Efficient and Robust Instrument Segmentation in 3D Ultrasound Using Patch-of-Interest-FuseNet with Hybrid Loss

Hongxu Yang\textsuperscript{a,}, Caifeng Shan\textsuperscript{b}, Arthur Bouwman\textsuperscript{d}, Alexander F. Kolen\textsuperscript{c}, Peter H.N. de With\textsuperscript{a}

\textsuperscript{a} Eindhoven University of Technology, Eindhoven, the Netherlands
\textsuperscript{b} Shandong University of Science and Technology, Qingdao, China
\textsuperscript{c} Philips Research, Eindhoven, the Netherlands
\textsuperscript{d} Catharina Hospital, Eindhoven, the Netherlands

\textbf{Abstract}

Instrument segmentation plays a vital role in 3D ultrasound (US) guided cardiac intervention. Efficient and accurate segmentation during the operation is highly desired since it can facilitate the operation, reduce the operational complexity, and therefore improve the outcome. Nevertheless, current image-based instrument segmentation methods are not efficient nor accurate enough for clinical usage. Lately, fully convolutional neural networks (FCNs), including 2D and 3D FCNs, have been used in different volumetric segmentation tasks. However, 2D FCN cannot exploit the 3D contextual information in the volumetric data, while 3D FCN requires high computation cost and a large amount of training data. Moreover, with limited computation resources, 3D FCN is commonly applied with a patch-based strategy, which is therefore not efficient for clinical applications. To address these, we propose a POI-FuseNet, which consists of a patch-of-interest (POI) selector and a FuseNet. The POI selector can efficiently select the interested regions containing the instrument, while FuseNet can make use of 2D and 3D FCN features to hierarchically exploit contextual information. Furthermore, we propose a hybrid loss function, which consists of a contextual loss and a class-balanced focal loss, to improve the segmentation performance of the network. With the collected challenging ex-vivo dataset on RF-ablation catheter, our method achieved a Dice score of 70.5\%, superior to the state-of-the-art methods. In addition, based on the pre-trained model from ex-vivo dataset, our method can be adapted to the in-vivo dataset on guidewire and achieves a Dice score of 66.5\% for a different cardiac operation. More crucially, with POI-based strategy, segmentation efficiency is reduced to around 1.3 seconds per volume, which shows the proposed method is promising for clinical use.

© 2020 The Author(s). Published by Elsevier B.V.

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

Recently, image-based instrument detection in ultrasound-guided intervention has been studied, because it is radiation-free for both patients and surgeons. Ultrasound (US) provides a better characterization between instrument and tissue than conventional fluoroscopy, also avoiding contrast agents that may damage organs. Moreover, 3D US provides a better spatial description than traditional 2D X-ray and offers more clear visualization of the spatial relationships for a sonographer. Nevertheless, manually segmenting an instrument in the 3D US is always challenging and difficult because of the complex anatomical structure visualized in the US and the low-resolution imaging of the US in general. Therefore, to guide the intervention, an efficient and accurate instrument segmentation in the 3D US is highly desired for clinical practice.

1.1. Related Work

The commonly studied instrument detection methods can be classified into external device-based methods and image-based methods. By comparing these two approaches, device-based methods are not widely adopted in practice, such as active sensing (Guo et al., 2014) or robotic-based guidance (Nadeau et al., 2014). This is due to their requirement for extra equipment in the operation, extra cost for hospital and extra approval for manufacturing companies. Alternatively, image-based methods provide a more
convenient solution in 3D US with limited constraints in practice. Although, image-based instrument detection or segmentation in 3D US has been studied during past years, the studies in this area are still limited. From the view of methodology, these works can be classified into two categories: non-learning-based methods and learning-based methods.

Non-learning-based methods: Before the popularity of machine learning in computer vision area, traditional computer vision technologies were applied on 3D US to detect medical instrument by analyzing geometry and intensity properties of the instruments, such as shapes, intensity distribution, etc. Novotny et al. (2003) proposed to apply Principal Component Analysis (PCA) on a thresholded 3D US volumetric data, which was derived by applying cluster analysis to select the most possible region as the detected instrument. Barva et al. (2008) applied Radon Transformation on 3D US volume to accumulate intensity value, which was used to localize the straight electrode in 3D US. Similar to Radon Transformation, Zhou et al. (2007) proposed to detect the needle by line description in 3D space by 3D Random Hough Transformation. With a more advanced spatial and instrument model description in 3D space, Zhao et al. (2013) proposed to apply line filter in 3D US, which can roughly filter out needle like structure in 3D images. Then the 3D RANdom SAmple Consensus (RANSAC) algorithm is applied to select the most confidential region as the target instrument. At the same period, Cao et al. (2013) proposed to apply template matching on 3D US volume to detect the catheter with complex post-processing, which achieved successful detection results with a strong assumption of catheter direction in the images. In the above approaches, RANSAC with line filtering achieved a more promising performance, which is because of a better thresholding for 3D US and efficient model description by RANSAC. However, the above approaches still have limitations like: (1) With limited discriminating information by thresholding method, it is hard to extract accurate 3D regions for instrument detection. (2) Most of the above methods were validated on simulated 3D images or phantom data, which are significantly different than real clinical applications. (3) Strong assumption of instrument shape and direction, which leads to an unsatisfied generalization of their proposed methods. Therefore the above methods were not fully make use of instrument-related information.

Learning-based methods: This approach has been studied in recent years, which classifies voxels into category of instrument or non-instrument. Handcrafted features were proposed by considering Frangi vesselness filter (Uherčík et al., 2013), Gabor filterbank (Pourtoherian et al., 2017), time-domain statistical feature (Beigi et al., 2017) or multi-definition features (Yang et al., 2019d), which achieved reasonable instrument segmentation results albeit with complex post-processing. However, these methods are less robust or partly inefficient when US images are recorded from a complex anatomical environment, due to the voxel-based processing. Recently, deep learning, such as convolutional neural networks (CNNs) or fully convolutional neural networks (FCNs), had been intensively studied and applied for medical imaging-related areas (Litjens et al., 2017). CNNs are applied as a classifier to distinguish the category of the voxels in the 3D US, which are used to segment medical instruments in the 3D US image. A voxel-of-interest-based CNN pipeline (Yang et al., 2019c) was proposed to segment the instrument for cardiac intervention. The Frangi vesselness filter (Frangi et al., 1998) is firstly applied to select the the possible voxels belonging to the instrument globally, and then a CNN is subsequently applied to classify the remaining voxels. This method avoids iterative voxel prediction on the full volume and achieves an inference time of 10 seconds on the average per volume. However, this efficiency is still far from real-time clinical application. Later, slice-based semantic segmentation was applied on 3D US images to segment the instrument efficiently (Pourtoherian et al., 2018; Yang et al., 2019a). Nevertheless, this 2D approach has limited performance due to the slice-based strategy, which hampers the 3D information usage. Alternatively, patch-based 2.5D (Yang et al., 2019b) and 3D (Yang et al., 2019e) semantic segmentation methods were proposed to segment the instrument in 3D US. Similar to voxel-based methods, straightforward iterative patch-based prediction on a full volume requires considerable computation time, which is not attractive for real-time applications (typically spend more than 10 seconds per volume). Furthermore, the segmentation performances (Yang et al., 2019b; 2019e) are not optimized because of their limited information usage by a single network design with limited training samples. At the same time, Arif et al. (2019) proposed to directly apply a drastically simplified 3D UNet on liver US volume to detect the needle. Although their method achieved promising results in the application of 3D liver US, its stability and accuracy are still need to be exploited in a more challenging dataset and recording conditions, such as cardiac ultrasound. There are some studies on instrument detection or segmentation in 2D US (Hachhalilouglu et al., 2015; Hatt et al., 2015; Mwikirize et al., 2019), but they are focusing on in-plane instrument, being different than the 3D US-based tasks. By comparing the literatures in the instrument detection, non-learning-based methods are usually worse than the learning-based method due to limited model description and challenging US images.

