Real-time device detection with rotated bounding boxes and its clinical application

YingLiang Ma[1], Sandra Howell[2], Aldo Rinaldi[2], Tarv Dhanjal[3] and Kawal S. Rhode[2]

[1] School of Computing Sciences, University of East Anglia, UK
[2] School of Biomedical Engineering and Imaging Sciences, King's College London, UK
[3] Warwick Medical School, The University of Warwick, UK
yingliang.ma@uea.ac.uk

Abstract. Interventional devices and insertable imaging devices such as transesophageal echo (TOE) probes are routinely used in minimally invasive cardiovascular procedures. Detecting their positions and orientations in X-ray fluoroscopic images is important for many clinical applications. Nearly all interventional devices used in cardiovascular procedures contain a wire or wires and are inserted into major blood vessels. In this paper, novel attention mechanisms were designed to guide a convolution neural network (CNN) model to the areas of wires in X-ray images. The first attention mechanism was achieved by using multi-scale Gaussian derivative filters in the first convolutional layer inside the proposed CNN backbone. By combining these multi-scale Gaussian derivative filters together, they can provide a global attention on the wire-like or tube-like structures. Furthermore, the dot-product based attention layer was used to calculate the similarity between the random filter output and the output from the Gaussian derivative filters, which further enhances the attention on the wire-like or tube-like structures. By using both attention mechanisms, a high-performance CNN backbone was created, and it can be plugged into light-weighted CNN models for multiple object detection. An accuracy of 0.88±0.04 was achieved for detecting an echo probe in X-ray images at 58 FPS, which was measured by intersection-over-union (IoU). Based on the detected pose of the echo probe, 3D echo can be fused with live X-ray images to provide a hybrid guidance solution. Codes are available at https://github.com/YingLiangMa/AttWire.

Keywords: Rotated Object Detection, X-ray Imaging, Attention CNN.

1 Introduction

Minimally invasive cardiovascular procedures are routinely carried out to treat diseases such as coronary heart diseases, valvular heart disease, congenital heart diseases and more. The procedure is usually guided using X-ray fluoroscopy and interventional devices and insertable imaging devices are routinely used during the procedure. Real-time object detection for medical devices is one of the most important tasks in hybrid guidance systems as well as robotic procedure systems. Hybrid guidance systems for minimally invasive cardiovascular procedures involves fusing information from Magnetic

Resonance Imaging (MRI) images, CT images or real-time 3D transesophageal echo (TOE) with X-ray fluoroscopy [1][2]. Device detection can facilitate the hybrid guidance system using both 3D echo volumes and X-ray fluoroscopic images, and the real-time registration is achieved by detecting the pose of the TOE probe in X-ray images [3]. Real-time device detection also facilitates motion compensation and automatic registration in MRI or CT based hybrid guidance systems [4]. Furthermore, knowing the locations of devices may allow procedures with complete or shared autonomy with robots in the near future.

The detection of the TOE probe and interventional devices in X-ray images has been previously studied. Existing methods can be divided into two categories: traditional computer vision techniques [5][6] and learning-based methods [7][8][9][10]. Methods based on traditional computer vision techniques are prone to errors due to image artifacts and the presence of other similar objects. Although learning-based methods have demonstrated a great potential to detect devices robustly, they relied on manual feature selection. Therefore, these methods are not easily transferred to other target devices or detect multiple devices at the same time.

In recent years, state-of-the-art multiple object detection methods have been developed to detect and identify common objects (e.g. vehicles, people, animals and more) [11]. The majority of these methods use axis-aligned bounding boxes to locate the target objects. However, our proposed method requires rotated bounding boxes as medical devices in X-ray images often have arbitrary orientations and rotated bounding boxes are more accurate to determine their locations. Furthermore, applications such as the hybrid guidance using echo and X-ray images requires the orientation of the TOE probe in X-ray images. Few deep-learning based object detection methods using rotated bounding boxes have been developed and they are mainly in the domain of satellite image analysis [12]. However, all existing methods do not meet our requirements for device detection. First, existing methods do not optimize for grayscale X-ray images and lack attention mechanisms for our target objects. Secondly, existing methods do not have sufficient accuracy, robustness or speed to be used in our applications. Therefore, we designed a convolution neural network (CNN) from scratch to achieve our requirements and also to take advantage of additional information available in X-ray images. Many interventional devices contain a wire and wire mesh and are inserted into major blood vessels. Insertable imaging devices are tube-like structures. Therefore, novel attention mechanisms using trainable pre-defined filters and an attention layer were designed to guide our CNN models to the areas of wires in the X-ray images.

