A multiple catheter tips tracking method in X-ray fluoroscopy images by a new lightweight segmentation network and Bayesian filtering

Hui Tang, Hao Kai Li, Chun Feng Yang, Jean-louis Dillenseger, Gouenou Coatrieux, Juan Feng, Shou Jun Zhou, Yang Chen

A multiple catheter tips tracking method in X-ray fluoroscopy images by a new lightweight segmentation network and Bayesian filtering

Hui Tang[1, 2, 3], Hao Kai Li[1], Chun Feng Yang[1, 2, 3, *], Jean-Louis Dillenseger[4, 5], Gouenou Coatrieux [6], Juan Feng[7], Shou Jun Zhou[8, *], Yang Chen [1, 2, 3]

1 School of Computer Science and Engineering, Southeast University, Nanjing, China

2 The Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, School of Computer Science and Engineering, Southeast University, Nanjing, China

3 Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China

4 Centre de Recherche en Information Biomédicale Sino-Francais, INSERM, University of Rennes 1, Rennes 35042, France

5 Univ Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France

6 Mines-Telecom Telecom Bretagne, INSERM U1101 La TIM, Brest, France

7 Shanghai United Imaging Company, Shanghai, China

8 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

Abstract

During percutaneous coronary intervention (PCI), the guiding catheter plays an important role. Tracking the catheter tip placed at the coronary ostium in the X-ray fluoroscopy sequence can obtain image displacement information caused by heart beating, which can help dynamic coronary roadmap (DCR) overlap on X-ray fluoroscopy images. Due to the low exposure dose, the X-ray fluoroscopy is noisy and low contrast, which causes some difficulties in tracking. We developed a new catheter tip tracking framework in this paper. First, a lightweight efficient catheter tips segmentation network is proposed and boosted by a self-distillation training mechanism. Then, the Bayesian filtering post-processing method is used to consider the sequence information to refine the single image segmentation results. By separating the segmentation results into several groups based on connectivity, our framework can track multiple catheter tips. The proposed tracking framework is validated on a clinical X-ray sequence dataset.

Keywords: X-ray fluoroscopy sequence, catheter tips segmentation network, self-distillation training, multi-objective Bayesian filtering

1 Introduction

Coronary artery disease is one of the leading causes of death in worldwide. Percutaneous coronary intervention (PCI) is nowadays the preferred minimally invasive surgery for the treatment of coronary artery disease due to its advantages low risk of intraoperative complications, quick recovery, and excellent curative effects. PCI contains two categories: the balloon angioplasty and the implantation of the intracoronary stent. In both ones, interventional cardiologists push forward a catheter through complex anatomical interventional pathways to the appropriate location based on the position of the catheter tip visible in a fluoroscopy sequence [1].

Dynamic Coronary Roadmapping (DCR) [2, 3] is an advanced technique that superimposes contrast-enhanced coronary angiographic images onto live X-ray fluoroscopy images. DCR significantly improves visualization and reduces contrast agent usage during procedures. However, in DCR implementation, accurate localization of the catheter tip is vital for achieving effective image registration and overlay. Due to the dynamic nature of the beating heart, the coronary arteries and catheter tip may experience motion-induced displacements during the cardiac cycle. Such displacements can result in misalignment between the angiographic image and the real-time fluoroscopy, compromising the precision and reliability of DCR. Tracking the catheter tip placed at the coronary ostium in the X-ray fluoroscopy sequence can obtain image displacement information caused by heart beating, which can help dynamic coronary artery roadmap overlap on X-ray fluoroscopy images. So, it is necessary to check if the catheter tip has been placed at the correct position and then, during the intervention, the catheter tips should be always tracked to give the position information.

In order to reduce ionizing radiation for the patient and interventional cardiologists during image acquisition, low power X-rays need to be used. The counterpart is that they produce noisy, low-contrast images, which makes it difficult to identify the position of the catheter tip. Therefore, it is imperative to accurately localize and track the catheter tip in the image sequence to perform guidance.

Figure 1 A typical frame of catheter tips segmentation in 2D X-ray fluoroscopy sequence. Left: the input image. Middle: position of the catheter tip. Right: the catheter tip is overlayed in red in the input image.

The target of pixel-wise segmentation is to get a binary mask of the region of interest. In this work, the target region is the tip of the guiding catheter, as shown in Figure 1. In the original image, the target catheter tip has a relatively low gray value. Many works have been proposed to segment and track the whole catheter in X-ray fluoroscopy images [4-8]. Only a few works are reported about the catheter tips segmentation. So, we propose a tracking framework only for the catheter tips in this paper. Teixeira et al. [9] regard the catheter tips detection task as a landmark localization task

and introduce a loss function called “Adaloss” which adapts the target precision during the training. However, this method cannot fully segment the catheter tip and the “Adaloss” can’t be used in the segmentation task. Ma et al. [2] track the catheter tip by concatenating the output of a convolutional neural network and the output of a particle filtering framework, but it doesn’t take the high resource occupancy of the network into account and cannot track multiple targets per frame.

Different from the normal segmentation tasks, the segmentation of the catheter tip is a challenging task for the following four reasons: (1) the source X-ray fluoroscopy images have disadvantages of low contrast, low signal-to-noise ratios, and non-uniform illumination; (2) the movement of the catheter tip is irregular between frames due to the heartbeat, which makes it hard to track; (3) the presence of other body tissues with similar characteristics is likely to cause false detection; (4) in order to apply the catheter tips segmentation method to the mobile medical equipment, it has high requirements in low computing resource occupancy and high inference speed.

To solve the four problems mentioned above, a novel automatic framework for catheter tips segmentation and tracking is proposed in this paper. In the catheter tips segmentation stage, a new convolutional neural network based on the encoder-decoder architecture is proposed. The lightweight encoder part of the proposed model is composed of the inverted residual convolution block used in MobileNetV3 [10], which can improve the inference speed with acceptable accuracy. To overcome image quality problems such as low signal-to-noise ratio, a pyramid-style feature combination step is designed in the decoder to enrich features with high-level context. In addition, the self-distillation training mechanism is used to reinforce representation learning of this network. Finally inspired by Ma et al. [2], an improved temporal tracking method based on Bayesian filtering is used in the catheter tip tracking stage. We expanded the work from only tracking a single catheter tip to the ability of tracking multiple catheter tips.