1.2. Proposed Method

To address above limitations in segmentation efficiency and accuracy, we propose a patch-of-interest-based FuseNet (POI-FuseNet) method for instrument segmentation in the 3D US. The proposed method consists of two main steps with coarse-to-fine pipeline, which is shown in Fig. 1: (1) Slice-based UNet for coarse segmentation and POI selection, and (2) FuseNet for a finer segmentation on the selected patches. The first step applies a Slice-based UNet on the 3D US image for a fast segmentation, which has a better architecture than the conventional slice-based segmentation strategy. The purpose of this network is to roughly segment the instrument in 3D US and select the potential regions as the input for the second step in an efficient way. As a consequence, the number of patches to be segmented by the second stage FuseNet can be drastically reduced for a faster execution than conventional exhaustive approaches. More specifically, to accelerate the prediction efficiency, we propose a skipping slicing strategy with spatial downsampling in 3D US, which therefore reduces the complexity of the slice-based prediction. As a result, the full-volume prediction time is significantly reduced compared to conventional approaches (Pourtoherian et al., 2018; Yang et al., 2019a). With the coarsely segmented instrument in 3D US, patches overlapped with the prediction result are extracted for the next step. The second stage is based on a patch-based FuseNet, which combines two individual networks, Direction-fused UNet and 3D Pyramid UNet, by feature level fusion, to better exploit 3D contextual information than the existing 2.5D or 3D methods. The Direction-fused UNet is constructed based on 2D UNet with direction-based feature operations, which can mainly exploit spatial information within 2D slices of 3D patch. Considering it cannot fully make use of the information of the input patch due to the 2D convolutions in 3D space, a 3D Pyramid UNet is introduced in parallel to exploit the 3D information within the 3D patch at higher information level. The two networks are complementary to each other. With the feature fusion, the proposed FuseNet can better explore the information in the patch for instrument segmentation. To better supervise the FuseNet, a hybrid loss function is introduced to allow the network to learn the pixel-level and image-level discriminat-
ing information simultaneously. With collected ex-vivo dataset for RF-ablation catheter under challenging conditions, we have thoroughly validated the proposed method for instrument segmentation and achieved the Dice score of 70.5%, which is better than the state-of-the-art methods by a large margin of at least 10% improvement. More specifically, the proposed FuseNet achieves around 3% higher DSC score than the backbone UNet. Meanwhile, the proposed hybrid loss further improves performance with 3% DSC. Furthermore, we conducted experiments on a in-vivo dataset collected for guidewire segmentation (in TAVI operation). Considering the limited images of in-vivo dataset, we performed fine-tuning on the pre-trained model from the ex-vivo dataset, which achieved the Dice score of 66.5% and is better than directly training from scratch. More crucially, the prediction time is reduced from around 40 seconds to around 1.3 seconds per volume, which shows a promising efficiency for clinical use.

Our contributions are summarized as threefold. First, we propose an efficient POI-based pipeline for the instrument segmentation task, to reduce the computational cost for a fine-grained patch-based segmentation. The pipeline adopts an efficient slice-based UNet, to coarsely select the POI. Second, a FuseNet is proposed based on Direction-fused UNet and 3D Pyramid UNet, which exploits inter-and intra-slice information with 3D contextual information for a better segmentation performance. Third, we propose a novel hybrid loss function, which allows the network to learn the pixel-level and image-level discriminating information simultaneously. Compared to our preliminary conference paper at MICCAI 2019 (Yang et al., 2019e), we have extended the segmentation pipeline by introducing a pre-processing network as the POI selection, which reduces operation cost of full 3D patch-based segmentation. With the coarse-to-fine strategy, the instrument segmentation efficiency is around 10 times faster than our original method. Moreover, we proposed a more advanced 3D segmentation network for patch-based segmentation task, which combines 2.5D and 3D information simultaneously and therefore achieves more robust results. With extended literature review on medical instrument detection in 3D US, we have conducted comprehensive experiments in this paper where 60% experiments are new with respect to the MICCAI paper. The proposed method is described in Section 2. The dataset and implementation details are shown in Section 3. Experimental results are presented in Section 4. Finally, discussion and conclusion are given in Section 5.

2. Our Methods

The proposed approach includes two-stage coarse-to-fine segmentation. The overview of the proposed POI-FuseNet is shown in Fig. 1. The input 3D US volume is first processed by a Slice-based UNet, to efficiently and coarsely segment the instrument in the 3D volume. The volume is divided into small patches without overlapping and the POIs, i.e., the patches overlapped with the coarse segmentation, are used as the input for the second-stage FuseNet segmentation. More details are discussed as follows.

2.1. Slice-based UNet for Coarse Segmentation and POI Selection

When applying 3D UNet to the whole image for ROI-based feature extraction and segmentation (Huang et al., 2018), the key challenge is the limited GPU memory for complex 3D operations. Besides, the instrument has a large variance in length and location inside the 3D space, which is typically ranging from 9 to 100 voxels. As a consequence, it is challenging to apply the feature-map-based ROI in our task, which is designed for colorectal tumor segmentation (Huang et al., 2018). Alternatively, when applying the patch-based segmentation, the two-stage coarse-to-fine strategy is commonly adopted to avoid exhaustive segmentation in 3D space. Therefore, an efficient patch-of-interest selection is needed for the 3D patch-based network. The 2D slice-based UNet was originally proposed to segment the instrument (Pouraherian et al., 2018), which is however processes limited 3D spatial information and obtains worse performance than 3D UNets. Moreover, the iterative slice-by-slice prediction leads to computation redundancy and reduces time efficiency, especially when considering this approach as a pre-processing method to extract POIs. With these considerations, we propose a Slice-based UNet with downsampled prediction using spatial skipping slice extraction to improve the prediction efficiency and improve a coarse segmentation result.

The pipeline for the full-volume prediction is shown in Fig. 2. Considering a simple example, the input volume with the size of $M^3$ voxels is decomposed into 2D slices along the principal direction, which is denoted as set $\mathcal{M}$. This set is iteratively downsampled on the slice basis to a ratio of $K$, and called $\mathcal{M}_K$. For each 2D plane in $\mathcal{M}_K$, by extracting its two adjacent planes with spatial gap $d_1$ from the original set $\mathcal{M}$, a 3-channel image is constituted and used as the input for the Slice-based UNet. The Slice-based UNet is constructed based on the VGG16 encoder (Simonyan and Zisserman, 2014), since it was proven to be a successful backbone for instrument segmentation task in US images (Pouraherian et al., 2018; Yang et al., 2019a). As is shown in Fig. 2, Slice-based UNet has convolutional layers with 5-level Maxpooling. After the last Maxpooling layer, the subsequent convolutional layers have kernel numbers $1024$, $1024$, and $2$. After those layers, 4 deconvolutional layers are following, which all have equal kernel sizes of $2 \times 2$. Moreover, for each deconvolutional layer, an additional convolution operation is added to improve stability. To exploit more discriminating information at different scales, skipping connections are considered to construct the UNet structure. To further improve the performance of UNet, deep supervision (Dou et al., 2016) is employed at different feature scales at the decoder. With the proposed spatial downsample slicing strategy, the output volume is obtained with a faster prediction for coarse segmentation. To address the challenging case that the instrument crossing the slices with small footprint, an orthogonal slicing strategy along elevation and lateral direction is adopted in the coarse segmentation (Pouraherian et al., 2018; Yang et al., 2019a) as the complementary to the spatial stride $d_1$. Because more spatial information of the instrument can be observed, this is better than...
the case where only applying single direction slicing during the training.

As a consequence of deconvolution operation and input set \( \mathcal{M}_k \), the output prediction is downsampled to a ratio of \( K \) when compared to the original input image in each dimension. The downsampled prediction is upscaled to its original size by interpolation. Thresholding and connectivity analysis are applied to select the two largest connected components as the POI volume for patch extraction. Given the input image, the 3D volumetric data is divided into small non-overlapping patches with size \( (D + S)^3 \) voxels. By comparing the coarse segmentation and pre-allocated patches, patches containing coarse predictions are extracted as the input for the FuseNet. It is worth to mention that because of the zero-padding in convolution operations for FuseNet, a \( (D + S)^3 \)-voxels patch is actually extracted based on the patch above, where the \( S \) is the compensation for information leakage at the patch boundary, which is depicted in Fig. 3. In this paper, we select \( D = 32 \) and \( S = 16 \) based on experimental results, which yields a balance between efficiency and accuracy.

