Incorporating Temporal Prior from Motion Flow for Instrument Segmentation in Minimally Invasive Surgery Video

Yueming Jin[1(✉)], Keyun Cheng[1], Qi Dou[2], and Pheng-Ann Heng[1,3]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
ymjin@cse.cuhk.edu.hk
[2] Department of Computing, Imperial College London, London, UK
[3] T Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong, China

Abstract. Automatic instrument segmentation in video is an essentially fundamental yet challenging problem for robot-assisted minimally invasive surgery. In this paper, we propose a novel framework to leverage instrument motion information, by incorporating a derived temporal prior to an attention pyramid network for accurate segmentation. Our inferred prior can provide reliable indication of the instrument location and shape, which is propagated from the previous frame to the current frame according to inter-frame motion flow. This prior is injected to the middle of an encoder-decoder segmentation network as an initialization of a pyramid of attention modules, to explicitly guide segmentation output from coarse to fine. In this way, the temporal dynamics and the attention network can effectively complement and benefit each other. As additional usage, our temporal prior enables semi-supervised learning with periodically unlabeled video frames, simply by reverse execution. We extensively validate our method on the public 2017 MICCAI EndoVis Robotic Instrument Segmentation Challenge dataset with three different tasks. Our method consistently exceeds the state-of-the-art results across all three tasks by a large margin. Our semi-supervised variant also demonstrates a promising potential for reducing annotation cost in the clinical practice.

1 Introduction

With advancements of robot-assisted minimally invasive surgery, enhancing automatic context awareness of the surgical procedure is important for improving surgeon performance and patient safety. Segmentation of the surgical instrument plays a fundamental role for various further tasks including tool pose estimation, tracking and control. In addition, for augmented reality, referring a segmentation mask can prevent the overlay of rendered tissue from occluding instruments.

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11768, pp. 440–448, 2019.
https://doi.org/10.1007/978-3-030-32254-0_49

However, accurate instrument segmentation from surgical videos is very challenging, due to the complicated scene, blur from instrument motion, inevitable visual occlusion by blood or smoke, and various lighting conditions. Recognizing the instrument in greater details, e.g., separating its different parts or specifying its sub-type, is even harder given the limited inter-class variance.

To meet these challenges, early methods use hand-crafted features from color and texture, with machine learning models such as random forests and Gaussian mixture model [3,13]. Later, convolutional neural network (CNN) based methods have demonstrated new state-of-the-art on instrument segmentation. The ToolNet [5] uses a holistically-nested fully convolutional network, imposing multi-scale constraint of predictions. Laina et al. [9] propose a multi-task CNN to concurrently regress the segmentation and localization. Milletari et al. [11] use residual CNN and integrate multi-scale features of a frame via LSTM. Shvets et al. [16] design a skip-connection model trained with transfer learning, winning the 2017 EndoVis Challenge [2]. The existing works treat sequential data as static image, and perform segmentation purely using visual cues in single frame.

With the sequential nature, temporal information actually can provide valuable clues for video analysis, and has demonstrated benefit in other surgical tasks, e.g., workflow recognition [7,18], instrument detection [15], and pose estimation [1]. These methods either implicitly learn spatio-temporal features in a network (generally with LSTM), or straightforwardly take the optical flow map as an extra input channel to a network. In addition, they only need to produce coarse predictions rather than pixel-level dense segmentation. How to more interpretably utilize time cues and more explicitly incorporate it into a network, are of large importance to achieve an accurate segmentation.

We propose a novel framework integrating a prior derived from **m**otion **f**low into a **t**emporal **a**ttention **p**yramid network (named MF-TAPNet) for automatic instrument segmentation in minimally invasive surgery video. Our method uses the inherent temporal clues from the instrument motion to boost results. Specifically, we propagate the prediction mask of the previous frame, via optical flow in an unsupervised way, and infer a reliable prior indicating the instrument's location and shape in the current frame. Next, we make explicit use of this temporal prior, by incorporating it at the bottleneck layer of a segmentation network as an initial attention map, and evolve a pyramid of attention modules. In this way, the sequential dynamics and the attention network can complement and progressively highlight discriminative features (or suppress irrelevant regions). As an exciting additional usage, our method enables semi-supervised learning at periodically unlabeled video, simply by propagating the prior in reverse direction. We evaluate our method on three different tasks of 2017 MICCAI EndoVis Challenge. Our MF-TAPNet consistently outperforms the leaderboard methods at all tasks. Our semi-supervised setting also achieves promising results only requiring labeling 50% frames, which endorses potential value in clinical practice.

