



Lightweight Double Attention-Fused Networks for Intraoperative Stent Segmentation

Yan-Jie Zhou^{1,3(✉)}, Xiao-Liang Xie^{1,3}, Zeng-Guang Hou^{1,2,3}, Xiao-Hu Zhou¹,
Gui-Bin Bian¹, and Shi-Qi Liu¹

¹ State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
zhouyanjie2017@ia.ac.cn

² CAS Center for Excellence in Brain Science and Intelligence Technology,
Beijing 100190, China

³ School of Artificial Intelligence, University of Chinese Academy of Sciences,
Beijing 100049, China

Abstract. In endovascular interventional therapy, the fusion of preoperative data with intraoperative X-ray fluoroscopy has demonstrated the potential to reduce radiation dose, contrast agent and processing time. Real-time intraoperative stent segmentation is an important pre-requisite for accurate fusion. Nevertheless, this task often comes with the challenge of the thin stent wires with low contrast in noisy X-ray fluoroscopy. In this paper, a novel and efficient network, termed Lightweight Double Attention-fused Network (LDA-Net), is proposed for end-to-end stent segmentation in intraoperative X-ray fluoroscopy. The proposed LDA-Net consists of three major components, namely feature attention module, relevance attention module and pre-trained MobileNetV2 encoder. Besides, a hybrid loss function of both reinforced focal loss and dice loss is designed to better address the issues of class imbalance and misclassified examples. Quantitative and qualitative evaluations on 175 intraoperative X-ray sequences demonstrate that the proposed LDA-Net significantly outperforms simpler baselines as well as the best previously-published result for this task, achieving the state-of-the-art performance.

Keywords: Stent segmentation · Intraoperative X-ray fluoroscopy · Convolution neural networks

1 Introduction

Abdominal aortic aneurysm (AAA) has been the most common aneurysm, which is usually asymptomatic until it ruptures, with an ensuring mortality 85% to 90% [1]. Clinical evidence-based research shows a lower perioperative morbidity and mortality, and similar long-term survival, for endovascular aortic repair (EVAR) compared with open repair of suitable AAAs. Meanwhile, recent technological

advances in EVAR make it the treatment of choice for most AAA patients [2]. However, due to the complexity of the EVAR, long-term radiation and large doses of contrast agents are usually required during the intervention, which will lead to common complications for patients, such as renal insufficiency. It is therefore of special concern to reduce the procedure time of EVAR.

Fusion of preoperative data with intraoperative X-ray fluoroscopy to guide the intervention has been proved to reduce contrast agent and radiation dose [3]. The preoperative data is obtained by 3D computed tomography (CT). However, the fusion may become inaccurate due to patient motion and deformation of the vessels caused by interventional instruments. To avoid repeated use of contrast agent, comparing the stent segmentation with preoperative information can assess and monitor the quality of the current fusion throughout the intervention [4]. Hence, real-time and accurate intraoperative stent segmentation is imperative. Nevertheless, fully automatic stent segmentation is not straightforward for the following reasons: (1) The morphological variation of stents in different interventions affects visual features such as shape and size. (2) The low ratio of stent wire pixels to background pixels results in class imbalance. (3) The contrast agent, artifacts from the spine and wire-like structures such as guidewire interfere with the classification accuracy of edge pixels of stents.

Although the guidewire segmentation [5] and catheter segmentation [6] in X-ray fluoroscopy have received widespread interest, less attention has been spent on stent segmentation. Previously, Demirci *et al.* [7] proposed a model-based method that relies on Hessian-based filtering for preprocessing. Although this method can directly recover the shape of the stent in 3D, it needs to define the model of the stent in advance and is limited to a certain stent shape. Recently, deep learning has achieved promising results in medical image segmentation [8, 9] and provide a data-driven approach to address stent segmentation. Breininger *et al.* [4] presented a fully convolutional network with a contraction and expansion path to segment aortic stents. However, due to the utilization of residual units as its backbone, the real-time requirements were not met.

