

End-to-End Real-time Catheter Segmentation with Optical Flow-Guided Warping during Endovascular Intervention

Nguyen, A., Kundrat, D., Dagnino, G., Chi, W., Abdelaziz, M. E. M. K., Guo, Y., Ma, Y., Kwok, T. M. Y., Riga, C. & Yang, G-Z.

Author post-print (accepted) deposited by Coventry University's Repository

Original citation & hyperlink:

Nguyen, A, Kundrat, D, Dagnino, G, Chi, W, Abdelaziz, MEMK, Guo, Y, Ma, Y, Kwok, TMY, Riga, C & Yang, G-Z 2020, End-to-End Real-time Catheter Segmentation with Optical Flow-Guided Warping during Endovascular Intervention. in 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 9967 - 9973, International Conference on Robotics and Automation, Paris, France, 31/05/20.

<https://dx.doi.org/10.1109/ICRA40945.2020.9197307>

DOI 10.1109/ICRA40945.2020.9197307

ISBN 978-1-7281-7395-5

ESSN 2577-087X

Publisher: IEEE

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Copyright © and Moral Rights are retained by the author(s) and/ or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This item cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder(s). The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

This document is the author's post-print version, incorporating any revisions agreed during the peer-review process. Some differences between the published version and this version may remain and you are advised to consult the published version if you wish to cite from it.

End-to-End Real-time Catheter Segmentation with Optical Flow-Guided Warping during Endovascular Intervention

Anh Nguyen¹, Dennis Kundrat¹, Giulio Dagnino¹, Wenqiang Chi¹, Mohamed E. M. K. Abdelaziz¹, Yao Guo¹, YingLiang Ma², Trevor M. Y. Kwok³, Celia Riga³, and Guang-Zhong Yang^{1,4}

Abstract—Accurate real-time catheter segmentation is an important pre-requisite for robot-assisted endovascular intervention. Most of the existing learning-based methods for catheter segmentation and tracking are only trained on small-scale datasets or synthetic data due to the difficulties of ground-truth annotation. Furthermore, the temporal continuity in intraoperative imaging sequences is not fully utilised. In this paper, we present FW-Net, an end-to-end and real-time deep learning framework for endovascular intervention. The proposed FW-Net has three modules: a segmentation network with encoder-decoder architecture, a flow network to extract optical flow information, and a novel flow-guided warping function to learn the frame-to-frame temporal continuity. We show that by effectively learning temporal continuity, the network can successfully segment and track the catheters in real-time sequences using only raw ground-truth for training. Detailed validation results confirm that our FW-Net outperforms state-of-the-art techniques while achieving real-time performance.

I. INTRODUCTION

In cardiovascular surgery, endovascular intervention offers many advantages compared to the traditional open surgical approaches, including smaller incisions, less trauma for patients, local instead of general anesthesia, stability, and more importantly, reduced risks for patients who have comorbidities [1]. Endovascular intervention involves the manipulation of catheters and guidewires to reach target areas in the vasculature to deliver a treatment (e.g. stenting, ablation or drug delivery [2]). Such tasks require a high level of technical skills to avoid damage to the vessel wall, which could result in perforation and hemorrhage, or dissection and organ failure, all of which can be fatal. Despite their relative advantages, endovascular procedures still present some limitations such as limited sensory feedback, misalignment of visuo-motor axes, and the need for high dexterity from the operators [3]. Robotics and computer assistance have been integrated into the clinical workflow to provide augmentation of surgical skills in terms of enhanced dexterity and precision [4]–[12].

As a pre-requisite to robot-assisted intervention, the task of catheter segmentation can provide essential visual or haptic feedback for the surgeons. For example, in [10], a vision-based force sensing is developed based on the tip

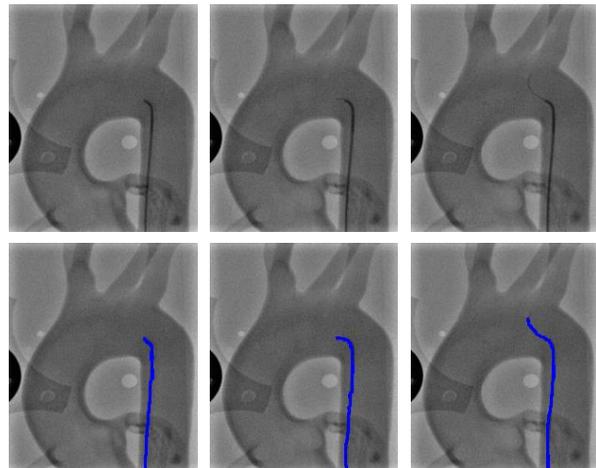


Fig. 1. Catheter segmentation in 2D X-ray fluoroscopy sequences. **Top row:** The X-ray images of the catheter advancing within an aortic phantom. **Bottom row:** An illustration of segmented results.

position of catheter and the vasculature. However, in routine practice, damages to the vessels are generated not only by the contact of the catheter tip with the vessel wall, but by the contacts between the entire catheter and the endothelial wall. Therefore, delineation and tracking of the entire catheter are essential. However, autonomous catheter segmentation is not a trivial task for two main reasons. Firstly, in the X-ray image, catheters can be easily confused with other similar linear structures like blood vessels due to its low contrast. Secondly, during clinical trials, catheters and guidewires can have a sudden and large deformation movement. This leads to the fact that traditional methods [13]–[16] based on primitive features of catheter appearance have limited generality and would not be able to segment catheters in real-time and dynamic surgical environments.