### 2.2. **Patch-based FuseNet for Fine Segmentation**

In this section, a novel FuseNet for 3D segmentation is proposed, which is an extended segmentation network based on our previous work (Yang et al., 2019b; 2019e), but with more advanced network structure and design. The overview of the proposed FuseNet is shown in Fig. 4. The methods in our previous studies only considered a simple and straightforward network design. As a result, 3D spatial information cannot be fully exploited in the limited dataset. Alternatively, the proposed FuseNet consists of two individual UNet variants with different spatial operations for 3D patch segmentation: a semi-3D Direction-fused UNet and a full 3D Pyramid-UNet, which are shown in Fig. 5 and Fig. 6, respectively. Intuitively, the DF-UNet exploits intra-slice information by using 2D feature extractor and utilizes the inter-slice information with high-level tensor operations. Moreover, the pre-trained parameters could make use of prior-knowledge from nature images to boost the segmentation performance. Nevertheless, this network cannot fully analyze 3D contextual information due to its 2D feature extractor. In contrast, the 3D Pyramid-UNet exploits the 3D contextual information in a more straightforward way. Meanwhile, it also preserves more low-level information than the standard UNet. However, this Pyramid-UNet may not be properly trained with the limited data. More crucially, it is challenged by higher GPU memory cost and limited field-of-view (compared to DF-UNet). To address these limitations, the FuseNet is proposed with high-level feature fusion. By doing so and with the end-to-end jointly training, the high-level representative features from two branches are jointly optimized for a better segmentation. Nevertheless, the FuseNet comes with the complex 2D-3D tensor operation, which leads to a longer predictions time than the straightforward and simplified 3D UNet. Moreover, by integrating two different networks on a single GPU, the overall complexity would be increased due to unparalleled network queue. By applying the coarse-to-fine strategy, this efficiency problem is addressed while the segmentation accuracy of the FuseNet can be preserved. As a result, the overall prediction efficiency is improved with more than 30 times acceleration. The details of the Direction-fused UNet and 3D Pyramid UNet are discussed below.

**Direction-fused UNet:** For an input patch with size \( (D + S)^3 \), it is decomposed into \( D + S \) planes along each axis. For each plane, a 3-channel image is formed based on the adjacent images of the actual plane with spatial gap \( d_2 \) voxels along the axis, which is shown in Fig. 5. As a result, the patch leads to \( D + S \) different 3-channel images in each direction (padding is applied at the boundary plane). Then, each image is processed by the 2D UNet, which is based on a VGG16 encoder and a customized decoder. The encoder is based on the VGG16 network while removing the dense
Fig. 4. Overview of FuseNet, which consists of Direction-Fused UNet, Pyramid UNet and a feature-fusion part. The feature maps from two different branches are concatenated and processed by external convolutional operations to generate the output prediction.

Fig. 5. Overview of Direction-fused UNet, which is constructed based on a 2D UNet and feature space operations. (a) The input 3D patch is decomposed into adjacent slices in three different directions simultaneously, which are processed by the 2D UNet to generate slice-based feature maps. The slice-based feature maps are then permute to form feature maps in 3D space. (b) 2.5D image with stride $d_2$ as the input for 2D UNet.
connections. Then, the output of the encoder is filtered by 3 convolutional layers with filter numbers 1024, 1024, and 2. The decoder stage includes 4 deconvolution layers with masks of 2 × 2, 2 × 2, 2 × 2 and 4 × 4 pixels, respectively. Furthermore, after each deconvolution layer, an extra convolution operation is included to smooth the features, which is to avoid checkerboard artifacts from transposed convolutional operations (Odena et al., 2016). We considered transposed layer instead of upsampling by interpolation, since we failed to obtain the network convergence with interpolation method. To reduce the cost on the GPU memory, we perform summation instead of concatenation for skipping connections, therefore the GPU memory usage is reduced for FuseNet during the end-to-end jointly training. As shown in Fig. 5, the images are processed by the proposed UNet and its output features are stacked based on the plane’s original position, to construct feature maps along three axes together with high-dimensional transposition. Furthermore, feature maps from different axes are accumulated to form a fused feature map. Finally, the final prediction of DF-UNet is obtained by applying a 3D convolution and a sigmoid layer. For our original design, the direction-based operation for DF-Unet was iteratively applied to extract features for each slice, which is time consuming and memory demanding. Alternatively, in this paper, to improve the efficiency, the slices from each direction are formed into a mini-batch, which is then simultaneously processed by 2D UNet for feature map generation. As a result, compared to our original method (Yang et al., 2019b), this new design reduces the prediction time from more than 5 minutes to around 30 seconds per volume.

**Pyramid UNet:** The Pyramid UNet is proposed based on a customized 3D UNet, which has a simpler network architecture and avoids overfitting (Yang et al., 2017). As for a feature pyramid-based network, as it goes deeper, the discriminating information at low-level vanishes and degrades the segmentation performance. Even though the UNet employs skipping connections to preserve the low-level information between encoder and decoder, it still cannot adequately exploit the information at different image scales (He et al., 2016; Lin et al., 2017a). To preserve more detailed information at different image scales, we consider to introduce pyramid input branches. The pyramid inputs at different scales preserve low-level information within the encoding stage, which potentially compensates the information vanish during the feature extraction of UNet. Furthermore, to better supervise and synchronize the features at different decoder scales, we also use deep supervision (Dou et al., 2016), but introduce an extra convolutional block for better stability. By introducing the pyramid inputs and outputs to the UNet, the proposed network potentially preserves more information at different feature scales than the standard UNet for US images. As depicted in the Fig. 6, the network has 32 kernels at the very beginning, which is gradually doubled as the network goes deeper.

**Feature Fusion:** Based on the Direction-fused UNet and the Pyramid UNet, the feature fusion is performed to combine the features from different feature extractors. As shown in Fig. 4, feature maps extracted from two networks, e.g. denoted by dark yellow squares, are concatenated prior to convolution operations, which followed by two convolutional layers with the filter number of 24 and 12. The final prediction of FuseNet comes from the feature fuse layer, which is indicated as red block in Fig. 4. Although (Li et al., 2018) proposed a H-DenseUNet, which integrates a 2D DenseNet and a 3D DenseNet for tumor segmentation in 3D CT images, their method did not consider the spatial correlation in 2D DenseNet. Therefore the spatial information usage is limited when compared to our proposed Direction-fused UNet. Moreover, they concatenated the 2D prediction with the original image as the input mask for 3D DenseNet, which unfortunately degrades the segmentation in our case.

### 2.3. Hybrid Loss for Patch-based Segmentation

To better supervise the FuseNet and to enforce it to learn more contextual information rather than a conventional voxel-based difference, such as cross-entropy and Dice loss, we propose a hybrid loss function. It includes a contextual loss (CL) and a class-balanced focal loss (FL). As for a predicted patch and corresponding ground truth, which are denoted as \( \hat{Y} \) and \( Y \), the hybrid loss function is defined as

\[
Loss(\hat{Y}, Y) = Loss_{FL}(\hat{Y}, Y) + Loss_{CL}(\hat{Y}, Y),
\]

(1)
where $\text{Loss}_{\text{fl}}$ denotes the class-balanced focal loss and $\text{Loss}_{\text{cL}}$ is the contextual loss. For each predicted output of the FuseNet, Eqn. (1) is applied with the same weight as unity, except the outputs from deep supervision in the Pyramid U-Net, which are assigned as 0.6 and 0.4 for middle and deepest layer, respectively.

Conventionally, networks are learned by employing a voxel-based loss function, such as cross-entropy or Dice loss, which ignores the high-level difference between prediction and ground truth at a global level. To allow the network to learn a better contextual representation or so-called high-level feature representation, we propose a contextual loss, which formulates the contextual difference in a high-level feature space. The prediction and ground truth are processed by a contextual encoder, which is depicted in Fig. 7, to generate a high-level feature vector in latent space, which are denoted as $S_{\gamma}$ and $S_{\gamma}$, respectively. As a consequence, the contextual loss $\text{Loss}_{\text{cL}}$ is characterized by

$$
\text{Loss}_{\text{cL}}(\hat{Y}, Y) = ||\text{CE}(\hat{Y}) - \text{CE}(Y)||_2 = ||S_{\gamma} - S_{\gamma}||_2, \tag{2}
$$

where $|| \cdot ||_2$ is the L2 distance and $\text{CE}(\cdot)$ represents the context encoder in Fig. 7.