To address above-mentioned concerns, the Lightweight Double Attention-fused Networks (LDA-Net) is proposed for real-time stent segmentation in intraoperative X-ray fluoroscopy. Firstly, aggregation for multi-scale features is conducive to capturing the shape and size features of stents at different scales. Hence, the feature attention module is employed to fuse different scale dense features. Secondly, the relevance attention module is designed in gating to disambiguate irrelevant and noisy responses in skip connections. Thirdly, the pre-trained MobileNetV2 encoder can reduce network parameters and improve model processing speed while ensuring performance. Additionally, the designed hybrid loss function with dice loss to address extreme class imbalance and reinforced focal loss to force model to focus on the pixels easily misclassified.

Our main contributions can be summarized as follows: (1) To the best of our knowledge, this is the first real-time approach that achieves fully automatic stent segmentation at the inference rate of 12.6 FPS in intraoperative X-ray fluoroscopy. (2) The designed double attention modules and hybrid loss improve

model sensitivity to stent wire pixels without requiring complicated heuristics. (3) The proposed LDA-Net achieves the state-of-the-art segmentation performance on three different datasets, namely SeTaX, PUGSeg and NLM Chest.

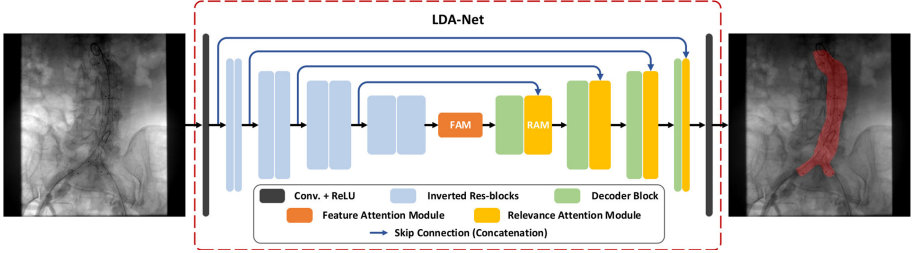


Fig. 1. An overview of the Lightweight Double Attention-fused Networks (LDA-Net). It contains double attention modules, namely feature attention module (FAM) and relevance attention module (RAM).

2 Method

In this section, we first present a general architecture of our proposed network and then introduce the designed double attention modules, namely feature attention module (FAM) and relevance attention module (RAM). Finally, we describe the hybrid loss function of both reinforced focal loss and dice loss.

2.1 Lightweight Double Attention-Fused Networks

The architecture of the proposed LDA-Net is shown in Fig. 1. The proposed network takes the original intraoperative X-ray images as input and outputs the predicted mask for stent without any post-processing. The network is a novel encoder-decoder structure, where the pre-trained MobileNetV2 [10] is employed as the backbone in the encoder stage. The depth-wise separable convolutions in the MobileNetV2 replace the standard convolutional layers, thereby reducing considerable computational burden. The FAM is utilized to gather dense pixel-level feature from the output of MobileNetV2.

Each decoder block in decoder consists of transposed convolution and batch normalization, aims to recover the resolution of the feature map from 16×16 to 512×512 . In order to highlight salient features useful for the stent wire and disambiguate irrelevant and noisy responses in skip connections, the RAMs are designed and employed in the decoder stage.

Feature Attention Module. To gather precise dense pixel-level features, the FAM is integrated into the network, as shown in Fig. 2(a). The FAM combines features from three different scales by U-shape architecture. In order to better extract the context from different level scales, the 3×3 , 5×5 and 7×7 convolutions are utilized in the structure of FAM, respectively. Because of the low resolution of the high-level feature maps, using large kernel size does not increase computational complexity by much [11]. This structure gradually integrates the information of different scales through up and down sampling, which can integrate the adjacent scales of context features more accurately. Then, after passing through a 1×1 convolution of the original features from the encoder part, multiply the pixel-wisely by the different level attention feature. Specifically, the adaptive average pooling is used to improve model performance further.

