Medical Image Analysis 69 (2021) 101894

Contents lists available at ScienceDirect

Medical Image Analysis

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

High-resolution 3D abdominal segmentation with random patch network fusion

Yucheng Tang[a,*], Riqiang Gao[a], Ho Hin Lee[a], Shizhong Han[b], Yunqiang Chen[b], Dashan Gao[b], Vishwesh Nath[a], Camilo Bermudez[c], Michael R. Savona[d], Richard G. Abramson[d], Shunxing Bao[a], Ilwoo Lyu[a], Yuankai Huo[a], Bennett A. Landman[a,c,d]

[a] Dept. of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235, USA
[b] 12 Sigma Technologies, San Diego, CA 92130, USA
[c] Dept. of Biomedical Engineering, Vanderbilt University, Nashville, TN 37235, USA
[d] Radiology, Vanderbilt University Medical Center, Nashville, TN 37235, USA

ARTICLE INFO

Article history:
Received 20 March 2020
Revised 4 November 2020
Accepted 5 November 2020
Available online 16 December 2020

Keywords:
Abdominal organ segmentation
3D CT
Coarse to fine
High resolution
Network fusion

ABSTRACT

Deep learning for three dimensional (3D) abdominal organ segmentation on high-resolution computed tomography (CT) is a challenging topic, in part due to the limited memory provide by graphics processing units (GPU) and large number of parameters and in 3D fully convolutional networks (FCN). Two prevalent strategies, lower resolution with wider field of view and higher resolution with limited field of view, have been explored but have been presented with varying degrees of success. In this paper, we propose a novel patch-based network with random spatial initialization and statistical fusion on overlapping regions of interest (ROIs). We evaluate the proposed approach using three datasets consisting of 260 subjects with varying numbers of manual labels. Compared with the canonical "coarse-to-fine" baseline methods, the proposed method increases the performance on multi-organ segmentation from 0.799 to 0.856 in terms of mean DSC score (p-value < 0.01 with paired t-test). The effect of different numbers of patches is evaluated by increasing the depth of coverage (expected number of patches evaluated per voxel). In addition, our method outperforms other state-of-the-art methods in abdominal organ segmentation. In conclusion, the approach provides a memory-conservative framework to enable 3D segmentation on high-resolution CT. The approach is compatible with many base network structures, without substantially increasing the complexity during inference.

Given a CT scan with at high resolution, a low-res section (left panel) is trained with multi-channel segmentation. The low-res part contains down-sampling and normalization in order to preserve the complete spatial information. Interpolation and random patch sampling (mid panel) is employed to collect patches. The high-dimensional probability maps are acquired (right panel) from integration of all patches on field of views.

1. Introduction

Computed tomography (CT) of the abdomen is an essential clinical tool in diagnostic investigation and efficient quantitative measurement for internal organs, bones, soft tissue and blood vessels (Hsieh et al., 2019). CT allows for identification of structures in possible abnormalities and tumors. To explore complicated spatial relationship between abdominal organs and tissue structures, multi-organ segmentation on CT scans has been widely studied (Wolz et al., 2019; Xu et al., 2015). Manual annotations are regarded as gold standard (Simpson et al., 2019), but these are time and resource intensive. To reduce the manual efforts, atlas-based methods (Aljabar et al., 2009; Liu et al., 2017; Wang et al., 2012; Xu et al., 2014) and deep models (Ronneberger et al., 2015; Long et al., 2015; Zhang et al., 2018) have been proposed to achieve quantitative organ segmentation from the clinically acquired CT scans automatically (Xu et al., 2016).

CNN based models are widely explored for medical image analysis. From the perspective of computation resource-accuracy tradeoffs, 2D approaches take separated slices for training result in lacking spatial information, but faster at approximately batch size of 8 and 700 iterations per minute (Lai et al., 2015). The tri-planar architecture (Moeskops et al., 2016) performs better than 2D with

* Corresponding author.
E-mail address: yucheng.tang@vanderbilt.edu (Y. Tang).

https://doi.org/10.1016/j.media.2020.101894
1361-8415/

advantage of three views for each voxel at approximately batch size of 3 and 200 iteration per minute. The 3D architecture needs scans to be either 1) patched or 2) down-sampled, it is the slowest way to train at approximately 80 iterations per minute and one patch per iteration. Recently, CNN methods have been explored to 3D segmentation, which perform abdominal segmentation with 3D volumes, like 3D U-Net (Çiçek et al., 2016) or V-Net (Milletari et al., 2016). However, we cannot directly fit the clinically acquired high resolution CT (e.g., 0.8 mm or higher isotropic voxel size) to such networks due to the memory restriction of prevalent GPU. In this context, (de Brebisson et al., 2015) proposed a network to learn 2D or 3D patches along with volume coordinates for 3D segmentation. One key observation is that the patch-based methods (Coupé et al., 2011; Bai et al., 2013; Asman et al., 2013; Zhang et al., 2012; Yang et al., 2016) in high-resolution approaches tend to underperform given a lack of broad spatial context. Huo et al. (2019) demonstrated that the patch-based method for whole brain segmentation, to deal with the local anatomical variation based on registered atlases.

Unlike brain segmentation, abdomen CT do not have a well-established registration method for standard space due to larger variations in soft tissues among subjects. Thus, removing spatial context of a high dimensional volume by cropping images leads to a loss of relevant knowledge for abdominal segmentation. A second way for implementing 3D training is to down-sample the image to low resolution volume (Çiçek et al., 2016). However, this approach introduces fuzzy interpolation operations that will break biologic structures in medical images. Holger et al. proposed hierarchical method (Roth et al., 2017), which introduced a coarse-to-fine strategy that significantly improved the performance of pancreas segmentation. Holger et al. also proposed the multi-scale pyramid network (Roth et al., 2018a) that extend the hierarchical strategy to multi-stage learning. The input images are scaled at different levels, and predictions by last level can be selectively emphasized. However, the performance of scaled images may miss voxels due to inaccurate bounding box predicted by lower level models. Additionally, the output segmentations from upper levels present higher resolution but it still needs to be up-sampled to original space.

Currently, most prevailing deep learning frameworks on medical image segmentation are focused on similar backbones: FCN (Long et al., 2015), U-Net (Ronneberger et al., 2015) and Fast R CNN (Girshick et al., 2015). In practice, the tri-planar architecture aims to collect combinations of three-view slices for each voxel, and a 3D approach employs a 3D CT scan represented by a sequence of 2D slices. One of the first 3D models was introduced by Urban et al. (2014) to segment brain tumors with varying size. The intuition was followed by multi-scan, multi-path models (Kamnitsas et al., 2015; Chen et al., 2018) to capture subsampled features of the image. To exploit 3D context and to cope with limitation of computational resource, researchers investigated hierarchical frameworks. They attempt to extract features at multiple resolution levels. Roth et al. (2015) proposed a hierarchical architecture to perceive multi-scale information in pancreas segmentation. Chen et al. (2018) aims to simulate human behaviors and generalize RNN to employ 3D context. These approaches provide handling of different field of views at multiple levels, which reduces problems in both spatial context and low-resolution segmentation.

More works have been done on coarse-to-fine methods on abdominal organ segmentation. Li et al. (2018) proposed hybrid densely connected UNets, which learns 2D intra-slice features in the first stage then concatenates 3D contexts in the second stage. However, the connected 3D contexts are still down-sampled volumes, which limits in preserving high-resolution details. Zhou et al. (2017) developed an FCN based fix-point model to learn both the rough pancreas location and fine segmentation. But it only considered coarse-to-fine regions regardless of overlap regions, which is not optimized for spatial predictions. Similarly, Roth et al. (2017) proposed a hierarchical method, which introduced a coarse-to-fine strategy that significantly improved the performance of pancreas segmentation. However, by constraining rough pancreas locations, the method might be vulnerable to lose information. Roth et al. (2018a) extended the coarse-to-fine method to multi-scale pyramid networks. The input images are scaled at different levels, and predictions by last level can be selectively emphasized. However, the performance of scaled images may still miss voxels due to inaccurate bounding box predicted by lower-level models. Zhu et al. (2018) proposed an effective sliding window approach, performing 3D pancreas segmentation in two stages from both entire CT volume and sub-volumes. In addition, Zhu et al. (2018) used the expanding bounding box to improve the robustness for covering target regions. This approach adjusts many outliers in the experiment, but it still may suffer from catastrophic failures in the coarse stage. In summary, current state-of-the-art coarse-to-fine method for abdominal organ segmentation still needs down-sampling in the training and testing and might be vulnerable to failure localization in the coarse stage. To address these issues, we focus on proposing effective patch-based method without isotropic interpolation in the fine stage, while protecting patches from losing target information. Herein, we propose a concise coarse-to-fine framework named random patch network fusion (RPNF), design to alleviate the difficulties for 3D multi-organ segmentation. The method presents two advantages comparing to state-of-the-art methods.

To deal with the anatomical variance from medical images, patches are widely employed to handle the high dimensionality issue (Coupé et al., 2011; Bai et al., 2013; Zhang et al., 2012; Huo et al., 2019; Wang et al., 2013; Schlegl et al., 2019; Wang et al., 2019a). Schlegl et al. (2017) proposed a patch-based method on retina images. For medical image segmentation, 3D patch-based methods are used in many applications (Eskildsen et al., 2012; Asman et al., 2015; Wang et al., 2019b). In these methods, patches are represented as structural tiling architectures. Each individual region within patch follows fixed pattern over a pre-defined cropping displacement. Ding et al. (2016) proposed translational data augmentation, which employed shifts at each sampling point. The resulting performance exploited advantages of successive shifts and yields to final result with concatenation. However, these methods are dependent on manually defined landmarks or extra labels. To break the fixed definition, Cheheb et al. (2017), Liang et al. (2001) evaluated the benefit of random features for patch-based segmentation, which provided a robustness analysis to patch-based methods. Coupe et al. proposed an ensemble method (Coupé et al., 2019) based on a large number of CNNs processing for different brain areas. The assembleNet (Coupé et al., 2019) introduces sharing of knowledge among multiple U-Nets and assembles result with high-resolution predictions by majority voting.

