (MLD). From the results reported in Table 6, we can see that each component of the proposed approach helped to improve the segmentation results. More specifically, when HDC and DUC were replaced by strided convolutions and transposed convolutions respectively, worst results were observed. Replacement of transposed convolutions by DUC helped to improve the

Table 6
Quantitative comparison between the proposed approach and two other variants on the MICCAI 2018 multi-modality IVD localization and segmentation challenge data. The best result in each column is highlighted with bold font.

Methods	MDSC(%)	MASD (mm)	MLD (mm)
The present approach	**90.88**	**0.57**	**0.78**
The first variant	85.98	1.05	0.93
The second variant	81.99	1.35	0.92

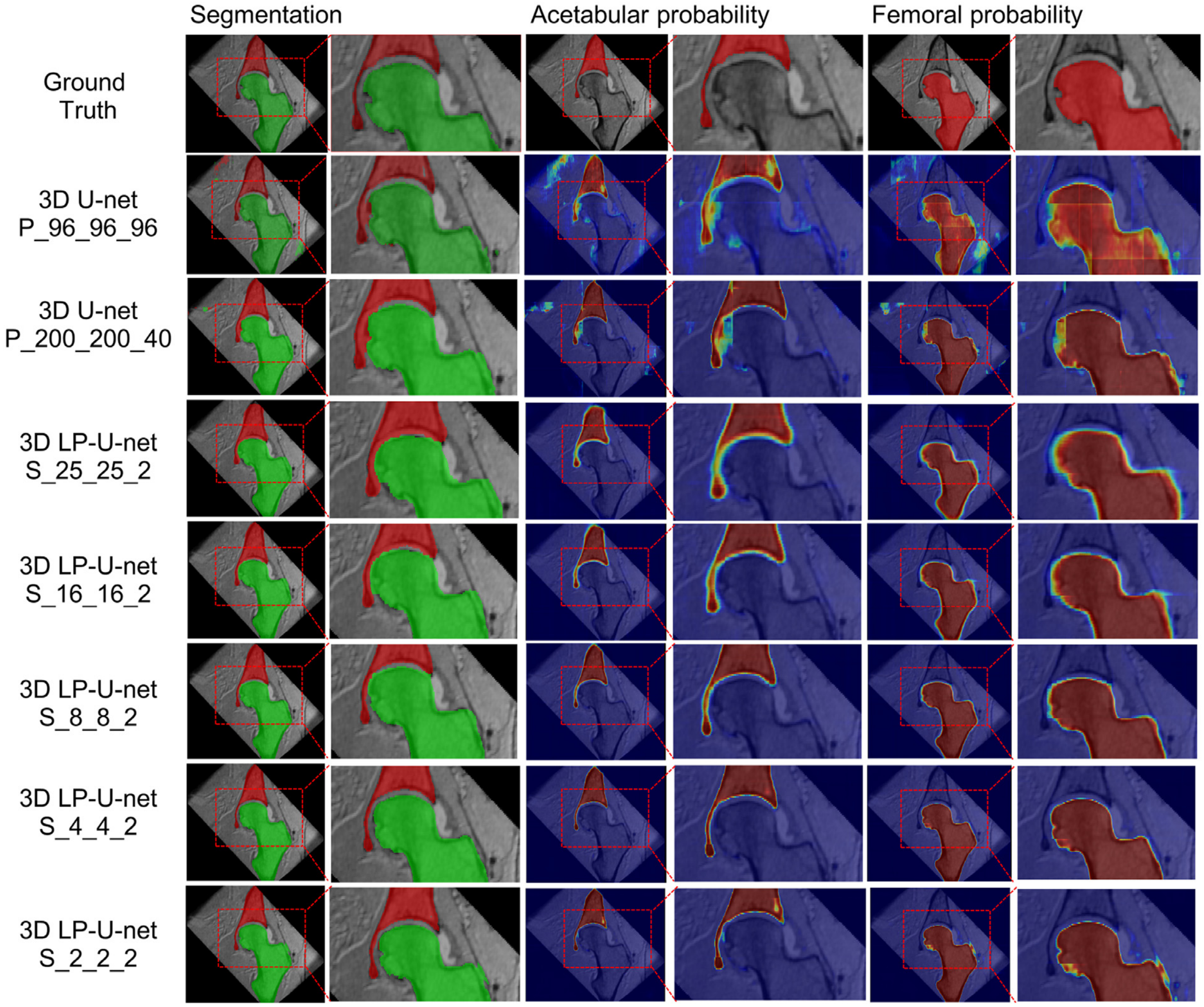
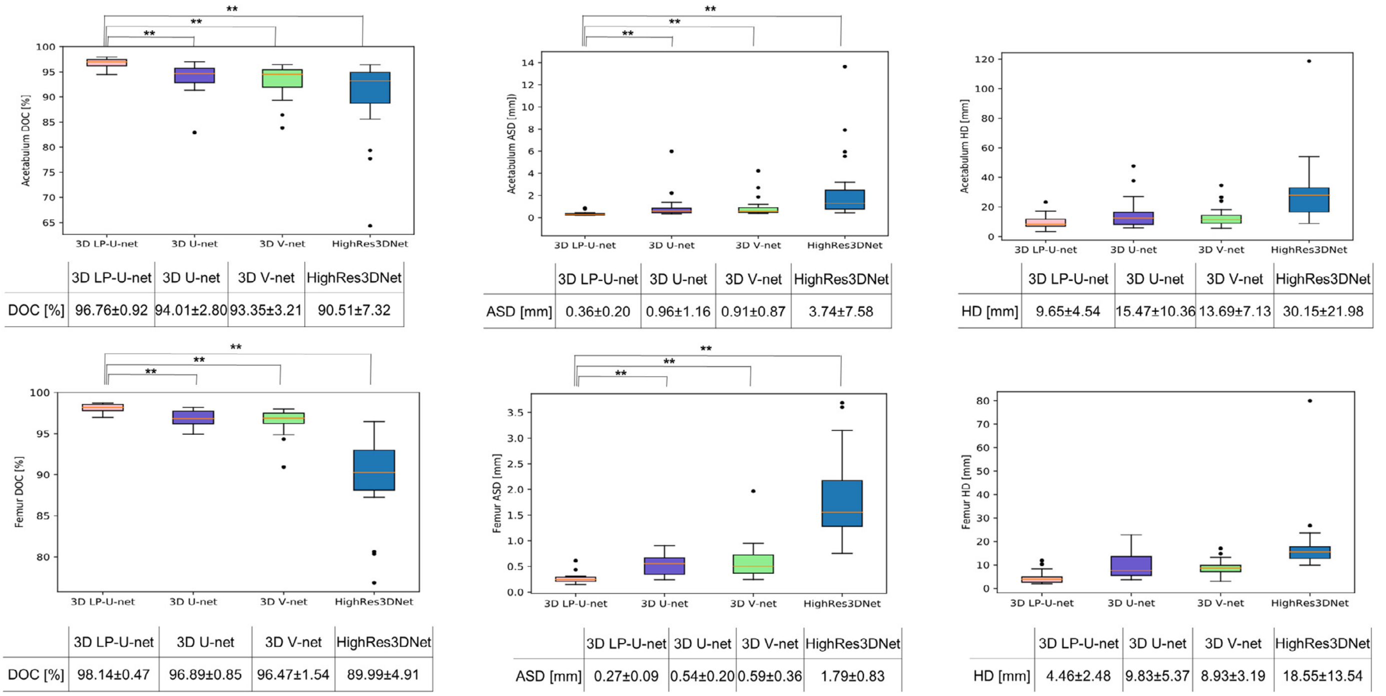


