Medical Image Analysis 70 (2021) 101920

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Challenge Report

Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge

Tobias Roß[a,b,1,*,*], Annika Reinke[a,b,1], Peter M. Full[b,c], Martin Wagner[d], Hannes Kenngott[d], Martin Apitz[d], Hellena Hempe[a], Diana Mindroc-Filimon[a], Patrick Scholz[a,e], Thuy Nuong Tran[a], Pierangela Bruno[a,v], Pablo Arbeláez[f], Gui-Bin Bian[g,h], Sebastian Bodenstedt[i,j,k], Jon Lindström Bolmgren[l], Laura Bravo-Sánchez[f], Hua-Bin Chen[g,h], Cristina González[f], Dong Guo[m], Pål Halvorsen[n,o], Pheng-Ann Heng[p], Enes Hosgor[l], Zeng-Guang Hou[g,h], Fabian Isensee[b,c], Debesh Jha[n,q], Tingting Jiang[r], Yueming Jin[p], Kadir Kirtac[l], Sabrina Kletz[s], Stefan Leger[i,j,k], Zhixuan Li[r], Klaus H. Maier-Hein[c], Zhen-Liang Ni[g,h], Michael A. Riegler[n], Klaus Schoeffmann[s], Ruohua Shi[r], Stefanie Speidel[i,j,k], Michael Stenzel[l], Isabell Twick[l], Gutai Wang[m], Jiacheng Wang[t], Liansheng Wang[t], Lu Wang[m], Yujie Zhang[t], Yan-Jie Zhou[g,h], Lei Zhu[p], Manuel Wiesenfarth[u], Annette Kopp-Schneider[u], Beat P. Müller-Stich[d], Lena Maier-Hein[a]

[a] *Computer Assisted Medical Interventions (CAMI), German Cancer Research Center, Im Neuenheimer Feld 223, 69120, Heidelberg, Germany*
[b] *University of Heidelberg, Germany, Seminarstraße 2, 69117 Heidelberg, Germany*
[c] *Division of Medical Image Computing (MIC), Im Neuenheimer Feld 223, 69120 Heidelberg, Germany*
[d] *Department for General, Visceral and Transplantation Surgery, Heidelberg University Hospital, Im Neuenheimer Feld 110, 69120 Heidelberg, Germany*
[e] *HIDSS4Health – Helmholtz Information and Data Science School for Health, Im Neuenheimer Feld 223, 69120 Heidelberg, Germany*
[f] *Universidad de los Andes, Cra. 1 No 18A - 12, 111711 Bogotá, Colombia*
[g] *University of Chinese Academy Sciences, 52 Sanlihe Rd., Beijing, China*
[h] *State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, 100864 Beijing, China*
[i] *National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany: German Cancer Research Center, Im Neuenheimer Feld 460, 69120 Heidelberg, Germany*
[j] *Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Fetscherstraße 74, 01307 Dresden, Germany*
[k] *Helmholtz Association/Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Bautzner Landstraße 400, 01328 Dresden, Germany*
[l] caresyntax, *Komturstraße 18A, 12099 Berlin, Germany*
[m] *School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Shahe Campus:No.4, Section 2, North Jianshe Road, 610054 | Qingshuihe Campus:No.2006, Xiyuan Ave, West Hi-Tech Zone, 611731, Chengdu, China*
[n] *SimulaMet, Pilestredet 52, 0167 Oslo, Norway*
[o] *Oslo Metropolitan University (OsloMet), Pilestredet 52, 0167 Oslo, Norway*
[p] *Department of Computer Science and Engineering, The Chinese University of Hong Kong, Chung Chi Rd, Ma Liu Shui, Hong Kong, China*
[q] *Department of Informatics, UIT The Arctic University of Norway, Hansine Hansens vei 54, 9037 Tromsø, Norway*
[r] *Institute of Digital Media (NELVT), Peking University, 5 Yiheyuan Rd, Haidian District, 100871 Peking, China*
[s] *Institute of Information Technology, Klagenfurt University, Universitätsstraße 65-67, 9020 Klagenfurt, Austria*
[t] *Department of Computer Science, School of Informatics, Xiamen University, 422 Siming South Road, 361005 Xiamen, China*
[u] *Division of Biostatistics, German Cancer Research Center, Im Neuenheimer Feld 581, Heidelberg, Germany*
[v] *Department of Mathematics and Computer Science, University of Calabria, 87036 Rende, Italy*

ARTICLE INFO

Article history:
Received 20 May 2020
Revised 22 September 2020
Accepted 24 November 2020
Available online 28 November 2020

ABSTRACT

Intraoperative tracking of laparoscopic instruments is often a prerequisite for computer and robotic-assisted interventions. While numerous methods for detecting, segmenting and tracking of medical instruments based on endoscopic video images have been proposed in the literature, key limitations remain to be addressed: Firstly, *robustness*, that is, the reliable performance of state-of-the-art methods

* Corresponding author.
E-mail address: t.ross@dkfz-heidelberg.de (T. Roß).
[1] Contributed equally to this paper.

https://doi.org/10.1016/j.media.2020.101920

Keywords:
Multi-instance instrument
Minimally invasive surgery
Robustness and generalization
Surgical data science

when run on challenging images (e.g. in the presence of blood, smoke or motion artifacts). Secondly, *generalization*; algorithms trained for a specific intervention in a specific hospital should generalize to other interventions or institutions.

In an effort to promote solutions for these limitations, we organized the *Robust Medical Instrument Segmentation (ROBUST-MIS) challenge* as an international benchmarking competition with a specific focus on the robustness and generalization capabilities of algorithms. For the first time in the field of endoscopic image processing, our challenge included a task on binary segmentation and also addressed multi-instance detection and segmentation. The challenge was based on a surgical data set comprising 10,040 annotated images acquired from a total of 30 surgical procedures from three different types of surgery. The validation of the competing methods for the three tasks (binary segmentation, multi-instance detection and multi-instance segmentation) was performed in three different stages with an increasing domain gap between the training and the test data. The results confirm the initial hypothesis, namely that algorithm performance degrades with an increasing domain gap. While the average detection and segmentation quality of the best-performing algorithms is high, future research should concentrate on detection and segmentation of small, crossing, moving and transparent instrument(s) (parts).

1. Introduction

Minimally invasive surgery has become increasingly common over the past years (Siddaiah-Subramanya et al., 2017). However, issues such as limited view, a lack of depth information, haptic feedback and increased difficulty in handling instruments have increased the complexity for the surgeons. Surgical data science applications (Maier-Hein et al., 2017) could help the surgeon to overcome those limitations and to increase patient safety. These applications, e.g. surgical skill assessment (Law et al., 2017; Lin et al., 2019), augmented reality (Wang et al., 2017; Burström et al., 2019), assistance robots (Amini Khoiy et al., 2016; Zhang and Gao, 2020), vision-based force estimation (Su et al., 2018) or depth enhancement (De Paolis and De Luca, 2019), are often based on the segmentation and/or tracking of medical instruments during surgery. Currently, commercial tracking systems usually rely on optical or electromagnetic markers and, therefore, also require additional hardware (Bianchi et al., 2019; Zhou et al., 2019), which are expensive, need extra space and require technical knowledge. Alternatively, with the recent success of deep learning methods in the medical domain (Esteva et al., 2019) and first surgical data science applications (Fawaz et al., 2019; Nguyen et al., 2019), video-only based approaches offer new opportunities to handle difficult image scenarios such as bleeding, light over-/underexposure, smoke and reflections (Bodenstedt et al., 2018). Video-only based approaches offer new opportunities to handle difficult image scenarios such as bleeding, light over-/underexposure, smoke and reflections (García-Peraza-Herrera et al., 2016; Kurmann et al., 2017; Laina et al., 2017; Pakhomov et al., 2019; Zhao et al., 2019). In turn, the tracking information may directly affect the instructions provided to the surgeon to navigate the surgical instruments. Furthermore, unreliable algorithms potentially reduce the acceptance on the part of the surgical team, and thus, the chances for translation into the clinical routine (Panch et al., 2019; Qayyum et al., 2020).

As validation and evaluation of image processing methods is usually performed on the researchers' individual data sets, finding the best algorithm suited for a specific use case is a difficult task. Consequently, reported publication results are often difficult to compare (Ioannidis, 2005; Armstrong et al., 2009). In order to overcome this issue, we can implement *challenges* to find algorithms that work best on specific problems. These international benchmarking competitions aim to assess the performance of several algorithms on the same data set, which enables a fair comparison to be drawn across multiple methods (Maier-Hein et al., 2018; 2019).

One international challenge which takes place on a regular basis is the Endoscopic Vision (EndoVis) Challenge[2]. It hosts sub-challenges with a broad variety of tasks in the field of endoscopic image processing and and has been held annually at the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI) since 2015 (exception: 2016). However, data sets provided for instrument detection/tracking/segmentation in previous EndoVis editions (e.g., (Allan et al., 2019, 2020)) comprised a relatively small number of cases (between ∼500 to ∼4,000) and generally represented best cases scenarios (e.g. with clean views, limited distortions in videos) which did not comprehensively reflect the challenges in real-world clinical applications. Although these competitions enabled primary insights and comparison of the methods, the information gained on robustness and generalization capabilities of methods were limited.

To remedy these issues, we present the Robust Medical Instrument Segmentation (ROBUST-MIS) challenge 2019, which was part of the 4th edition of EndoVis at MICCAI 2019. We introduced a large data set comprising more than 10,000 image frames for instrument segmentation and detection, extracted from daily routine surgeries. The data set contained images which included all types of difficulties and was annotated by medical experts according to a pre-defined labeling protocol and subjected to a quality control process. The challenge addressed methods with a projected application in minimally invasive surgeries, in particular the tracking of medical instruments in the abdomen, with a special focus on the generalizibility and robustness. This was achieved by introducing three stages with increase in difficulty in the test phase. To emphasize the robustness of methods, we used a ranking scheme that specifically measures the worst-case performance of algorithms.

Section 2 outlines the challenge design as a whole, including the data set. The results of the challenge are presented in Section 3 with a discussion following in Section 4. The appendix includes challenge design choices regarding the organization (see Appendix A), the labeling and submission instructions (see Appendix B and Appendix C), the rankings across all stages (see Appendix D) and the complete challenge design document (see Appendix F).

2. Methods

The ROBUST-MIS 2019 challenge was organized as a sub-challenge of the Endoscopic Vision Challenge 2019 at MICCAI 2019

[2] https://endovis.grand-challenge.org/.