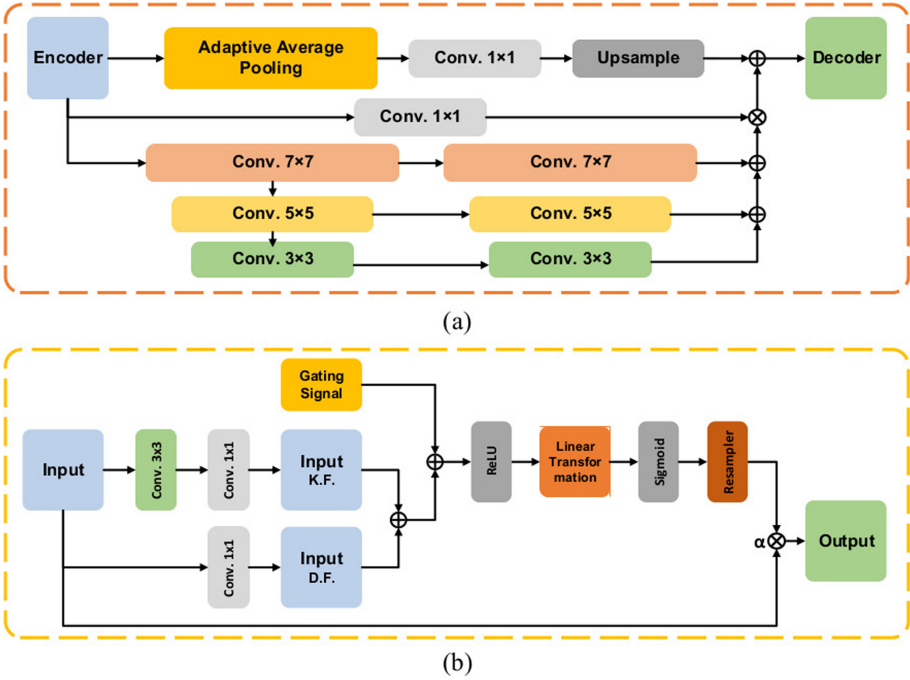


Fig. 2. (a) Schematic of the Feature Attention Module (FAM). (b) Schematic of the Relevance Attention Module (RAM).

Relevance Attention Module. In order to capture a sufficiently large receptive domain to obtain semantic context information, our designed RAMs are integrated into the LDA-Net. Compared with two-stage networks, RAM gradually suppresses the feature responses of irrelevant background regions without the necessity of region of interest (ROI). As shown in Fig. 2(b), the input of RAMs

can be divided into two parts. The first part is to obtain the key feature map (K.F.) by a series of convolution 3×3 , 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 RAMs is the element-wise multiplication of input feature maps and attention coefficients: $\hat{x}_{i,c}^l = x_{i,c}^l \cdot \alpha_i^l$. The attention coefficient $\alpha_i \in [0, 1]$ identifies image salient regions to preserve the activation relevant to the stent wire. In the default setting, a single scalar attention value is calculated for each pixel vector $x_{i,c}^l$, where F_l corresponds to the number of feature maps in layer l . The gating vector g_i is used for each pixel i to determine the focus regions. The gating vector consists of contextual information and removes lower-level feature responses as recommended in [12]. Additive attention is employed to obtain the gating coefficient [13]. The RAM is represented as follows:

$$\alpha_i^l = \sigma_2(\psi^T(\sigma_1(W_x^T x_i^l + W_g^T g_i + b_g)) + b_\psi) \quad (1)$$

where σ_1 and σ_2 represent the ReLU activation and sigmoid activation respectively. The W_x and W_g correspond to the weights of linear transformation, and b_g and b_ψ are the bias. In order to reduce trainable parameters and computational complexity of RAMs, linear transformation ($1 \times 1 \times 1$ convolution) is performed without any spatial support, and the input feature map is down-sampled to the resolution of the gating signal. Grid re-sampling of the attention coefficients is employed by trilinear interpolation. The designed RAMs are merged into our network to highlight salient features useful for the stent wire. 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 edge misclassified pixels prediction.