Herein, we propose a concise coarse-to-fine framework named random patch network fusion (RPNF), design to alleviate the difficulties for 3D multi-organ segmentation. The method presents two advantages comparing to state-of-the-art methods. 1) it enables segmentation in original CT resolution without image scaling in the input and output. 2) it performs robustness to save the catastrophic failures from the coarse stage. The method enables segmentation in original CT resolution by a two-stage cascade design. The proposed strategy is built on the concept that the performance of a higher resolution level in hierarchical model is indicative of the low-resolution level in hierarchy. To validate the proposed strategy, experiments on baselines methods are performed, including low-resolution (Çiçek et al., 2016), high-resolution (Coupé et al., 2011) and multi-scale pyramid models (Roth et al., 2018a). For the family of patch-based method, we evaluate different strate-

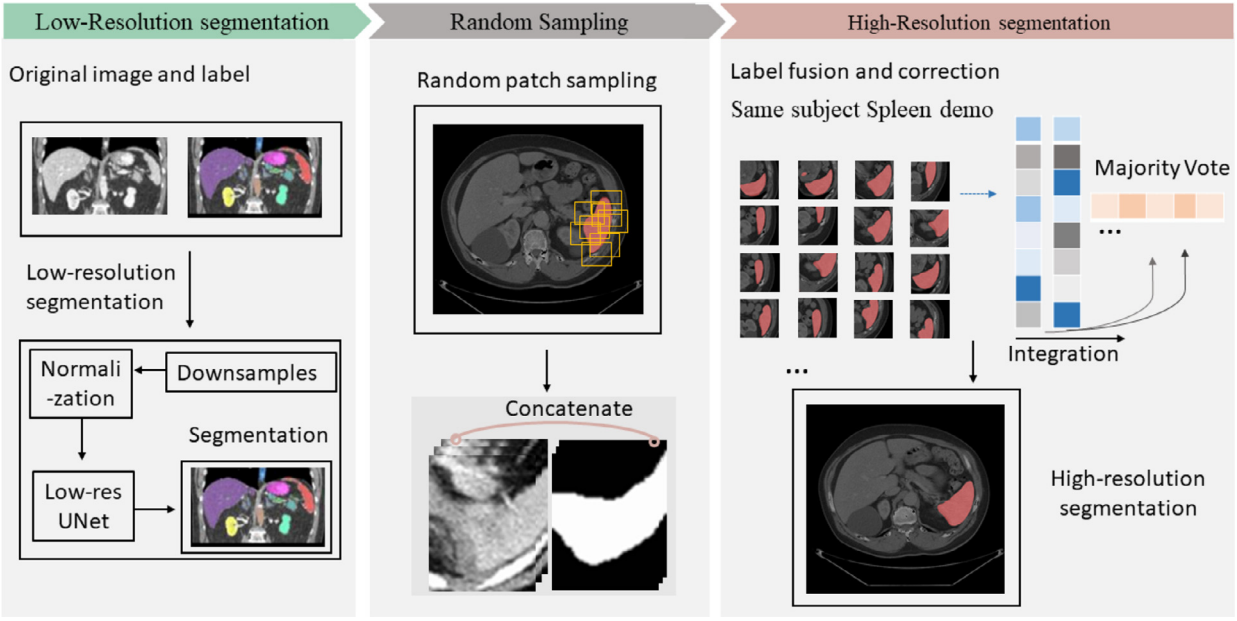

Fig. 1. Method framework. Given a CT scan with at high resolution of ~0.8 x ~ 0.8 x ~2 mm, a low-res section (left panel) is trained with multi-channel segmentation. The low-res part contains down-sampling and normalization in order to preserve the complete spatial information. After the coarse segmentations are acquired from low-res UNet, we interpolate the mask to match the image's original resolution. Next, random patch sampling (mid panel) is employed to collect patches, and patches are concatenated with corresponding coarse segmentation masks. Finally, we trained a patch-based high-res (right panel) segmentation model, the high-dimensional probability maps are acquired from integration of all patches on field of views. Majority vote is used to merge estimates into a final segmentation.

gies including structural tiling, random shifting and combined approaches. We perform sensitivity analysis in terms of patch numbers, as well as the ablation studies on behalf of averaged coverages per voxel and the effect of variant patch size. We present our study on the dataset from "Multi-Atlas Labeling Beyond the Cranial Vault" (BTCV) Challenge of MICCAI 2015. For external validation, we evaluate our method on two cohorts that were excluded from training, the ImageVU pancreas with 40 subjects and HEM1538 with 82 subjects.

To summarize, the contributions of this work are:

(1) We proposed a new coarse-to-fine framework termed 'random patch network fusion' by introducing randomly localized patches between first and second stage.
(2) We show that our proposed method can be implemented to predict original space segmentation in second level model.
(3) We provide large-scale validations on analyzing patch-based strategies and comparing them with our method, supporting that patch-based method plus random shifting could boost 3D segmentation performance.

2. Theory

The proposed method for abdominal segmentation consists of three main components: (1) a 3D multi-organ U-Net that produces coarse, preliminary segmentations, (2) a random patch sampling process which imposes constraints of the field of view, and (3) a second stage model followed by statistical fusion to achieve final segmentation (Fig. 1). The approach combines convolutional neural networks, hierarchical models and statistical fusion. For training the subject image i, given Hounsfield Units of voxels, the goal of the random patch network fusion algorithm is to estimate the segmentation S using observed labels s' from voters V_i. Consider the framework as a hierarchical model with two stages l and h. At each level, let $S_m = (S_l, S_h)$ be the mapping vector that corresponds to labels at each level of segmentation. Let $s \in L$, $Y_m = (Y_l, Y_h)$ be the collection of ground truth at the m level of hierarchy. The entire problem definition of our goal is to estimate the segmentation such that:

$$f(V_{ij} = s' \mid Y_i = s,\ S,\ \theta) \tag{1}$$

where voters V for each voxel j observes label s' given the ground truth Y, hierarchical model S, the parameters of each model θ.

2.1. Stage 1: preliminary segmentation

The labeled data i represents the original resolution CT scan. x is the down-sampled volume using tri-linear interpolation. Consider the segmentation network parameterized by θ. Low-resolution training aims to:

$$\underset{\theta_l,\ Y_l}{\operatorname{argmin}} L_{D_l}(\theta_l) \tag{2}$$

where $L_{D_l}(\theta_l)$ is the Multi-Sourced Dice Loss (MSDL) (Tang et al., 2019). MSDL was proposed as a way of evaluating datasets with varying labels with a single score by extending the Dice loss (Sudre et al., 2017) to adapt unbalanced multi-organ segmentation,:

$$L_{D_l} = -\frac{2}{A} \frac{\sum_{a=0}^{A} w \sum_{i=1}^{M} \sum_{j=1}^{N} Y_{ij} P_{ij} + \in}{\sum_{a=0}^{A} w \sum_{i=1}^{M} \sum_{j=1}^{N} Y_{ij}^2 + \sum_{a=0}^{A} w \sum_{i=1}^{M} \sum_{j=1}^{N} P_{ij}^2 + \in} \tag{3}$$

where A denotes the number of anatomies and w represents the variance to different label set properties in given image dimension of M and N. Y is the voxel value and P are the predicted probability maps. A small number, $\in$, was used in computing the prediction and voxel value correlation to prevent discontinuities. MSDL was iteratively optimized, and $P_{ij}$was computed by the softmax of the probability of voxel j in image i to anatomy.

2.2. Stage 2: random patch sampling

We proposed an approach that is inspired by hierarchical algorithms (Asman and Landman, 2014) and random sampling

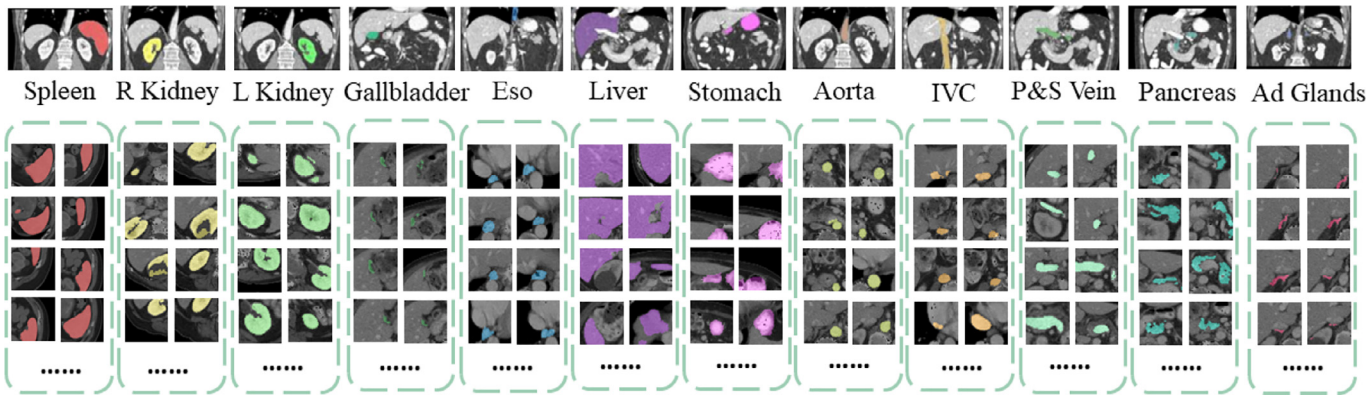

Fig. 2. Representative random patches for 12 abdominal organs of a single subject. The patch size is 128 × 128 × 48 and 8 samples are shown for each anatomy. Patch size defines the volume of field of view corresponding to organs. Large organs like spleen, liver and stomach cannot be covered until a number of patches are sampled, the patch of 128 × 128 × 48 covers most regions of mid-sized organ (kidney, pancreas, and portal & splenic vein), while small anatomies (adrenal glands, gallbladder, and vessels) can be covered by single patch with above size. The patch size effect is explored in an ablation study.

(Cheheb et al., 2017). We randomly select predicted voxels in the coarse segmentation mask according to the distribution. Using the selected voxel as indices' center, we place a bounding box as the local field of view. In order to introduce randomness, we also add a random shift to all axes' direction by the distribution. The distance of shifting is given by Gaussian random number generator, the mean and variance of the norm is determined by the mean distance of centers indices (e.g. the spleen patches have the mean distances of 4.2 and variance of 2.3 in x axes direction, the shifting distance to x direction is generated by the Gaussian random number generator). Patches are cropped according to bounding boxes as second stage model inputs as shown in the middle panel of Fig. 1. The strategy crops CT scan at original resolution without resampling, and it builds the hierarchy of non-linear features from random patches regardless of 3D contexts. The method employs detail context at original resolution and incorporates advantages of data augmentation with shifting.

2.3. Stage 3: label fusion

After separating full spatial context to k randomly selected subspaces, patches will overlap with each other. The overlapped region could provide more than one segmentation label for a voxel. Herein, except placing patches back to original coordinate space, it is required to summarize a single label given a vector of class labels from n candidates. In this work, we implement label fusion with majority vote algorithm, which fuses n segmentations from network predictions to a single label. The final segmentation label for voxel j in image i is acquired by:

$$S_{ij} = \text{argmax} \frac{1}{n} \sum_{m=1}^{n} p(a|s', j) \quad (4)$$

where $p(as', j) = 1$ if s'aquals to anatomy class a and 0 otherwise. We ignore the voters outside the image space, related values are excluded in the label fusion. For voxels with equal number of voters, we label the voxel randomly to either be target or background. uncertainty.

3. Methods

The 3D abdominal segmentation task involves segmentation of 12 abdomen structures with highly deformable volume and shape. Anatomies present high class-imbalance, which involve large organs (spleen, liver, stomach and kidneys), vessels (aorta, portal and splenic vein, and inferior vena cava (IVC)), and small anatomies (esophagus, gallbladder, pancreas and adrenal glands). The details of each dataset are provided below. a) BTCV dataset: We perform de-identified data acquired from the Vanderbilt University Medical Center (VUMC) under IRB approval. We retrieved 100 subjects with 12 labeled anatomies, labels are annotated by experts. We integrate all 100 subjects in this study, the in-plane pixel dimension of each scan varies from 0.4 to 1.2 mm. Each volume is preprocessed by excluding outlier intensities beyond −1000 and 1000 HU. The slice thickness ranges from 1 to 6 mm. Each CT scan consists 80 to 225 slices of 512 × 512 pixels. 100 CT scans are independent from subjects and with contrast enhancement in portal venous phase. Part of the dataset is released in the MICCAI 2015 Multi-Atlas Labeling Challenge, which contains 30 scans with 3779 axial slices (Zhou et al., 2019). The 12 organs were outlined manually by interpreters under supervision of clinical radiologists (MD) from Vanderbilt University Medical Center (> 10 years of experience in abdomen radiology). For each organ, the interpreter was instructed to verify the segmentation slice-by-slice in all axial, sagittal, and coronal views. To avoid inter-rater variability and perform reproducibility, we have independent observers perform manual segmentations on the same dataset.