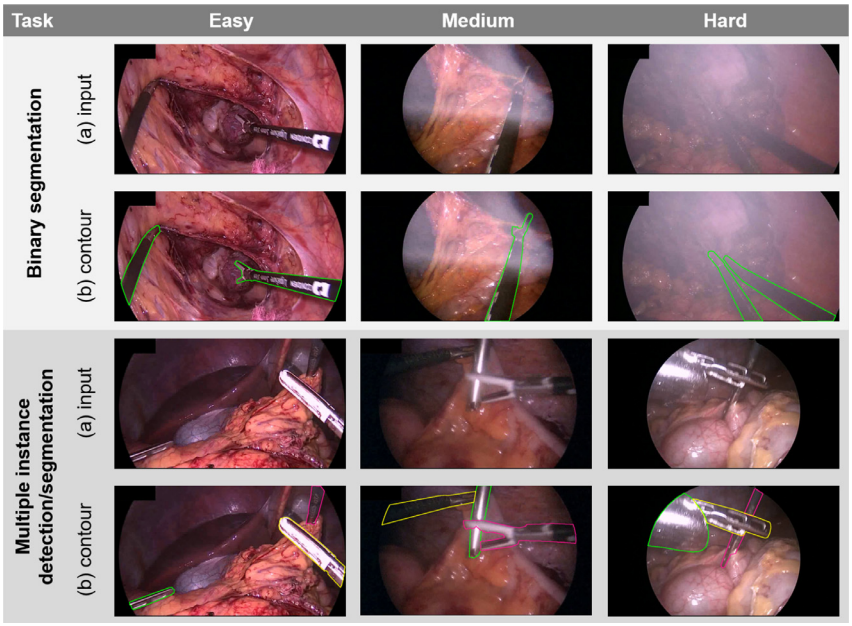


Fig. 1. Various levels of difficulty represented in the challenge data for the binary segmentation (two upper rows) and multi-instance detection/segmentation tasks (two lower rows). Input frames (a) are shown along with the reference segmentation masks for all tasks. The latter are shown as contours (b).

in Shenzhen, China. Details of the challenge organization can be found in Appendix A and Appendix F. The objective of the challenge, the challenge data sets and the assessment method used to evaluate the participating algorithms are presented in the following.

2.1. Mission of the challenge

The goal of the ROBUST-MIS 2019 challenge was to benchmark algorithms designed for instrument detection and segmentation in videos of minimally invasive surgeries. Specifically, we were interested in (1) identifying robust methods for instrument detection and segmentation, (2) assessing the generalization capabilities of the methods proposed and (3) identifying the image properties (e.g. smoke, bleeding, motion artifacts) that make images particularly challenging. The challenges' metrics and ranking schemes were designed to assess these properties (see Section 2.3).

The challenge was divided into three different tasks with separate evaluations and leaderboards (see Fig. 1). For the binary segmentation task, participants had to provide precise contours of instruments, using binary masks, with '1' indicating the presence of a surgical instrument in a given pixel and '0' representing the absence thereof. Analogously, for the multi-instance segmentation task, participants had to provide image masks by allotting numbers '1', '2', etc. which represented different instances of medical instruments. In contrast, the multi-instance detection task merely required participants to detect and roughly locate instrument instances in video frames in which the location could be represented by arbitrary forms, such as bounding boxes.

As detailed in Section 2.3, the generalizability and performance of all participating algorithms was assessed in three stages with increasing levels of difficulty:

- **Stage 1:** Test data was taken from the procedures (patients) from which the training data were extracted.
- **Stage 2:** Test data was taken from the exact same type of surgery as the training data but from procedures (patients) not included in the training
- **Stage 3:** Test data was taken from a different but similar type of surgery (and different patients) compared to the training data.

Before the algorithms were submitted to the challenge, participants were only informed of the surgery types for stages 1 and 2 (rectal resection and proctocolectomy, see Section 2.2.1). For the third stage, the surgery type (sigmoid resection) was referred to as *unknown surgery* to enable the generalizability to be tested.

2.2. Challenge data set

2.2.1. Data recording

All data was recorded with a Karl Storz Image 1 laparoscopic camera (Karl Storz SE & Co. KG, Tuttlingen, Germany), with a 30° optic lens. The Karl Storz Xenon 300 was used as a light source. Data acquisition was executed during daily routine procedures at the Heidelberg University Hospital, Department of Surgery in the integrated operating room (Karl Storz OR1 FUSION®). Whenever parts of the video showed the outside of the abdomen, these frames were manually excluded for the purpose of anonymization. To reduce storage and memory usage, image resolution was reduced from 1920 × 1080 pixels (HD) in the primary video to 960 × 540. Videos from 30 minimally invasive surgical procedures taken in three different types of surgery, namely 10 *rectal resection* procedures, 10 *proctocolectomy* procedures and 10 procedures of *sigmoid resection* procedures, served as a basis for this challenge. A total of 10,040 images were extracted from these 30 procedures according to the procedure summarized in Section 2.2.2.

2.2.2. Data extraction

The frames were selected according to the following procedures: Initially, whenever the camera was outside the abdomen, the corresponding frames were removed to ensure anonymization. Next, all videos were sampled at a rate of 1 frame/sec, eliciting 4,456 extracted frames. To increase this number, additional frames were extracted during the surgical phase transitions, resulting in a total of 10,040 frames. Labels for the surgical phases were available from the previous challenge *EndoVis Surgical Workflow Analysis in the SensorOR*[3]. All of these frames were annotated as described in 2.2.3.

[3] https://endovissub2017-workflow.grand-challenge.org/.

Table 1
Case distribution of the data with frames per stage and surgery. Empty frames (ef) were classed as the % of frames in which an instrument did not appear.

PROCEDURE	TRAINING	TESTING		
		Stage 1	Stage 2	Stage 3
proctocolectomy	2,943 (2% ef.)	325 (11% ef.)	225 (11% ef.)	0
rectal resection	3,040 (20% ef.)	338 (20% ef.)	289 (15% ef.)	0
sigmoid resection*	0	0	0	2,880 (23% ef.)
TOTAL	5,983 (17% ef.)	663 (15% ef.)	514 (13% ef.)	2,880 (23% ef.)

* *unknown surgery*

2.2.3. Label generation

As stated in the introduction, a labeling mask was created for each of the 10,040 extracted endoscopic video frames. The assignment of instances was done per frame, not per video. The instrument labels were generated according to the following procedure: First, the company Understand AI[4] performed initial segmentations on the extracted frames. Following this, the challenge organizers analyzed the annotations, identified inconsistencies and agreed on an annotation protocol (see Appendix B). A team of 14 engineers and four medical students reviewed all of the annotations and, if necessary, refined them according to the annotation protocol. In ambiguous or unclear cases, a team of two engineers and one medical student generated a consensus annotation. For quality control, a medical expert went through all of the refined segmentation masks and reported potential errors. The final decision on the labels was made by a team comprised of a medical expert and an engineer.

2.2.4. Training and test case definition

A training case comprised a 10 second video snippet in the form of 250 endoscopic image frames and a reference annotation for the last frame. For training cases, the entire video was provided as context information along with information on the surgery type. Test cases were identical in format but did not include a reference annotation.

For the division of the data into training and test data, in accordance with the described testing scheme, all sigmoid resection procedures were reserved for stage 3. The two shortest videos per procedure (20%) were selected from the remaining 20 videos for stage 2 in order to have as much training data as possible. Finally, every 10th annotated frame from the remaining 16 videos was used for stage 1 testing. All other frames were released as training data.

No validation cases for hyperparameter tuning were provided by the organizers; hence, it was up to the challenge participants to split the training cases into training and validation data. In summary, this led to a case distribution as shown in Table 1.

2.3. Assessment method

2.3.1. Metrics

The following metrics[5] were used to assess performance:

- Binary Segmentation: Dice Similarity Coefficient (*DSC*) (Dice, 1945) and Normalized Surface Dice (*NSD*)[6] (Nikolov et al., 2018),
- Multi-instance Detection: F1-score (other name for the *DSC*)(Dice, 1945),
- Multi-instance Segmentation: Multi-instance Dice Similarity Coefficient (*MI_DSC*) and multi-instance Normalized Surface Dice (*MI_NSD*).

The *DSC* is a widely used overlap metric for segmentation (Cardoso, 2018; Everingham et al., 2015) and detection challenges (e.g., the Cerebral Aneurysm Detection (CADA)[20]). It is defined as the harmonic mean of precision and recall:

$$DSC(Y,\hat{Y}) := \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|}, \tag{1}$$

where Y denotes the reference annotation and $\hat{Y}$ the corresponding prediction of an image frame.

The *NSD* served as a distance-based measurement for assessing performance. In contrast to the *DSC*, which measures the overlap of volumes, the *NSD* measures the overlap of two surfaces (mask borders) (Nikolov et al., 2018). Furthermore, the metric uses a threshold that is related to the inter-rater variability of the annotators. In our case, the inter-rater variability was computed by a pairwise comparison of a total of 5 annotators over $n = 20$ training images, which resulted in a threshold of $\tau := 13$. Further analysis revealed that thresholds above 10 had no effect on rankings.

According to the challenge design, the indices of instrument instances between the references and predictions did not necessarily match. The only requirement was that each instance was assigned a unique instrument index. Thus, all multi-instance tasks required the prediction and references to be matched, which was computed by applying the Hungarian algorithm (Kuhn, 1955).

To compute the *MI_DSC* and *MI_NSD*, matches of instrument instances were computed. Afterwards, the resulting performance scores for each instrument instance per image have been aggregated by the mean. The choice of the metrics *(MI_)DSC* and *(MI_)NSD* were based on the Medical Segmentation Decathlon challenge (Cardoso, 2018) for the binary segmentation and the multi instance tasks.

Finally, the *F1-score* for the detection task requires the definition of true positives (TP), false negatives (FN) and false positives (FP), where $F1(Y,\hat{Y}) := \frac{2TP}{2TP+FN+FP}$. The assignment of matching candidates was done using the Hungarian algorithm. For this purpose, the intersection over union (*IoU*) was computed for each possible pair of reference and prediction instances, which simply measures the overlap of two areas, divided by their union:

$$IoU(Y,\hat{Y}) := \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|}, \tag{2}$$

where in both cases Y denotes the reference annotation and $\hat{Y}$ the corresponding prediction of an image frame. Similar to the *MI_DSC* computation, the Hungarian algorithm (Kuhn, 1955) was used to assign matching pairs of references and predictions. Assigned pairs of references and predictions $(Y,\hat{Y})$ were defined as TP if their $IoU(Y,\hat{Y}) > \xi := 0.3$. Reference instances without or with a smaller

[4] https://understand.ai.