The loss function, such as Dice or cross-entropy, is typically applied for segmentation tasks in medical imaging. However, it is not optimal when segmented objects have large size variations and imbalanced class distributions in the ground truth (Lin et al., 2017b). Moreover, when the instrument has a small size in 3D space and hard-classified boundary voxels are more critical than easily-classified voxels at the center part of the instrument, the commonly used loss functions may not be optimal, especially for POI-based task, which requires a more accurate segmentation. Therefore, we adopt the class-balanced focal loss function, which is based on binary cross-entropy and F-score loss (Isensee et al., 2019; Yang et al., 2019a). The proposed focal loss is defined as

$$
\text{Loss}_{\text{FL}}(\hat{Y}, Y) = \left(1 - \gamma \right) \left(1 + \beta^2 \right)^{N} \sum_{i=1}^{N} y_{i} \hat{y}_{i} + \beta^2 \sum_{i=1}^{N} y_{i} y_{i} \hat{y}_{i} + \sum_{i=1}^{N} y_{i} \hat{y}_{i} + \gamma \sum_{i=1}^{N} \omega_{\text{i}} \left(1 - y_{i}\right) \gamma \log(\hat{y}_{i}) \left(1 - \hat{y}_{i}\right)^{2} \gamma \log(\hat{y}_{i}), \tag{3}
$$

where $y_{i}$ denotes an instrument voxel from the ground truth, $\hat{y}_{i}$ represents the voxel’s prediction probability for the instrument class, while $y_{i}$ and $\hat{y}_{i}$ are non-instrument voxels and their corresponding prediction probability, respectively. Parameters $\beta$ and $\omega_{\text{i}}$ are controlling the weight between different classes, which are calculated as the square root of the inverse of the class ratio. Parameters $\gamma$ and $\sigma$ are controlling the slope of the loss curve, which are empirically set to $\gamma = 0.3$ and $\sigma = 2$, respectively.

3. Experiments and Implementation

3.1. Datasets and evaluation metric

**ex-vivo RF-ablation catheter dataset:** The ex-vivo dataset consists of ninety-two 3D cardiac US images from eight porcine hearts. During the recording, the hearts were placed in a water tank with an RF-ablation catheter ( Biosense Webster, NAVISTAR®DS, 7 Fr $\approx 2.3$ mm; Boston Scientific, Blazer™™ II XP,10 Fr $\approx 3.3$ mm) inside the heart chambers. The US probes were placed next to the heart to capture the images containing the instrument. Our data collection setup and example 2D B-mode images are shown in Fig. 8. The dataset includes volumes of size ranging from $120 \times 69 \times 92$ to $294 \times 283 \times 202$ voxels, in which the voxel size was isotropically resampled to the range of 0.4–0.7 mm. The datasets were manually annotated by clinical experts to generate the binary segmentation mask as the ground truth.

**in-vivo TAVI guidewire dataset:** The collected in-vivo dataset includes eighteen volumes from two TAVI operations. During the recording, the sonographer recorded images from different locations of the chamber without any influence on the procedure. The volumes were recorded with a mean volume size of $201 \times 202 \times 302$, where the volume voxel size was resampled to 0.6 mm. The applied instrument in the in-vivo dataset is a guidewire (0.889 mm). The images were manually annotated by clinical experts to generate the binary segmentation mask as the ground truth. Example images are shown in Fig. 9 (e) and (f). The study was approved by the Medical Research Ethics Committee United (MEC-U; study ID:non-WMO 2017–106). The study was classified as non-WMO by the MEC-U based on the retrospective design. Therefore obtaining an informed consent was not deemed necessary as it conforms to the Dutch Agreement on Medical Treatment Act. All data were analyzed anonymously.

**Evaluation Metrics:** To evaluate the performance of the proposed method, we consider the Dice score (DSC) and Hausdorff Distance (HD) as the evaluation metrics for different scenarios. As for DSC, it is defined as:

$$
\text{DSC} = \frac{2 \text{TP}}{2 \text{TP} + \text{FP} + \text{FN}}, \tag{4}
$$

where the TP denotes true positive, FP is false positive and FN is false negative. The Hausdorff Distance (HD) (Taha and Hanbury, 2015) is more sensitive to voxel mismatch between annotation voxels and the prediction voxels, and is defined as:

$$
\text{HD}(A, B) = \max(d(A, B), d(B, A)). \tag{5}
$$

In Eqn. (5), parameter $d(A, B)$ is the directed Hausdorff Distance, which is specified by:

$$
d(A, B) = \max_{a \in A} \min_{b \in B} ||a - b||. \tag{6}
$$

where the $A$ and $B$ denote voxel groups from the ground truth and the predicted results, respectively, and $a$ and $b$ are voxels from the corresponding groups.

As for coarse segmentation, DSC is used to evaluate the capacity of slice-based U-Net, which means that for a higher DSC value a better POI selection can be achieved with fewer outliers. As for patch-based segmentation, DSC and HD are used to measure the
network performances under different settings for the instrument segmentation task. Moreover, the average prediction time is also considered for the pipeline comparison.

3.2. Training Procedures

The proposed method has separate networks for coarse and fine segmentation tasks, thus the training procedures are described in the sequel. For ex-vivo dataset, the images were randomly divided into 62/30 volumes for training/testing. Considering the limited data in the in-vivo dataset, a 3-fold cross-validation was applied with fine-tuning, which was based on the pre-trained ex-vivo model for the RF-ablation catheter.

Training for coarse segmentation: To train the Slice-based UNet, each annotated instrument voxel in the ground truth is used as the center of the input sliced image, which introduces a translation invariance in a natural way to facilitate instrument segmentation. Non-instrument slices, i.e., slices using non-instrument voxel as the center point, are downsampled to the same size as the instrument voxels to generate some images without instrument inside. The network is initialized based on a pre-trained VGG16 encoder, and is trained by the Adam optimizer with learning rate of 0.00001 using the Dice Loss. The ex-vivo dataset was trained based on above description while the in-vivo dataset was trained with learning rate as 0.00001 for 2000 iterations based on the pre-trained ex-vivo model. During the training, rotation, mirror, contrast transformation were applied on-the-fly to augment the images. Meanwhile, to learn case that the instrument is crossing the slices, the slice sampling is randomly applied along elevation or lateral directions, i.e. orthogonal strategy. It is worth to mention that the downsample strategy is only considered in the testing stage to accelerate the inference efficiency, since it would degrade the information usage if we apply it in the training phase.

Training for fine segmentation: The training patches were selected from instrument voxels (Yang et al., 2019e), where an instrument voxel is used as the patch center. With input patches and Adam optimizer for training, the Direction-fused UNet (DF-UNet) and the Pyramid-UNet are initially trained separately with mini-batch size 4 and 8, respectively. More specifically, the learning rate for DF-UNet is 0.0001 while it is set to 0.001 for the Pyramid-UNet. There are three epochs for each individual training. Based on the pre-trained networks, i.e., DF UNet and Pyramid UNet, the feature-fusion part is then jointly trained with a learning rate of 0.00001 for one epoch, which finally generates the feature-fuse output. The network is trained first by a standard the Dice Loss to converge. Then, the parameters are fixed to warm up the contextual encoder for 3,000 iterations, after which the whole networks are jointly trained by the proposed hybrid loss function until converge. The ex-vivo dataset was trained based on above description while the in-vivo dataset was trained with learning rate as 0.00001 for 2000 iterations based on the pre-trained ex-vivo model. During the training, rotation, mirror, contrast transformation were applied on-the-fly to augment the images.

4. Experimental Results

In this section, we thoroughly validate the proposed POI-FuseNet with respect to accuracy and efficiency. Meanwhile, several ablation studies are also considered to validate the proposed components.

