Scalable joint detection and segmentation of surgical instruments with weak supervision

Ricardo Sanchez-Matilla[1], Maria Robu[1],
Imanol Luengo[1], and Danail Stoyanov[1,2]

[1] Digital Surgery, a Medtronic company, London, UK
[2] Wellcome / EPSRC Centre for Interventional and Surgical Sciences,
University College London, London, UK

Abstract. Computer vision based models, such as object segmentation, detection and tracking, have the potential to assist surgeons intraoperatively and improve the quality and outcomes of minimally invasive surgery. Different work streams towards instrument detection include segmentation, bounding box localisation and classification. While segmentation models offer much more granular results, bounding box annotations are easier to annotate at scale. To leverage the granularity of segmentation approaches with the scalability of bounding box-based models, a multi-task model for joint bounding box detection and segmentation of surgical instruments is proposed. The model consists of a shared backbone and three independent heads for the tasks of classification, bounding box regression, and segmentation. Using adaptive losses together with simple yet effective weakly-supervised label inference, the proposed model use weak labels to learn to segment surgical instruments with a fraction of the dataset requiring segmentation masks. Results suggest that instrument detection and segmentation tasks share intrinsic challenges and jointly learning from both reduces the burden of annotating masks at scale. Experimental validation shows that the proposed model obtain comparable results to that of single-task state-of-the-art detector and segmentation models, while only requiring a fraction of the dataset to be annotated with masks. Specifically, the proposed model obtained 0.81 weighted average precision (wAP) and 0.73 mean intersection-over-union (IOU) in the Endovis2018 dataset with 1% annotated masks, while performing joint detection and segmentation at more than 20 frames per second.

Keywords: Instrument detection · instrument segmentation · multi-task learning · semi-supervised learning.

1 Introduction

Detection of surgical instruments in minimally invasive surgery video frames allows automatic generation of offline surgical analytics, that can provide valuable information for improving surgical procedures [1]. Additionally, surgical instrument detection can provide real-time decision support during the surgery and notify preventable risks during computer assisted interventions [2].

Accurate models are required to successfully use decision support systems during surgical procedures. Current machine learning approaches typically estimate the location and type of surgical instruments via either bounding boxes *detection* [3] [4] or semantic *segmentation* [5] [6]. Tool detection models generally rely on annotated bounding boxes during training. This has a major limitation for instrument detection as the annotated bounding boxes include a high number of background pixels due to the elongated dimensions of the surgical instruments, which might impede a model from learning discriminative features of the instruments. Alternatively, segmentation models directly estimate the probability of each pixel to belong to a specific instrument type by relying on fine-grained pixel-wise segmentation mask annotations. While masks solve the aforementioned challenge faced by bounding boxes, the annotation cost significantly grows up to almost two orders of magnitude for annotating masks with respect to only annotating frame-level labels or bounding boxes [7]. In practice, the annotation of datasets with masks at scale could be unfeasible, which can prevent models from achieving the generalisation and robustness required to be applied in real-world applications.

To address some of the challenges above and leverage the strengths of both workstreams, a multi-task model is proposed that jointly learns to estimate bounding boxes and masks for surgical instruments. The model aggregates information from the multiple tasks by using a shared backbone as encoder, while having a head for each task: instrument classification, bounding box regression and segmentation. While the classification and regression heads allow to localise and classify surgical instruments using scalable annotations, the segmentation head achieves the detailed pixel-wise annotations. To alleviate the burden of expensive pixel-wise annotation on large datasets, we introduce a training framework that accounts for missing masks and uses a weakly-supervised loss computed on frame-level labels which can be freely obtained from the bounding box annotations. Experimental results show that our model achieves detection and segmentation performance on par with fully supervised alternatives, while requiring as little as 1% of the masks in training.

2 Related work

Surgical tool identification and localisation is an active research field, which has resulted in multi-centre collaborations releasing novel surgical datasets to encourage the research community to design models to advance segmentation quality [8], model robustness and generalisation [9].