2.2 Hybrid Loss

In the task of stent segmentation, the thin stent wire results in class imbalance. Meanwhile, due to the contrast agent, artifacts from the spine and wire-like structures, the edge pixels of the stent turn into the misclassified samples. The huge number of easy and background samples tend to overwhelm the training. Dice loss performs relatively better than cross entropy loss when the training samples are highly imbalanced [9]. However, dice loss fails to capture the pixels on the border which are difficult to classify. The modulating factor in focal loss can automatically reduce the weight of easy examples in the training process and quickly focus the model on misclassified examples [14]. To this end, we design a hybrid loss function of both reinforced focal loss and dice loss to better address the issues of class imbalance and misclassified examples. The hybrid loss function is formulated as follows:

$$L = L_{R-Focal} + \lambda L_{Dice} \quad (2)$$

$$L_{R-Focal} = \begin{cases} -\alpha(1 - p_i)^\gamma \log p_i & y_i = 1 \\ -p_i^\gamma \log(1 - p_i) & y_i = 0 \end{cases} \quad (3)$$

where L_{Dice} is the dice loss function. y_i is the label of the i_{th} pixel, 1 for stent wire, 0 for background and p_i is the prediction probability of the i_{th} pixel. The weighting factor α and the modulating factor γ are tunable within the range of $\alpha, \gamma \geq 0$. And we have strengthened the role of weighting factor α to increase the weight contribution of the stent wire, thus solving the extreme class imbalance more efficiently. Besides, λ is also a hyper-parameter coordinating the balance between reinforced focal loss and dice loss, which is set to 0.75 in this work.

3 Experiments

In this section, quantitative and qualitative evaluations for the proposed LDA-Net are carried out on three different datasets, namely SeTaX, PUGSeg and NLM Chest X-ray Database.

3.1 Datasets

SeTaX is an intraoperative stent dataset based on 2D X-ray fluoroscopy, which is provided by Peking Union Medical College Hospital. This dataset consists of 1269 images (20 patients) in training set, 381 images (6 patients) in testing set and 254 images (4 patients) in validation set. Each image has 512×512 pixels. **PUGSeg** is an interventional tool dataset containing various stiff guidewires, which are provided by Shanghai Huadong Hospital and Peking Union Medical College Hospital. It consists of 1585 images for training, 476 images for testing and 317 images for validation. Each image has a resolution of 512×512 pixels. **NLM Chest X-ray Database** is the standard digital image database for tuberculosis [15]. 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 contains 336 cases with tuberculosis and 326 normal cases.

3.2 Implementation Details

The proposed framework was implemented on PyTorch library (version 0.4.1) with one NVIDIA TITAN Xp (12 GB). To ensure the validity of the experimental evaluation, the patient data of the training set, validation set and testing set are independent of each other. 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, the poly learning rate policy is employed, the learning rate is multiplied by the factor of 0.9 when the validation accuracy was saturated. Moreover, we set the batch size of 8, and 180 epochs was used for each model training.

We report mean precision, sensitivity and F_1 -Score to evaluate the segmentation performance. The mean processing time is calculated to verify the real-time performance. To obtain the processing time, we load the sequence into the proposed framework and compute each frame parallelly offline. The total processing time T can be computed after getting the results of all the frames (N). Therefore, we obtain the inference rate N/T frames per second (FPS) and processing time $1000 \times T/N$ ms.

Table 1. Ablation study on SeTaX. BaseNet is the regular U-Net. BCE represents Binary Cross Entropy Loss. DL and FL represent Dice Loss and Focal Loss respectively. DRF represents the proposed hybrid loss with both reinforced Focal Loss and Dice Loss.

