Real-time surgical instrument tracking in robot-assisted surgery using multi-domain convolutional neural network

Liang Qiu, Changsheng Li, Hongliang Ren ✉

Department of Biomedical Engineering, National University of Singapore, Singapore 117575, Singapore
✉ *E-mail: ren@nus.edu.sg*

Published in Healthcare Technology Letters; Received on 13th September 2019; Accepted on 2nd October 2019

Image-based surgical instrument tracking in robot-assisted surgery is an active and challenging research area. Having a real-time knowledge of surgical instrument location is an essential part of a computer-assisted intervention system. Tracking can be used as visual feedback for servo control of a surgical robot or transformed as haptic feedback for surgeon–robot interaction. In this Letter, the authors apply a multi-domain convolutional neural network for fast 2D surgical instrument tracking considering the application for multiple surgical tools and use a focal loss to decrease the effect of easy negative examples. They further introduce a new dataset based on m2cai16-tool and their cadaver experiments due to the lack of established public surgical tool tracking dataset despite significant progress in this field. Their method is evaluated on the introduced dataset and outperforms the state-of-the-art real-time trackers.

1. Introduction: Minimally invasive surgery has attracted broad attention in the surgical practice, which can easily access the small surgical site inside the human body and is less painful to the patients. However, manipulation of the surgical instruments in the restricted operating space has brought new problems, such as limited field of view of the endoscope, reduction of surgeon’s dexterity and lacking perception of force feedback. In recent years, robot-assisted intervention has been introduced into the study and clinic, which has provided great help to operation. Surgeons can teleoperate a robot to control articulated instruments with master manipulators, and high precision and dexterity can be obtained at the same time. Nevertheless, the operation in the complex and volatile surgical environment still poses a great challenge, so more tracking information should be collected to reduce uncertainty. The real-time pose of a surgical tool can be used to constrain the dynamic motion and provide haptic feedback for human–robot interaction. Besides, it can further help to realise the autonomous navigation of surgical instruments with servo control.

Marker-based optical tracking system and magnetic tracking system are two mainly available commercial tracking systems for surgical navigation. The magnetic tracking system uses magnetic transmitters to create a magnetic field to detect the pose of the sensors, which can avoid the occlusion problem [1]. However, it easily suffers from electromagnetic interference and its effective working space is limited. Marker-based optical tracking systems can be divided into two parts: infrared-based tracking and image-based tracking. The infrared-based tracking uses an infrared camera to localise reflective spheres attached to surgical tools with high precision, but the price is relatively high [2]. The image-based tracking utilises an ordinary camera and designed markers to acquire the location of instruments, which should take biocompatibility into consideration [3]. Recently, visual simultaneous localisation and mapping has been used for localisation of the endoscope without the aid of other equipment [4], and the registration between preoperative and intraoperative information has effectively improved the accuracy and the success rate of the surgery.

In order to realise the localisation completely relied on an image-based method without modification to the surgical setup, the surgical tool tracking or detection on 2D images is an essential step. Recently, a novel 6D object pose estimation algorithm was proposed [5]. It can exploit a denoising autoencoder to obtain 3D orientation estimation just utilising rendered 3D model views without the existence of pose-annotated training data. Then 3D translation estimation is involved in the framework considering the pinhole camera model, so knowing the camera intrinsic parameters and target locations with 2D bounding boxes is necessary here. Kurmann *et al.* [6] used a convolutional neural network (CNN) architecture to simultaneously recognise multiple instruments and estimate the positions of 20 joints in 2D images. Laina *et al.* [7] proposed a method that combines the surgical instrument segmentation and localisation together into a one deep learning architecture to realise the surgical instrument tracking, which indeed can provide more abundant medical information, but its robustness still needs to be enhanced in some challenging situations such as illumination variation, deformation and occlusion. Jin *et al.* [8] leveraged the region-based CNN to detect surgical tools and assess the operative skill based on their introduced new dataset named m2cai16-tool-location. All the methods are using object detection algorithms for surgical tool tracking, and there is no such large surgical dataset with annotation on sequential frames either. Therefore, we propose to use a deep neural network to track the surgical tool on our own dataset. It should be robust to various challenging in vivo scenes, such as deformation, motion blur, scale variation, occlusion, in-plane rotation etc.

