Efficient Medical Instrument Detection in 3D Volumetric Ultrasound Data

Hongxu Yang, Caifeng Shan, Alexander F. Kolen, Peter H. N. de With

***Abstract*—Ultrasound-guided procedures have been applied in many clinical therapies, such as cardiac catheterization and regional anesthesia. Medical instrument detection in 3D Ultrasound (US) is highly desired, but the existing approaches are far from real-time performance. Our objective is to investigate an efficient instrument detection method in 3D US for practical clinical use. We propose a novel Multi-dimensional Mixed Network for efficient instrument detection in 3D US, which extracts the discriminating features at 3D full-image level by a 3D encoder, and then applies a specially designed dimension reduction block to reduce the spatial complexity of the feature maps by projecting from 3D space into 2D space. A 2D decoder is adopted to detect the instrument along the specified axes. By projecting the predicted 2D outputs, the instrument is detected or visualized in the 3D volume. Furthermore, to enable the network to better learn the discriminative information, we propose a multi-level loss function to capture both pixel- and image-level differences. We carried out extensive experiments on two datasets for two tasks: (1) catheter detection for cardiac RF-ablation and (2) needle detection for regional anesthesia. Our experiments show that our proposed method achieves a detection error of 2-3 voxels with an efficiency of about 0.12 sec per 3D US volume. The proposed method is 3-8 times faster than the state-of-the-art methods, leading to real-time performance. The results show that our proposed method has significant clinical value for real-time 3D US-guided intervention.**

***Index Terms*—Medical instrument detection, 3D ultrasound, Multi-dimensional Mixed Network, multi-level loss.**

I. Introduction

ADVANCED clinical imaging modalities, such as fluoroscopy and ultrasound, are widely applied during minimally invasive therapy or intervention surgery. Example procedures include biopsies, cardiac intervention and regional anesthesia, all of which require manipulation of an instrument inside the human body, to reach the target area and perform the task. Among the various imaging modalities, ultrasound (US) is the most popular choice to guide the instrument, because of its real-time radiation-free image visualization capabilities for both tissue anatomy and instrument. Moreover, the US imaging system provides mobility and incurs low costs for the intervention at the hospital. Unfortunately, US imaging suffers challenges of lower image resolution, low image contrast of tissue, image artifacts and appearance distortion, which requires a highly experienced sonographer to interpret the US data. Furthermore, traditional 2D US requires complicated manipulation for coordination alignment between the US probe and instrument, which introduces extra training and effort for clinical specialists to find the instrument. As a consequence, clinical experts might focus too much on instrument detection rather than the operation itself.

In recent years, 3D US has become matured and adopted, providing a potential solution to overcome the limitations in 2D US imaging for computer-assisted instrument detection, which can reduce the complexity of manual operation and improve the operational efficiency. This is because the 3D US transducer provides a large field of view that includes the instrument inside the volumetric data. Automatic detection of an instrument in 3D US introduces the following applications: (1) based on the detection, a smart region-of-interest selection can be achieved, which could automatically demonstrate the target area and improve the efficiency; (2) any sudden movement of the US probe or human body will not lead to repeatedly and manually tuning for the sonographer to find the instrument; (3) a better spatial relationship between instrument and tissue can be obtained using advanced visualization technologies, such as using augmented reality for modeling the organ and instrument during a cardiac intervention. Two examples of 3D US-guided operations are shown in Fig. 1: (a) is an example of cardiac catheterization, and (b) is an example of regional anesthesia.

A. Related work

The commonly studied instrument detection approaches are divided into image-based detection methods and external devices-based methods. By comparing these two approaches, external devices, such as active sensing [1] and robotic-based guidance [2], are not widely adopted in practice. This is due to their requirement for additional equipment in the operation room, extra cost and regulation approval. As a consequence, image-based instrument detection methods provide a more convenient solution in 3D US, as they are potentially less constrained in practice. Although image-based medical instrument detection in 3D US for computer-assisted operation has been studied during past years, the research and works in this area are still limited when compared to the fast development in medical image analysis [3]. From the literature, the medical instrument detection in 3D US can be mainly classified into three categories: non-learning-based methods, handcrafted features based learning methods, deep learning methods.

Non-learning-based methods: Before the popularity of machine learning-based methods, traditional computer vision

Hongxu Yang and Peter H. N. de With are with the Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands.

Caifeng Shan is with College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China. The research was done while he was with Philips Research, Eindhoven, The Netherlands.(email: caifeng.shan@gmail.com)

Alexander F. Kolen is with Philips Research, Eindhoven, The Netherlands.

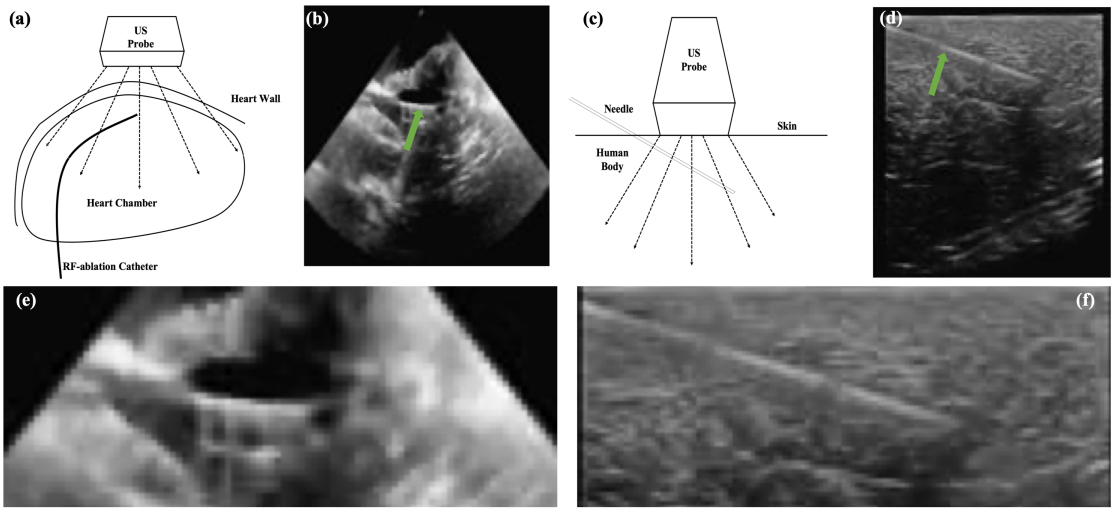

Fig. 1. Examples of two different 3D-US guided operations: (a)(b) cardiac catheterization with RF-ablation catheter; (c)(d) regional anesthesia with needle. Green arrows are pointing to the instruments, (e)(f) enlarged medical instrument in (b)(d).

technologies were applied on 3D US volumes to detect the instrument for intervention therapy. Ayvaci *et al.* [4] proposed to apply an energy function on Gaussian filtered images, which segments a needle in transrectal ultrasound images. The segmented needle is fused with MR images together with EM tracker to guide the operation. Zhao *et al.* [5] extensively compared several methods on biopsy needle detection in 3D US, which considers Principal Component Analysis (PCA) [6], random Hough transform (RHT) [7], parallel integral projection (PIP) [8] and ROI-based RANdom SAmple Consensus (RANSAC) and Kalman filter (ROI-RK). Although ROI-RK demonstrates a promising performances in the reported cases, the lack of discriminating information and challenging US image quality make their methods worse than learning-based approaches [9], [10].

