IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 19, 2019, accepted October 24, 2019, date of publication October 29, 2019,
date of current version November 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2950263

Real-Time Tracking of Guidewire Robot Tips Using Deep Convolutional Neural Networks on Successive Localized Frames

IHSAN ULLAH, PHILIP CHIKONTWE, AND SANG HYUN PARK, (Member, IEEE)
Department of Robotics Engineering, Daegu Gyeonbuk Institute of Science and Technology, Daegu 42988, South Korea

Corresponding author: Sang Hyun Park (shpark13135@dgist.ac.kr)

This work was supported in part by the Robot Industry Fusion Core Technology Development Project through the Korea Evaluation Institute of Industrial Technology (KEIT) funded by the Ministry of Trade, Industry, and Energy of Korea (MOTIE) under Grant 10052980, and in part by the Daegu Gyeongbuk Institute of Science and Technology (DGIST) Research and Development Program of the Ministry of Science, and the Information and Communication Technology (ICT) under Grant 19-RT-01.

ABSTRACT Studies are proceeded to stabilize cardiac surgery using thin micro-guidewires and catheter robots. To control the robot to a desired position and pose, it is necessary to accurately track the robot tip in real time but tracking and accurately delineating the thin and small tip is challenging. To address this problem, a novel image analysis-based tracking method using deep convolutional neural networks (CNN) has been proposed in this paper. The proposed tracker consists of two parts; (1) a detection network for rough detection of the tip position and (2) a segmentation network for accurate tip delineation near the tip position. To learn a robust real-time tracker, we extract small image patches, including the tip in successive frames and then learn the informative spatial and motion features for the segmentation network. During inference, the tip bounding box is first estimated in the initial frame via the detection network, thereafter tip delineation is consecutively performed through the segmentation network in the following frames. The proposed method enables accurate delineation of the tip in real time and automatically restarts tracking via the detection network when tracking fails in challenging frames. Experimental results show that the proposed method achieves better tracking accuracy than existing methods, with a considerable real-time speed of 19ms.

INDEX TERMS Convolutional neural networks, micro-robot tracking, guidewire tracking, patch-wise segmentation.

I. INTRODUCTION

Cardiac catheterization is a procedure used to diagnose and treat cardiovascular conditions. During cardiac catheterization, physicians insert a guidewire into an artery or vein and then transport stent via the guidewire under fluoroscopic guidance. However, placing the guidewire is complex and requires high expertise to control and navigate as the blood vessels to which the guidewire should be inserted are not visible without a contrasting agent. Moreover, the narrowed or blocked blood vessels are not visible even when the contrast agent is used. Consequently, conventional cardiac catheterization requires long treatment time, high concentration, and many contrast medications. Thus, there is a demand to develop localization technology [1] for an autonomous guidewire that can alleviate the potential for injury and radiation exposure or discomfort to physicians and patients.

Recently, small robotic guidewires and catheters have been developed for precise localization of target areas during interventions. Most methods [2]–[5] perform tracking of the guidewire by employing sensing systems with manual operation. For example, the catheter system [2], such as Amigo, was designed and evaluated so that users could place the robot on desired locations, but requires manual operation of pacing thresholds and endocardial electrograms. To address this, Borgstadt *et al.* [3] proposed closed-loop control for use with multiple sensors i.e., electromagnetic pose sensors (EPS) and stereo imaging in place of fluoroscopy for real-time catheter localization consisting two particle filters. The first filter (PF1) uses EPS for the measurement update, while the second (PF2) uses an imaging system with the outputs of the filters combined at each time step to produce the overall state estimate. However, guidewire tip

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammed Nabil El Korso.

localization with sensors alone is difficult and highly dependent on the sensing performance. Recently, Zhao *et al.* [6] proposed a guidewire navigation prototype using CNNs which estimate appropriate action under a closed-loop control. However, the system still requires accurate guidewire extraction. Consequently, to address these challenges, image analysis based methods using probabilistic frameworks [4], marginal space learning [5] and spline models [7] have been proposed for tracking the guidewire. However, these methods still need manual or semi-automatic initialization and the localization is often inaccurate even with multiple sensors.

In this paper, a novel method is proposed to reliably track the robot tip through image analysis in video sequences captured by an optical zoom camera. Ultimately, the position of the micro robot guidewire tip should be tracked in X-ray images, but guidewire tip tracking in video is also important especially in the system development stage to assess whether the control system works properly [8], [9]. Moreover, given the difficulty of X-ray data acquisition for developing catheter tracking algorithms, we first perform tracking studies in camera images that are easy to acquire with the intention of extending the approach to X-ray images. Although robust tracking in video sequences can be useful, tracking guidewire robot tips in real time remains inherently challenging as the tip is often very small, thin and appears with ambiguous noise in the background. Also, non-rigid abrupt motion of the guidewire makes tracking more difficult. To address these challenges, a new tracker which consists of two deep learning networks is proposed; a detection network for the localization of the robot tip using a bounding box and a segmentation network to segment the tip near the localized regions. The segmentation network is trained with localized guidewire tip patches extracted from consecutive frames, which makes the tip segmentation robust by effectively learning spatial and motion features near target object [10], [11].

The proposed method enables accurate delineation of the small tip in real time due to the low computational cost of patch-wise predictions, and addresses tracking failures in frames with abrupt motion by restarting the localization via the detection network, alleviating the need for manual correction. The key contributions are as follows:

1) To the best of our knowledge, this work is the first to apply a deep learning-based tracking method to track the guidewire tip in real cardiac robotic systems. Our proposed method does not require multiple sensors, heuristic manual tuning or post-processing steps [12].
2) To effectively track a very small object, a patch-wise segmentation strategy is proposed to consider both motion and spatial features across adjacent previous frames. Unlike conventional bounding box-based tracking methods [13]–[18], our method conducts precise pixel-wise predictions in real time by constraining the area in which the segmentation needs to be performed.
3) Our method effectively reduces tracking failures by adaptively using the detection network with respect to the output of segmentation network.

A preliminary version of this work has been presented at a conference [19]. Herein, (i) the proposed method is significantly improved by using patch-wise U-net which can consider previous adjacent frame information for efficient prediction, (ii) extensive quantitative and qualitative results are reported to confirm the effectiveness of the proposed method over the preliminary version, (iii) ablation studies are carried out to assess the effect of tracking performance as the number of previous frames is varied, (iv) the tracking time and the amount of tracking failures are significantly reduced by utilizing previous frame information.

