Deep Segmentation and Registration in X-Ray Angiography Video

Athanasios Vlontzos
athanasios.vlontzos14@imperial.ac.uk

Krystian Mikolajczyk
k.mikolajczyk@imperial.ac.uk

Dept. of Electrical and Electronic
Engineering
Imperial College
London, UK

Abstract

In interventional radiology, short video sequences of vein structure in motion are captured in order to help medical personnel identify vascular issues or plan intervention. Semantic segmentation can greatly improve the usefulness of these videos by indicating exact position of vessels and instruments, thus reducing the ambiguity. We propose a real-time segmentation method for these tasks, based on U-Net network trained in a Siamese architecture from automatically generated annotations. We make use of noisy low level binary segmentation and optical flow to generate multi class annotations that are successively improved in a multistage segmentation approach. We significantly improve the performance of a state of the art U-Net at the processing speeds of 90fps.

1 Introduction

X-ray angiography is the most used imaging modality to visualise blood vessels for interventional purposes such as stenting of stenosed vessels or for diagnostic purposes such as assessment of myocardial perfusion or stenosis grading. To minimise ionising radiation exposure of the patient and medical personnel during image acquisition, low power X-Rays are used resulting in noisy and low contrast images. In the context of diagnosis, the main object of interest is the vascular tree, its branchings and variations in thickness. It is therefore necessary to accurately highlight the vessels in consecutive frames to reduce the noise and improve contrast. In addition, in interventional procedures, identifying interventional instruments (catheter, wires) is also needed in order to better plan and control their positioning. Efficiently discriminating between instruments and vessels as well as other anatomical structures that may have similar appearance is crucial during the interventions. Figure 1(a-c) shows an example of an angiogram sequence. Note large non-rigid motion between frames as well as the ambiguity between vessels and the catheter. Figure 1(e) shows a frame from a different sequence of the same patient but taken at different scan and angle and (f) shows a different patient. There is a significant difference in vessel as well as catheter locations in all three sequences, which we consider as independent examples. Figure 1(d) shows the ground truth segmentation of the first frame.

CNN based methods provide state of the art segmentation results but need to be trained with a large number of examples with ground truth segmentations that are typically obtained by manual annotation by experts. The presence of noise and low contrast make this task
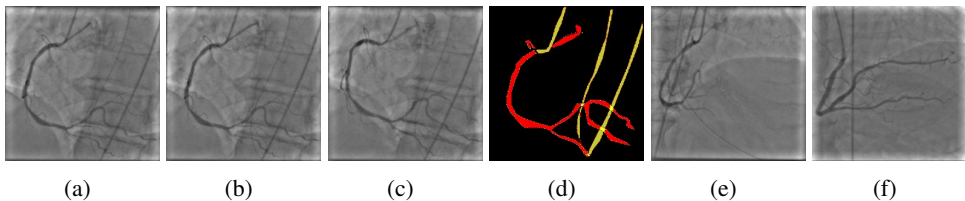
Figure 1: Frames from a video angiogram. From left to right: (a)-(c) three consecutive frames showing significant non rigid motion due to heart beating, (d) three class segmentation into background (black), vessels (red), and catheter (yellow); (e) frame from the same patient but a different sequence at different angle; (f) frame from a different patient.

particularly challenging and methods with handcrafted features lead to inaccurate object detection. Neural networks were demonstrated to achieve outstanding performance in vision tasks when trained from well annotated and large datasets. This is in contrast to unsupervised training where limited success has been achieved so far. We make a step towards that direction by developing an approach to automatically generate annotated examples by exploiting the knowledge of the application context and the data.

In this paper we present a multistage architecture based on Convolutional Neural Networks for semantic segmentation of instruments and vessel tree. We propose a method to automatically generate label proposals that are successively refined in a multistage process. This data is then used to train a CNN in a way that exploits spatial and temporal continuity. This results in a network capable of generating accurate segmentations. In summary, our main contributions are:

- We introduce a new approach for weakly supervised training of segmentation network applied to video angiograms. We propose a method that exploits low-level segmentations and optical flow to generate annotations for multi-class video segmentation.