Fig. 6. Qualitative comparison of the segmentation results of 3D LP-U-Net with different shuffling factors and the 3D U-net with different patch sizes, where "3D_U-net P_X_Y_Z" means the results obtained from the 3D U-net with a patch size of ($X \times Y \times Z$) and "3D-LP-U-net S_x_y_z" means the results obtained from the 3D LP-U-net with a shuffling factor of (x, y, z) and a fixed patch size of ($400 \times 400 \times 80$).



	3D LP-U-net	3D U-net	3D V-net	HighRes3DNet
DOC [%]	96.76±0.92	94.01±2.80	93.35±3.21	90.51±7.32

	3D LP-U-net	3D U-net	3D V-net	HighRes3DNet
ASD [mm]	0.36±0.20	0.96±1.16	0.91±0.87	3.74±7.58

	3D LP-U-net	3D U-net	3D V-net	HighRes3DNet
HD [mm]	9.65±4.54	15.47±10.36	13.69±7.13	30.15±21.98

	3D LP-U-net	3D U-net	3D V-net	HighRes3DNet
DOC [%]	98.14±0.47	96.89±0.85	96.47±1.54	89.99±4.91

	3D LP-U-net	3D U-net	3D V-net	HighRes3DNet
ASD [mm]	0.27±0.09	0.54±0.20	0.59±0.36	1.79±0.83

	3D LP-U-net	3D U-net	3D V-net	HighRes3DNet
HD [mm]	4.46±2.48	9.83±5.37	8.93±3.19	18.55±13.54

Fig. 7. Boxplots showing segmentation accuracy of the proposed 3D LP-U-net and three state-of-the-art methods when evaluated on 25 T1 hip MR images with limited field of view via a standard 5-fold cross validation study. **indicates significant accuracy improvement with significance level of 0.01.

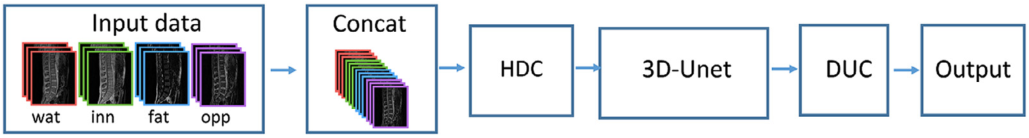


Fig. 8. A schematic view of how to apply the present approach to segment multi-modality MR images.

MDSC from 81.99% to 85.98%. Further replacing strided convolutions by HDC boosted the performance of the present approach to a MDSC of 90.88%.

4.4. *Main study*

Based on the findings from the ablation studies, we chose the 3D LP-U-net for our remaining studies. We compared the performance of the 3D LP-U-net with other state-of-the-art methods when evaluated on four different datasets as described below.

4.4.1. *Validation on the 25 hip MR images with limited field of view*

We conducted a standard 5-fold cross validation study on the 25 T1 hip MR images with limited field of view such that each image was used one time as the testing data. We used the same data augmentation strategy and the same training strategy as we used in the ablation study. In this study, for the 3D LP-U-net, we chose a fixed patch size of $(400 \times 400 \times 80)$ and a fixed shuffling factor of (4, 4, 2). We compared the performance of the 3D LP-U-net with state-of-the-art methods such as the 3D U-net (Çiçek et al., 2016), the 3D V-net (Milletari et al., 2016), and the HighRes3dNet (Li et al., 2017). For the 3D U-net and the 3D V-net, the chosen patch size is $(200 \times 200 \times 40)$ while for the HighRes3DNet, the patch size was chosen to be $(200 \times 200 \times 20)$. The top row of Fig. 7 shows boxplots for overall DOC, ASD and HD of all four methods for segmenting the acetabulum. For all metrics, the 3D LP-U-net achieved the best results. More specifically, an average DOC of 96.76 ± 0.92%, 94.01 ± 2.80%, 93.35 ± 3.21% and 90.51 ± 7.32% was found for the 3D LP-U-net, the 3D U-net, the 3D V-net and the HighRes3DNet, respectively. The 3D LP-U-net showed significantly higher accuracy than all other three methods ($p < 0.01$). For ASD, the same significance was also observed. The bottom row of Fig. 7 shows the comparison results for the proximal femur. An average DOC of 98.14 ± 0.47%, 96.89 ± 0.85%, 96.47 ± 1.54% and 89.99 ± 4.91% was found for the 3D LP-U-net, the 3D U-net, the 3D V-net and the HighRes3DNet, respectively. The 3D LP-U-net showed significantly higher accuracy than all other three methods ($p < 0.01$) when segmenting the proximal femur.

4.4.2. *Validation on the MICCAI 2015 IVD localization and segmentation challenge data*

We conducted experiments on the MICCAI 2015 IVD localization and segmentation challenge data (Zheng et al., 2017), which contains 25 3D T2-weighted MR images. The resolution of all images were resampled to 2 mm × 1.25 mm × 1.25 mm. The size of the images is between $39 \times 305 \times 305$ and $48 \times 304 \times 304$ voxels. Each image contains at least 7 IVDs T11-S1. These 25 MR images were divided into three non-overlapped subsets as training data (15 3D MR images), Test1 data (5 3D MR images) and Test2 data (the remaining 5 3D MR images). All methods were trained on the training data and then separately evaluated on the two testing datasets. Manual segmentation was used as the reference for all evaluations.