The remainder of the paper is organized as follows. First, we revisit previous works in the field of intervention tool tracking in Sec. II. We then present our method in Sec. III and provide a thorough evaluation in Sec. IV. Finally, we present conclusions in Sec. VI.

II. RELATED WORK

We address recent advances related to catheter tracking in two facets, namely hand engineered feature-based methods and deep learning-based approaches that make use of deep convolutional neural networks (CNN).

A. CONVENTIONAL FEATURE-BASED METHODS

Several catheter or guidewire tracking methods have been proposed for image-guided navigation, although there are not robotic systems. For example, Franken *et al.* [20] extracted local image features using the Hessian matrix and enhanced elongated structures to localize the catheter. However, the implementation was too slow for clinical use, as computational cost was relatively high, implementation on a graphical processing unit (GPU) was suggested. Ma *et al.* [21] used a blob extraction method to detect all possible catheter electrode candidates, later choosing the best with a certain criteria. Ma *et al.* [22] enhanced the visibility of wire-like structures using multiscale vesselness enhancement filters, and then used the k-nearest neighbor algorithm to distinguish the target wires from other wire-like artifacts. Palti-Wasserman *et al.* [23] used a modified Laplacian filter to enhance the guidewire and tracked the guidewire via Hough transform. De Buck *et al.* [24] proposed a method for catheter tip detection using a fixed template-based registration in conjunction with a Kalman filter. Fallavollita *et al.* [25] proposed a method to measure the location of catheter tip electrodes. A convex hull algorithm was employed to reconstruct a 3D model of the left ventricle using aligned reference mappings between catheters and estimated tip centroids of electrodes. However, since the guidewire has unique characteristics such as non-rigidity, thinness and complicated motion, conventional tracking methods based on filtering or thresholding were inefficient to achieve robust guidewire tracking
results.

Generally, active contour-based methods achieved superior performance by utilizing internal curve models compared to the filter based methods [20], [21], [23]. For example, Slabaugh *et al.* [26] modeled the geometry of the guidewire as a spline with a length constraint term in the B-spline energy equation to retain a prior length. Later, Hoffmann *et al.* [12] improved catheter extraction by utilizing enhancement filters, followed by a spline fitting methodology, starting from the guidewire shape in the previous frame. However, these methods require manual annotation of the first frame in the image sequence and often incur a drift problem when the result in the previous frame is bad. To address the drifting problem, Chang *et al.* [7] proposed to optimize spline fitting with respect to control points with an equidistance prior to better fit complex curves. However, the method is sensitive to control point detection, requiring re-initialization for incorrect detection. In contrast, Mountney *et al.* [27] proposed a learning based method which identifies the needle in an X-ray image for tracking. Nevertheless, this approach might not cope with small and flexible objects, such as tracking the guidewire in interventional procedures. Also, these methods may heavily rely on intensity gradients, and thus are easily attracted to image noise as well as other wire-like structures in fluoroscopy.

B. DEEP LEARNING-BASED METHODS

CNNs have been validated to be very effective in object detection and tracking [28], [29]. Wang *et al.* [30] proposed a guidewire tracking method using region proposal networks (RPN). Compared to feature-based methods, RPN has several advantages; improved feature representation with proposals that adapt well to object variability such as diverse scales and aspect ratios. However, the method evaluates many regions of interest (ROIs), which is time consuming and does not consider the information of the previous frames. Baur *et al.* [31] proposed a method for catheter detection and depth estimation based on CNN. Notably, this method is also computationally expensive and the net response for an image at full scale (512×512) took approximately 1000 ms. More importantly, the majority of recent methods usually address the guidewire detection problem only without precise segmentation of the guidewire tip which is necessary for robot control. Recently, Chen and Wang [32] and Ambrosini *et al.* [33] introduced segmentation approaches using a U-net architecture [34] for tracking the guidewire in ultrasound images and X-ray fluoroscopic images, respectively. These methods achieved significant improvements compared to the feature-based methods in terms of guidewire extraction. However, localization of the small guidewire tip has often been unstable without temporal features of adjacent frames. To accurately track the guidewire tip, a two-step unified framework is proposed to incorporate both detection and segmentation for tracking. The proposed framework effectively segments the guidewire tip in real time by constraining the search space and considering temporal features between adjacent frames useful for precise tracking.

Algorithm 1 Tracking Procedure of the Proposed Method

1: Start from an initial frame.
2: Predict bounding box using Faster RCNN (Sec. III-A)
3: Extract center of mass of the bounding box
4: **Iterate** 5 $\sim$ 12 steps until the last frame,
5: Extract patch and perform segmentation using the patch-wise U-net (Sec. III-B)
6: Separate connected components using a contouring algorithm [35]
7: **If** no segmentation component is generated, go to 2
8: **Else** choose the closest component from the center of tip predicted in the previous frame. If there are more than two connected components, delete other components except the closest one.
9: Reconstruct the segmentation result to the frame
10: **If** it is the last frame, go to 13.
11: **Otherwise** the next frame is input.
12: Extract center of mass of the segmentation in the previous frame and go to 5
13:**Done**

III. TRACKING BY PATCH-WISE U-NET SEGMENTATION

The proposed tracker consists of a detection network and a segmentation network. Given the initial frame as input, the detection network predicts the bounding box locations containing the guidewire robot tip. A patch is extracted from the center of the bounding box, thereafter tip segmentation is performed via the segmentation network. Finally, the output patch is reconstructed to the original image size. Given the next frame as input, we perform segmentation on a patch centered on the segmented tip extracted in the previous frame. In the segmentation network, the label of the current patch is estimated by using both the image patch and its segmentation label predicted in the previous frames, as well as the image patch in the current frame. This process is repeated until the last frame.

Given the segmentation prediction, morphological processing [35] based on a contouring algorithm is employed, where each connected component is a closed loop of 2D points. However, if the number of components are more than two, we only select the component closest to the center of mass of the segmented tip in the previous frame and delete the rest. If no component is generated, we predict the bounding box using the detection network in a similar fashion with the initial frame. Algorithm 1 presents the overall steps of the proposed method. Also, the flowchart of our proposed method is shown in Fig. 1.