- We demonstrate that a convolutional network can be trained from automatically segmented sequences with noise, and can improve the quality of these unsupervised segmentations.

- We improve the performance of the state of the art segmentation network U-Net [19] by 20%, processed with 90fps.

2   Related work

Segmentation.   Image segmentation has been a well researched topic, with new techniques arising frequently. Pixel-wise segmentation strives to create pixel masks (labels) that correspond to objects of interest in the image. One of the most common variational approaches from [10] estimates the background using an incremental version of EM and then subsequently subtracts it from the images. Many variational methods rely on the calculation of a Hessian matrix to acquire a vesselness index, i.e. the probability of a given pixel belonging to a vessel. In [9],[8] the authors construct a Hessian filter to enhance and segment coronary arteries, based on the Hessian matrix and the eigenvalues of the image. In [21], a decimation free directional filter bank is used to decompose, analyse and then average the images

to produce a final vesselness metric that is less sensitive to noise. In [8] the Hessian filter is enhanced with a pipeline consisting of a guided filter, Canny edge detection and block searching in order to determine the vessels of the angiography X-Rays; [16] further extends this approach to address its diaphragm misclassification. An interesting variational method comes from [12], which split the images into three layers; the breathing layer, the quasi-static layer and the vessel layer based on the movement of each of the layers. The layers are then processed using morphological operators and principal component analysis to segment the vessels. A Polygonal Path Image was introduced in [3] to unify local and global curvature structures with controllable smoothness to highlight guide wires.

Due to the data hungry nature of CNN approaches and the lack of annotated fluoroscopy data, the related work has been focusing on modalities like Magnetic Resonance Imaging (MRI), Computer Tomography (CT) or Positron Emission Tomography (PET). One of the most popular and effective architectures for medical image segmentation has been U-Net as introduced in [19] or 3D-Unet [4], where a symmetric convolutional network is segmenting and reconstructing images. A residual neural network type architecture with two streams, a low resolution and a high resolution one, was investigated in [13] to segment brain lesions from MRI volumes. In [17], an extension of the well known GrabCut [2] was introduced for pixel wise segmentation initialised by a bounding box of a of neo-natal brain lesion.

Tracking and registration. Tracking of blood vessels in angiography is very challenging since vessels appear and disappear as contrast agent is injected and vessel shapes change because of the heart and lung motion. Due to changes in appearance, template trackers (e.g. [6]) are less robust. Model based methods such as [14] regard blood vessels as active curves and deform the curves in the current frame to match the curves in the next frame. They are sensitive to noise and large change of the vessel shape. This problems led to classification-based tracking where pixels are classified in every frame as belonging to objects or background instead of transforming the appearance of objects from frame to frame. Such a tracker is robust against appearance changes and intermittent object presence. Classification-based trackers can be static e.g. [1] where training is performed prior to tracking or adaptive, e.g. [5] where a classifier is built during tracking. Since our goal is to create a real-time pipeline, our preference is with static-type tracking using a convolutional neural network which can be very fast in inference mode while it can generalise well so that shape variability is accounted for. A multi-resolution registration algorithm was proposed in [23] where the mask image is decomposed to coarse and fine sub-image blocks iteratively and each block is rigidly registered to the live image. An iterative refinement algorithm was introduced in [22] for registration in DSA. Nonrigid motion is iteratively modelled using thin-plate spline (TPS) calculated from a set of corresponding interest points. The iterative nature of such registration algorithms leads to high computational time, which makes them difficult to use in real-time clinical applications. In [18] the authors use a prior CT Scan to perform a 3d to 2d registration of the coronary artery. The process however is not real time.