[5] The implementation of all metrics can be found here: https://phabricator.mitk.org/source/rmis2019/.

[6] https://github.com/deepmind/surface-distance.

[20] https://cada.grand-challenge.org

prediction than ξ were defined as FN. All instances that could not be assigned to a reference instance were assigned to FP.

2.3.2. Rankings

Separate rankings for accuracy and robustness were computed for stage 3 of the challenge in order to address multiple aspects of the challenge purpose. To investigate accuracy, a significance ranking[7] as recently applied in the MSD (Cardoso, 2018) and described in Algorithm 1 was computed. The robustness ranking specifically

Algorithm 1 Ranking scheme for the binary and multi-instance segmentation tasks.

Let $T = \{t_1, \ldots, t_N\}$ be the test cases for the given task.
for all participating algorithms a_i **do**
  Determine the performance $m(a_i, t_j)$ of algorithm a_i for each test case t_j
  if $m(a_i, t_j) == N/A$ **then**
    $m(a_i, t_j) = 0$
  end if
  Aggregate metric values $m(a_i, t_j)$ with the following two aggregation methods:

1. **Accuracy:** Compute the *significance ranking*. For each pair of algorithms, perform one-sided Wilcoxon signed rank tests with a significance level of $\alpha = 0.05$ to assess differences in the metric values. The accuracy rank $r_a(a_i)$ for algorithm~a_i is based on the number of significant test results for each algorithm (Maier-Hein et~al., 2018; Cardoso, 2018).
2. **Robustness:** Compute the *5% percentile* of all $m(a_i, t_j)$ to get the robustness rank $r_r(a_i)$ for algorithm~a_i.

end for

focused on the worst case performance of methods. For this reason, the 5% percentile was computed instead of aggregating metric values with the mean or median. The computation of the *F1-score* naturally included a ranking as the TP, FN, FP were aggregated across all test cases. This led to a global metric value for each participant which was used to create the ranking. Please note both that the number of test cases and the number of algorithms were generally differed for each task and stage. For the binary and multi-instance segmentation tasks, the rankings were computed for both metrics, namely *(MI_)DSC* and *(MI_)NSD*, as shown in Algorithm 1.

These procedures produced nine rankings in total, namely four separate rankings (accuracy and robustness ranking for the *(MI_)DSC* and the *(MI_)NSD*) for the binary and the multi-instance segmentation task respectively and one ranking for multi-instance detection. In every ranking scheme, missing cases were set to the worst possible value, namely 0 for all metrics.

2.3.3. Statistical analyses

The stability of the rankings was investigated via bootstrapping as this approach was identified as appropriate for quantifying ranking variability (Maier-Hein et al., 2018). The analysis was performed using the R package *challengeR* (Wiesenfarth et al., 2019b; 2019a). The package was further used to create plots that visualize (1) the absolute frequency of test cases in which each algorithm achieved the different ranks and (2) the bootstrap results for each algorithm.

[7] Please note that an algorithm A with a higher rank (according to the significance ranking) than algorithm B did not necessarily perform significantly better than algorithm B, as detailed in Wiesenfarth et al. (2019b).

2.3.4. Further analyses

Expert baseline Given the imperfect reference (no perfect ground truth) resulting from human annotation, it is typically difficult to determine a plausible upper bound (optimal) performance. To address this knowledge gap, one additional labeling expert, a medical student with six years of experience in labeling (henceforth denoted 'expert') annotated all images from stage 2. Inspired by a human vs. algorithms analysis for natural image multi-label classification from Shankar et al. (2020), we used the additional data in two principal ways. Firstly, we considered the expert as an additional team and generated new rankings for both the binary and the multi-instance segmentation task using the *(MI_)DSC*. Secondly, we analyzed his performance as a function of the number of instruments present in the image. *Worst case analysis* The influence of the image artifacts and the size and number of instruments were analyzed. For this purpose, the 100 cases with the worst performance were analyzed to investigate which image artifacts cause the main failures of the algorithms.

3. Results

In total, 75 participants registered on the Synapse challenge website (Roß et al., 2019b) before the submission deadline. Aside from one team that decided to be excluded from the rankings, all teams with a working docker[8] submission were included in this paper. Their participation over the three challenge tasks and the total amount of submissions is summarized in Table 2.

3.1. Method descriptions of participating algorithms

In the following, the participating algorithms are briefly summarized based on a description provided by the participants upon submission of the challenge results. Further details can be found in Table 3.

Team caresyntax*: Single network fits all*

The *caresyntax* team's core idea for multi-instance segmentation was to apply a Mask R-CNN (He et al., 2017) based on a single network with shared convolutional layers for both branches. They hypothesized that it would help the network to generalize better if it was only provided with limited training data. The team decided to use a pre-trained version of the Mask R-CNN without including any temporal information from the videos. In their results, they reported that their approach outperformed a U-Net-based model by a significant margin. The team worked out that tuning pixel-level and mask-level confidence thresholds on the predictions played an important role. Furthermore, they acknowledged the importance that the training set size had for improved predictions, both qualitatively and quantitatively. The team participated in all three tasks using the same method. They produced the same output for the multi-instance segmentation and detection tasks and binarized the output of the multi-instance segmentation for the binary segmentation task.

Team CASIA_SRL*: Dense pyramid attention network for robust medical instrument segmentation*

The *CASIA_SRL* team proposed a network named Dense Pyramid Attention Network (Ni et al., 2020) for multi-instance segmentation. They mainly focused on two problems: Changes in illumination and surgical instruments scale changes. They proposed that an attention module should be used, which was able to capture second-order statistics, with the goal of covering semantic dependencies between pixels and capturing the global context

[8] https://www.docker.com/.

Table 2
Overview of selected participating teams over the three tasks, namely binary segmentation (BS), multi-instance detection (MID) and multi-instance segmentation (MIS).

Team identifier	BS	MID	MIS	Affiliations
caresyntax	x	x	x	[1] caresyntax, Berlin, Germany
CASIA_SRL	x		x	[1] University of Chinese Academy Sciences, Beijing, China [2] State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
Djh	x			[1] SimulaMet, Oslo, Norway [2] Arctic University of Norway (UiT), Tromsø, Norway [3] Oslo Metropolitan University (Oslomet), Oslo, Norway
fisensee	x	x	x	[1] University of Heidelberg, Germany [2] Division of Medical Image Computing (MIC), German Cancer Research Center, Heidelberg, Germany
haoyun	x			[1] Department of Computer Science, School of Informatics, Xiamen University, Xiamen, China and School of Mechanical [2] Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
NCT	x			[1] National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, German [2] Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany [3] Helmholtz Association/Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany
SQUASH	x	x	x	[1] Institute of Information Technology, Klagenfurt University, Austria
Uniandes	x	x	x	[1] Universidad de los Andes, Bogotá, Colombia
VIE	x	x	x	[1] Institute of Digital Media (NELVT), Peking University, Peking, China
www	x	x	x	[1] Department of Computer Science, School of Informatics, Xiamen University, Xiamen, China [2] Department of Computer Science Engineering, The Chinese University of Hong Kong, Hong Kong, China
valid submissions	10	6	7	
invalid submissions	2	1	1	
TOTAL	12	7	8	

(Ni et al., 2020). As the scale of surgical instruments constantly changes as they move, the team introduced dense connections across scales to capture multi-scale features for surgical instruments. The team did not use the provided videos to complement the information contained in the individual frames. The team participated in the binary and multi-instance segmentation tasks. They produced the same output for the multi-instance segmentation and detection tasks and binarized the output of the multi-instance segmentation for the binary segmentation task.

Team Djh*: A RASNet-based deep learning approach for the binary segmentation task*

The *Djh* team only participated in the binary segmentation task. They used the Refined Attention Segmentation Network (Ni et al., 2019) and put a large amount of effort into data augmentation and hyperparameter tuning. Their motivation for using this architecture was its U-shape design which consists of contracting and expanding paths like the ResUNet++ (Jha et al., 2019). The RASNet is able to capture low-level and higher-level features. The team did not use the videos provided to complement the information contained in the individual frames.

Team fisensee*: OR-UNet*

Team *fisensee's* core idea was to optimize a binary segmentation algorithm and then adjust the output with a connected component analysis in order to solve the multi-instance segmentation and detection tasks (Isensee and Maier-Hein, 2020). Inspired by the recent successes of the nnU-Net (Isensee et al., 2018), the authors used a simple established baseline architecture (the U-Net (Ronneberger et al., 2015)) and iteratively improved the segmentation results through hyperparameter tuning. The method, referred to as optimized robust residual 2D U-Net (OR-UNet), was trained with the sum of *DSC* and cross-entropy loss and a multi-scale loss. During training, extensive data augmentation was used to increase robustness. For the final prediction, they used an ensemble of eight models. They hypothesized that ensembles perform better than a single network. In their report, the team wrote that they attempted to use the temporal information by stacking previous frames but did not observe a performance gain. Additionally, they noticed that in many cases, instruments did not touch thus they used a connected component analysis (Shapiro, 1996) to separate instrument instances.

Team haoyun*: Robust medical instrument segmentation using enhanced DeepLabV3+*

The *haoyun* team only participated in the binary segmentation task. They based their work on the DeepLabV3+ (Chen et al., 2018) architecture in order to focus on high-level information. To enrich the receptive fields, they used a pre-trained ResNet-101 (He et al., 2016) with dilated convolutions as encoder. To train their network, the team combined the *DSC* with the focal loss (Lin et al., 2017) in order to focus more on less accurate pixels and challenging images. In addition, the team used a 5-fold cross validation to improve both generalization and stability of the network. They did not use the provided videos to complement the information contained in the individual frames.

Team NCT*: Robust medical instrument segmentation in robot-assisted surgery using deep convolutional neuronal network*

The *NCT* team only participated in the binary segmentation task. They used a TernausNet with a pre-trained VGG16 network (Iglovikov and Shvets, 2018) as TernausNet had already showed promising results in two previous MICCAI EndoVis segmentation challenges from 2017 and 2018 (Allan et al., 2019). The team did not use the provided videos to complement the information contained in the individual frames.

