24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

# A Lightweight Recurrent Attention Network for Real-time Guidewire Segmentation and Tracking in Interventional X-ray Fluoroscopy

**Yan-Jie Zhou**[1,3] and **Xiao-Liang Xie**[1] and **Gui-Bin Bian**[1] and **Zeng-Guang Hou** [1,2,3]

**Abstract.** In endovascular surgery and cardiology, interventional therapy is currently the treatment of choice for most patients. Robust guidewire detection in 2D X-ray fluoroscopy can greatly assist physicians in interventional therapy. Nevertheless, this task often comes with the challenge of the extreme foreground-background class imbalance caused by the slenderer guidewire structure compared to other interventional tools. To address this challenge, a novel efficient network architecture, termed Fast Recurrent Attention Network (FRA-Net), is proposed for fully automatic mono-guidewire and dual-guidewire segmentation and tracking. The main contributions of the proposed network are threefold: 1) We propose a novel attention module that improves model sensitivity to guidewire pixels without requiring complicated heuristics. 2) We design a recurrent convolutional layer that ensures better feature representation. 3) Focal Loss is reinforced to better address the problems of extreme class imbalance and misclassified examples. Quantitative and qualitative evaluation on various datasets demonstrates that the proposed network significantly outperforms simpler baselines as well as the best previously-published result for this task, achieving the state-of-the-art performance. To the best of our knowledge, this is the first end-to-end approach capable of real-time segmenting and tracking mono-guidewire and dual-guidewire in 2D X-ray fluoroscopy.

## 1 Introduction

In endovascular surgery and cardiology, endovascular aneurysm repair (EVAR) and percutaneous coronary intervention (PCI) are the primary treatments for abdominal aortic aneurysm (AAA) and coronary heart disease (CHD), respectively. Among them, AAA has been the most common aneurysm, which is usually asymptomatic until it ruptures, with an ensuring mortality 85% to 90% [15]. In addition, CHD has become the largest cause of death worldwide [23]. Thus, it is imperative to improve the success rate of EVAR and PCI.

During endovascular interventional therapy, physicians insert guidewire to vessel and place stent to target stenosis using guidewire guidance. The most critical procedure is to judge the relative positions of the guidewire and the lesion in 2D X-ray fluoroscopy [6]. Hence, the shape and position of the guidewire obtained by robust and accurate segmentation and tracking are of great help for the interventional therapy.

[1] State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[2] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, Email: zhouyanjie2017@ia.ac.cn

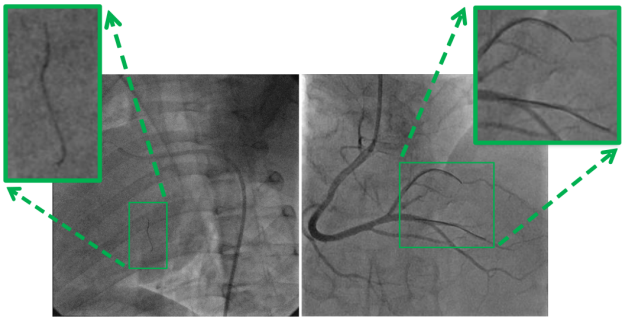

**Figure 1.** The mono-guidewire and dual-guidewire in 2D X-ray fluoroscopy with low SNR. The contrast agents (right) and guidewire-like structures such as vertebrae contours (left) are major challenges for the task of guidewire segmentation and tracking.

**Challenge:** Fully automatic segmentation of guidewire is tough. As shown in Figure 1, the main reasons can be summarized as follows: (1) The X-ray images have low signal-noise-ratio (SNR) and background noise greatly interferes with the segmentation of guidewire. (2) The extreme foreground-background class imbalance is produced by the low ratio of guidewire pixels to the pixels of the background. (3) The edge pixels of the guidewire are misclassified examples due to the guidewire-like structures (e.g. surgical wire, ribs and vertebrae contours) and contrast agents [30].

**Prior Art:** As far as we know, there is a few research focus on guidewire segmentation. Most of researches focus on the guidewire tracking. Traditional tracking methods of interventional instruments are based on spline fitting [4, 10], which are difficult in complex background. And in these methods, the first frame of the fluoroscopy sequence needs to be initialized manually and the instruments between two consecutive frames cannot be significantly deformed.

Then, some learning-based tracking methods were proposed, in which specific hand-crafted features were utilized. In [29], segment-like features (SEGlets) were introduced to overcome large deformations between successive frames. Pauly *et al.* intuitively proposed the local mean orthogonal profiles as features of the original image [24]. The relationship between tracking errors and features was learned by regression methods. Hand-crafted features and intuitively designed tend to have poor generalization and robustness, especially in noisy environments.

In recent years, convolutional neural networks (CNNs) have achieved promising results in this field [3, 31, 33]. Ambrosini *et al.* attempted to segment the whole catheter and guidewire. How-

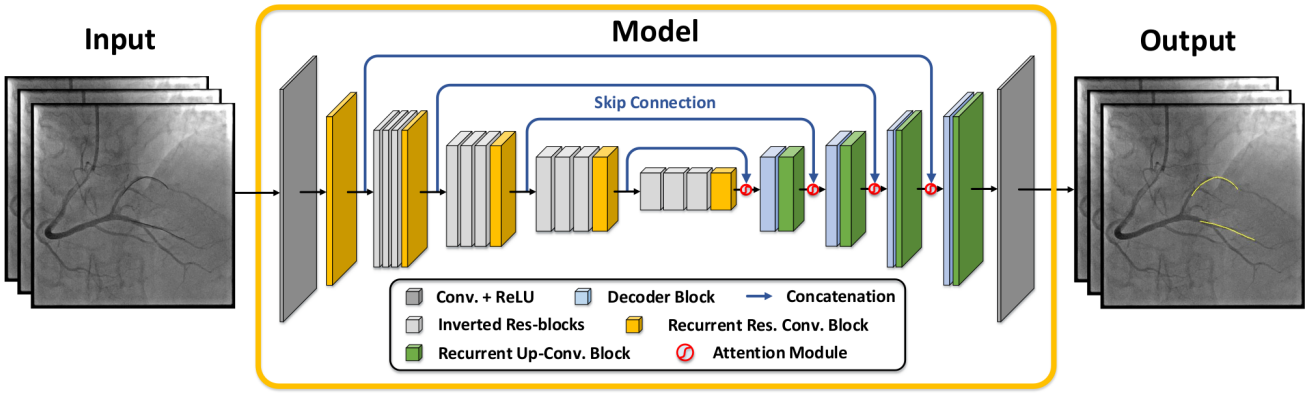


**Figure 2.** The architecture of FRA-Net. The RCLs instead of regular forward convolutional layers are utilized in both encoder and decoder. The AMs filter the features propagated through the skip connections. The mask of the model output is superimposed with the input image.