3 Weakly supervised segmentation

In this section we present our approach for generating segmented sequences that are then used to train an automatic segmentation method. There is a large number of available unannotated X-ray angiography sequences. These sequences are captured after a catheter is introduced and contrast agent injected. In the first few frames, only the catheter is visible, with

vessels appearing once contrast agent is injected and propagates through the vessels. The goal is to classify pixels into three labels i.e. *background*, *vessels* with contrast agent, and a *catheter*.
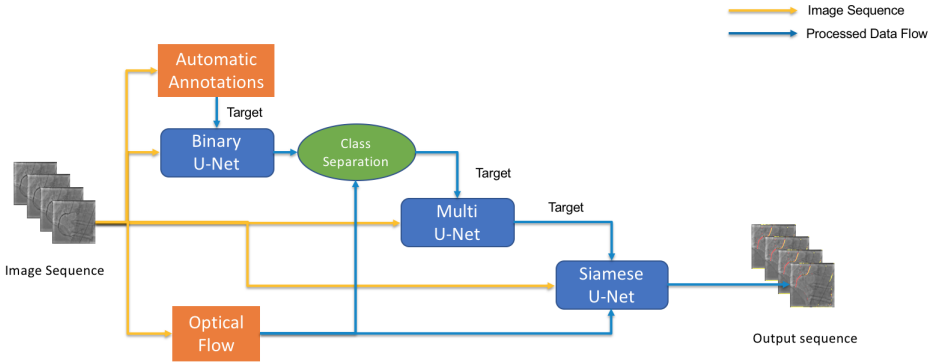

Figure 2: Proposed training pipeline. Input: sequence of X-Ray images; Preprocessing (orange): automatic annotations and optical flow estimation; CNN based segmentation (blue), binary, multi-class and custom Siamese U-Net; Optical Flow (green) estimation for class separation. Each step of the pipeline uses the output of the previous as its target.

The proposed pipeline for annotating data and training Siamese U-Net is illustrated in Figure 2. We first employ a simple low-level morphological operator to generate binary masks for training a binary U-Net that discriminates between background and other labels. Such trained U-Net outputs segmentation masks with significantly reduced noise compared to the low level processing. Motion maps, generated by coarse-to-fine optical flow [15], provide additional information for more accurate segmentation and discrimination between catheter and contrast agent. The binary segmentation masks are then converted into three class labels using initial frames and their motion maps. This data is then used to train a multi class U-Net, which further improves multi-class segmentation. Finally, a Siamese U-Net is trained using image sequences, motion maps and refined multi-class segmentation masks. The approach trains a multi-class segmentation convolutional network with an unsupervised process, otherwise weakly supervised if a small amount of manually annotated data is used to pre-train or fine-tune the network in addition to the large volume of data that is automatically labelled. In the following we discuss each of the components in more details.

3.1   U-Net segmentation

Our segmentation pipeline relies on U-Net approach [19] at various processing stages, we therefore briefly explain its architecture here. U-Net consists of two streams; a convolutional (downstream) and a de-convolutional (upstream) path. The convolutional path is based upon traditional convolutional neural network layers including spatial convolutions, max pooling and activation functions (ReLU). The outputs of consecutive layers are down-sampled with a stride of 2 while increasing the number of features at each level by 2. There are 4 levels with 3 convolutional layers each, that generate a compressed representation of 1024 features. This is followed by a de-convolutional stream which up-converts the feature maps while concatenating them with the corresponding ones from the convolutional stream using skip

connections to preserve details. ReLU layers are used after each deconvolution. This architecture was designed to learn a segmentation process of still images and training is typically performed with binary cross-entropy loss between original images and their segmentation masks.

3.2 Multi-class separation

Background and motion maps are used to infer the three label segmentations masks from a video sequence.

Background segmentation is performed for every input frame. Since interventional tools and vessels with contrast agent appear darker than the surrounding pixels, we employ the black top-hat morphological operator [11], defined as the difference between the closing of the image and the original one to produce approximate binary masks. The operator is used with a 9×9 structuring element. We eliminate small artefacts produced by the morphological operation using connected components analysis. The generated data still contains some noise as well as irregular vessel and catheter boundaries. Moreover there are some significant parts of vessels missing in such segmentations. We observed that a U-Net trained with such data can significantly reduce the occurrence of noise while successfully adding the missing parts. Figure 4 shows an example of such segmentation. We therefore improve the binary segmentation by using U-Net pretrained with the initial binary masks.