Table 3
Overview of submitted methods. Abbreviations are as follows: Stochastic gradient descent (SGD) (Kiefer et al., 1952), adaptive moment estimation (Adam) (Kingma and Ba, 2014).

Team	Basic architecture	Video data used?	Additional data used?	Loss functions	Data augmentation	Optimizer
caresyntax	Mask R-CNN (He et al., 2017) (backbone: ResNet-50 (He et al., 2016))	No	ResNet-50 pre-trained on MS-COCO (Lin et al., 2014)	Smooth L1 loss, cross entropy loss, binary cross entropy loss	Applied in each epoch: Random flip (horizontally) with probability 0.5	SGD (Kiefer et al., 1952)
CASIA_SRL	Dence Pyramid Attention Network (Ni et al., 2020) (backbone: ResNet-34 (He et al., 2016))	No	ResNet-34 backbone pre-trained on ImageNet (Russakovsky et al., 2015)	Hybrid loss: cross entropy $-\alpha \log(Jaccard)$	Data augmented once before training: Random rotation, shifting, flipping	Adam (Kingma and Ba, 2014)
Djh	RASNet (Ni et al., 2019)	No	ResNet50 (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015)	*DSC* coefficient loss	Applied on the fly on each batch: Crop (random and center), flip (horizontally and vertically), scale, cutout, greyscale	Adam (Kingma and Ba, 2014)
fisensee	2D U-Net (Ronneberger et al., 2015) with residual encoder	No	No	Sum of *DSC* and cross-entropy loss	Randomly applied on the fly on each batch: Rotation, elastic deformation, scaling, mirroring, Gaussian noise, brightness, contrast, gamma	SGD (Kiefer et al., 1952)
haoyun	DeepLabV3+ (Chen et al., 2018) with ResNet-101 (He et al., 2016)) encoder	No	ResNet-101 pre-trained on ImageNet (Russakovsky et al., 2015)	Logarithmic *DSC* loss	Applied on the fly on each batch: Flip (vertically), crop (random)	Adam (Kingma and Ba, 2014)
NCT	TernausNet (Iglovikov and Shvets, 2018), replaced ReLU with eLU (Clevert et al., 2015)	No	VGG16 pre-trained on ImageNet (Russakovsky et al., 2015)	Weighted binary cross entropy in combination with Jaccard Index	Applied on the fly on each batch: Flips (horizontally and vertically), rotations of $[-10, 10]^\circ$, image contrast manipulations (brightness, blur, motion-blur)	Adam (Kingma and Ba, 2014)
SQUASH	Mask R-CNN (He et al., 2017) (backbone: ResNet-50 (He et al., 2016))	*Yes, to* estimate the probability that last frame of video shows instrument instance	No	ResNet-50: Focal loss, Mask R-CNN: Mask R-CNN loss + cross entropy loss	35% of total input for classification: Gaussian blur, sharpening, gamma contrast enhancement; additional 35% of images: Mirroring (along x- and y-axes); minority class: Translation (horizontally); non-instrument image frames are not processed	SGD (Kiefer et al., 1952)
Uniandes	Mask R-CNN (He et al., 2017) (backbone: ResNet-101 (He et al., 2016))	Yes, for data augmentation	Pre-trained on MS-COCO (Lin et al., 2014)	Standard Mask R-CNN loss functions	Applied on the fly on each batch: Random flips (horizontally), propagation of annotation backwards to previous video frames	SGD (Kiefer et al., 1952)
VIE	Mask R-CNN (He et al., 2017) (backbone: ResNet-50 (He et al., 2016))	Yes, calculating the optical flow over 5 frames	No	RPN class loss, MASK R-CNN loss	Applied on the fly on each batch: Image resizing (1024x1024), bounding boxes, label generation	N/A
www[9]	Mask R-CNN (He et al., 2017) (backbone: ResNet-50 (He et al., 2016))	No	Pre-trained[9] on ImageNet (Russakovsky et al., 2015)	Smooth L1 loss, focal loss, binary cross entropy loss	Applied on the fly on each batch: Random flip (horizontally and vertically), rotations of $[0, 10]^\circ$	Adam (Kingma and Ba, 2014)

Team SQUASH*: An ensemble of models, combining image frame classification and multi-instance segmentation*

Team *SQUASH's* hypothesis was that they could increase the robustness and generalizability of all challenge tasks simultaneously by using multiple recognition task training. In training their method from scratch, they assumed that the network capabilities were fully utilized to learn detailed instrument features. Based on a ResNet50 (He et al., 2016), the team used the video data provided and built a classification model in order to predict all instrument frames in a sequence of video frames. On top of this classification model, they built a segmentation model by employing a Mask R-CNN (He et al., 2017) to detect multiple instrument instances in the image frames. The segmentation model was trained by leveraging the preliminary trained classification model on instrument images as a feature extractor to deepen the learning of the task of instrument segmentation. Both models were combined in a two-stage framework to process a sequence of video frames. The team reported that their method had trouble dealing with instrument occlusions, but on the other hand, they were surprised to find that it handled reflections and black borders well.

Team Uniandes*: Instance-based instrument segmentation with temporal information*

Team *Uniandes* based their multi-instance segmentation approach on the Mask R-CNN (He et al., 2017). For training purposes, they created an experimental framework with a training and validation split as well as supplementary metrics in order to identify the best version of their method and gain insight into the performance and limitations. Data augmentation was performed by calculating the optical flow with a pre-trained FlowNet2 (Ilg et al., 2017) and using the flow to map the reference annotation on to the previous frames. However, they did not find significant benefits in using the augmentation technique. The team participated in all three tasks. They produced the same output for the multi-instance segmentation and detection tasks and binarized the output of the multi-instance segmentation for the binary segmentation task. The team observed that their approach was limited in terms of finding all instruments in an image frame, but once an instrument was found it was segmented with a high *DSC* score. Although the team achieved good metric scores they stated that they fell short in segmenting small or partial instruments and instruments covered by smoke.

Team VIE*: Optical flow-based instrument detection and segmentation*

The *VIE* team approached the multi-instance segmentation task with an optical flow-based method. Their hypothesis was that the detection of moving parts in the image enables medical instruments to be detected and segmented. For their approach, they calculated the optical flow over the last five frames of a case by using the OpenCV[9] library and concatenated the optical flow with the raw image as input for a Mask R-CNN (He et al., 2017). The team assumed that this would reduce most of unnecessary clutter segmentation. The team participated in all three tasks. They produced the same output for the multi-instance segmentation and detection tasks and binarized the output of the multi-instance segmentation for the binary segmentation task. The team hypothesized that the temporal data could have been used more effectively.

Team www*: Integration of Mask R-CNN and DAC block*[9]

Team *www* proposed that a framework based on Mask R-CNN (He et al., 2017) to handle the three tasks in the challenge. Based on the observation that the instruments have variable sizes, their idea was to enlarge the receptive field and tune the anchor size for the Mask R-CNN. In addition, the team integrated DAC blocks (Gu et al., 2019) into the framework to collect more information. The team participated in all three tasks. They produced the same output for the multi-instance segmentation and detection tasks and binarized the output of the multi-instance segmentation for the binary segmentation task. The team reported that including temporal information might have helped to improve their performance.[10]

3.2. Individual performance results for participating teams

The teams' individual performances in both segmentation tasks are presented in Fig. 2 and Table 4. The dot- and boxplots show the metric values for each algorithm over all test cases in stage 3.

3.3. Challenge rankings for stage 3

As described in Section 2.3.2, an accuracy and a robustness ranking were computed for both metrics of the segmentation tasks (resulting in 4 rankings for each task). These are shown in Tables 5 and 7. For the multi-instance detection task, the *F1-score* was computed for each participant (see Table 6). The metric computation already included aggregated values, therefore only one ranking was computed for this task.

To provide deeper insight in the ranking variability, ranking heatmaps (see Fig. 3) and blob plots (see Fig. 4) were computed for all rankings of both segmentation tasks. Ranking heatmaps were used to visualize the challenge assessment data (Wiesenfarth et al., 2019b). Blob plots were used to visualize ranking stability based on bootstrap sampling (Wiesenfarth et al., 2019b).

The computed rankings for the remaining stages are given in Appendix D.