Recently, machine learning, especially deep learning has been widely adopted as a novel approach for medical image segmentation [17]–[19]. The effectiveness of deep learning comes from the ability to handle a large amount of multimodal input data [20]–[22]. However, this advantage becomes a potential problem in catheter segmentation since it is not easy to create large-scale datasets with pixel-wise labels. This is because the annotation task requires a certain amount of medical expertise, while manually labeling is very tedious, especially for objects with elongated structures such as catheters and guidewires. Due to these challenges, recent

¹Hamlyn Centre for Robotic Surgery, Imperial College London, UK
a.nguyen@imperial.ac.uk

²School of Computing, Electronics and Mathematics, Coventry University, UK

³Department of Surgery and Cancer, Imperial College London, St Mary's Hospital, UK

⁴Institute of Medical Robotics, Shanghai JiaoTong University, China

deep learning methods for catheter segmentation mainly train on a very small dataset [23] [24], use synthetic data [25], or create ground-truth based on a particular observation about pixel intensity [26]. These assumptions technically limit the power of deep learning and the generality of the methods.

In this paper, we propose Flow-guided Warping Net (FW-Net), a new end-to-end framework for catheter segmentation in 2D X-ray fluoroscopy sequences (Fig. 1). Our hypothesis is that deep network can be trained using the raw ground-truth while the overall accuracy can be improved by effectively learning the temporal continuity from X-ray sequences. In particular, we first create the raw ground-truth using a vision-based approach [27], then we design FW-Net with three modules: *i*) a segmentation network, *ii*) a flow network, and *iii*) a novel flow-guided warping function. We train FW-Net on raw ground-truth data and employ the flow-guided warping function to learn the temporal continuity between consecutive X-ray frames. This will encourage the network to predict based on both the raw ground-truth and sequential information, hence potentially improve the accuracy.

The rest of the paper is organized as follows. We review the related work in Section II, then describe the data collection process on our robotic platform in Section III. In Section IV, we present the new end-to-end architecture for effectively segmenting the catheter from raw ground-truth. The experimental results are presented in Section V. Finally, we conclude the paper and discuss the future work in Section VI.

II. RELATED WORK

Recently, there has been an increasing effort in segmenting catheters and guidewires from X-ray images. These methodologies can be divided into two main categories: vision-based approach and learning-based approach.

Traditional methods for catheter segmentation mainly used primitive image level cues such as pixel intensity, texture, or histogram [13]–[16], [28], [29]. In [30], the authors introduced a method based on Hough transform for detecting supporting device position in adult chest X-ray. Similarly, Kao et al. [31] proposed a system to detect endotracheal tubes on pediatric chest X-ray image using local features and multiple thresholds. Keller et al. [32] introduced a semi-automated method for catheter detection and tracking using prior information from users input. Mercan et al. [33] proposed to use local and global curvature features with controllable smoothness for guidewire segmentation. More recently, the authors in [27] used the multiscale vessel enhancement filter and adaptive binarization technique for detecting catheters and guidewires in real-time. A major drawback of all methods based on thresholding techniques is they do not generalize well and are very sensitive to a particular input X-ray data.

Machine learning techniques are also widely used for catheter segmentation and tracking [34]–[37]. With the rise of deep learning, methods based on Convolutional Neural Networks (CNN) are adapted for catheter segmentation [38] [39]. Early work in [40] used a simple neural

network to detect chest tubes then post-processed the results using a curve fitting technique to connect discontinued segments. In [17] [18], the state-of-the-art U-Net and V-Net architecture were introduced for data-driven medical image segmentation. Ambrosini et al. [41] presented an adaptive U-Net architecture for catheter segmentation in X-ray sequences. Vlontzos and Mikolajczyk [26] segmented the catheter from X-ray angiography video with a deep network and the ground-truth created by a carefully manual thresholding. Unberath et al. [42] presented a framework for simulating fluoroscopy and digital radiography from CT scans, then detecting anatomical landmarks with a deep network. The authors in [23] [24] used CNN with multihead for stent segmentation in X-ray fluoroscopy images. More recently, in [25] a scale-recurrent network was used to detect catheters in synthetic X-ray data.

While deep learning-based approaches can learn meaningful features from input data, applying deep learning to catheter segmentation problem is not straightforward due to the lack of real X-ray data, and the tediousness when manually labeling ground-truth. In this work, we propose to learn from raw ground-truth data and encode the temporal consistency between neighborhood X-ray frames. This will help the network rely more on the temporal information to segment the catheter in X-ray sequences.

III. DATA COLLECTION

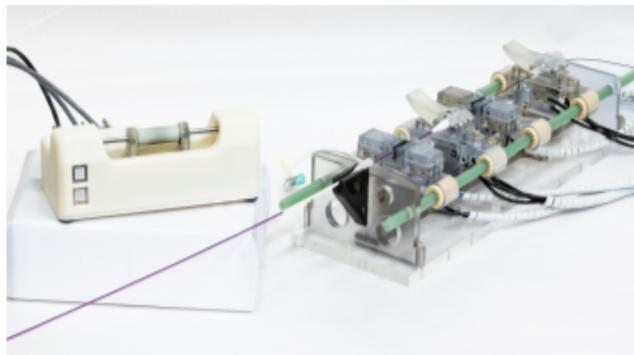


Fig. 2. CathBot robotic platform for fluoroscopy and MR-guided endovascular interventions. **Left:** Master device. **Right:** MR-safe slave robot.