2 Method

2.1 Image acquisition and image synthesis

10,072 X-ray images were acquired in 43 different clinical cases using a mono-plane X-ray system at *** Hospital. There were 6,533 images from 9 transcatheter aortic valve replacement (TAVR) procedures, 250 images from one atrial fibrillation (AF) procedure guided by X-ray and transesophageal echo images and 3,289 images from 33 standard AF procedures.

As 4,789 images out of total 10,072 images do not contain the TOE probe, a method of image synthesis has been developed to automatically insert an image patch of a TOE probe. It is based on Poisson image editing (PIE) [13][14], which blends an image patch into the context of a destination image. The blending was achieved via solving the equation (1).

$$\min_{f_{in}} \iint_{\Omega} |\nabla f_{in} - v|^2 \text{ with } f_{in}|_{\partial\Omega} = f_{out}|_{\partial\Omega} \tag{1}$$

where ∇ is the gradient operator. The goal of eq. (1) is to find the intensity values f_{in} within the masked area (Ω) of image patch matching with the surrounding values f_{out} of the destination image. A binary mask will be used to create the masked area (Ω), which is the loose selection of the blending object. $\partial\Omega$ is the border of the masked area and v is the image gradient within the masked area. Figure 1 gives an example of PIE.

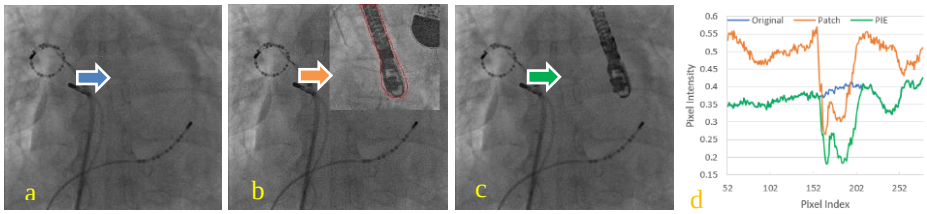

Fig. 1. An example of PIE. (a) The original image. (b) Overlay the image patch with the original image. The red contour is the border of the masked area ($\partial\Omega$). (c) Image after applying PIE. (d) The intensity profiles. Arrows in the images indicate the location of the intensity profiles.

40 image patches were extracted from 40 image sequences which contain the TOE probe. Data augmentation techniques were used to increase the variation of the pose of the TOE probe in X-ray images. Random rotations and translations were applied to the extracted image patches. There are restrictions applied to random rotations and translations to ensure the generated image are anatomically correct.

2.2 The attention backbone

Our clinical applications require real-time object detection while maintaining high accuracy and robustness. To achieve this goal, attention mechanisms were designed to take advantage of additional information about the location and structure of medical devices. Many devices contain a wire and wire mesh or tube-like structures. To guide the attention of our CNN models, the multi-scale Gaussian derivative filters were used in the first convolution layer to enhance the visibility of wire-like or tube-like objects [15]. This process involves the calculation of a 2x2 Hessian matrix, and it is computed at every image pixel [16]. The Hessian matrix H consists of second order derivatives that contain information about the local curvature. H is defined such as:

$$H = \begin{bmatrix} L_{xx}(x,y;s) & L_{xy}(x,y;s) \\ L_{yx}(x,y;s) & L_{yy}(x,y;s) \end{bmatrix} \tag{3}$$

Where $L_{xy}(x, y; s) = \frac{\partial^2 L(x,y;s)}{\partial x \partial y}$ and the other terms are defined similarly. Here, $L_{xy}(x, y; s) = L_{yx}(x, y; s)$. $L(x, y; s)$ is an image smoothed by a Gaussian filter of the appropriate scale s. $L(x, y; s)$ is computed as $L(x, y; s) = L(x, y) * G(x, y; s)$, where $*$ is the convolution operator and the Gaussian filter $G(x, y; s) = \frac{1}{2\pi s} e^{-(x^2+y^2)/2s}$. Therefore, eq. (3) can be converted to

$$H = \begin{bmatrix} L(x,y) * G_{xx}(x,y;s) & L(x,y) * G_{xy}(x,y;s) \\ L(x,y) * G_{yx}(x,y;s) & L(x,y) * G_{yy}(x,y;s) \end{bmatrix} \quad (4)$$

Where $G_{xx}(x, y; s)$, $G_{yy}(x, y; s)$ and $G_{xy}(x, y; s)$ are Gaussian derivatives and are often known as Laplacian of Gaussians (LoG). In practice, we just pre-compute the masks of these Gaussian derivatives, convolve with the input image. By combining these multi-scale Gaussian derivative filters together, they can provide a global attention on the wire-like or tube-like structures.