ever, due to the disparate materials and diameter variance between the catheter and the guidewire, and training in the same network, the reported error by the guidewire is significantly higher than that of the catheter [3]. A recent approach based on cascaded CNN in [31] was designed for segmenting the guidewire, using Faster R-CNN to detect the target region where the guidewire is located firstly, and then using Deep-Lab network to achieve the segmentation of guidewire in the cropped region. Whereas cascaded frameworks lead to excessive and redundant use of computational resources and model parameters, which results in slow processing speed at 0.25 seconds per frame. Although the above methods provide promising initial results, there are miscellaneous shortcomings to be addressed.

**Approach:** To overcome such issues, we propose a novel efficient network (FRA-Net) for fully automatic segmentation and tracking of various guidewires in interventional X-ray fluoroscopy. The proposed network has a novel encoder-decoder architecture, which combines the advantages of attention mechanism, recurrent convolutional neural networks (RCNN) as well as the pre-trained components of MobileNetV2. The improvements between the proposed network with respect to the regular U-Net [26] are threefold. Firstly, the attention module (AM) allows attention coefficients to be more specific to guidewire regions compared to gating based on a global feature vector. This improves the precision and sensitivity precision of the model for dense label prediction with minimal computational cost. Secondly, the feature accumulation method with respect to different time-steps in recurrent convolutional layers (RCLs) guarantees better and stronger feature representation, which helps to extract very low-level features. Last but not least, the pre-trained components of MobileNetV2 in encoder can reduce network parameters and improve model processing speed while ensuring performance. In addition, Focal Loss [19] is reinforced to effectively address the problems of extreme class imbalance and misclassified examples. Experiment results indicate that our proposed approach can significantly improve the segmentation performance, compared to other state-of-the-art approaches. Further analysis also indicates that each individual component of our proposed network contributes to the overall performance improvement.

**The key contributions and novelties can be concluded as follows:**

- As far as we know, this is the first fully automatic approach that achieves real-time segmentation and tracking of various guidewires at the inference rate of 15 FPS.
- The proposed novel network significantly outperforms simpler baselines as well as the best previously-published result, achieving the state-of-the-art performance on various datasets.
- The proposed attention module and designed recurrent convolution layer are effective and can be merged into other encoder-decoder architectures.
- The proposed approach can extend to other applications (e.g. guidewire endpoint localization & visibility enhancement).

## 2 Proposed Network Architecture

The architecture of the proposed FRA-Net is shown in Figure 2. The proposed model has an encoder-decoder architecture. The encoder starts with a convolution on $512 \times 512$ input grayscale images with a kernel of size $7 \times 7$ and a stride of 2. The spatial max-pooling is then performed in the area of $3 \times 3$ with a stride of 2. The latter part of the encoder consists of components of MobileNetV2 pre-trained on ImageNet and recurrent residual convolutional blocks. The key building components in the MobileNetV2 network are inverted residual block [28], which is illustrated in Figure 3. The depth-wise separable convolutions replace the standard convolutional layers in the residual block, thereby reducing considerable computational cost. Compared with the pre-trained backbones ResNet-101 (45M), ResNet-50 (24M) and VGG-16 (34M) utilized in encoder [9], MobileNetV2 (2.4M) greatly reduces the parameters of the network and improves the processing speed while ensuring performance. Recurrent residual convolutional blocks will be introduced in later subsections.

Each decoder block in decoder consists of transposed convolution and batch normalization, followed by the recurrent up-convolutional blocks, aims to recover the resolution of the feature map from $16 \times 16$ to $512 \times 512$. The details of the decoder block are shown in Figure 3. In addition, the decoder blocks are connected to the corresponding encoder blocks through the skip connections, and the AMs in decoder highlight salient features useful for the guidewire which are passed through the skip connections. AM will be introduced in later subsections. After decoder block, we obtain the final segmentation mask by a $1 \times 1$ convolution.

As mentioned in the introduction, the proposed network consists of three major components: (1) The RCLs used for feature accumulation guarantee better and stronger feature representation. (2) The AM highlights salient features which are passed through the skip connections. (3) The augmented Focal Loss is utilized for addressing the problems of extreme foreground-background class imbalance and misclassified examples.

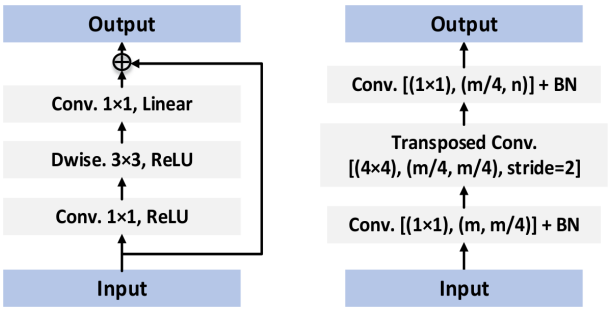


**Figure 3.** The inverted residual block (left) and decoder block (right). The parameters in the convolution block indicate the kernel sizes.

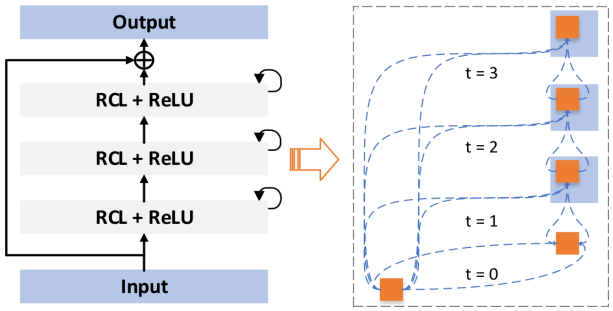


**Figure 4.** The recurrent residual convolutional block (left) and unfolded recurrent convolutional unit for t = 3 (right).