(1) HEM1538 dataset: We retrieved 82 splenomegaly subjects substantially acquired with clinically trials. 117 splenomegaly CT scans are included and used as external validation (Huo et al., 2018). Splenomegaly indicates the enlargement of spleen with different levels of red blood cell destruction and inflation. These scans have large variance of spleen shape, which size varies from 143 cubic centimeter (cc) to 3045 cc. Each CT volume consists of 60 to 200 slices of 512 × 512 pixels, with resolution of ([0.59 × 0.59] mm to [0.98 × 0.98] mm). The slice thickness ranged from 1 mm to 2 mm. For each case, the spleen is manually annotated and reviewed by a radiologist.
(2) ImageVU pancreas:

A total of 40 subjects were selected and retrieved from Vanderbilt University Medical Center (VUMC). The dataset is collected from a group of 40 outliers out of 598 retrieved subjects. These outlier-guided subjects were a collection of studies that evaluated with benefits for rare/in-frequent population. The pancreas was manually traced for each subject under a soft tissue window.

(1) Patches for organs

To illustrate the random patch definition, we show the sampled patch field of views in Fig. 2, which are acquired from BTCV dataset, 12 annotated anatomical structures are shown.

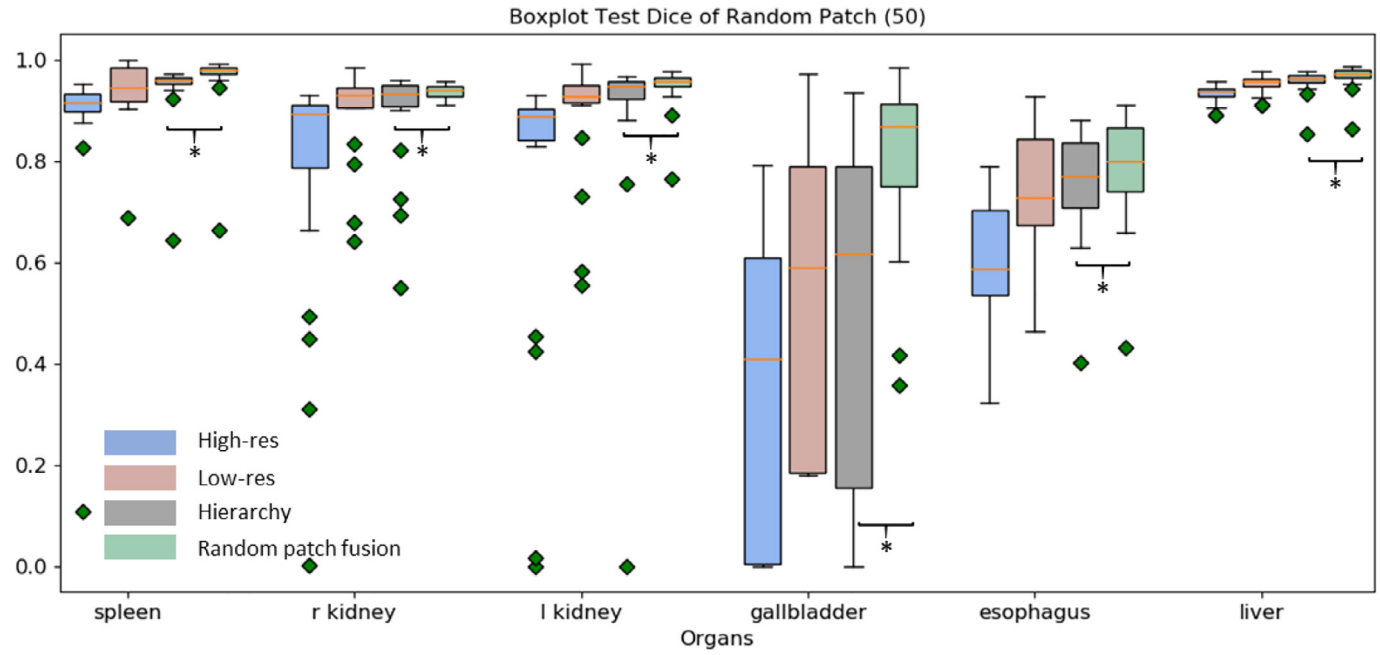

Fig. 3. Quantitative results from the testing cohort: spleen to liver (50 patches used). We compare our random patch network fusion method with three baseline approaches (high-res, low-res and hierarchical framework). The high-res method presents result with large variance and outliers in boxplot due to limited field of view in each patch. The low-resolution segmentation performs better than high-res method in mean DSC, which indicates complete spatial information is essential in abdominal organ segmentation. The hierarchical approach increases training resolution in the second step and achieved higher DSC. Hierarchical method's performance is limited when bounding box is inaccurate from previous levels. Our method achieves overall highest result compared to hierarchical method with significant improvement, "*" indicates statistically significant ($p<0.01$ from paired t-test). The random patch fusion framework employs advantages from both low-res and high-res settings, and it achieves segmentation without resample postprocessing. In boxplot, small anatomies (gallbladder, esophagus) presents higher improvements than large organs (spleen, kidneys and liver), which presents higher median DSC, smaller variance and fewer outliers.

3.1. Preprocessing and body part regression

We processed CT scan with soft tissue window with range of [−175, 250] HU. Intensities were normalized to [0,1]. Clinically acquired CT scans can exhibit large variance in volume size, we adopt a critical pre-processing with body part regression (Yan et al., 2018; Rousseeuw et al., 2005). The body part regression helps to remove slices on inconsistency volumes, and to localize anatomical regions automatically. We used the pre-trained model from unsupervised regression network (Yan et al., 2018) to navigate slices in abdomen region (scalar reference index ranges from −6 to 5).

3.2. Baseline architectures

We compared our random patch network fusion framework with a series of state-of-the-art approaches, including 1) low-resolution model on down-sampling images to fit maximum GPU memory, 2) high-resolution architecture with complete tiling patches, and 3) hierarchy with two-level pyramids.

3.3. Low-resolution architecture

The 3D U-Net is trained on images with finest resolution to house the maximum GPU memory. Each scan is down-sampled from [512, 512] to [168, 168] and normalized to consistent voxel resolution of [2 × 2 × 6]. The output and ground truth labels are compared using MSDL. We ignored the background loss in order to increase weights for anatomies. The final segmentation maps are up-sampled to original space with nearest interpolation (Olivier et al., 2012) in order to spatially align with CT resolution. This approach is trained end-to-end, and the resulting segmentation is summarized in Figs. 3 and 4. Qualitative results are shown in Fig. 5 and Fig. 6. The low-resolution framework incorporated down-sampled volume which lacks detail structures of anatomies, but it preserves complete spatial context in CT scan.

3.4. High-resolution architecture

The image is normalized to 1 mm isotropic resolution with dimension of 512 × 512. Since the high-resolution volume cannot be fed into GPU given structure of 3D U-Net, we employed $\boldsymbol{k}$ patches tile to cover full CT space. The patch number $\boldsymbol{k}$ for each image is based on equal distribution that continually tiles in x, y, and z axes. Each image is split to 32 patches along with the dimension, each patch covers a subspace. To maximize the usage of GPU memory, we use patch volume with [168 × 168 × 64] voxels. The subspace can be presented by coordinate $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ and patch size $(\mathbf{d_x}, \mathbf{d_y}, \mathbf{d_z})$.

$$\boldsymbol{\psi}_{\mathbf{k}} = [\mathbf{x_k} : \mathbf{x_k} + \mathbf{d_x}, \mathbf{y_k} : \mathbf{y_k} + \mathbf{d_y}, \mathbf{z_k} : \mathbf{z_k} + \mathbf{d_z}] \quad (5)$$

Patches are extracted without overlaps, each patch is padded to fixed size once it exceeded the volume dimension. For 3D U-Net, we adjust the decoder section upon original 3D U-Net implementation to be compatible with 12 labels prediction. 12 output channels are employed in the de-convolutional layers in the model. We also presented the overlapped patch strategies analyzed in Fig. 7. The half-overlapped patches covered half volume of subspaces, one-third overlapped patches cover one-third volume of subspace, etc. The effect of mean number of coverages per voxel is analyzed in Section 4.2. The high-resolution method is evaluated end-to-end, and final segmentation masks are acquired by tiling ordered patches.

3.5. Hierarchy with multi-scale pyramid network

To effectively segment an image at higher resolution, we compare our method with the multi-scale auto-context pyramid ap-

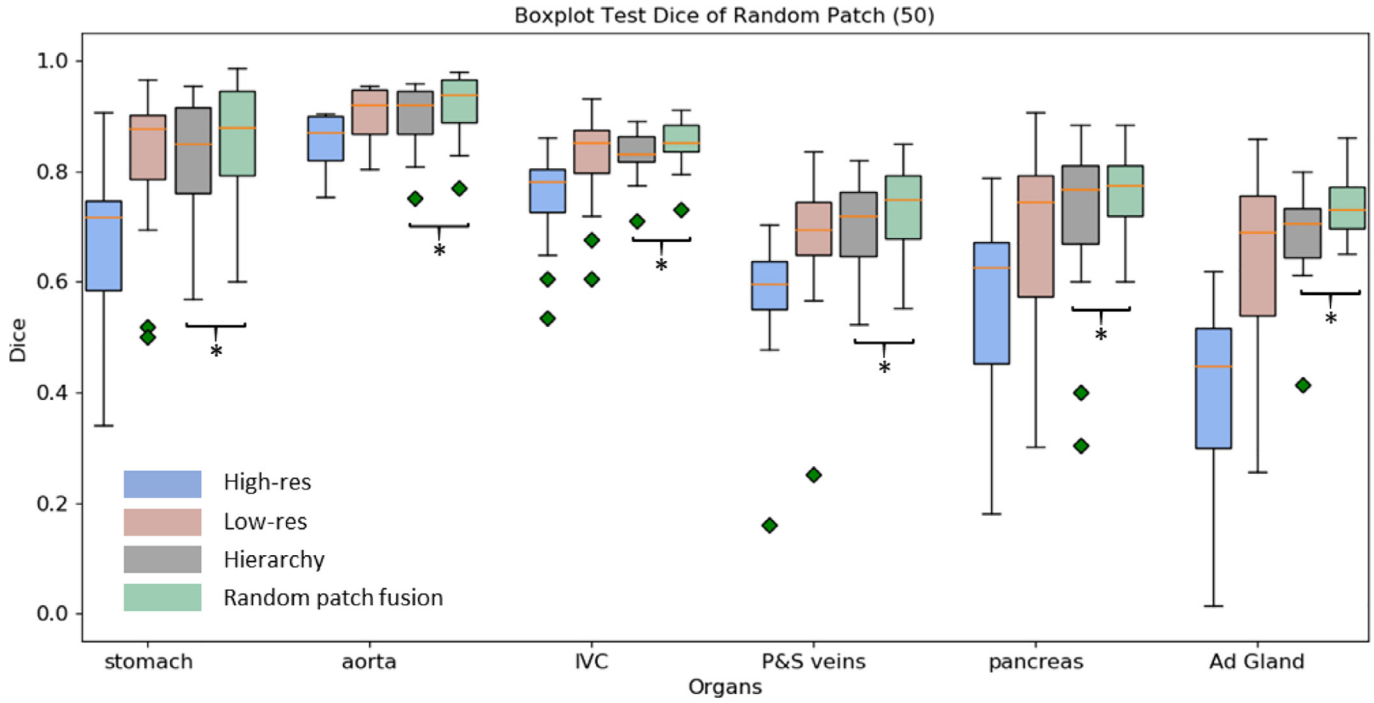

Fig. 4. Quantitative result for the testing cohort: stomach to adrenal glands (50 patches used). "*" indicates our method outperforms hierarchical method by statistically significant improvement (p<0.01 from paired t-test).

proach (Roth et al., 2018a). The method both captured spatial information at lower resolution down-sampled images while learned accurate segmentation from higher resolution in multiple levels. $\mathbf{F} = \{(\mathbf{f_m}(\mathbf{X_m}, \boldsymbol{\theta_m})),\ \mathbf{m} = 1, \ldots \mathbf{M}\}$, with $\mathbf{m}$ the order of levels in the approach. $\mathbf{X_m}$ is the subspace at level $\mathbf{m}$. In the first level, the 3D U-Net is trained the same as low-res network, which employed lowest resolution to fit largest amount of spatial information. In the next levels, it uses the predicted segmentation masks as an input channel to the next network. The succeeded input volume is cropped according to bounding box define by predicted segmentation map in level $\mathbf{m} - 1$. And down-sampled by a factor of $\mathbf{d_m} = \mathbf{d_{m1}}/2$. The previous level's segmentation is up-sampled by 2 in order to align with higher resolution levels. 3D U-Net at each level is optimized using Dice loss as the same wit low-resolution training. The predicted segmentation masks and cropped images are concatenated as the next level input. The final segmentations are acquired by interpolating the last level prediction and match the cropped images to original coordinates. In our implementation, we trained the pyramid models with two levels.