2 Method

Figure 1 presents our proposed MF-TAPNet, which incorporates motion flow based temporal prior to a designed attention pyramid network for accurate surgical instrument segmentation from video. We elaborate each component in this section.

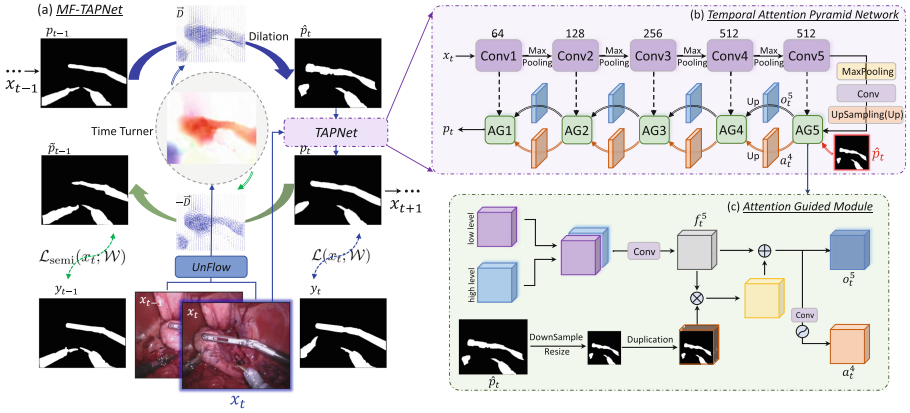

Fig. 1. Illustration of the proposed (a) MF-TAPNet for surgical instrument segmentation based on motion flow, with architecture of (b) temporal attention pyramid network and (c) attention guided module presented in detail.

2.1 Unsupervised Temporal Propagation via Motion Flow

In surgical video, instruments performed by surgeons, usually have obvious and rich motion information. Such valuable temporal inherence in the sequential data is unexplored in previous works on instrument segmentation. We propose a novel temporal prior propagation strategy, named as ***time turner***, to take advantage of such domain knowledge to a large extent.

Intuitively, we argue that the motion derived from the raw images (frames in video data) also applies to their corresponding instrument masks. This generally shares the spirit with atlas-based segmentation, but here, our "deformation field" is the motion flow between sequential frames in video. In practice, we derive the apparent instrument movement using optical flow which is de-facto for motion analysis. More specifically, we use Unflow [10], a recent state-of-the-art method, to obtain a map $\vec{\boldsymbol{D}}$ of displacement vector between adjacent frame pair of (x_{t-1}, x_t), showing the motion magnitude and orientation at each pixel. In the map $\vec{\boldsymbol{D}}$, each displacement vector $\vec{\boldsymbol{d}} = [d_a, d_b]$ directs a position from frame x_{t-1} to x_t. In our intuition, their instrument masks also follow the same location shift with such motion. Given the mask prediction p_{t-1} (output from a

network given input of x_{t-1}), we can propagate it with $\vec{\boldsymbol{D}}$ to infer a prediction for x_t. Formally, with denoting $\boldsymbol{u}_{t-1} = [u_a, u_b]$ as the position of one value in p_{t-1}, we infer its position in next frame as $\boldsymbol{u}_t = \boldsymbol{u}_{t-1} + \vec{\boldsymbol{d}} = [u_a + d_a, u_b + d_b]$. With operation for all positions, we obtain the inferred mask prediction for x_t, which is the referred concept of *temporal prior* in this paper. We further enhance it using morphological dilation to relieve the effects of camera zoom, The finally obtained temporal prior is denoted by $\hat{p}_t$, which is very informative and of high-quality regarding the location and shape of instrument in frame x_t. Note that in multi-class segmentation, we sum the probabilities of all positive classes and get p_{t-1} as a 2D map indicating non-background probability, therefore $\hat{p}_t$ is also a 2D map accordingly. Our prior can be obtained via propagating the prediction map by computing optical flow, no matter the instrument motion is large or mild compared with background motion. Therefore, it can be well generalizable to some unusual yet extreme conditions in surgical video, such as video clips with the large camera motion, still instruments and no instrument.