## 2.1 Recurrent Residual Convolutional Block

RCNN and its variants have shown superior performance on different object recognition tasks [18]. According to the improved-residual network [1, 2], the recurrent residual convolution operation can be proved mathematically. The operations of the RCL are performed with respect to the discrete time steps. As shown in Figure 4, we assume the $x_l$ is the input of $l^{th}$ layer of the recurrent residual convolutional block and a pixel is located at $(i, j)$ in an input sample on the $k^{th}$ feature-map. In addition, the output of the network $O^l_{ijk}(t)$ which is at the time step t. The output can be formulated as follows:

$$O^l_{ijk}(t) = (w^c_k)^T * x^{c(i,j)}_l(t) + (w^r_k)^T * x^{r(i,j)}_l(t-1) + b_k \quad (1)$$

where $x^{c(i,j)}_l(t)$ and $x^{r(i,j)}_l(t-1)$ are the inputs to the regular forward convolutional layer and for the $l^{th}$ RCL respectively. The $w^c_k$ and $w^r_k$ are the weights of regular forward convolutional layer and RCL of the $k^{th}$ feature-map respectively, and $b_k$ is the bias. The output $O^l_{ijk}(t)$ is fed into the standard ReLU activation function. And the final outputs $x_{l+1}$ of the recurrent convolutional unit pass through the residual unit. It can be calculated as follows:

$$x_{l+1} = x_l + F(x_l, w_l) = x_l + \max(0, O^l_{ijk}(t)) \quad (2)$$

Here, $x_l$ represents the input samples of the recurrent residual convolutional block. $F(x_l, w_l)$ is the output of the $l^{th}$ layer of the recurrent convolutional block. The output $F(x_l, w_l)$ is utilized for down-sampling and up-sampling layers in the encoding and decoding units of the FRA-Net respectively. The final output $x_{l+1}$ is the input of immediate succeeding sub-sampling or up-sampling layers. In order to extract the features of the lower layers, we further deepen the recurrent residual convolution blocks, each containing three RCL layers. The pictorial representation of the unfolded RCL layers with respect to time-step is shown in Figure 4. Here $t = 3(0 \sim 3)$, refers to the recurrent convolutional operation that includes one single convolution layer followed by three subsequential recurrent convolutional layers. As mentioned in introduction, the feature accumulation based on different time-steps ensures a better and stronger feature representation. Therefore, it helps to extract very low-level features that are essential for guidewire segmentation and tracking.

## 2.2 Attention Module

In order to capture a sufficiently large receptive domain to obtain semantic context information, the feature-map grid is progressively down-sampled in the standard CNN architecture. However, it is still difficult to reduce false-positive predictions for tiny objects with large shape variability. To improve accuracy, the current segmentation frameworks [27, 16] rely on the previously additional object localization models to simplify the task to separate localization and subsequent segmentation steps. Here, we demonstrate that the same goal can be achieved by integrating the proposed AMs into a standard CNN model. This does not require training multiple models and a large number of additional model parameters. Compared with cascaded CNN, AM gradually suppresses the feature responses of irrelevant background regions without the necessity of region of interest (ROI) [21].

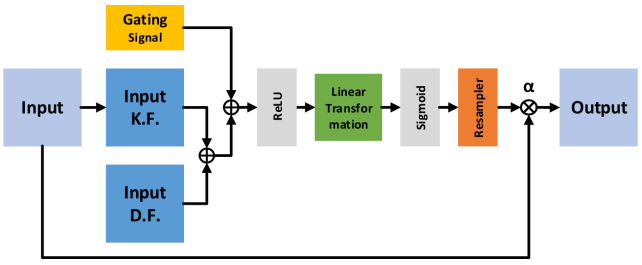


**Figure 5.** Schematic of the proposed attention module (AM). Input features are scaled with attention coefficients computed in AM.

As shown in Figure 5, the attention coefficient $\alpha_i \in [0, 1]$ identifies image salient regions to preserve the activation relevant to the guidewire. The input of AMs can be divided into two parts. The first part is to obtain the key feature map (K.F.) by a series of convolution $3 \times 3$, BN and ReLU. Another part is to directly adjust the feature map (D.F.) to the universal. Then making the summation of two parts to enhance the nonlinearity. The output of AMs is the element-wise multiplication of input feature-maps and attention coefficients: $\widehat{x}^l_{i,c} = x^l_{i,c} \cdot \alpha^l_i$. In the default setting, a single scalar attention value is calculated for each pixel vector $x^l_{i,c}$, where $F_l$ corresponds to the number of feature-maps in layer $l$. The gating vector $g_i$ is utilized for each pixel $i$ to determine the attention regions. Additive attention [5] is utilized to obtain the gating coefficient. Although additive attention is computationally more expensive, experiments have shown that it has higher accuracy than multiplication attention [20]. The formula of AM is represented as follows:

$$\alpha^l_i = \sigma_2(\psi^T(\sigma_1(W^T_x x^l_i + W^T_g g_i + b_g)) + b_\psi) \quad (3)$$

where $\sigma_1$ and $\sigma_2$ correspond to the ReLU activation and sigmoid activation respectively. The $W_x$ and $W_g$ are the weights of linear transformation, and $b_g$ and $b_\psi$ are the bias. In order to reduce the number of trainable parameters and the computational complexity of

AMs, linear transformation is performed without any spatial support ($1 \times 1 \times 1$ convolution), and the input feature-map is down-sampled to the resolution of the gating signal. Grid re-sampling of the attention coefficients is performed by trilinear interpolation to avoid the regional aliasing effect. The proposed AMs are merged into our network to highlight salient features useful for the guidewire, as shown in Figure 2. The information extracted from coarse scale is utilized in gating to disambiguate irrelevant and noisy responses in skip connections, thereby improving the accuracy and sensitivity of the model for foreground pixels prediction.