3.6. Implementation details

We adopt 3D U-Net as the segmentation model, which contains encoder and decoder paths with four levels resolution. It employs deconvolution to up-sample the lower level feature maps to the higher space of images. This process enables the efficient denser pixel-to-pixel mappings. Each level in the encoder consists two $3 \times 3 \times 3$ convolutional layers, followed by rectified linear units (ReLU) and a max pooling of $2 \times 2 \times 2$ and strides of 2. In the decoder, the transpose convolutions of $2 \times 2 \times 2$ and strides of 2 are used. And followed by two $3 \times 3 \times 3$ convolutions, followed with ReLU. 3D U-Net employs skip connectors from layers of same level in the decoder to provide higher-resolution features to the decoder part. The last layer is a $1 \times 1 \times 1$ convolution that set the number of output channels to the number of class labels. We used Multi-sourced Dice Loss and Dice Loss for multi-organ segmentation and single class segmentation respectively.

The baseline low-resolution multi-organ segmentation uses the largest volume size of $168 \times 168 \times 64$ in order to fit maximum memory of a normal 12GB GPU under architecture of 3D U-Net. The volume size is also employed in baseline hierarchical method for training the first level model. For patch-based segmentation, we firstly chose the medium size of [128,128,48] for experiments, the effect of different size of patch is evaluated in ablation study, presented in Figs. 8 and 9.

To fairly compare methods, the same 3D U-Nets is used with same hyper-parameters except input dimension and channels. We use batch size of 1 for all implementations. We used Instance Normalization, which is agnostic to batch size. We adopted ADAM algorithm with SGD, momentum=0.9. The learning rates is set to 0.001 and it reduced by a factor of 10 every 10 epochs after 50th epoch. Implementations are performed using NVIDIA Titan X GPU 12 G memory and CUDA 9.0. Training, validation and testing are executed on a Linux workstation with Intel Xeon CPU, 32GB of RAM. The code of all experiments including baseline methods are implemented in python 3.6 with anaconda3. Networks and frameworks are implemented in Pytorch 1.0.

3.7. Experimental design

We conducted experiments on three perspective of analysis to evaluate the effectiveness of different approaches. First, we compared state-of-the-art methods with RPNF on multi-organ segmentation to provide effectiveness. Then, in order to prove robustness and sensitivity, we did three ablation studies to validate the effect of 1) patch-based strategies, 2) number of random parches per scan and 3) patch size. Last, we tested the trained model on two external datasets to provide stability of the RPFN. All segmentation comparisons are assessed the average DSC score across 12 non-background labels. The claim of statistical significance is evaluated by paired t-test ($\mathbf{p} < 0.01$).

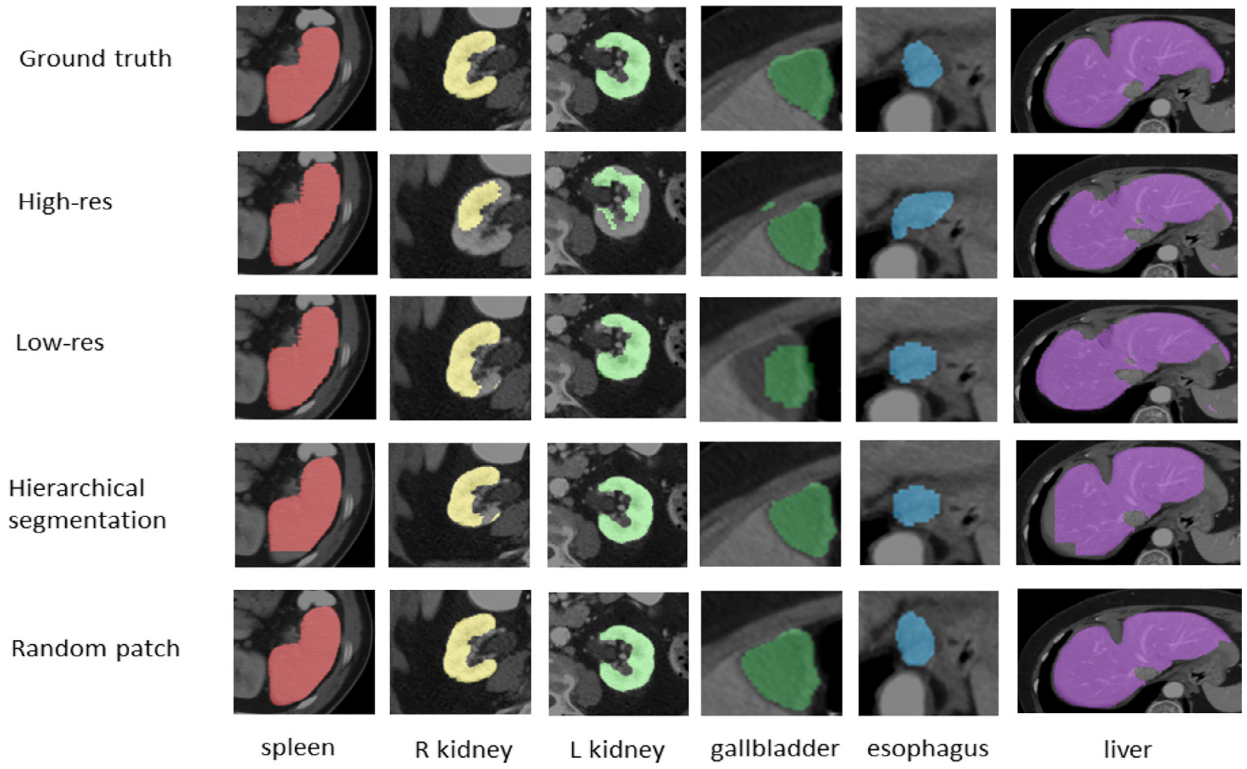

Fig. 5. Same subject qualitative result of our method compared to baseline approaches (spleen to liver). Second and third row presents direct high-resolution and low-resolution segmentation, mis-predictions are shown due to limited field of view, and resampling respectively. The hierarchical method presents smoother boundaries but suffers from truncation due inaccurate bounding box from first step. Our random patch fusion method presents complete segmentation masks with smoother boundaries among structures.

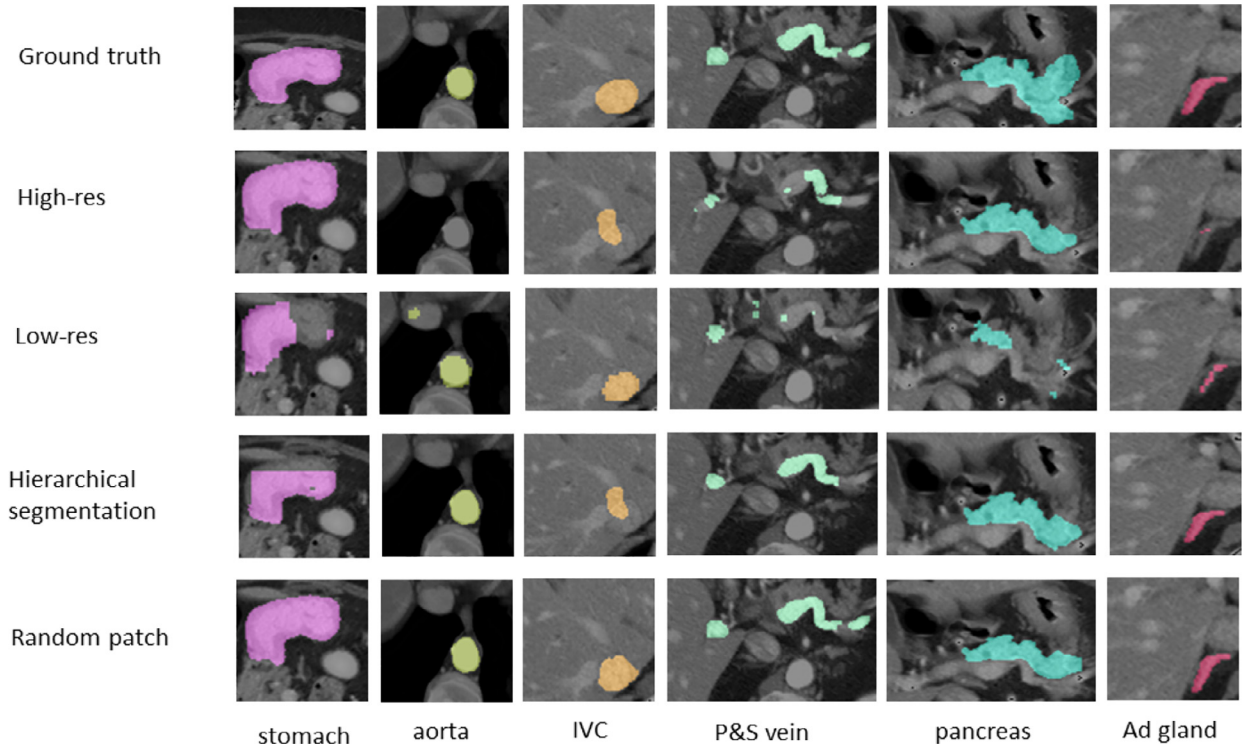

Fig. 6. The same subject qualitative result of our method compared to baseline approaches from stomach to adrenal glands.

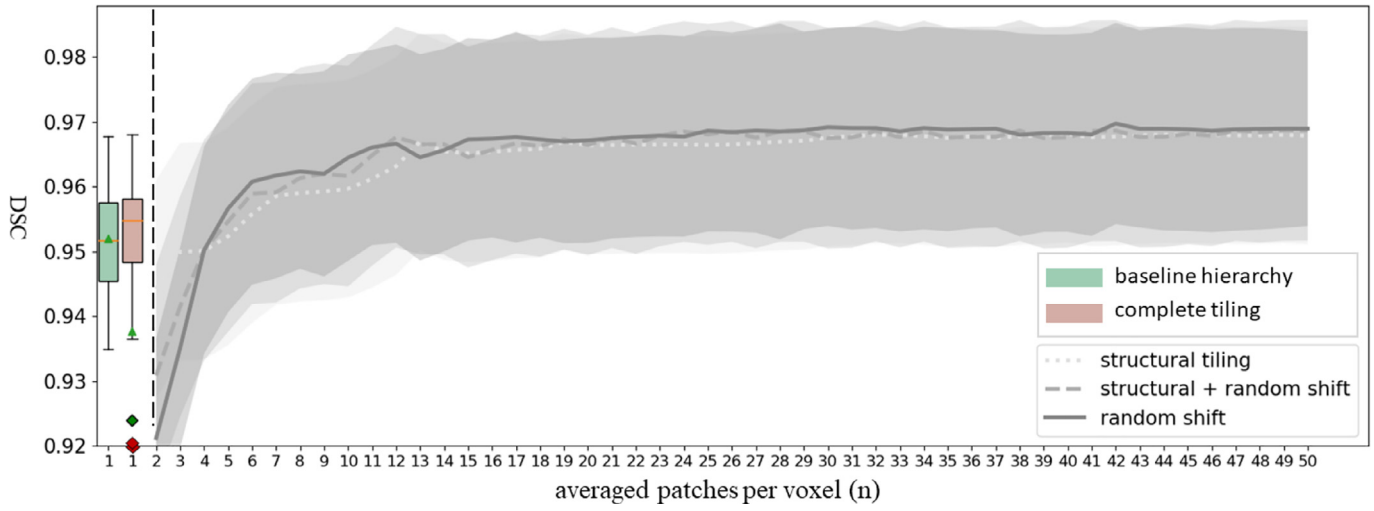

Fig. 7. Boxplot and uncertainty curves on patch strategies. The boxplots on the left presents the DSC coefficients on testing scans of baseline hierarchy method compared to the complete tiling. Complete tiling shows less variance than baseline hierarchy method. Red diamonds present outliers in complete tiling, and baseline hierarchy shows better DSC (green triangles) than complete tiling. Uncertainty plots show means/standard deviations comparison of structural tiling, structural tilling plus random shift and only random shift methods along with averaged patches per voxel. This presents DSC of each experiment with averaged covered voxels from 2 to 50.