Handcrafted features based learning methods: Uherčík *et al.* [11] applied Frangi features to classify instrument voxels using support vector machine (SVM). The model-fitting based on RANSAC was applied to localize straight line instruments. Meanwhile, Zhao *et al.* [12] applied the same learning method to track the needle using an ROI-based Kalman filter. However, these approaches only considered a pre-defined Frangi feature as the discriminating information, which is less sensitive to diameter variation, but also with limited information representation capacity. Hatt *et al.* [13] had proposed to apply Gabor features for learning-based segmentation to guide the beam steering during the intervention. Recently, Pourtaherian *et al.* [9] have studied instrument detection algorithms in 3D US. Their method distinguishes the candidate instrument-like voxels by incorporating the Gabor feature. This feature introduces more discriminating information on the distribution of local orientations. After the voxel-based classification, the instrument is localized with a pre-defined semantic model. Pourtaherian *et al.* performed an experiment on catheter detection in an *in-vitro* dataset, which shows the necessity for further validation on *ex-vivo* or *in-vivo* datasets. Yang *et al.* [10] employed more discriminating features for a supervised learning method with a multi-scale approach, to capture more contextual information. Although they achieved satisfying performance on different *ex-vivo* and *in-vivo* datasets, the limited capacity of handcrafted features leads to outliers after segmentation, which requests a complex model-fitting or post-processing to finally detect the catheter. Furthermore, handcrafted feature designing requires experience and effort, which is then gradually replaced by the deep learning.

Deep learning methods: Convolutional neural networks (CNNs) have achieved significant success in different recognition tasks in the medical imaging area [3]. Researchers have proposed medical instrument detection methods using this deep learning approach in many different applications and modalities. For example, Ambrosini *et al.* [14] proposed to use a 2D fully convolutional network (FCN) to segment complex cardiac catheters under 2D X-ray fluoroscopy for real-time applications. Mwikirize *et al.* [15], [16] employed transfer learning-based FCN to detect the needle under 2D US for real-time applications. Although these methods achieve promising performances, their 2D methods are not suitable for instrument detection in 3D volumetric data, since these 2D networks cannot handle the complex 3D information with limited a field of view, or they request a manual preselection of an image slice from the 3D volumetric data. To exploit more spatial information in 3D US, tri-planar CNN methods for voxel-wise classification were introduced for instrument segmentation [17]. However, these approaches require the network to iteratively predict all the voxels in 3D US, leading to a high computation cost, which is therefore not suitable for real-time applications. Although Yang *et al.* [18] proposed a pre-filtering-based acceleration method, the 10-seconds prediction time is still too long for a real-time application, which is typically around 5-10 frames per second in 3D US-guided operation for most clinical scenarios. Slice-based FCN [17], [19] was proposed to segment an instrument in 3D US by decomposing the volume into adjacent slices using a transfer-learned 2D FCN. Although their methods achieve impressive

segmentation results for instrument segmentation, the capacity of 3D contextual information extraction was limited due to the slice-based strategy. To overcome the 3D information leakage, Yang *et al.* [20] proposed a patch-based 3D UNet to segment a cardiac catheter in 3D US, which achieved the satisfied performance. Nevertheless, iteratively patch-based operation in 3D volume hampers the interpretation of the contextual and semantic information of the whole image. Instead of patch-based approaches, Arif *et al.* [21] proposed a full-3D CNN method for instrument segmentation with 3D UNet as a backbone, which shows an impressive performance on a challenging dataset. However, the true 3D operations at both encoder and decoder sides complicated network structure and can be easily constrained by the GPU memory size.

B. Our work

With the above considerations, such as network complexity, full-image information usage and time efficiency, we propose a novel Multi-dimensional Mixed Network (MixDNet) to detect the medical instrument in 3D volumetric data at full image level while achieving high detection efficiency. The full-3D US volume is first processed by a compact 3D encoder, which extracts the 3D semantic features from the whole image. Then, the 3D features are processed by our proposed dimension-reduction module, which applies the multi-dimensional reduction in the spatial/channel domain and converts the 3D features into a 2D format. Based on the 2D features reduced along one of the specified axes, a compact 2D decoder is introduced to generate the instrument's prediction in a 2D plane. With the predicted instrument through two orthogonal directions, the instrument can be detected in the reconstructed 3D volumetric data. With the proposed structure, 3D semantic features are extracted and compacted by the dimension reduction module, which simplifies the decoder part and omits the spatial information redundancy. Thereby, the overall complexity of the network is reduced, and therefore can be trained more easier. To validate our method, we perform extensive experiments on two different datasets for different clinical tasks, i.e. for RF-ablation surgery for cardiac catheterization and regional anesthesia. The results showed a detection accuracy with an end-point error of around 2-3 voxels and an instrument axes mismatch around 5-7 degrees. More crucially, the proposed method can detect the instrument in 3D US volume within 0.12 seconds, which is 3-8 times faster than the state-of-the-art methods. These results provide similar or better accuracy than the state-of-the-art methods and satisfy the real-time requirement for 3D US-guided operations. The results show that our proposed method is suitable for real-time clinical applications, thereby potentially improving the treatment outcome.

With the proposed MixDNet, this paper presents the following contributions. (1) We propose a novel multi-dimensional hybrid structure for instrument detection in 3D US. With this approach, network complexity is reduced and overfitting can be better avoided when compared to the traditional full-3D networks. (2) The obtained structure is based on a specifically designed dimension-reduction block, which reduces the spatial information from 3D to 2D and extracts the most relevant instrument information along the reduced direction. (3) To train the CNN, we propose a multi-level loss function, which can allow the network to learn the information at pixel-level and image-level simultaneously. The remainder of this paper describes the details of our proposed method in Section II. The experiments, results and discussions are discussed in Section III and IV, respectively. Section V provides conclusions for the paper.

II. METHODS

In this section, we propose a framework to detect the medical instrument in 3D volumetric B-mode ultrasound images using the Multi-dimensional Mixed Network (MixDNet). The proposed pipeline is described in Fig. 2, in which the input 3D volume is processed by MixDNet. The output of the MixDNet is the instrument skeleton estimation projected along the axes. Based on the estimated skeleton in axial, lateral or elevation direction, the instrument is detected and visualized in 3D space or a 2D plane. The details of MixDNet are shown in the Section II.A. To train the MixDNet, a hybrid multi-level loss is introduced in Section II.B, which is based on the annotation of the instrument skeleton. With the predicted instruments in the 2D planes, the instrument is obtained in 3D space by projecting planes, which is shown in Section II.C.