`Training and testing.` We compared the performance of the 3D LP-U-net with the top-5 state-of-the-art methods described in Zheng et al. (2017). In the training phase, we chose a fixed patch size of $32 \times 288 \times 288$ voxels and a fixed shuffling factor of (1, 2, 2) for the 3D LP-U-net in order to incorporate as large as possible context information. Each patch was normalized as zero mean

Table 7
Accuracy (DOC, %) comparison between the 3D LP-U-net and the state-of-the-art methods as described in Zheng et al. (2017) on the MICCAI 2015 IVD localization and segmentation challenge dataset.

Methods	Test1 results (%)	Test2 results (%)	Overall (%)
3D LP-U-net	**92.4** ± **1.5**	**92.1** ± **1.7**	**92.2** ± **1.7**
UNILJU	91.5 ± 2.3	92.0 ± 1.9	91.8 ± 2.1
UNIBE	89.8 ± 2.9	91.2 ± 2.0	90.5 ± 2.6
UNIEXE	89.8 ± 3.6	90.2 ± 2.6	90.0 ± 3.1
Sectra	90.0 ± 2.6	90.0 ± 2.2	90.0 ± 2.4
UNICHK	88.4 ± 3.7	88.9 ± 3.4	88.6 ± 3.5

Table 8
Quantitative comparison between the proposed approach and the top-5 methods as described in Zeng et al. (2018) on the MICCAI 2018 multi-modality IVD localization and segmentation challenge data. The best result in each column is highlighted with bold font.

Methods	MDSC(%)	MASD (mm)	MLD (mm)
The present approach	**90.88**	**0.57**	0.78
Changliu	90.64	0.60	**0.77**
Gaoyunhe_cuhk	90.58	0.61	0.78
Ucsf_Claudia	89.71	0.74	0.86
Livia	89.67	0.65	0.96
Wanghuan	88.77	0.82	0.92

and unit variance before fed into the network. The 3D LP-U-net was trained from scratch. All parameters were initialized from a Gaussian distribution ($\mu = 0,\ \sigma = 0.01$) and then updated by the stochastic gradient descent (SGD) algorithm (momentum = 0.9, weight decay = 0.005). We trained the 3D LP-U-net for 10,000 iterations. The initial learning rate was set as 1×10^{-3} and halved by every 3,000 iterations.

Table 7 shows the accuracy comparison between 3D LP-U-net and the state-of-the-art methods as described in Zheng et al. (2017). For both testing datasets, the 3D LP-U-net achieved consistently better results than other state-of-the-art methods. The lower standard deviation of DOC shows that the 3D LP-U-net is the most stable and robust across all different IVD cases. The results that we obtained on the MICCAI 2015 IVD localization and segmentation challenge dataset prove the effectiveness of our approach.

4.4.3. Validation on the MICCAI 2018 multi-modality IVD localization and segmentation challenge data

We used the same setup as we used in the ablation study on multi-modality IVD MR images. We compared the results achieved by our approach with the results achieved by the teams participating the MICCAI 2018 multi-modality IVD localization and segmentation challenge. Table 8 shows the accuracy comparison between our approach and the top-5 methods as described in Zeng et al. (2018). Our approach achieved comparable or better results than other state-of-the-art methods. Please note that the best results during the challenge were achieved by team ChangLiu (Liu and Zhao, 2018). Their method was based on an ensemble of 2.5D multi-scale fully convolutional networks with additional post-processing, which involved a complicated non-rigid registration between the output from the ensemble and a best-fit template picked from the training data. In contrast, our method directly outputs the segmentation results and no non-rigid registration is needed, which is a clear advantage.

4.4.4. Validation on NIH pancreas CT dataset

We verified our approach on the NIH pancreas CT dataset (Roth et al., 2015) as well, which contains 82 contrast-enhanced abdominal CT volumes provided by an experienced radiologist. The size of CT volumes is between 181 × 512 × 512 and 466 × 512 × 512 voxels and their spatial resolutions are $w \times h \times d$ where $d = 1.0$ mm and $w = h$ ranges from 0.5 mm to 1.0 mm. For the data pre-processing, we simply truncated the raw intensity values to be in $[-300, 300]$ and added a random noise in the range of $[-3, 3]$. Following the training protocol (Roth et al., 2015), we conducted a 4-fold cross validation in a random split from 82 patients for training and testing folds, where each testing fold had 22, 20, 20, and 20 cases, respectively.

Table 9
Accuracy (DOC, %) comparison between 3D LP-U-net and the state-of-the-arts on the NIH pancreas segmentation dataset. The best result in each column is highlighted with bold font.

Approach	Average	Max	Min
Roth et al. (2015)	71.42 ± 10.11	86.29	23.99
Zhou et al. (2017)	82.37 ± 5.68	90.85	62.43
Cai et al. (2017)	82.4 ± 6.7	90.1	60.0
Roth et al. (2018a)	81.27 ± 6.27	88.96	50.69
Yu et al. (2018)	**84.50** ± **4.97**	**91.02**	62.81
Our approach	83.0 ± 5.85	90.31	**68.39**

`Training and testing.` We implemented a two-stage pipeline consisting of a coarse stage and a fine stage. In the coarse stage, we first trained a deep segmentation network to locate the rough region of the pancreas from a whole CT volume. The goal of the fine stage is then to train another deep segmentation network to further refine the results. In both stages, we used the 3D LP-U-net as the segmentation networks. During the training phase of the coarse stage, we chose a fixed patch size of 480 × 480 × 64 voxels and a fixed shuffling factor of (4, 4 1). Each patch was normalized as zero mean and unit variance before fed into the network. The 3D LP-U-net in the coarse stage was trained from scratch. All parameters were initialized from a Gaussian distribution ($\mu = 0$, $\sigma = 0.01$) and then updated by the stochastic gradient descent (SGD) algorithm (momentum = 0.9, weight decay = 0.005). We trained the 3D LP-U-net for 20,000 iterations. The initial learning rate was set as 2×10^{-3} and halved by every 3,000 iterations. During testing phase, we adopted a sliding and stitching strategy. More specifically, we extracted overlapped sub-volume patches and fed them into the trained network to get prediction probability maps. For the overlapped voxels, the final probability maps were calculated as the average of the probability maps of the overlapped patches, which were then used to derive the final segmentation results by setting a threshold of 0.3. During the training phase of the fine stage, all training images were cropped by the bounding box calculated from the associated ground truth segmentation plus a padding of 20 voxels along all three spatial axes. Then all cropped images were resampled to a fixed size of 196 × 128 × 128 voxels. We further conducted following data augmentations: rotate each volume randomly around the Z axis in the range of -15^o and 15^o and scale each volume randomly in the range of 0.94 and 1.06 times. For the 3D LP-U-net in the fine stage, we chose a fixed patch size of 176 × 112 × 96 voxels and a fixed shuffling factor of (2, 2, 1). Each patch was normalized as zero mean and unit variance before fed into the network. The same parameter initialization strategy as we used in the coarse stage was also used here. The 3D LP-U-net in the fine stage was then optimized by SGD algorithm (momentum = 0.9, weight decay = 0.005). We trained the 3D LP-U-net in the fine stage for 60,000 iterations. The initial learning rate was set to 1×10^{-3} and halved by every 6,000 iterations. During testing phase, starting from the rough segmentation results obtained in the coarse stage, we first computed a bounding box and then added a padding of 20 voxels along each direction in order to crop a sub-volume containing the pancreas region from the whole CT volume. The cropped sub-volume was then resampled to a fixed size of 196 × 128 × 128. A similar sliding and stitching strategy as