In conclusion, the main contributions are:

1) A new lightweight segmentation network architecture is proposed to accurately segment the catheter tip in the X-ray fluoroscopy image. It includes a lightweight encoder and a multi-scale feature fusion decoder.

2) A self-distillation mechanism is applied in the training phase of the catheter tips segmentation network. It improves the representation learning of itself without bringing any computational cost during the inference phase.

3) A tracking method based on Bayesian filtering is applied and improved. The improved method considers the convolutional neural network’s segmentation result of the catheter tips as the likelihood terms of Bayesian filtering and uses multi-objective Bayesian filtering to track multiple catheter tips simultaneously.

The paper is organized as follows. In Section 2, the three contributions mentioned above will be separately introduced in detail: lightweight catheter tips segmentation network architecture, self-distillation training mechanism, and multi-objective Bayesian filtering. Section 3 is the configuration related to the experiments, including the dataset, data argument method, comparison work, implementation details, and evaluation metrics. Moreover, experiments on the three innovation points verify the effectiveness of each innovation point. Section 4 is the summarization of this work.

2 Methods

The proposed catheter tips segmentation method contains three main parts: network architecture, training mechanism, and post-processing. Lightweight network architecture and the training mechanism are proposed to segment the catheter tips. The trained network is then light enough to be used on mobile medical equipment. Nevertheless, post-processing is still needed to reduce the false segmentation rate of the model which is quite high because of the intravascular contrast agent.

2.1 Catheter Tip Segmentation Network

The proposed catheter tips segmentation network is an encoder-decoder structure comprised of four encoders E1~E4 and one decoder D, as shown in Figure 2. To build a mobile model, MobileNetV3 [10] is chosen as the lightweight encoder and connected with one selected decoder. The encoder is built on efficient building blocks which introduce lightweight attention modules based on squeeze and excitation (SE) [11] into the linear bottleneck and inverted residual structure. This block is named Mobile Block, which has two types: with and without a shortcut. The detail structure of the Mobile Block with a shortcut is illustrated in Figure 3. The Mobile Block without a shortcut has a down-sampling operation since the size of the input is not equal to the output. The other components of the two types are the same: a 1 x 1 expansion convolution layer followed by a depth-wise separable convolution layer, a squeeze and excitation layer before a 1 x 1 projection layer.

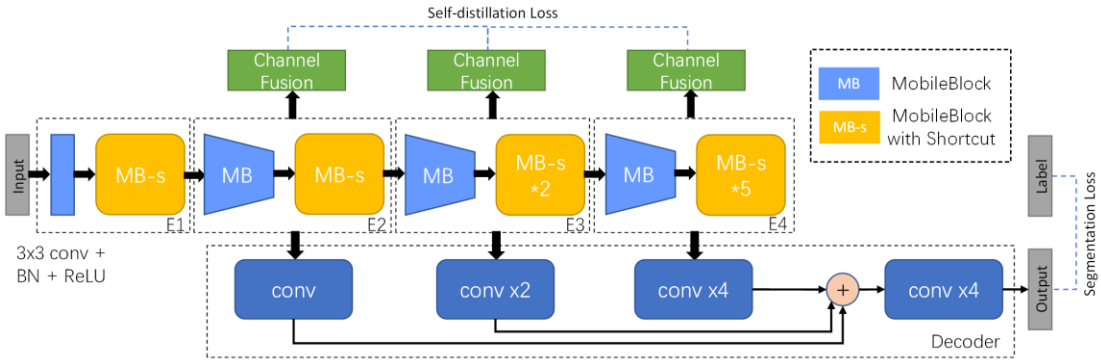


Figure 2 The architecture of the proposed catheter tips segmentation network. '*' means multiple blocks concatenation. 'x2', 'x4' means bilinear up-sampling operation.

Mobile Block

The detailed structure of the Mobile Block with a shortcut is shown in Figure 3. The first 1 x 1 convolution layer expands the number of input channels, which is opposite to the residual block [12] that decreases the number of input channels first. The second layer is called a depth-wise convolution, it performs lightweight filtering by applying a single convolutional filter to each input channel which is responsible for computing the feature inside per channel. The third layer is the squeeze and excitation module [11], which captures attention from the expanded representation. The last layer is a 1 x 1 convolution layer, called a point-wise convolution, which computes the linear combinations between each channel and decreases the number of input channels. Finally, the

input and output are connected with a residual connection only if they have the same number of channels. This inverted residual structure maintains a compact representation at the input and the output while expanding to higher-dimensional feature space and extracting attention by the SE module internally to increase the expressiveness of nonlinear channel transformations.

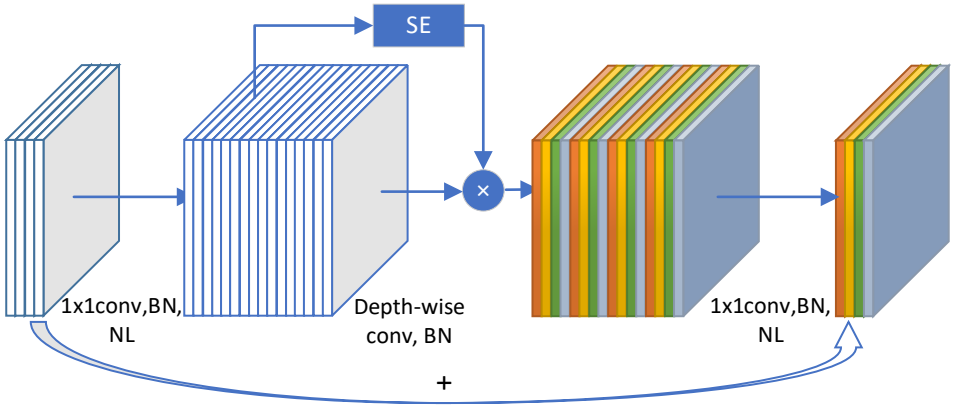


Figure 3 The architecture of the Mobile Block with a shortcut (MB-s).

Encoder

The encoder is built on a combination of the Mobile Block. As shown in Figure 2, the encoder of the catheter tips segmentation network contains an initial fully convolution layer transforming from 1 to 16 channels, with stride 2, followed by 12 Mobile Blocks and those blocks are divided into four parts: E1~E4. With the exception of the first encoder, E2~E4 starts with a Mobile Block, with stride 2, which halves the length and width of the input. The encoder structure is a lightweight version model proposed by Howard et al. [10], it uses Network Architecture Search (NAS) to optimize each network block and search per layer for the number of filters, which gives full play to the feature extraction ability of the encoder.