For the detection network, Faster RCNN [36] is employed to learn the relationship between the images and their corresponding bounding boxes that include the robot tip. At the same time, a modified U-net architecture serves as the segmentation network to model the relationship between image patches, tip segmentation of the previous frames along with

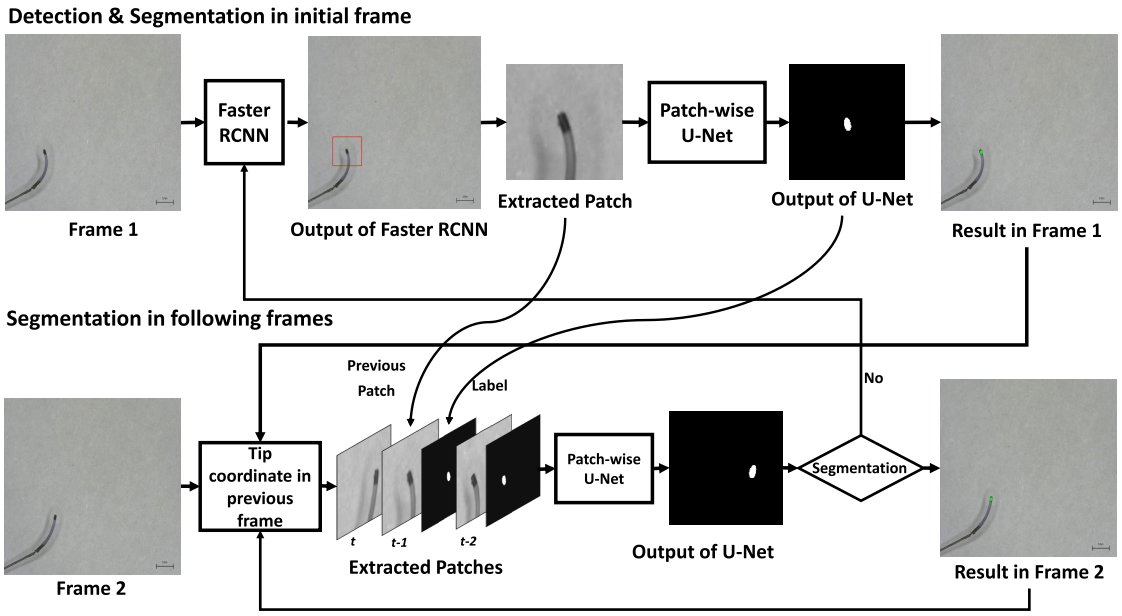

FIGURE 1. Overview of our proposed tracking method.

the current image patch and its corresponding tip segmentation. Moreover, the detection network can be replaced with other networks such as Yolo [37], RCNN [38], Fast RCNN [39] instead of Faster RCNN [36]. Similarly, the segmentation network can equally be replaced by FCN [40], DilatedNet [41], DeepLab-v3 [42] instead of U-Net [34]. The details of proposed networks are described in the subsections.

A. DETECTION NETWORK

Segmentation of a small tip from the whole image is often unstable. Thus, we localize the desired area by using a bounding box-based detection method that can reflect the overall characteristics around the tip. We adopt Faster RCNN [36] for this task given its shown good performance in various applications. In particular, the detection framework consists of a region proposal network (RPN), region of interest (ROI) pooling, region proposal layer, and bounding box regression modules. The RPN module is a fully convolutional network which can predict object bounds and scores at each position simultaneously. This module acts nearly in a cost-free way by sharing full-image convolutional features with a detection network based on a deep residual network with 50 layers (ResNet-50) [43]. Rectangular object proposals generated by RPN are fed to the fully connected layers for bounding box classification and regression. Regression towards the bounding boxes is achieved by comparing proposals relative to reference boxes.

B. SEGMENTATION NETWORK

We propose a segmentation network to achieve guidewire tip segmentation given a patch size of 160×160. The patch size was set so that the change of the tip across adjacent frame appears within the patch in all training samples. To utilize the spatial and temporal features between previous adjacent frames, we train the network with multiple patches along with segmentation masks extracted from previous frames. Specifically, if the adjacent frames are not considered, we extract an image patch and its corresponding segmentation at each frame in the training data and train the segmentation network. In the testing stage, the segmentation of a patch at t frame, centered on the center of mass of the tip segmentation in $t-1$ frame is estimated. On the other hand, if we consider the problem of estimating the segmentation at frame t with a previous frame, we train the segmentation network using three concatenated images (i.e., image patches at $t, t-1$ frames and a label patch at $t-1$ frame) as input and the corresponding label at t frame as output. In testing, the segmentation at t frame is estimated using the image patches at $t, t-1$ frames and the predicted label patch at $t-1$ frame. If n previous frames are considered, $(2 * n + 1)$ images are concatenated and then used as the input of the segmentation network (see Fig. 2).

The proposed segmentation network consists of a contracting path and expanding path similar to U-net [34]. In the contracting path, 3×3 convolution layers (unpadded), each followed by rectified linear units (ReLU) and down-sampling 2×2 max pooling operations with stride 2 are repeated. In the expansive path, up-sampling of features is followed by 2×2 convolutions (upconvolution) that expand the feature maps. Further, we concatenate these features with the corresponding feature maps of the contracting path followed by two repeated 3×3 convolutions each with ReLU. At the final layer, a 1×1 convolution is used to map each component feature to the desired number of classes, i.e., two classes: one for the robotic tip and another for background.

For training, the mini-batch size was set as 16 with Adam optimizer [44] used to minimize the Dice error between the prediction and the ground truth. The learning rate was

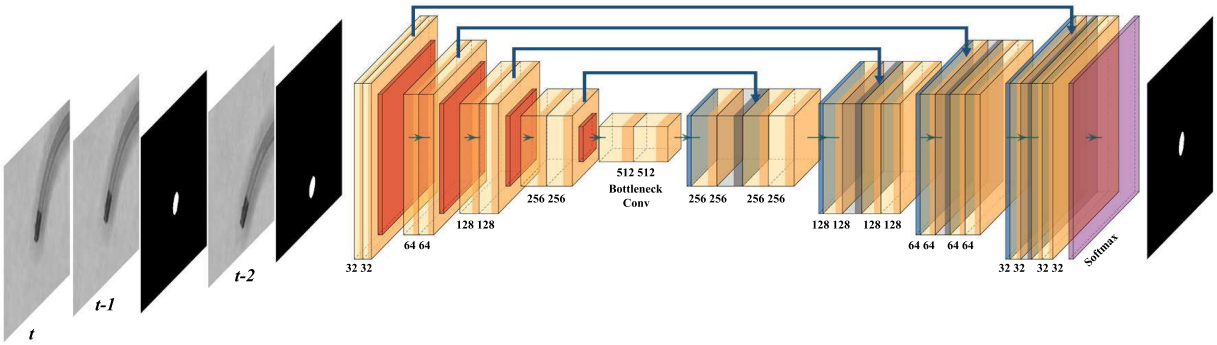

FIGURE 2. The proposed segmentation network with successive localized frames for guidewire tip segmentation at frame t. If two previous frames are used, the image patches at frame t, previous frames $t-1$, $t-2$, and the tip segmentations generated for frame $t-1$ and $t-2$ concatenated and passed through the network to predict the tip segmentation at frame t.