3.4. Comparison across all stages

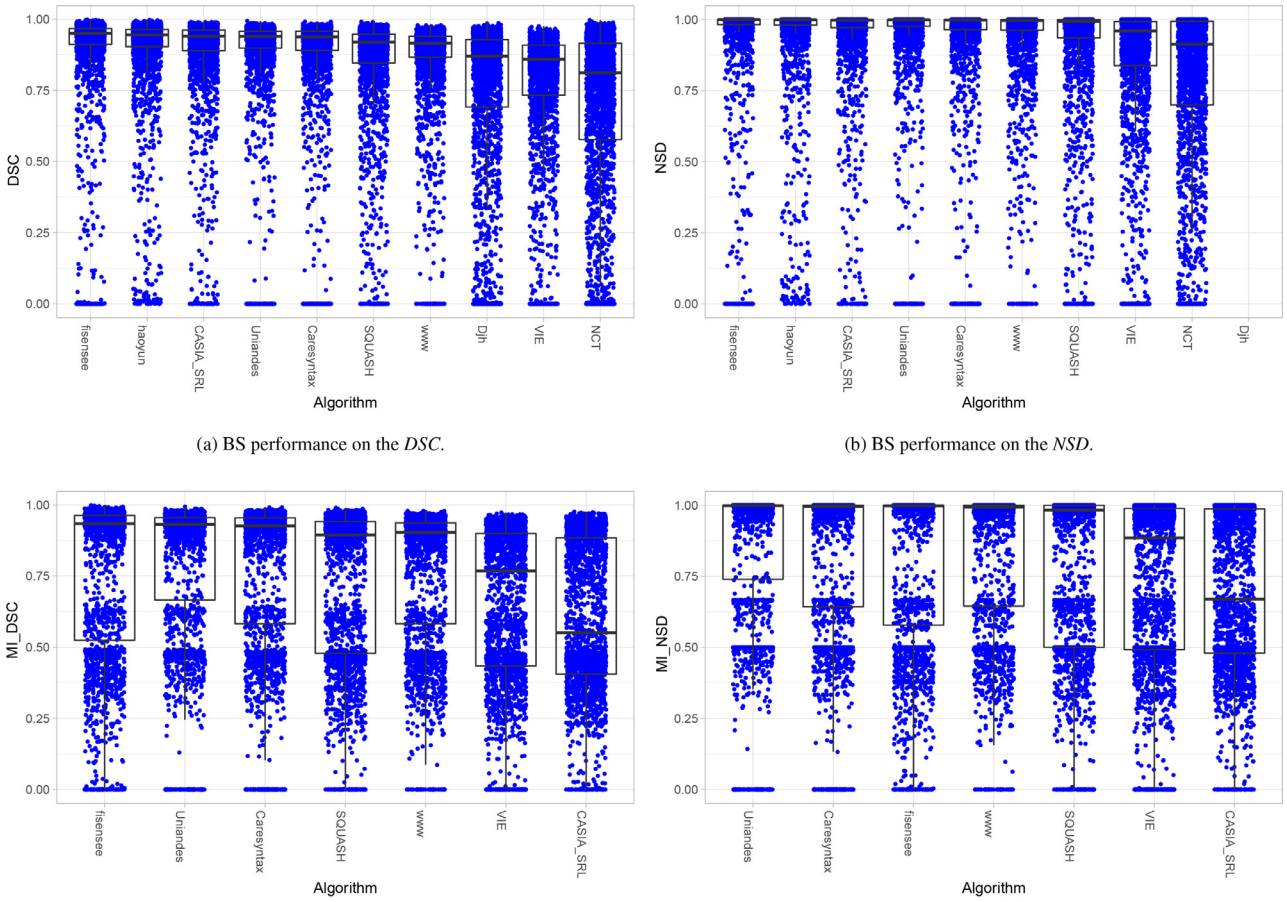
Fig. 5 shows the comparison of the average *(MI_)DSC* performances of the participating algorithms over the three evaluation stages (see Section 2) for both segmentation tasks. For this purpose, boxplots were generated for both tasks over the average metric values per team. A clear performance drop is visible in line with the increasing difficulty of the stages: Average performance produces median values of 0.88 (min: 0.73, max: 0.92) for the binary segmentation task and 0.80 (min: 0.65, max: 0.84) for the multi-instance segmentation task for stage 1. For stage 2, the median metric values decrease to 0.87 (min: 0.76, max: 0.90) and 0.78 (min: 0.64, max: 0.84) and finally, the performance for stage 3 resulted in a median of 0.85 (min: 0.69, max: 0.89) and 0.76 (min: 0.60, max: 0.80).

3.5. Further analysis

Expert baseline Only the rankings of the *(MI_)DSC* metrics were used to compare the algorithms' performances with that of a human annotator, as similar results were obtained for the *(MI_)NSD*. As images in stage 2 contain only a maximum of three instrument instances, the analysis can only show differences for n instances, where $n \in \{1, 2, 3\}$. In both tasks, the expert is the winner for both rankings. Team *fisensee* shares the first rank with the expert in the accuracy rankings for the binary segmentation task and the

[9] https://opencv.org/.

[10] Please note that this team used data from the EndoVis 2017 challenge (Allan et al., 2019) to visually check their performance on a different medical data set. The participation policies (see Appendix A) prohibit the use of other medical data for algorithm training or hyperparameter tuning. The challenge organizers defined this case as a grey zone but noted that the team may have had a competitive advantage in terms of performance generalization.



(a) BS performance on the *DSC*.

(b) BS performance on the *NSD*.

(c) MIS performance on the *MI_DSC*.

(d) MIS performance on the *MI_NSD*.

Fig. 2. Dot- and boxplots showing the individual performances of algorithms on the binary segmentation (BS; top) and multi-instance segmentation (MIS; bottom) tasks. The (multi-instance) Dice Similarity Coefficient (*(MI_)DSC*; left) and the (multi-instance) Normalized Surface Distance (*(MI_)NSD*; right) were used as metrics.

Table 4
Quantitative results of all participating methods for all three stages for the tasks binary and multi-instance segmentation. The metrics are *DSC* for the binary and *MI_DSC* for the multi-instance segmentation task. The table contains the mean, median and the 5th (Q05), 25th (Q25), 75th (Q75) and 95th (Q95) quantile for each metric.

Binary instance segmentation

Team	Stage 1						Stage 2						Stage 3					
	Mean	Median	Q5	Q25	Q75	Q95	Mean	Median	Q05	Q25	Q75	Q95	Mean	Median	Q5	Q25	Q75	Q95
CASIA_SRL	0.90	0.95	0.70	0.91	0.96	**0.98**	0.89	0.95	0.43	0.91	**0.97**	**0.98**	0.88	0.94	0.50	0.89	0.96	**0.98**
caresyntax	0.89	0.94	0.69	0.91	0.96	**0.98**	0.88	0.95	0.36	0.91	0.96	**0.98**	0.85	0.94	0.00	0.89	0.96	0.97
Djh	0.81	0.90	0.08	0.81	0.94	0.96	0.79	0.90	0.03	0.78	0.94	0.97	0.75	0.87	0.00	0.69	0.93	0.96
NCT	0.73	0.87	0.04	0.62	0.94	0.97	0.76	0.86	0.11	0.68	0.94	0.97	0.69	0.81	0.00	0.58	0.92	0.97
SQUASH	0.88	0.93	0.55	0.88	0.95	0.97	0.85	0.93	0.34	0.87	0.95	0.97	0.83	0.92	0.22	0.85	0.95	0.97
Uniandes	0.90	0.94	0.71	0.91	0.96	0.97	0.89	0.95	0.41	0.92	0.96	0.97	0.87	0.94	0.28	0.90	0.96	0.97
VIE	0.79	0.87	0.30	0.76	0.92	0.95	0.77	0.87	0.00	0.74	0.91	0.95	0.76	0.86	0.00	0.73	0.91	0.94
fisensee	**0.92**	**0.96**	**0.76**	**0.93**	**0.97**	**0.98**	**0.90**	**0.96**	**0.54**	**0.93**	**0.97**	**0.98**	0.88	**0.95**	0.34	**0.91**	**0.97**	**0.98**
haoyun	0.90	0.95	0.64	0.91	0.96	**0.98**	0.89	0.95	0.42	0.91	**0.97**	**0.98**	**0.89**	0.94	**0.52**	0.90	0.96	**0.98**
www	0.88	0.92	0.68	0.88	0.94	0.96	0.86	0.92	0.37	0.88	0.94	0.95	0.85	0.91	**0.52**	0.86	0.94	0.95
expert	-	-	-	-	-	-	***0.91***	***0.96***	***0.73***	***0.93***	***0.97***	***0.98***	-	-	-	-	-	-

Multi-instance segmentation

Team	*Stage 1*						*Stage 2*						*Stage 3*					
	Mean	Median	Q5	Q25	Q75	Q95	Mean	Median	Q05	Q25	Q75	Q95	Mean	Median	Q5	Q25	Q75	Q95
CASIA_SRL	0.65	0.69	0.24	0.44	0.91	0.96	0.64	0.68	0.18	0.43	0.91	0.96	0.60	0.55	0.19	0.41	0.88	0.95
caresyntax	0.82	0.93	0.32	0.83	0.96	0.97	**0.80**	**0.94**	0.32	0.68	0.96	**0.98**	0.77	**0.93**	0.00	0.58	0.95	0.97
SQUASH	0.78	0.90	0.32	0.60	0.94	0.97	0.75	0.91	0.26	0.48	0.95	0.97	0.73	0.89	0.22	0.48	0.94	0.97
Uniandes	**0.84**	**0.94**	**0.40**	**0.88**	0.96	0.97	0.84	**0.94**	**0.39**	**0.88**	0.96	0.97	**0.80**	**0.93**	0.26	**0.67**	0.95	0.97
VIE	0.67	0.81	0.16	0.45	0.90	0.95	0.65	0.77	0.00	0.43	0.90	0.95	0.65	0.77	0.00	0.43	0.90	0.94
fisensee	0.80	**0.94**	0.32	0.62	**0.97**	**0.98**	**0.80**	**0.94**	0.28	0.61	**0.97**	**0.98**	0.76	**0.93**	0.17	0.52	**0.96**	**0.98**
www[10]	0.81	0.90	0.37	0.79	0.94	0.96	0.78	0.91	0.30	0.63	0.94	0.96	0.76	0.89	**0.31**	0.58	0.93	0.95
expert	-	-	-	-	-	-	***0.88***	***0.95***	***0.47***	***0.91***	***0.97***	**0.98**	-	-	-	-	-	-

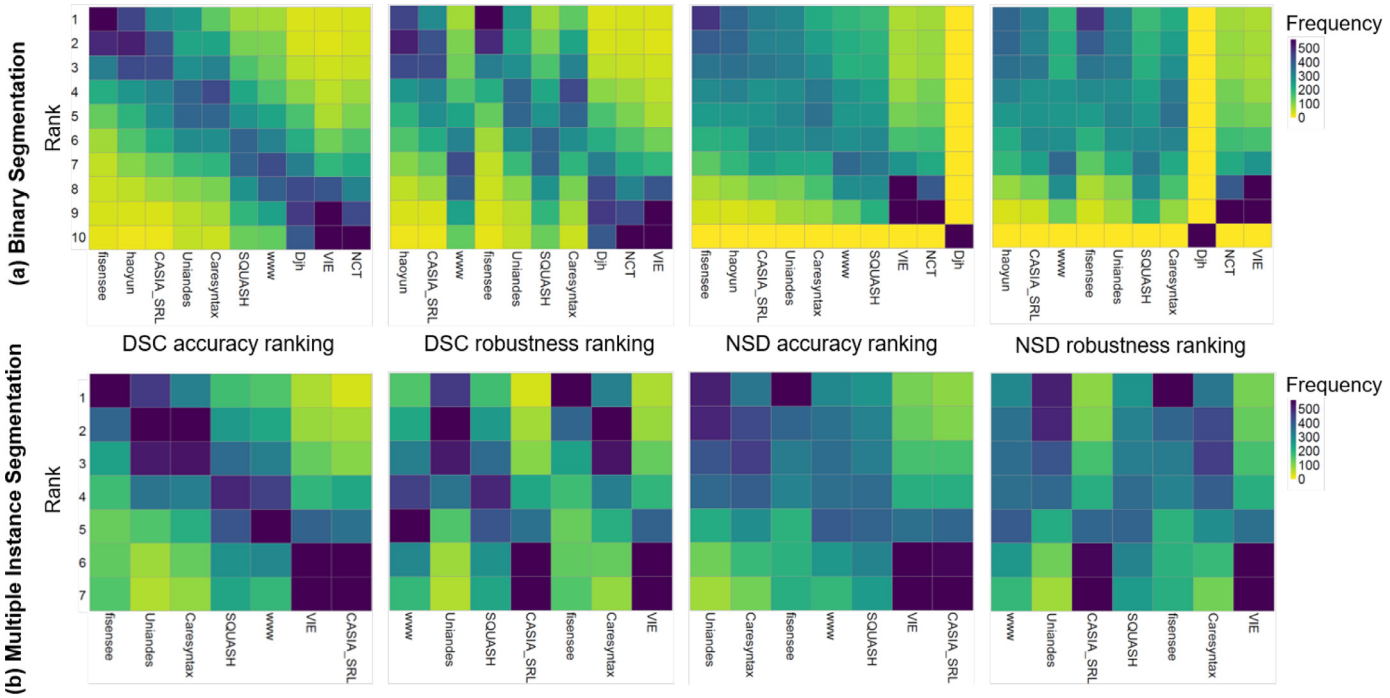


Fig. 3. Ranking heatmaps for the four rankings in the binary segmentation and multi-instance segmentation tasks. Each cell (i, A_j) shows the absolute frequency of test cases in which algorithm A_j achieved rank i. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