Decoder

The proposed decoder module is illustrated in the lower part of Figure 2. According to DeepLabV3+ [13], not all the features of the encoder modules are essential to contribute to the decoder module. Therefore, three main features from all the mobile blocks are extracted. Specifically, the outputs of E2 ~ E4 are considered as low-level, middle-level, and high-level features respectively. The decoder is designed as an efficient feature up-sampling module to fuse all level features from the encoder. In order to reduce the computation complexity, a 1 x 1 convolution followed by a bilinear up-sampling operation is firstly applied to reduce channels and match the size of the low-level feature maps. Then, all the three-level features are added together and up-sampled by a factor of 4 to make the final prediction.

Segmentation loss

The segmentation loss L_{seg} is the standard Dice loss [14]. The Dice loss aims at increasing the intersection of union between the predicted catheter tip image y and ground truth catheter tip mask y', defined as:

$$L_{seg}(y, y') = 1 - \frac{2\sum_k {y'}_k y_k}{\sum_k {y'}_k + \sum_k y_k} \quad (1)$$

where ${y'}_k$ and y_k are respectively the pixels of the mask y' and the output image y.

2.2 Self-distillation Training Mechanism

Compared to the normal knowledge distillation mechanism [15], self-distillation (SD) allows the catheter tips segmentation network to the reinforce representation learning of itself without the need of a teacher network. In our specific case, the SD training mechanism performs layer-wise and top-down feature distillation to enhance the representation learning process. More specifically, it exploits the output of the channel fusion modules which are linked out from the output of E2 ~ E4, as shown in Figure 2.

Activation maps generation

To generate the distillation targets for lower layers, each layer needs to be converted to an activation map with the same size. The activation generation mapping function is defined as $\mathrm{G}: R^{C_t \times H_t \times W_t} \rightarrow R^{H_t \times W_t}$, where C_t, H_t, and W_t denote the size of the channel, height, and width, respectively. The mapping function computes statistics of features values across the channel dimension, which is represented by "Channel Fusion" part in Fig. 2. It is defined as follows:

$$\mathrm{G}_{sum}^{2}(x_t) = \sum_{i=1}^{C_t} |x_{ti}|^2 \tag{2}$$

where $x_t \in R^{C_t \times H_t \times W_t}$ denotes the t-th output of the three encoders, x_{ti} denotes the i-th slice of x_t in the channel dimension. From the network structure illustrated in Figure 2, we can get that the range of t is 1 to 3.

Self-distillation for training

The layer-wise distillation loss is defined as follows:

$$L_{distill}(x_t, x_{t+1}) = \sum_{t=1}^{2} L_d\big(\Phi(x_t), \Phi(x_{t+1})\big) \tag{3}$$

where $\Phi(.) = B(\mathrm{G}_{sum}^{2}(.))$ denotes the activation maps generation and $B(.)$ denotes a conditional Bilinear up-sampling, that is, if the size is not matched then the up-sampling is use. Here, L_d is typically defined as a L_2 loss.

The total training loss is defined as follows:

$$L = L_{seg}(y, y') + \alpha L_{distill}(x_t, x_{t+1}) \tag{4}$$

where the parameter α is a coefficient which balances the influence of segmentation loss and distillation loss during the training phase, y is the segmentation map produced by the network and y' is the binary mask of the catheter tip.

2.3 Multi-objective Bayesian Filtering Tracking Method

In the above segmentation task, only the image characteristics are considered without any context information between the neighbor frames in the video so that the false segmentation is inevitable. In order to solve this problem, a tracking method based on Bayesian filtering is applied as post-process to reduce the false detection rate and improve the robustness of the whole process. Bayesian Filtering tracking is a temporal tracking method which uses information from the previous frames. It can put additional constraints in the model which makes the tracking more robust.

Theory of Bayesian Filtering

Bayesian filtering is a state-space approach aiming at estimating the true state of a system that changes over time from a sequence of noisy measurements made on the system [2]. So, it is applied to remove false segmentations by considering them as noise. The formula can be written as:

$$P(\mathbf{x}_t \mid \mathbf{z}_{0:t}) \propto P(\mathbf{z}_t \mid \mathbf{x}_t)P(\mathbf{x}_t|\mathbf{z}_{0:t-1}) \tag{5}$$

where $\mathbf{x}_t$ denotes the pixel coordinates of the catheter tip in the *t*-th frame and $\mathbf{z}_t$ denotes the *t*-th image in the X-ray sequence.

The posterior probability $P(\mathbf{x}_t \mid \mathbf{z}_{0:t})$ means the estimation of $\mathbf{x}_t$ in the t-th frame based on the set of all available observations $\mathbf{z}_0, \ldots, \mathbf{z}_t$ up to frame *t*. According to Eq. (5), the posterior probability depends linearly on the likelihood $P(\mathbf{z}_t \mid \mathbf{x}_t)$ and the prior probability $P(\mathbf{x}_t \mid \mathbf{z}_{0:t-1})$. According to the total probability formula, the prior probability $P(\mathbf{x}_t \mid \mathbf{z}_{0:t-1})$ can be written as:

$$P(\mathbf{x}_t \mid \mathbf{z}_{0:t-1}) = \sum_k P(\mathbf{x}_t \mid \mathbf{x}_{t-1})P(\mathbf{x}_{t-1}^k \mid \mathbf{z}_{0:t-1}) \tag{6}$$

where *k* denotes the segmented multiple catheter tips in the *t-1*th frame. The posterior probability $P(\mathbf{x}_t \mid \mathbf{z}_{0:t})$ can be simplified as a belief probability, denoted by $Bel(\mathbf{x}_t)$. Since $\mathbf{x}_t$ represents the coordinates of the catheter tip. Eq. (5) can be rewritten as:

$$Bel(\mathbf{x}_t) \propto P(\mathbf{z}_t \mid \mathbf{x}_t) \sum_k P(\mathbf{x}_t \mid \mathbf{x}_{t-1})Bel(\mathbf{x}_{t-1}^k) \tag{7}$$

In our application case, the segmentation network takes the *t*-th frame in the X-ray image sequence as the input and outputs a probability diagram with the same size. The output result indicates the probabilities of each location to belong to the catheter tip. This probability map can be considered as the likelihood $P(\mathbf{z}_t \mid \mathbf{x}_t)$ in Eq. (7). The probability $Bel(\mathbf{x}_{t-1})$, also the $P(\mathbf{x}_{t-1} \mid \mathbf{z}_{0:t-1})$ in Eq. (5), denotes the previous result of the Bayesian filtering. The probability $P(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ represents the mapping of coordinates from the previous frame to the current frame. Here it is computed by the optical flow method [16]. In the summation term, the integral of the product of the two probabilities indicates that the model estimates the current catheter tip coordinates based on the result of the last filtering and the coordinate mapping from the previous frame to the current frame.