Catheter transfer. The binary masks do not distinguish between vessels and catheters. According to standard medical practice, the angiography recording is started before introducing the contrast agent, since the latter is absorbed within a few seconds. Thus the first few frames, typically the second, as well as their binary segmentations can be used to discriminate between background and the catheter. In order to transfer the catheter label from the initial frames to those with contrast agent and visible blood vessels we use dense optical flow vectors efficiently calculated with coarse to fine approach [15]. Motion maps are then used to transform the binary mask of frame f_0 with the catheter only, onto frame f_t that contains both catheter and vascular tree. We denote the result of the non-rigid warping f_{0t}. The segmentation labels that appear in both, f_t and f_{0t}, correspond to the catheter while the binary labels in f_t that find no correspondence in f_{0t} are classified as contrast agent (blood vessel). Thus, we generate multi-class labels using the binary masks and the optical flow. Similarly to the processing of binary masks we use this data to train a multi-class U-Net and apply it to improve the segmentation results. The multi-class U-Net is based upon the earlier binary block with the substitution of the last convolution layer with a $256 \times 256 \times 3$ layer. In addition we use categorical cross entropy (cf. Equation 1) as the loss function as opposed to the binary cross entropy from the earlier version. The input to the multi-class U-Net are the raw frames and the optical flow class separated masks as target labels.

3.3 Siamese U-Net

U-Net as proposed in [19] and discussed in Section 3.1 is designed to segment static images. We extend this approach to make use of the temporal information captured by dense motion maps. Inspired by the success of Siamese architecture in vision related tasks such as match-
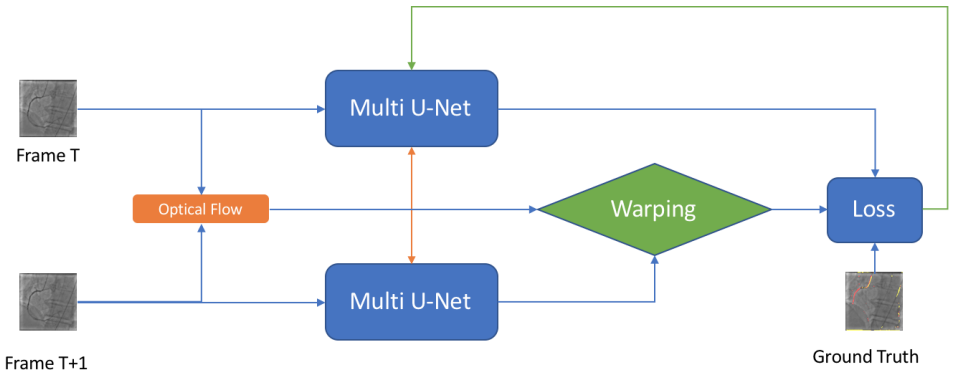
Figure 3: Siamese U-Net; Input: pair of frames f_t and $f_{t+\Delta t}$ and optical flow between the frames; Multi-Class U-Nets in Siamese setting sharing weights. Warping of segmentation mask of frame $f_{t+\Delta t}$ according to Optical Flow. Categorical cross-entropy loss uses segmentation masks of both, f_t and $f_{t+\Delta t}$, to update the weights.

ing two images, we train a Siamese network that analyses pairs of frames from different time instances and enforces temporal consistency between their segmentations.

The architecture is illustrated in Figure 3 and consist of two U-Nets that share their weights. At each training iteration the Siamese U-Net takes two frames, f_t and $f_{t+\Delta t}$, producing segmentations s_t and $s_{t+\Delta t}$. The dense motion map generated by optical flow between the two frames is used to transform $s_{t+\Delta t}$ into $s'_{t+\Delta t}$. In case of errorless optical flow and spatial segmentation, s_t and $s'_{t+\Delta t}$ should be identical. The multi-class cross entropy loss is

$$\ell(f_t, f_{t+\Delta t}, y_t) = -\sum_c y_{t,c}[\log(s_{t,c}) + \log(s'_{t+\Delta t,c})] \qquad (1)$$

where $y_{t,c}$ are automatically generated segmentation labels as discussed in Section 3.2 and index c corresponds to one of the three classes in the segmentation and ground truth masks i.e. background, catheter, and vessels.