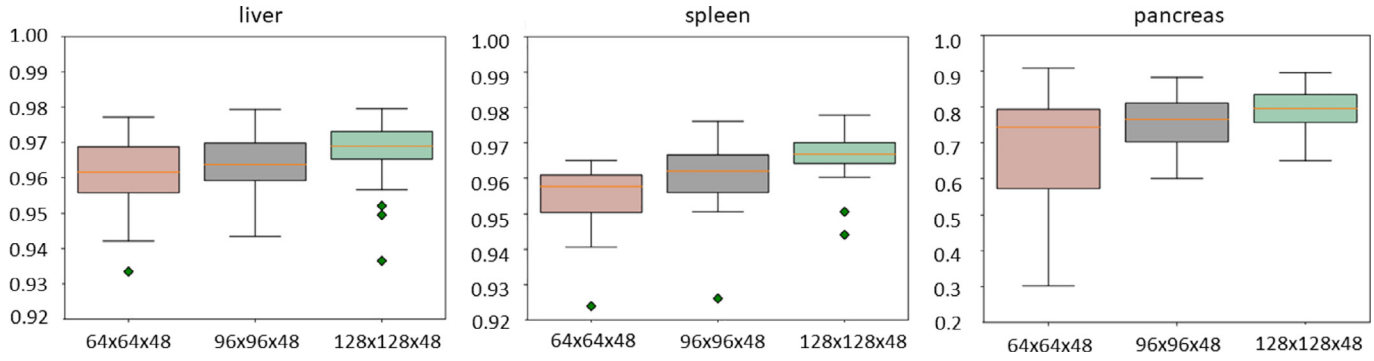

Fig. 8. Boxplot on three different patch size along x-y axes. The ablation study conducted on three abdominal organs (spleen, liver and pancreas). Patch size range from small (64 × 64 × 48), medium (96 × 96 × 48) and large (128 × 128 × 48). The boxplots show that larger patch sizes perform better than smaller patch sizes.

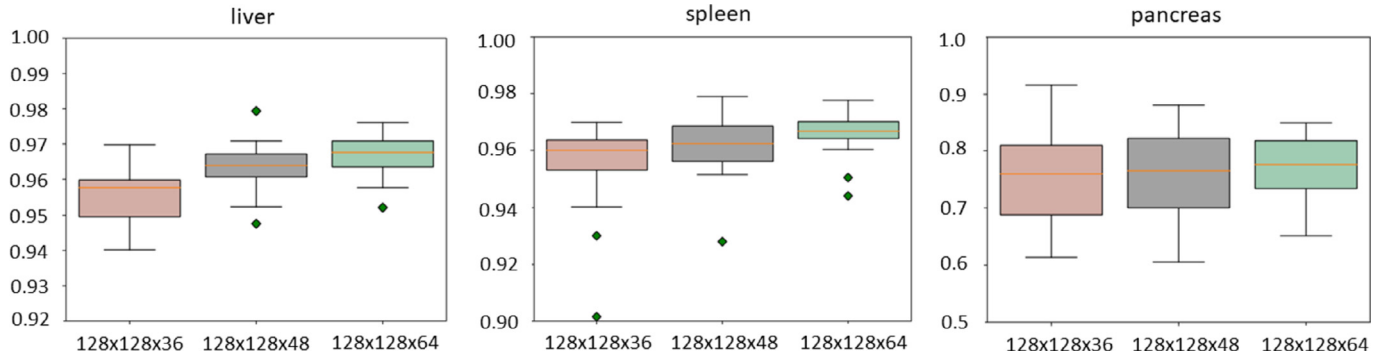

Fig. 9. Boxplot on three different patch size along the z-axis. The experiments are conducted on spleen, liver and pancreas with patch size ranging from small (128 × 128 × 36), medium (128 × 128 × 48) and maximum (128 × 128 × 64). The boxplots also present that larger number of slices perform better than less in the volume.

3.8. Random patch network fusion

To perform best effectiveness of the proposed method, we implemented experiments with maximum number (50) of patches for evaluating performance of baselines and RPNF. We implemented experiments with 5 fold cross validation on BTCV dataset. To perform standard five-fold cross validation, we split 100 scans into five complementary folds, each of which contains 20 cases. For each fold evaluation, we use 4 folds as training and testing on the remaining cases.

We compared RPNF with three baseline architectures (low-resolution, high-resolution and hierarchy) with same dataset and parameters on task of multi-organ segmentation. Briefly, we first trained the low-resolution approach, which has been shown its capability on 3D multi-organ segmentation with full spatial contexts. Second, we trained the 3D networks with structural tiles without overlapping to evaluate the high-resolution method. In this setting,

Table 1
Mean DSC and variance of 12 abdominal organs compared with our method and three baseline approaches on BTCV miccai2015 challenge testing cohort. Our method presented significant improvement compare to Hierarchical method. (p-value < 0.01 with paired t-test). Note: Bold values indicates best mean DSC of each organ.

Organ	High-resolution	Low-resolution	Hierarchy	RPFN (ours)
1.spleen	0.8732 ± 0.0316	0.9382 ± 0.0244	0.9422 ± 0.0048	0.9635 ± 0.0050
2.right kidney	0.7675 ± 0.0617	0.8996 ± 0.0180	0.8810 ± 0.0233	0.9310 ± 0.0231
3.left kidney	0.7579 ± 0.0812	0.8893 ± 0.0141	0.8872 ± 0.0435	0.9453 ± 0.0210
4.gallbladder	0.3565 ± 0.0896	0.5394 ± 0.0896	0.5014 ± 0.1178	0.8263 ± 0.0348
5.esophagus	0.6079 ± 0.0139	0.7481 ± 0.0140	0.7582 ± 0.0120	0.7881 ± 0.0120
6.liver	0.9321 ± 0.0039	0.9558 ± 0.0031	0.9557 ± 0.0072	0.9656 ± 0.0670
7.stomach	0.6641 ± 0.0247	0.8298 ± 0.0158	0.8267 ± 0.0118	0.8567 ± 0.0118
8.aorta	0.8540 ± 0.0024	0.9063 ± 0.0241	0.9032 ± 0.0326	0.9232 ± 0.0321
9.IVC	0.7528 ± 0.0068	0.8285 ± 0.0068	0.8328 ± 0.0192	0.8528 ± 0.0186
10.P&S veins	0.5778 ± 0.0123	0.6817 ± 0.0145	0.6979 ± 0.0672	0.7279 ± 0.0670
11.pancreas	0.5581 ± 0.0265	0.6765 ± 0.0265	0.7209 ± 0.0205	0.7608 ± 0.0535
12.Ad gland	0.3956 ± 0.0295	0.6321 ± 0.0295	0.6897 ± 0.0642	0.7356 ± 0.0367

local patches are cropped by pre-defined coordinates, patch size remains the same as low-resolution training ($168 \times 168 \times 64$). To evaluate the coarse-to-fine methods, we trained the two-level pyramid networks as the third baseline. In the first level, the low-dimensional representation is used for computational efficiency. Then, the input volume is cropped according to bounding box define by predicted segmentation map in the first level. The volume is down-sampled by a factor of $\mathbf{d}_2 = \mathbf{d}_1/2$, which means pyramid networks use higher resolution patches in the second level hierarchy. In this experiment, we evaluate the proposed method on all of 12 abdominal anatomies. We aim to show the method can be applied to multiple organs with variant sizes (large organ such as liver, and small anatomy such as adrenal gland). Thus, we claim the effectiveness of RPNF on the representative multi-organ dataset.

3.9. Ablation study

In this section, we evaluated the effect of three key factors that influence the performance. To simplify the evaluation on patch strategies and numbers, we conducted experiments on spleen segmentation as the representative task of abdomen segmentation. On evaluation of patch size, we performed the experiments on three representative organs (liver, spleen and pancreas). We performed experiments on the same BTCV dataset with 80 training scans, the withheld 20 cases are split with 10 for validation and 10 for testing.

Effect of patch-based strategies: For comparing patch-based strategies, we implemented methods of structural tiling, structural plus randomness and pure randomness. Briefly, we first employed complete structural tiling. Similar to high-resolution training, we cropped the image with fixed coordinates. In the second strategy, we start from the structural bounding boxes in the first method, then randomly shift each box in three directions (x, y and z). Last, we perform pure random selection of patches instead of predefined bounding boxes. To perform a fair comparison, the same 3D segmentation network with the same parameters are used in experiments. To be specific, the patch size $= 128 \times 128 \times 48$, batch size $= 1$, optimizer $=$ "Adam", and learning rate $= 0.001$. The preprocessing remains the same for different strategies.

Effect of average number of coverages per voxel: To further evaluate patch-based methods, we conducted a large-scale of experiments on number of coverages per voxel. We designed experiments on use of average number of coverages per voxel from 1 to 50. The structural tiling method is implemented by increasing overlapped region with 1/2, 1/3, 1/4, ... 1/50. The structural plus randomness is performed by shifting overlapped patches randomly. Last, the evaluation on the pure randomness strategy is achieved by increasing number of random patches until reach the same average coverage per voxel as other experiments. We showed the analysis in Fig. 7.

Effect of patch size: While patch-based studies achieved promising results, there are rare work focused on effect of patch size. Following the current prevailing GPU memory, previous studies (Çiçek et al., 2016) implemented patch with dimension for maximizing usage of memory (for example: $128 \times 128 \times 96$ vol typically occupies 12 GB with 3D U-Net). To better understand the effect of patch size, we evaluated 3D segmentation on three representative abdominal organs, spleen, liver and pancreas. We first employed three different sizes for evaluating dimension of x-y axes. Volume size varies from $64 \times 64 \times 48$, $96 \times 96 \times 48$ to $128 \times 128 \times 48$. Then we performed training on different volume length ($128 \times 128 \times 36$, $128 \times 128 \times 48$ and $128 \times 128 \times 64$). Except patch size, other settings remain the same as random patch network fusion.

3.10. Validation on external datasets

To show the stability of the proposed method, we validate the trained model on two independent cohorts (HEM1538 and ImageVU pancreas). We implemented the model of RPNF and baseline approaches on all subjects in these two unseen datasets. For evaluation of HEM1538, we aim to present the stability of our model, which trained on normal spleen and test on splenomegaly cases. For ImageVU pancreas, we compare the performance on outlier-guided study with academic controlled dataset. The same preprocessing and patch selections are used in the validation correspond to each model. Except the proposed PRNF method, the output segmentation volumes from other models were resampled back to the original image space.