4.1. Ablation Studies

Coarse Segmentation Performance of Slice-based UNet: Several performance comparisons of Slice-based UNet are conducted in this section. First, the variations of spatial gap \(d_1\) between adjacent slices are validated from 0 to 5. Second, the variations of downsampling ratio \(K\) are tested, which is assigned as 0.25, 0.5, and 1.0, respectively. The networks of ex-vivo dataset were initialized based on the pre-trained VGG16 network for ImageNet while the networks of in-vivo dataset were initialized based on their corresponding ex-vivo models. The results are summarized by barplots in Fig. 9. Meanwhile, the inference efficiency for downsampling ratio \(K = 0.25, 0.5, 1.0\) are around 0.2 sec., 0.6 sec, and 2.6 sec., respectively (hybrid calculation with CPU and GPU, the most time-consuming part is CPU-based slicing). We have observed that a larger spatial gap \(d_1\) provides a higher performance, which is because more spatial correlations are captured by the stride of slices. However, a too large stride \(d_1\) may degrade the performance due to spatial decorrelation. In terms of downsampling ratio, \(K = 0.5\) provides a better trade-off between efficiency and performance. Although \(K = 0.25\) provides a higher segmentation efficiency for 3D US, detailed spatial information is missing during the slicing, which therefore leads to unacceptable performance. As a consequence, we have experimentally selected hyper-parameters \(d_1 = 2\) and \(K = 0.5\) for the coarse segmentation method, to select the patch-of-interest for further experiments. Because the length of the instrument inside 3D US volumes varies, the number of selected patches can range from 2 to 8, which is counted by automatically matching the coarse prediction to the pre-allocated patches in 3D US images. Based on experimental statistics, the average number of selected patches from coarse segmentation is around 5, which is depicted in Fig. 1 as selected POI.
Ablation Studies of Proposed DF-UNet: The ablation studies of Direction-fused UNet (DF-UNet) on the variation of spatial stride $d_2$ between adjacent slices are validated from 0 to 5. Moreover, we also validated whether DF-UNet achieves a better performance than UNet without DF, i.e., only a single branch of three in Fig. 5 (a). The results are depicted in Fig. 10, which are trained by a standard Dice Loss. The networks of ex-vivo data were trained with an initialized pre-trained VGG16 since it provides higher performance than when trained from scratch (w/o TF). Similar to the slice-based approach, the networks of in-vivo dataset were initialized based on their corresponding ex-vivo models. From the results, DF-UNet achieves a better performance than train from scratch or only considering a single direction by using same $d_2$, which shows a more powerful ability of transferring the knowledge of pre-trained ex-vivo dataset. Moreover, with a larger spatial gap, a better 3D space information description can be achieved. As for in-vivo dataset, spatial gap $d_2$ has less influences and variance than ex-vivo dataset, which is because of less image variation for TAVI images. Furthermore, the proposed DF-UNet is consistently better than train from scratch and single direction fusion in both datasets. Based on the results, the spatial stride is experimentally selected as $d_2 = 3$ for further feature fusion.

Ablation Studies of Proposed Pyramid-UNet: Ablation studies of the proposed Pyramid-UNet are depicted in this section. Specifically, to validate the effectiveness of the components of Pyramid-UNet, following configurations are depicted. (1) Compact-UNet trained by Dice Loss (3D Pyramid-UNet without multiple input/output). (2) Atrous Spatial Pyramid Pooling (ASPPv1) with dilation rate [1, 2, 4, 8] based on the encoder of our proposed Compact-UNet under guidance of Deeplab v3+ (Chen et al., 2018), which is trained by Dice Loss. (3) With two 3D convolution layers from Compact-UNet, ASPP is directly applied on feature maps of original image size, which means the ASPP applied on low-level directly. The dilation rate is same as ASPPv1 and is trained by Dice Loss. The model is denoted as ASPPv2. (4) Our proposed 3D Pyramid-UNet trained by Dice Loss. The results are summarized in Table 1. The basic backbone Compact-UNet provides the baseline performance for the ablation studies. Based on this architecture, the pyramid input-output structure is introduced to exploit the multi-scale features, which extract more discriminating information at different image scales. As a result, the segmentation performance is improved. As shown in Table 1, the Pyramid-UNet outperforms the Compact-UNet. As depicted in Fig. 11, Compact-UNet generates a noisier feature map where more outliers and blurry boundaries are obtained compared to the Pyramid-UNet. Compared to ASPP networks, the proposed Pyramid-UNet has better performance than both ASPP structures in our case. This is because the Pyramid-UNet structure exploits richer complex feature relationship at different image scale in both input and output branches. Initially, ASPP was proposed after a complex and proper encoder for images, such as VGG encoder or ResNet (He et al., 2016). However, in our approach, the compact network encoder leads to a less complex and discriminating feature maps for ASPP, which cannot represent sufficient information for further steps. It is worth to mention that similar
to DF-UNet, in-vivo dataset obtains better performance based on the pre-trained model from ex-vivo, which improves DSC from 60% to 64.5%.

Ablation Studies of Proposed FuseNet: Moreover, ablation studies on POI-FuseNet were also performed to validate the effectiveness of its components, which are gradually introduced to construct the proposed method. (1) Direction-fuse UNet trained by Dice Loss (DF-UNet) based on above configurations. (2) Pyramid-UNet trained by Dice Loss. (3) EnsembleNet trained by Dice Loss, which is our proposed FuseNet without feature fusion layer, i.e., the averaged output from two individually trained networks without joint training. (4) FuseNet trained by Dice Loss. (5) FuseNet trained by Contextual Loss (CL) under the guidance of Dice. Specifically, we failed to obtain the result solely using CL without Dice, which shows the CL is a kind of compensation of pixel-level loss in high-level space. (6) FuseNet trained by Focal Loss (FL). (7) FuseNet trained by Hybrid Loss (HL). The results are shown in Table 2. From the results, the DF-UNet and Pyramid-UNet obtained similar performances, although with different architecture and information extraction steps. More specifically, DF-UNet decomposes the 3D information into 2D slices with tensor operations. With this intra-slice feature extraction strategy, the DF-UNet could exploit semantic information within the slices and fuse them in the semi-3D operation. Nevertheless, the DF-UNet cannot exploit 3D information due to its nature design. Furthermore, the DF-UNet is much time consuming due to the complex 2D-3D transformations in the network design. In contrast, Pyramid-UNet directly extracts the 3D information from 3D space, which exploits semantic information in a straightforward manner. However, this network may not fully exploit information with limited training samples in the complex 3D space. These two individual networks could be complementary to each other. By integrating these two networks with the feature-level fusion, the proposed FuseNet achieved better performance than each individual network. From the demonstration in Fig. 11, fused features from two subnetworks compresses backgrounds, such as tissue and chambers, while improving the confidence of the instrument related voxels. Intuitively, this feature fusion could make the network to automatically synchronize and select the most discriminating information in high-level feature space. Moreover, the end-to-end jointly training of FuseNet enhances the capacity of the semantic information usage, which uses Pyramid-UNet to help DF-UNet with a better inter-slice and intra-slice feature extraction under 3D contextual guidance. Meanwhile, a larger field-of-view and better initialized feature representation from DF-UNet could guide the Pyramid-UNet to learn the information in 3D patch. As a result, the proposed feature fusion achieves

### Table 1
Ablation studies of our proposed Pyramid-UNet, measured by the Dice Score (DSC), Hausdorff Distance (HD) and prediction time, which are expressed by their mean ± std.

<table>
<thead>
<tr>
<th>Method</th>
<th>RF-ablation Catheter ex-vivo</th>
<th>TAVI Guidewire in-vivo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSC (%) HD (voxels) Time (sec.)</td>
<td>DSC (%) HD (voxels) Time (sec.)</td>
</tr>
<tr>
<td>Compact-UNet w Dice</td>
<td>62.2 ± 20.0 13.3 ± 15.6 ~10</td>
<td>63.8 ± 9.2 9.8 ± 5.5 ~12</td>
</tr>
<tr>
<td>ASPV1 w Dice</td>
<td>63.8 ± 16.7 11.8 ± 18.6 ~11</td>
<td>63.6 ± 9.0 8.9 ± 4.6 ~13</td>
</tr>
<tr>
<td>ASPV2 w Dice</td>
<td>57.8 ± 21.7 14.8 ± 19.3 ~17</td>
<td>62.6 ± 9.5 8.5 ± 3.4 ~19</td>
</tr>
<tr>
<td>Pyramid-UNet w Dice</td>
<td>65.8 ± 18.9 11.3 ± 13.8 ~11</td>
<td>64.5 ± 8.3 8.8 ± 3.2 ~12</td>
</tr>
</tbody>
</table>

### Table 2
Ablation studies of our proposed FuseNet, measured by the Dice Score (DSC), Hausdorff Distance (HD) and prediction time, which are expressed by their mean ± std.