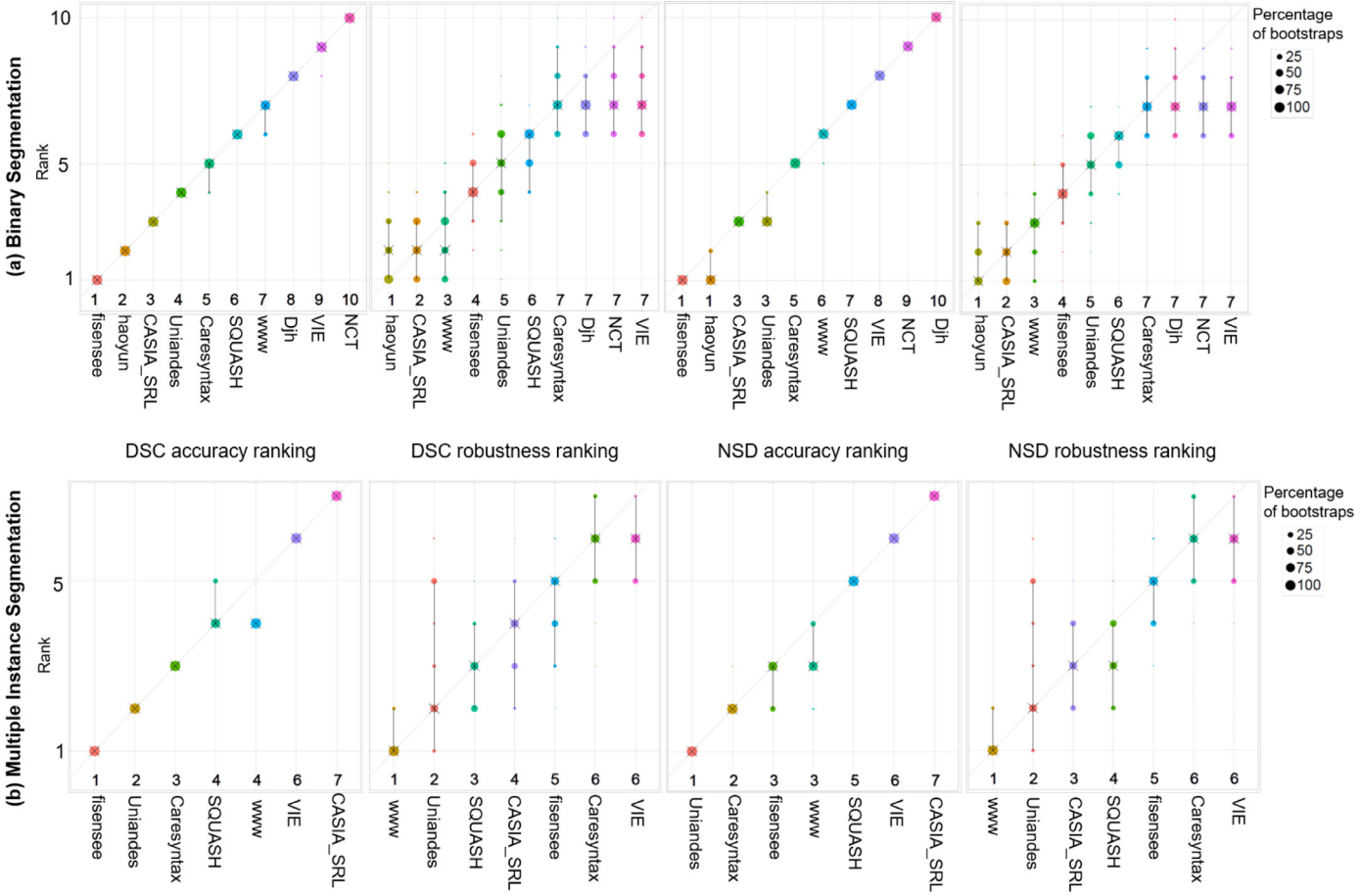


Fig. 4. Blob plots for the four rankings in the binary segmentation and multi-instance segmentation tasks. Blob plots are used to visualize ranking stability based on bootstrap sampling. Algorithms are color-coded, and the area of each blob at position (A_i, rank j) is proportional to the relative frequency A_i of the achieved rank j across $b = 1000$ bootstrap samples. The median rank for each algorithm is indicated by a black cross. 95% bootstrap intervals across bootstrap samples are indicated by black lines. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

Table 5
Binary segmentation: Rankings for stage 3 of the challenge. The upper part of the table shows the Dice Similarity Coefficient (*DSC*) rankings and the lower part shows the Normalized Surface Distance (*NSD*) rankings (accuracy rankings on the left, robustness rankings on the right). Each ranking contains a team identifier, either a proportion of significant tests divided by the number of algorithms (prop. sign.) for the accuracy ranking or an aggregated *DSC*/*NSD* value (aggr.*DSC*/*NSD* value) and a rank.

DSC: ACCURACY RANKING			DSC: ROBUSTNESS RANKING		
Team identifier	Prop. Sign.	Rank	Team identifier	Aggr. *DSC* Value	Rank
fisensee	**1.00**	**1**	***haoyun***	**0.52**	**1**
haoyun	0.89	2	*CASIA_SRL*	0.50	2
CASIA_SRL	0.78	3	*www*[10]	0.49	3
Uniandes	0.67	4	*fisensee*	0.34	4
caresyntax	0.56	5	*Uniandes*	0.28	5
SQUASH	0.44	6	*SQUASH*	0.22	6
www[10]	0.33	7	*caresyntax*	0.00	7
Djh	0.22	8	*Djh*	0.00	7
VIE	0.11	9	*NCT*	0.00	7
NCT	0.00	10	*VIE*	0.00	7
NSD: ACCURACY RANKING			*NSD*: ROBUSTNESS RANKING		
Team identifier	Prop. Sign.	Rank	Team identifier	Aggr. *NSD*Value	Rank
haoyun	**0.89**	**1**	***haoyun***	**0.63**	**1**
fisensee	**0.89**	**1**	*CASIA_SRL*	0.62	2
CASIA_SRL	0.67	3	*www*[10]	0.57	3
Uniandes	0.67	3	*fisensee*	0.45	4
caresyntax	0.56	5	*Uniandes*	0.32	5
www[10]	0.44	6	*SQUASH*	0.26	6
SQUASH	0.33	7	*caresyntax*	0.00	7
VIE	0.22	8	*Djh*	0.00	7
NCT	0.11	9	*NCT*	0.00	7
Djh	0.00	10	*VIE*	0.00	7

Table 6
Multi-instance detection: Ranking for the mean average precision (*mAP*) in stage 3 of the challenge.

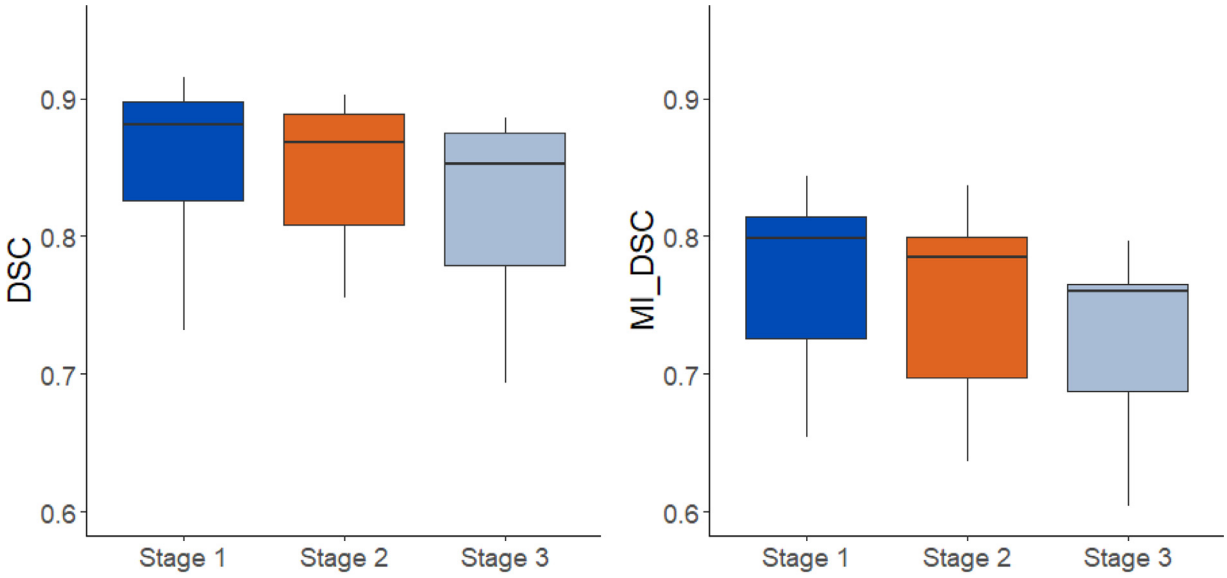
Team identifier	F1-score	Rank
Uniandes	0.91	**1**
www[10]	0.90	2
caresyntax	0.89	3
SQUASH	0.86	4
fisensee	0.86	5
VIE[9]	0.82	6

multi-instance segmentation task for frames with 1 instrument. The mean segmentation accuracy per instrument instance can be seen in Fig. 6.