A. CathBot

In this work, we collect sequences of X-ray data during the intervention using the CathBot [5] robot. CathBot (Fig. 2) comprises a versatile master-slave setup and navigation framework. Unlike previous platforms, the robot can be safely integrated and used in Magnetic Resonance (MR) environments thanks to pneumatic actuation and additive manufacturing. The master robot is an intuitive human machine interface (HMI) which mimics the human motion pattern (i.e. grasping the instrument followed by insertion/retraction and/or rotation) and provides haptic feedback to the users generated by the navigation systems as described in [3], [10]. Motions are mapped to the 4-DOF MR-safe slave

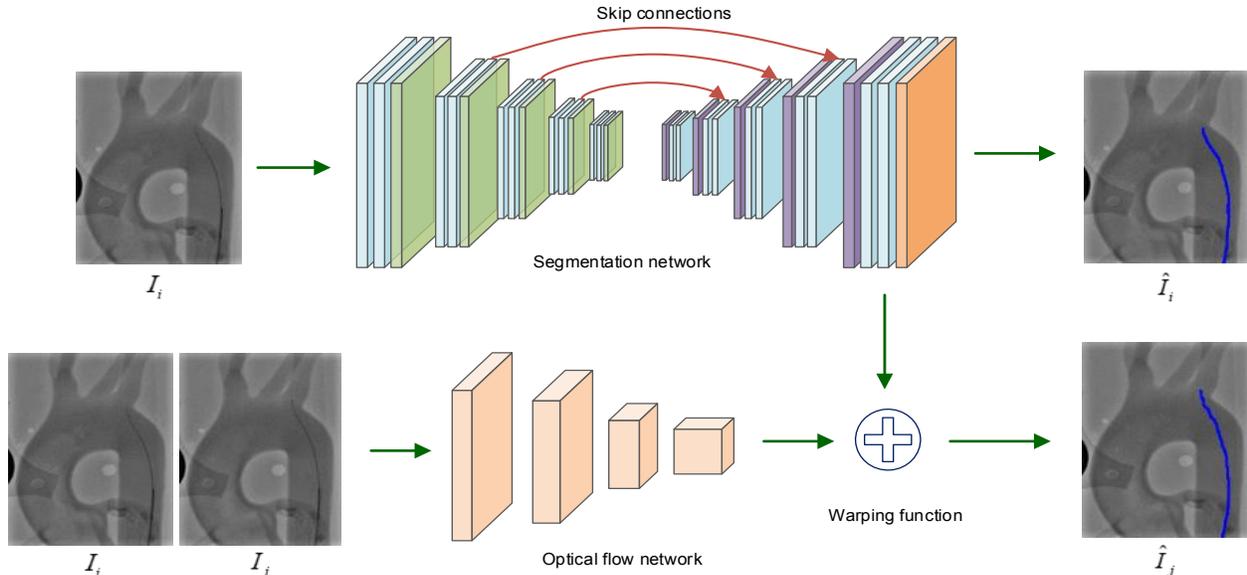


Fig. 3. An overview of our FW-Net architecture. The network consists of three modules: a segmentation network with encoder-decoder architecture and skip connections, a flow network to extract optical flow information from two neighborhood frames, and a flow-guided warping function to learn the frame-to-frame temporal continuity.

robot, capable of manipulating off-the-shelf catheters and guidewires.

B. X-ray Data Collection

A vascular soft silicone phantom (Elastrat, Geneva, Switzerland) of a normal adult human aortic arch was placed underneath an X-ray imaging system to simulate a patient lying on the angiography table to undergo an endovascular procedure. The phantom was connected to a pulsatile pump to simulate normal human blood flow and optimize the level of realism for tool-tissue interactions. A professional surgeon was asked to cannulate three arteries by manipulation of wire and catheter. Namely, the left subclavian (LSA), left common carotid (LCCA) and right common carotid (RCCA) arteries. The cannulation was performed in two scenarios: manual and robot assisted. During each maneuver, fluoroscopy was activated by the operator using a pedal. Real-time video stream of the surgical scene was acquired using an image grabber (DVI2USB3, Epiphan Video, Ottawa, Canada) from a vascular imaging system - in this study we have used a fluoroscopic system for interventional radiology procedure (Innova 4100 IQ GE Healthcare). The video stream was acquired on a workstation (Windows 7, Intel i7-6700, 3.4GHz, 16GB RAM) and digitalized into image sequence for image processing.

IV. METHODOLOGY

Our goal is to segment the catheters and guidewires in X-ray fluoroscopy sequences using the raw ground-truth created by [27]. Since the selected ground-truth annotation method does not take into account the temporal continuity, which is the key information from the X-ray sequences, we construct a unified framework to effectively learn this information.

Towards this end, we propose FW-Net, a new end-to-end architecture to effectively segment the catheter in X-ray sequences using a novel flow-guided warping function. The overall architecture of our proposed approach is illustrated in Fig. 3.