Multi-objective tracking

There are two steps in Bayesian filtering: prediction and update. In the prediction stage, the prior probability $P(\mathbf{x}_t \mid \mathbf{z}_{0:t-1})$ is replaced by the locations of the catheter tip in the *t*-th frame which are predicted based on the previous *t*-1 frames. In the update stage, the posterior probability $Bel(\mathbf{x}_t)$ is generated according to Eq. (7), which multiplies the prior probability of the prediction stage by the current (*t*-th frame) probability map.

In our specific clinical data and unlike the situation handled by Ma H. et al. [2] with only one catheter tip in the frames, there are multiple catheter tips in one frame. To be compatible with the multiple catheter tips situation, the segmentation results of the first frame are separated into several groups according to the connectivity information. Each group stands for one possible catheter tip. For each group, Bayesian filtering is applied to update the information of the current frame, based on the following rules.

The output of the segmentation network is separated into two classes: active and inactive ones. In the prediction stage, all groups of the catheter tip locations will be predicted. In the update stage, each predicted location will be searched in a limited distance and updated to the nearest location

according to the output of the segmentation network. Among the updated locations, the ones belonging to the active group will be displayed in the current *t*-th frame and the times of being updated will be calculated for further usage. The ones belonging to the inactive group or without enough updated times keep the current state and wait for being updated or activated in the next update stage. After each update phase, the active and inactive groups are re-divided according to the updated times of the locations. The multi-objective Bayesian filtering tracking method of catheter tip is summarized in Algorithm 1.

Algorithm 1 multi-objective Bayesian filtering tracking method of catheter tip

structure {**x** = coordinate, times=0, active_flag=false}, denotes one tip
Input: image sequence $\mathbf{z}_{0:T}$, the segmentation model *Seg*, number of frames *T*
Initialize:
Connection domain analysis of $Seg(\mathbf{z}_0) \rightarrow \mathbf{x}_0^{0:k}$
Put the initial results to the inactive set $\mathbf{x}_0^{0:k} \rightarrow S_{inactive}$
The number of targets: $N_s = length(S_{inactive}) + length(S_{active})$
for *t*=1:*T* **do**
  Compute optical flow O_{t-1} from $\mathbf{z}_{t-1}$ to $\mathbf{z}_t$, which is $P(\mathbf{x}_t \mid \mathbf{x}_{t-1})$
  Calculate the probability map from $Seg(\mathbf{z}_t)$, denote as map_t
  for $k = 1{:}N_s$ **do**
    Predict the current connection domain according to the optical flow $O_{t-1}(\mathbf{x}_{t-1}^{k-1}) \rightarrow \mathbf{x}_t^k$
    if $map_t(\mathbf{x}_t^k) > 0.5$ **then**
      Update the current tracking times $\mathbf{x}_t^k.times + 1 \rightarrow \mathbf{x}_t^k.times$
    end if
    if $\mathbf{x}_t^k.times > t/3$ **then**
      Put the current instance to the active set $\mathbf{x}_t^k \rightarrow S_{active}$
      Show the result
    else
      Put the current instance to the inactive set $\mathbf{x}_t^k \rightarrow S_{inactive}$
    end if
  end for
end for

3 Experiments and Results

3.1 Dataset

The data used in this paper are provided by the Shanghai United Imaging Company. They consist of 9884 anonymized clinical X-ray fluoroscopic images selected from 173 angiogram sequences. The size of the images is 512 x 512 and the pixel intensity range is [0, 255]. The labels of the catheter tips are manually annotated by two experienced cardiac radiologists with double-checking.

The dataset is divided into three sets: the training dataset with 7885 images, the validation dataset with 1072 images and the test dataset with 927 images (around 8:1:1). To avoid the

inappropriate dataset division, the above three datasets are divided manually to ensure that images from the same sequence (patient) are put in the same set. We also use an image transformation for data enhancement to overcome the over-fitting problem. In the training phase, both the input images and the corresponding label are randomly flipped and rotated between -15° and 15°. Each training round stopped at the 200th epoch.

3.2 Baseline and Implementation Details

There is no network model dedicated to segment the catheter tip. So, five representative segmentation networks: U-net [17], Attention U-net [18], DeepLabV3+ [13], DFA-net [19] and HRNet [20] are chosen to be compared with the proposed lightweight catheter tip segmentation network. U-net is a classical medical image segmentation network that is widely used as the baseline of other work. Attention U-net is a variant of U-net, which has better feature extraction ability than U-net. DeepLabV3+ is a well-known image segmentation network dealing with natural light scenes. It combines spatial pyramid pooling module into the encoder-decoder structure which made it state-of-the-art at that time. DFA-net is a lightweight semantic segmentation network with a balance of speed and accuracy. HRNet achieved the state-of-the-art on several challenging datasets in computer vision with an encoder-decoder architecture.

The proposed method is implemented in Python and all segmentation networks are implemented by PyTorch library (version 1.6.0) with one NVIDIA RTX 2080ti GPU. In the training phase of all segmentation networks, all networks are trained using Adam optimization with weight decay set as 10^{-4}. The base learning rate is set as 10^{-3} and the learning rate per iteration is calculated by "poly" learning rate policy. Specifically, the base learning rate is multiplied by $(1-\frac{iter}{max\_iter})^{0.9}$ per iteration.

3.3 Evaluation Metrics

Two evaluation metrics *Dice* and *Precision* coefficient are chosen to measure the accuracy of the segmentation results, which are defined as following.

$$Dice = \frac{2\|X \cap Y\|}{\|X\|+\|Y\|} = \frac{2TP}{2TP+FP+FN} \tag{8}$$

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

where TP (True Positive) denotes the number of correctly predicted catheter tip pixels, FP (False Positive) denotes the false positive pixels, and FN (False Negative) indicates the false negative pixels. The $Dice$ coefficient is used to measure the similarity between the segmentation result X and the ground truth mask Y. Since the catheter tip segment task is sensitive to the false segmentation due to the influence of contrast agents, the $Precision$ coefficient is used to evaluate the segmentation accuracy.