<table>
<thead>
<tr>
<th>Method</th>
<th>RF-ablation Catheter ex-vivo</th>
<th>TAVI Guidewire in-vivo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSC (%) HD (voxels) Time (sec.)</td>
<td>DSC (%) HD (voxels) Time (sec.)</td>
</tr>
<tr>
<td>DF-UNet w Dice</td>
<td>64.2 ± 18.4 11.2 ± 13.8 ~28</td>
<td>64.1 ± 7.9 8.2 ± 2.9 ~32</td>
</tr>
<tr>
<td>Pyramid-UNet w Dice</td>
<td>65.8 ± 18.9 11.3 ± 13.8 ~11</td>
<td>64.5 ± 8.3 8.8 ± 3.2 ~12</td>
</tr>
<tr>
<td>EnsembleNet w Dice</td>
<td>64.1 ± 18.4 11.2 ± 13.8 ~39</td>
<td>63.2 ± 8.3 8.7 ± 3.3 ~44</td>
</tr>
<tr>
<td>FuseNet w Dice</td>
<td>67.7 ± 15.9 10.1 ± 13.0 ~41</td>
<td>65.0 ± 8.3 8.3 ± 3.2 ~47</td>
</tr>
<tr>
<td>FuseNet w CL</td>
<td>68.9 ± 14.7 8.8 ± 10.2 ~41</td>
<td>65.8 ± 8.2 8.1 ± 3.0 ~47</td>
</tr>
<tr>
<td>FuseNet w FL</td>
<td>69.1 ± 11.1 9.0 ± 9.1 ~41</td>
<td>65.9 ± 8.0 8.0 ± 3.0 ~47</td>
</tr>
<tr>
<td>FuseNet w HL</td>
<td>70.5 ± 9.2 7.3 ± 3.9 ~41</td>
<td>66.5 ± 7.5 8.2 ± 2.9 ~47</td>
</tr>
</tbody>
</table>
better performance than the straightforward prediction ensemble, i.e. directly averages the predictions from two individual networks. Compared to a standard Dice loss trained FuseNet, our proposed hybrid loss can further improve the segmentation performance in both datasets. More specifically, when compared to the FuseNet with Dice loss, both contextual loss (CL) and focal loss (FL) can improve the segmentation results in different aspects. As for CL, it encourages the contextual-level consistency between the prediction and annotation in the high-level latent space. In contrast, the FL addresses the extremely imbalanced class distributions (instrument voxels are around 1% of the patch voxels) and focuses on the hard-classified voxels in 3D space, which leads to a better performance. Based on these two losses, the proposed hybrid loss (HL) outperforms the individual CL and FL, which also provides a smaller standard deviation. It is worth to mention that the HL with contextual encoder is only considered in the training stage that the extra prediction complexity is not introduced during the testing phase. By comparing ex-vivo and in-vivo datasets, the hybrid loss has more influence on ex-vivo, which is explained by a larger image variance within the dataset. As the performance of the network improving, the prediction time is also increased, which is because of the complexity of the proposed network architecture.

Ablation Studies of Proposed POI-FuseNet: The ablation studies of the proposed patch-of-interest strategy are demonstrated in Table 3. Specifically, three different K values are validated. (1) Our proposed POI-FuseNet by HL with downsample ratio 0.25 (POI-FuseNet w HL, K=0.25), (2) Our proposed POI-FuseNet by HL with downsample ratio 0.5 (POI-FuseNet w HL, K=0.5) and (3) Our proposed POI-FuseNet by HL without downsample (POI-FuseNet w HL, K=1.0). With the introduction of POI selection by coarse segmentation, the total prediction time per volume is drastically accelerated because of preventing the iteratively patch-based prediction. More specifically, as the downsampling ratio of Slice-based UNet increasing, the performance of POI-FuseNet is improved with cost of time efficiency. There is little difference in performance with K = 0.5 or K = 1.0 while the performance is degraded in K = 0.25 in ex-vivo dataset. This is because some catheters are partly missing from the slices due to larger sampling value, and therefore only part of the catheter can be segmented. As a consequence, higher K value provides a better generalization for the POI-FuseNet. Although the coarse segmentations have different performances due to K values, the final prediction performances of POI-FuseNet have not much difference. Finally, the proposed FuseNet is trained on in-vivo dataset by fine-tuning the parameters of the ex-vivo model, which achieved 3-5% higher Dice score than training from scratch. Nevertheless, the overall framework of the proposed POI-FuseNet cannot be trained in a end-to-end style, which is limited by the coarse-to-fine strategy. This limitation is due to the large memory capacity requirement for complex 3D US images and limited training datasets, when compared to the state-of-the-art object segmentation methods in the current computer vision area. As a result, the feature maps from the full image-level cannot be used for POI purpose, such as in Mask R-CNN (He et al., 2017). To further validate the complexity of the models, we measured the number of trainable parameters for Slice-based UNet, DF-UNet, Pyramid-UNet and FuseNet, and also their floating-point operations (FLOPs). The results are shown in Table 4. It can be observed that although the network-wise complexity is increased by feature fusion in the FuseNet, the overall efficiency can be drastically improved with the efficient POI pre-selection (from more than 40 seconds to around 1 second as shown in Table 3). It is worth to mention that because of 2D slicing strategy and 2D operations, the 2D Slice-based UNet obtains a much faster segmentation while it has similar model size with DF-UNet. Nevertheless, the DF-UNet costs much higher FLOPs due to complex spatial operations. Although the overall FLOPs are increased for the POI-based FuseNet, the time efficiency is improved by coarse-to-fine strategy and proper network design, since the prediction time is also related to pipeline design, mini-batch size and GPU memory. These results show our method has more potential for clinical use.

Ablation Studies of Different Patch Processing: Beside the above ablation studies in network configurations, we also validated our proposed patch processing strategy, as mentioned in Section 2.2, which considers the zero-padding during the convolutional operations. With fixed patch size, i.e., D + S = 48, where the parameter D is the non-overlapped patch size while P is the extending parameter to compensate for the information leakage. The total patch size is experimentally chosen to balance the network complexity, size of the instrument in 3D US and prediction efficiency (Yang et al., 2019e). As shown in the Table 5, we experimentally compared different combinations of D and S in terms of DSC and HD to validate the significant influence of these parameters. Based on the observation, patch with S = 0 generates a worse segmentation result w.r.t full volume cases, which is because the patch boundaries are affected by padding operations during the convolution operation. Even the skipping connections are applied to compensate the information leakage. In contrast, for setting with S = 32, it provides slightly better performance than the case of S = 16, while significantly degrading the prediction efficiency (even compared to POI-based condition, which is at least two times slower). As a conclusion, a proper combination strategy of patch-based segmentation should be considered, which provides stable results for semantic segmentation.

4.2. Performance comparison with learning-based methods

We have compared our method with the learning-based state-of-the-art instrument segmentation methods on our challenging datasets, such as Gabor feature with SVM classifier (GF-SVM)
Table 5
Segmentation performances under different combination of parameters D and S, measured by the Dice Score (DSC) and Hausdorff Distance (HD), which are expressed by their mean ± std. The (D,S) represent the combination of different value of D + S = 48. Bold values denote the setting reported in this paper.