A. Segmentation Network

Our specific segmentation task is to compute a binary mask separating the foreground (i.e., catheter and guidewire) from the background for every X-ray frame of the video. Inspired by the effectiveness of deep neural networks in image segmentation, we build our segmentation branch based on encoder-decoder architecture [17] [43]. To improve the real-time performance of the network, we use big convolution kernels with large strides to extract features from the input X-ray frame. Since the convolution operation is comparably cheap with a small number of channels as in X-ray images, using big kernels does not significantly increase the computational costs. Furthermore, we combine large strides with skip connections as in U-Net architecture [17] to maintain low-level features during the decoding process.

Specifically, the input of the segmentation network is the RGB X-ray image of size (256×256) pixels. The encoder network has 5 ResNet blocks [44] to extract the depth features from input images. Each ResNet block consists of a convolutional layer, ReLU, skip links and pooling operations. The output map after each ResNet block in the encoder network has the size of 128, 64, 32, 16, and 8 respectively. Each decoder block is associated with an encoder. In each decoder block, the encoder feature map is upsampled using the deconvolutional operation. Finally, a 2 classes soft-max layer is used at the end of the decoder network to classify

the background and foreground for all pixels in the current X-ray frame.

Unlike the traditional image segmentation problem, in catheter segmentation, the imbalance between the foreground and background regions is strongly significant since the foreground only occupies a small portion number of pixels. To overcome this problem, we employ the weighted version of pixel-wise cross-entropy loss function \mathbb{E} as in [45]. The segmentation loss is defined as follows:

$$\mathcal{L}_s = -(1-w) \sum_{i,j \in \text{fg}} \log \mathbb{E}(y_{ij} = 1; \theta) - w \sum_{i,j \in \text{bg}} \log \mathbb{E}(y_{ij} = 0; \theta) \quad (1)$$

where i and j are the pixel location of the foreground fg and the background bg, respectively; y_{ij} denotes the binary prediction of each pixel of the input image, w is the foreground-background pixel-number ratio, and θ represents the network parameters.

B. Optical Flow Network

Extracting optical flow is a fundamental task in video analysis. Traditional methodologies for this problem have been studied for decades and mainly used variational approaches which address small displacements [46]. Recently, deep learning has been exploited for learning optical flow. In this work, we adopt the simple version of FlowNet [47], a state-of-the-art deep neural-based architecture as our flow network. To decrease the computational complexity, we reduce the number of convolutional kernels in each layer of FlowNet by half and hence reduce the overall complexity to one fourth.

In practice, we stack two neighborhood X-ray images (I_i, I_j) together and feed them through a deep network to extract the flow motion. Note that, the I_i frame is also the input frame for segmentation network. Since the computed optical flow is aligned with the segmentation output, their shared feature map information can be combined later naturally to generate the segmented map for I_j . Specifically, our flow network has a sequence of 6 convolutional layers to estimate the flow motion from consecutive video frames. All convolutional layers have the stride of 2. Compared to the segmentation network, the flow network is simpler with fewer parameters.

C. Flow-Guided Warping Function

Unlike the traditional image segmentation problem, where the temporal information is not available, in video segmentation, temporal consistency across frames is the key to success. Our observation is that the consecutive X-ray frames are highly similar. This similarity is even stronger in the deep feature maps since they encode high level semantic concepts from these frames [48]. We exploit the similarity by warping the deep features from segmentation network with the flow motion from flow network.

As motivated by Zhu et al. [12], given a reference frame I_i and a neighbor frame I_j , a flow motion field $M_{i \rightarrow j} = \mathcal{F}(I_i, I_j)$ is estimated by a flow network \mathcal{F} (e.g., FlowNet).

The feature maps on the reference frame are warped to the neighbor frame according to the optical flow. The warping function is defined as:

$$f_{i \rightarrow j} = \mathcal{W}(f_i, M_{i \rightarrow j}) = \mathcal{W}(f_i, \mathcal{F}(I_i, I_j)) \quad (2)$$

where $f_{i \rightarrow j}$ denotes the feature maps warped from previous frame I_i to frame I_j . $\mathcal{W}(\cdot)$ is the bilinear warping function applied on all the locations for each channel in the feature maps. $\mathcal{F}(I_i, I_j)$ is the flow field estimated by the flow network, which maps a location $\mathbf{p} = (p_x, p_y)$ in frame I_i to the location $\mathbf{p} + \delta\mathbf{p}$ in frame I_j .

Since the feature maps has several channels, the warping is performed in each channel as:

$$\varphi_{i \rightarrow j}(\mathbf{p}) = \sum_{\mathbf{q}} \mathbf{K}(\mathbf{q}, \mathbf{p} + \delta\mathbf{p}) \varphi_j(\mathbf{q}) \quad (3)$$

where \mathbf{q} denotes all spatial locations in the feature maps, and \mathbf{K} indicates the bilinear interpolation kernel.