### 2.3 Augmented Focal Loss

In the task of guidewire segmentation, the slender guidewire structure results in extremely imbalanced ratio, especially the mono-guidewire ($1:1000$). Meanwhile, due to the influence of guidewire-like structures and contrast agents, the edge pixels of the guidewire turn into the misclassified examples. The huge number of easy and background examples tend to overwhelm the training. The two-stage cascade network has been shown effective to alleviate the overwhelming effect of easy samples in many computer vision tasks [25]. Whereas it leads to excessive and redundant use of computational resources and model parameters.

To this end, we reinforce the Focal Loss to better address the problems of extreme class imbalance and misclassified examples. It is the dynamically scaled cross entropy loss. The scaling factor can automatically reduce the weight of easy examples in the training process and quickly focus the model on misclassified examples. The augmented Focal Loss is formulated as follows:

$$Loss = \begin{cases} -\alpha(1-p_i)^{\gamma}\log p_i & y_i = 1 \\ -p_i^{\gamma}\log(1-p_i) & y_i = 0 \end{cases} \quad (4)$$

where $y_i$ is the label of the $i_{th}$ pixel, 1 for guidewire, 0 for background and $p_i$ is the final mask probability of the $i_{th}$ pixel. The weighting factor $\alpha$ and the modulating factor $\gamma$ are tunable within the range of $\alpha, \gamma \geq 0$. Whether it is the foreground class or the background class, the loss contribution from easy examples can be decreased by $\gamma$. Moreover, we have strengthened the role of weighting factor $\alpha$ to increase the weight contribution of the guidewire more efficiently, thus solving the extreme class imbalance. Moreover, $\alpha$ and $\gamma$ are hyper-parameters, and their optimal combination will be verfied in experiments.

## 3 Materials and Implementation

### 3.1 Data Acquisition

There are no public datasets for the guidewire currently. Hence, we establish a new dataset called MDGSeg based on 2D X-ray fluoroscopy, which is provided by Shanghai Huadong Hospital and Peking Union Medical College Hospital. The Innova 3100-IQ digital flat-panel angiography instrument (GE Healthcare) is utilized to acquire clinical sequences of the most representative 30 patients. The details of the MDGSeg dataset are illustrated in Table 1.

In the process of data annotation, a new annotation tool has been designed specifically for labeling the guidewire, which is shown in Figure 6. We first determine the area where the guidewire is located and enlarge the image of the area through the bounding box. Then we mark some points on the guidewire and fit a spline by these points. For each image, it is manually annotated by two technical experts with +5 years experience in medical imaging. When the average error distance of the center line of the guidewire labelled by the two is less than 0.5 pixels, the label is considered valid. After annotation, each 2D X-ray image in sequences is a binary image with size of $512 \times 512$. In binary images, the pixel value of the guidewire is 1, or else 0.

**Table 1.** The details of the MDGSeg dataset. The dataset contains 3239 images of 180 sequences from 30 patients.

| NO. | Guidewire Type | Patient | | Sequence | | Image | |
|---|---|---|---|---|---|---|---|
| | | Train | Test | Train | Test | Train | Test |
| 1 | Mono-guidewire | 10 | 2 | 54 | 18 | 920 | 460 |
| 2 | Dual-guidewire | 8 | 2 | 45 | 15 | 656 | 328 |
| 3 | Stiff Guidewire | 6 | 2 | 36 | 12 | 585 | 290 |
| - | **Total** | 24 | 6 | 135 | 45 | 2161 | 1078 |

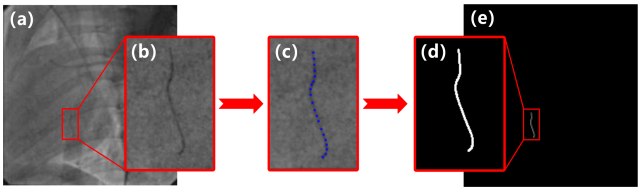


**Figure 6.** Process of data annotation. (a) Determine the coarse bounding box. (b) The guidewire is enlarged. (c) Mark the points along the guidewire. (d) Spline fitting for those points. (e) Obtain the groundtruth mask.

### 3.2 Implementation Details

The proposed network was implemented on PyTorch library (version 0.4.1) with one NVIDIA TITAN Xp (12 GB). In the training phase, two independent sequences in the training set were split as validation set to prevent over-fitting due to insufficient data. To shorten the training cycle, transfer learning was used as the backbone architecture instead of learning from scratch [22]. Stochastic gradient descent (SGD) was used as optimizer with an initial learning rate of 0.001, weight decay of 0.0005 and momentum of 0.9. To find the optimal performance, we reduced the learning rate by the factor of 2 when the validation accuracy was saturated. Moreover, we set the batch size of 32, and 300 epochs was used for each model training.

## 4 Experiments and Results

In this section, several ablation studies are first given to prove the effectiveness of the proposed modules. And then to further evaluate the proposed approach, we apply our network on two different datasets. The first dataset is our own dataset MDGSeg, which consists of patients data of PCI and EVAR. And the other dataset is a publicly available challenge dataset which will be introduced in later subsections.

### 4.1 Error Metrics

We report three main metrics including precision, sensitivity and $F_1$-Score (a higher value is better) to evaluate the segmentation performance. The definition of these metrics are clarified in [29]. The processing time is utilized to evaluate the real-time performance of the proposed approach. To calculating the processing speed, we load the sequence into the proposed model and compute each frame parallelly offline. The total processing time T can be computed after getting the