In summary, our main contributions in this Letter are as follows:

(i) A new surgical tool tracking (STT) dataset is introduced with the bounding box as the ground-truth annotation in sequential frames. It will be made public available after accepted.

(ii) Multi-domain CNN is applied to surgical tool tracking and we optimise a multi-task loss by reducing the effect of easy negative examples into consideration. Our tracking method is demonstrated to have an improved performance against state-of-the-art real-time trackers on the STT dataset.

2. Dataset: Tracking the surgical tool during the operation poses a great challenge because of variable surgical scenes and poses of surgical tools. Besides, these surgical tools are easily missing in the frame due to the limited field of view of endoscope and occlusion that occurs frequently with tissue deformation.

However, the surgical tool detection or tracking datasets for public use are limited, which is preventing the faster improvement of computer-assisted intervention system. JIGSAWS [9] and m2cai16-tool [10] are open to public use without tool location

doi: 10.1049/htl.2019.0068

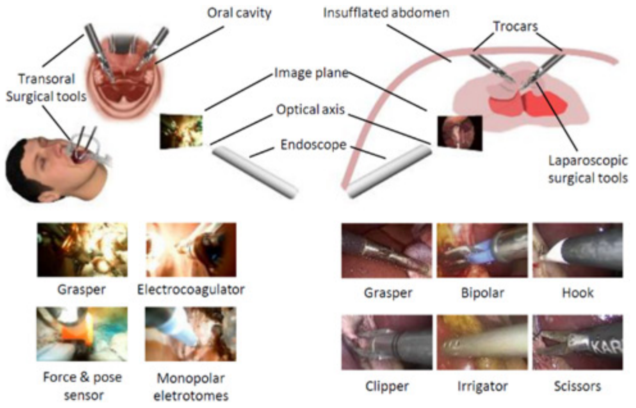

Fig. 1 *Our dataset is made based on cadaver experiment videos for transoral surgery and selected laparoscopic surgery videos in the m2cai16-tool dataset. The corresponding examples of surgical tools are shown here*

annotation. Recently, m2cai16-tool-location is introduced, which is an extension of m2cai16-tool with spatial bounds of tools. This dataset is mainly for tool identification and localisation, while not for typically surgical tool tracking problem. Besides, Sarikaya *et al.* [11] provided a surgical tool detection dataset called ATLAS Dione for public use. However, ATLAS Dione is mainly built on a phantom setting, which is still quite different from the real surgical scenes, and the type of daVinci surgical tool in the dataset is single, which is not universally suitable. We, therefore, collect and build a new dataset called STT dataset with sequential frame annotations using bounding boxes, which is a core contribution of our work.

Our dataset has two sources, as shown in Fig. 1, one is m2cai16-tool (m2cai-tool-tracking sub-dataset) which is for laparoscopic surgery and the other is our cadaver experiment for robot-assisted transoral surgery (robot-assisted-tracking sub-dataset). Forty-five videos are collected with 22,591 frames annotated with the bounding box, indicating the locations of targeted surgical instruments including (i) grasper, electrocoagulator, force/pose sensor and monopolar eletrotomes in our cadaver experiment videos and (ii) grasper, bipolar, hook, clipper, irrigator and scissors in the m2cai16-tool dataset. Particularly, each video is collected in different surgical scenes and conditions and the test dataset containing 17 videos is also selected considering the balance of surgical tool categories. Our dataset generation process is strictly according to online tracking benchmark (OTB) [12] standard, which is a mainstream object tracking benchmark in computer vision community and could be transferred to surgical tool tracking field. We made the bounding-box annotations under the guidance of surgeons who also joined our cadaver experiments. Besides, we made the annotations frame by frame manually and put our best effort to make them as accurate as possible. Considering different challenges, we carefully classify it into eight different categories, such as illumination variation, background clutter, deformation, occlusion, in-plane rotation, scale variation, out-of-plane rotation and motion blur. Each frame provided in our dataset is in JPEG format with size 1920 × 1080 pixels and the corresponding annotations are provided in the OTB format. See Table 1 for more details of our dataset.