Proposed research directions to tackle surgical tool identification and localisation include semantic segmentation and tool detection. Segmentation models are able to segment instruments against background (binary segmentation) [10], tool types (semantic segmentation) [6] or tools instances (instance segmentation) [5]. Most relevant, Gonzalez et al. [5] proposed to segment entire instruments instances instead of pixel-wise segmentation to achieve state-of-the-art results. A comprehensive overview of the latest segmentation models is available in [9].

Recent detection models [3] [4] are anchor-based and are composed of a convolutional backbone with ResNet architecture, that generates feature maps at different scales, and two task-specific heads that perform object classification and bounding box regression from the feature pyramid. This approach faces an extreme foreground-background class imbalance during training. This can be handled by using the focal loss [3], a variation of the cross-entropy loss function that down-weights the loss assigned to well-classified examples. EfficientDet [4] proposed to jointly scale up model width, depth and resolution to meet real-time requirements without sacrificing detection accuracy. The model computes a feature pyramid using EfficientNet [11] as backbone. EfficientDet proposed a weighted bi-directional feature pyramid network (BiFPN) to efficiently leverage the multi-scale feature information for object detection.

More complex approaches have also been proposed. Joint detection and segmentation can be learnt jointly [12]. A semi-supervised object segmentation model can rely on a single manual bounding box initialisation to produce class-agnostic object masks and rotated bounding boxes with a fully-convolutional siamese model [13]. Multi-task models using weak supervision can jointly detect and segment with a weakly-supervised cyclic policy can be used to complement the learning of both tasks simultaneously [14]. In the same line of work, other works show that a weakly-supervised convolutional model can be used to estimate the presence and localisation of surgical instruments using only frame-label annotations [15], [16]. However, the performance of these weakly-supervised models is still far from that of the fully supervised ones.

3 Proposed model

3.1 Joint detection and segmentation

Let $\mathbf{x} \in \{0, 255\}^{W,H,C}$ be an RGB image with width W, height H and $C = 3$ colour channels. Let $D(\cdot) : \mathbf{x} \rightarrow (\mathbb{B}^{N,4}, \mathbb{C}^{N}, \mathbb{M}^{W,H,M})$ be a joint detection and segmentation model that localises and classifies surgical instruments within $\mathbf{x}$ which outputs are a set of bounding boxes ($\mathbb{B}^{N,4}$), their corresponding estimated classes ($\mathbb{C}^{N}$), and a segmentation mask ($\mathbb{M}^{W,H,M}$), with N being the number of detected instruments and M the total number instrument types in the dataset.

The problem is formulated as a multi-task learning problem. The proposed model, depicted in Fig. 1, is composed of three main components, namely, backbone, feature fusion module, and three heads - one for each task: localisation, classification, and segmentation. The shared backbone acts as a joint representation learning module whose aim is to learn multi-level feature representations suitable for all the tasks. Having $\mathbf{x}$ as input, the backbone $\beta(\cdot) : \mathbf{x} \rightarrow \mathbb{P}$ generates a pyramid of features at S scales $\mathbb{P} = (p_s)_{s=1}^{s=S}$. The feature pyramid is fed to a bi-directional feature pyramid network (BiFPN) [4] that fuses the features across scales while maintaining their number and resolution $\gamma(\cdot) : \mathbb{P} \rightarrow \hat{\mathbb{P}}$. The heads guide the learning of the backbone and the feature fusion modules to learn more discriminative and complementary features to further improve the three