<table>
<thead>
<tr>
<th>(D,S)</th>
<th>RF-ablation Catheter ex-vivo</th>
<th>TAVI Guidewire in-vivo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSC (%) HD (voxel) Time (sec.)</td>
<td>DSC (%) HD (voxel) Time (sec.)</td>
</tr>
<tr>
<td>(16,32)</td>
<td>70.7 ± 9.0 7.2 ± 4.2 ~ 288</td>
<td>67.0 ± 7.7 8.1 ± 3.2 ~ 443</td>
</tr>
<tr>
<td>(32,16)</td>
<td>70.5 ± 9.3 7.3 ± 3.9 ~ 41</td>
<td>66.5 ± 7.5 8.2 ± 2.9 ~ 47</td>
</tr>
<tr>
<td>(48,0)</td>
<td>65.3 ± 13.7 10.8 ± 11.1 ~ 13</td>
<td>65.2 ± 8.4 8.8 ± 3.5 ~ 15</td>
</tr>
</tbody>
</table>

Table 6
Segmentation performances for different methods, measured by the Dice Score (DSC), Hausdorff Distance (HD) and prediction time, which are expressed by their mean ± std. All the methods are validated on our datasets (~ means failed to calculate the result due to memory overflow).

<table>
<thead>
<tr>
<th>Method</th>
<th>RF-ablation Catheter ex-vivo</th>
<th>TAVI Guidewire in-vivo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSC (%) HD (voxels) Time (sec.)</td>
<td>DSC (%) HD (voxels) Time (sec.)</td>
</tr>
<tr>
<td>GF-SVM (Pour taherian et al., 2017)</td>
<td>3.3 ± 8.5 -</td>
<td>1.0 ± 1.7 -</td>
</tr>
<tr>
<td>MF-AdaB (Yang et al., 2019d)</td>
<td>36.5 ± 19.0 19.1 ± 8.5 ~ 100</td>
<td>37.6 ± 23.3 21.9 ± 18.2 ~ 100</td>
</tr>
<tr>
<td>ShareFCN (Pour taherian et al., 2018)</td>
<td>52.8 ± 21.0 15.6 ± 16.7 ~ 3</td>
<td>55.9 ± 12.1 11.6 ± 7.8 ~ 4</td>
</tr>
<tr>
<td>VOI-ConvNet (Yang et al., 2019c)</td>
<td>58.5 ± 10.7 11.5 ± 7.7 ~ 10</td>
<td>58.6 ± 7.9 11.0 ± 5.1 ~ 11</td>
</tr>
<tr>
<td>Prenatal-UNet (Yang et al., 2017)</td>
<td>24.6 ± 24.9 38.3 ± 22.3 ~ 2</td>
<td>53.2 ± 14.7 18.8 ± 11.1 ~ 2</td>
</tr>
<tr>
<td>ACNN (Oktay et al., 2017)</td>
<td>61.0 ± 18.3 14.6 ± 13.9 ~ 10</td>
<td>63.9 ± 9.7 8.5 ± 4.7 ~ 12</td>
</tr>
<tr>
<td>FuseNet</td>
<td>70.5 ± 9.3 7.3 ± 3.9 ~ 41</td>
<td>66.5 ± 7.5 8.2 ± 2.9 ~ 47</td>
</tr>
<tr>
<td>POI-FuseNet</td>
<td>70.5 ± 9.2 7.3 ± 4.1 ~ 13</td>
<td>66.0 ± 8.3 8.2 ± 2.9 ~ 1.4</td>
</tr>
</tbody>
</table>

Fig. 12. Instrument segmentation results of different methods on example US volumes. Top row: ex-vivo, bottom row: in-vivo. (a) Segmentation results displayed in 2D slices with corresponding original image, ground truth and segmentations. Green: ground truth. Red: segmentation. (b) Segmentation results displayed in 3D US volumes, where 3D US volume is blue, the ground truth of instrument is green, and the instrument segmentation is red. All the images are corresponding to above 2D B-mode images.

(Pour taherian et al., 2017), multi-definition and multi-scale feature fusion with AdaBoost classifier (MF-AdaB) (Yang et al., 2019d), orthogonal slice-based ShareFCN (Pour taherian et al., 2018), Voxel-of-interest-based ConvNet (VOI-ConvNet) for voxel-based classification (Yang et al., 2019c). Furthermore, other US-based segmentation methods, like Prenatal-UNet (Yang et al., 2017) and Compact-UNet with Anatomically Constrained (Oktay et al., 2017), i.e. ACNN, are also considered. The results are shown in Table 6, where DSC is Dice Score and HD is Hausdorff Distance. Meanwhile, some results are shown in Fig. 12, which qualitatively demonstrates the results of voxel-based classification, slice-based segmentation and patch-based segmentation on our datasets.

From the table, several observations and conclusions can be made, which we have clustered into five topics.

1. **Comparison with handcrafted features:** As for handcrafted features with voxel-based classification, the performances are worse than deep learning methods, which are due to limited 3D information representation for information description, especially for Gabor filterbank method. First, it is a single-scale feature, which means it focuses on a specific spatial resolution. However, this cannot handle our dataset with different spatial resolutions and instrument diameters. Second, the Gabor feature mainly focuses on boundary contrast at the edge in a homogeneous or semi-homogeneous background, which works
properly in needle segmentation for anesthesia by 3D US. Nevertheless, in our case, the boundary of a catheter can be blurred with a lower contrast with anatomical background, which leads to significant difficulties for using a single-scale Gabor feature with a linear support vector machine. Third, the experimental settings are clearly different. The original paper applied a leave-one-out cross-validation method with image-specific threshold to obtain the highest performance. However, in our case, we applied a dataset-level threshold, which leads to much a lower performance than the original reported value. In contrast, the multi-scale approach with multi-definition features achieves much better performance around 36% DSC. Because of the multi-scale and definition features, more instrument-related information can be described from different viewpoints. Meanwhile, the non-linear Adaptive Boosting classifier provides a non-linear decision boundary than a simpleLSVM that a higher performance is obtained. However, handicrafted feature design is strongly relying on experience and instrument model estimated, which therefore is limits the information usage from training images. As a comparison, the proposed POI-FuseNet can extract much more discriminating information by considering data-driven method with task specified network design.

(2) Comparison to VOI-ConvNet: Compared to VOI-ConvNet, which is based on a pre-filtered voxel selection, our proposed patch-based semantic segmentation method portrays a better performance. Because the VOI-ConvNet method can degrade the true positive rate by introducing an imperfect voxel selection, whereas the POI-based method can overcome this degradation. More specifically for VOI-selection, the interested voxels are obtained from the remaining voxels after a Frangi filtering, which cannot fully preserve the instrument voxels by a simple filtering, even a CNN is followed to classify the voxels’ category (Yang et al., 2019c). In contrast, the POI-based approach selects the possible regions in 3D volume using patches, which contains full instrument-related voxels for a second time segmentation by FuseNet. As a consequence, the instrument voxels are re-calculated by a more accurate ConvNet with higher accuracy.

(3) UNet structure: Compared to a more complex and generalized UNet (Yang et al., 2017), our proposed Compact-UNet achieved better performance. It is because of a smaller input size, simpler architecture, and task specified design. Although the Prenatal-UNet achieved a fast prediction time as a result of larger patch size, i.e., 64³ voxels, it is much more hard to be trained for our instrument segmentation task, since the instrument occupies a small volume in a large patch space. From the table, our POI-FuseNet obtained higher accuracy and efficiency than the Prenatal-UNet.

(4) Contextual description: When compared to ACNN, which employs anatomical constrained knowledge to formulate the contextual information, our proposed method achieves a much better performance. It is because the design of ACNN includes a fixed pre-trained shape description encoder for ground truth, which is not suitable for prediction with large location and value variation. Actually, from the design of ACNN, it is used for anatomical structure segmentation with a fixed global location and size, which is easy to be learned and described by an Auto-Encoder. However, in our case, the instrument is located at any position of the input patches. Moreover, the prediction of the input patch is ranging from 0 to 1 instead of a fixed integer value. This variation leads to the encoder in ACNN cannot perfectly represent the contextual information difference between the ground truth and prediction. As a consequence, it cannot improve performance when compared with our jointly trained approach. The jointly training procedure enables to adaptively to learn the contextual information with varying instrument location and values.

(5) Information exploitation in 2D and 3D: Patch-based semantic segmentation approaches, i.e., Pyramid-UNet and DF-UNet, achieve a higher performance than SOTAs because of richer spatial information usage. Moreover, DF-UNet obtains better performance than Compact-UNet, since the parameters are initialized from the pre-trained model, which is shown in ablation studies. With extracted feature maps from two independent networks, FuseNet obtains more accurate segmentation results. It is explained by better hierarchical exploitation of contextual information among voxels.