4   Experimental results

In this section we evaluate our proposed approach and discuss results. We note that we have evaluated the pipeline both in its unsupervised capacity as well as the weakly supervised by injecting a small amount of manually annotated images for finetuning.

Dataset   The dataset consists of anonymised fluoroscopy X-Rays of 26 different patients. The images were acquired during stent placement using a General Electric Innova 2000 system and stored according to standard medical protocol in DICOM format. In total the dataset includes 36000 frames corresponding to 365 distinct video sequences with an average of 98 frames each. Different sequences of the same patient were taken at different angles and stages of the procedure therefore they differ significantly as shown in Figure 1(c)(e)(f). Each frame is rescaled from 512×512 to 256×256 due to memory constraints. We randomly sampled 3000 pairs of frames not used in the test set, and used them to automatically generate annotations for training as discussed in Section 4.1 and Section 4.2. Furthermore, two
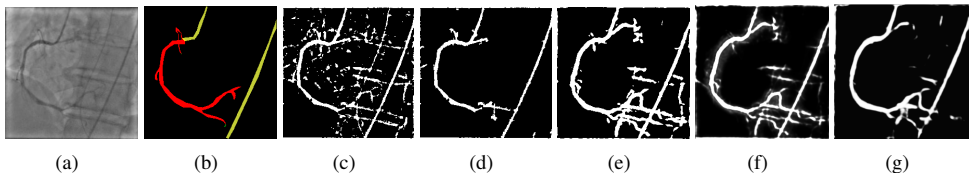
Figure 4: Segmentation process. From left to right: (a) original frame, (b) ground truth of three colour labels, (c) Black top-hat based background segmentation, (d) Mask filtered with connected components, (e) U-Net mask, (f) Siamese U-Net mask, (g) Fine-tuned U-Net.

sequences of 85 frames each, were manually annotated to serve as our test set for quantitative evaluation. Finally, we randomly selected 85 training frames for which we performed manual annotation using MIT's CSAIL LabelMe tools [20]. These images were then augmented 10-fold by applying random rotations within range $[-45°, 45°]$, and used in data augmentation experiments.

4.1 Qualitative evaluation

Using the approach discussed in section 3.2 we automatically generated annotations for 3000 randomly selected frames from sequences different than those with manual ground truth. In order to highlight the difference between the manual and automatic segmentation we present both in figure 4. The ground truth mask includes three labels i.e. background (black), vessel tree (red), and catheter (yellow). Black top-hat based segmentation mask contains significant noise, boundary irregularities and large parts of vessels as well as catheter are missing. The masks filtered with connected components seem to reduce the amount of noise but also further remove some small valid segments. This is crucial when comparing quantitative results of this simple segmentation method with and without connected component filter in Section 4.2.
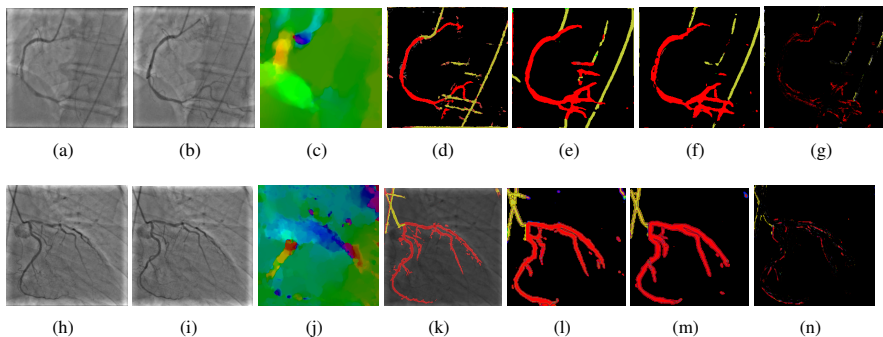

Figure 5: Multi-class segmentation results. (a) original frame at t, (b) original frame at $t + 1$, (c) optical flow (HSV, hue=angle, value=magnitude), (d) optical flow based class separation, (e) multi-class U-Net generated mask, (f) Siamese U-Net mask, (g) difference between warped mask $t + 1$ and mask $t + 1$. (h) to (n) results from a different sequence.