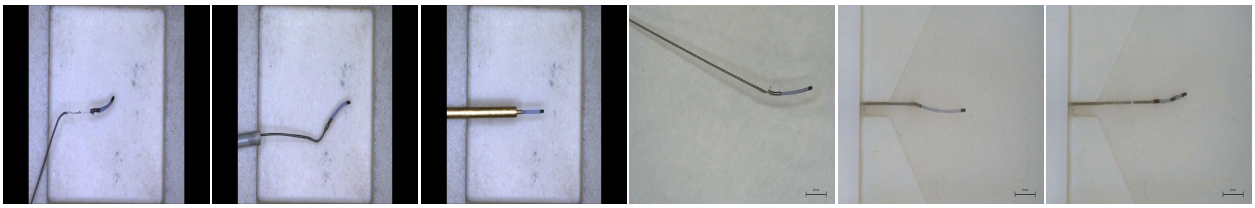

FIGURE 3. Dataset samples containing multiple type of guidewire robots.

set as 0.00001 with 200 epochs in total. The method was implemented using Keras with a Tensorflow backend and all training and testing performed on a workstation with NVIDIA Titan XP GPU.

IV. EXPERIMENTS AND RESULTS

A. DATA SET

For evaluation of the proposed method, we acquired 11 videos consisting a total of 11,884 frames collected from the DGIST-ETH Microrobot Research Center (DEMRC) using a VZM 600i Zoom Imaging Lens (Edmund Optics Inc., USA) linked to a camera (Grasshopper; Point Gray Research, Inc., Canada). The video sequences contain the guidewire moving in different directions with respect to the changes of magnetic fields in eight coils of the robotic system [9]. Moreover, the dataset has multiple types of guidewire robots with cluttered backgrounds and fast motion captured from top views (see Fig. 3).

To train the proposed tracker, ground truth annotations are created in a semi-automatic manner. A Particle Filter (PF) based tracking algorithm [45] is employed with the allocation of the tip location in the initial frames. Whenever tracking fails, we manually allocated the tip location, and then performed tracking. In order to generate a binary mask, we extracted the tip by conducting otsu thresholding [46] near the particle positions obtained by the PF algorithm. If the segmentation result is not accurate, manual correction is performed. Moreover, bounding boxes centered on the center of mass of the binary masks are generated to train the detection network.

B. EVALUATION SETTINGS

We divided 11 videos into two groups (5 and 6 videos, respectively), then performed 2-fold cross-validation. We first trained a model with 4 training videos and a single video for validation in the first fold, then applied the trained model to the remaining 6 videos in the second fold. Next, we trained a model with 5 videos and a single video for validation in the second fold, thereafter applied the model to 5 videos in the first fold. Evaluation was performed with the average accuracy scores on all 11 videos.

The proposed method was compared to the conventional feature based detection methods such as Boosting [13], MIL [14], KCF [15], TLD [16], MEDIANFLOW [47], and CSRT [17], as well as deep learning-based detection methods; Faster RCNN [36] and GOTURN [18], including the segmentation method U-Net [34]. Since most detection methods such as Boosting [13], MIL [14], KCF [15], TLD [16], MEDIANFLOW [47], CSRT [17] and GOTURN [18] predict a bounding box instead of the tip binary mask, we extracted the binary mask using otsu thresholding in the bounding box and compared against the ground truth masks. Moreover, for the methods which required the initial target object location such as the feature based detection methods and GOTURN [18], we allocated the tip position manually in the initial frame. In the case of feature based-methods, evaluation was performed on the entire set (11 videos) without cross validation given offline training is not required.

Further, additional experiments with data augmentation are included to confirm the robustness of the model to appearance and pose variations. The augmented data samples are

TABLE 1. Tracking performances of comparison methods.

Methods	Dice	IoU	CLE	Failures	Length	Time(ms)
Boosting [13]	8.91	6.84	330.88	811.45	102.82	13.8
KCF [15]	3.27	3.19	312.68	747.82	22.18	**3.1**
MIL [14]	31.71	15.50	141.04	683.64	85.55	56.3
CSRT [17]	29.71	21.78	173.05	636.45	279.27	21.7
TLD [16]	46.69	15.68	29.83	326.45	191.55	316.2
MEDIANFLOW [47]	74.87	58.20	317.45	545.82	184.55	7.2
GOTURN [18]	68.11	27.47	56.47	317.45	126.36	38.3
Faster RCNN [36]	72.0	64.0	7.54	705.64	6.09	371.8
U-Net [34]	88.07	85.07	26.8	119.45	612.64	71.0
Proposed (Single)	96.04	85.13	0.37	0.027	914.64	19
Proposed (Double)	97.12	84.40	0.38	0.09	1014.91	19
Proposed (Triple)	98.51	86.99	**0.36**	**0**	**1080.36**	19
Proposed (Augmented)	**98.58**	**87.02**	**0.36**	**0**	**1080.36**	19

created using transformations [48] such as scaling (from 0.5 to 1.5 ratio), horizontal and vertical flips, blurring with gaussian filters, elastic deformations with different scaling factors and elasticity coefficients [49] as well as rotations (90, 180, 270 degrees) on each input image.