3.11. Evaluation metrics

We used the Dice similarity coefficients (DSC) as the measurement for our method and baseline approaches,

$$DSC = \frac{2|A \cap M|}{|A| + |M|} = \frac{2|TP|}{2|TP| + |FP| + |FN|} \tag{6}$$

where TP is true positive, FP is false positive, FN is false a negative. The statistical measurement between methods were evaluated by paired t-test and the difference was significant when $p < 0.05$.

4. Results

4.1. Random patch network fusion

In Fig. 3, the quantitative boxplot shows the proposed random patch network fusion method with 12 anatomies' labels achieved

superior performance compared with baseline methods on metric of DSC scores. Table 1 reports mean DSC scores and standard deviation. As shown in Table 1, our proposed framework achieves state-of-the-art method "Hierarchy" (Roth et al., 2017) by a large margin. For large organs, our RPFN achieves 0.963 against 0.942 (spleen), 0.965 against 0.955 (liver), 0.856 against 0.826 (stomach), which are around advancement of 1.5%. In comparison of middle sized organ, our methods achieve 0.931 vs 0.881 (right kidney), 0.945 vs 0.887 (left kidney), 0.788 vs 758 (esophagus), 0.923 vs 0.903 (aorta), 0.853 vs 0.833 (IVC), 0.761 vs 0.721 (pancreas), which are around 3% advancement. Regarding of small tissues, our method achieves 0.826 vs 0.501 (gallbladder), 0.728 vs 0.698 (portal and splenic vein), 0.736 vs 0.690 (adrenal glands), which increases by a large margin.

Fig. 3 indicates our method achieved significant improvement with paired t-test of p-value < 0.001, compared with performance of the state-of-the-art two-level hierarchy. In Fig. 3, the DSC scores of high-resolution methods are the lowest since local patches result in holistic information. Unlike patch-based method in brains (Huo et al., 2019), abdomen CTs do not have registration step to alleviate the bias and variance in patients, which indicates intensities in patches are not scaled and normalized. Herein, we observed that soft structures such as stomach and pancreas show large std (DSC score) in Figs. 3 and 4.

A similar result happens in large structure, liver and spleen, since the segmented tiled patches contain outliers. On performance of low-resolution model, we see an improvement for all structures compare to high-resolution model, which indicates that spatial contexts are essential for 3D segmentation. As full context provides complete shape and background knowledge to training model, the low-resolution model shows smaller standard deviation in Table 1. The limitation of the low-resolution method comes from the tri-linear and nearest interpolation during downsample-upsample steps. Small structures, gallbladder, adrenal glands are with limited number of voxels in low-resolution volume, the predicted segmentation will not memorize the shape structure after up-sampling with nearest interpolation. In hierarchy approach, we implemented multi-scale pyramid network with two levels, the result present in gray boxes in Figs. 3 and 4. The hierarchical approach shows general better DSC scores than low-res model by incorporating spatial context in first level and higher resolution context in second level. The final segmentation result relies on the bounding box predicted by outputs in previous level. We observe that cases may miss part of structure due to cropping with inaccurate bounding box, as presented in qualitative result (Fig. 5). The uncertainty of boundaries results in amounts of outliers especially in segmenting soft structures (such as stomach and pancreas). Herein, the boost in DSC score for these structures are marginal for the hierarchical approach. The random patch network fusion presents overall higher DSC scores on all structures. We see a high improvement of ~30% DSC score for gallbladder. Adrenal glands also present large improvements, these small structures benefit greatly since a single random patch will cover entire structure, and the random patch works as data augmentation scheme while benefits with assembled result with label fusion. The random patch fusion network utilized advantages from all the baseline approaches, 1) complete spatial context in low-resolution model, 2) detailed feature of structures in high-resolution model and 3) coarse attention mechanism provided by multi-scale architecture. Additionally, the second step of our method predicts masks in original CT space, which indicates no re-sampling step needed for final segmentation result.

4.2. Ablation study

4.2.1. Effect of patch-based strategies

Hierarchy vs complete tiling: We present the results of 3D spleen segmentation by two patch-based baselines in left panel of Fig. 6, which is the two-level hierarchy method and complete tiling. The two-level hierarchy approach used the same patch configuration in the second step as complete tiling. Patch size of [128,128,48] is used for both experiments. The implementations are conducted with averaged patches per voxel of 1, which indicates no overlapped patches. In Fig. 7, we observed similar scenario as high-resolution experiment, the mean DSC score is lower than hierarchical method. This effect is probably due to complete tiling act only on local patches, which lacks holistic context. The unsatisfied performance of the tiling patches presents two reasons. First, compare with hierarchical method, complete tiling contains unrelated patches without target. Second, the large variance of target present unbalanced intensity distribution.

Structural tiling vs random shift: In this section, We evaluated translational data augmentation techniques. The right panel of Fig. 7 compares structural tiling, structural tiling plus random shift, and only random shift strategies. The structural tiling means adjacent patches cropped along axes such as complete tilling. With the increasing of average covered voxel, we shift the tiled patches to be overlapped by half, 1/3, 1/4, etc. The structured shifting was implemented in three dimensions (x, y and z axes) in order to balance spatial context for augmentation. The gray dash line in Fig. 7 indicates performance of structural tiling and random shift. From tiled patches, we implemented Gaussian random shift upon structural patches. This method involves moving the image randomly along the x, y, and z direction, which enables network to ignore absolute location of targets.

4.2.2. Effect of average number of coverages per voxel

We point out that patch-based approaches' performance is partially influenced by number of overlapping region of interests. To pin-point the gain of increasing number of patches, we conducted large-scale of experiments on spleen by using coverage of 2 per voxel to 50 for each patch strategy. For structural tiling, the number of coverages per voxel is calculated by overlapping tiles. For random shifting, we count the mean coverages per voxel as the same with structural tiling. In Fig. 7. These competitors perform decently when number of coverages reach 10 and more. Curves represent the mean DSC of each experiment, while the shading area indicates variance.

Interestingly, the performance of random shift is higher than structural tiling when n is less than 15. For n larger than 15, the effect of random shift is less influential compared with pure structural tiling. Herein, we conclude that translational data augmentation is comparable to random shift effect when averaged patches per voxel reaches a large number.

4.2.3. Effect of patch size

Fig. 7 shows the results on dimension of 64/96/128. With the increasing of the scheduled dimension, all models perform better DSC, as indicated by larger spatial context. Big patches contain broader spatial context with consistent intensity distribution and trace of boundaries. Then, we conducted experiment on increasing of slice numbers. With the fixed x-y dimension of 128, we changed volume length from 36 to 64 shown in Fig. 8, it's no supervise that the performance follows similar result in x-y dimension. We conclude that 3D U-Net is capable to capture local features in patch with larger size, with the current computational resource, larger patch is better than small patches in segmentation metric.

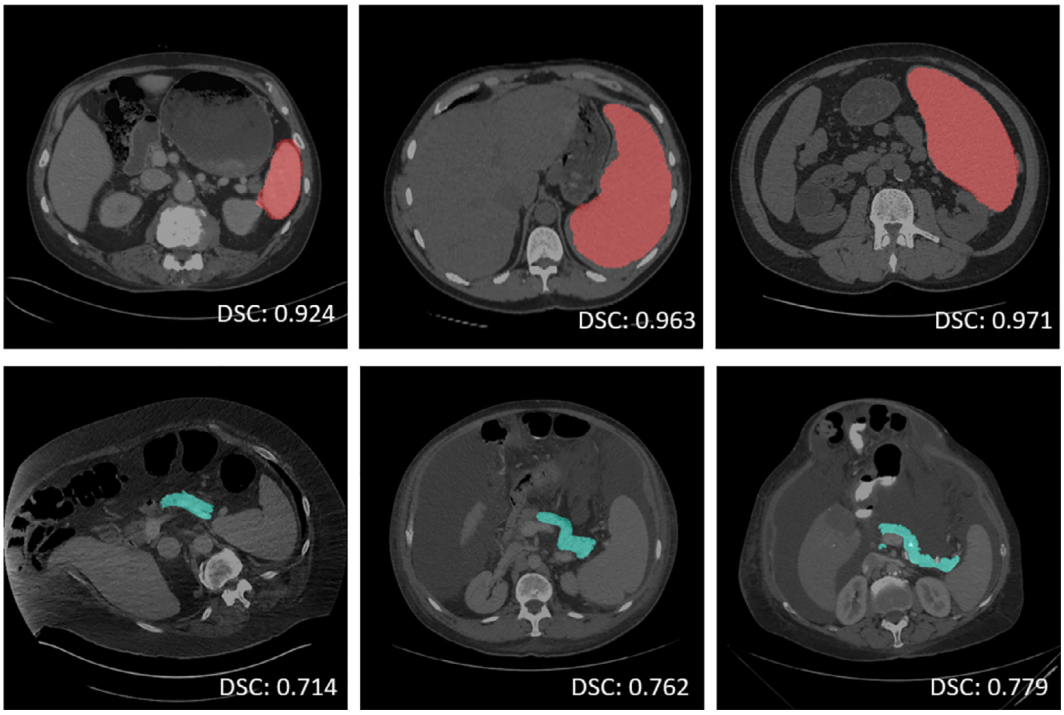

Fig. 10. Qualitative result of three representative subjects. From low to high, we show the segmentation result evaluated by our method. The testing performance on external datasets (top: HEM1538-splenomegaly, bottom: ImageVU-pancreas outliers).

Table 2
Segmentation performance of models trained on BTCV dataset in Mean DSC and variance, tested on HEM1538 and ImageVU pancreas, the proposed method is compared with baselines (p-value $<$ 0.01 with paired *t*-test between random patch network fusion and two-level hierarchy).

	HEM1538 (spleen)		ImageVU (pancreas)	
	Vol DSC	Std	Vol DSC	Std
High-resolution	0.9134	0.0283	0.5046	0.0711
Low-resolution	0.9268	0.0211	0.5537	0.0513
Two-level Hierarchy	0.9493	0.0189	0.5782	0.0548
Random Patch network fusion	0.9672	0.0143	0.6019	0.0423

Note: Bold cases indicates best mean DSC.

4.2.4. Validation on external datasets

In this evaluation, we investigated two clinical scenarios instead of research subjects. We adopt the model trained on the BTCV dataset and tested on external cohort with HEM1538 (splenomegaly) and ImageVU (pancreas). which are manually labeled by experts. Results are shown in Table 2 and Fig. 10.

HEM1538: The quantitative performance on HEM1538 is presented in Table 2. We implemented the same comparison experiments with low-resolution, high-resolution and multi-scale hierarchy. The mean Dice similarity coefficient (DSC) is calculated for all testing scans in HEM1538. The multi-scale hierarchy method achieves best DSC among baseline approaches and was used as a reference method. Our proposed random patch network fusion models perform better than multi-scale hierarchy as presented in boxplot with significant improvement ($\mathbf{p} < 0.01$, paired *t*-test). As HEM1538 is a pathology cohort with extra large spleens, we prove that RPFN could effectively preserves the stability on generalizing knowledge from normal spleen to splenomegaly.

ImageVU pancreas: In this study, we introduced an outlier-guided cohort since clinically acquired scans contain hard cases among population. The quantitative result is presented in Table 2, the mean volume DSC showed the detailed measurement for all methods, which showed that the proposed RPFN with 50 patches and majority vote achieves superior performance compared with the two-level multi-scale hierarchy with ($p < 0.01$, paired *t*-test with mean DSC).