**Table 2. Quantitative Comparison:** Varying $\alpha$ and $\gamma$ for Augmented Focal Loss.

| Method | | Precision | Sensitivity | $F_1$-Score |
|---|---|---|---|---|
| Focal loss | | $0.914 \pm 0.025$ | $0.927 \pm 0.024$ | $0.918 \pm 0.013$ |
| GHM-C | | $0.939 \pm 0.024$ | $0.909 \pm 0.012$ | $0.923 \pm 0.016$ |
| $\alpha = 50$ | $\gamma = 1.5$ | $0.896 \pm 0.027$ | $0.926 \pm 0.021$ | $0.917 \pm 0.019$ |
| $\alpha = 50$ | $\gamma = 2.0$ | $0.938 \pm 0.018$ | $0.909 \pm 0.023$ | $0.921 \pm 0.011$ |
| $\alpha = 75$ | $\gamma = 2.0$ | $0.914 \pm 0.015$ | $0.932 \pm 0.009$ | $0.924 \pm 0.023$ |
| $\alpha = 75$ | $\gamma = 2.5$ | $0.928 \pm 0.014$ | $0.917 \pm 0.019$ | $0.926 \pm 0.009$ |
| $\alpha = 100$ | $\gamma = 2.5$ | $0.929 \pm 0.022$ | $\mathbf{0.949 \pm 0.017}$ | $\mathbf{0.940 \pm 0.011}$ |
| $\alpha = 100$ | $\gamma = 3.0$ | $\mathbf{0.943 \pm 0.009}$ | $0.911 \pm 0.016$ | $0.927 \pm 0.027$ |
| $\alpha = 125$ | $\gamma = 3.0$ | $0.937 \pm 0.012$ | $0.887 \pm 0.021$ | $0.912 \pm 0.019$ |

results of all the frames (N). Therefore, we obtain the final processing speed $N/T$ FPS and processing time $1000 \times T/N$ ms.

To show results more clearly, the mean of distances from ground truth to segmentation results, is also evaluated and called as guidewire special precision (GSP). The distances in this paper are all in image pixels (px), which is formulated as follows:

$$d_{spe}(D_G, D_s, s) = \min_s(\|D_G(s) - D_s(s)\|) \tag{5}$$

where $D_S(s)$ is segmentation results and $D_G(s)$ is the ground truth.

## 4.2 Analysis of Augmented Focal Loss

To evaluate the effectiveness of the augmented Focal Loss on our approach, we apply our model on two different loss function. The first is the regular Focal Loss [19], which yields the optimal solution at $\alpha = 0.25$ and $\gamma = 2$. The second is the gradient harmonizing mechanism (GHM), which is currently the state-of-the-art method of solving class imbalance. The GHM is to make a balanced cumulative contribution of samples of various difficulty types by weighting the gradients, which are generated by different samples and changing their contribution [17]. Therefore, we utilize the regular Focal Loss and GHM classification loss (GHM-C) as the experimental baselines.

The augmented Focal Loss has two hyperparameters $\alpha$ and $\gamma$. The experimental results indicate when weighing factor $\alpha$ and modulating factor $\gamma$ are 100 and 2.5 respectively (as shown in Table 2), the model has the optimal performance. The mean $F_1$-score, precision and sensitivity are respectively 0.940, 0.929 and 0.949 wherein the $F_1$-Score is improved by 2.41% and 1.84% over baseline respectively.

As shown in Figure 8, some of the guidewire pixels in the segmentation results of the regular Focal Loss are missing due to the influence of extreme class imbalance and guidewire-like structures (e.g. vertebrae). The influence of contrast agents results in the misclassification of background pixels and the missing of guidewire pixels in the segmentation results of GHM-C. In contrast, the segmentation results of the augmented Focal Loss are more smooth and accurate.

## 4.3 Ablation Experiments

In this section, we conduct extensive ablation studies to validate the effectiveness of the proposed AM and RCL. We follow the previous protocols. Augmented Focal Loss and the pre-trained MobileNetV2 are utilized as the base loss function and the backbone network, respectively. Then we verify the efficiency of the pre-trained MobileNetV2 for improving processing speed.

**Table 3. Quantitative comparison:** Ablation experiments.

| Method | $F_1$-Score | Time (ms) |
|---|---|---|
| FRA-Net | $0.940 \pm 0.011$ | $66.5 \pm 1.6$ |
| without AM | $0.908 \pm 0.023$ | $60.6 \pm 2.3$ |
| without RCL | $0.910 \pm 0.015$ | $58.4 \pm 1.9$ |
| ResNet-34 | $0.929 \pm 0.018$ | $133.5 \pm 4.2$ |
| ResNet-50 | $0.940 \pm 0.021$ | $139.6 \pm 2.7$ |
| ResNet-101 | $0.944 \pm 0.013$ | $172.2 \pm 3.4$ |
| VGG-11 | $0.925 \pm 0.017$ | $129.6 \pm 1.9$ |
| VGG-16 | $0.934 \pm 0.021$ | $155.3 \pm 3.8$ |

To evaluate the contribution of the AM and RCL on our approach, we remove the AM and recurrent convolution block from the original model separately and train them. To verify the improvement in processing speed brought by the pre-trained MobileNetV2, we replace the backbone of the original network with wildly-used backbones ResNet and VGGNet. As shown in Table 3, it clearly demonstrates the improvement in accuracy brought by the proposed AM and RCL, and the promotion in processing speed brought by the pre-trained MobileNetV2. The MobileNetV2 as the backbone of the network processes one image much faster than other heavy backbones.

To further verify the robustness of our proposed network, the class activation map (CAM) is utilized to visualize the discriminative region of the network for the test frames. The global average pooling outputs the spatial average of the feature map. The predicted class scores are mapped back to the previous convolutional layer to generate the CAMs [32]. The CAM highlights the discriminative regions of the specific class. As shown in Figure 7, most of the discriminative regions are concentrated around the guidewire, indicating that the network has learned robust discriminative ability.

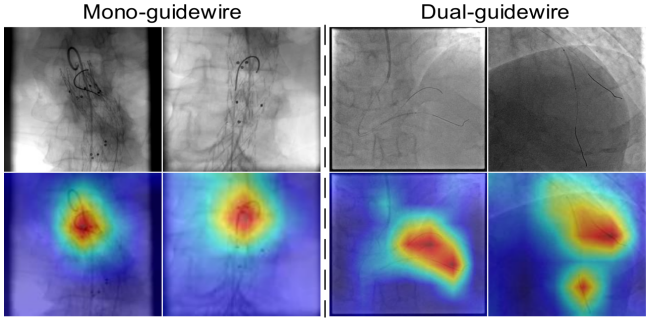


**Figure 7.** The class activation maps (CAMs) of typical test frames. The maps highlight the discriminative regions of the guidewire.