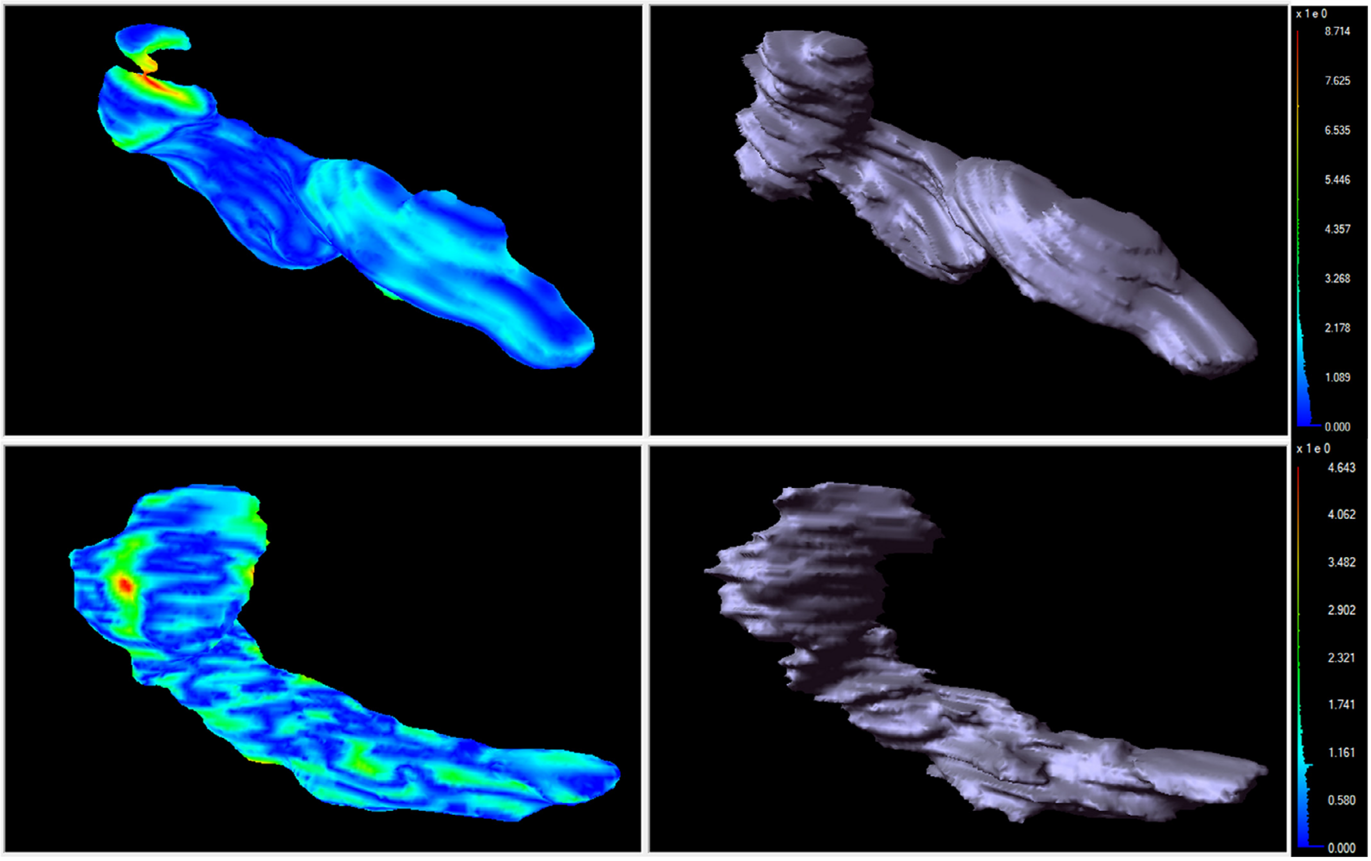


Fig. 9. Two pancreas segmentation examples. Top row: a bad case where the structure is under-segmented; bottom row: a good case. In both rows, the left column shows the automatic segmentation results with color-coded segmentation errors, the middle column shows the ground truth segmentation and the right column shows the color map.

we used in the testing phase of the coarse stage was also applied here to get the final segmentation results by setting a threshold of 0.5. The obtained segmentation results were then resampled back to the original space.

Table 9 shows the accuracy comparison between our approach and previous state-of-the-art methods when evaluated on the NIH pancreas CT dataset. Our approach achieved a comparable average DOC with the state-of-the-art methods. Although the average DOC and the maximum DOC achieved by our approach are slightly worse than those achieved by Yu et al. (2018), our approach achieved much better result in the worst case (68.39% by our approach vs. 62.81% by previous state-of-the-art), which guaranteed the reliability of our approach in clinical applications. Fig. 9 shows two segmentation examples.

5. Analysis of the HDC

In this section, we experimentally analyzed the effectiveness of the HDC when it was used in the 3D LP-U-net for semantic segmentation of volumetric images. To conduct the experiments, we used the same dataset as we used in the ablation study on hip MR images with limited field of view. We took the testing data as our validation data to analyze the learning process and to compare the segmentation results. The 3D U-net was used as the baseline model.

5.1. Learning curves

We first analyzed the learning process of the proposed 3D LP-U-net with a fixed patch size of (400 × 400 × 80) but different shuffling factors and the associate baseline model, which was the 3D U-net, with different patch sizes. As shown in Fig. 10, in all cases, as the training loss goes down, the validation loss decreases consistently, demonstrating that there is no serious over-fitting for all models even with such small datasets. Please note that in order to better understand the results, the loss that we draw here equals to (the number of classes - the sum of DOC), which means for a perfect segmentation, the loss will be 0. Thus, the smaller the loss, the better the segmentation results.

From Fig. 10, we observe that due to the smaller patch size allowed by the 3D U-net, its learning curves are not smooth. Furthermore, the 3D U-net with large patch size has lower losses on both training and validation datasets than the one with small patch size, demonstrating the importance of using large patch size.