Tracking accuracy was measured by Dice score, Intersection over Union (IoU), and central location error (CLE). The Dice score is a widely used overlap measure for pairwise comparison of binary segmentations of the foreground with the ground truth. Formally, it is represented as:

$$Dice = \frac{2 \times (A \cap B)}{A + B} \times 100 \tag{1}$$

where A is the ground truth and B is the predicated mask. Dice coefficient ranges from 0 to 1, where 1 means complete overlap. Further, the IoU metric, also referred to as the Jaccard index, is another metric to quantify the percent overlap between the target mask and our predicted output. It measures the number of similar pixels between the target and prediction masks divided by the total number of pixels present across both masks as:

$$IoU = \frac{A \cap B}{|A| \cup |B|} \times 100 \tag{2}$$

CLE is an evaluation metric to measure the Euclidean distance between the predicated center position and the ground truth center position of the guidewire tip. The center position was extracted from the binary mask. Formally, CLE is defined as:

$$CLE = \sqrt{(x_p - x_g)^2 + (y_p - y_g)^2} \tag{3}$$

where x_p, y_p are the predicted coordinates and x_g,y_g are the ground truth coordinates. To demonstrate the effectiveness of the proposed method, we also evaluated the tracking robustness by measuring the tracking length and failure rates. For the tracking length, we counted how long the tracker consistently tracks the objects across the sequences. As for failure rate, we report the total number of failures for each algorithm averaged over the total number of frames in the sequences. If the center of mass distance of a tracked object from ground-truth is larger a pre-defined threshold, we considered this as a tracking failure and resumed tracking with re-initialization of the center location until the last frame. The threshold was heuristically set as 10 in our experiments.

Finally, we also measured tracking time per image.

C. QUANTITATIVE RESULTS

1) TRACKING ACCURACY

Table 1 shows the accuracy scores of several benchmarked methods, including the proposed methods on the test sequences according to three key metrics i.e. Average Dice, IoU and CLE, respectively. The distributions of tracking accuracy scores are also shown with box plots in Fig. 4. In the majority of the test sequences, our method achieved the highest scores in terms of Dice and IoU, as well as the lowest scores for CLE (lower is better) among the comparative methods.

KCF [15] reports the lowest performance at 3.27% and 3.19% for Dice and IoU, respectively. Although this method requires the lowest computation, it fails to successfully track the guidewire with reference to the central location as is evident from the large CLE score of 312.68. KCF [15] tracker was proposed to augment the principle ideas of Boosting [13] and MIL [14] based trackers by exploiting mathematical properties that make tracking faster and report tracking failures better than the previous. However, it fails in tracking small object tips such as the guidewire, given its performance is not reliable. More especially, these methods fail to recover from occlusion.

On the contrary, MEDIANFLOW [47] reports the best performance relative to all the feature-based methods including one of the learning-based methods (GOTURN [18]), i.e. 74.87%, 58.20% and 37.31% for Dice, IoU and CLE, respectively. This tracker tracks a given object by considering both the forward and backward directions in time as well

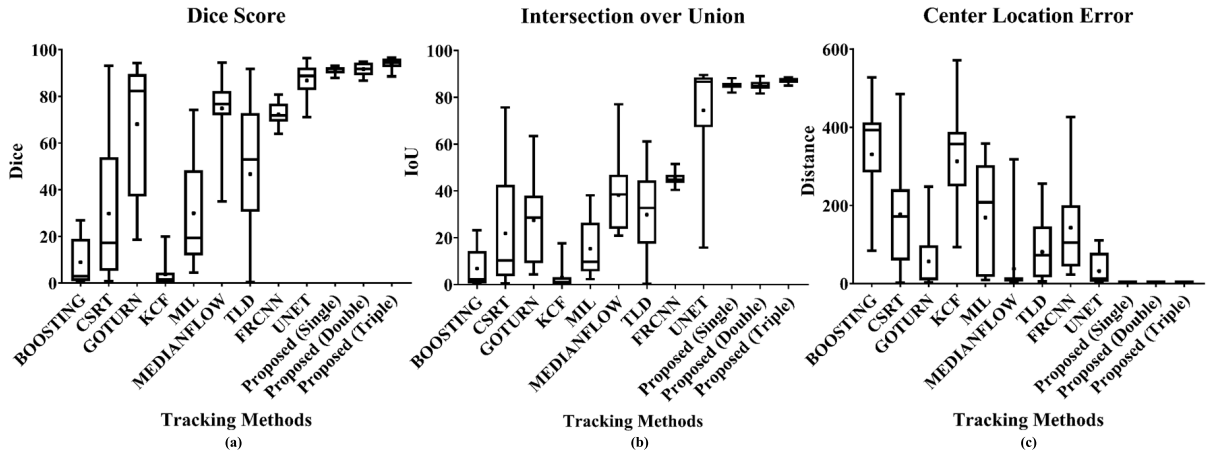

FIGURE 4. Box plots of the (a) Dice score, (b) Intersection over Union and (c) center location error over 11 test video sequences. The top, center and bottom lines of each box represent upper quartile, median, and lower quartile scores, respectively. The dot shows the mean values. The upper and lower whiskers represent the maximum and minimum scores, respectively.

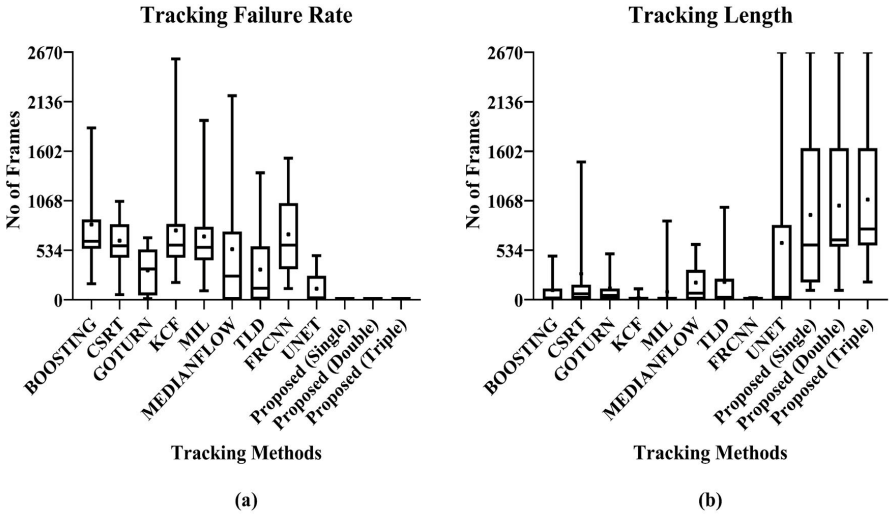

FIGURE 5. Box plots of the (a) Tracking Failure and (b) Tracking Length scores. The top, center,dot and bottom lines of each box represent upper quartile, median,mean, and lower quartile scores, respectively. The upper and lower whiskers represent the maximum and minimum scores, respectively.

as measures the differences between these trajectories for accurate tracking across initial and current frames. Accurate tracking is achieved via the selection of reliable trajectories and detection of failures by minimizing the forward-backward error. However, this tracker fails under large motion and scale variations. In several sequences, the guidewire exhibits abrupt motion in-turn leading to failure. Moreover, MEDIANFLOW [47] implicitly assumes a point-based representation (a set of points initialized on a rectangular grid within an initial bounding box) where the target object is composed of small rigid patches; thus, when the object does not satisfy this assumption, point voting fails and rejects several points in the target frame, consequently increasing the error rate.