Method	Backbone	Loss	F_1 -Score	Time (ms)
BaseNet	MobileNetV2	DRF	0.898 ± 0.009	62.6 ± 1.5
BaseNet+FAM	MobileNetV2	DRF	0.946 ± 0.023	73.5 ± 1.9
BaseNet+RAM	MobileNetV2	DRF	0.925 ± 0.011	68.5 ± 0.9
LDA-Net	MobileNetV2	DRF	0.969 ± 0.015	79.6 ± 1.3
LDA-Net	ResNet-50	DRF	0.970 ± 0.018	143.6 ± 1.7
LDA-Net	ResNet-101	DRF	0.973 ± 0.016	174.2 ± 2.3
LDA-Net	VGG-11	DRF	0.955 ± 0.008	142.4 ± 2.5
LDA-Net	VGG-16	DRF	0.964 ± 0.014	158.9 ± 1.9
LDA-Net	MobileNetV2	BCE	0.835 ± 0.009	–
LDA-Net	MobileNetV2	DL	0.916 ± 0.018	–
LDA-Net	MobileNetV2	FL	0.932 ± 0.012	–

3.3 Results on SeTaX

Ablation Study. To evaluate the contribution of different modules on our approach, we conduct experiments with different settings. As shown in Table 1, the double attention modules improve the model performance significantly. In details, we first conduct BaseNet with FAM, which improves the performance from 0.898 to 0.946. Then, we implement BaseNet with RAM, which yields 0.227 improvement in mean F_1 -Score. When both of FAM and RAM are integrated into BaseNet, the mean F_1 -Score reaches 0.969 and improves by 7.91 % over baseline. Specifically, it can be seen from the processing time that double attention modules do not bring much computational burden.

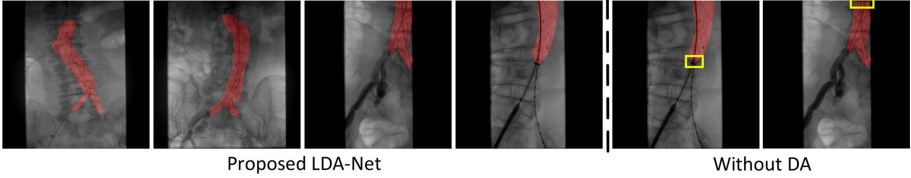
To verify the performance of backbone and loss function, we first replace the backbone of the original network with widely-used backbones ResNet and VGGNet. As shown in Table 1, it clearly demonstrates the promotion in processing speed brought by the pre-trained MobileNetV2, reducing mean processing time from 174.2 ms for ResNet-101 to 79.6 ms. Then, we employ our model on three different loss function, which are Binary Cross Entropy Loss, Dice Loss and Focal Loss respectively. Every baseline loss function is set with the best hyper-parameters. The hyper-parameter settings of our proposed hybrid loss function are as follows: $\lambda = 0.75$, $\alpha = 100$ and $\gamma = 2.5$. As shown in Table 1, our proposed hybrid loss outperforms the other three baseline loss functions remarkably.

Comparing with the State-of-the-art. To demonstrate the advantage of our proposed approach, we compare it with three widely-used networks (U-Net, LinkNet and TeraNet), two attention-based networks (Attention U-Net and CS-Net) and a previously-proposed approach on SeTaX. It is worth noting that

Table 2. Quantitative comparison with state-of-the-art approaches on SeTaX.