4.3. Comparison of state-of-the-art methods

4.3.1. Coarse-to-fine methods

Our model is compared with other state-of-the-art coarse-to-fine networks. The results are in Table 3. Roth et al. (2017) used hierarchical 3D fully convolutional networks (FCN) with two stages. Zhou et al. (2017) developed a fix-point model for small organ segmentation. Li et al. combined 2D and 3D FCNs for hierarchically aggregating volumetric contexts. Zhu et al. (2018) proposed a novel 3D coarse-to-fine framework that achieved promising result on pancreas segmentation. For each method, we trained 12 models for 12 organs. Comparing with these state-of-the-art coarse-to-fine methods, our work achieves a consistent higher DSC. The average Dice of our method is 0.8564, compared to 0.7920 (Roth et al.2017), 0.8138 (Zhou et al.2018), 0.8176 (Li et al., 2018), 0.8328 (Zhu et al., 2018), respectively.

4.3.2. Patch selection methods

We implemented different patch selection strategies used in abdominal organ segmentation. To fairly conduct the evaluation, we used 3D UNet as the segmentation model for all methods. All experiments are implemented using the same BTCV dataset on 12 organs. As shown in Table 4, we compared five different strategies with our method in addition to the ablation study. The evaluation metric employed in these experiments includes the mean surface distance, Dice scores, and Hausdorff distance. In comparison of single-stage model, patches selected with overlap perform better than tiles without overlap. In comparison of two-stage model, we evaluated the fine-stage performance using fine-scaled, sliding window, and our random patch method. The average Dice of our method is 0.8564, compared to 0.8297 (Zhu et al., 2018), 0.7991 (Roth et al., 2018b), respectively.

Table 3
Comparison of coarse-to-fine methods between our proposed approach and state-of-the-art methods. The evaluation is conducted on BTCV testing dataset in terms of mean DSC.

Methods	spleen	R Kid	L Kid	Gall	Eso	liver	Sto	aorta	IVC	Vein	Pan	AG	All
Roth et al. (Roth et al. 2017)	.926	.884	.889	.531	.724	.953	.819	.884	.823	.687	.720	.664	.792
Zhou et al. (Zhou et al. 2018)	.941	.918	.932	.603	.753	.964	.842	.907	.820	.689	.722	.675	.814
Li et al. (Li et al., 2018)	.957	.917	.924	.636	.760	.963	.840	.901	.821	.697	.726	.669	.817
Zhu et al. (Zhu et al., 2018)	.961	.928	.932	.693	.772	.964	.849	.913	.837	.698	.762	.684	.833
Ours	.963	.931	.945	.826	.788	.966	.857	.923	.853	.728	.760	.736	.856

R Kid: right kidney, L Kid: left kidney, Gall: gallbladder, Eso: esophagus, Sto: stomach, IVC: inferior vena cava, Vein: portal and splenic veins, Pan: pancreas, AG: adrenal gland.

Table 4
Fine stage performance comparison with state-of-the-art methods on patch selection strategies using same backbone network (3D UNet). The evaluation is performed on BTCV testing data on 12 abdominal organs in terms of mean and variance.

Methods	Mean Surface Distance Vein Pan AG	Average Dice	Hausdorff Distance
Local patches (tiling no overlap)	6.6129± 3.1458	0.6748± 0.0670	52.1484± 31.9348
Local patches (tiling 1/2 overlap)	5.5195± 3.0981	0.7075± 0.0664	47.2357± 26.7541
Kim et al. (kim et al., 2020) (uniform crop)	5.4912± 3.0385	0.7493± 0.0659	45.0924± 27.1705
Roth et al. (Roth et al., 2018b) (fine-scaled)	4.6011± 2.5651	0.7991± 0.0623	38.5917± 20.6583
Zhu et al. (Zhu et al., 2018) (sliding window)	1.8143± 1.0359	0.8297± 0.0617	24.5591± 17.4515
Ours (random patch)	1.4237± 0.5916	0.8564± 0.0608	18.9862± 12.4169

Table 5
Average time cost per CT volume in the testing phase on different multi-organ segmentation models, where mean DSC is the average Dice score across 12 organs on BTCV testing data.

Methods	Mean DSC	Testing Time (s)
Roth et al. (Roth et al., 2018b)	0.7920±0.0652	304
Zhou et al. (Zhou et al., 2017)	0.8138±0.0644	312
Zhu et al. (Zhu et al., 2018)	0.8328±0.0631	294
Ours ($N = 25$)	0.8492±0.027	297
Ours ($N = 50$)	0.8564±0.0608	308

As shown in Table 4, our method outperforms other patch selection strategies in terms of three evaluation metrics. We observed that the fine-scaled and the sliding window method perform well when the first stage predictions are relatively good. But the fine-stage is sensitive to catastrophic failures predicted from coarse stage. The Zhu et al. (2018) method outperforms Roth et al. (2018b) mainly because it involves the operation of expanding box ($n = 12$). Our method achieved the improvement may be due to the smoothing effect introduced by random patches, it could save the catastrophic failures in the first stage. Fig. 5 shows visualization of sample results of our method compared to Roth et al. (2018b). We could observe that the fixed patches may be vulnerable to the error field of view given by first-stage segmentation.

4.3.3. Comparison of time efficiencies with different methods

We discuss the average time cost of our proposed method against other coarse-to-fine methods. The number of patches used in the approach matters the overall testing time. Here, we choose $n = 50$ and $n = 25$ for discussing the concern of accuracy-time trade-off. In experiments of Roth et al. (2018b), Zhou et al. (2017),Zhu et al. (2018), we trained 12 models for 12 abdominal organs. The time cost evaluation is calculated after acquiring the final multi-organ segmentation output (including post-processing steps reported in each method). In implementing Zhu et al. (2018), We choose the overlap size $n = 6$ and 12 as noted in the study. Experimental results are shown in Table 5. Zhu et al. (2018) is the most efficient. Our method achieves comparable time efficiency on $N = 25$. When N is larger, the performance improves but the testing time also increases which is reasonable. We also observe that, in the testing phase of coarse-to-fine methods, the time of loading models composes the most part. Overall, coarse-to-fine methods take more than double seconds in the testing phase due to the loading of multiple models, and the automatic algorithms take much less time than radiologists, which presents the clinical significance of the work.

4.3.4. Comparison with different medical image segmentation methods

We discussed different prevalent methods on medical image segmentation methods on the task of 12 abdominal organ segmentation. We used the same data split configuration during experiments. The results are shown in Table 6. In comparison of 2D methods, the basic model 2D UNet (Ronneberger et al., 2015) and ResNet (He et al., 2016) suffer from worse DSC of small organs such as adrenal glands and gallbladder. Mask R-CNN and DeepLab V3 (Chen et al., 2017) achieves higher performance because the localization effect in the framework. In experimental results of 3D methods. We observe low performance due to the severely downsampled volume of CT images in 3D UNet (Cicek et al., 2016), V-Net (Milletari et al., 2016), and 3D FCN (Chen et al., 2016). While nnUNet (Isensee et al., 2018) benefits from the cascaded framework that incorporates many ensembled predictions in the outputs. In comparison of 2D/3D hybrid networks, former methods achieve comparable results, these studies utilized both 2D and 3D context in a single network. Compare with above current state-of-the-art medical image segmentation methods, coarse-to-fine frameworks achieve consistent higher DSC in the task, probably due to the effective combination of low-resolution context and high-resolution contexts.

4.3.5. Comparison with Multi-Atlas abdomen labeling challenge leaderboard

The result of top teams on the leaderboard are listed in Table 7. Note that “try” team leads the top several rankings. It is also noticeable that some latest work achieved high performance such as Zhou et al. (2019) that achieves mean DSC of 0.850 are not reported on the leaderboard. Compare with the leaderboard, our method outperforms other state-of-the-art methods, and achieve the highest mean DSC and the best Hausdorff distance performance in the standard competition.

Table 6
Evaluation of different medical image segmentation methods on the BTCV testing dataset in multi-organ segmentation (12 organs). The evaluation is performed in terms of mean DSC and across all organs.

2D methods		3D methodsVein Pan AG		Hybrid and 2.5D methods	
2D UNet (Ronneberger et al., 2015) Resnet	0.4935	3D UNet (Cicek et al. 2016)	0.5381	H-denseUNet (Li et al., 2018)	0.8172
ResNet (He et al., 2016)	0.5328	V–Net (Milletari et al., 2016)	0.5284	AH–Net. (Liu et al., 2018)	0.7947
Mask R–CNN (He et al., 2017)	0.7032	3D FCN (Chen et al., 2016)	0.5406	UMCT (Xia et al. 2018)	0.7984
DeepLab V3 (Chen et al., 2018)	0.8015	nnUNet (Isensee et al., 2018)	0.7934	OAN-RC (Wang et al. 2019)	0.7885
Ours ($N = 25$)	0.8492				
Ours ($N = 50$)	0.8564				

Table 7
Leaderboard of Multi-Atlas abdomen labeling challenge (mean).

Team	Mean Surface DistanceVein Pan AG	Average Dice	Hausdorff Distance
Try-1	1.3522	0.84056	20.3802
Try-2	1.4088	0.83626	20.0736
Path	2.9252	0.777832	32.6082
Ours	1.4237	0.85641	18.9862

5. Conclusion and discussion

In the work, we revisit the challenging whole volume based 3D abdominal segmentation. Due to limitations in low-resolution, high-resolution and hierarchy approaches under restricted GPU memory, we explored the usage of randomly selected patches to the hierarchical method. First, we provided a 3D coarse multi-organ segmentation using 3D U-Net. Then, we implemented the random sampling to crop the context around target for removing fixed pattern in patches. Next, we trained a high resolution fine-tuning network to compensate the shape and boundary structure for patches. Last, we employed majority vote mechanism to fuse the full segmentation mask for CT scan. Moreover, we conducted exhaustive experiments on comparison of different strategies of patch-based methods, we demonstrated that translational data augmentation and random sampling both provided boosted performance in comparison to simply adding more patches in 3D CNN. Additionally, we did large scale of experiments on effect of average covered patches per voxel, which we conclude that more patches generously perform better than less number of patches. However, majority vote works differently on variant structures, too much patches only increase the computational time in some anatomies. Besides, we deployed the trained models on two unseen datasets, we show that the method can be generalized to outlier cases and pathological cohort.

In this study, the proposed random patch network fusion enables the training to address the memory issue for high dimensional 3D abdominal segmentation. In this study, 50 random patches are used for segmenting variant isotropic resolution at ~ $[0.8 \times 0.8 \times 2]$ for CT scans. It could be possible that GPU memory would be large enough for housing entire abdomen CT in the future. However, the local-global feature trade-off games would still exist. Herein, the random patch network fusion technique could be a good choice for such scenarios.

The major limitation of the proposed method is that the computational time would be linearly accumulated with the increasing number of random patches. Besides, the majority vote algorithm is not time efficient when applied to voxel-wise voting. Another disadvantage of majority vote is that when voxel is rarely covered by voters, the voxel is vulnerable to be miss-labeled. Therefore, it's appealing to investigate better statistical fusion algorithm with more efficient time and space complexity, while perverse stability for removing outlier labels.

Another limitation in this work is that the hierarchical labeling framework still failed to mimic doctor's process for identifying structures. In the future, the hierarchical labeling could be investigated in clinical inspired approaches. Instead of simply transferring low-resolution feature to high resolution model, we could also pass anatomies' features to next hierarchy. For example, radiologist would first find portal and splenic vein before identifying pancreas. If we could transfer the correlated feature as a prior in different levels' hierarchy, the performance would be potentially improved. In addition, the previous work (Ni et al., 2019) indicated a high-dimensional data often suffer redundancy (e.g., not every voxel in 3D volume is useful). Mining the boundary of organ versus backgrounds or other tissues could leads to a more efficient model. We posit that it is worth additional study of the patch selection strategy around boundaries. Future work could study the efficacy of boundary patches and inner voxel patches.