3. Approach: Due to the application for multiple surgical tools, our approach for surgical tool tracking is based on a tracker named RT-MDNet, which means a real-time multi-domain convolutional neural network [13]. The architecture of this network is made up of several shared layers and multiple branches of domain-specific layers. The network can be trained specifically for each domain (each instrument type) and the generic target representation can be obtained in the shared layers, as shown in Fig. 2. To make the training process more effective, we exploit a focal loss to decrease the effect of easy negative examples.

Table 1 Properties of our dataset

Data sources	Surgical tool types	No. of videos	No. of frames
m2cai-tool-tracking	bipolar	4	
	clipper	2	
	grasper	3	
	hook	4	
	irrigator	4	
	scissor	4	
robot-assisted-tracking	grasper/electrocoagulator/ sensor/monopolar eletrotomes	7	
total		28	12,347
test set		17	10,244
total		45	22,591

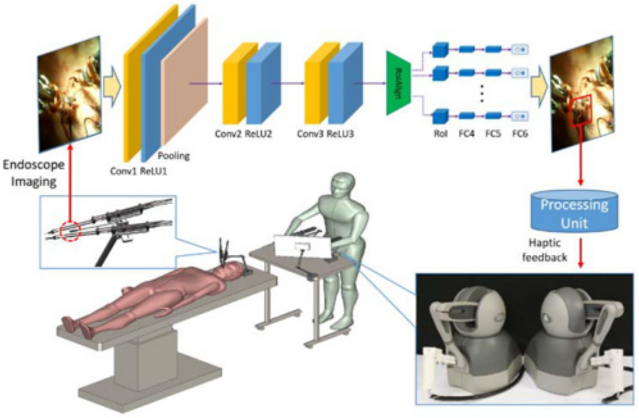

Fig. 2 *Our robot-assisted surgery framework which exploits multi-domain CNN to track surgical tools with endoscopic images as input and surgical tool location as output. The output information with bounding boxes will be further utilised in the processing unit to provide 6D pose estimation, which will provide more benefits for surgical tool navigation*

3.1. Network architecture: The RT-MDNet is an improved version of the MDNet. It can take the sequential images obtained from the endoscope mounted on the surgical robot as inputs and the outputs are the 2D bounding boxes indicating surgical tool locations, which can be used for automatic visual servo control and surgeon–robot interaction. The whole architecture of the network is made up of three convolutional layers (conv1–conv3), an improved Region of Interest Alignment (RoIAlign) layer and three fully connected layers (fc4–fc6). Firstly, the convolutional feature maps of the input images are extracted by the fully convolutional layers. Then, all the feature maps will be put into a RoIAlign layer to obtain targeted surgical tool representations. RoIAlign [14] was designed by exploiting bilinear interpolation to enhance the quantisation of the feature map, but it may also fail when the size of RoI is too large. In order to avoid obtaining the coarse extracted features and improve the representation quality of the RoIs, an improved adaptive RoIAlign layer is designed by exploiting a denser feature map from fully convolutions and enlarged receptive field of every activation. After that, fc4 and fc5 will accept the refined RoI representation as an input to classify between the surgical tool and background. The domain-specific layer fc6 with D branches (each branch corresponding to each surgical tool in a specific condition) tries to perform the multi-domain learning during the training stage and will be fine tuned with the initial frame during the testing.

doi: 10.1049/htl.2019.0068

3.2. Loss function: The real-time MDNet accelerates the procedure of accurate feature extraction with the help of the improved RoIAlign technique and dilated convolutions. The discriminative feature learning has been enhanced to distinguish multi-domain foreground objects compared to the original MDNet [15] which only considers the distinction between the foreground and background. The output score is denoted by f^d which is a concatenation of all the activation from the last fully connected layers ($fc6^1 - fc6^D$):

$$f^d = [\phi^1(x^d; R), \phi^2(x^d; R), \ldots, \phi^D(x^d; R)] \in \mathbb{R}^{2\times D}, \tag{1}$$

where x^d and R denote the input image in domain d and the corresponding bounding box, respectively, and ϕ is the mapping from the input image to 2D binary classification score.