A. Construction of MixDNet

The proposed MixDNet is depicted in Fig. 3. In contrast with a standard encoder-decoder architecture like 3D UNet [22] or Feature Pyramid Nets [23], we propose a hybrid dimension architecture, which consists of a 3D encoder, 3D to 2D information-reduction layer and a 2D decoder. For a 3D US volume, a 3D encoder is applied for high-level feature extraction. The encoded information is processed by a projection layer to extract the most discriminating information along the principal axes, which extracts the information through the dimensions and channels. Then, the compressed features are processed by a 2D decoder, which decodes the projected features to generate the instrument skeleton along the axes. More specifically, there are two individual branches of dimension reduction to extract the dimension information along the axial or side direction simultaneously, i.e. axial and lateral (or elevation) direction (following the nature of the US cone) in the Fig. 3. To reduce the complexity of the network, reduction blocks at the same image size are shared. Moreover, the 2D decoder parameters are shared in different directions.

Considering the limited GPU memory size of hardware and the complex 3D convolutional kernels, we have designed a compact 3D encoder to avoid GPU memory overflow and network overfitting. Specifically, the 3D encoder includes a stack of 3D convolutional and Maxpooling layers. For each convolutional layer, ReLU and and Instance Normalization layers are followed to accelerate the convergence.

As for the 3D-to-2D reduction module, it is a spatial-channel based attentional module, which can extract the most relevant information along a specific dimension, while reducing the size of feature mappings. More specifically, we

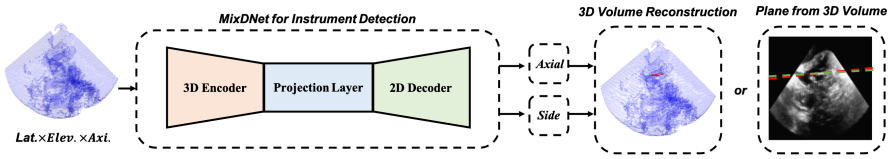

Fig. 2. Pipeline of the proposed medical instrument detection method, where the green dashed line is the ground truth, while the red dashed line indicates the detected instrument (or its axis). The 3D ultrasound volume is processed by MixDNet to generate the instrument prediction on the projected planes along axial, lateral or elevation directions. Based on the prediction on the projected planes, the detected instrument is reconstructed and visualized in a 3D volume or visualized in a 2D plane by slice selection.

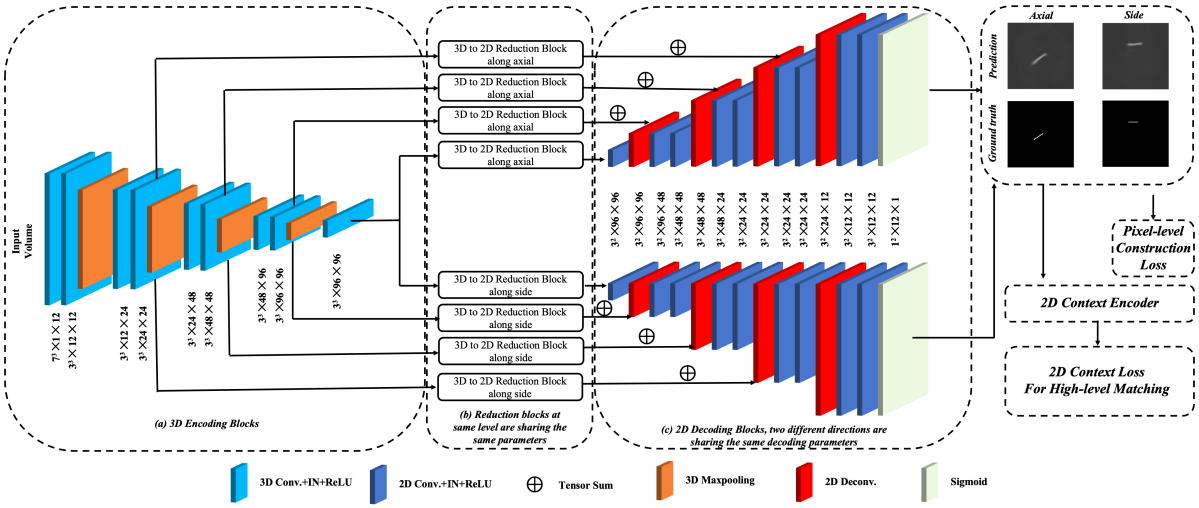

Fig. 3. Illustration of the proposed Multi-dimensional Mixed Network (MixDNet). The network consists of three main parts: (a) information Encoder in 3D space; (b) dimension reduction along the axial and lateral (side) axes; (c) information Decoder in 2D space, which generates the projected prediction of the instrument in 2D space. Based on the prediction from the 2D decoder, the multi-level loss is applied to supervise the network.

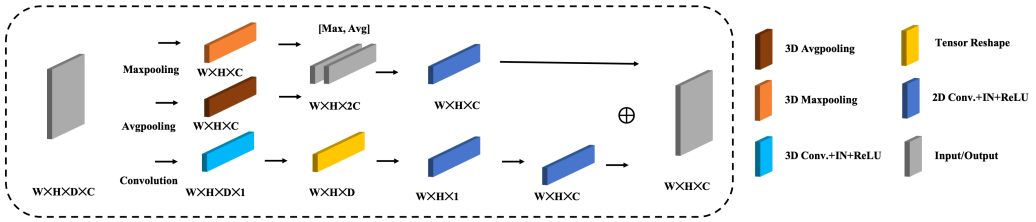

Fig. 4. Flow of the dimension-reduction module. The input feature map is processed by three different branches along one of three axes (along D axis in this figure): Maxpooling, AvgPooling and Convolution operations. The feature maps processed by Maxpooing and Avgpooling are concatenated and processed by a convolution operation, which is then fused with the Convolution branch by summation. The block color is the operation applied on the feature maps while the size of the result is shown under the operation.

design a dimension-reduction (or dimension-projection) block based on three different operations along one of three principal axes, which is depicted in Fig. 4. The block extracts the first-order statistics along the interested axis, by maximizing and averaging all possible discriminative information. Then, the output tensors are concatenated and processed by a convolution operation to reduce the channel size. Meanwhile, a series of convolution operations are applied on the input feature maps. First, the channel information is compressed, which is then followed by two convolutions to obtain the dimension-reduced feature map. Finally, the feature maps are accumulated to obtain the final result. This approach is attention block associated with spatial and channel information, but consisting of different ways to summarize them. As for the Avgpooling-based branch, it can summarize all information along one dimension, while the Maxpooling operation would only focus on the maximized signal responses and ignore some minor information. As for the Convolution-based branch, it can summarize the channel-based information, which acts as a compensation for the above non-parametric approaches. As a consequence, the information can be summarized properly and the spatial dimensions are reduced. As shown in Fig. 3, for each feature scale, a scale-specific dimension-reduction block is designed to fit the convolutional channels. In this paper, the side view is chosen to be the lateral direction, because of the fixed pose between instrument and tissue in the datasets.