As for the learning-based methods, U-Net [34] reports the lowest CLE score and largely outperforms the counterpart learning-based methods i.e. Faster RCNN [36] and GOTURN [18]. Our initial assumption was that U-Net [34] may be effective at segmenting small objects across frames compared to previous methods. Despite the reasonable performance, owing to the nature of the architecture, it is not built to recover from failures. Thus, by combining both detection and segmentation in a single framework to better address failures, our method reports 96.04% (+7.97) when only a single patch/frame is considered during inference, marking a huge improvement over the approaches that use either detection or segmentation models only. Moreover, to better handle the challenge of abrupt motion we encode robustness to large displacements by using multiple frames i.e. double or triple. Notably, optimal performance is achieved with 3 frames as using more than 3 frames results in suboptimal performance due to the catheter tip moving out scope of the patch region and as a result some patches among successive frames may not contain any catheter tip. The accuracy

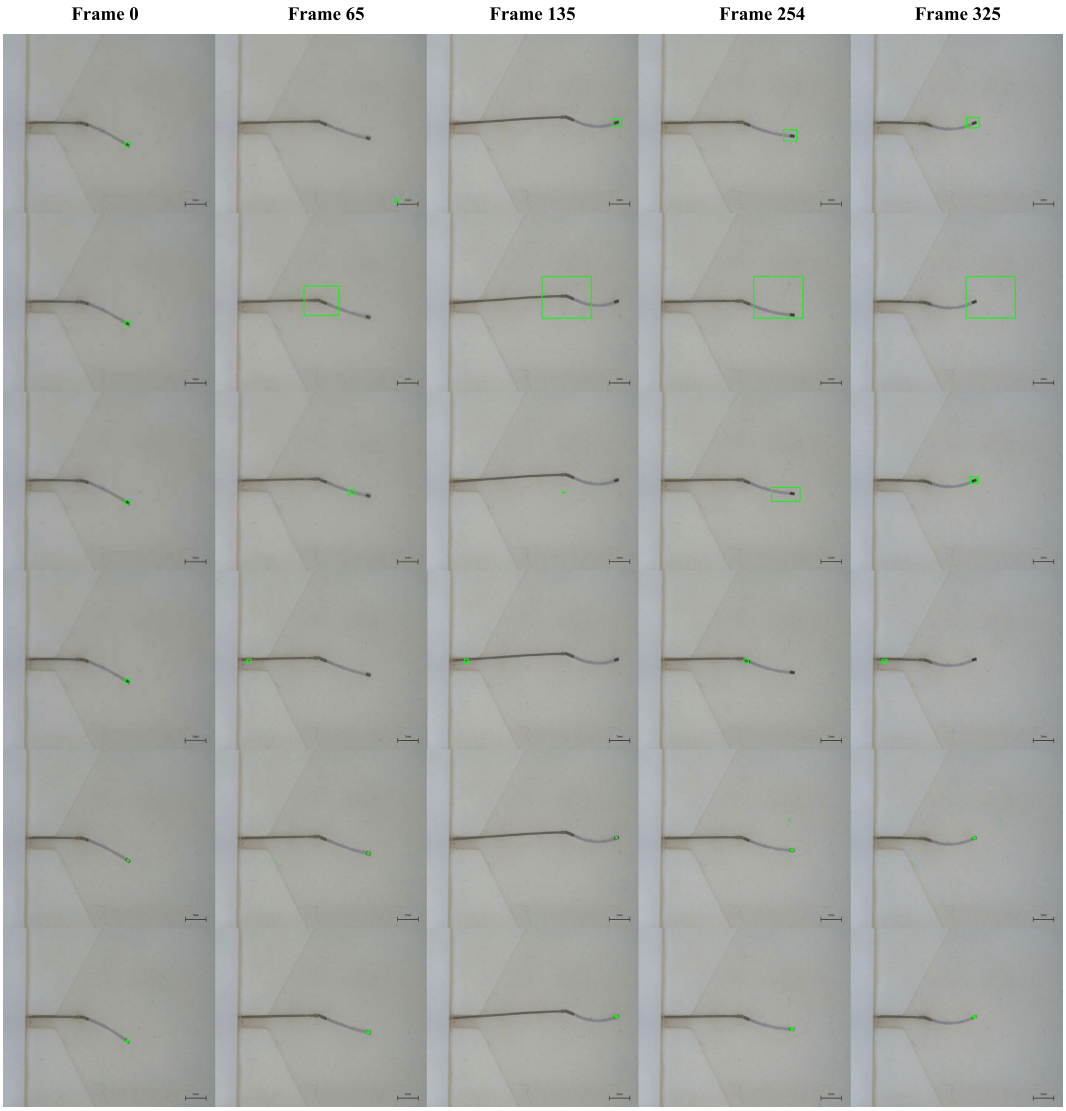

FIGURE 6. From top to bottom: inputs, TLD [16], MEDIANFLOW [47], GOTURN [18], FRCNN [36], and U-Net [34] and the proposed method. Intermediate results highlighting failure cases across different methods.

of the model trained with the augmented data was slightly better.

2) TRACKING ROBUSTNESS

Fig. 5 (a) and (b) present the tracking failure rate and length, respectively. Indeed, most conventional methods failed in many frames, and thus required re-initialization that resets the position with respect to the ground-truth. Notably, only TLD [16] and MEDIANFLOW [47] were able to consider the deformation of the guidewire tip exhibiting fewer failures compared to Faster-RCNN [36] and GOTURN [18]. Due to the small size of the guidewire tip and non-rigid shape, most trackers failed to successfully recover from failure, especially in sequences with an abrupt motion (see Section 4.4 for sequences with such motion). On the other hand, U-Net [34] was consistent across all sequences with little to no failure reported. We hypothesize the segmentation stage of the proposed method was significant in avoiding failure by considering the small size of the tip. Moreover, we complement this setup with auto re-initialization and ensure the central location of the target in all time-steps.