The network is trained by optimising the loss function composed of standard cross-entropy (CE) loss for binary classification and instance embedding loss for distinguishing multi-domain target instances. The softmax function $\sigma_{\text{inst}}(\cdot)$ and the loss function L_{inst} for instance embedding are formulated as

$$[\sigma_{\text{inst}}(f^d)]_{ij} = \exp(f^d_{ij}) \bigg/ \sum_{k=1}^{D} \exp(f^d_{ik}), \tag{2}$$

$$L_{\text{inst}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D}[y_i]_{+d}\cdot \log([\sigma_{\text{isnt}}(f^d)]_{+d}), \tag{3}$$

where $y_i \in \{0, 1\}^{2\times D}$ is defined as the class label in a one-hot encoding format and the sign + means only positive examples are used.

However, due to the class imbalance between limited positive examples and quite a few negative examples, which may hold a dominant position in CE loss, the gradient update can be in an inappropriate direction leading to an unsatisfactory training model. Given a class imbalance, we take focal loss [16] into consideration and add a modulating factor $(1 - [\sigma_{\text{cls}}(f^{\hat{d}(k)})]_{c\hat{d}(k)})$ to the CE loss function as shown below:

$$[\sigma_{\text{cls}}(f^d)]_{ij} = \exp(f^d_{ij}) \bigg/ \sum_{k=1}^{2} \exp(f^d_{kj}), \tag{4}$$

$$L_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2}[y_i]_{c\hat{d}(k)}(1 - [\sigma_{\text{cls}}(f^{\hat{d}(k)})]_{c\hat{d}(k)})\log([\sigma_{\text{cls}}(f^{\hat{d}(k)})]_{c\hat{d}(k)}) \tag{5}$$

where $\sigma_{\text{cls}}(\cdot)$ and L_{cls} are defined as the binary classification softmax function and the modified loss function, respectively, and the network update is based on a mini-batch collected from the domain $\hat{d}(k) = (k \bmod D)$ in the kth iteration.

Then the network is trained by optimising a multi-task function for each frame:

$$L = L_{\text{cls}} + \alpha L_{\text{inst}}, \tag{6}$$

where α indicates the hyperparameter to balance the importance of two components in the loss function. The advantage of our designed loss function will be shown in Section 4, and evaluation on the STT dataset will be provided as well.

3.3. Online tracking: Our online tracking method is almost according to the pipeline of MDNet. At testing stage, the multiple branches fc61–fc6D will be replaced with a single initialised layer fc6 and the fully connected layers fc4–fc6 will be fine tuned to customise the new test sequence by using the initial frame with annotated bounding box as a ground truth.

The network will be updated with the long-term, short-term update methods in the rest of frames in order to improve the robustness and adaptiveness [15]. Long-term updates are used regularly by collecting positive examples, while short-term updates are conducted when tracking fails, which means the estimated scores are below a threshold. Considering the redundance and irrelevance of the negative examples over the long period, we only consider the negative examples during the short term. Actually, to obtain the target state based on an input image, a group of samples $x^1, x^2, \ldots, x^N$ is selected around the previous target position by utilising a Gaussian distribution. The optimal target state x^* can be obtained by

$$x^* = \arg\max_{x^i} f^+(x^i), \tag{7}$$

where $f^+(x^i)$ denotes the positive score of the ith sample.

Besides, we also exploit the bounding-box regression technique [17] to obtain tight bounding box, which usually means better localisation accuracy. Consider its time-consuming problem for an online update, a linear regression model is trained only in the first frame and it will be used to adjust the estimated bounding boxes obtained from (6) in the following frames if they satisfy $f^+(x^i) > 0.5$ in our application.

4. Experiments and evaluation

4.1. Implementation: The implementation details of our method are similar to RT-MDNet. The first three convolutional layers (conv1–conv3) are initialized with the weights transferred from VGG-M [18] network which is pretrained on ImageNet [19], while the following fc4–fc6 are randomly initialised.