Since we employ end-to-end training, the backpropagation of $\varphi_{i \rightarrow j}$ with respect to φ_i and flow $\delta\mathbf{p}$ is derived as:

$$\begin{aligned} \frac{\partial \varphi_{i \rightarrow j}(\mathbf{p})}{\partial \varphi_i(\mathbf{q})} &= \mathbf{K}(\mathbf{q}, \mathbf{p} + \delta\mathbf{p}) \\ \frac{\partial \varphi_{i \rightarrow j}(\mathbf{p})}{\partial \mathcal{F}(I_i, I_j)(\mathbf{p})} &= \sum_{\mathbf{q}} \frac{\partial \mathbf{K}(\mathbf{q}, \mathbf{p} + \delta\mathbf{p})}{\partial \delta\mathbf{p}} \varphi_i(\mathbf{q}) \end{aligned} \quad (4)$$

Intuitively, the warping function $\mathcal{W}(\cdot)$ combines the features of the segmentation network with the output of the flow network in the same region of the reference frame I_i , then generate the segmentation for that region in the neighbor frame I_j . This warping process provides more diverse information on the same image region, such as deformation and varied illuminations while effectively use the temporal information from the flow. We also note that the flow network cannot generate the semantic segmentation by itself since it only predicts the displacement by optical flow. Therefore, we need to combine the flow network with the segmentation network using the warping function $\mathcal{W}(\cdot)$ to generate the segmentation map for the neighbor frame.

Training The network is end-to-end trained using stochastic gradient descent (SGD) with a fixed 0.001 learning rate and 0.9 momentum. In each mini-batch, a pair of nearby video frames (I_i, I_j) with $0 \leq j - i \leq 6$, are randomly sampled. The total loss is the combination of two cross-entropy losses as follows:

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_w \quad (5)$$

where \mathcal{L}_s is the loss of segmentation network to generate segmented map for I_i , and \mathcal{L}_w is the loss for generating segmented map for I_j . λ is the hyperparameter and is empirically set to 0.4.

In practice, we implement our method using the Tensorflow library [49]. The network is trained from scratch until convergence with no further reduction in training loss. The training time is approximately 2 days on an NVIDIA GTX 2080 GPU on a dataset with more than 20,000 X-ray frames.

TABLE I
DICE SCORES OVER THE TESTING SET

	Training?	Temporal?	FPS	<i>Dice</i>
TCF [27]	No	No	10 (CPU)	0.796
U-Net [17]	Yes	No	2 (GPU)	0.677
Adaptive U-Net [41]	Yes	Yes	8 (GPU)	0.745
Siamese U-Net [26]	Yes	Yes	90 (GPU)	0.768
FW-Net (ours)	Yes	Yes	15 (GPU)	0.821

V. EXPERIMENTS

A. Experimental Setup

Dataset We perform 28 clinical trials using the CathBot robot, resulting in 28 X-ray videos. Each video describes the movement of the catheter and guidewire in each trial and is approximately 2 – 5 minutes long. We extract X-ray frames from each video at 8 frames per second. In total, our new X-ray dataset has 25,271 frames from 28 sequences. We resize all the frames to (256 × 256) pixels before using them in our network. The raw pixel-wise ground-truth of the frames is created using the method in [27]. Due to the nature of the technique in [27], both the catheter and guidewire are considered as one class in our experiment. For quantitative evaluation, we manually label 2 sequences with approximately 1,000 frames for testing, and use all frames from the other videos for training. We notice all labels for training are created automatically by [27], and no further manual human correction is needed.

Metric As the standard practice in binary segmentation, we use the *Dice* metric to evaluate the segmentation results. The *Dice* index is defined between the ground-truth mask X and the predicted mask Y as follows:

$$Dice = \frac{2||X \cap Y||}{||X|| + ||Y||} = \frac{2TP}{2TP + FP + FN} \quad (6)$$

where TP denotes the true positive number of labeled pixels, FP indicates the false positive pixels, and FN is the false negative pixels.

Baseline We compare our results (FW-Net) with the following state-of-the-art methods: U-Net [17], Siamese U-Net [26], Adaptive U-Net [41], and TCF [27]. We note that TCF [27] does not require training since it used vision-based technique, while all other methods are trained on the same training set with raw ground-truth. Within the deep learning methods, the U-Net architecture does not take into account the sequential information, while our network, Siamese U-Net [26], and Adaptive U-Net [41] exploit the use of temporal information.

B. Results

Table I summarizes segmentation results of our method and the baselines on the testing set. The results clearly show that our FW-Net consistently improves over the state-of-the-art. In particular, FW-Net achieves the *Dice* score of 0.821, which is a concrete improvement over the second-best method. It is worth noting that FW-Net outperforms the original approach U-Net approach by a large margin of

14.3%. This result is explainable since the U-Net architecture is designed for an individual frame and does not take into account the temporal information, while our FW-Net is designed to learn the frame-to-frame temporal continuity effectively from X-ray sequences.

We also observe a significant improvement of our FW-Net over Adaptive U-Net and Siamese U-Net, which are the deep learning-based methods exploit the temporal information. It shows that our proposed flow-guided warping method can encode the temporal information more successfully than Adaptive U-Net (which only trains the video frame sequentially) or Siamese U-Net (which relies heavily on data augmentation). We also found that all networks exploit temporal information achieve better results than the original U-Net. However, since all the network are trained using the raw ground-truth, other deep networks except our FW-Net cannot outperform the classical TCF method.

Table I also provides the intuitive inference time of all methods in frame per second (FPS). Overall, our FW-Net achieves a speed of 15 FPS on NVIDIA GTX 2080 GPU, which is reasonable for real-time applications. Within deep learning-based methods, Siamese U-Net has the fastest inference time at 90 FPS. However, here we notice that all deep learning methods need to use GPU for real-time performance, while the TCF [27] method can achieve 10 FPS on a core i7 CPU. A visualization of the segmented results of all methods can be found in Fig. 4. More qualitative results can be found in our supplemental video.