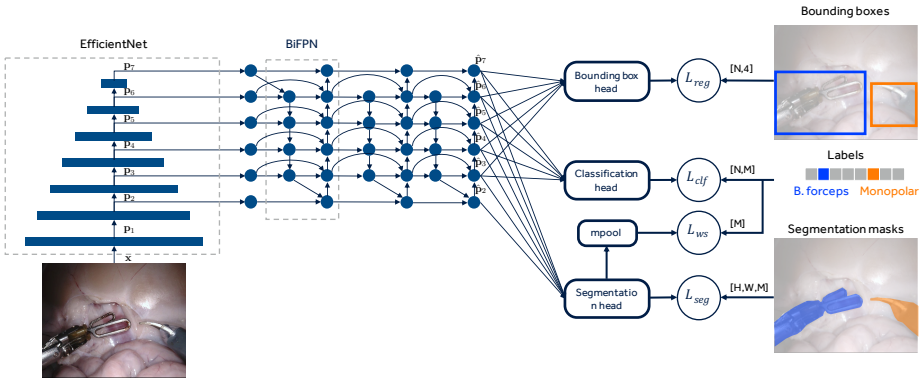

Fig. 1. Proposed multi-task model composed of an EfficientNet backbone, a set of bi-directional feature pyramid network (BiFPN), and the proposed three heads for the tasks of bounding box regression, bounding box classification, and segmentation. The model is trained using four loss functions named regression L_{reg}, classification L_{clf}, weak supervision L_{ws} and segmentation L_{seg}. The model requires bounding box and label annotations for every frame, and segmentation masks for a reduced number of frames. *mpool* is the global maxpool operation. The text on the most-right arrows indicates the shape of the annotations, where N is the number of instruments in a given frame, M is the total number instrument types in the dataset, and W, H are the dimensions of the input frame.

tasks while adapting the generated features for task-specific problems. In our implementation, we use the localisation and classification heads proposed in [4].

The following subsections describe the proposed segmentation head as well as the mechanism that allows learning to segment with only a fraction of annotated masks via weak supervision.

3.2 Segmentation head

The segmentation head aims to generate a mask $\mathbb{M}^{W,H,M}$ from the fused feature pyramid $\hat{\mathbb{P}}$. The segmentation head architecture (Fig. 2) is composed of three main steps: feature upsampling and concatenation, convolutional block, and upsampling.

To make use of the information contained in the multiple scales of the fused feature pyramid, $\hat{\mathbb{P}}$, the $S-2$ smallest feature maps $(\hat{p}_s)_{s=2}^{s=S}$ are first upsampled to the resolution of $\hat{p}_2$ using bi-linear interpolation. Then, the $S-1$ feature maps are concatenated

$$\tilde{\mathbb{P}} = (\hat{p}_2, U_2(\hat{p}_3), U_2(\hat{p}_4), U_2(\hat{p}_5), U_2(\hat{p}_6), U_2(\hat{p}_7)) \tag{1}$$

where $U_2(\cdot)$ is the bilinear interpolation operation that upsamples the feature map to the resolution of $\hat{p}_2$, and $(\cdot, \ldots, \cdot)$ represents the concatenation operation. A convolutional block is then applied to achieve a feature map with M channels

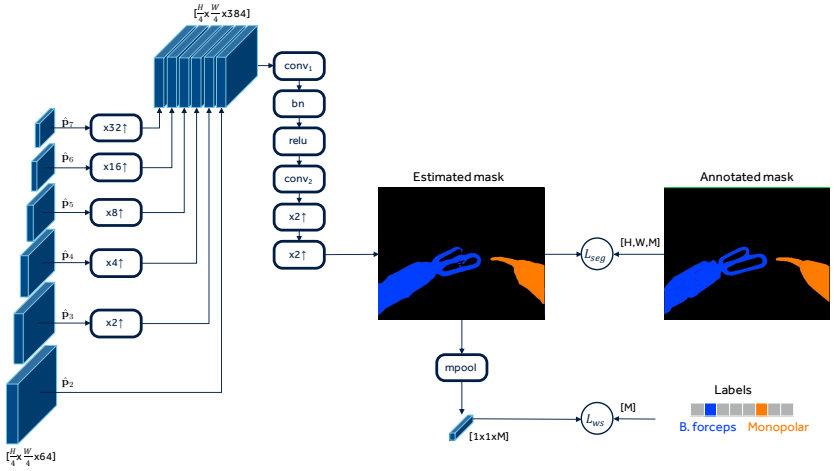

Fig. 2. Diagram of the segmentation head with the proposed weak-supervision module. L_{seg} is the cross entropy segmentation loss that is computed for a given frame when the mask is annotated. L_{ws} is the cross entropy weak supervision loss computed at every frame. *mpool* is the global maxpool operation. The text on the most-right arrows indicates the shape of the annotations, where W, H are the dimensions of the input frame, and M is the total number instrument types in the dataset.

as

$$\tilde{\mathbb{M}} = conv_2(relu(bn(conv_1(\tilde{\mathbb{P}})))), \tag{2}$$

where $conv_1(\cdot)$ is a 2D convolution with kernel (1x1) and S − 1x64 channels that fuses the features with different resolutions, $bn(\cdot)$ is a batch normalisation layer, $relu(\cdot)$ is the Rectified Linear Unit (ReLU) operation, and $conv_2(\cdot)$ is a 2D convolution with kernel (1x1) and M channels to reduce the number of channels to the number of instrument types in the dataset, M.

