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

Yan-Jie Zhou1,3 and Xiao-Liang Xie1 and Gui-Bin Bian1 and Zeng-Guang Hou1,2,3

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.

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.
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 $512\times512$ input grayscale images with a kernel of $7\times7$ and a stride of 2. The spatial max-pooling is then performed in the area of $3\times3$ 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 $16\times16$ to $512\times512$. 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 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.
2.2 Attention Module

for guidewire segmentation and tracking. Therefore, it helps to extract very low-level features that are essential for accurate localization. Different time-steps ensures a better and stronger feature representation. As mentioned in introduction, the feature accumulation based on different residual convolutional blocks, each containing three RCL layers. To extract the features of the lower layers, we further deepen the recurrent convolutional unit pass through the residual unit. It can be calculated as follows:

\[ x_{t+1} = x_t + F(x_t, w_t) = x_t + \max(0, O_{ijk}(t)) \]

Here, \( x_t \) represents the input samples of the recurrent residual convolutional block. \( F(x_t, w_t) \) is the output of the \( l^{th} \) layer of the recurrent convolutional block. The output \( F(x_t, w_t) \) is utilized for downstreaming and sub-sampling layers in the encoding and decoding units of the FRA-Net respectively. The final output \( x_{t+1} \) is the input of the recurrent convolutional block. The output \( O_{ijk}(t) \) is fed into the standard ReLU activation function. And the final outputs are the weights of regular forward convolutional layer and regular backward convolutional layer and for the recurrent residual convolutional block. The output is fed into the standard ReLU activation function. And the final outputs are the weights of regular forward convolutional layer and regular backward convolutional layer. The output can be formulated as follows:

\[ O_{ijk}(t) = (w_k^T)^2 x_i^{(k;i)}(t) + (w_k^T)^2 x_i^{(k;i)}(t - 1) + b_k \]

where \( x_i^{(k;i)}(t) \) and \( x_i^{(k;i)}(t - 1) \) are the inputs to the regular forward convolutional layer and for the \( l^{th} \) RCL respectively. The \( w_k^T \) and \( w_k^T \) 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 \( O_{ijk}(t) \) is fed into the standard ReLU activation function. And the final outputs are the weights of regular forward convolutional layer and regular backward convolutional layer. The output can be formulated as follows:

\[ x_{t+1} = x_t + F(x_t, w_t) = x_t + \max(0, O_{ijk}(t)) \]

Here, \( x_t \) represents the input samples of the recurrent residual convolutional block. \( F(x_t, w_t) \) is the output of the \( l^{th} \) layer of the recurrent convolutional block. The output \( F(x_t, w_t) \) is utilized for downstreaming and sub-sampling layers in the encoding and decoding units of the FRA-Net respectively. The final output \( x_{t+1} \) is the input of the recurrent convolutional block. The output \( O_{ijk}(t) \) is fed into the standard ReLU activation function. And the final outputs are the weights of regular forward convolutional layer and regular backward convolutional layer. The output can be formulated as follows:

\[ O_{ijk}(t) = (w_k^T)^2 x_i^{(k;i)}(t) + (w_k^T)^2 x_i^{(k;i)}(t - 1) + b_k \]

where \( x_i^{(k;i)}(t) \) and \( x_i^{(k;i)}(t - 1) \) are the inputs to the regular forward convolutional layer and for the \( l^{th} \) RCL respectively. The \( w_k^T \) and \( w_k^T \) 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 \( O_{ijk}(t) \) is fed into the standard ReLU activation function. And the final outputs are the weights of regular forward convolutional layer and regular backward convolutional layer. The output can be formulated as follows:

\[ x_{t+1} = x_t + F(x_t, w_t) = x_t + \max(0, O_{ijk}(t)) \]

Here, \( x_t \) represents the input samples of the recurrent residual convolutional block. \( F(x_t, w_t) \) is the output of the \( l^{th} \) layer of the recurrent convolutional block. The output \( F(x_t, w_t) \) is utilized for downstreaming and sub-sampling layers in the encoding and decoding units of the FRA-Net respectively. The final output \( x_{t+1} \) is the input of the recurrent convolutional block. The output \( O_{ijk}(t) \) is fed into the standard ReLU activation function. And the final outputs are the weights of regular forward convolutional layer and regular backward convolutional layer. The output can be formulated as follows:

\[ x_{t+1} = x_t + F(x_t, w_t) = x_t + \max(0, O_{ijk}(t)) \]

where \( x_t \) represents the input samples of the recurrent residual convolutional block. \( F(x_t, w_t) \) is the output of the \( l^{th} \) layer of the recurrent convolutional block. The output \( F(x_t, w_t) \) is utilized for downstreaming and sub-sampling layers in the encoding and decoding units of the FRA-Net respectively. The final output \( x_{t+1} \) is the input of the recurrent convolutional block. The output \( O_{ijk}(t) \) is fed into the standard ReLU activation function. And the final outputs are the weights of regular forward convolutional layer and regular backward convolutional layer. The output can be formulated as follows:

\[ x_{t+1} = x_t + F(x_t, w_t) = x_t + \max(0, O_{ijk}(t)) \]

where \( x_t \) represents the input samples of the recurrent residual convolutional block. \( F(x_t, w_t) \) is the output of the \( l^{th} \) layer of the recurrent convolutional block. The output \( F(x_t, w_t) \) is utilized for downstreaming and sub-sampling layers in the encoding and decoding units of the FRA-Net respectively. The final output \( x_{t+1} \) is the input of the recurrent convolutional block. The output \( O_{ijk}(t) \) is fed into the standard ReLU activation function. And the final outputs are the weights of regular forward convolutional layer and regular backward convolutional layer. The output can be formulated as follows:

\[ x_{t+1} = x_t + F(x_t, w_t) = x_t + \max(0, O_{ijk}(t)) \]

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

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 \( x_t \) 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 \( O_{ijk}(t) \) which is at the time step \( t \). The output can be formulated as follows:

\[ O_{ijk}(t) = (w_k^T)^2 x_i^{(k;i)}(t) + (w_k^T)^2 x_i^{(k;i)}(t - 1) + b_k \]

In the default setting, a single scalar attention value \( \alpha_i \) is utilized for the feature accumulation based on different residual convolutional 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 subsequent 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].

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:

\[ x_{t+1} = x_t \odot \alpha_i \]

In the default setting, a single scalar attention value is calculated for each pixel vector \( x_i \). 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:

\[ \alpha_i = \sigma_2(\psi^T(\sigma_1(W_{x}^T x_i + W_{y}^T g_i + b_y)) + b_y) \]

where \( \sigma_1 \) and \( \sigma_2 \) correspond to the ReLU activation and sigmoid activation respectively. The \( W_x \) and \( W_y \) are the weights of linear transformation, and \( b_y \) and \( b_y \) are the bias. In order to reduce the number of trainable parameters and the computational complexity of

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

Figure 5. Schematic of the proposed attention module (AM). Input features are scaled with attention coefficients computed in AM.
AMs, linear transformation is performed without any spatial support (1 × 1 × 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:

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

(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 verified 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

<table>
<thead>
<tr>
<th>NO.</th>
<th>Guidewire Type</th>
<th>Patient</th>
<th>Sequence</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mono-guidewire</td>
<td>10</td>
<td>2</td>
<td>54</td>
</tr>
<tr>
<td>2</td>
<td>Dual-guidewire</td>
<td>8</td>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>Stiff Guidewire</td>
<td>6</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>24</td>
<td>6</td>
<td>135</td>
</tr>
</tbody>
</table>

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 patient 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>$F_1$-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focal loss</td>
<td>0.914 ± 0.025</td>
<td>0.927 ± 0.024</td>
<td>0.915 ± 0.013</td>
</tr>
<tr>
<td>GHM-C</td>
<td>0.939 ± 0.024</td>
<td>0.909 ± 0.012</td>
<td>0.923 ± 0.016</td>
</tr>
<tr>
<td>$\alpha = 50$</td>
<td>$\gamma = 1.5$</td>
<td>0.896 ± 0.027</td>
<td>0.926 ± 0.021</td>
</tr>
<tr>
<td>$\alpha = 75$</td>
<td>$\gamma = 2.0$</td>
<td>0.938 ± 0.018</td>
<td>0.909 ± 0.023</td>
</tr>
<tr>
<td>$\alpha = 100$</td>
<td>$\gamma = 2.5$</td>
<td>0.914 ± 0.015</td>
<td>0.932 ± 0.009</td>
</tr>
<tr>
<td>$\gamma = 3.0$</td>
<td>0.928 ± 0.014</td>
<td>0.917 ± 0.019</td>
<td>0.926 ± 0.009</td>
</tr>
</tbody>
</table>

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

Table 3. Quantitative comparison: Ablation experiments.

<table>
<thead>
<tr>
<th>Method</th>
<th>$F_1$-Score</th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRA-Net</td>
<td>0.940 ± 0.011</td>
<td>66.5 ± 1.6</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>0.929 ± 0.018</td>
<td>133.5 ± 4.2</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>0.940 ± 0.021</td>
<td>139.6 ± 2.7</td>
</tr>
<tr>
<td>ResNet-101</td>
<td>0.944 ± 0.013</td>
<td>172.2 ± 3.4</td>
</tr>
<tr>
<td>VGG-11</td>
<td>0.925 ± 0.017</td>
<td>129.6 ± 1.9</td>
</tr>
<tr>
<td>VGG-16</td>
<td>0.934 ± 0.021</td>
<td>155.3 ± 3.8</td>
</tr>
</tbody>
</table>

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:

$$d_{spx}(D_G, D_s, s) = \min_s ||D_G(s) - D_s(s)||$$  \hspace{1cm} (5)

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.

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.
### 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.

---

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

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

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

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.

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