Besides the above two coefficients, the number of trainable parameters in the segmentation network and the inference time are also used to evaluate the network efficiency. The runtime is obtained by calculating the average inference time of 100 input samples.

To evaluate the tracking algorithm on the sequence, the average false detection rate on the image-wise and the sequence-wise are estimated to measure the algorithm's performance. The

intersection-over-union (IoU) between tracking results and labels is chosen to judge whether a catheter tip is correctly detected. Tracking results whose IoUs are less than 0.5 are considered as FP. Then, the average false detection rate per image f_i and per sequence f_s are defined as:

$$f_i = \frac{N_{FP}}{N_{total}} \tag{9}$$

$$f_s = \frac{\sum_{i=0}^{n} \frac{N_{FP}}{N_i}}{N_s} \tag{10}$$

where N_{FP} is the number of images that have false positive detection; N_{total} is the number of images; N_i is the number of images of the i-th sequence; N_s is the number of sequences.

3.4 Experimental results

The decoder architecture

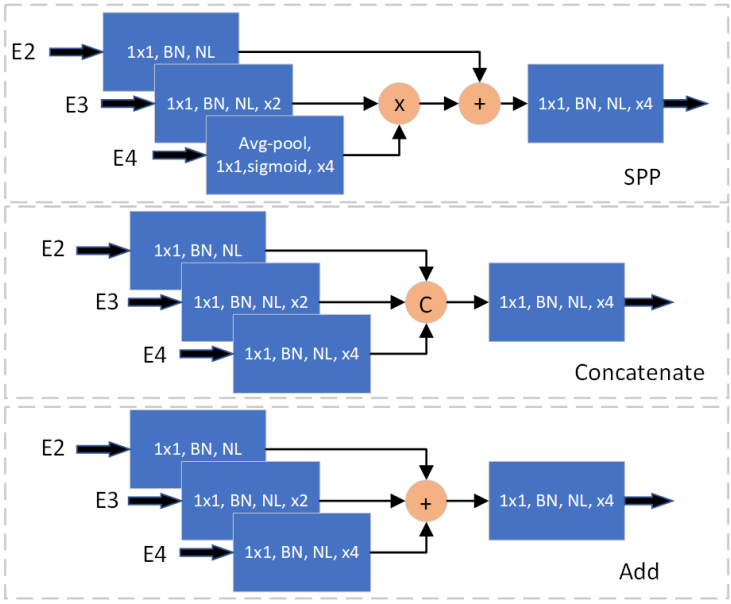


Figure 4 Three different decoder structures. From top to bottom: Spatial Pyramid Pool (SPP), Concatenate, Add.

For the decoder, three commonly used architectures are chosen for comparison. They are Lite R-ASPP decoder (Spatial Pyramid Pool, SPP) used in the model MobileNetV3 [10], "Concatenate" used in the model U-net [17] and "Add" used in the model DFA-net [19]. The structures of the three decoders are shown in Figure 4. The selection decision of the decoder is made according to the experimental results.

All of the decoder architectures take the same multi-level features provided by encoder blocks as input and combine features with different operations. Table 1 summarizes the experimental results with the three decoder architectures. It can be observed that the results obtained with the addition of multi-level features (Add) are better than the results of the other two for all the metrics. The catheter tip is in a rectangular ROI with a low gray value in X-ray images which means that the feature is simple. Combining features by addition not only makes full use of information at the channel level, but also reduces the number of parameters and the runtime.

The Dice and Precision results of the segmentation on the validation dataset with the three decoders are shown in Figure 5. It can be seen that the Add decoder (pink in Figure 5) achieves the best performance throughout the training phase. This reliable performance indicates that the selected decoder architecture is suitable for the catheter tips segmentation task.

Table 1 Performance of different decoder architectures.

Decoder architecture	Dice (%)	Precision (%)	Parameters (million)	Runtime (ms)
SPP	74.27	78.07	1.4	9.21
Concatenate	75.68	80.45	**1.2**	8.91
Add	**77.47**	**81.82**	**1.2**	**8.72**

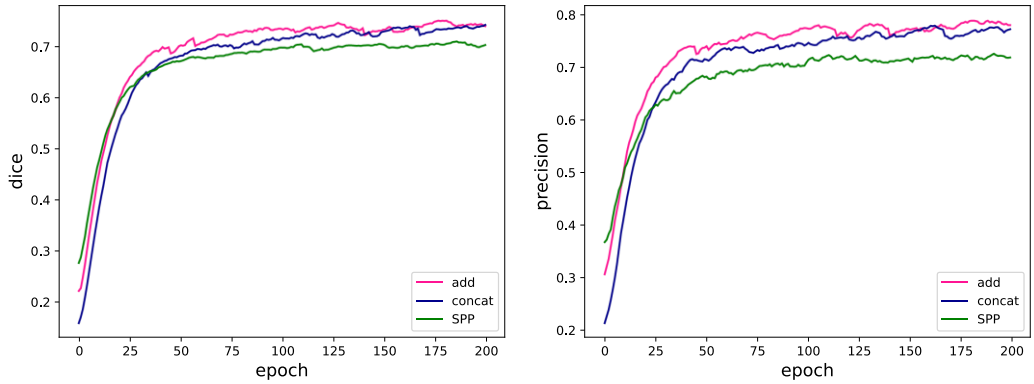


Figure 5 Comparison of the segmentation performance with different architectures. left: the evolution of the dice coefficient value with increasing number of epochs, right: the evolution of the precision coefficient value with increasing number of epochs.

The self-distillation training mechanism

The proposed catheter tip segmentation network is trained on 300 epochs with and without self-distillation. As recommended by Hou et al. [21], we added the self-distillation training mechanism at the 200-th epoch, which achieves higher performance than adding it at the beginning of the training phase.

Table 2 Performance of different weighting factor settings.