To conclude, our FW-Net can effectively learn the temporal continuity and significantly improves over the state of the art. Our method is also end-to-end and does not require data augmentation or any extra post-processing. The inference time of FW-Net is 15 FPS on a GPU which allows it to be used in wide range clinical applications. More details about our project can be found at <https://sites.google.com/site/cathetersegmentation/>.

VI. CONCLUSIONS AND FUTURE WORK

We propose FW-Net, an end-to-end and real-time deep learning framework for catheter and guidewire segmentation in 2D X-ray fluoroscopy sequences. Our FW-Net consists of three components to effectively learn the temporal information: a segmentation network, a flow network, and a novel flow-guided warping function. We showed that by learning the temporal continuity, the segmentation result can be improved even when training with the raw ground-truth data. The experimental results demonstrate that our FW-Net not only achieves state-of-the-art results, but also has real-time performance. Hence, the proposed approach can be integrated to robotic control frameworks or considered for generation of haptic feedback with deployment to various endovascular applications.

Since we use a vision-based method to automatically generate ground-truth with only binary segmentation mask, our FW-Net is currently tested with the binary segmentation problem. In the future, we would like to explore the ability of FW-Net in multiclass segmentation problem with X-ray

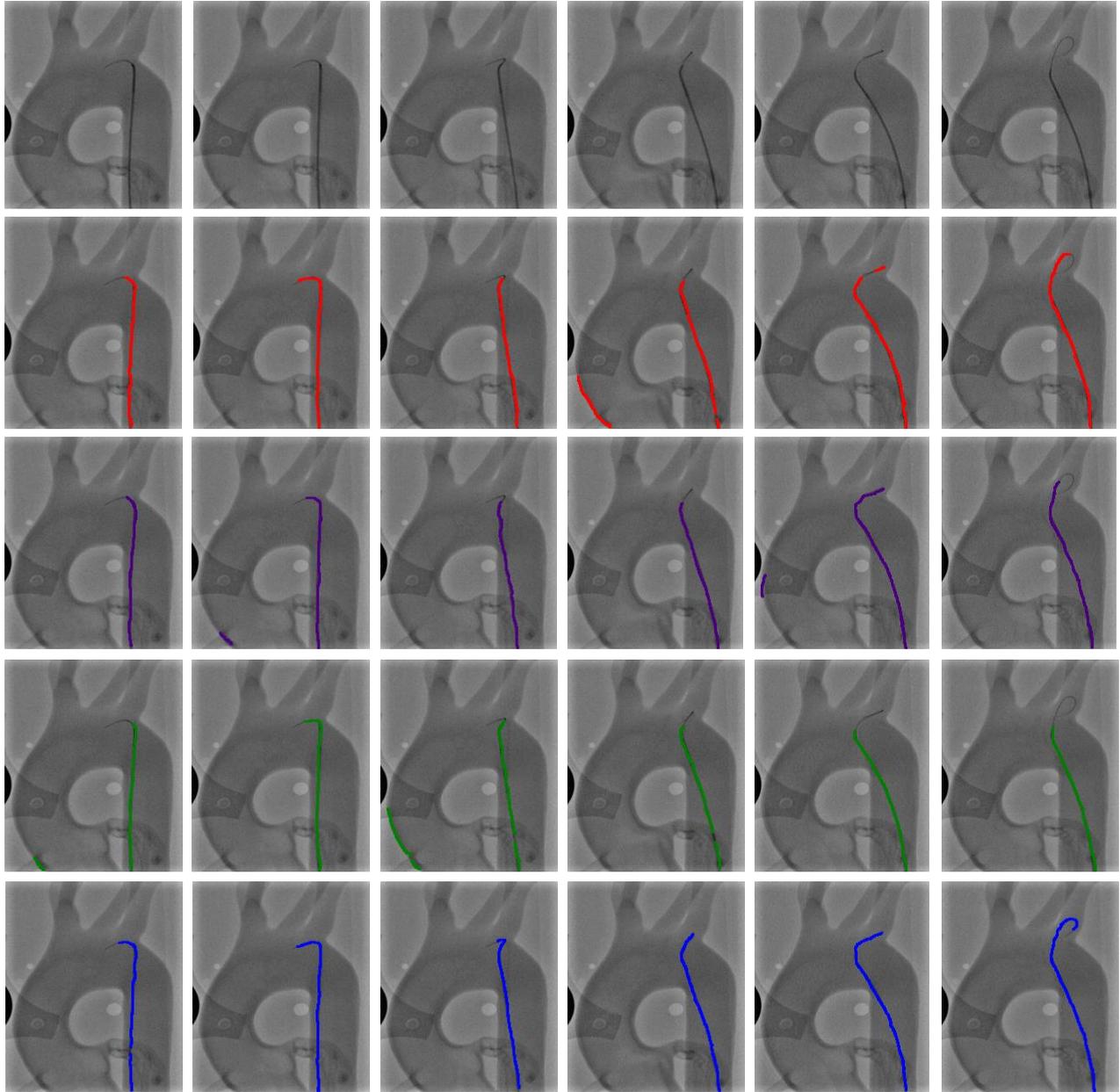


Fig. 4. A visualization of the segmentation results of an X-ray image sequence. Top row - original X-ray images; Second row - TCF [27]; Third row: Adaptive U-Net [41]; Fourth row: Siamese U-Net [26]; Fifth row: Our FW-Net. Compared to other methods, our FW-Net shows less over-segmented regions as well as less disconnected segments.

images, where we can have more classes such as catheter, guidewire, blood vessel. This will allow FW-Net to become more useful in clinical scenarios. This further motivates application to closed-loop control with robotic platforms [5] that facilitate individual manipulation of catheters and guidewires. The proposed methodology will be prospectively fused with advanced user assistance to incorporate the entire interaction of endovascular instruments and vascular structures for adaptive generation of haptic feedback [10]. Finally, the contribution bears great potential for integration into a novel skill assessment framework with image-based metrics in endovascular surgery.