Finally, $\tilde{\mathbb{M}}$ is upsampled to generate masks with same dimensions than the input images

$$\mathbb{M} = U_0(\tilde{\mathbb{M}}), \tag{3}$$

where $U_0(\cdot)$ is the bilinear interpolation operation that upsamples the feature map to the resolution of the input image $\mathbf{x}$.

3.3 Semi-supervised learning with weak supervision

When the annotated mask, $\bar{\mathbb{M}}$, is available for a given frame, we use the cross-entropy loss function ($L_{CE}(\cdot,\cdot)$) for training the segmentation head. However, as not all samples will have an annotated mask, in each batch we weight the cross-entropy loss by the ratio of samples with annotated masks A to the total number of samples within the batch B as

$$L_{seg} = \frac{A}{B} L_{CE}(\mathbb{M}, \bar{\mathbb{M}}). \tag{4}$$

Thus, batches with fewer annotated masks have a lower weight.

In addition, we enable the training of the segmentation head in the absence of annotated masks. We reduce the estimated mask with a global max pooling into a vector that is supervised with frame-level annotations (presence of the instruments in each frame) [15] as

$$\mathbb{O} = mpool(\mathbb{M}), \tag{5}$$

where $mpool(\cdot)$ is the 2D maxpool operation with kernel size (H, W) that generates a vector $\mathbb{O} \in \mathbb{R}^{1,1,M}$. The information within $\mathbb{O}$ estimates the presence/absence of each instrument type within the frame. These outputs indicate the presence of a given instrument type within the frame. Note that these annotations, which are cheap to generate, are already available to the model within the bounding box annotations.

The weakly-supervised loss is the cross entropy between $\mathbb{O}$ and the instrument type frame-level multi-label annotations, $\bar{\mathbb{O}}$ as

$$L_{ws} = L_{CE}(\mathbb{O}, \bar{\mathbb{O}}). \tag{6}$$

Note that we compute $L_{ws}(\cdot)$ for all frames, regardless of whether their mask is provided or not.

In conclusion, the full loss used to train the backbone, BiFPN, and heads is

$$L = w_{reg} \cdot L_{reg} + w_{clf} \cdot L_{clf} + w_{seg} \cdot L_{seg} + w_{ws} \cdot L_{ws}, \tag{7}$$

where (L_{reg}, L_{clf}) is the focal loss [3], and w_{reg}, w_{clf}, w_{seg}, and w_{ws} are weights of regression, classification, segmentation and weak supervision losses that tune the contribution of each loss.

4 Experimental setup

4.1 Dataset

We validate the performance of the proposed model in EndoVis2018 dataset released as part of the Robotic Scene Segmentation Challenge [17]. The dataset is composed of 15 sequences of 149 frames each. We use the annotated masks of the instrument type provided by [5]. We automatically generate bounding boxes from the masks by selecting the minimum and maximum values for the vertical and horizontal coordinates of each mask. We split the data in training and validation sets as done by [5]. Sequences 5, 9, and 15 compose the validation set and the remaining ones the training set. The sequence *seq2* is discarded from either of the sets as it contains frames with two instances of the same instrument type, and therefore, we cannot automatically generate bounding boxes.