3000 pairs of frames with their optical flow from the automatically generated dataset were used to train the binary U-Net and the Siamese U-Net models presented in Sections 3.1 and 3.3. As illustrated in Figure 4(f) the Siamese U-Net is more accurate than the baseline U-Net. U-Net fine-tuned with manually annotated and augmented data provides masks with less noise and smooth object boundaries, although the same segments are successfully extracted in both, unsupervised Siamese U-Net and the fine-tuned one.

Figure 5 compares multi-class segmentation obtained with optical flow, multi-class U-Net and Siamese U-Net. All methods were trained with automatically generated annotations without fine-tuning. Both methods provide visually comparable segmentations therefore in the following section we perform a quantitative analysis of the segmentation accuracy.

Some of the errors appearing in the segmentation (Figure 5(g),5(n)) can be attributed to the use of simple interpolation layers instead of transposed convolutions. Despite the use of skip connections between encoding and decoding sections of the U-Net, the use of simple up sampling layers introduces loss of details and thickening of the vessels by a few pixels. We believe that use of transposed convolutions would improve results slightly. This will be the subject of future study.

4.2   Quantitative evaluation

In this section we present quantitative results for our proposed segmentation approach. The comparison is done using the dice index defined between segmentation masks X and Y

$$Dice = \frac{2\|X \cap Y\|}{\|X\| + \|Y\|} = \frac{2TP}{2TP + FP + FN} \qquad (2)$$

where TP stands for the true positive number of labeled pixels, FP is the false positive number and FN are false negative pixels. Dice index, however, is best suited for binary segmentation, therefore, to address this shortcoming we report per class dice index in addition to an overall score related to the binary segmentation by the proposed methods. In the multi-class case we define the binary mask as the union of the catheter and contrast agent masks. Table 1 shows the results for various stages of the binary segmentation pipeline noting the variant of the approach. Interestingly, according to the dice index, the black top-hat segmentation produces more accurate masks than the connected components filter. As discussed in Figure 4(d) the connected components filter also removes some true positive segments in the mask. Unsupervised binary U-Net significantly improves the results. The proposed Siamese U-Net provides the best results with dice score of 0.98. Note that the multi-class labels are converted here to binary masks of background and the remaining two labels considered as one. We further evaluate our models in multi-class segmentation task and report the dice scores for each label in table 2. Again, our proposed Siamese U-Net consistently improves upon the state of the art U-Net from [19]. We test two models trained with data augmented with different transformations i.e. small random rotation & translation in Augm_1 vs. rotation only in Augm_2. The latter seems to be more suited for the type of transformations that the input sequences typically include. Our best performing Siamese U-Net outperforms the original approach from [19] by more than 20%.

Architecture	Dice
Black top-hat + CC	0.77
Black top-hat	0.85
U-Net	0.92
U-Net + CC	0.89
Siamese U-Net	0.95
Siamese U-Net + CC	**0.98**

Table 1: Binary segmentation results. CC stands for connected component postprocessing. Best dice score is obtained by the proposed Siamese U-Net trained in unsupervised way without data augmentation.

Architecture	Catheter	Vessels
U-Net	0.49	0.30
U-Net + CC	0.44	0.36
Siamese U-Net + CC	0.51	0.50
Siamese U-Net + Augm_1	0.61	0.49
Siamese U-Net + Augm_2	**0.69**	**0.54**

Table 2: Dice scores for multi-class segmentation. Augm_1 is augmentation with translations and rotations while Augm_2 is with rotations only. Our Siamese U-Net with data augmentation significantly improves the results upon the original U-Net [19].

5 Efficiency

The proposed method was implemented in Python and the Keras framework within Tensorflow and tested with a single NVidia Titan-X GPU. In inference mode, the processing speed of U-Net trained in Siamese setup is 90 frames per second, which is suitable for real time segmentation of angiography sequences. Training time ranges from 20 min up to 8 hours when all 3000 automatically generated annotations are used for training.