Weighting factor	Dice (%)	Precision (%)
0	77.47	81.82
0.001	77.03	81.38
0.0007	76.77	81.33
0.0005	**77.65**	**83.38**
0.0003	77.48	82.36
0.0001	77.35	83.00

Some experiments were done to further investigate the weighting factor α in the loss function. When α equals to zero, it means that the self-distillation is not taken into account. The test interval of α is determined by the order of magnitude of the two losses L_{seg} and $L_{distill}$ in Eq. (4). Table 2 shows that with the increasing weighting factor of the self-distillation loss, the performance of the model fluctuates but reaches a peak on both Dice and Precision when the factor is 0.0005. With the

best experimental factor of 0.0005, the self-distillation training mechanism increases the precision of the lightweight catheter tip segmentation network from 81.82% to 83.38% with the same model size and inference time. This proves that the self-distillation training mechanism improves the representation learning of itself without bringing any additional computational cost.

The proposed catheter tip segmentation network

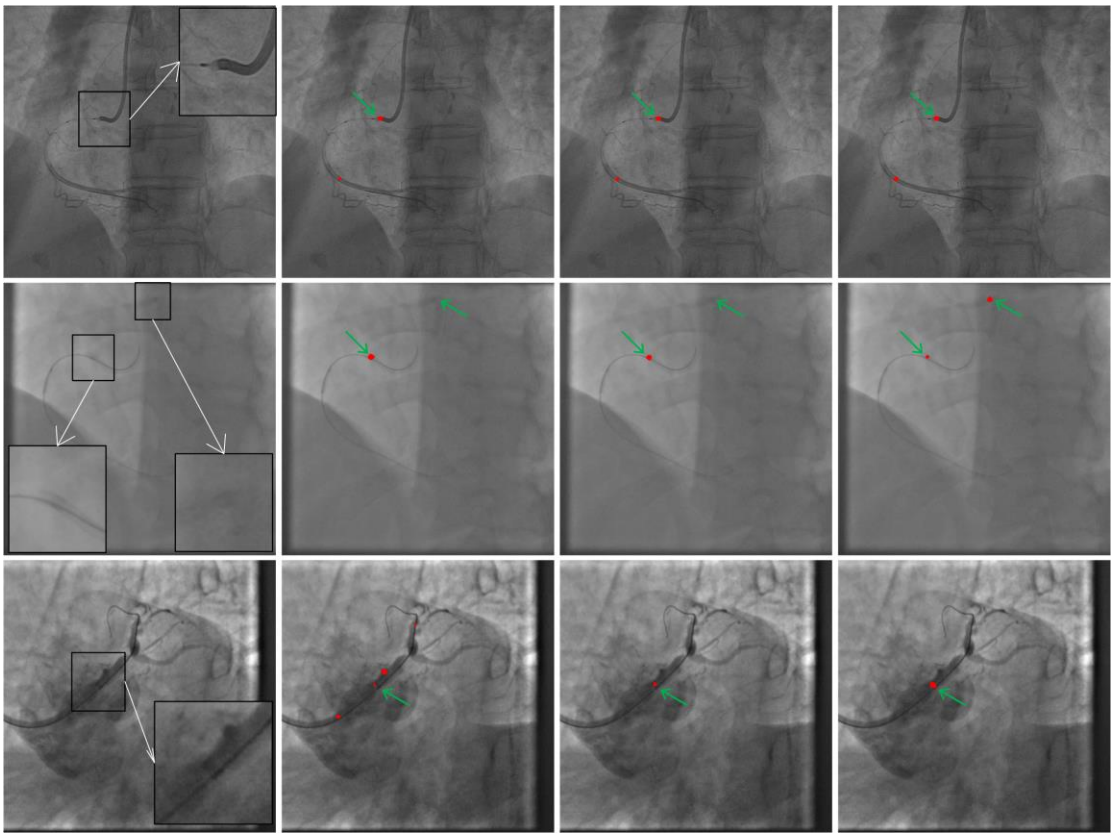
Several segmentation networks listed in Section 3.2 are tested for comparison in this section. All segmentation networks are trained under the same configurations, such as dataset, number of training epochs, loss function, and so on. As shown in Table 3, in terms of accuracy (Dice and Precision), the proposed method is at the same level with DeepLabV3+ [13]. With the self-distillation training mechanism (SD), the proposed method achieved the best accuracy among all the tested models. Because the proposed network structure has less parameters, it has better runtime than all the other tested models. With self-distillation training, the accuracy of the proposed model is improved without any increase on parameter number.

Table 3 Performance of different network architectures. According to the number of parameters, the comparison studies are divided into two groups: large models (the first group) and small models (the second group).

Models	Dice (%)	Precision (%)	Parameters (million)	Runtime (ms)
U-Net [17]	76.05	81.35	31.0	62.79
Attention U-Net [18]	76.40	79.58	34.9	83.87
DeepLabV3+ [13]	**77.72**	82.93	40.5	77.48
HRNet-big [20]	76.56	81.24	65.6	54.92
DFA-Net [19]	71.40	78.26	1.8	47.33
HRNet-small [20]	75.05	80.91	9.9	19.64
Proposed net without SD	77.47	81.82	**1.2**	**8.72**
Proposed net with SD	77.65	**83.38**	**1.2**	**8.72**

Some visual segmentation results of some cases generated by U-Net, DeepLabV3+, and the proposed work are shown in Figure 6. The first column is the original X-ray fluoroscopy image. In each image the regions of interest which contain the catheter tip are placed in the irrelevant area. The original image in the first row only includes one catheter tip, but all the tested networks tell two catheter tips. One possible reason could be that the characteristics of the catheter tip are too simple to be distinguished. So, it is necessary to use post-processing to eliminate such false detections. The original image in the second row includes two catheter tips in which the upper one is more difficult to recognize. The U-net (the second column) and the DeepLabV3+ (the third column) only tell the other catheter tip and lose the upper one. The proposed network can accurately detect both the two ones, which indicates the power of the proposed network. The original image in the last row includes one catheter tip. All the three networks tell the correct one, but U-net detects some redundant regions with similar features, which indicates that its generalization ability is the worst of the three. It is in accordance with the numerical results shown in Table 3. DeepLabV3+ correctly detect the target area, but the segmentation results are incomplete because of the lack of a well-designed decoder architecture. Overall, the proposed network performs the best in the single image catheter tips

segmentation task. It has a strong generalization ability to identify target with insignificant features and the decoder architecture is also well designed for the target task.