From qualitative illustrates in Fig. 12, voxel-based classification method, i.e. MF-AdaB and VOI-ConvNet, generates non-smooth surface, which comes from voxel-by-voxel classification with limited field-of-view. As for slice-based segmentation, i.e. ShareFCN, it generates a smoother surface and boundary, which is because of semantic information usage. However, when compared to the proposed patch-based segmentation, i.e. the proposed method, the slice-based method has worse performance due to limited degraded spatial information than 3D space.

In terms of the prediction efficiency, our proposed POI-FuseNet achieved prediction efficiency of ~1.3 seconds per volume, which includes ~0.6 sec. POI preprocessing and 0.7 sec. patch-based finer segmentation. Because the patch-based segmentation re-calculates the voxels category, our proposed method does not hamper the segmentation accuracy when compared to the VOI-based CNN (Yang et al., 2019c). From our experiments, voxel-based methods, such as GF-SVM, MF-AdaB, or LateCNN, consumed more than 100 seconds, which is due to the iteratively voxel-by-voxel calculations. As for patch-based methods on full volume, they take around 10-50 seconds to obtain the final prediction (based on the architecture of the networks), which is still far from real-time implementations. In contrast, our proposed POI-FuseNet keeps the segmentation accuracy by the re-calculation of voxels, but also improve the segmentation efficiency. All the GPU-based methods are measured on GeForce GTX 1080Ti on Python 3.6 with TensorFlow 1.10.

It is worth to mention that Arif et al. (2019) proposed to segment the needle from full volumetric data. However, from our challenging dataset, we failed to obtain a successful segmentation result, which might be explained by much more simple network architecture and more challenging task for cardiac ultrasound imaging. Moreover, when compared to our preliminary study (Yang et al., 2019e), which applied morphology operations to connect the closest component and is therefore time-consuming for the post-processing, our proposed POI-FuseNet avoids this complicated post-processing, so that it is more efficient and robust. Otherwise, the morphology operations in 3D space would take more than 10 seconds of processing time for each data volume.

4.3. Performance comparison with non-learning-based methods

Beside the comparisons to learning-based methods, we also perform comparison to non-learning-based methods: Principal Component Analysis on thresholded 3D US (PCA) by Novotny et al. (2003), Parallel Integral Projection (PIP) by Barva et al. (2008), Random Hough Transformation (RHT) by Zhou et al. (2007) and line filter-based RANSAC (Line-RANSAC) algorithm by Zhao et al. (2013). Since above methods using different approaches than direct segmentation, we adopt detection of success volume over total testing volume and detection error as the evaluation metrics, i.e. success rate and endpoints error in voxel, instead of DSD and HD. Specifically, The endpoints error is defined as the average distance of two endpoints on the instrument skeleton from the ground truth, i.e. tip and tail point, to the instrument axis obtained from detection. The experimental results are shown in Table 7. From the table, the proposed deep learning-
Table 7

<table>
<thead>
<tr>
<th>Method</th>
<th>RF-ablation Catheter ex-vivo</th>
<th>TAVI Guidewire in-vivo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Success rate (%)</td>
<td>EE (voxels)</td>
</tr>
<tr>
<td>PCA (Novotny et al., 2003)</td>
<td>23.3%</td>
<td>3.7</td>
</tr>
<tr>
<td>RHT (Zhou et al., 2007)</td>
<td>80%</td>
<td>9.6</td>
</tr>
<tr>
<td>PIP (Barva et al., 2008)</td>
<td>3.3%</td>
<td>13.2</td>
</tr>
<tr>
<td>Line-RANSAC (Zhao et al., 2013)</td>
<td>76.7%</td>
<td>4.0</td>
</tr>
<tr>
<td>Proposed</td>
<td>100%</td>
<td>1.8</td>
</tr>
</tbody>
</table>

based method shows 100% successful rate with the lowest axis error while traditional computer vision techniques have less detection rate with higher errors. The reasons are explained as follows. First, non-learning methods segment the images with simple and straightforward thresholding approaches (PIP does not apply thresholding). These approaches cannot extract accurate instrument-related voxels and omit background voxels, i.e., voxels from tissue and chambers. Second, non-learning-based methods are focusing on post-processing to localize the instrument, which heavily relies on the assumption: background is not complex while instrument has higher intensity distribution than the background. This also explain why these methods achieved promising results on simulated or phantom images. However, real tissue-based images for cardiac US is much more challenging for non-learning based methods, which therefore achieved much lower successful rate. Third, it is worth to discuss about PIP method, which relies on parallel intensity projection for thin instrument detection. As can be seen, it is almost failed in our case. This is because the cardiac instrument has similar intensity distribution with heart tissue, as shown in Fig. 8. More crucially, the heart tissue occupies much more space than the instrument that the PIP method would automatically converge to the direction with tissue passing through it, such as the heart wall of Fig. 8 (e). The experiments with non-learning-based methods further demonstrates the importance of the segmentation stage, which promises a robust and accurate instrument detection. It is worth to mention that with a thicker instrument, it is much easier to detect the catheter than guide-wire from 3D US. Moreover, for catheter with different diameter, a thicker one would much more easier to be detected by non-learning-based method, such as Line-RANSAC or RHT. However, the complex deep learning method could provide a more generalized results with sufficient training images.

5. Conclusions and Future Works

In this paper, we propose an efficient and accurate coarse-to-fine instrument semantic segmentation method for 3D cardiac US images with high efficiency and accuracy. Our method contains the following contributions. First, we proposed a POI-based pipeline for instrument segmentation in the 3D US, which reduces the computation complexity and maintains the segmentation performance for our challenging task. The proposed pipeline is based on a POI selector, which can efficiently select the most interesting regions in the 3D US, thereby improving the segmentation speed for real-time applications. Second, we proposed a FuseNet, which fuses multi-defined features from 2.5D/3D feature extractors, which improves the segmentation in complex 3D US volumetric data. With the proposed feature extractors, the FuseNet can extract direction-fused and full 3D spatial features, which leads to better information usage than solely considering a 3D UNet. Third, a hybrid loss function is proposed to guide the networks to learn the pixel-level and image-level discriminating information simultaneously, which therefore improves the segmentation performances. Challenging 3D US datasets for the ex-vivo RF-ablation procedure and in-vivo TAVI operation were collected to thoroughly validate our proposed method, which achieved higher performances than the state-of-the-art instrument segmentation methods by a large margin. It is worth to mention that the proposed two-stage coarse-to-fine approach has a more flexible and suitable pipeline for our challenging task. Because the length of the instrument has a large variation, it leads to difficulty for extracting the POI-based feature region by a feature attentional pipeline (Huang et al., 2018).

Compared to previous studies for instrument segmentation in the 3D US, our proposed method achieves a higher accuracy with a considerably higher prediction efficiency, which offers a promising value for clinical implementation. From the discussion with clinical experts, they are satisfied with the result, as the segmentation is already clinically useful to facilitate physicians to find the target quickly and make operation easier. Nevertheless, some aspects of our method still need to be discussed. (1) To train the network, the voxel-level annotation is required, which is extremely challenging for low-quality 3D ultrasound images. Therefore, it is difficult to build a system for clinical usage with large-scale image-based supervised learning. (2) For each 3D US data containing the instrument, the instrument must be appeared and can be visualized in the B-mode US image for successful segmentation. However, it is not always the case during real clinical usage. As the relative pose between the US probe and instrument changes, the instrument may become invisible because of acoustic reflection. It seriously hampers the robustness of the proposed method. (3) With the validation on an in-vivo dataset, our proposed method presents a significant value for clinical applications. Nevertheless, further study on extended in-vivo data is required to support further evaluation, which is considered as future work.

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.

CRediT authorship contribution statement

Hongxu Yang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. Caifeng Shan: Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision. Arthur Bouwmans: Investigation, Resources, Data curation. Alexander F. Kolen: Conceptualization, Investigation, Resources, Data curation, Supervision, Project administration, Funding acquisition. Peter H.N. de With: Conceptualization, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.
Acknowledgements

This research was conducted in the framework of “Impulse-2 for the healthcare flagshiptoultrasound” at Eindhoven University of Technology in collaboration with Catharina Hospital Eindhoven and Royal Philips.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.medima.2020.101842

References