**Table 4. Quantitative comparison:** The segmentation metrics of different approaches on MDGSeg dataset.

| Method | Fully Auto. | Mean ± Std (px) | Med (px) | $F_1$-Score | Time (ms) |
|---|---|---|---|---|---|
| U-Net [26] | √ | 1.624 ± 1.439 | 1.598 | 0.909 | 102.4 |
| TernausNet [14] | √ | 0.873 ± 0.647 | 0.899 | 0.926 | 130.5 |
| LinkNet [8] | √ | 1.106 ± 0.755 | 1.182 | 0.914 | 112.8 |
| ITT [10] | × | 4.726 ± 4.011 | 2.996 | 0.847 | 59.7 |
| DT [11] | × | 5.378 ± 4.824 | 2.545 | 0.864 | 111.1 |
| GE [13] | × | 4.961 ± 4.118 | 3.232 | 0.902 | \ |
| SEG [29] | √ | 2.245 ± 2.196 | 1.359 | 0.923 | 500 |
| **Ours** | √ | **0.568 ± 0.436** | **0.342** | **0.940** | **66.5** |

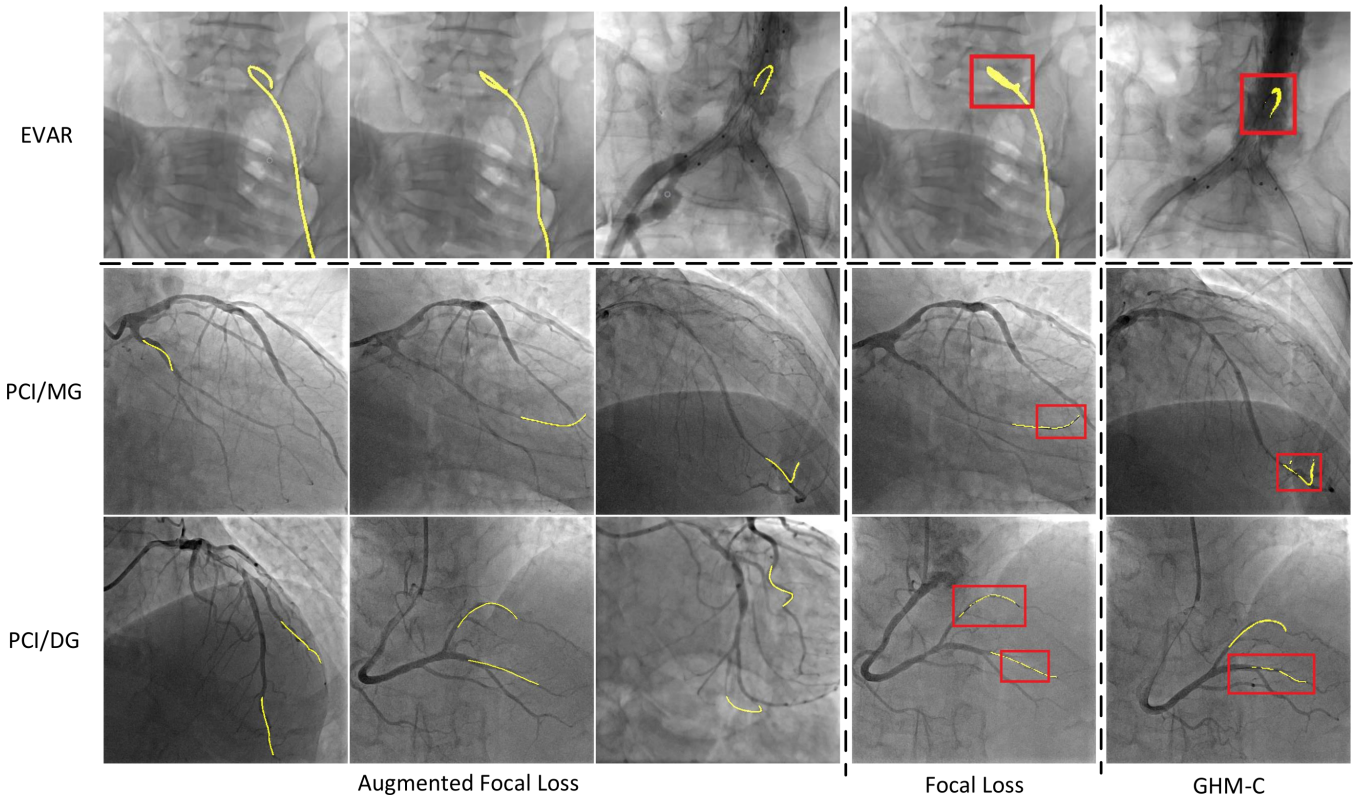


**Figure 8. Qualitative comparison:** The various guidewire segmentation results of typical test frames by different approaches. Stiff guidewire in EVAR (top), mono-guidewire (MG) in PCI (mid) and dual-guidewire (DG) in PCI (bottom).

## 4.4 Comparison with State-of-the-Art Methods

To demonstrate the advantage of our proposed approach, we compare it with the widely-used networks (U-Net, TernausNet and LinkNet) and four other previously proposed approaches on the same dataset (MDGSeg). It is worth noting that we implement other approaches with best parameters. As shown in Table 4, it clearly demonstrates that our approach achieves better accuracy than other state-of-the-art approaches in terms of mean and median GSP, $F_1$-Score and processing time. As can be seen in Figure 8, the proposed approach is robust to all kinds of guidewires in different interventional treatments, and the segmentation results are accurate without any post-processing.

To further analyze the generalization capability of our proposed approach, we compare it with well-known medical segmentation networks [26, 8, 14, 21] on six different patient sequences in the test set. Among them, TernausNet is an improved U-Net with the VGG-11 encoder pre-trained on ImageNet [14]. LinkNet is the network that can obtain accurate instance-level predictions [8]. And the patient sequences in the test set consist of the samples of mono-guidewire, dual-guidewire and stiff guidewire. Quantitative comparison of different patient sequences are shown in Figure 9. As shown in Figure 9 (a), the proposed approach has good generalization performance and has achieved excellent $F_1$-Score in different patients and different guidewire sequences. As can we seen from the Figure 9 (b), our approach is far superior to other medical segmentation networks and achieves the state-of-the-art segmentation performance.