Based on the reduced feature maps from dimension-

reduction blocks, a 2D decoder is designed to formulate the output, which describes the instrument skeleton along the axes. The decoder consists of 2D convolutional layers, followed by a ReLU and Instance Normalization. De-convolutional layers are applied to upsample the the feature maps. More details of network are shown in Fig. 3.

The motivations of the proposed structure are explained in following: (1) Conventional 3D U-Net structure requests a complex encoder and decoder, which thereby increases the memory usage for limited GPU hardware. Moreover, with the input volume size increased, GPU memory can easily overflow with a larger mini-batch size, which therefore increases the difficulties to train the network. (2) In the proposed structure, the decoder part is simplified from 3D space to 2D space, which is based on the prior-knowledge of instrument shape in 3D images and to omit the decoder redundancy. More crucially, the dimension-reduced outputs drastically reduce the class imbalance phenomena, which makes it easier to train the network. To overcome it and apply the detection on whole volume, we designed such structure for efficient detection purpose.

B. Multi-level loss function

The input of the proposed network is a full-3D B-mode volumetric image, while the output is in 2D planes, indicating the instrument skeleton projection along the axes, see Fig. 5. To constrain the MixDNet to generate the correct skeleton in 2D projected images, we design a multi-level loss function, which is generally formulated by

$$Loss(\hat{Y}, Y) = \alpha Loss_{\text{pixel-level}}(\hat{Y}, Y) + \beta Loss_{\text{image-level}}(\hat{Y}, Y), \tag{1}$$

where the $\hat{Y}$ is the network prediction and Y is ground truth. Loss component $Loss_{pixel-level}$ focuses on the prediction of the projected instrument skeleton in a 2D plane, while the loss component $Loss_{image-level}$ concentrates on a high-level description of the skeleton in the 2D image. Parameters α and β are weight parameters to balance the losses. More specifically, component $Loss_{\text{pixel-level}}$ is defined as a weighted binary cross entropy (BCE), specified by

$$Loss_{\text{pixel-level}}(\hat{Y}, Y) = \sum_{j=1}^{N} w_j^i y_j^i log(\hat{y_j}^i) + \sum_{j=1}^{N} w_j^n y_j^n log(\hat{y_j}^n), \tag{2}$$

where N denotes the number of pixels for each 2D prediction or ground truth image, i represents the instrument skeleton pixels and n represents the group of the non-instrument pixels. The class weight parameter w is a hyper-parameter to control the weight between two different classes, which is employed because of the extremely imbalance of classes in the ground-truth images. Moreover, deep supervisions [24] are employed at the 2D decoding blocks with weight 0.1, which are applied after dimension-reduction operations in Fig. 3.

Besides the pixel-level loss, we also define an image-level loss, which enforces the MixDNet to learn high-level information to properly match the predictions and ground truth in 2D planes. As described in Fig. 5, the constructed projection images, together with corresponding ground-truth images are processed by a shared contextual encoder (CE) to generate the high-level descriptor [20], which can describe the input images in a latent space. For descriptor of each view, its corresponding loss function is defined as the distance between the descriptor of prediction and its corresponding ground truth, leading to

$$Loss_{\text{image-level}}(\hat{Y}, Y) = ||\text{CE}(\hat{Y}) - \text{CE}(Y)||_2, \tag{3}$$

where function CE$(\cdot)$ denotes the contextual encoder net for latent space projection, $||\cdot||_2$ is the norm-2 distance. As a consequence, Eqn. (3) also holds for any other predictions along different axes. It is worth to mention that the CE$(\cdot)$ is projection function, which projects complex information into a latent high-level space. The $Loss_{\text{image-level}}$ measures the similarity between two images in the latent space, and therefore can be sensitive to the shape and location difference at the contextual view. Based on above definitions for $Loss_{\text{pixel-level}}$ and $Loss_{\text{image-level}}$, the overall loss function on two individual axes can be formulated as a summation, based on the predictions and ground-truth pairs, i.e. $\{\hat{Y}_{\text{axial}}, Y_{\text{axial}}\}$ and $\{\hat{Y}_{\text{side}}, Y_{\text{side}}\}$.

C. Instrument Detection based on 2D Projections

The MixDNet generates the estimated instrument skeleton along different axes as 2D predicted images, i.e. $I_{\text{2d-axial}}$ and $I_{\text{2d-side}}$, as shown in Fig. 3. Based on the 2D predictions, the instrument in the 3D volume is obtained by replicating the 2D images along the feature map's direction of reduction. As a result, the 3D volume with the detected instrument is obtained as follows

$$I_{3d} = Rep(I_{\text{2d-axial}}, \theta_{\text{axial}}) + Rep(I_{\text{2d-side}}, \theta_{\text{side}}), \tag{4}$$

where the $Rep(\cdot, \theta)$ is the replication operation of 2D prediction along the specific direction θ, such as θ_{side} for side-view direction. Based on the reconstructed volume, a simple threshold and RANSAC model-fitting are applied on the sparse volume to find the instrument. Another choice to detect the instrument from 2D planes is plane extraction, which is inspired by clinical usage. In practice, sonographers prefer to automatically visualize the plane containing the instrument, i.e. the instrument axis is in-plane, which can avoid a complex plane tuning to find the instrument in 3D volumetric data. Exploiting the natural properties of ultrasound imaging, i.e. the propagation of sound waves propagate always along the direction of the axial direction of the ultrasound probe, the instrument detection can be formulated by two steps: (1) extract the plane containing the instrument along the axial direction of the probe, (2) based on the prediction alongside the view axis, the instrument in the extracted plane is detected. These steps are demonstrated in Fig. 6 as an example.

III. Materials and Evaluation Metrics

A. Datasets and Preprocessing

To validate our proposed instrument detection method, we have collected two different datasets for different ultrasound-guided operation tasks: RF-ablation operation for cardiac intervention and needle interventions for regional anesthesia.

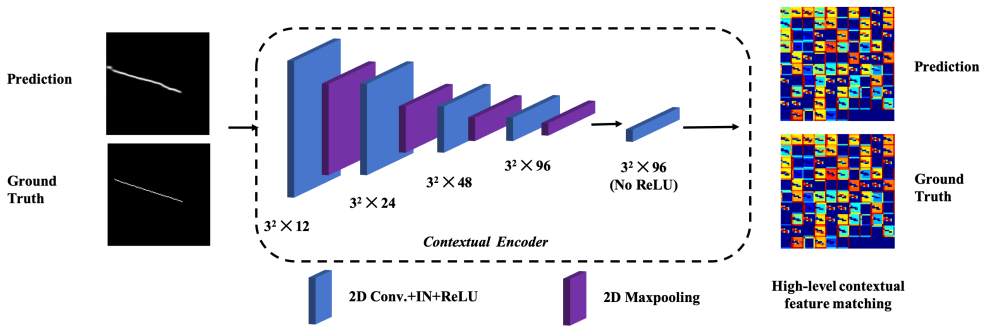

Fig. 5. Overview of the contextual encoder for image-level loss construction. The ground truth and prediction are processed by an encoder to generate high-level feature maps, which are matched to measure the high-level similarity in an encoded feature space. The high-level feature maps for ground truth and prediction are represented by heat maps.