3) PROCESSING TIME

We report the processing time per image of each tracker on the right side of Table 1. Conventional trackers KCF [15] and MEDIANFLOW [47] are fast, but the accuracy scores were limited. GOTURN [18] was the fastest among the previous deep learning-based methods, but the accuracy was low. On the other hand, our patch-wise segmentation method shows considerable speed up (less than 20ms) by reducing the input feature size. Moreover, no significant change was noted among the proposed methods (i.e., single, double, and triple) since the size of input features was considerably small even with multiple adjacent frames.

D. QUALITATIVE RESULTS

Fig. 6 illustrates the performance of several methods on a test video where many failures were incurred. Most methods did track the object correctly across several frames, but tracking accuracy and robustness scores were lower especially in the case of MEDIANFLOW [47] which fails to correctly estimate the bounding box after an initial failure. In some cases, the predicted bounding box showed incorrect scaling, further highlighting the challenge of scaling also observed by GOTURN [18] which falsely detected the wire as a tip (Fig. 6, row 3). Moreover, GOTURN [18] and Faster-RCNN [36] often exhibited drifting, resulting in failure to recover from false detections. Notably, U-Net [34] showed consistent tracking in the first few sequences, however, several predicted false positives later affect tracking in the next frames. Moreover, when failure occurs there is a large error in terms of central location. Despite the failures incurred by prior methods, our proposed approach shows consistent performance under the challenging scenarios. Based on the results, we can confirm the effectiveness of our proposed patch-wise tip segmentation strategy. Further, our methods show more consistent tracking compared to bounding-box regression methods in small search windows without sacrificing efficiency.

V. DISCUSSION

The proposed methods considering single and double previous frames failed once in some test sequences with an average tracking failure rate of 0.027 and 0.09 points in both cases. Interestingly, the proposed method with triple frames successfully tracked the guidewire tip with 100% accuracy in all the test sequences, because multiple patches better encode large motion displacements whereas augmentation facilitates more scale and pose invariance. The proposed method also emphasizes efficient feature learning by reducing the search window centered on the target object. Considering the tracking length; methods with low failures report relatively prolonged tracking times. Overall, the proposed method consistently outperforms all methods with considerable time, asserting our initial hypothesis that using features in successive localized frames plays a crucial role in improving performance emphasizing less errors across time-steps.

We also confirm the performance of the proposed method when trained with extensive data augmentation. The proposed network was trained with 11,884 frames, which is considerably sufficient and included several rotational and translation variations. Thus, it was robust to variations during inference. Further, additional experiments with data augmentation show that the proposed method was robust to both scaled and low-quality images as well as robust towards pose variance. The proposed model trained with augmented data achieved the best accuracy in terms of dice score and IoU.

Despite the considerable performance of the proposed method across different settings; we highlight a few drawbacks and limitations that require careful attention; First of all, robot control experiments were not considered in this study; thus, it would be beneficial to assess the impact of varied control settings as a function of model performance for both training and inference. In addition, this work largely focuses on experimental settings based on camera (natural) images and does not consider X-Ray fluoroscopic images as well as how the proposed method can be applied in the later setting given the domain shift between fluoroscopy and camera images. In future, we plan to address these drawbacks and conduct robot control experiments that move the guidewire tip to a desired location without sensors. Furthermore, we will extend our method to automatically track robot guidewire tips in X-ray images by adopting the domain transfer learning.

VI. CONCLUSION

In this paper, a deep learning-based tracking method is proposed for localizing small guidewire robot tips in video sequences. The proposed tracker consisting of a detection and segmentation network accurately delineates small tips by utilizing both spatial and temporal features. Compared to previous guidewire tracking methods, the proposed method does not require multiple sensors, heuristic manual tuning, and post-processing steps. Further, we show that pixel-wise predictions enable accurate guidewire tip localization.

REFERENCES

[1] N. Dey, A. S. Ashour, F. Shi, and R. S. Sherratt, "Wireless capsule gastrointestinal endoscopy: Direction-of-arrival estimation based localization survey," *IEEE Rev. Biomed. Eng.*, vol. 10, pp. 2–11, 2017.

[2] E. M. Khan, W. Frumkin, G. A. Ng, S. Neelagaru, F. M. Abi-Samra, J. Lee, M. Giudici, D. Gohn, R. A. Winkle, J. Sussman, B. P. Knight, A. Berman, and H. Calkins, "First experience with a novel robotic remote catheter system: Amigo mapping trial," *J. Intervent. Cardiac Electrophysiol.*, vol. 37, no. 2, pp. 121–129, 2013.

[3] J. A. Borgstadt, M. R. Zinn, and N. J. Ferrier, "Multi-modal localization algorithm for catheter interventions," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2015, pp. 5350–5357.

[4] P. Wang, T. Chen, Y. Zhu, W. Zhang, S. K. Zhou, and D. Comaniciu, "Robust guidewire tracking in fluoroscopy," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 691–698.

[5] A. Barbu, V. Athitsos, B. Georgescu, S. Boehm, P. Durlak, and D. Comaniciu, "Hierarchical learning of curves application to guidewire localization in fluoroscopy," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[6] Y. Zhao, S. Guo, Y. Wang, J. Cui, Y. Ma, Y. Zeng, X. Liu, Y. Jiang, Y. Li, L. Shi, and N. Xiao, "A CNN-based prototype method of unstructured surgical state perception and navigation for an endovascular surgery robot," *Med. Biol. Eng. Comput.*, vol. 57, no. 9, pp. 1875–1887, 2019.

[7] P.-L. Chang, A. Rolls, H. De Praetere, E. V. Poorten, C. V. Riga, C. D. Bicknell, and D. Stoyanov, "Robust catheter and guidewire tracking using B-spline tube model and pixel-wise posteriors," *IEEE Robot. Autom. Lett.*, vol. 1, no. 1, pp. 303–308, Jan. 2016.

[8] A. K. Hoshiar, S. Jeon, K. Kim, S. Lee, J.-Y. Kim, and H. Choi, "Steering algorithm for a flexible microrobot to enhance guidewire control in a coronary angioplasty application," *Micromachines*, vol. 9, no. 12, p. 617, 2018.

[9] S. Jeon, A. K. Hoshiar, S. Kim, S. Lee, E. Kim, S. Lee, K. Kim, J. Lee, J.-Y. Kim, and H. Choi, "Improving guidewire-mediated steerability of a magnetically actuated flexible microrobot," *Micro Nano Syst. Lett.*, vol. 6, no. 1, 2018, Art. no. 15.