2.2 Temporal Prior Driven Attention Pyramid Network

Way of incorporating the temporal prior $\hat{p}_t$ provided by the *time turner* is crucial for taking great advantage of it. In this regard, we design a **t**emporal **a**ttention **p**yramid **net**work (TAPNet) which consists of multi-stage attention guided (AG) modules. It injects the prior at the encoder-decoder bottleneck and progressively learns attention guide-map pyramid in coarse-to-fine, see Fig. 1(b). The temporal prior serves as initialization of the series of attentions, and forms the essential focus throughout the pyramid. Some previous methods may also use multi-stage attention, however, most works implicitly learn attention maps from home-grown features within a network [4,12]. Our TAPNet is explicitly driven by the distinct temporal prior, making the model precisely focus on the instrument regions and hence the benefit of attention pyramid is maximized.

We first elaborate the operation inside an AG module in Fig. 1(c), with example of the most coarse one (AG5) where prior $\hat{p}_t$ is incorporated. In the segmentation task, we use skip connection to concatenate low/high-level features, followed by 1×1 convolution producing f_t^5. We first downsample $\hat{p}_t$, and then duplicate it to the same channels as f_t^5. Next, we conduct element-wise multiplication between f_t^5 and the processed $\hat{p}_t$, to extract features from those spatial locations recognized in temporal prior. The result is resummed with f_t^5, outputting a representation with enlarged instrument-related activation and necessary visual context. It is forwarded to a 3×3 convolution and a Sigmoid function to generate the attention map for next stage AG. Formally, for the i-th AG, output o_t^i and attention map a_t^{i-1} for its following module are obtained with:

$$o_t^i = f_t^i + a_t^i \odot f_t^i, \;\; a_t^{i-1} = \mathit{Sigmoid}\,(\mathrm{Conv}(o_t^i; \omega)). \tag{1}$$

Both o_t^i, a_t^{i-1} are upsampled by interpolation before forwarding to the next stage. Overall, we stack 5 AG modules in pyramid to gradually decode coarse features guided by attention maps, and finally obtain the dense prediction p_t for frame x_t.

With denoting the label mask of frame x_t by y_t, we adopt weighted cross-entropy loss for multi-class segmentation, computing from all pixels in frame x_t:

$$\mathcal{L}(x_t; \mathcal{W}) = \sum -\alpha \cdot \log \mathcal{P}(y_t | x_t, \hat{p}_t), \text{ where } \hat{p}_t = \vec{\boldsymbol{D}}(p_{t-1} | x_{t-1}, x_t). \quad (2)$$

Similarly, prediction p_t of frame x_t is also used to infer $\hat{p}_{t+1}$ for its future frame x_{t+1}. To the end, a beneficial circulation for the entire network training is formed, to sequentially produce accurate segmentation masks of the entire surgical video. For the very beginning frame x_0, its prior is set as zero, but this rare case cannot affect learning. The optical flow at *time turner* is precomputed, so with p_{t-1} from the network, we can compute $\hat{p}_t$ in real-time during training.

2.3 Semi-supervision via Reverse Time Turner

Annotating medical data is time-consuming and laborious, especially for surgical video with high frequency. Excitingly, our method enables semi-supervised learning with fewer annotations using *time turner*. This is achieved by leveraging the sequential consistency to transfer the prediction of unlabeled frame to that of the adjacent frame whose label is available for loss calculation.