The training process and hyperparameter in our case are given as follows:

Offline pretraining: For each iteration, we collect examples for each minibatch from a single domain. We define the positive and negative bounding boxes based on Intersection over Union (IoU). If IoU of an example is larger than 0.7, it is treated as a positive one, while the IoU of negative example is usually lower than 0.5 in our case. Besides, the hyperparameter α in (5) is set to 0.2.
Online training: For the first frame, we need to fine tune the offline pertained. Here, we collect 500 positive and 5000 negative examples. For the rest frames, 50 positive examples with >0.7 IoU and 200 negative examples with <0.3 IoU are collected. Besides, we conduct the long-term update for every 10 frames.
Optimisation (stochastic gradient descent): For learning rate, 0.0001 is set for offline training with 800 epochs, while 0.0003 is for fine tuning. Weight decay is set to 0.0005 and momentum is set to 0.9.

4.2. Evaluation: We perform ablation studies about loss function on the STT dataset and further compare our method with the other five real-time trackers: SiamFC [20], DSST [21], BACF [22], ECO-HC [23] and RT-MDNet on the STT dataset and its two sub-datasets. All these compared real-time trackers are all the state-of-the-art trackers published in the last 3 years.

The standard one-pass evaluation (OPE) approach presented in a tracking benchmark [16] is followed, which includes the precision plot (centre location error) and the success plot (bounding-box overlap ratio) metrics. The precision plot is generated by measuring the frame rates of successfully tracked targets within different centre location error thresholds. The threshold used for ranking is set to 50 pixels which is different from other ordinary datasets such as OTB (20 pixels) [16] because the resolution of our dataset (1920 × 1080) is twice larger than the others. The success plot metric considers the bounding-box overlap ratio between the ground truth and the predicted result. The ranking order in the success plot is determined by a criterion called area under curve score [16].

Figs. 3*a* and *b* show that our method has an improvement over RT-MDNet, which means our loss function effectively decreases the bad impact from a large amount of easy negative examples and makes the network mainly focus on hard examples. Besides, our tracker surpasses all the other real-time trackers on STT dataset as well, as shown in Figs. 3*c*–*h*. In Fig. 4, we can further see that our tracker has good performances in various challenging scenarios: illumination variation, background clutter, deformation, occlusion, in-plane rotation, scale variation, out-of-plane rotation and motion blur, which also indicates that our tracker has good robustness. The quantitative comparison including frame-per-second (FPS) is presented in Table 2. Our method runs with 14 FPS on average which decreases heavily compared with running on other ordinary datasets [10] due to the much larger resolution of our input images. Specifically, multiple convolutional layers and fully connected layers in the network are applied to extracting feature maps, whose computational cost increases along with the input image resolution going up. Besides, the improved RoIAlign technique which is used to alleviate the ineffectiveness of target localisation due to coarse feature map also increases the computational complexity because it enlarges the receptive field and requires computing a denser convolutional feature map. The other trackers are similarly affected by a similar reason to a certain extent.

Notably, efficient convolution operator hand-crafted feature version of ECO ((ECO)-HC) which is real-time variant of ECO also performs well on our dataset and its tracking speed performance is particularly outstanding. Although ECO-HC runs twice faster, our tracker still has almost 2% higher success rate and precision. The speed of our method is almost the same with ECO-HC when tested on OTB, but decreases a lot when the input resolution increases twice larger in our dataset due to multi-convolutions with high-resolution images and improved RoIAlign technique to improve the representation quality of RoIs. Furthermore, our method outperforms the other algorithms (SiamFC, DSST, BACF and RT-MDNet) much more considering precision and success plots.