Worst case analysis For further analyses, we investigated the image frames that produced the 100 best or worst metric values of participating teams. This investigation revealed the strengths and weaknesses of the proposed methods. In general, algorithm performance drops with the number of instruments in the image as illustrated in Fig. 7. The algorithms succeeded in images containing reflections, blood, different illuminations and in finding the

Table 7
Multi-instance segmentation: Rankings for stage 3 of the challenge. The upper part of the table shows the multi-instance Dice Similarity Coefficient (*MI_DSC*) rankings and the lower part shows the multi-instance Normalized Surface Distance (*MI_NSD*) rankings (accuracy rankings on the left, robustness rankings on the right). Each ranking contains a team identifier, either a proportion of significant tests divided by the number of algorithms (prop. sign.) for the accuracy ranking or an aggregated *MI_DSC*/*MI_NSD* value (aggr. *MI_DSC*/*MI_NSD* value) and a rank.

MI_DSC: ACCURACY RANKING			MI_DSC: ROBUSTNESS RANKING		
Team identifier	Prop. Sign.	Rank	Team identifier	Aggr. *MI_DSC*Value	Rank
fisensee	**1.00**	**1**	***www***[10]	**0.31**	**1**
Uniandes	0.83	2	*Uniandes*	0.26	2
caresyntax	0.67	3	*SQUASH*	0.22	3
SQUASH	0.33	4	*CASIA_SRL*	0.19	4
www[9]	0.33	4	*fisensee*	0.17	5
VIE	0.17	6	*caresyntax*	0.00	6
CASIA_SRL	0.00	7	*VIE*	0.00	6
MI_NSD: ACCURACY RANKING			*MI_NSD*: ROBUSTNESS RANKING		
Team identifier	Prop. Sign.	Rank	Team identifier	Aggr. *MI_NSD*Value	Rank
Uniandes	**1.00**	**1**	***www***[10]	**0.35**	**1**
caresyntax	0.67	2	*Uniandes*	0.29	2
fisensee	0.50	3	*CASIA_SRL*	0.27	3
www[9]	0.50	3	*SQUASH*	0.26	4
SQUASH	0.33	5	*fisensee*	0.16	5
VIE	0.17	6	*caresyntax*	0.00	6
CASIA_SRL	0.00	7	*VIE*	0.00	6



(a) BS (b) MIS

Fig. 5. Boxplots of the variance across all test images for the (a) binary segmentation task with the Dice Similarity Coefficient (*DSC*) and (b) the multi-instance segmentation task with the Multi-instance Dice Similarity Coefficient (*(MI_)DSC*) for stages 1 to 3. The boxplots show the average algorithm performances (mean over all participant predictions per image) per image.

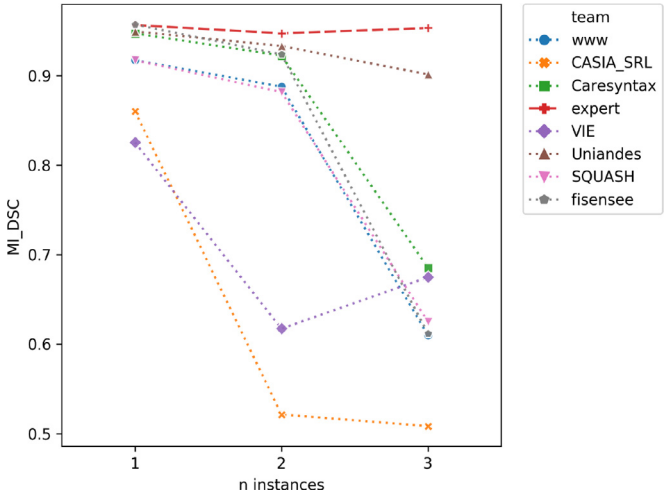


Fig. 6. *Median MI_DSC* as a function of the number of instruments in the image for stage 2 of the test data for the multi-instance segmentation task. It shows the performance of all algorithms in comparison to the human expert. Clearly, all algorithms' performance drops with the number of visible instruments in the image while the experts performance stays constant..

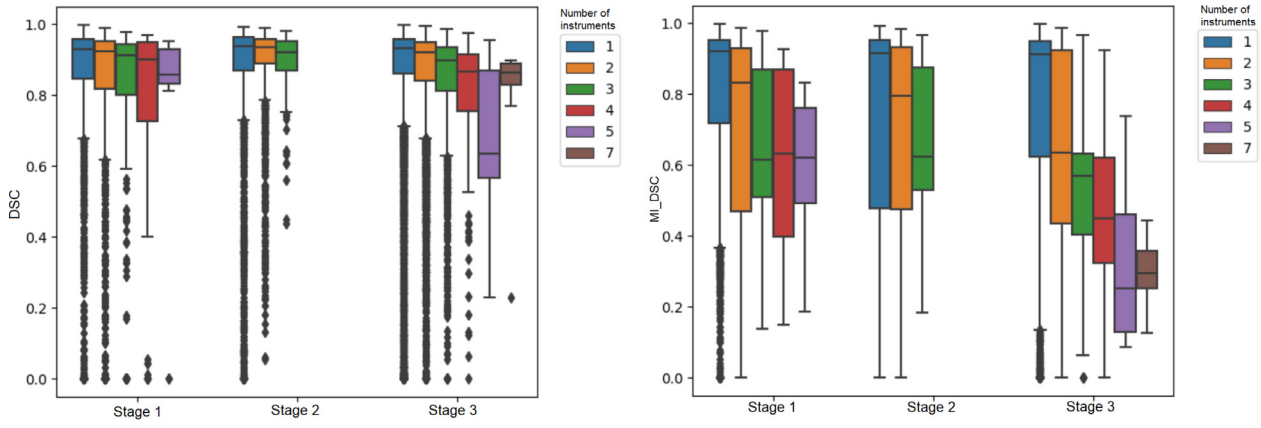
inside of the trocar (see Fig. 8). Problems still arose in image frames which contained small and transparent instruments. False positives (mainly objects that were not defined as instruments) turned out to be a problem for all tasks. Furthermore, algorithm performance was poor for images with instruments, close to another as well as crossing, partially hidden or moving instruments, instruments close to the image border and images containing smoke (see Fig. 9 and 10).

4. Discussion

We organized the first challenge in the field of surgical data science that (1) included tasks on multi-instance detection/tracking and (2) placed particular emphasis on the robustness and generalization capabilities of the algorithms. The key insights are:

1. Competing methods: These state-of-the-art methods are exclusively based on deep learning with a specific focus on U-Nets (Ronneberger et al., 2015) (binary segmentation) and Mask R-CNNs (He et al., 2017) (multi-instance detection and segmentation). For binary segmentation, the U-Net and the new DeepLabV3 architecture yielded an equally strong performance. For the multi-instance segmentation, a U-Net in combination with a connected component analysis was a strong baseline, but a Mask R-CNN approach was more promising overall, especially in terms of robustness.
2. Performance:
   (a) Binary segmentation: The mean performances of the winning algorithms for the accuracy ranking (*DSC* of 0.88) and the robustness ranking (*DSC* of 0.89) were similar to that of the previous winners of binary segmentation challenges (winner of the EndoVis Instrument Segmentation and Tracking Challenge 2015[11]: *DSC* of 0.84; winner of the EndoVis 2017 Robotic Instrument Segmentation Challenge (Allan et al., 2019): *DSC* of 0.88). Given the high complexity of ROBUST-MIS' data in comparison to previously released data sets, we attribute the fact that the performances are similar to the high amount of training data.
   (b) Multi-instance detection: The top three algorithms achieved *F1-score* $\geq$ 0.89 for stage 3. The winning algorithms featured very high accuracy, robustness and generalization capabilities. The few failure cases were related to the detection of small instruments, instruments close to another or instruments close to the image border.
   (c) Multi-instance segmentation: The mean *MI_DSC* scores for the winning algorithm of the accuracy ranking were
   - 0.82 for cases with one instrument instance,
   - 0.71 for cases with two instrument instances,
   - 0.62 for cases with three instrument instances,
   - 0.45 for cases with more than three instrument instances.

[11] https://endovissub-instrument.grand-challenge.org/.



(a) BS (b) MIS

Fig. 7. Boxplots of mean (multi-instance) Dice Similaritiy Coefficient (*(MI_)DSC*) values of participating algorithms for the binary and multi-instance segmentation tasks for stages 1 to 3 stratified by the number of instruments in the video frames.

Multi-instance segmentation in endoscopic video data, therefore cannot be regarded as a solved problem.

3. Generalization: All participating methods for the binary segmentation tasks had a satisfying generalization capability over all three stages, with a median drop from 0.88 (stage 1) to 0.85 (stage 3; 3%). The generalization capabilities for the multi-instance segmentation were slightly worse, with a median drop form 0.80 (stage 1) to 0.76 (stage 3; 5%).
4. Robustness: The most successful algorithms are robust to reflections, blood and smoke. The segmentation of small, close positioned, transparent, moving, overlapping and crossing instruments, however, remains a great challenge that needs to be addressed.

The following sections provide a detailed discussion on the challenge infrastructure (Section 4.1.1), challenge data (Section 4.1.3), challenge methods (Section 4.2.1) and challenge results (Section 4.2.2).

4.1. Challenge design

In this section, we discuss the infrastucture and the data of our challenge.

4.1.1. Challenge infrastructure

We decided to use Synapse[12] as our challenge platform as it is the underlying platform of the well-known and DREAM challenges[13], and, as such, provides a complete and easy to use environment for both challenge participants and organizers. Furthermore, in addition to helping organizers monitor on how a challenge should be structured, it also helps them to follow current best practices by relying on docker submissions. However, while the overall experience with Synapse was very good, downloading the data was a problem due to slow download rates, which were dependent on the global download location and the size of the data set (about 400 GB). Unlike the data download, the docker upload was very quick and easy to follow.

The submission of docker containers and complete evaluation is already in common usage in other disciplines (e.g. CARLA[14]). However, most of the very recent challenges in the biomedical image analysis community still use plain results submissions (e.g. BraTS[15], KiTS2019[16], PAIP 2019[17]). We believe that using dockers for the evaluation is the best way as it can help (1) to avoid test data set overfitting and (2) to prevent potential instances of fraud such as manually labeling the test data (Reinke et al., 2018). However, using docker containers also means more work for the individual participants (in creating of the docker containers) and for the organizers. In addition to providing the Computing Processing Unit (CPU) and Graphics Processing Unit (GPU) resources, they have to provide support for docker related questions and must have a strategy for dealing with invalid submissions (e.g. allowing re-submission). In our challenge for example, submitted dockers were run on a small proportion of the training set to check whether the submissions worked. For five participants, the first submission failed. They were allowed to re-submit but we manually checked whether the network parameters had changed.