With a video having T frames as $\boldsymbol{x} = \{x_0, x_1, \ldots, x_{T-1}\}$, we assume that $\boldsymbol{x}$ is labeled with intervals, e.g., only $\{x_0, x_2, x_4, \ldots\}$ being labeled. This is a reasonable setting in clinical practice because it is easier for surgeons to perform low hertz labeling. The whole data therefore consists of labeled subset $\mathcal{V} = \{x_k\}_{k=2n}$ and unlabeled subset $\mathcal{U} = \{x_k\}_{k=2n+1}$. If frame x_t is unlabeled, we simply execute our *time turner* in a reverse direction, to transfer its prediction p_t into $\tilde{p}_{t-1}$ which is corresponding to frame x_{t-1}. Note that a reverse flow $-\vec{\boldsymbol{\mathcal{D}}}$ is easily obtained with element-wise negative of $\vec{\boldsymbol{\mathcal{D}}}$, without extra computation, see green arrow in Fig. 1(a). The $\tilde{p}_{t-1}$ is expected to well overlap with label of x_{t-1}, given inherent motion consistency. Hence, we can borrow y_{t-1} to calculate semi-supervised cross entropy for x_t, as:

$$\mathcal{L}_{\text{semi}}(x_t; \mathcal{W}) = \sum -\beta \cdot (y_{t-1} \cdot \log \tilde{p}_{t-1}), \text{ where } \tilde{p}_{t-1} = -\vec{\boldsymbol{D}}(p_t | x_{t-1}, x_t). \quad (3)$$

Overall, the training uses the supervised loss in Eq. (2) if a frame x_t is labeled, otherwise it uses the semi-supervised loss in Eq. (3). By encouraging the temporal consistent predictions, the bi-directional use of *time turner* effectively benefits the network learning. Our semi-supervision enabled by motion flow is inherently general and can be applicable for other medical video analysis tasks.

3 Experiments

Dataset and Evaluation Metrics. We validate the proposed framework on the public dataset of Robotic Instrument Segmentation from the 2017 MICCAI EndoVis Challenge [2]. It consists of 10 video sequences of abdominal porcine procedures. Each video contains 300 frames obtained at sampling frequency of

Table 1. Comparison of instrument segmentation results on three tasks (mean±std).

Methods	Task1: Binary segmentation		Task2: Part segmentation		Task3: Type segmentation	
	IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)
U-Net [14]	75.44 ± 18.18	84.37 ± 14.58	48.41 ± 17.59	60.75 ± 18.21	15.80 ± 15.06	23.59 ± 19.87
TernausNet [16]	83.60 ± 15.83	90.01 ± 12.50	65.50 ± 17.22	75.97 ± 16.21	33.78 ± 19.16	44.95 ± 22.89
U-NetPlus [6]	83.75 ± 15.36	90.19 ± 11.77	65.75 ± 16.74	76.25 ± 15.54	34.19 ± 15.06	45.32 ± 19.86
PlainNet	81.86 ± 15.85	88.96 ± 12.98	64.73 ± 17.39	73.53 ± 16.98	34.57 ± 21.93	44.64 ± 25.16
TAPNet	84.01 ± 16.93	90.46 ± 13.56	65.84 ± 16.91	76.12 ± 16.75	34.23 ± 19.63	45.50 ± 22.55
MF-TAPNet (Ours)	**87.56 ± 16.24**	**93.37 ± 12.93**	**67.92 ± 16.50**	**77.05 ± 16.17**	**36.62 ± 22.78**	**48.01 ± 25.64**
MF-TAPNet (50%)	79.31 ± 17.13	87.18 ± 13.68	56.01 ± 15.59	68.13 ± 15.44	28.47 ± 23.41	38.39 ± 25.88
Semi-MF-TAPNet (50%)	**80.03 ± 16.87**	**88.07 ± 13.15**	**56.72 ± 16.12**	**68.51 ± 16.11**	**30.04 ± 19.79**	**41.01 ± 23.81**

2 Hz and a high resolution of 1280×1024. Specifically, 8×225-frame videos are used for training, while the remaining 8×75-frame videos and another 2×300-frame videos are used for testing; the ground-truth of test data is held-out by challenge organizer. There are three sub-tasks, i.e. binary instrument (2 classes), instrument part (4 classes), instrument type (8 classes), gradually fine-grained segmentation of an instrument. The challenge report [2] describes more details of the difficulties in the tasks. For direct and fair comparison, we follow the same evaluation manner as TernausNet [16] (challenge winner), by using 4-fold cross-validation with the same splits of 8×225 released training data. We also use the same evaluation metrics as [16], i.e., (1) mean intersection-over-union (IoU), which is also used in MICCAI EndoVis Challenge to evaluate participants, and (2) Dice coefficient (Dice), which is another common metric for segmentation.