The architecture of the attention backbone is illustrated in figure 3 and 15 LoG filters are used in the first convolution layer to provide the first attention mechanism. Among 15 filters, there are five groups, and each group contains three LoG filters with the same scale factor s, which are defined as $G_{xx}(x, y; s)$, $G_{yy}(x, y; s)$ or $G_{xy}(x, y; s)$. To accommodate different sizes of objects on the wires, five different scale factors were used in five groups of LoG filters. To calculate the scale factor s_0 for object size r_0, we use $s_0 = ((r_0 - 1)/3)^2$. This equation is motivated by the "3σ" ($s_0 = \sigma^2$) rule that 99% of energy of the Gaussian is within three standard deviations. The final multiscale s_0 is in the range of $0.11 \leq s_0 \leq 9$ and it is based object size from 2 to 10 (Unit is in image pixels) in an image with a 200x200 resolution. The second attention mechanism is achieved by a dot-product based attention layer [17], which calculates the similarity between the random filter output and the output from LoG filters (figure 3). The attention layer further enhances the attention on the wire-like or tube-like structures.

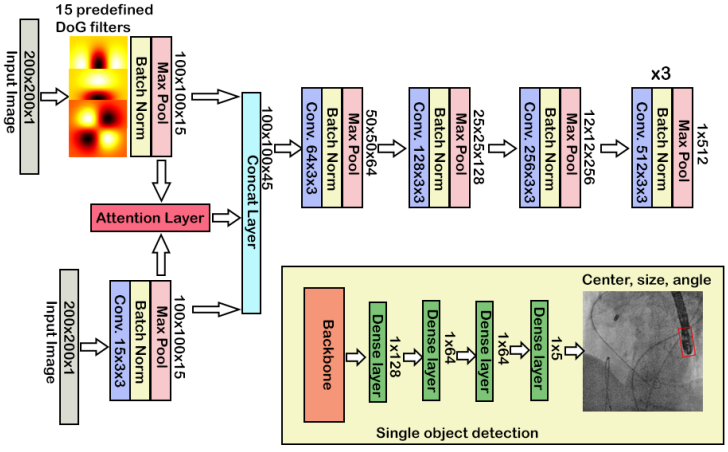

Fig. 3. Attention backbone and an example of its usage in single object detection.

2.3 Single object detection

Firstly, a customized CNN was designed for single object detection using the proposed attention backbone. As shown in figure 3, the localization of a rotated bounding box is achieved by the output of the final dense layer, which provides five parameters: center (x, y), size (w, h) and angle (δ).

Two object detectors were trained, one for the TOE probe and the other one for the transcatheter aortic valve (before deployment). A modulated rotation loss function was designed and adapted from [18]. It minimizes the difference between the predicted values and ground truth values. All five parameters which define the rotated bonding box were normalized between 0 and 1 to avoid errors caused by objects on different scales. The loss function is defined as:

$$l_{cp} = |x_g - x_p|/W_{img} + |y_g - y_p|/H_{img} \tag{5}$$

$$l_{mr} = min\begin{cases} l_{cp} + |w_g - w_p|/W_{img} + |h_g - h_p|/H_{img} + |\delta_g - \delta_p|/90^{\circ} \\ l_{cp} + |w_g - h_p|/W_{img} + |h_g - w_p|/H_{img} + \left|90^{\circ} - |\delta_g - \delta_p|\right|/90^{\circ} \end{cases} \tag{6}$$

where l_{cp} is the central point loss, (x_p, y_p) is the predicted center point and (x_g, y_g) is the ground-truth center point. Eq. (6) is for the exchangeability of height and width.

As shown in figure 4, the activation maps of selected layers were visualized to illustrate the model attentions in the aortic valve detector. The model global attention is clearly on the wires in the first layer and then enhanced by the attention layer. Finally, the model shifts the attention to the local areas of the target object (the aortic valve) in the final convolution layer.