(a) Original images (b) U-net (c) DeepLabV3+ (d) Proposed method

Figure 6 The results of different segmentation network. (a) X-ray fluoroscopy images, the magnified areas indicate the correct location of the catheter tips. (b) to (d) are segmentation results of U-net, DeepLabV3+, and the proposed method. The red areas are the segmented catheter tips. The green arrows point at the ground-truth locations of the catheter tips.

The Bayesian Filtering based post-processing Tracking Method

After segmentation, the Bayesian filtering is used to further refine the segmentation results. To show the effect of the Bayesian filtering based post-processing tracking method, DeepLabV3+ [13] is chosen for comparison because it has the best segmentation performance among other large networks. The numerical metrics are the average false detection rate per image (Eq. (9)) and per sequence (Eq. (10)). As shown in Table 4, the false detection on image-wise is reduced by about 5 times and that on sequence-wise is reduced by about 4 times. This means that the introduction of temporal information can significantly improve the accuracy of the segmentation.

Table 4 Average false detection rate on image-wise and sequence-wise before and after Bayesian filtering.

methods	Image-wise (%)		Sequence-wise (%)	
	Before	After	Before	After
DeepLabV3+	21.29	**4.54**	18.45	**4.11**
Proposed network with SD	10.60	**2.36**	8.72	**2.08**

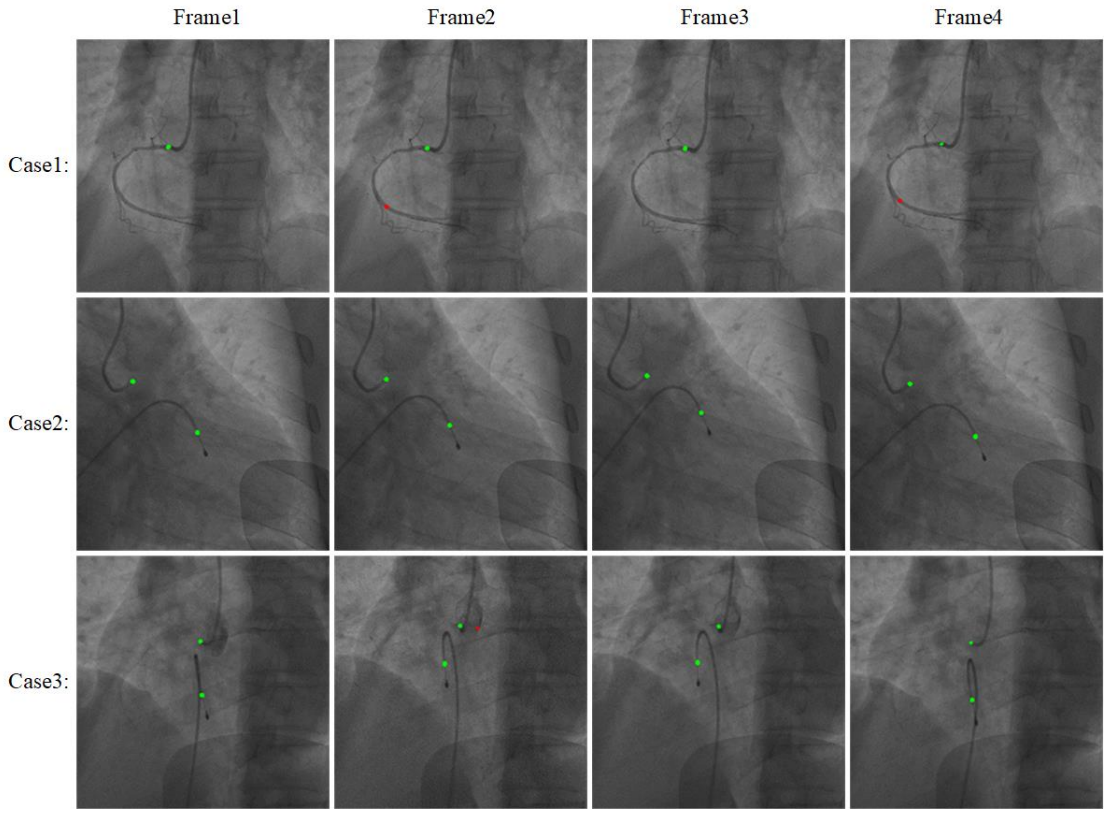


Figure 7 Tracking results of the catheter tip after applying the multi-objective Bayesian filtering method. Each row is the result of one case. The areas in red or green are both the results of segmentation network. The green area is the final positive results of the proposed tracking method. The red area is positive results of the segmentation but negative results of the tracking method, so the red areas are mis-segmented regions excluded through the tracking method. There is one catheter tip in Case 1, two catheter tips in Case 2 and Case 3.

Figure 7 shows some qualitative results of the multi-objective Bayesian filtering tracking method. There is one catheter tip in Case 1, two tips in Case 2 and Case 3. The segmentation results are first marked in red and then when the segmented tips are confirmed by the tracking method, they will be marked in green. So the mis-segmented regions are marked in red. It can be seen in Case 1, which is from the same case as the first row of Figure 6, that the false segmentation result (red area) is corrected after applying the post-processing tracking method (Frame 2 and Frame 4). From Case 2 and Case 3, we can see that the improved Bayesian filtering method is able to track multiple catheter tips and exclude the mis-segmented regions in the cases with multiple catheter tips.

4 Conclusion and discussions

In this paper, a new method is proposed and validated to segment the catheter tip in X-ray image sequences. The method consists of two steps. First, a segmentation network is designed and trained using the self-distillation training mechanism to segment the catheter tips. Second, the segmentation result is refined by the Bayesian filtering post-processing method that merges the information from the previous images in the same sequence. The proposed method is validated on several clinical X-ray fluoroscopy sequences. The experiment results show that compared to some state-of-art segmentation models, the proposed method achieved the best segmentation accuracy

with the least runtime. The Bayesian filtering post-processing method can reduce the false detection rate.

Following the injection of contrast agents, the catheter tip segmentation network often produces numerous false positive detections, obscuring the true catheter tip target. Involving the temporal information to infer the current position of the catheter tip may potentially enable concurrent contrast agent injection and catheter manipulation, thereby shortening the surgical process. Although the proposed catheter tip segmentation and tracking method gives some positive results, there is still some limitations in real applications. The X-ray images are two-dimensional, so the projection direction may be parallel to the catheter tip. In this case, the catheter tip appears as a ring style shape in the X-ray images. We didn't involve this kind of cases in the training set, so the model cannot be used in this special situation, which is the most severe restriction associated with the dataset.