ACKNOWLEDGMENT

We would like to thank A. Vlontzos for the useful discussion. This research is supported by the UK Engineering and Physical Science Research Council (EP/N024877/1) and the Wellcome Trust.

REFERENCES

- [1] N. Simaan, R. M. Yasin, and L. Wang, "Medical technologies and challenges of robot-assisted minimally invasive intervention and diagnostics," *Annual Review of Control, Robotics, and Autonomous Systems*, 2018.
- [2] H. Rafii-Tari, C. J. Payne, and G.-Z. Yang, "Current and emerging robot-assisted endovascular catheterization technologies: a review," *Annals of Biomedical Engineering*, 2014.

- [3] M. Benavente Molinero, G. Dagnino, J. Liu, W. Chi, M. Abdelaziz, T. Kwok, C. Riga, and G. Yang, "Haptic Guidance for Robot-Assisted Endovascular Procedures: Implementation and Evaluation on Surgical Simulator," in *IROS*, 2019.
- [4] Y. Thakur, J. S. Bax, D. W. Holdsworth, and M. Drangova, "Design and performance evaluation of a remote catheter navigation system," *IEEE Transactions on Biomedical Engineering*, 2009.
- [5] M. E. M. K. Abdelaziz, D. Kundrat, M. Pupillo, G. Dagnino, T. MY, W. C. Kwok, V. Groenhuis, F. J. Siepel, C. Riga, S. Stramigioli, *et al.*, "Toward a versatile robotic platform for fluoroscopy and mri-guided endovascular interventions: A pre-clinical study," in *IROS*, 2019.
- [6] G.-B. Bian, X.-L. Xie, Z.-Q. Feng, Z.-G. Hou, P. Wei, L. Cheng, and M. Tan, "An enhanced dual-finger robotic hand for catheter manipulating in vascular intervention: A preliminary study," in *ICIA*, 2013.
- [7] W. Chi, G. Dagnino, T. Kwok, A. Nguyen, D. Kundrat, E. M. K. Abdelaziz, Mohamed, C. Riga, C. Bicknell, and G.-Z. Yang, "Collaborative robot-assisted endovascular catheterization with generative adversarial imitation learning," in *ICRA*, 2020.
- [8] Y. Zhao, S. Guo, Y. Wang, J. Cui, Y. Ma, Y. Zeng, X. Liu, Y. Jiang, Y. Li, L. Shi, *et al.*, "A cnn-based prototype method of unstructured surgical state perception and navigation for an endovascular surgery robot," *Medical & Biological Engineering & Computing*, 2019.
- [9] R. J. Varghese, A. Nguyen, E. Burdet, G.-Z. Yang, and B. P. Lo, "Nonlinearity compensation in a multi-dof shoulder sensing exosuit for real-time teleoperation," *arXiv preprint arXiv:2002.09195*, 2020.
- [10] G. Dagnino, J. Liu, M. E. Abdelaziz, W. Chi, C. Riga, and G.-Z. Yang, "Haptic feedback and dynamic active constraints for robot-assisted endovascular catheterization," in *IROS*, 2018.
- [11] W. Chi, J. Liu, H. Raffi-Tari, C. Riga, C. Bicknell, and G.-Z. Yang, "Learning-based endovascular navigation through the use of non-rigid registration for collaborative robotic catheterization," *International Journal of Computer Assisted Radiology and Surgery*, 2018.
- [12] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *CVPR*, 2017.
- [13] A. Brost, R. Liao, J. Hornegger, and N. Strobel, "3-d respiratory motion compensation during ep procedures by image-based 3-d lasso catheter model generation and tracking," in *MICAI*, 2009.
- [14] A. Brost, R. Liao, N. Strobel, and J. Hornegger, "Respiratory motion compensation by model-based catheter tracking during ep procedures," *Medical Image Analysis*, 2010.
- [15] 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*, 2012.
- [16] S. A. Baert, M. A. Viergever, and W. J. Niessen, "Guide-wire tracking during endovascular interventions," *TMI*, 2003.
- [17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICAI*, 2015.
- [18] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision (3DV)*, 2016.
- [19] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, *et al.*, "Deepgeos: a deep interactive geodesic framework for medical image segmentation," *TPAMI*, 2018.
- [20] A. Nguyen, Q. D. Tran, T.-T. Do, I. Reid, D. G. Caldwell, and N. G. Tsagarakis, "Object captioning and retrieval with natural language," in *ICCVW*, 2019.
- [21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.
- [22] A. Nguyen, T.-T. Do, I. Reid, D. G. Caldwell, and N. G. Tsagarakis, "V2cnet: A deep learning framework to translate videos to commands for robotic manipulation," *arXiv:1903.10869*, 2019.
- [23] K. Breininger, T. Würfl, T. Kurzendorfer, S. Albarqouni, M. Pfister, M. Kowarschik, N. Navab, and A. Maier, "Multiple device segmentation for fluoroscopic imaging using multi-task learning," in *Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis*, 2018.