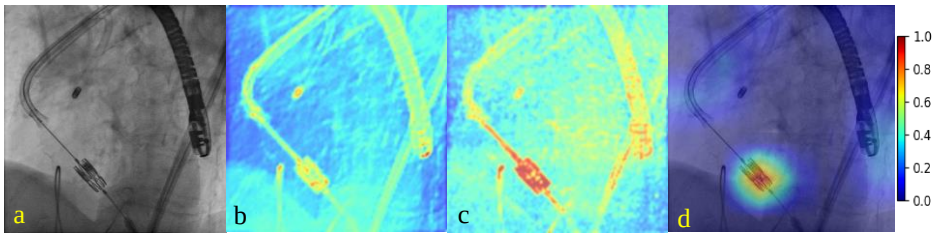

Fig. 4. (a) The original image. (b) The activation map from the convolution layer with 15 LoG filters. (c) The activation map from the attention layer. (d) The activation map from the last convolution layer in the attention backbone.

2.4 Multiple object detection

Inspired by the CenterNet [19], a one-stage multiple-object detector was designed by using a similar attention backbone proposed in this paper. The one-stage detector can achieve a higher inference speed and it is suitable for real-time applications. The proposed detector is a light-weight CNN model and only contains 3.7M trainable parameters for two-class object detector. As shown in figure 5, the proposed CNN model consists of down-sampling layers and up-sampling layers. The model has four outputs. The

first one is the center-point heatmap and it is used to localize the center point *(x, y)* of the rotated bounding box. The second output is used to determine the object size *(w, h)*. The third output is the rotate angle *(δ)* of the bounding box. The fourth output is the offset output, and it is used to recover from the discretization error caused due to the downsampling of the input. For example, in our model, the input image resolution is 256x256 and the image resolution after the last up-sampling layer is 128x128. If the ground truth of center point is (x_g, y_g) in center heatmap output, the corresponding ground truth of center point in the input image is $(2x_g + \varepsilon_x, 2y_g + \varepsilon_y)$. Both ε_x and ε_y are discretization errors and they are either 0 or 1 in our model.

The proposed CNN model not only outputs a rotated bounding box for each object but also outputs a confidence value. Therefore, the model can predict whether the target object exists in the image or not. The model also can achieve multiple object detection as it has multiple channels of center heatmaps and each channel can localize the center points of one class of objects. The ground-truth heatmap for center points is not defined as either 0 or 1 because locations near the target point should get less penalization than locations far away. Therefore, Gaussian heatmap $e^{\frac{\|P-P_g\|^2}{2\sigma^2}}$ was used and P is the predicted center point and P_g is the ground truth. σ is set to 1/3 of the radius, which is determined by the size of objects. Focal loss [20] is used in the output for center-point heatmap and it is mainly to solve the problem of imbalanced classification in target detection. The loss functions for the remaining outputs are L1 loss function. Figure 6 presents some results of center-point heatmaps and detection results.

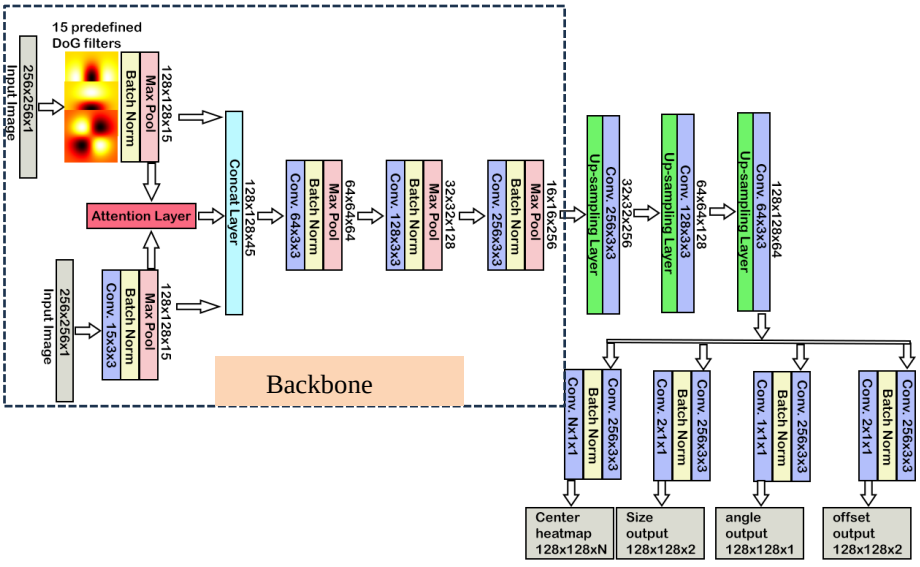

Fig. 5. The CNN architecture for multiple object detection. N is the number of classes.