In summary, the proposed random patch network fusion achieved consistent superior segmentation performance compare with other labeling frameworks, since it led to a better balance between performance and computational cost compared to other patch-based and multi-stage approaches. The balanced configurations are fulfilled by introducing 1) two-stage hierarchical levels, 2) randomly localized patches, 3) network label fusion. Our method presented positive result of 3D abdominal segmentation in variety of structures and datasets. We hope random patch network fusion will be useful with other context tasks that involve hierarchical labeling design.

Author statement

Yucheng Tang: Conceptualization, Methodology, Experiment, Software, writing, revision
Riqiang Gao: Code review, writing, data inspection.
Ho Hin Lee: Code review, Experiment, data inspection.
Shizhong Han: Conceptualization, Visualization, Investigation.
Yunqiang Chen: Conceptualization, Supervision.
Dashan Gao: Conceptualization, Software, Validation.
Vishwesh Nath: Writing- Reviewing and Editing.
Camilo Bermudez: Writing- Reviewing and Editing.
Michael R. Savona: Writing- Reviewing and Editing.
Richard G. Abramson: Writing- Reviewing and Editing.
Shunxing Bao: Writing- Reviewing and Editing.
Ilwoo Lyu: Writing- Reviewing and Editing.
Yuankai Huo: Methodology, Writing- Reviewing and Editing.
Bennett A. Landman: Conceptualization, Methodology, software, Writing- Reviewing and Editing

Declaration of Competing Interest

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

Acknowledgements

This research is supported by Vanderbilt-12Sigma Research Grant, NSF CAREER 1452485, NIH 1R01EB017230 (Landman). This study was in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. The imaging dataset(s) used for the analysis described were obtained from ImageVU, a research resource supported by the VICTR CTSA award (ULTR000445 from NCATS/NIH). Clinical trial was supported by TG Therapeutics.

References

Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. Neuroimage 46 (3), 726–738 1053-8119.

Asman, A.J., Landman, B.A., 2014. Hierarchical performance estimation in the statistical label fusion framework. Med. Image Anal. 18 (7), 1070–1081 ISSN: 1361-8415.

Asman, A.J., Huo, Y., Plassard, A.J., Landman, B.A., 2015. Multi-atlas learner fusion: an efficient segmentation approach for large-scale data. Med. Image Anal. 26 (1), 82–91 ISSN: 1361-8415.

Asman, A.J., Landman, B.A., 2013. Non-local statistical label fusion for multi-atlas segmentation. Med. Image Anal. 17 (2), 194–208 ISSN: 1361-8415.

Bai, W., Shi, W., O'regan, D.P., Tong, T., Wang, H., Jamil-Copley, S., Peters, N.S., Rueckert, D.A., 2013. p.robabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images. IEEE Trans. Med. Imaging 32 (7), 1302–1315 0278-0062.

Cheheb, I., Al-Maadeed, N., Al-Madeed, S., Bouridane, A., Jiang, R., 2017. Random sampling for patch-based face recognition. In: 2017 5th International Workshop on Biometrics and Forensics (IWBF), pp. 1–5 ISBN: 1509057919..

Chen, J., Yang, L., Zhang, Y., Alber, M., Chen, D.Z., 2016. Combining Fully Convolutional and Recurrent Neural Networks for 3d Biomedical Image Segmentation. Advances in neural information processing systems, pp. 3036–3044.

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

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O, 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Conference Name: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 424–432.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L, 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 54 (2), 940–954 ISSN: 1053-8119.

Coupé, P., Mansencal, B., Clément, M., Giraud, R., de Senneville, B.D., Ta, V.T., Lepetit, V., Manjon, J.V. AssemblyNet: a novel deep decision-making process for whole brain MRI segmentation. arXiv preprint arXiv:1906.01862. (2019).

de Brebisson, A., Montana, G., 2015. Deep neural networks for anatomical brain segmentation. In: Conference Name: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 20–28.

Ding, J., Chen, B., Liu, H., Huang, M., 2016. Convolutional neural network with data augmentation for SAR target recognition. In: IEEE Geoscience and Remote Sensing Letters, 13, pp. 364–368 ISSN: 1545-598X.

Eskildsen, S.F., Coupé, P., Fonov, V., Manjón, J.V., Leung, K.K., Guizard, N., Wassef, S.N., Østergaard, L.R., Collins, D.L., 2012. BEaST: brain extraction based on nonlocal segmentation technique. NeuroImage 59 (3), 2362–2373 ISSN: 1053-8119.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision.

Huo, Y., Xu, Z., Moon, H., Bao, S., Assad, A., Moyo, T.K., Savona, M.R., Abramson, R.G., Landman, B.A., 2018. Synseg-net: synthetic segmentation without target modality ground truth. IEEE Trans. Med. Imaging 38 (4), 1016–1025 ISSN: 0278-0062.

Huo, Y., Xu, Z., Xiong, Y., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S.M., Cutting, L.E., Landman, B.A, 2019. 3d whole brain segmentation using spatially localized atlas network tiles. Neuroimage 194, 105–119 ISSN: 1053-8119.

Isensee, Fabian, Petersen, Jens, Klein, Andre, Zimmerer, David, Jaeger, Paul F, Kohl, Simon, Wasserthal, Jakob, Koehler, Gregor Norajitra, Tobias Wirkert, Sebastian. nnu-net: self-adapting framework for U-net-based medical image segmentation. arXiv preprint arXiv:1809.10486. (2018).

Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., Glocker, B., 2015. Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. Ischemic Stroke Lesion Segmentation 13, 46.

Lai, M. Deep learning for medical image segmentation. arXiv preprint arXiv:1505.02000. (2015).

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-.W., Heng, P.-.A., 2018. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. In: IEEE Trans. Med. Imaging, 37, pp. 2663–2674.

Liang, L., Liu, Ce., Xu, Ying-Qing., Guo, B., Shum, H., 2001. Real-time texture synthesis by patch-based sampling. ACM Trans. Graph. (ToG) 20 (3), 127–150 ISSN: 0730-0301.

Liu, J., Huo, Y., Xu, Z., Assad, A., Abramson, R.G., Landman, B.A., 2017. Multi-atlas spleen segmentation on CT using adaptive context learning. Int. Soc. Opt. Photon. 10133.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., pp. 3431–3440.

Liu, Siqi, Xu, Daguang, Zhou, S Kevin, Pauly, Olivier, Grbic, Sasa, Mertelmeier, Thomas, Wicklein, Julia, Jerebko, Anna, Cai, Weidong, Comaniciu, Dorin, 2018. 3d anisotropic hybrid network: transferring convolutional features from 2d images to 3d anisotropic volumes. International Conference on Medical Image Computing and Computer-Assisted Intervention.

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

Moeskops, P., Wolterink, J.M., van der Velden, B., Gilhuijs, K.G., Leiner, T., Viergever, M.A., Išgum, I., 2016. Deep learning for multi-task medical image segmentation in multiple modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 478–486.

Ni, Tianwei, Xie, Lingxi, Zheng, Huangjie, Fishman, Elliot K, Yuille, Alan L., 2019. Elastic boundary projection for 3D medical image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2109–2118.

Olivier, R., Cao, H. Nearest neighbor value interpolation. arXiv preprint arXiv:1211.1768. (2012).

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.

Roth, H.R., Lu, L., Farag, A., Shin, H., 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, pp. 556–564.

Roth, H.R., Oda, H., Hayashi, Y., Oda, M., Shimizu, N., Fujiwara, M., Misawa, K., Mori, K. Hierarchical 3D fully convolutional networks for multi-organ segmentation. arXiv preprint arXiv:1704.06382. (2017).

Roth, H.R., Shen, C., Oda, H., Sugino, T., Oda, M., Hayashi, Y., Misawa, K., Mori, K., 2018. A multi-scale pyramid of 3d fully convolutional networks for abdominal multi-organ segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 417–425.

Rousseeuw, P.J., Leroy, A.M, 2005. Robust Regression and Outlier Detection. Publisher: John wiley & sons, p. 589 ISBN: 0471725374.

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. f-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. 54, 30–44 ISSN: 1361-8415.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International Conference on Information Processing in Medical Imaging., pp. 146–157.

Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. Journal: arXiv preprint arXiv:1902.09063. (2019).

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J., 2017. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. Book: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 240–248.

Tang, Y., Huo, Y., Xiong, Y., Moon, H., Assad, A., Moyo, T.K., Savona, M.R., Abramson, R., Landman, B.A., 2019. Improving splenomegaly segmentation by learning from heterogeneous multi-source labels. Soc. Opt. Photon. Med. Imaging 10949, 1094908.

Urban, G., Bendszus, M., Hamprecht, F., Kleesiek, J, 2014. Multi-modal brain tumor segmentation using deep convolutional neural networks. MICCAI BraTS (Brain Tumor Segmentation) Challenge. In: Proceedings, Winning Contribution, pp. 31–35.

Wang, H., Suh, J.W., Das, S., R., Pluta, J., B., Craige, C., Yushkevich, P.A., 2012. Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. 35 (3), 611–623 0162-8828.

Wang, H., Yushkevich, P., 2013. Multi-atlas segmentation with joint label fusion and corrective learning—an open source implementation. J.: Front. Neuroinform. 7, 27 ISSN: 1662-5196.

Wang, C., Zhao, Z., Ren, Q., Xu, Y., Yu, Yi, 2019a. Dense U-net based on patch-based learning for retinal vessel segmentation. Entropy 21 (2), 168.

Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L., 2019b. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Med. Image Anal. 55, 88–102 ISSN: 1361-8415.

Wolz, R., Chu, C., Misawa, K., Fujiwara, M., Mori, K., Rueckert, D., 2019. Automated abdominal multi-organ segmentation with subject-specific atlas generation. IEEE Trans. Med. Imaging 32 (9), 1723–1730 0278-0062.

Xu, Z., Li, Bo., Panda, S., Asman, A.J., Merkle, K.L., Shanahan, P.L., Abramson, R.G., Landman, B.A., 2014. Shape-constrained multi-atlas segmentation of spleen in CT. Int. Soc. Opt. Photon. Med. Imaging 9034, 903446.

Xu, Z., Gertz, A.L., Burke, R.P., Bansal, N., Kang, H., Landman, B.A., Abramson, R.G., 2016. Improving spleen volume estimation via computer-assisted segmentation on clinically acquired CT scans. Acad. Radiol. 23 (10), 1214–1220 1076-6332.

Yan, K., Lu, L., Summers, R.M., 2018. Unsupervised body part regression via spatially self-ordering convolutional neural networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1022–1025 ISBN: 1538636360.

Yang, W., Gao, R., Xu, Y., Sun, X., Liao, Q., 2016. Discriminative patch-based sparse representation for face recognition. In: IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), pp. 1–4 ISBN: 1509027084.

Zhang, D., Guo, Q., Wu, G., Shen, D., 2012. Sparse patch-based label fusion for multi-atlas segmentation. In: Conference Name: International Workshop on Multimodal Brain Image Analysis, pp. 94–102.

Zhang, Z., yang, L., Zheng, Y., 2018. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., pp. 9242–9251.

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. International Conference on Medical Image Computing and Computer-Assisted Intervention.

Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E., Yuille, A. Prior-aware neural network for partially-supervised multi-organ segmentation. arXiv preprint arXiv:1904.06346 . (2019).

Zhu, Z., Xia, Y., Shen, W., Fishman, E., Yuille, A., 2018. A 3D coarse-to-fine framework for volumetric medical image segmentation. International Conference on 3D Vision (3DV).