[10] S. H. Park, S. Lee, I. D. Yun, and S. U. Lee, "Hierarchical MRF of globally consistent localized classifiers for 3D medical image segmentation," *Pattern Recognit.*, vol. 46, no. 9, pp. 2408–2419, 2013.

[11] S. H. Park, S. Lee, I. D. Yun, and S. Lee, "Structured patch model for a unified automatic and interactive segmentation framework," *Med. Image Anal.*, vol. 24, no. 1, pp. 297–312, 2015.

[12] M. Hoffmann, A. Brost, M. Koch, F. Bourier, A. Maier, K. Kurzidim, N. Strobel, and J. Hornegger, "Electrophysiology catheter detection and reconstruction from two views in fluoroscopic images," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 567–579, Feb. 2015.
[13] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proc. Brit. Mach. Vis. Conf.*, vol. 1, no. 5, 2006, p. 6.
[14] B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 983–990.
[15] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
[16] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.
[17] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6309–6318.
[18] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 FPS with deep regression networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 749–765.
[19] I. Ullah, P. Chikontwe, and S. H. Park, "Guidewire tip tracking using U-Net with shape and motion constraints," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Feb. 2019, pp. 215–217.
[20] E. Franken, P. Rongen, M. van Almsick, and B. ter Haar Romeny, "Detection of electrophysiology catheters in noisy fluoroscopy images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2006, pp. 25–32.
[21] Y. Ma, A. P. King, N. Gogin, C. A. Rinaldi, J. Gill, R. Razavi, and K. S. Rhode, "Real-time respiratory motion correction for cardiac electrophysiology procedures using image-based coronary sinus catheter tracking," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2010, pp. 391–399.
[22] 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," *Med. Phys.*, vol. 45, no. 11, pp. 5066–5079, 2018.
[23] D. Palti-Wasserman, A. M. Brukstein, and R. P. Beyar, "Identifying and tracking a guide wire in the coronary arteries during angioplasty from X-ray images," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 2, pp. 152–164, Feb. 1997.
[24] S. de Buck, J. Ector, A. La Gerche, F. Maes, and H. Heidbuchel, "Toward image-based catheter tip tracking for treatment of atrial fibrillation," in *Proc. Workshop Cardiovascular Intervent. Imag. Biophys. Modelling (CI2BM09–MICCAI)*, London, U.K., Sep. 2009, p. 8.
[25] P. Fallavollita, P. Savard, and G. Sierra, "Fluoroscopic navigation to guide RF catheter ablation of cardiac arrhythmias," in *Proc. 26th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 1, Sep. 2004, pp. 1929–1932.
[26] G. Slabaugh, K. Kong, G. Unal, and T. Fang, "Variational guidewire tracking using phase congruency," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2007, pp. 612–619.
[27] P. Mountney, A. Maier, R. I. Ionasec, J. Boese, and D. Comaniciu, "Method and system for obtaining a sequence of X-ray images using a reduced dose of ionizing radiation," U.S. Patent 9 259 200, Feb. 16, 2016.
[28] G. Zhu, F. Porikli, and H. Li, "Robust visual tracking with deep convolutional neural network based object proposals on pets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 26–33.
[29] G. Ning, Z. Zhang, C. Huang, X. Ren, H. Wang, C. Cai, and Z. He, "Spatially supervised recurrent convolutional neural networks for visual object tracking," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2017, pp. 1–4.
[30] 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 *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 3169–3175.
[31] C. Baur, S. Albarqouni, S. Demirci, N. Navab, and P. Fallavollita, "CathNets: Detection and single-view depth prediction of catheter electrodes," in *Proc. Int. Conf. Med. Imag. Augmented Reality*. Cham, Switzerland: Springer, 2016, pp. 38–49.
[32] S. Chen and S. Wang, "Deep learning based non-rigid device tracking in ultrasound image," in *Proc. 2nd Int. Conf. Comput. Sci. Artif. Intell.*, 2018, pp. 354–358.
[33] 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 *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 577–585.
[34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[35] S. Suzuki and K. Be, "Topological structural analysis of digitized binary images by border following," *Comput. Vis., Graph., Image Process.*, vol. 30, no. 1, pp. 32–46, 1985.
[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[37] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.
[38] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[39] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.
[40] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[41] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: https://arxiv.org/abs/1412.7062
[42] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: https://arxiv.org/abs/1706.05587
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[45] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund, "Particle filters for positioning, navigation, and tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 425–437, Feb. 2002.
[46] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
[47] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2756–2759.
[48] A. Buslaev, A. Parinov, E. Khvedchenya, V. I. Iglovikov, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," 2018, *arXiv:1809.06839*. [Online]. Available: https://arxiv.org/abs/1809.06839
[49] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. ICDAR*, vol. 3, Aug. 2003, pp. 958–963.

IHSAN ULLAH received the B.S. degree in computer science from Abdul Wali Khan University, Pakistan, in 2015, and the M.S. degree in computer science from Chonbuk National University, South Korea, in 2018. He is currently pursuing the Ph.D. degree with the Department of Robotics Engineering, Daegu Gyeonbuk Institute of Science and Technology, Daegu, South Korea.

His research interests include deep learning, computer vision, and medical image processing.

PHILIP CHIKONTWE received the B.S. degree in computer science from Abdelhamid Mehri Constantine University, Algeria, in 2015, and the M.S. degree in computer science from Chonbuk National University, South Korea, in 2018. He is currently pursuing the Ph.D. degree with the Department of Robotics Engineering, Daegu Gyeonbuk Institute of Science and Technology (DGIST), Daegu, South Korea.

His research interests include medical image analysis, computer vision, and machine learning.

SANG HYUN PARK received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2008, and the Ph.D. degree in electrical and computer engineering from Seoul National University, Seoul, in 2014.

From 2014 to 2016, he was a Postdoctoral Fellow of the Image Display, Enhancement, and Analysis (IDEA) Laboratory, Department of Radiology, University of North Carolina at Chapel Hill (UNC-CH), NC, USA. From 2016 to 2017, he was a Postdoctoral Fellow of the SRI International at Menlo Park, CA, USA. Since 2017, he has been an Assistant Professor with the Daegu Gyeonbuk Institute of Science and Technology (DGIST), Daegu, South Korea. His research interests include medical image analysis, computer vision, and machine learning.