6 Conclusions

We have introduced an approach for weakly-supervised training of a multi-class segmentation network in medical application of angiogram sequences. Our pipeline exploits automatic low level binary segmentations as well as motion vectors from optical flow to generate data annotations for multi-class segmentations thus avoiding tedious manual segmentation of many training examples. We show that automatic annotations that include some noise can still be used to train a CNN approach and lead to improved results. We also demonstrated how to enforce the temporal consistency between segmented frames within a siamese training architecture. Multi-class U-Net trained in siamese settings significantly improves the accuracy of segmentations compared to the original U-Net approach. Our network achieves processing speed of 90 frames per second, far beyond the speed capabilities of most of the existing interventional systems which makes it suitable for a real time applications. More experiments with different type of medical data using similar training setup will further validate this approach.

References

[1] S. Avidan. Support vector tracking. *TPAMI*, 2004.

[2] H.R. Fazlali B. Felfelian, N. Karimi, SMR Soroushmehr, S. Samavi, B. Nallamothu, and K. Najarian. Vessel segmentation in low contrast x-ray angiogram images. *ICIP*, 2016.

[3] V. Bismuth, R. Vaillant, H. Talbot, and L. Najman. Curvilinear structure enhancement with the polygonal path image - application to guide-wire segmentation in x-ray fluoroscopy. *MICCAI*, 2012.

[4] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *arXiv:1606.06650 [cs.CV]*, 2016.

[5] R. Collins, Y. Liu, and M. Leordeanu. Online selection of discriminative tracking features. *TPAMI*, 2005.

[6] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *TPAMI*, 2003.

[7] C.Rother, V. Kolmogorov, and A. Blake. Grubcut: interactive foreground extraction using iterated graph cuts. *ACM TOG*, 2004.

[8] H.R. Fazlali, N. Karimi, S.M.R. Soroushmehr, S. Sinha, S. Samavi, B. Nallamothu, and K. Najarian. Vessel region detection in coronary x-ray angiograms. *ICIP*, 2015.

[9] A.F. Frangi, W.J. Niessen, K.L. Vincken, and M.A. Viergever. Multiscale vessel enhancement filtering. *MICCAI*, 1998.

[10] N. Friedman and S. Russel. Image segmentation in video sequences: A probabilistic approach. *arXiv:1302.1539 [cs.CV]*, 2013.

[11] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Pearson, 4th edition, 2018.

[12] H.Ma, A. Hoongendoorn, E. Regar, W.J. Niessen, and T. van Walsum. Automatic online layer separation for vessel enhancement in x-ray angiograms for percutaneous coronary interventions. *MIA*, 2017.

[13] K. Kamnitsas, E. Ferrante, S. Parisot, C. Ledig, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker. Deepmedic for brain tumor segmentation. *BRATS Challenge*, 2016.

[14] I. Nwogu and L. Lorigo. Fast temporal tracking and 3d reconstruction of a single coronary vessel. *ICIP*, 2007.

[15] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning feautres by watching objects move. *CVPR*, 2017.

[16] S. Petkov, X. Carrillo, P.Radeva, and C.Gatta. Diaphragm border detection in coronary x-ray angiographies: New method and applications. *CMIG*, 2014.

[17] M. Rajchl, M. C.H. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. De-modaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, and D. Rueckert. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *arXiv:1605.07866 [cs.CV]*, 2016.

[18] D. Rivest-Henault, H. Sundar, and M. Cheriet. Nonrigid 2d/3d registration of coronary artery models with live fluoroscopy for guidance of cardiac interventions. *TMI*, 2012.

[19] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

[20] B. Russel and A. Torralba. Labelme - open annotation tool. URL http://labelme.csail.mit.edu/Release3.0/.

[21] P.T. Truc, M.A. Khan, Y.K. Lee ans S. Lee, and T.S. Kim. Vessel enhancement filter using directional filter bank. *CVIU*, 2009.

[22] J. Wang and JQ Zhang. An iterative refinement dsa image registration algorithm using structural image quality measure. In *IIH-MSP*, 2009.

[23] J. Yang, Y. Wang, S. Tang, S. Zhou, Y. Liu, and W. Chen. Multiresolution elastic registration of x-ray angiography images using thin-plate spline. *TNS*, 2007.