In addition, Heidbuchel *et al*. mentioned that to reduce the radiation intake of physicians, the C-arm system operates at a low frame rate ($6 \sim 12$ FPS) [12]. The average processing time per image of our proposed network is about 66.5 ms (15 FPS), which enables accurate real-time segmentation and tracking.

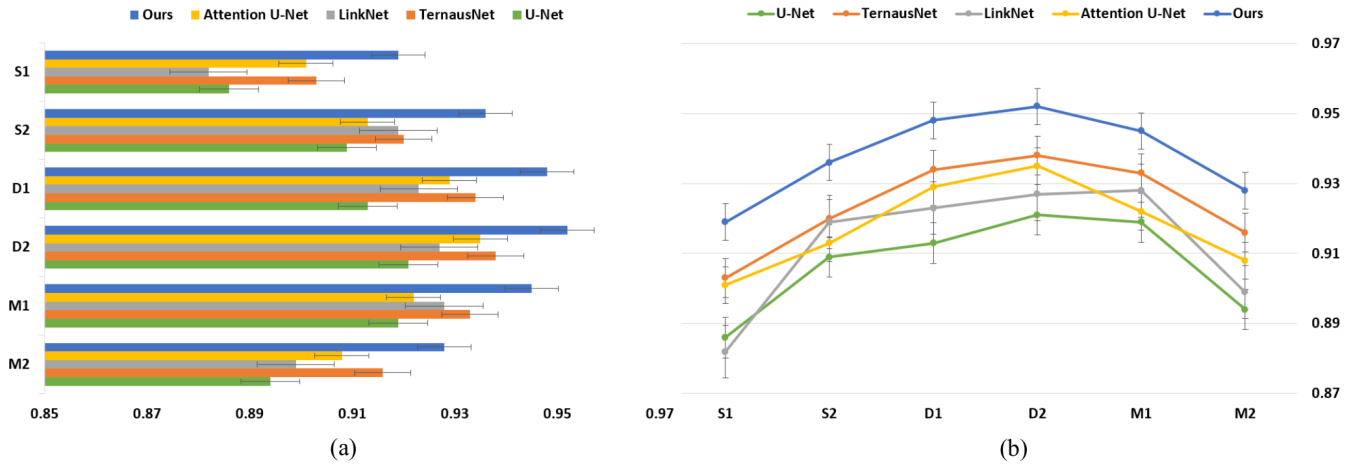


**Figure 9. Quantitative comparison:** The $F_1$-Score of different approaches for six patient sequences on the test set. M1, M2: Patients treated with mono-guidewire in PCI. D1, D2: Patients treated with dual-guidewire in PCI. S1, S2: Patients treated with stiff guidewire in EVAR.

**Table 5. Quantitative comparison:** The $F_1$-Score of different approaches on NLM Chest X-ray Database. This five sequences on the test set contain the cases with manifestation of tuberculosis and the normal cases.

| Method | Sequence 1 | Sequence 2 | Sequence 3 | Sequence 4 | Sequence 5 | Mean $F_1$-Score |
|---|---|---|---|---|---|---|
| U-Net [26] | 0.899 | 0.907 | 0.864 | 0.882 | 0.887 | 0.888 |
| LinkNet [8] | 0.902 | 0.915 | 0.872 | 0.893 | 0.896 | 0.896 |
| TernausNet [14] | 0.910 | 0.922 | 0.898 | 0.909 | 0.912 | 0.910 |
| **Ours** | 0.953 | 0.966 | 0.932 | 0.941 | 0.944 | **0.947** |

## 4.5 Validation on NLM Chest X-ray Database

We further validate our proposed approach on Chest X-ray dataset, which is the standard digital image database for tuberculosis [7]. The chest X-rays are from out-patient clinics, and were captured as part of the daily routine using Philips DR Digital Diagnose systems. This dataset consists of 336 cases with manifestation of tuberculosis and 326 normal cases. We select five sequences in the database containing tuberculosis and normal cases as the test set.

We conduct the experiment in leave-one-out manner. We visualize typical test samples in Figure 10 to make a qualitative comparison. The proposed approach can capture better contour which is usually considered as hard regions compared with the U-Net. As can be seen in Table 5, the quantitative comparison (the proposed network can improve the segmentation performance by about 3.7 % in terms of mean $F_1$-Score) also indicates the success of our proposed approach.

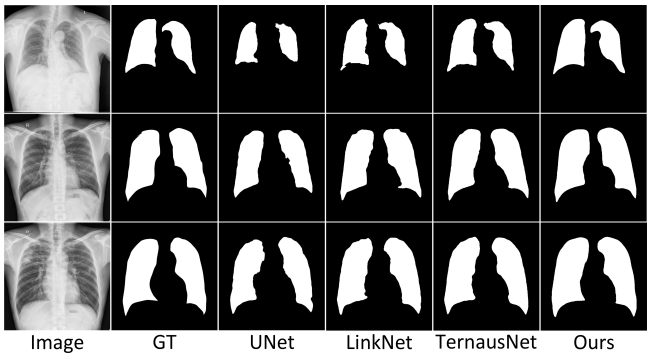


**Figure 10. Qualitative comparison:** The segmentation results of different approaches on NLM Chest X-ray Database.

## 5 Conclusions and Future Work

In this paper, we proposed a novel network, FRA-Net, to address the challenging task of real-time segmentation and tracking of various guidewires in interventional X-ray fluoroscopy. Quantitative and qualitative evaluation on MDGSeg dataset and NLM Chest X-ray database demonstrates that our approach achieve significant improvement in terms of both accuracy and robustness. Extensive ablation experiments prove the effectiveness of our proposed modules (AM and RCL) and augmented Focal Loss. By integrating these components into the network, our proposed model completely address extreme class imbalance and misclassified examples, achieving the state-of-the-art performance. Moreover, the inference rate of our model is approximately 15 FPS, which is promising for real-time assisting physicians in completing endovascular interventional therapy.

As future work, we would like to concentrate on applying the approach proposed in this paper to the vascular interventional surgery robot to achieve the computer-assisted treatment.

## ACKNOWLEDGEMENTS

The authors would like to thank B. Liu, Z. Lai (Peking Union Medical College Hospital) and X. Qu, S. Guan (Huadong Hospital Affiliated to Fudan University) for their help on the experiments. This work is partially supported by the National Key Research and Development Plan of China (Grant 2019YFB1311700), the National Natural Science Foundation of China (Grants 61533016, U1613210, 61421004).