When comparing the learning curves of the 3D LP-U-net and the 3D U-net in Fig. 10, clear distinctions can be observed. First, due to the usage of large patch size, the learning curves of 3D LP-U-net are quite smooth. More importantly, the 3D LP-U-net not only converges much faster than the 3D U-net but also produces much lower losses on both training and validation datasets. It is also interesting to observe that for the 3D LP-U-net, in general, the bigger the shuffling factors, the larger the converged losses but the best results were obtained when the shuffling factor was (4, 4, 2). Such a qualitative observation was consistent with the quantitative results shown in Table 2. These results also demonstrate that the proposed HDC can effectively speed up the training procedure by overcoming optimization difficulties via learning better context features from larger patches. Moreover, due to the periodic shuffling operation, most of the subsequent computations of the 3D

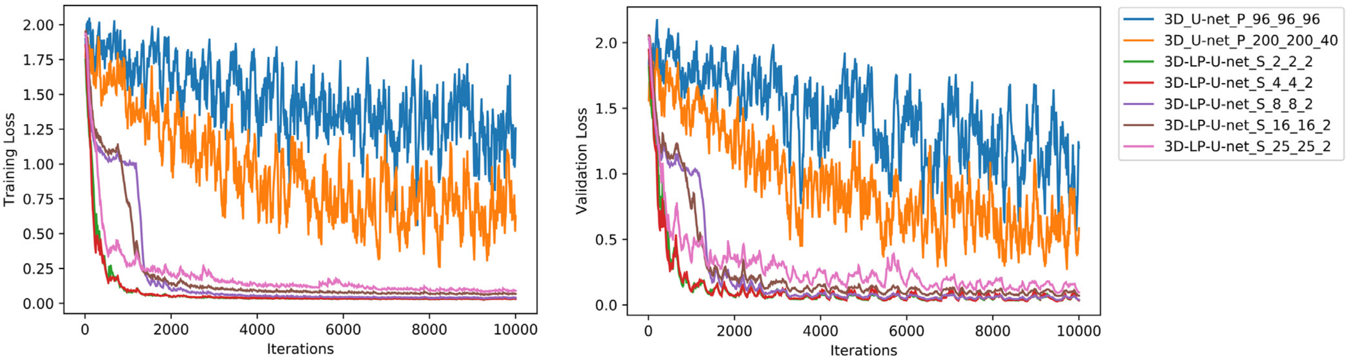


Fig. 10. Comparison of learning curves of the proposed 3D LP-U-Net with a fixed patch size of (400 × 400 × 80) but different shuffling factors and the 3D U-net with different patch sizes. The left images shows the learning curves of the training data and the right image shows the learning curves of the validation data, where "3D_U-net_P_X_Y_Z" means the results obtained from the 3D U-net with a patch size of ($X \times Y \times Z$) and "3D-LP-U-net_S_x_y_z" means the results obtained from the 3D LP-U-net with a shuffling factor of (x, y, z).

LP-U-net are done in the low-resolution space, leading to reduced computation time in both training and test stages.

5.2. Visualization of the 3D kernels of HDC

Next, we visualized the 3D kernels of HDC for a better understanding how HDC works. HDC consists of a periodic down-shuffling operation followed by conventional 3D convolutions with a kernel size of 3 × 3 × 3 in the low-resolution space. Direct visualization of the 3D kernels in the low-resolution space will not provide too much insight as the kernels are small. In order to better visualize the behavior of HDC, we can up-shuffle the 3D kernels in the low-resolution space to the high-resolution space. For example, if the down-shuffling factor is (25, 25, 2), then the 3D kernels in the high-resolution space will have a size of (75, 75, 6). Fig. 11 shows 20 out of 64 learned 3D kernels in the high-resolution space when the down-shuffling factor is (25, 25, 2). As shown in this figure, the weights of HDC have a strong similarity to designed features such as log-Gabor filters (Wang et al., 2008) or wavelets (Muraki, 1995). It is worth to mention that despite each kernel is independent in the low-resolution space, our independent kernels are actually smooth in the high-resolution space due to the periodic shuffling operations.

6. Discussion

Automated segmentation of volumetric medical images is a challenging task. The goal of the present study is to develop and validate a fully automatic deep learning segmentation pipeline that can address this challenge such that it can be used to segment 3D MRI and CT data of organs or structures of different volume sizes. In this paper, we presented a simple yet effective holistic decomposition convolution for improving semantic segmentation systems. The HDC consists of a periodic down-shuffling operation followed by conventional 3D convolutions. It can be directly applied to the input data and has the advantage of significantly reducing the size of the data for sub-sequential processing while using all the information available in the input irrespective of the down-shuffling factors. To achieve volumetric dense prediction at the output, we used a previously introduced dense upsampling convolution. Experimentally, we showed that HDC and DUC were network agnostic and could be combined with different FCNs for an improved performance. Based on results obtained from the ablation study, we chose the 3D LP-U-net as the segmentation approach. We then conducted extensive validation studies to evaluate the performance of the present approach when it was applied to four typical yet challenging volumetric image segmentation tasks such as segmentation of 3D T1 hip MR images with limited field of views, IVD segmentation and localization from multi-modality MR images, IVD segmentation from 3D T2 MR spine images and pancreas segmentation from 3D abdominal CT scans. The experimental results demonstrated that the present pipeline was able to accurately segment 3D MRI and CT data of structures of both large and small sizes, and could be applied to both mono-modality and multi-modality images.

In comparison with the state-of-the-art methods, the present approach achieved comparable or better results. For example, based on the focused shape models, Chandra et al. (2014) reported a mean absolute surface distance of 0.55 ± 0.18 mm and 0.75 ± 0.20 mm for femoral head and acetabular bone segmentation respectively in 35 3D unilateral MR datasets acquired from 25 subjects with different field of views. In comparison, as shown in Fig. 7, the 3D LP-U-net achieved a mean absolute surface distance of 0.27 ± 0.09 mm and 0.36 ± 0.20 mm for the proximal femur and the acetabulur bone segmentation respectively. By concentrating on segmenting the proximal femur from 3D MR images, Deniz et al. (2018) showed an average DOC of 95.0 ± 2.0%. In contrast, as shown in Fig. 7, our approach could be used to segment both the proximal femur and the acetabulum in 3D MR images and achieved respectively a mean DOC of 98.14 ± 0.47% and 96.76 ± 0.92% for these two bony structures. When the present approach was applied to the IVD segmentation task, as shown in Table 7, the 3D LP-U-net achieved better results than the top-5 methods as described in Zheng et al. (2017) when all methods were evaluated on the MICCAI 2015 IVD localization and segmentation challenge datasets. It is worth to mention that the 3D LP-U-net outperforms the method from the team UNICHK, which is a deeply supervised 3D segmentation network, by nearly 3.6% in terms of average DOC, which is a large improvement. We also showed that our approach could be applied to multi-modality image segmentation tasks. As shown in Table 8, our approach achieved comparable or better results than the top-5 methods as described in Zeng et al. (2018) when all methods were evaluated on the MICCAI 2018 IVD localization and segmentation challenge datasets. Furthermore, the cross-validation experiments that we conducted on the NIH pancreas CT dataset demonstrated the superior performance of the present approach. As shown in Table 9, our approach achieved better results than most of the state-of-the-art methods. For example, based on a two-stage approach combining random forest regression based pancreas localization with holistically-nested CNNs on three orthogonal views, Roth et al. (2018a) reported a mean DOC of 81.27 ± 6.27% when their approach was evaluated on the NIH pancreas CT dataset using 4-fold cross-validation. In Zhou et al. (2017), pancreas segmentation was