Besides, the qualitative evaluation of the six trackers is shown with remarkable selected example frames which indicate that our method outperforms the state-of-the-art on the STT dataset in Fig. 5. The first four rows of the images show testing performances on m2cai-tool-tracking sub-dataset while the last two rows are from

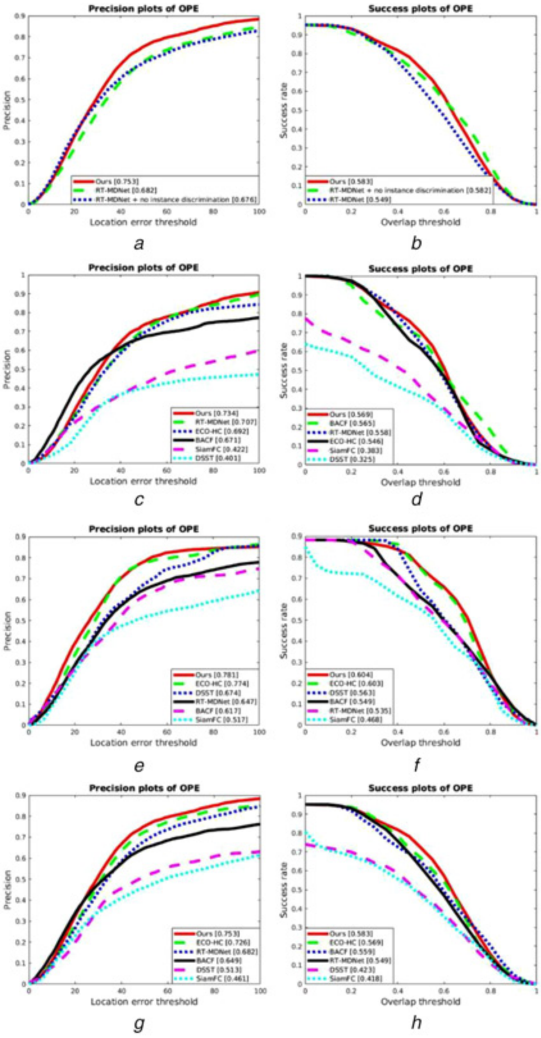

Fig. 3 *Precision and success plots using OPE*
a, *b* Ablation study: our method compares with RT-MDNet and the corresponding version without instance embedding loss on our STT dataset
c–*h* Show quantitative results of six real-time trackers on m2cai-tool-tracking sub-dataset, robot-assisted-tracking sub-dataset and STT dataset

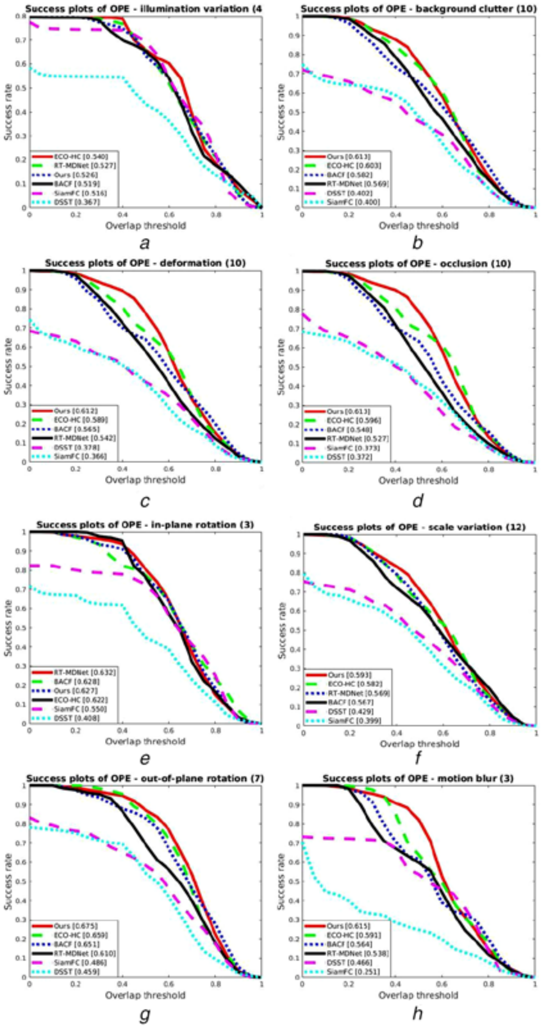

Fig. 4 *Success plots of six real-time trackers over eight tracking challenges*
a Illumination variation
b Background clutter
c Deformation
d Occlusion
e In-plane rotation
f Scale variation
g Out-of-plane rotation
h Motion blur

doi: 10.1049/htl.2019.0068

Table 2 Quantitative comparisons of six real-time trackers on STT dataset

Trackers	SiamFC	DSST	BACF	ECO-HC	RT-MDNet	Ours
succ (%)	25.1	46.6	56.4	59.1	53.8	61.5
prec (%)	48.6	45.9	65.1	65.9	61.0	67.5
FPS	19.4	11.5	9.9	32.6	13.9	14.0