4.2 Performance metrics

We evaluate detection using the mean Average Precision weighted (wAP) by the number of samples in the validation set. As proposed by [5], segmentation is evaluated using IOU averaged over the number of classes, M, and over the number of images, K:

$$IOU = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{M}\sum_{m=1}^{M}\frac{\mathbb{M}_{k,m}\cap\bar{\mathbb{M}}_{k,m}}{\mathbb{M}_{k,m}\cup\bar{\mathbb{M}}_{k,m}}\right). \tag{8}$$

4.3 Implementation details

As backbone, we use EfficientNet-D0 [11] pre-trained on ImageNet. We modify the BiFPN layer [4] to aggregate $S = 6$ feature scales instead of five for improved segmentation accuracy. The five smallest scales are used for both the regression and classification heads. Images are downscaled to 512x512 pixels and data augmentation that includes geometrical and colour transformations is used. A sampler to balance the number of different instruments types in each epoch is used. All models are trained for 150 epochs. We report the results on the validation set obtained in the last epoch. SGD optimiser with momentum and *1Cycle* learning scheduler [18] with cosine decay and a maximum learning rate of $5e^{-4}$ is used. Each batch contains 32 samples. The proposed loss (Eq. 7) weights are empirically set to: $w_{reg} = 1$, $w_{clf} = 5$, $w_{seg} = 700$, and $w_{ws} = 5$. These parameters encourage all losses to be in a similar range. These settings remain fixed for all the experiments.

4.4 Results

Ablation study. We first study how the proposed joint detection and segmentation model compares against the only-detection and only-segmentation alternatives. Results in the supplementary material indicate that performing detection and segmentation jointly slightly improves detection while maintaining similar segmentation performance when 100% of the masks are available. However, the segmentation-only model performance rapidly degrades when fewer masks are available during training. For instance, the performance drops from an IOU of 0.821 to 0.651 when reducing the masks from 100% to 20%, and to 0.544 when further reduced to 1%. Secondly, we perform an ablation study to understand how the presence/absence of the weakly supervised loss impacts the performance with limited availability of annotated segmentation masks. Results in the supplementary material show that when the weakly supervised loss is present the performance approximately remains stable, even when reducing the number of masks up to 1%. For instance, the performance between using 100% or 1% of the masks varies in 0.01, and 0.09 points for wAP and IOU, respectively.

Comparison against state of the art. The proposed model obtain competitive segmentation results against fully-supervised state-of-the-art alternatives while only requiring a 1% of annotated masks (Table 1). In terms of

Table 1. Comparison of the proposed model against state-of-the-art detection and segmentation models. The proposed model performance is evaluated for a range of different availability of masks during training. When available, mean and standard deviation (mean±std) over three trainings with different random seeds are reported. KEY: *, sequence seq2 is used in the validation set.

Model	Task		Weak superv.	% annotated masks	Detection	Segmentation
	Det.	Segm.			wAP	IOU
[4] EfficientDet	✓			-	0.808±0.008	-
[19] TernausNet*		✓		100%	-	0.399
[20] MFTAPnet*		✓		100%	-	0.391
[5] ISINet*		✓		100%	-	0.710
[6] HRNet*		✓		100%	-	0.714±0.019
[6] HRNet		✓		100%	-	0.738
Proposed	✓	✓	✓	100%	0.827±0.007	0.822±0.015
Proposed	✓	✓	✓	20%	0.817±0.003	0.800±0.014
Proposed	✓	✓	✓	5%	0.808±0.026	0.791±0.022
Proposed	✓	✓	✓	1%	0.813±0.016	0.728±0.006
Proposed	✓	✓	✓	0%	0.786±0.021	0.345±0.012

detection, the proposed model outperforms by up to 3% with respect to the detection-only model. We also study how reducing the number of segmentation masks available during training (100%, 20%, 10%, 5%, 1%, and 0% of the total number of training samples) affects the detection and segmentation performance of the proposed model. During the different training setups for different mask ratios, frames with masks are sampled equally spaced across the dataset. Results show that the proposed model outperforms the rest of the alternatives while only requiring 5% of masks. Even with only 1% of the masks available, the proposed model obtain competitive results when compared with fully-supervised alternatives. When no masks are available (0%), the model solely relies on the weakly-supervised module for learning to segment, and both detection and segmentation performance significantly drops (see the last column of Fig. 3). Three visual segmentation samples, one per each sequence of the validation set, are displayed in Fig. 3 for models trained using 100%, 20%, 5%, and 0% of annotated masks. The estimated masks maintain the quality even when the available masks are reduced to 5%. Some classification errors are observed in the second sequence when limited masks are used. When no masks are used during training (0%) the estimated masks tend to only focus on representative parts of the instrument. Additional visual examples are available in the supplementary material.