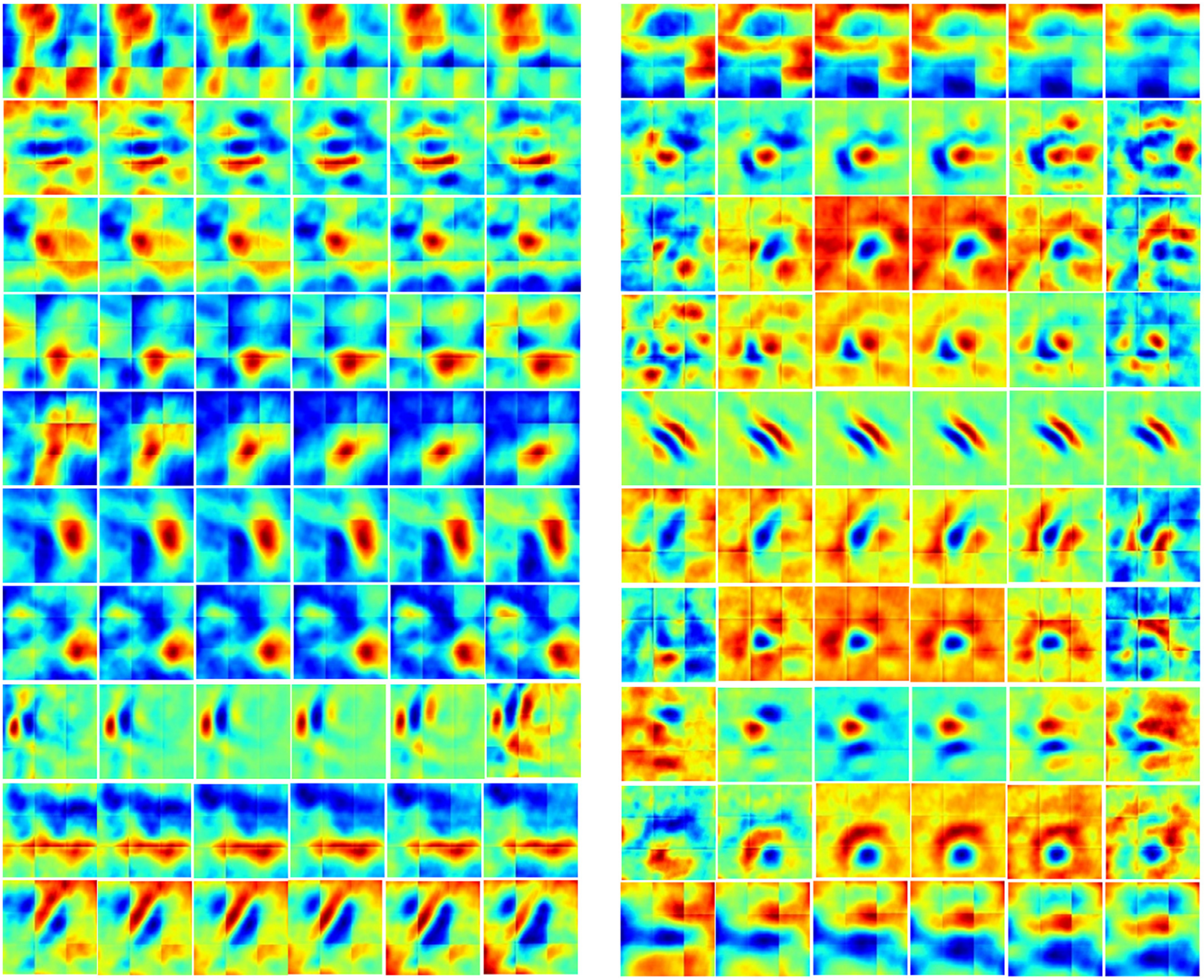


Fig. 11. Visualization of the 20 out of 64 learned 3D kernels in the high-resolution space when the shuffling factor is (25, 25, 2). Each row presents two kernels of size 75 × 75 × 6 expanded along the third dimension as six 75 × 75 maps.

performed slice by slice by applying an iterative fixed model. When this method was evaluated on the NIH pancreas CT dataset, a mean DOC of 82.37 ± 5.68% was reported. In comparison, with a coarse-to-fine strategy, our approach achieved a mean DOC of 83.0 ± 5.85%. Another 2D CNN-based method was presented in Yu et al. (2018). Following a coarse-to-fine strategy and by carefully designing a recurrent saliency transformation network to incorporate additionally multi-stage visual cues, their method achieved a better average DOC than our approach. Nonetheless, as shown in Table 9, our approach achieved much better result in the worst case than the method presented in Yu et al. (2018).