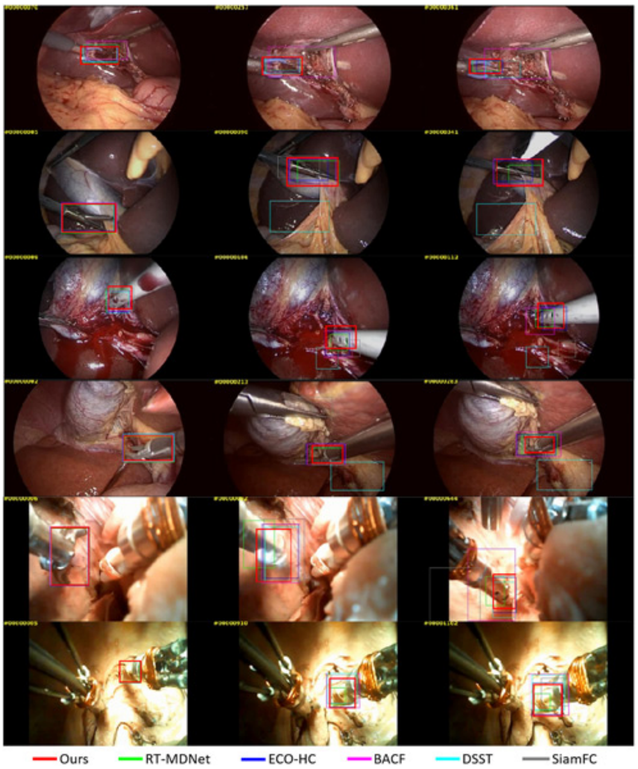

Fig. 5 *Qualitative evaluation of six real-time trackers with example frames shows that our method outperforms the state-of-the-art on STT dataset*

our cadaver experiments. Each row includes three sample images from a test sequence. From each case, our tracker can effectively detect the location of the target surgical tools with higher accuracy and tighter bounding boxes, while the other trackers have larger tracking errors with more invalid information, even total failure in some cases.

Although our algorithm has good performance over eight tracking challenges, it cannot work well when the surgical tool is totally out of sight, and tracking will be lost sometimes. Such relocalisation problem is much more challenging and complex, which is beyond the scope of this Letter.

5. Conclusion: In this Letter, we applied the multi-domain CNN to surgical instrument tracking. We design a novel multi-task loss by taking the reducing effect of easy negative samples and discriminating instances across domains into consideration together. Besides, we introduced a surgical tool tracking dataset called STT with bounding box as the ground-truth annotation. Our experiments demonstrate that our tracking method has a better performance against other state-of-the-art real-time trackers on our dataset. Especially, our tracker has almost 2% higher success rate and precision than ECO-HC which has the highest tracking speed, and has relatively better performance in most challenging cases. However, the speed of our method decreases a lot when the input resolution increases twice larger in our dataset compared with OTB due to multi-convolutions with high-resolution images and improved RoIAlign technique to enhance the representation quality of RoIs. In the future, we will try to solve this problem by modifying the network architecture to reduce computational cost. Furthermore, 3D translation information will be estimated based on endoscope calibration and 2D tracking with bounding boxes, and 3D orientation estimation will also be obtained by utilising the technique provided by Sundermeyer *et al.* [5]. Then, 6D pose estimation will be implemented to enhance surgical navigation.