Implementation Details. We reduce the resolution to 640×512 to save memory. We train models using an Adam optimizer [8], with learning rates initialized as 3e−5, 3e−5 and 2e−5 respectively for binary, part and type segmentation tasks. Our framework is implemented in PyTorch with 4 NVIDIA Titan Xp GPUs for training. The multiple GPUs enable the network to be trained at batch size of 8. The backbone of our network is VGG11 [17] with 5 scales of downsampling, and deeper networks did not yield much better results in experiments, so we stick to VGG11 for the sake of real-time efficiency during surgery. The code is available at https://github.com/keyuncheng/MF-TAPNet.

Comparison with State-of-the-Art Methods. We first compare our method with the state-of-the-art results in challenge on three tasks. Table 1 lists the performance of U-Net [14] (results quoted from [16]), TernausNet [16], and latest reported U-NetPlus [6] (an enhanced U-Net with batch normalized encoders and nearest neighbor interpolation). We see that MF-TAPNet consistently outperforms all other methods across all three tasks. Our IoU exceeds the challenge winner by 3.96% at binary segmentation, 2.42% at part segmentation, and 2.84% at type segmentation. Though [16] and [6] develop advanced strategies to enhance a network, our method is superior by using temporal prior to explicitly provide a reliable guidance, which helps the network learn to focus on regions of interest.

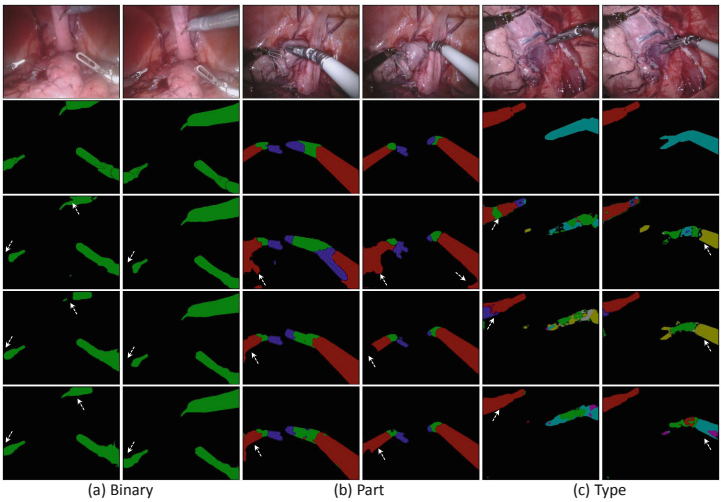

Fig. 2. Typical results for instrument (a) binary segmentation (instrument and background tissues), (b) part segmentation (shaft, wrist and jaws), (c) type segmentation (different yet looking quite similar instruments). From top to bottom, for each task, we present two continuous video frames and their corresponding ground truth, with segmentation results using PlainNet, TAPNet and our proposed MF-TAPNet.

The improvement is more obvious for binary segmentation, because our prior also aggregates probabilities from all positive classes. Under such homologous guidance, we achieve the highest Dice score 93.37%, which is useful for context-aware robot-assisted surgery. Meanwhile, there still exists a big room to boost type segmentation performance, even though we set the highest among existing methods. This reflects the natural great challenge in this task, i.e. the extremely similar appearance (shape and intensity value) in different fine-grained types. Our method is extensible for further improvement by inferring a multi-class prior, while such extension is limited for other methods. Last, as our method relies on the temporal consistency, its efficacy may degrade when unexpected motion appears, resulting in slightly higher standard deviations. This can be alleviated as advancements of more stable surgical robots.