More importantly, the present approach can facilitate the application of 3D deep neural networks to medical image analysis tasks. It is worth to note that the medical images differ from natural images in that a large part of the medical imaging data available in clinical routine is in 3D, e.g, magnetic resonance imaging (MRI) data, computed tomography (CT) data and data generated by many other modalities. To deal with these data, researchers either use 2D networks, which perform image analysis by slicing a 3D volume into 2D slices, or train a 3D network using patch-based training and testing strategies. Although most of the 2D methods use strategies to fuse the output from different 2D views to obtain 3D segmentation results, they inevitably lose some 3D contexts, which is important for capturing the discriminative features for the target task. In the latter case, the size of the input patch is usually small if no specialized hardware with large GPU memory is used, which may introduce discontinuity artifacts (Roth et al., 2018b). Previously, in order to use large patch in 3D networks, one would have to use specialized hardware with large GPU memory. For example, Deniz et al. (2018) used an input data size of 256 × 256 × 48 voxels by training their 3D U-net on a server using a NVIDIA Tesla P100 GPU card with 16 GB GPU memory. Using a DeepLearning BOX (univ) including four NVIDIA Quadro P6000 GPUs with 24 GB each, Roth et al. (2018b) fed patches with a fixed size of 120 × 120 × 120 to their 3D U-net. In contrast, running on a single NVIDIA GTX 1080 Ti graphics card with 11 GB GPU memory, our approach allows for using patches with a size as large as 400 × 400 × 80. Additional advantage of the present approach is that it is computationally more efficient than other state-of-the-art 3D networks. As shown in Table 3, it took 294 min to train the 3D U-net when the input size was chosen to be 200 × 200 × 40. To achieve similar accuracy, we could use 3D LP-U-net with a shuffling factor of (16, 16, 2) and a fixed patch size of 400 × 400 × 80. To train such a network, it only took 121 min. Although strided convolution can also be used to downsample data, as shown in Table 6, it leads to worse results than our approach, especially in segmenting small size of objects such as IVDs.

In summary, we presented a novel deep segmentation pipeline that achieved state-of-the-art performance in four typical yet challenging segmentation tasks. Our approach is not only accurate but also computationally efficient. Extensive experiments conducted on both in-house and open datasets confirmed the efficacy of the present approach.

Declaration of Competing Interest

None.

Acknowledgments

This study was partially supported by a start-up funding from Shanghai Jiao Tong University, China Grant No. WF220882002 and the Swiss National Science Foundation via project 205321_163224/1.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2019.07.003.

References

Aitken, A., Ledig, C., Theis, L., Caballero, J., Wang, Z., Shi, W., 2017. Checkerboard artifact free sub-pixel convolution: a note on sub-pixel convolution, resize convolution and convolution resize. arXiv:1707.02937v1.

Arezoomand, S., Lee, W.-S., Rakhra, K.S., Beaulé, P.E., 2015. A 3d active model framework for segmentation of proximal femur in mr images. Int. J. CARS 10 (1), 55–66.

Ayed, I.B., Punithakumar, K., Garvin, G., Romano, W., Li, S., 2011. Graph cuts with invariant object-interaction priors: application to intervertebral disc segmentation. In: Biennial International Conference on Information Processing in Medical Imaging. Springer, pp. 221–232.

Cai, J., Lu, L., Xie, Y., Xing, F., Yang, L., 2017. Pancreas segmentation in mri using graph-based decision fusion on convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 674–682.

Castro-Mateos, I., Pozo, J., Eltes, P., Del Rio, L., Lazary, A., Frangi, A., 2014. 3D segmentation of annulus fibrosus and nucleus pulposus from t2-weighted magnetic resonance images. Phys. Med. Biol. 59, 7847–7864.

Chandra, S.S., Xia, Y., Engstrom, C., Crozier, S., Schwarz, R., Fripp, J., 2014. Focused shape models for hip joint segmentation in 3d magnetic resonance images. Med. Image Anal. 18 (3), 567–578.

Chen, C., Belavy, D., Yu, W., Chu, C., Armbrecht, G., Bansmann, M., Felsenberg, D., Zheng, G., 2015. Localization and segmentaiton of 3d intervertebral discs in mr images by data driven estimation. IEEE Trans. Med. Imaging 34 (8), 1719–1729.

Chen, F., Liu, J., Zhao, Z., Zhu, M., Liao, H., 2017a. 3D feature-enhanced network for automatic femur segmentation. IEEE J. Biomed. Health Inf.

Chen, H., Dou, Q., Wang, X., Qin, J., Cheng, J.C., Heng, P.-A., 2016a. 3d fully convolutional networks for intervertebral disc localization and segmentation. In: International Conference on Medical Imaging and Virtual Reality. Springer, pp. 375–382.

Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.-A., 2017b. Voxresnet: deep voxelwise residual networks for brain segmentation from 3d mr images. NeuroImage.

Chen, H., Qi, X., Cheng, J.-Z., Heng, P.-A., et al., 2016b. Deep contextual networks for neuronal structure segmentation.. In: AAAI, pp. 1167–1173.

Chevrefils, C., Cheriet, F., Aubin, C.-É., Grimard, G., 2009. Texture analysis for automatic segmentation of intervertebral disks of scoliotic spines from mr images. IEEE Trans. Inf. Technol.Biomed. 13 (4), 608–620.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Proc. MICCAI, 9901, pp. 424–432.

Deniz, C.M., Xiang, S., Hallyburton, R.S., Welbeck, A., Babb, J.S., Honig, S., Cho, K., Chang, G., 2018. Segmentation of the proximal femur from mr images using deep convolutional neural networks. Sci. Rep. 8 (1), 16485.

Dong, S., Luo, G., Wang, K., Cao, S., Mercado, A., Shmuilovich, O., Zhang, H., Li, S., 2018. Voxelatlasgan: 3d left ventricle segmentation on echocardiography with atlas guided generation and voxel-to-voxel discrimination. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 622–629.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.-A., 2017. 3D deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. 41, 40–54.

Gilles, B., Magnenat-Thalmann, N., 2010. Musculoskeletal mri segmentation using multi-resolution simplex meshes with medial representations. Med. Image Anal. 14, 291–302.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31.

Huang, S.-H., Chu, Y.-H., Lai, S.-H., Novak, C.L., 2009. Learning-based vertebra detection and iterative normalized-cut segmentation for spinal mri. IEEE Trans. Med. Imaging 28 (10), 1595–1605.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.

Karasawa, K., Oda, M., Kitasaka, T., Misawa, K., Fujiwara, M., Chu, C., Zheng, G., Rueckert, D., Mori, K., 2017. Multi-atlas pancreas segmentation: atlas selection based on vessel structure. Med. Image Anal. 39, 18–28.

Krizhevsky, A., ISutskever, Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105.

Kronman, A., Joskowicz, L., 2016. A geometric method for the detection and correction of segmentation leaks of anatomical structures in volumetric medical images. Int. J. Comput. Assisted Radiol. Surg. 11 (3), 369–380.

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T., 2017. On the compactness, efficiency, and representation of 3d convolutional networks: Brain parcellation as a pretext task. In: Proc. IPMI, 10265, pp. 348–360.

Li, X., Dou, Q., Chen, H., Fu, C.-W., Qi, X., Belavỳ, D.L., Armbrecht, G., Felsenberg, D., Zheng, G., Heng, P.-A., 2018. 3D multi-scale fcn with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality mr images. Med. Image Anal. 45, 41–54.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, C., Zhao, L., 2018. Intervertebral disc segmentation and localization from multi–modality mr images with 2.5 d multi-scale fully convolutional network and geometric constraint post-processing. In: International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging. Springer, pp. 144–153.