## REFERENCES

[1] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, and Tarek M Taha, 'Inception recurrent convolutional neural network for object recognition', *arXiv preprint arXiv:1704.07709*, (2017).

[2] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari, 'Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation', *arXiv preprint arXiv:1802.06955*, (2018).

[3] Pierre Ambrosini, Daniel Ruijters, Wiro J Niessen, Adriaan Moelker, and Theo van Walsum, 'Fully automatic and real-time catheter segmentation in X-ray fluoroscopy', in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 577–585. Springer, Cham, (2017).

[4] Shirley AM Baert, Max A Viergever, and Wiro J Niessen, 'Guide-wire tracking during endovascular interventions', *IEEE Transactions on Medical Imaging*, **22**(8), 965–972, (2003).

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, 'Neural machine translation by jointly learning to align and translate', *arXiv preprint arXiv:1409.0473*, (2014).

[6] Dominique B Buck, Joost A Van Herwaarden, Marc L Schermerhorn, and Frans L Moll, 'Endovascular treatment of abdominal aortic aneurysms', *Nature Reviews Cardiology*, **11**(2), 112, (2014).

[7] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald, 'Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration', *IEEE Transactions on Medical Imaging*, **33**(2), 577–590, (2013).

[8] Abhishek Chaurasia and Eugenio Culurciello, 'Linknet: Exploiting encoder representations for efficient semantic segmentation', in *Proceedings of the 2017 IEEE Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–4. IEEE, (2017).

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, (2016).

[10] Hauke Heibel, Ben Glocker, Martin Groher, Marcus Pfister, and Nassir Navab, 'Interventional tool tracking using discrete optimization', *IEEE Transactions on Medical Imaging*, **32**(3), 544–555, (2013).

[11] Tim Hauke Heibela, Ben Glockera, Martin Grohera, Nikos Paragios, and Nikos Komodakis, 'Discrete tracking of parametrized curves', in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1754–1761. IEEE, (2009).

[12] Hein Heidbuchel, Fred H M Wittkampf, Eliseo Vano, Sabine Ernst, and Richard Schilling, 'Practical ways to reduce radiation dose for patients and staff during device implantations and electrophysiological procedures', *Europace*, **16**(7), 946–964, (2014).

[13] Nicolas Honnorat, Régis Vaillant, and Nikos Paragios, 'Guide-wire extraction through perceptual organization of local segments in fluoroscopic images', in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 440–448. Springer, Cham, (2010).

[14] Vladimir Iglovikov and Alexey Shvets, 'Ternausnet: U-Net with VGG11 encoder pre-trained on ImageNet for image segmentation', *arXiv preprint arXiv:1801.05746*, (2018).

[15] K Craig Kent, 'Abdominal aortic aneurysms', *New England Journal of Medicine*, **371**(22), 2101–2108, (2014).

[16] Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi, 'Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers', *Medical Image Analysis*, **51**, 21–45, (2019).

[17] Buyu Li, Yu Liu, and Xiaogang Wang, 'Gradient harmonized single-stage detector', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8577–8584, (2019).

[18] Ming Liang and Xiaolin Hu, 'Recurrent convolutional neural network for object recognition', in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3367–3375. IEEE, (2015).

[19] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, 'Focal loss for dense object detection', in *International Conference on Computer Vision (ICCV)*, pp. 2999–3007. IEEE, (2017).

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning, 'Effective approaches to attention-based neural machine translation', *arXiv preprint arXiv:1508.04025*, (2015).

[21] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, and Matthew Lee, 'Attention U-Net: learning where to look for the pancreas', *arXiv preprint arXiv:1804.03999*, (2018).

[22] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic, 'Learning and transferring mid-level image representations using convolutional neural networks', in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1717–1724. IEEE, (2014).

[23] World Health Organization et al., 'Global status report on noncommunicable diseases 2014', Technical report, World Health Organization, (2014).

[24] Olivier Pauly, Hauke Heibel, and Nassir Navab, 'A machine learning approach for deformable guide-wire tracking in fluoroscopic sequences', in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 343–350. Springer, (2010).

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, 'Faster r-cnn: Towards real-time object detection with region proposal networks', in *Advances in Neural Information Processing Systems*, pp. 91–99, (2015).

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, 'U-Net: Convolutional networks for biomedical image segmentation', in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. Springer, Cham, (2015).

[27] Holger R Roth, Le Lu, Nathan Lay, Adam P Harrison, Amal Farag, Andrew Sohn, and Ronald M Summers, 'Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation', *Medical Image Analysis*, **45**, 94–107, (2018).

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, 'Mobilenetv2: Inverted residuals and linear bottlenecks', in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520. IEEE, (2018).

[29] Alessandro Vandini, Ben Glocker, Mohamad Hamady, and Guang-Zhong Yang, 'Robust guidewire tracking under large deformations combining segment-like features (SEGlets)', *Medical Image Analysis*, **38**, 150–164, (2017).

[30] Wen Wu, Terrence Chen, Peng Wang, Shaohua Kevin Zhou, Dorin Comaniciu, Adrian Barbu, and Norbert Strobel, 'Learning-based hypothesis fusion for robust catheter tracking in 2D X-ray fluoroscopy', in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1097–1104. IEEE, (2011).

[31] Yu-Dong Wu, Xiao-Liang Xie, Gui-Bin Bian, Zeng-Guang Hou, Xiao-Ran Cheng, Sheng Chen, Shi-Qi Liu, and Qiao-Li Wang, 'Automatic guidewire tip segmentation in 2D X-ray fluoroscopy using convolution neural networks', in *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, (2018).

[32] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, 'Learning deep features for discriminative localization', in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929. IEEE, (2016).

[33] Yan-Jie Zhou, Xiao-Liang Xie, Gui-Bin Bian, Zeng-Guang Hou, Yu-Dong Wu, Shi-Qi Liu, Xiao-Hu Zhou, and Jia-Xing Wang, 'Fully automatic dual-guidewire segmentation for coronary bifurcation lesion', in *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, (2019).