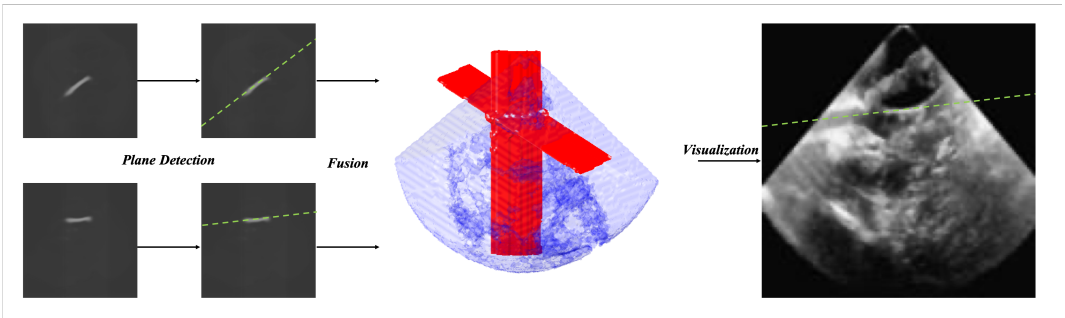

Fig. 6. Plane containing the instrument is extracted from 3D volumetric data based on the 2D prediction along axes.

RF-ablation Catheter Dataset: We have collected 94 3D cardiac US volumetric data from eight porcine hearts. During the recording, the heart was placed in a water tank with the RF-ablation catheter for Electrophysiology (with diameter of 2.3-3.3 mm) inside the heart chambers. The phase-array US probe (X7-2t with 2500 elements by Philips Medical Systems, Best, Netherlands) was placed next to the interested chambers to capture the images containing the catheter, which was monitored by US console (EPIQ 7 by Philips). The obtained volumetric images are re-sampled to have a volume size of $160 \times 160 \times 160$ voxels (where padding is applied at the boundary to make the volume such that it has equal sizes in each direction), which leads to a voxel size ranging from 0.3-0.8 mm. Because of the limited free space in the heart chamber, only the head of the catheter can be captured, which yields a straight and short instrument in the volumetric data, due to the flexible nature of the catheter. An example 2D slicing image is shown in Fig. 1 (e). More specifically, the catheter exhibits a higher contrast against water inside the cardiac chambers. However, it become especially challenging when the chambers collapse and attaching the catheter, which hampers the visibility and contrast of the instrument. As a consequence, we tried to fill sufficient water into the chambers to avoid collapsing during the recording.

Anesthesia Needle Dataset: A dataset based on needle usage is collected by a motorized VL13-5 linear-array (VL13-5 with 192 elements by Philips Medical Systems, Best, Netherlands) from chicken breast, , which was monitored by US console (iU22 xMATRIX by Philips). The dataset includes 20 volumetric images with two different types of needles: 17G (diameter of 1.47 mm) and 22 G (diameter of 0.72 mm). For each type of needle, 10 images were collected. To ensure the image independence, needles were inserted into different location of the chicken breast. The images are isotropically re-sampled to obtain a voxel size of 0.3 mm, which leads to a volume size of $128 \times 128 \times 128$ voxels. An example of a 2D slicing image is shown in Fig. 1 (f). As a contrast to catheter dataset, needle has a clear contrast to surrounding tissues because of different material and medium.

Annotation: The ground-truth binary skeleton mask for both datasets are generated by connecting the annotated instrument endpoints, which are annotated by two clinical experts. This annotation strategy can reduce the annotation difficulty and effort in the 3D US volumetric data compared to voxel-level annotation [9], [21]. Although [21] proposed to use dilation operation to instrument voxels, the deformation and blurry boundaries of the instrument lead to the voxel-category uncertainties in the automatically generated ground truth for network learning. However, with skeleton-based training giving a sparse annotation only, it is more challenging for the network to learn the semantic information when compared to the dense annotation-based instrument segmentation.

B. Implementation Details and Training Process

Considering the limited dataset and GPU memory (11G for an NVidia 1080 Ti GPU), the proposed MixDNet has 12 convolutional kernels in the first layer, which are gradually doubled after each maxpooling operation, except for the deepest level where two convolutional layers are applied with

kernel size 96. As a consequence, the MixDNet has number of hyper-parameters that is around 1.1 M, which is smaller than a standard 3D UNet or similar architectures (commonly around 5-10 million).

We have trained the network using stochastic gradient descent update with the Adam optimizer. As for the catheter dataset, the initial learning rate is set to 0.001 for a mini-batch size equal to 4. The learning rate is reduced for every 100 epochs by a factor of 0.1 and the training is terminated after 200 epochs. During the experiment, 64 volumes are randomly selected as the training data and 30 volumes are used as the testing images. With respect to the needle dataset, due to the limited amount of images for training, five-fold cross-validation is applied. This means 20 images are randomly divided in to two parts: 16 images are used as the training dataset while the rest as the testing image. The procedure repeated five times to obtain the overall performance by the average. The initial learning rate is set to be 0.001 for a mini-batch with size 8 and is terminated after 500 epochs. During the training, we apply mirroring on elevation/lateral directions and rotation along the axial direction to preserve the global shape information of US cone. Shifting within 16 pixels along the elevation or lateral directions are randomly applied on-the-fly during the training stage. Moreover, intensity jittering, image resizing from 0.8 to 1.2 times is also randomly applied. During the training, the total number of observed images for the network are 12,800 and 8,000, for the catheter dataset and the needle dataset after the data augmentation, respectively. Data augmentation could facilitate the network to learn more invariant information of the dataset and avoid overfitting [25]. To achieve the best performance of the proposed structure, cross-validation is applied on the training images to find the best overall hyper-parameters in both catheter and needle dataset. The parameters α and β were empirically selected as 1 and 0.01, respectively. The proposed method is implemented in Python 3.7 with TensorFlow 1.10.

C. Evaluation Metrics

Because the ground truth of our datasets is an instrument skeleton indicated by a line having a diameter of one voxel, standard evaluation metrics such as the Dice score, are not feasible to evaluate our method. Instead, the following metrics are considered to evaluate the performance of our proposed method. Moreover, the prediction time per volume is evaluated in the experiments.