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

Liu, J., Chen, F., Pan, C., Zhu, M., Zhang, X., Zhang, L., Liao, H., 2018b. A cascaded deep convolutional neural network for joint segmentation and genotype prediction of brainstem gliomas. IEEE Trans. Biomed. Eng. 65 (9), 1943–1952.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proc. CVPR, pp. 3431–3440.

Luo, G., An, R., Wang, K., Dong, S., Zhang, H., 2016. A deep learning network for right ventricle segmentation in short-axis mri. In: 2016 Computing in Cardiology Conference (CinC). IEEE, pp. 485–488.

Michopoulou, S.K., Costaridou, L., Panagiotopoulos, E., Speller, R., Panayiotakis, G., Todd-Pokropek, A., 2009. Atlas-based segmentation of degenerated lumbar intervertebral discs from mr images of the spine. IEEE Trans. Biomed. Eng. 56 (9), 2225–2231.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of 2016 International Conference on 3D Vision (3DV). IEEE, pp. 565–571.

Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J., Išgum, I., 2016. Automatic segmentation of mr brain images with a convolutional neural network. IEEE Trans. Med. Imaging 35 (5), 1252–1261.

Muraki, S., 1995. Multiscale volume representation by a dog wavelet. IEEE Trans.Visual. Comput.Graphics 1 (2), 109–116.

Neubert, A., Fripp, J., Engstrom, C., Schwarz, R., Lauer, L., Salvado, O., Crozier, S., 2012. Automated detection, 3d segmentation and analysis of high resolution spine mr images using statistical shape models. Phys. Med. Biol. 57, 8357–8376.

Norman, B., Pedoia, V., Majumdar, S., 2018. 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 172322.

Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M., 2013. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: MICCAI 2013;16(Pt 2), pp. 246–253.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Proc. MICCAI, 9351, pp. 234–241.

Roth, H., Lu, L., Lay, N., Harrison, A., Farag, A., Sohn, A., Summers, R., 2018a. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Med. Image Anal. 45, 94–107.

Roth, H., Oda, M., Shimizu, N., Oda, H., Hayashi, Y., Kitasaka, T., Fujiwara, M., Misawa, K., Mori, K., 2018b. Towards dense volumetric pancreas segmentation in ct using 3d fully convolutional networks. In: Medical Imaging 2018: Image Processing, 10574. International Society for Optics and Photonics, p. 105740B.

Roth, H.R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E.B., Summers, R.M., 2015. Deeporgan: multi-level deep convolutional networks for automated pancreas segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 556–564.

Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M., 2014. A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 520–527.

Roth, H.R., Oda, H., Zhou, X., Shimizu, N., Yang, Y., Hayashi, Y., Oda, M., Fujiwara, M., Misawa, K., Mori, K., 2018. An application of cascaded 3d fully convolutional networks for medical image segmentation. Comput. Med. Imaging Graphics 66, 90–99.

Shi, R., Sun, D., Qiu, Z., Weiss, K.L., 2007. An efficient method for segmentation of mri spine images. In: Complex Medical Engineering, 2007. CME 2007. IEEE/ICME International Conference on. IEEE, pp. 713–717.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub--pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883.

Shi, Z., Zeng, G., Zhang, L., Zhuang, X., Li, L., Yang, G., Zheng, G., 2018. Bayesian voxdrn: a probabilistic deep voxelwise dilated residual network for whole heart segmentation from 3d mr images. In: Proc. MICCAI, 11073, pp. 569–577.

Shimizu, A., Kimoto, T., Kobatake, H., Nawano, S., Shinozaki, K., 2010. Automated pancreas segmentation from three-dimensional contrast-enhanced computed tomography. Int. J. Comput. Assisted Radiol. Surg. 5 (1), 85.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: CVPR 2015. IEEE, pp. 1–9.

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G., 2018. Understanding convolution for semantic segmentation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1451–1460.

Wang, W., Li, J., Huang, F., Feng, H., 2008. Design and implementation of log-gabor filter in fingerprint image enhancement. Pattern Recognit. Lett. 29 (3), 301–308.

Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D., 2013. Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE Trans. Med. Imaging 32 (9), 1723–1730.

Xia, Y., Chandra, S.S., Engstrom, C., Strudwick, M.W., Crozier, S., Fripp, J., 2014. Automatic hip cartilage segmentation from 3d mr images using arc-weighted graph searching. Phys. Med. Biol. 59, 7245–7266.

Xia, Y., Fripp, J., Chandra, S.S., Schwarz, R., Engstrom, C., Crozier, S., 2013. Automated bone segmentation from large field of view 3d mr images of the hip joint.. Phys. Med. Biol. 21, 7375–7390.

Xie, S., Tu, Z., 2015. Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1403.

Yu, F., Koltun, V., Funkhouser, T., 2017a. Dilated residual networks. In: Proc. CVPR, pp. 636–644.

Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.-A., 2017b. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans. Med. Imaging 36 (4), 994–1004.

Yu, Q., Xie, L., Wang, Y., Zhou, Y., Fishman, E.K., Yuille, A.L., 2018. Recurrent saliency transformation network: incorporating multi-stage visual cues for small organ segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8280–8289.

Zeng, G., Belavy, D., Li, S., Zheng, G., 2018. Evaluation and comparison of automatic intervertebral disc localization and segmentation methods with 3d multi--modality mr images: a grand challenge. In: International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging. Springer, pp. 163–171.

Zeng, G., Yang, X., Li, J., Yu, L., Heng, P.-A., Zheng, G., 2017. 3d u-net with multi--level deep supervision: fully automatic segmentation of proximal femur in 3d mr images. In: Prod. MLMI@MICCAI, 10541, pp. 274–282.

Zeng, G., Zheng, G., 2017. Dsms-fcn: a deeply supervised multi-scale fully convolutional network for automatic segmentation of intervertebral disc in 3d mr images. In: International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging. Springer, pp. 148–159.

Zheng, G., Chu, C., Belavỳ, D.L., Ibragimov, B., Korez, R., Vrtovec, T., Hutt, H., Everson, R., Meakin, J., Andrade, I.L., et al., 2017. Evaluation and comparison of 3d intervertebral disc localization and segmentation methods for 3d t2 mr data: a grand challenge. Med. Image Anal. 35, 327–344.

Zhou, Y., Xie, L., Shen, W., Wang, Y., Fishman, E.K., Yuille, A.L., 2017. A fixed–point model for pancreas segmentation in abdominal ct scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 693–701.