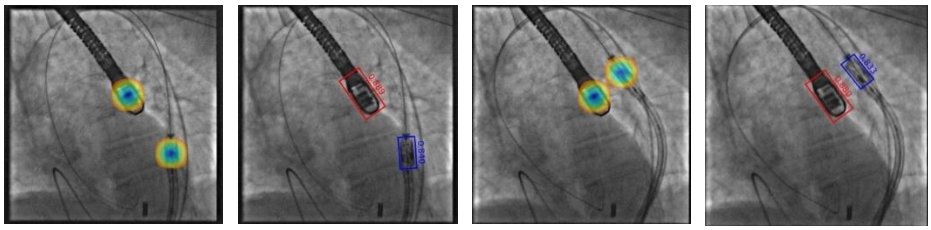

Fig. 6. Center point heatmaps and object detection with confidence values.

2.5 Clinical application: fusing 3D echo with live X-ray images

The object detector can detect the location as well as in-plane rotation (figure 7(a)) and scale of the TOE probe. There are two additional rotations: roll and pitch (figure 7b), which are out-of-plane. Roll and pitch angles could not be detected by our object detector. A template library was developed to detect both out-of-plane rotation angles and it is a comprehensive collection of images of the TOE probe in different roll and pitch rotation angles. These images are created from a digitally reconstructed radiography (DRR) model of the TOE probe. As the TOE probe is sitting inside the oesophagus during the procedure, the probe is not free to move in all directions. Our template library only covers the pitch angle from -45° to 45° and the roll angle from -90° to 90°. The angle interval is 2°. Therefore, the number of images in the library is 4050 images ($4050 = (180/2) \times (90/2)$). The normalized cross correlation is used to compute the similarity between the detected probe image patch and an image from the template library. A real-time performance can be achieved by using a GPU-based implementation.

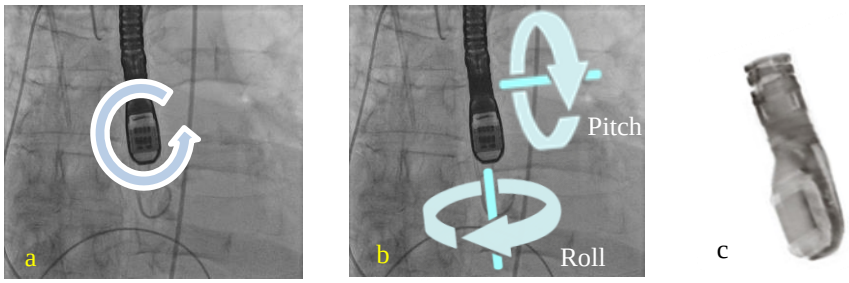

Fig. 7. (a) In-plane rotation. (b) Roll and pitch (out-of-plane). (c) The DRR model

The 3D TOE image volume can be visualized in the 2D X-ray fluoroscopic image by aligning the TOE and X-ray system coordinate systems. The transformation matrix, $T_{TOE\_to\_Xray}$, which transforms from 3D TOE image space to 2D X-ray image space consists of a rigid body transformation matrix T_{rigid} and a projection matrix T_{proj}. It can be computed as $T_{TOE\_to\_Xray} = T_{proj}T_{rigid}$. The projection matrix transforms from 3D X-ray C-arm space to 2D X-ray image space. This can be calculated by using the intrinsic parameters of the X-ray system [21]. T_{rigid} can be decomposed into two matrices ($T_{rigid} = T_{model\_to\_C-arm}T_{TOE\_to\_model}$). Where $T_{TOE\_to\_model}$ transforms from

3D TOE model space to 3D X-ray C-arm space. This matrix is generated by the probe detection algorithm and probe image matching method that positions the 3D TOE model in C-arm space. $T_{TOE\_to\_model}$ relates the position of the 3D TOE images to the position of the 3D TOE model. This is the TOE probe calibration matrix and is calculated pre-procedurally using a specifically designed calibration phantom and the calibration method can be in [22].