Average Hausdorff Distance: The ground truth of instrument skeleton is a line with one voxel diameter in the 3D space. As a result, the Average Hausdorff Distance (AHD) is considered as an evaluation metric ([26]), because it is more sensitive to voxel mismatch between annotation voxels and the prediction voxels. Moreover, the AHD is less sensitive to outliers than a standard Hausdorff Distance, since the detected skeleton of the instrument is sometimes thicker than the ground truth, as shown in Fig. 5. The AHD is defined by:

$$\mathrm{AHD}(A,B) = \max(d(A,B), d(B,A)). \tag{5}$$

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

$$d(A,B) = \frac{1}{N}\sum_{a\in A} \min_{b\in B}||a-b||, \tag{6}$$

where the A and B denotes voxel groups from the ground truth and the predicted result, while a and b are voxels of them, respectively. Parameter N represents the size of group A.

Axis Localization Error: The axis localization error is defined with two different metrics: endpoints error (EE) and the orientation error (OE). The EE is defined as the maximum distance of two endpoints on the instrument skeleton from the ground truth, i.e. tip and tail point, to the instrument axis obtained from 3D reconstruction (conservative measure). Similarly, the OE is defined as the angle difference between the detected instrument axis to the ground-truth skeleton.

IV. Results

This section is organized as follows: (1) an ablation study of different loss types and effectiveness of going from 3D to the 2D dimension-reduction module; (2) a performance comparison with different state-of-the-art (SOTA) medical instrument detection methods in 3D US. In our experiments, we considered the RANSAC-based model-fitting as post-processing step to detect the instrument, which introduces randomness in the results. We have performed detection session five times and chosen the worst-case results of each method. The proposed method and the SOTA methods are compared based on above.

A. Ablation Study

We have performed two ablation studies to validate our proposed method. First, we validate our proposed method with different types of loss functions, i.e. only standard BCE loss, only with pixel-level loss (weighted BCE), only with image-level loss and the hybrid loss (see Eqn. (1)) as proposed with MixDNet. Second, with the proposed multi-level loss function, we have performed another ablation study on the effectiveness of different 2D dimension-reduction methods, which are discussed in Section II for each individual branch and their ensemble, i.e. Maxpooling, Avgpooling, Convolution, Concatenation of Maxpooling/Avgpooling (denoted as [Max,Avg]), and the proposed method. Meanwhile, we also considered the comparison between true 3D network and our proposed MixDNet, i.e. excluding the dimensional reduction blocks and using full 3D convolutions in the decoder (due to the limited GPU memory for full-volume operations on an NVidia 1080Ti with 11 G memory, we apply a mini-batch size equal to unity). All above ablation studies are summarized in Table I and Table II.

From the results in Table I, it shows that the multi-level loss can provide a better performance in both datasets. However, the network cannot learn the meaningful semantic information with only image-level loss. Because the randomly initialized contextual encoder cannot generate a correct feature representation of both ground truth and prediction from the untrained network, this fails to guide and constrain the segmentation network to learn meaningful knowledge after training iterations.

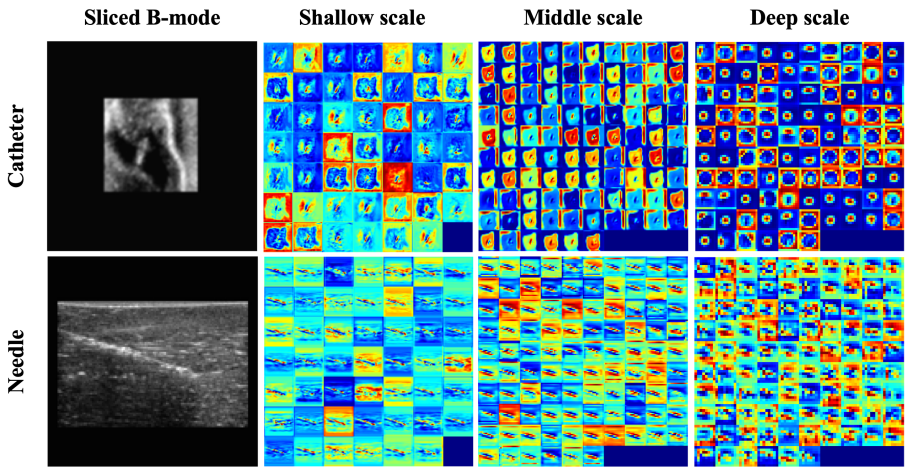

Fig. 7. Heat maps of feature activation after dimension-reduction blocks at shallow, middle and deep scales of the network, which are depicted in Fig. 3 as summation symbols. Feature maps are re-scaled for visualization. The results show the dimensional-reduction blocks extracts the most instrument-relevant information from 3D tensors. The top-row images are from the catheter dataset and the bottom-row images are from the needle data.

TABLE I

ABLATION STUDY ON DIFFERENT LOSS TYPES FOR TWO DATASETS USING OUR PROPOSED NETWORK, WHICH ARE EVALUATED BY AVERAGE HAUSDORFF DISTANCE (AHD), ENDPOINTS ERROR (EE) AND ORIENTATION ERROR (OE) USING MEAN±STD. FAILED MEANS THAT WE FAILED TO OBTAIN THE RESULTS. $*$ MEANS A STATISTICAL DIFFERENCE UNDER THE METRIC WITH SIGNIFICANCE LEVEL AS 0.05.

Loss	Catheter		
	AHD (voxel)	EE (voxel)	OE (degree)
BCE	7.2 ± 8.6	7.5 ± 19.0	16.1 ± 19.2
pixel-level loss	2.6 ± 1.1	2.6 ± 0.7	10.5 ± 6.5
image-level loss	Failed	Failed	Failed
Proposed	$\mathbf{2.4 \pm 0.9}$	$\mathbf{2.3 \pm 0.5}$	$\mathbf{7.3 \pm 2.1}^*$
Loss	Needle		
	AHD (voxel)	EE (voxel)	OE (degree)
BCE	16.4 ± 19.6	11.4 ± 19.8	23.3 ± 35.7
pixel-level loss	5.2 ± 6.7	2.9 ± 1.0	5.5 ± 3.7
image-level loss	Failed	Failed	Failed
Proposed	$\mathbf{3.2 \pm 2.2}^*$	$\mathbf{2.5 \pm 0.6}$	$\mathbf{5.0 \pm 4.9}$

TABLE II

ABLATION STUDY ON DIFFERENT DIMENSION-REDUCTION MODULES FOR TWO DIFFERENT DATASETS, WHICH ARE EVALUATED BY THE AVERAGE HAUSDORFF DISTANCE (AHD), ENDPOINTS ERROR (EE) AND ORIENTATION ERROR (OE) USING MEAN±STD. $*$ MEANS A STATISTICAL DIFFERENCE UNDER THE METRIC WITH SIGNIFICANCE LEVEL AS 0.05.