Effectiveness of Temporal Prior and Motion Flow. We investigate effectiveness of key components in our MF-TAPNet. Table 1 also lists the results of three ablation settings: (1) a plain encoder-decoder as baseline (PlainNet); (2) our TAPNet, but directly use the previous frame's prediction p_{t-1} as temporal prior; (3) our entire framework at fully-supervised learning. The network backbone is unchanged for different settings for clear comparison. We observe that TAPNet performs better than PlainNet, especially for binary segmentation (1.50% higher Dice) and part segmentation (2.59% higher Dice). This shows that, explicitly incorporating a temporal prior can provide powerful guidance,

even with the rough prediction from previous frame. Accordingly, our TAPNet can pyramidally refine the guidance and gradually concentrate on segmenting attentive objects. Here, our TAPNet can already achieve comparable results with the state-of-the-art methods. More importantly, our MF-TAPNet further largely increases performances for all tasks (averagely 2.67% IoU). It demonstrates that after involving motion dynamics derived in *time turner*, the prior presents much higher quality and can convey more accurate shape and location in current frame. Some visual results are shown in Fig. 2. MF-TAPNet can achieve complete and consistent segmentations, and largely suppress the irrelevant and incorrect regions.

Semi-supervised Variant Enabled by Time Turner. We conduct experiment with the variant of semi-supervised learning. Our setting is that the data are labeled at an interval of 2, resulting in 50% frames having labels. In Table 1, we see that our semi-supervised loss (i.e., approximating the prediction of an unlabeled frame towards a reasonable label) can better confront the performance drop at sparse annotation, compared with an ordinary training of MF-TAPNet with 50% labeled data. This is a bonus from our *time turner* with interpretable meanings, and a simply reverse execution invokes a promising potential to reduce annotation cost which is very valuable in clinical surgery.

4 Conclusion

We propose a novel framework to incorporate temporal information pyramidally in a network for automatic instrument segmentation from robot-assisted surgery. Our method consistently outperforms the state-of-the-art methods across all the three tasks on the 2017 MICCAI EndoVis Challenge dataset, by a large margin. Our temporal prior enables semi-supervised learning simply by reverse execution. The achieved outstanding results, and demonstrated potentials for extension and label efficiency, endorse a promising value of our method in clinical intervention.

Acknowledgments. The work was partially supported by HK RGC TRS project T42-409/18-R, HK RGC project CUHK14225616, and CUHK T Stone Robotics Institute, CUHK. Yueming Jin is funded by the HK Ph.D. Fellowship.

References

1. Allan, M., Ourselin, S., et al.: 3-D pose estimation of articulated instruments in robotic minimally invasive surgery. IEEE TMI **37**(5), 1204–1213 (2018)
2. Allan, M., Shvets, A., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
3. Bouget, D., Benenson, R., et al.: Detecting surgical tools by modelling local appearance and global shape. IEEE TMI **34**(12), 2603–2617 (2015)

4. Chen, J., et al.: Multiview two-task recursive attention model for left atrium and atrial scars segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 455–463. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_51
5. García-Peraza-Herrera, L.C., Li, W., et al.: ToolNet: holistically-nested real-time segmentation of robotic surgical tools. In: IEEE/RSJ IROS, pp. 5717–5722 (2017)
6. Hasan, S., Linte, C.A.: U-NetPlus: a modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instrument. arXiv preprint arXiv:1902.08994 (2019)
7. Jin, Y., Dou, Q., et al.: SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. IEEE TMI **37**(5), 1114–1126 (2018)
8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Laina, I., et al.: Concurrent segmentation and localization for tracking of surgical instruments. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 664–672. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_75
10. Meister, S., Hur, J., Roth, S.: UnFlow: unsupervised learning of optical flow with a bidirectional census loss. In: AAAI (2018)
11. Milletari, F., Rieke, N., Baust, M., Esposito, M., Navab, N.: CFCM: segmentation via coarse to fine context memory. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 667–674. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_76
12. Oktay, O., Schlemper, J., et al.: Attention U-Net: learning where to look for the pancreas. MIDL (2018)
13. Rieke, N., Tan, D.J., et al.: Real-time localization of articulated surgical instruments in retinal microsurgery. Med. Image Anal. **34**, 82–100 (2016)
14. 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
15. Sarikaya, D., Corso, J.J., Guru, K.A.: Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. IEEE TMI **36**(7), 1542–1549 (2017)
16. Shvets, A.A., Rakhlin, A., et al.: Automatic instrument segmentation in robot-assisted surgery using deep learning. In: ICMLA, pp. 624–628 (2018)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Twinanda, A.P., Shehata, S., et al.: EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE TMI **36**(1), 86–97 (2017)