3 Results

A total of 10,072 X-ray images (80% train, 10% validation and 10% testing) was used to train and test object detectors. Both validation and testing images are real images, and 4,789 synthetic images were only used in the training dataset. All models using different backbones were implemented in Keras with a Tensorflow (version 2.10) backend and were trained on a GPU farm (NVidia RTX 6000 Ada with 48G memory). The trained models were evaluated on an Intel i7 1.8GHz laptop with a NVidia T550 graphics card to test the inference speed. Table 1 and 2 show the comparison results of our approach (AttWire) with state-of-the-art backbones in single and multiple object detection. AP_{50} and AP_{75} are the average precisions, which are evaluated at IoU=0.5 and IoU=0.75. AP_{50} and AP_{75} in multiple object detection are the mean values of all objects. mAP is the mean value across different IoU thresholds (IoU thresholds from 0.5 to 0.95 with a step size of 0.05).

Table 1. Results for single object detection.

Target Object	Backbone	Parameters	FPS	IoU	AP_{50}	AP_{75}
TOE probe head	VGG16	17.1M	43	0.77±0.21	0.9	0.812
	ResNet-50	36.4M	31	0.83±0.14	0.977	0.794
	AttWire	6.8M	55	0.89±0.05	1.0	0.978
Aortic valve	VGG16	17.1M	52	0.81±0.11	0.972	0.743
	ResNet-50	36.4M	37	0.85±0.08	0.981	0.886
	AttWire	6.8M	59	0.93±0.04	1.0	1.0

Table 2. Results for multiple object detection.

Backbone	Parameters	FPS	IoU (TOE)	IoU (valve)	mAP	AP_{50}	AP_{75}
MobileNet	7.3M	53	0.79±0.09	0.77±0.17	0.546	0.999	0.603
ResNet-50	28.7M	41	0.81±0.07	0.80±0.15	0.618	0.998	0.729
DenseNet121	11.1M	34	0.81±0.07	0.79±0.16	0.584	0.997	0.644
AttWire	3.7M	58	0.88±0.04	0.87±0.11	0.779	1.0	0.922

Overall accuracy of fusing 3D echo with X-ray images was evaluated by using target registration error (TRE). TRE is defined as error distances between corresponding points in both X-ray and echo images. Although real-time synchronized visualization of the live data stream was possible during the clinical procedures, the post-procedure

analysis for this paper required that the recorded X-ray and echo data were synchronized manually, resulting in only approximately synchronized sequences. The manual synchronization was done through visual matching using landmarks such as catheters or artificial valves. Total 20 overlay views are created from 10 X-ray image sequences. Corresponding catheters were manually defined in the echo and X-ray views using spline curves. Equally spaced points along the echo curve were automatically defined as measurement points. The corresponding X-ray point was defined as the closest point on the X-ray curve. An example of these error measurements is given in figure 8(b). Overall, our method achieves a TRE of 2.5±1.2 mm at a speed of 32 FPS.

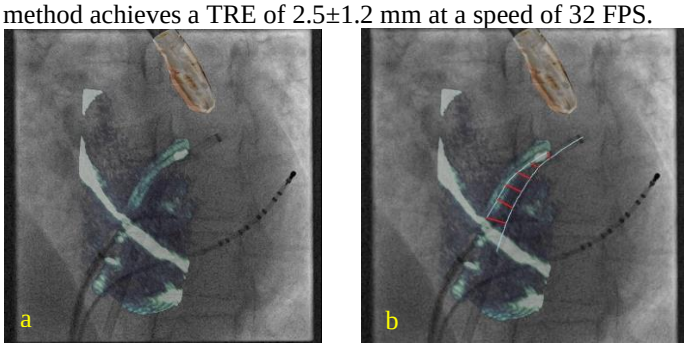

Fig. 8. An example of error measurement. (a) Echo X-ray overlay. (b) Error measurement. Red lines are the shortest distances.

4 Conclusion and Discussions

Clinical applications for minimally invasive heart procedures require highly robust and accurate algorithms for detecting interventional and imaging devices in real-time X-ray fluoroscopic images. In this paper, novel attention mechanisms were designed to guide the CNN model to the areas of wires in X-ray images. The attention-based backbones were implemented in both single and multiple object detection models and they outperform existing state-of-the-art and light-weight backbones by every metric. In addition, our single object detection framework has achieved above 0.97 in AP_{75} and more than 50 FPS. The proposed models for multiple object detection also can perform keypoint detection. With the attention mechanisms we designed, the framework could robustly localize the positions of electrodes on the catheter, and this will enable detecting catheters and devices simultaneously. The detection CNN model facilitates real-time fusion between X-ray fluoroscopy and 3D echo images. It could provide both detection of both TOE probe and other surgical devices.