Method	Precision	Sensitivity	F_1 -Score	Time (ms)
U-Net [8]	0.890	0.903	0.896	104.5
LinkNet [16]	0.914	0.932	0.924	179.3
TernausNet [17]	0.939	0.923	0.932	142.8
Attention U-Net [18]	0.951	0.940	0.945	125.4
CS-Net [19]	0.942	0.955	0.948	125.8
KBS [4]	0.960	0.934	0.945	750
LDA-Net	0.962	0.978	0.969	79.6

we implement other approaches with best parameters. As shown in Table 2, it clearly demonstrates that our approach achieves better accuracy than other existing approaches in terms of mean F_1 -Score and processing time. As can be seen in Fig. 3, the proposed approach is robust to all kinds of intraoperative stents in different interventions, and the segmentation results are accurate without any post-processing. Besides, mean processing time per image of our proposed network is about 79.6 ms (12.6 FPS), which meets real-time requirements [20].

**Fig. 3.** Visualization results on SeTaX. DA represents double attention modules.**Table 3.** Quantitative comparison on PUGSeg and NLM Chest X-ray Database.

Method	PUGSeg				NLM Chest X-ray Database			
	Seq. 1	Seq. 2	Seq. 3	Mean F_1	Seq. 1	Seq. 2	Seq. 3	Mean F_1
U-Net [8]	0.884	0.909	0.911	0.901	0.899	0.907	0.864	0.890
LinkNet [16]	0.889	0.918	0.917	0.908	0.902	0.915	0.872	0.896
TernausNet [17]	0.916	0.933	0.923	0.924	0.910	0.922	0.898	0.910
Att. U-Net [18]	0.928	0.941	0.945	0.938	0.931	0.945	0.916	0.931
CS-Net [19]	0.928	0.946	0.946	0.940	0.943	0.951	0.929	0.941
LDA-Net	0.938	0.955	0.961	0.951	0.956	0.968	0.934	0.953

3.4 Results on PUGSeg and NLM Chest X-Ray Database

To further verify the effectiveness of our proposed LDA-Net, we conduct experiments on PUGSeg and NLM Chest X-ray Database. As shown in Table 3, it clearly demonstrates that our proposed LDA-Net is superior to other widely-used networks and attention-based networks in terms of F_1 -Score. Visualization results of different approaches are shown in Fig. 4. Compared with U-Net and Attention U-Net, our proposed LDA-Net can capture better contours which are usually considered as hard samples, obtaining more accurate and smooth segmentation masks. The qualitative comparison on PUGSeg and NLM Chest X-ray Database also indicates the success of our proposed approach.

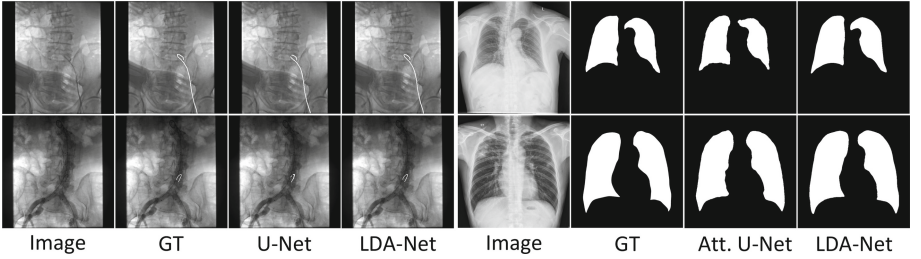


Fig. 4. Visualization results on PUGSeg and NLM Chest X-ray Database.

4 Conclusion

In this paper, we have proposed Lightweight Double Attention-fused Networks (LDA-Net) to address the challenging task of real-time stent segmentation in intraoperative X-ray fluoroscopy. Quantitative and qualitative evaluations on SeTaX, PUGSeg and NLM Chest X-ray database demonstrate that our approach achieves significant improvement in terms of both accuracy and robustness. The ablation experiments prove the effectiveness of double attention modules (FAM and RAM) and hybrid loss. By integrating these components into the network, our proposed LDA-Net effectually addresses the issues of class imbalance and misclassified examples, achieving the state-of-the-art performance. Specifically, the inference rate of our approach is approximately 12.6 FPS, which enables for real-time computer-assisted interventions.