Acknowledgement

This work was supported in part by the National Key R&D Project of China (2018YFA0704102, 2018YFA0704104 and 2022YFC2408500), in part by the National Natural Science Foundation of China (No. 81827805), in part by Natural Science Foundation of Guangdong Province (No. 2023A1515010673), and in part by Shenzhen Technology Innovation Commission (No. JCYJ20200109114610201, JCYJ20200109114812361, and JSGG20220831110400001), in part by the Shenzhen Engineering Laboratory for Diagnosis & Treatment Key Technologies of Interventional Surgical Robots (XMHT20220104009), and in part by the Key Research and Development Programs in Jiangsu Province of China under Grant BE2021703 and BE2022768. The authors thank the Shanghai United Imaging Company for the data support.

References

[1] S. Kato *et al.*, "Assessment of coronary artery disease using magnetic resonance coronary angiography: a national multicenter trial," *Journal of the American College of Cardiology,* vol. 56, no. 12, pp. 983-991, 2010.

[2] H. Ma, I. Smal, J. Daemen, and T. van Walsum, "Dynamic coronary roadmapping via catheter tip tracking in X-ray fluoroscopy with deep learning based Bayesian filtering," *Medical Image Analysis,* vol. 61, Apr 2020, Art no. 101634.

[3] Y. Zhu, Y. H. Tsin, H. Sundar, and F. Sauer, "Image-Based Respiratory Motion Compensation for Fluoroscopic Coronary Roadmapping," in *13th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Sep 20-24 2010, vol. 6363, in Lecture Notes in Computer Science, 2010, pp. 287-294.

[4] P. Ambrosini, D. Ruijters, W. J. Niessen, A. Moelker, and T. van Walsum, "Fully Automatic and Real-Time Catheter Segmentation in X-Ray Fluoroscopy," in *Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017*, Cham, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds., 2017 2017: Springer International Publishing, pp. 577-585.

[5] D. Kim, S. Park, M. H. Jeong, and J. Ryu, "Registration of angiographic image on real-time fluoroscopic image for image-guided percutaneous coronary intervention," *International Journal of Computer Assisted Radiology and Surgery,* vol. 13, no. 2, pp. 203-213, Feb 2018.

[6] Y. L. Ma *et al.*, "Real-time x-ray fluoroscopy-based catheter detection and tracking for cardiac electrophysiology interventions," *Medical Physics,* vol. 40, no. 7, Jul 2013, Art no. 071902, doi: 10.1118/1.4808114.

[7] X. L. Wu, J. Housden, Y. L. Ma, B. Razavi, K. Rhode, and D. Rueckert, "Fast Catheter Segmentation From Echocardiographic Sequences Based on Segmentation From Corresponding X-Ray Fluoroscopy for Cardiac Catheterization Interventions," *Ieee Transactions on Medical Imaging,* vol. 34, no. 4, pp. 861-876, Apr 2015, doi: 10.1109/tmi.2014.2360988.

[8] L. Yatziv, M. Chartouni, S. Datta, and G. Sapiro, "Toward Multiple Catheters Detection in Fluoroscopic Image Guided Interventions," *Ieee Transactions on Information Technology in Biomedicine,* vol. 16, no. 4, pp. 770-781, Jul 2012, doi: 10.1109/titb.2012.2189407.

[9] B. Teixeira, B. Tamersoy, V. Singh, and A. Kapoor, "Adaloss: Adaptive Loss Function for Landmark Localization," *Arxiv,* preprint Aug 02 2019, doi: arXiv:1908.01070.

[10] A. Howard *et al.*, "Searching for MobileNetV3," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, SOUTH KOREA, Oct 27-Nov 02 2019, in IEEE International Conference on Computer Vision, 2019, pp. 1314-1324, doi: 10.1109/iccv.2019.00140.

[11] J. Hu, L. Shen, G. Sun, and Ieee, "Squeeze-and-Excitation Networks," in *31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, Jun 18-23 2018, in IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132-7141, doi: 10.1109/cvpr.2018.00745.

[12] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun, and Ieee, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, Jun 27-30 2016, in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778, doi: 10.1109/cvpr.2016.90.

[13] L. C. E. Chen, Y. K. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *15th European Conference on Computer Vision (ECCV)*, Munich, GERMANY, Sep 08-14 2018, vol. 11211, in Lecture Notes in Computer Science, 2018, pp. 833-851, doi: 10.1007/978-3-030-01234-2_49.

[14] F. Milletari, N. Navab, S. A. Ahmadi, and Ieee, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *4th IEEE International Conference on 3D Vision (3DV)*, Stanford Univ, Stanford, CA, Oct 25-28 2016, in International Conference on 3D Vision, 2016, pp. 565-571, doi: 10.1109/3dv.2016.79.

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *Arxiv,* preprint Mar 09 2015, doi: arXiv:1503.02531.

[16] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *Ieee Transactions on Signal Processing,* vol. 50, no. 2, pp. 174-188, Feb 2002, doi: 10.1109/78.978374.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Munich, GERMANY, Oct 05-09 2015, vol. 9351, in Lecture Notes in Computer Science, 2015, pp. 234-241, doi: 10.1007/978-3-319-24574-4_28.

[18] O. Oktay *et al.*, "Attention U-Net: Learning Where to Look for the Pancreas," *Arxiv,* preprint May 20 2018, doi: arXiv:1804.03999.

[19] H. C. Li, P. F. Xiong, H. Q. Fan, J. Sun, and I. C. Soc, "DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation," in *32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, Jun 16-20 2019, in IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9514-9523, doi: 10.1109/cvpr.2019.00975.

[20] K. Sun, B. Xiao, D. Liu, J. D. Wang, and I. C. Soc, "Deep High-Resolution Representation Learning for Human Pose Estimation," in *32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, Jun 16-20 2019, in IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5686-5696, doi: 10.1109/cvpr.2019.00584.

[21] Y. N. Hou, Z. Ma, C. X. Liu, C. C. Loy, and Ieee, "Learning Lightweight Lane Detection CNNs by Self Attention Distillation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, SOUTH KOREA, Oct 27-Nov 02 2019, in IEEE International Conference on Computer Vision, 2019, pp. 1013-1021.