Method	Catheter		
	AHD (voxel)	EE (voxel)	OE (degree)
3D UNet	5.9 ± 6.7	5.3 ± 6.0	26.2 ± 26.2
Maxpooling	3.2 ± 2.1	2.8 ± 1.0	9.4 ± 4.5
Avgpooling	3.5 ± 6.6	2.7 ± 2.7	8.8 ± 4.4
Convolution	3.1 ± 1.3	3.0 ± 1.0	11.0 ± 5.6
[Max,Avg]	2.7 ± 1.3	2.4 ± 0.6	9.3 ± 4.2
Proposed	$\mathbf{2.4 \pm 0.9}$	$\mathbf{2.3 \pm 0.5}$	$\mathbf{7.3 \pm 2.1}^*$
Method	Needle		
	AHD (voxel)	EE (voxel)	OE (degree)
3D UNet	19.2 ± 20.2	8.9 ± 15.2	11.6 ± 11.3
Maxpooling	4.5 ± 7.0	2.5 ± 0.9	5.2 ± 3.9
Avgpooling	4.3 ± 5.0	2.7 ± 0.7	5.1 ± 3.5
Convolution	4.3 ± 4.6	2.8 ± 0.7	5.4 ± 3.2
[Max,Avg]	3.6 ± 2.4	2.7 ± 1.0	5.3 ± 3.7
Proposed	$\mathbf{3.2 \pm 2.2}$	$\mathbf{2.5 \pm 0.6}$	$\mathbf{5.0 \pm 4.9}$

As a consequence, the image-level loss can be only considered as complementary for the pixel-level loss. It is worth to mention that the BCE loss can lead to a failure in detection, due to extremely imbalanced classes, which underestimates the detection and generates an empty prediction. To further validate the proposed method, a paired t-test is performed with a significance level set to 0.05 based on the maximized value of 5 runs rather than multiple comparisons (same as the rest t-tests). As shown in Table I, the proposed multi-level loss does not have statistically significant performance improvement in most metrics. However, we do observe that on the catheter dataset it performs much better than the pixel-level loss and the standard BCE loss with the OE metric. On the needle dataset, the proposed multi-level loss performs better than others with the AHD metric.

As shown in Table II, for catheter detection, the proposed dimensional reduction significantly outperforms other modules except for [Max, Avg]. Compared to [Max, Avg] module, our proposal does not show statistically different results with the AHD and EE metrics, but performs better with the OE metric. For needle detection, most of the examined dimensional reduction modules do not show statistically significant difference, only 3D UNet performs much worse. Considering the needle dataset is with very limited data, more investigation needs to be performed on a larger dataset in future.

Example feature maps after the proposed dimension-reduction blocks are shown in Fig. 7 for two different datasets. As can be observed, from Shallow scale to Deep scale, the feature maps represent discriminating information from local texture to high-level locations. By comparing the instrument areas to B-mode slice, the instrument can be found with high contrast in feature maps. However, when it comes to black region in B-mode, i.e. empty area, the corresponding feature maps look rather noisy, which is because they are obtained by compressing the non-instrument information. This figure demonstrates that the proposed block can extract the discriminating information along the specific feature map axis. With further operations, the instrument skeleton is predicted

TABLE III
PERFORMANCE COMPARISONS WITH SOTA METHODS FOR CATHETER DETECTION, WHICH ARE EVALUATED BY AVERAGE HAUSDORFF DISTANCE (AHD), ENDPOINTS ERROR (EE), ORIENTATION ERROR (OE) AND INFERENCE TIME.

Method	Catheter			
	AHD (voxel)	EE (voxel)	OE (degree)	Time (sec.)
Handcrafted [9], [10]	6.6 ± 10.4	6.5 ± 7.8	17.0 ± 17.9	> 600
VOI-PatchCNN [18]	2.8 ± 1.5	2.8 ± 1.2	8.2 ± 3.2	~ 10
SliceFCN [19]	4.1 ± 5.3	3.7 ± 4.9	9.2 ± 5.3	~ 1.0
Pyramid-UNet [20]	2.6 ± 1.9	2.4 ± 1.0	7.5 ± 3.3	~ 48.0
Proposed	2.4 ± 0.9	2.3 ± 0.5	7.3 ± 2.1	~ 0.12

TABLE IV
PERFORMANCE COMPARISONS WITH SOTA METHODS FOR NEEDLE DETECTION, WHICH ARE EVALUATED BY AVERAGE HAUSDORFF DISTANCE (AHD), ENDPOINTS ERROR (EE), ORIENTATION ERROR (OE) AND INFERENCE TIME.

Method	Needle			
	AHD (voxel)	EE (voxel)	OE (degree)	Time (sec.)
Handcrafted [9], [10]	7.8 ± 8.7	7.6 ± 10.3	8.3 ± 9.0	> 120
PatchCNN [17]	4.9 ± 7.3	2.8 ± 0.7	5.9 ± 4.7	> 240
ShareFCN [17]	5.9 ± 8.3	2.6 ± 0.7	5.1 ± 2.1	~ 1.0
3D UNet [21]	19.2 ± 20.2	8.9 ± 15.2	11.6 ± 11.3	~ 0.2
Proposed	3.2 ± 2.2	2.5 ± 0.6	5.0 ± 4.9	~ 0.06

in 2D output images, of which examples are shown in Fig. 6.

B. Performance Comparison with SOTA

We have compared our proposed method to many different state-of-the-art (SOTA) medical instrument detection methods in 3D US with respect to AHD, EE, OE and inference time. For fair comparison, we implement all the SOTA methods on our datasets with *2 voxels dilation* on skeleton annotation instead of voxel-level accurate annotation [21], since we failed to obtain successful prediction due to extremely imbalanced skeleton annotation for CNN-based segmentation methods, or too much false positives after the handcrafted feature-based classification method. This also indicates the SOTA segmentation methods are not feasible for skeleton-based detection in 3D space while the proposed method can handle it. The detailed results are listed in Table III and Table IV.

As for catheter detection in 3D US, our proposed method is compared in Table III with several catheter detection methods in 3D US volumetric data, based on multi-scale and definition features with AdaBoost classifier (Handcrafted from [9], [10]), voxel-of-interest-based patch-wise CNN (VOI-PatchCNN from [18]), slice-based 2D FCN for 3D US (SliceFCN from [19]) and Pyramid UNet for patch-based segmentation ([20]). From the results, our method achieves a better detection accuracy with higher efficiency. It should be noticed that our proposed method is solely based on the annotated skeleton instead of voxel-level annotation, which provides less information than the SOTA methods. However, experimental results show our method achieved a better performance with more challenging training condition. Our approach is therefore more challenging than the reported SOTA references. Moreover, when considering the time efficiency, our proposed MixDNet architecture achieves a more than 8 times faster inference efficiency. The obtained fast and accurate results present a promising performance for the requirement of real-time applications.