The proposed model has 4.01M parameters and can perform detection and segmentation simultaneously while requiring only 5% and 1% more parameters than the only-detection and only-segmentation models, respectively. The proposed model obtains an inference speed of 22.4fps in an NVIDIA Quadro RTX 6000.

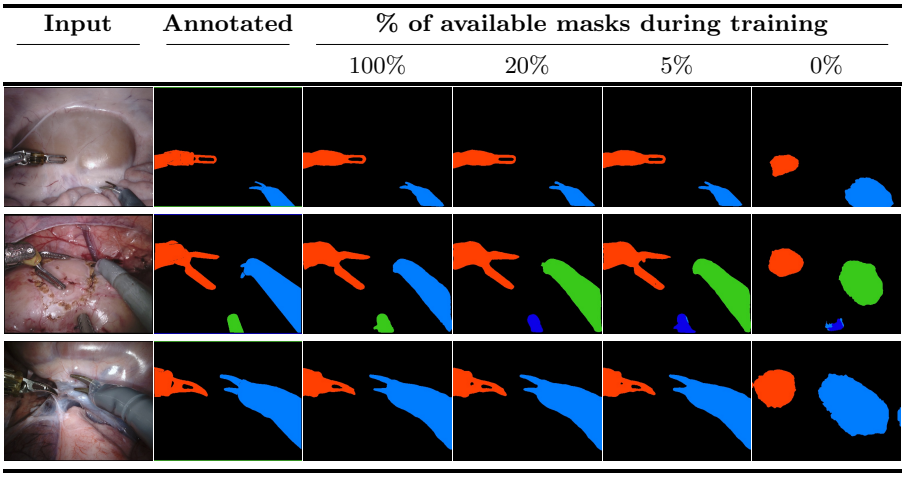

Fig. 3. Visual segmentation results of the proposed model with different percentage of available annotated masks during training. The colours encode the instrument type.

5 Conclusion

This work[3] proposed a multi-task model that jointly learns to detect and segment surgical instruments. A weakly-supervised adaptive loss is also proposed, that enables the learning of segmentation masks when only a fraction of masks are available during training by supervising the learning with frame-level annotations. Experimental results showed that the proposed model obtains comparable results to a fully-supervised alternative, while only requiring a 1% of the frames to have annotated masks.

Further investigation is required to understand how to effectively add temporal information and consistency to the model as well as how to further interrelate the learning of the multiple tasks.

References

1. A. Trehan, A. Barnett-Vanes, M. J. Carty, P. McCulloch, and M. Maruthappu, "The impact of feedback of intraoperative technical performance in surgery: a systematic review," *BMJ Open*, vol. 5, no. 6, 2015.
2. K. Jo, Y. Choi, J. Choi, and J. W. Chung, "Robust real-time detection of laparoscopic instruments in robot surgery using convolutional neural networks with motion vector prediction," *Applied Sciences*, vol. 9, no. 14, p. 2865, 2019.

[3] This work was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) at UCL (203145Z/16/Z), EPSRC (EP/P012841/1, EP/P027938/1, EP/R004080/1) and the H2020 FET (GA 863146). Danail Stoyanov is supported by a Royal Academy of Engineering Chair in Emerging Technologies (CiET18196) and an EPSRC Early Career Research Fellowship (EP/P012841/1).

3. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017.
4. M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.
5. C. González, L. Bravo-Sánchez, and P. Arbelaez, "Isinet: An instance-based approach for surgical instrument segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention* (A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, eds.), pp. 595–605, 2020.
6. K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *CoRR*, vol. abs/1904.04514, 2019.
7. H. Bilen, "Weakly supervised object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
8. M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen, A. Kori, V. Alex, G. Krishnamurthi, D. Rauber, R. Mendel, C. Palm, S. Bano, G. Saibro, C.-S. Shih, H.-A. Chiang, J. Zhuang, J. Yang, V. Iglovikov, A. Dobrenkii, M. Reddiboina, A. Reddy, X. Liu, C. Gao, M. Unberath, M. Kim, C. Kim, C. Kim, H. Kim, G. Lee, I. Ullah, M. Luna, S. H. Park, M. Azizian, D. Stoyanov, L. Maier-Hein, and S. Speidel, "2018 robotic scene segmentation challenge," 2020.
9. T. Ross, A. Reinke, P. M. Full, M. Wagner, H. Kenngott, M. Apitz, H. Hempe, D. M. Filimon, P. Scholz, T. N. Tran, P. Bruno, P. Arbeláez, G.-B. Bian, S. Bodenstedt, J. L. Bolmgren, L. Bravo-Sánchez, H.-B. Chen, C. González, D. Guo, P. Halvorsen, P.-A. Heng, E. Hosgor, Z.-G. Hou, F. Isensee, D. Jha, T. Jiang, Y. Jin, K. Kirtac, S. Kletz, S. Leger, Z. Li, K. H. Maier-Hein, Z.-L. Ni, M. A. Riegler, K. Schoeffmann, R. Shi, S. Speidel, M. Stenzel, I. Twick, G. Wang, J. Wang, L. Wang, L. Wang, Y. Zhang, Y.-J. Zhou, L. Zhu, M. Wiesenfarth, A. Kopp-Schneider, B. P. Müller-Stich, and L. Maier-Hein, "Robust medical instrument segmentation challenge 2019," 2020.
10. L. C. García-Peraza-Herrera, W. Li, C. Gruijthuijsen, A. Devreker, G. Attilakos, J. Deprest, E. Vander Poorten, D. Stoyanov, T. Vercauteren, and S. Ourselin, "Real-time segmentation of non-rigid surgical tools based on deep learning and tracking," in *Computer-Assisted and Robotic Endoscopy* (T. Peters, G.-Z. Yang, N. Navab, K. Mori, X. Luo, T. Reichl, and J. McLeod, eds.), pp. 84–95, 2017.
11. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, pp. 6105–6114, 2019.
12. J. Cao, Y. Pang, and X. Li, "Triply supervised decoder networks for joint detection and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7384–7393, 2019.
13. Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," *CoRR*, vol. abs/1812.05050, 2018.
14. Y. Shen, R. Ji, Y. Wang, Y. Wu, and L. Cao, "Cyclic guidance for weakly supervised joint detection and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

15. A. Vardazaryan, D. Mutter, J. Marescaux, and N. Padoy, "Weakly-supervised learning for tool localization in laparoscopic videos," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis* (D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi, V. Cheplygina, D. Mateus, E. Trucco, E. Granger, and P. Jannin, eds.), pp. 169–179, 2018.
16. C. I. Nwoye, D. Mutter, J. Marescaux, and N. Padoy, "Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos," *International journal of computer assisted radiology and surgery*, vol. 14, no. 6, pp. 1059–1067, 2019.
17. M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen, A. Kori, V. Alex, G. Krishnamurthi, D. Rauber, R. Mendel, C. Palm, S. Bano, G. Saibro, C.-S. Shih, H.-A. Chiang, J. Zhuang, J. Yang, V. Iglovikov, A. Dobrenkii, M. Reddiboina, A. Reddy, X. Liu, C. Gao, M. Unberath, M. Kim, C. Kim, C. Kim, H. Kim, G. Lee, I. Ullah, M. Luna, S. H. Park, M. Azizian, D. Stoyanov, L. Maier-Hein, and S. Speidel, "2018 robotic scene segmentation challenge," 2020.
18. L. N. Smith and N. Topin, "Super-convergence: Very fast training of residual networks using large learning rates," *CoRR*, vol. abs/1708.07120, 2017.
19. A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 624–628, 2018.
20. Y. Jin, K. Cheng, Q. Dou, and P.-A. Heng, "Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video," in *International Conference on Medical Image Computing and Computer Assisted Intervention* (D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, eds.), (Cham), pp. 440–448, Springer International Publishing, 2019.