Acknowledgments. This study was funded by EPSRC, UK (grant number EP/X023826/1).

References

1. Rhode K., et al.: Clinical applications of image fusion for electrophysiology procedures, ISBI 2012, pp. 1435–1438 (2012)

2. Housden R.J., et al.: Evaluation of a Real-Time Hybrid Three-Dimensional Echo and X-Ray Imaging System for Guidance of Cardiac Catheterisation Procedures, MICCAI 2012. LNCS vol. 7511, 25–32 (2012)
3. Ma, Y. et al.: Real-time registration of 3D echo to x-ray fluoroscopy based on cascading classifiers and image registration. Physics in Medicine Biology, **66**(5), (2021)
4. Panayiotou M. et al.: A statistical method for retrospective cardiac and respiratory motion gating of interventional cardiac X-ray images, Medical Physics, **41**(7), 071901–071913, (2014)
5. Mountney, P. et al: Ultrasound and fluoroscopic images fusion by autonomous ultrasound probe detection. MICCAI 2012, LNCS, vol. 7511(15) pp544-551 (2012)
6. Gao, G., et al: Registration of 3D trans-esophageal echocardiography to X-ray fluoros-copy using image-based probe tracking. Medical Image Analysis, 16 (1), 38-49 (2012)
7. Heimann T., Mountney P., John M. and Ionasec R.: Learning without Labeling: Domain Adaptation for Ultrasound Transducer Localization. MICCAI 2013, LNCS, vol. 8151(16) 49-56 (2013)
8. Hatt C. R., Speidel M. A., Raval, A. N.: Hough Forests for Real-Time, Automatic Device Localization in Fluoroscopic Images: Application to TAVR. MICCAI 2015, LNCS, vol. 9349 (18) 307-314 (2015)
9. Miao, S., Wang Z., and Liao R.: A CNN regression approach for real-time 2D/3D registration. IEEE transactions on medical imaging, **35**(5) 1352-1363 (2016)
10. D. Marc et al.: ConTrack: Contextual Transformer for Device Tracking in X-Ray, MICCAI 2023, LNCS, vol. 14228 pp 679–688 (2023).
11. E. Arkin et al.: A survey: object detection methods from CNN to transformer. Multimedia Tools and Applications, **82**, 21353–21383 (2023)
12. Long W., Yu . Yi F., Xinyu L.: A comprehensive survey of oriented object detection in remote sensing images. Expert Systems with Applications **224** (4), 1-16 (2023)
13. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH 2003, pp. 313–318 (2003)
14. Tan, J., Hou, B., Day, T., Simpson, J., Rueckert, D., Kainz, B.: Detecting outliers with poisson image interpolation. MICCAI 2021. LNCS, vol. 12905, pp. 581–591 (2021)
15. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: A Multiscale vessel enhancement filtering. MICCAI 1998. LNCS, vol. 1496, pp 130-137 (1998)
16. Ma Y., Alhrishy M., Narayan S.A., Mountney P., Rhode K.S.: A novel real-time computational framework for detecting catheters and rigid guidewires in cardiac catheterization procedures. Medical Physics, **45**(11), 5066–5079 (2018)
17. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient Attention: Attention with Linear Complexities. IEEE Winter Conference on Applications of Computer Vision (WACV), 3530-3538 (2018)
18. Qian, W., et al.: Learning modulated loss for rotated object detection. In Proceedings of the AAAI conference on artificial intelligence vol. **35**(3) pp. 2458-2466 (2021)
19. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850, (2019).
20. Lin, T., Goyal, P., Girshick, R., He, K. Dollár, P.: Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, **42**(2), 318-327 (2020)
21. Hawkes, et al.: The accurate 3-D reconstruction of the geometric configuration of vascular trees from X-ray recordings. Physics and Engineering of Medical Imaging, NATO ASI Series vol. **119** 250-256 (1987)
22. Housden R.J., et al.: Spatial compounding of trans-esophageal echo volumes using X-ray probe tracking, ISBI 2012, pp. 1092-1095 (2012)