As for the needle detection in 3D US, our proposed method is compared in Table IV with several needle detection methods in 3D US volumetric data, based on multi-scale and definition features with AdaBoost classifier (Handcrafted from [9], [10]), Patch-wise CNN ([17]), ShareFCN ([17]) and 3D UNet ([21]). From the results, the proposed method achieves better performance than state-of-the-art methods with higher efficiency. It is mentioned that because of the different tasks and datasets, we failed to obtain the results based on the structure of [21]. This is explained by a shallow network with just 4 kernels in the first layer, which cannot handle our complex dataset. In contrast to it, a more complex and shallower 3D UNet is mentioned and shown in the Table II and Table IV, which obtained a much worse performance due to an extremely class imbalance and skeleton annotation.

Moreover, paired t-tests are performed based on the metric of the endpoints-error (EE) with the significance level of 0.05. For the catheter dataset, the proposed method is statistically better than handcrafted, VOI-PatchCNN and SliceFCN methods, while there is no statistical difference between the Pyramid-UNet and our method. For the needle dataset, the proposed method is statistically better than handcrafted, PatchCNN and 3D UNet methods, while there is no statistical difference between the ShareFCN and our method. However, the proposed method achieves much faster inference efficiency than the state-of-the-art methods.

In terms of time efficiency, the inference time is around 0.12 sec. per volume for the catheter dataset and around 0.06 sec. per volume for the needle dataset on a GTX 1080Ti GPU. It is worth to mention that our method achieves around 6 and 3 seconds for catheter and needle, respectively, on a standard CPU (2.4-GHz quadcore 8th-generation i5 processor). This efficiency improvement indicates clearly lower hardware requirements for real-time application.

C. Discussion

The proposed method achieved better detection accuracy, yet more crucially a higher efficiency than the state-of-the-art methods. However, to apply our method in real-time applications, there are still few discussion aspects for it. Further validation on *in-vivo* datasets is still needed to support the clinical value for our proposed method, which we consider as future work. More specifically, there are some limitations for both validated datasets. Because the US images during the procedures are different than those in our datasets, these differences can introduce and pose challenges. As for the catheter dataset, the isolated hearts were placed in a water tank. The heart chambers were filled with water to mimic the intervention procedure, although it should be blood for a real heart. This difference can lead to different image noise and contrast level between the instrument and the surrounding background. For the needle dataset, the real procedure would introduce more complex subcutaneous tissue and vessel structure, which complicates the acquired images. As a consequence, to make our proposed method applicable in real clinical practice, we will conduct more studies on *in-vivo* data in the future work. Moreover, the proposed method is applied to static images instead of 3D US video, which is commonly used during interventions. Therefore, a further study in 4D US is necessary in the future.

V. Conclusion

3D Ultrasound-guided therapy has been widely used, but it is difficult for a sonographer to localize the instruments in 3D US, because of the complex manual handling of instruments and the finally obtained US imaging plane. Therefore, automated detection of medical instruments in 3D US is required to reduce the operation effort and thereby increase the efficiency. Nevertheless, the existing instrument detection methods in 3D US are not efficient enough for real-time applications. In this paper, we propose a novel method based on a full-image level CNN to detect the medical instrument in 3D US volumetric data, which achieves a similar or higher performance than state-of-the-art method with 3-8 times higher efficiency, thereby paving the way for real-time applications.

References

[1] W. Xia *et al.*, "In-plane ultrasonic needle tracking using a fiber-optic hydrophone," *Medical Physics*, vol. 42, no. 10, pp. 5983–5991, 2015.

[2] C. Nadeau *et al.*, "Intensity-based visual servoing for instrument and tissue tracking in 3d ultrasound volumes," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 1, pp. 367–371, 2014.

[3] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[4] A. Ayvaci *et al.*, "Biopsy needle detection in transrectal ultrasound," *Computerized Medical Imaging and Graphics*, vol. 35, no. 7-8, pp. 653–659, 2011.

[5] Y. Zhao *et al.*, "Evaluation and comparison of current biopsy needle localization and tracking methods using 3d ultrasound," *Ultrasonics*, vol. 73, pp. 206–220, 2017.

[6] P. M. Novotny *et al.*, "Tool localization in 3d ultrasound images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2003, pp. 969–970.

[7] W. Qiu *et al.*, "Needle segmentation using 3d hough transform in 3d trus guided prostate transperineal therapy," *Medical Physics*, vol. 40, no. 4, p. 042902, 2013.

[8] M. Barva *et al.*, "Parallel integral projection transform for straight electrode localization in 3-d ultrasound images," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 55, no. 7, pp. 1559–1569, 2008.

[9] A. Pourtaherian *et al.*, "Medical instrument detection in 3-dimensional ultrasound data volumes," *IEEE Transactions on Medical Imaging*, vol. 36, no. 8, pp. 1664–1675, 2017.

[10] H. Yang *et al.*, "Catheter segmentation in three-dimensional ultrasound images by feature fusion and model fitting," *Journal of Medical Imaging*, vol. 6, no. 1, p. 015001, 2019.

[11] M. Uherčík *et al.*, "Line filtering for surgical tool localization in 3d ultrasound images," *Computers in Biology and Medicine*, vol. 43, no. 12, pp. 2036–2045, 2013.

[12] Y. Zhao *et al.*, "Automatic needle detection and tracking in 3d ultrasound using an roi-based ransac and kalman method," *Ultrasonic Imaging*, vol. 35, no. 4, pp. 283–306, 2013.

[13] C. R. Hatt *et al.*, "Enhanced needle localization in ultrasound using beam steering and learning-based segmentation," *Computerized Medical Imaging and Graphics*, vol. 41, pp. 46–54, 2015.

[14] P. Ambrosini *et al.*, "Fully automatic and real-time catheter segmentation in x-ray fluoroscopy," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 577–585.

[15] C. Mwikirize *et al.*, "Signal attenuation maps for needle enhancement and localization in 2d ultrasound," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 3, pp. 363–374, 2018.

[16] ——, "Learning needle tip localization from digital subtraction in 2d ultrasound," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–10, 2019.

[17] A. Pourtaherian *et al.*, "Robust and semantic needle detection in 3d ultrasound using orthogonal-plane convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 9, pp. 1321–1333, 2018.

[18] H. Yang *et al.*, "Catheter localization in 3d ultrasound using voxel-of-interest-based convnets for cardiac intervention," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 6, pp. 1069–1077, 2019.

[19] ——, "Efficient catheter segmentation in 3d cardiac ultrasound using slice-based fcn with deep supervision and f-score loss," in *IEEE International Conference on Image Processing*, 2019.

[20] ——, "Transferring from ex-vivo to in-vivo: Instrument localization in 3d cardiac ultrasound using pyramid-unet with hybrid loss," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2019.

[21] M. Arif *et al.*, "Automatic needle detection and real-time bi-planar needle visualization during 3d ultrasound scanning of the liver," *Medical Image Analysis*, vol. 53, pp. 104–110, 2019.

[22] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2015, pp. 234–241.

[23] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[24] Q. Dou *et al.*, "3d deeply supervised network for automated segmentation of volumetric medical images," *Medical Image Analysis*, vol. 41, pp. 40–54, 2017.

[25] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[26] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, 2015.