6 References

[1] Chmarra M.K., Grimbergen C., Dankelman J.: 'Systems for tracking minimally invasive surgical instruments', *Minim Invasive Ther. Allied Technol.*, 2007, **16**, (6), pp. 328–340

[2] Brown A.J., Uneri A., De Silva T.S., ET AL.: 'Design and validation of an open-source library of dynamic reference frames for research and education in optical tracking', *J. Med. Imaging*, 2018, **5**, (2), p. 021215

[3] Zhang L., Ye M., Chan P.-L., ET AL.: 'Real-time surgical tool tracking and pose estimation using a hybrid cylindrical marker', *Int. J. Comput. Assist. Radiol. Surg.*, 2017, **12**, (6), pp. 921–930

[4] Qiu L., Ren H.: 'Endoscope navigation and 3D reconstruction of oral cavity by visual SLAM with mitigated data scarcity'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops, Salt Lake City, Utah, USA, 2018, pp. 2197–2204

[5] Sundermeyer M., Marton Z.-C., Durner M., ET AL.: 'Implicit 3D orientation learning for 6D object detection from RGB images'. Proc. of the European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 699–715

[6] Kurmann T., Neila P.M., Du X., ET AL.: 'Simultaneous recognition and pose estimation of instruments in minimally invasive surgery'. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Quebec, Canada, 2017, pp. 505–513

[7] Laina I., Rieke N., Rupprecht C., ET AL.: 'Concurrent segmentation and localization for tracking of surgical instruments'. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Quebec, Canada, 2017, pp. 664–672

[8] Jin A., Yeung S., Jopling J., ET AL.: 'Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks'. 2018 IEEE Winter Conf. on Applications of Computer Vision (WACV), Lake Tahoe, NV/CA, USA, 2018, pp. 691–699

[9] Gao Y., Vedula S.S., Reiley C.E., ET AL.: 'JHU-ISI gesture and skill assessment working set (JIGSAWS): a surgical activity dataset for human motion modeling'. MICCAI Workshop: M2CAI, Boston, MA, USA, 2014, vol. 3, p. 3

[10] 'Workshop and challenges on modeling and monitoring of computer assisted interventions'. Available at http://camma.u-strasbg.fr/m2cai2016/

[11] 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 Trans. Med. Imaging*, 2017, **36**, (7), pp. 1542–1549

[12] Wu Y., Lim J., Yang M.-H.: 'Online object tracking: a benchmark'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Portland, Oregon, USA, 2013, pp. 2411–2418

[13] Jung I., Son J., Baek M., ET AL.: 'Real-time MDNet'. Proc. of the European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 83–98

[14] He K., Gkioxari G., Dollár P., ET AL.: 'Mask R-CNN'. Proc. of the IEEE int. Conf. on Computer Vision, Venice, Italy, 2017, pp. 2961–2969

[15] Nam H., Han B.: 'Learning multi-domain convolutional neural networks for visual tracking'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, 2016, pp. 4293–4302

[16] Lin T.-Y., Goyal P., Girshick R., *et al.*: 'Focal loss for dense object detection'. Proc. of the IEEE Int. Conf. on Computer Vision, Venice, Italy, 2017, pp. 2980–2988

[17] Felzenszwalb P.F., Girshick R.B., McAllester D., *et al.*: 'Object detection with discriminatively trained part-based models', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, **32**, (9), pp. 1627–1645

[18] Chatfield K., Simonyan K., Vedaldi A., *et al.*: 'Return of the devil in the details: delving deep into convolutional nets', arXiv:1405.3531, 2014

[19] Russakovsky O., Deng J., Su H., *et al.*: 'Imagenet large scale visual recognition challenge', *Int. J. Comput. Vis.*, 2015, **115**, (3), pp. 211–252

[20] Bertinetto L., Valmadre J., Henriques J.F., *et al.*: 'Fully-convolutional siamese networks for object tracking'. European Conf. on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 850–865

[21] Danelljan M., Häger G., Khan F.S., *et al.*: 'Discriminative scale space tracking', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **39**, (8), pp. 1561–1575

[22] Kiani Galoogahi H., Fagg A., Lucey S.: 'Learning background-aware correlation filters for visual tracking'. Proc. of the IEEE Int. Conf. on Computer Vision, Venice, Italy, 2017, pp. 1135–1143

[23] Danelljan M., Bhat G., Shahbaz Khan F., *et al.*: 'Eco: efficient convolution operators for tracking'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Hawaii, USA, 2017, pp. 6638–6646

doi: 10.1049/htl.2019.0068