- [24] K. Breininger, S. Albarqouni, T. Kurzendorfer, M. Pfister, M. Kowarschik, and A. Maier, "Intraoperative stent segmentation in x-ray fluoroscopy for endovascular aortic repair," *International Journal of Computer Assisted Radiology and Surgery*, 2018.
- [25] X. Yi, S. Adams, P. Babyn, and A. Elnajmi, "Automatic catheter and tube detection in pediatric x-ray images using a scale-recurrent network and synthetic data," *Journal of Digital Imaging*, 2019.
- [26] A. Vlontzos and K. Mikolajczyk, "Deep segmentation and registration in x-ray angiography video," *arXiv:1805.06406*, 2018.
- [27] Y. Ma, M. Alhrishy, S. A. Narayan, P. Mountney, and K. S. Rhode, "A novel real-time computational framework for detecting catheters and rigid guidewires in cardiac catheterization procedures," *Medical Physics*, 2018.
- [28] W. Wu, T. Chen, P. Wang, S. K. Zhou, D. Comaniciu, A. Barbu, and N. Strobel, "Learning-based hypothesis fusion for robust catheter tracking in 2d x-ray fluoroscopy," in *CVPR*, 2011.
- [29] Y. Ma, N. Gogin, P. Cathier, R. J. Housden, G. Gijbsbers, M. Cooklin, M. O'Neill, J. Gill, C. A. Rinaldi, R. Razavi, *et al.*, "Real-time x-ray fluoroscopy-based catheter detection and tracking for cardiac electrophysiology interventions," *Medical Physics*, 2013.
- [30] C. Sheng, L. Li, and W. Pei, "Automatic detection of supporting device positioning in intensive care unit radiography," *International Journal of Medical Robotics and Computer Assisted Surgery*, 2009.
- [31] E.-F. Kao, T.-S. Jaw, C.-W. Li, M.-C. Chou, and G.-C. Liu, "Automated detection of endotracheal tubes in paediatric chest radiographs," *Computer Methods and Programs in Biomedicine*, 2015.
- [32] B. M. Keller, A. P. Reeves, M. D. Cham, C. I. Henschke, and D. F. Yankelevitz, "Semi-automated location identification of catheters in digital chest radiographs," in *Medical Imaging: Computer-Aided Diagnosis*, 2007.
- [33] V. Bismuth, R. Vaillant, H. Talbot, and L. Najman, "Curvilinear structure enhancement with the polygonal path image-application to guide-wire segmentation in x-ray fluoroscopy," in *MICAI*, 2012.
- [34] P. Wang, T. Chen, Y. Zhu, W. Zhang, S. K. Zhou, and D. Comaniciu, "Robust guidewire tracking in fluoroscopy," in *CVPR*, 2009.
- [35] B.-J. Chen, Z. Wu, S. Sun, D. Zhang, and T. Chen, "Guidewire tracking using a novel sequential segment optimization method in interventional x-ray videos," in *International Symposium on Biomedical Imaging (ISBI)*, 2016.
- [36] L. Wang, X.-L. Xie, G.-B. Bian, Z.-G. Hou, X.-R. Cheng, and P. Prasong, "Guide-wire detection using region proposal network for x-ray image-guided navigation," in *IJCNN*, 2017.
- [37] O. Pauly, H. Heibel, and N. Navab, "A machine learning approach for deformable guide-wire tracking in fluoroscopic sequences," in *MICAI*, 2010.
- [38] H. Yang, C. Shan, A. F. Kolen, and H. de With Peter, "Improving catheter segmentation & localization in 3d cardiac ultrasound using direction-fused fcn," in *ISBI*, 2019.
- [39] P. Zaffino, G. Pernelle, A. Mastmeyer, A. Mehrtash, H. Zhang, R. Kikinis, T. Kapur, and M. F. Spadea, "Fully automatic catheter segmentation in mri with 3d convolutional neural networks: application to mri-guided gynecologic brachytherapy," *Physics in Medicine & Biology*, 2019.
- [40] C. A. Mercan and M. S. Celebi, "An approach for chest tube detection in chest radiographs," *IET Image Processing*, 2013.
- [41] 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 *MICAI*, 2017.
- [42] M. Unberath, J.-N. Zaech, S. C. Lee, B. Bier, J. Fotouhi, M. Armand, and N. Navab, "Deepdr—a catalyst for machine learning in fluoroscopy-guided procedures," in *MICAI*, 2018.
- [43] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Detecting object affordances with convolutional neural networks," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2016.
- [44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [45] S. Xie and Z. Tu, "Holistically-nested edge detection," in *ICCV*, 2015.
- [46] J. Weickert, A. Bruhn, T. Brox, and N. Papenberg, "A survey on variational optic flow methods for small displacements," in *Mathematical Models for Registration and Applications to Medical Imaging*, 2006.
- [47] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *ICCV*, 2015.
- [48] D. Jayaraman and K. Grauman, "Slow and steady feature analysis: higher order temporal coherence in video," in *CVPR*, 2016.
- [49] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *Symposium on Operating Systems Design and Implementation*, 2016.