Acknowledgments. This work was supported in part by the National Key Research and Development Plan of China (2019YFB1311700), the National Natural Science Foundation of China (U1913210, U1613210, 61533016), the CAMS Innovation Fund for Medical Sciences (2018-I2M-AI-004), and the Youth Innovation Promotion Association of CAS (2020140).

References

1. Kent, K.C.: Abdominal aortic aneurysms. *N. Engl. J. Med.* **371**(22), 2101–2108 (2014)
2. Buck, D.B., Van Herwaarden, J.A., Schermerhorn, M.L., Moll, F.L.: Endovascular treatment of abdominal aortic aneurysms. *Nat. Rev. Cardiol.* **11**(2), 112 (2014)
3. Schulz, C.J., Schmitt, M., Böckler, D., Geisbüsch, P.: Fusion imaging to support endovascular aneurysm repair using 3D–3D registration. *J. Endovasc. Ther.* **23**(5), 791–799 (2016)
4. Breininger, K., Albarqouni, S., Kurzendorfer, T., Pfister, M., Kowarschik, M., Maier, A.: Intraoperative stent segmentation in X-ray fluoroscopy for endovascular aortic repair. *Int. J. Comput. Assist. Radiol. Surg.* **13**(8), 1221–1231 (2018). <https://doi.org/10.1007/s11548-018-1779-6>
5. Zhou, Y.-J., et al.: Real-time guidewire segmentation and tracking in endovascular aneurysm repair. In: Gedeon, T., Wong, K.W., Lee, M. (eds.) *ICONIP 2019. LNCS*, vol. 11953, pp. 491–500. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-36708-4_40
6. Ambrosini, P., Ruijters, D., Niessen, W.J., Moelker, A., van Walsum, T.: Fully automatic and real-time catheter segmentation in X-ray fluoroscopy. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) *MICCAI 2017. LNCS*, vol. 10434, pp. 577–585. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_65
7. Demirci, S., et al.: 3D stent recovery from one X-ray projection. In: Fichtinger, G., Martel, A., Peters, T. (eds.) *MICCAI 2011. LNCS*, vol. 6891, pp. 178–185. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23623-5_23
8. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015. LNCS*, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
9. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: *3DV*, pp. 565–571. IEEE (2016)
10. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: *CVPR*, pp. 4510–4520. IEEE (2018)
11. Li, H., Xiong, P., An, J., Wang, L.: Pyramid attention network for semantic segmentation (2018). arXiv preprint [arXiv:1805.10180](https://arxiv.org/abs/1805.10180)
12. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H.: Residual attention network for image classification. In: *CVPR*, pp. 3156–3164. IEEE (2017)
13. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation (2015). arXiv preprint [arXiv:1508.04025](https://arxiv.org/abs/1508.04025)
14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *ICCV*, pp. 2980–2988. IEEE (2017)
15. Candemir, S., Jaeger, S., Palaniappan, K., Musco, J.P.: Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Trans. Med. Imag.* **33**(2), 577–590 (2013)
16. Chaurasia, A., Culurciello, E.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: *VCIP*, pp. 1–4. IEEE (2017)
17. Iglovikov, V., Shvets, A.: Terausnet: U-Net with VGG11 encoder pre-trained on ImageNet for image segmentation (2018). arXiv preprint [arXiv:1801.05746](https://arxiv.org/abs/1801.05746)
18. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M.: Attention U-Net: learning where to look for the pancreas (2018). arXiv preprint [arXiv:1804.03999](https://arxiv.org/abs/1804.03999)

19. Mou, L., et al.: CS-Net: channel and spatial attention network for curvilinear structure segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 721–730. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_80
20. Heidebuchel, F., Wittkamp, F.H., Vano, E., Ernst, S., Schilling, R.: Practical ways to reduce radiation dose for patients and staff during device implantations and electrophysiological procedures. *Europace* **16**(7), 946–964 (2014)