4.1.2. Metrics and ranking

Following recommendations of the Medical Segmentation Decathlon (Cardoso, 2018), we decided to use two metrics for the segmentation task; an overlap measure (*DSC*) and a distance measure (*NSD*). We used a non-global *DSC* for the multi-instance segmentation, meaning that the *DSC* values of instrument instances were first averaged to get an image-based score before taking the mean over all images. Another option would have been to use a global *DSC* measure, which would compute the *DSC* score globally over the complete data set and all instrument instances. However, we decided to use the non-global metric to give higher weight to small instruments.

To put a particular focus on the robustness of the methods, we decided to compute a dedicated ranking for the 5% percentile performance of the methods, as summarized in Section 2.3.2. Given our previous work on ranking stability (Maier-Hein et al., 2018), it can be assumed that a ranking based on the 5% percentile would naturally lead to less robust rankings compared to an aggregation with the mean or the median. This is one possible explanation for the fact that the ranking stability for the robustness ranking was worse compared to that of the accuracy ranking, as shown in Fig. 4.

Initially, during the challenge event at the MICCAI conference, the mean average precision (*mAP*) (Everingham, Van Gool, Williams, Winn, Zisserman, 2010) metric was used (results are provided in Table D.1) to determine the best performing algorithm. However, due to an error in the implementation and missing con-

[12] https://www.synapse.org/.
[13] http://dreamchallenges.org/.
[14] https://carlachallenge.org/.
[15] http://braintumorsegmentation.org/.
[16] https://kits19.grand-challenge.org/rules/.
[17] https://paip2019.grand-challenge.org/.

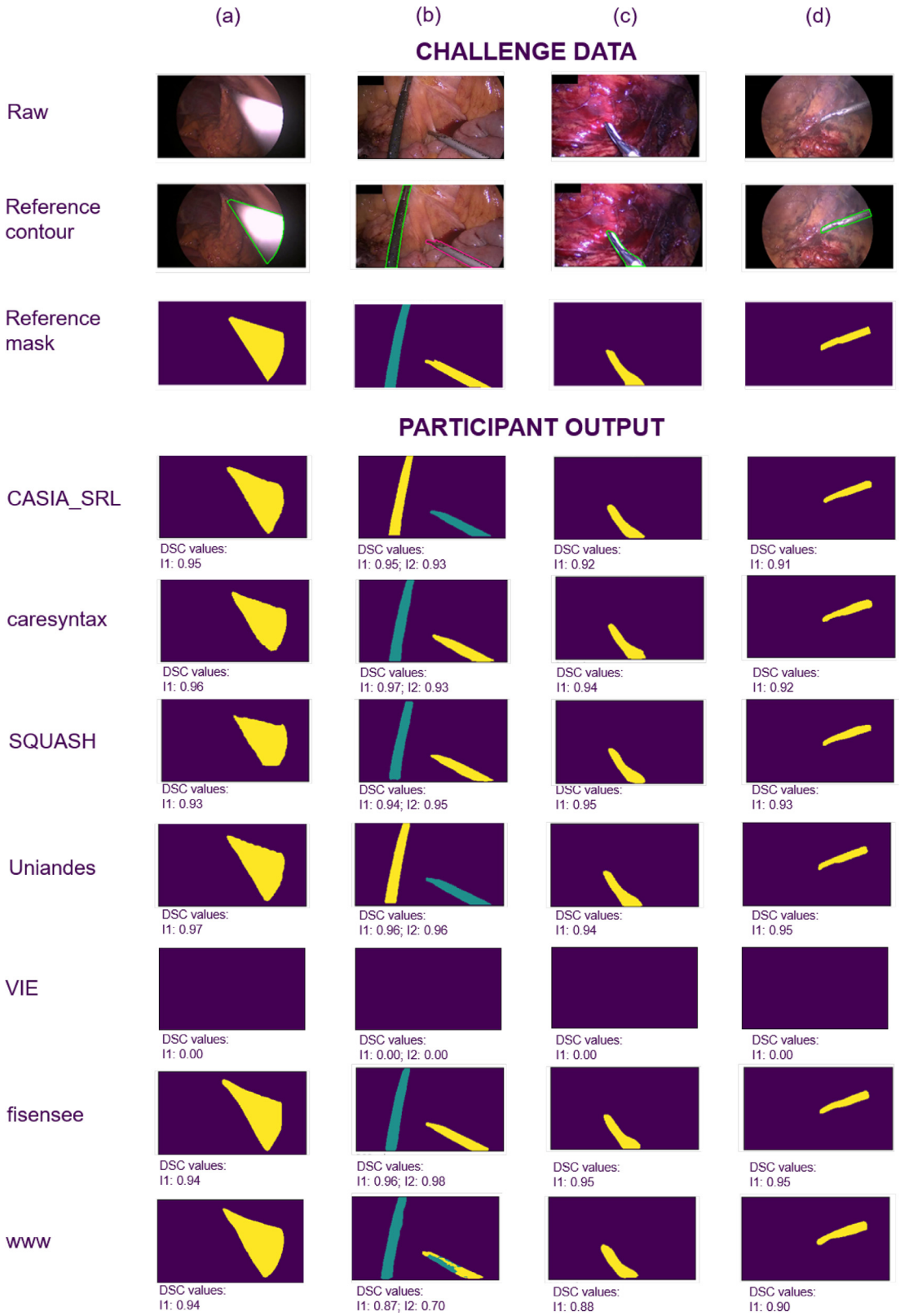


Fig. 8. Test cases with high corresponding algorithm performances. Each row shows the raw frame, the reference contours and mask as well as the algorithm output of the participating teams of the multi-instance segmentation (MIS) task for one representative frame. The columns represent image frames with (a) overexposure, (b) clearly separated instruments, (c) blood and reflections, (d) smoke, respectively. For participants, the Dice Similarity Coefficient (DSC) values are provided per instrument instance (I_i).

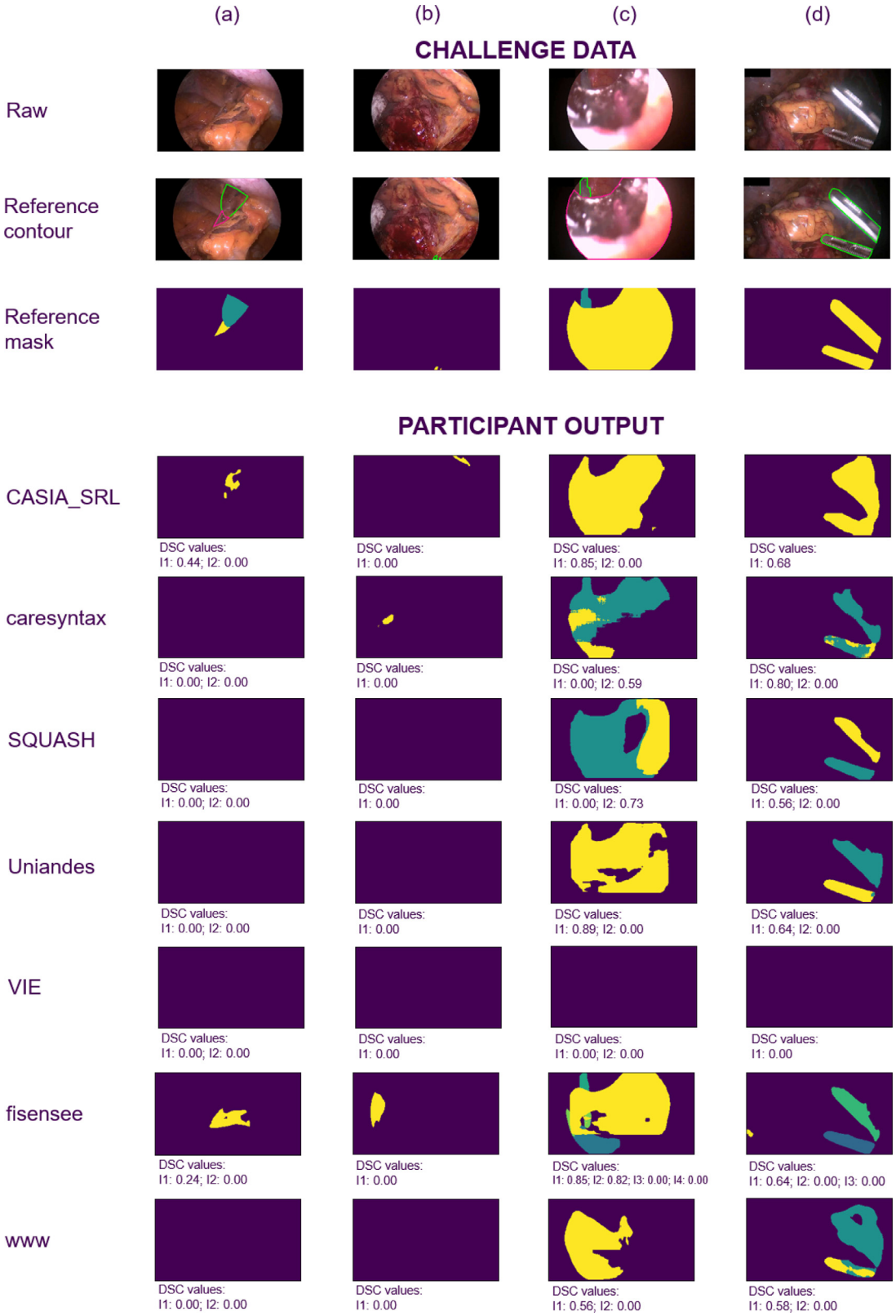


Fig. 9. Test cases with low corresponding algorithm performances. Each row shows the raw frame, the reference contours and mask as well as the algorithm output of the participating teams of the multi-instance segmentation (MIS) task for one representative frame. The columns represent image frames with (a) transparency, (b) small instruments, (c) overlapping instruments, (d) instruments near the border, respectively. For participants, the Dice Similarity Coefficient (DSC) values are provided per instrument instance (I_i).

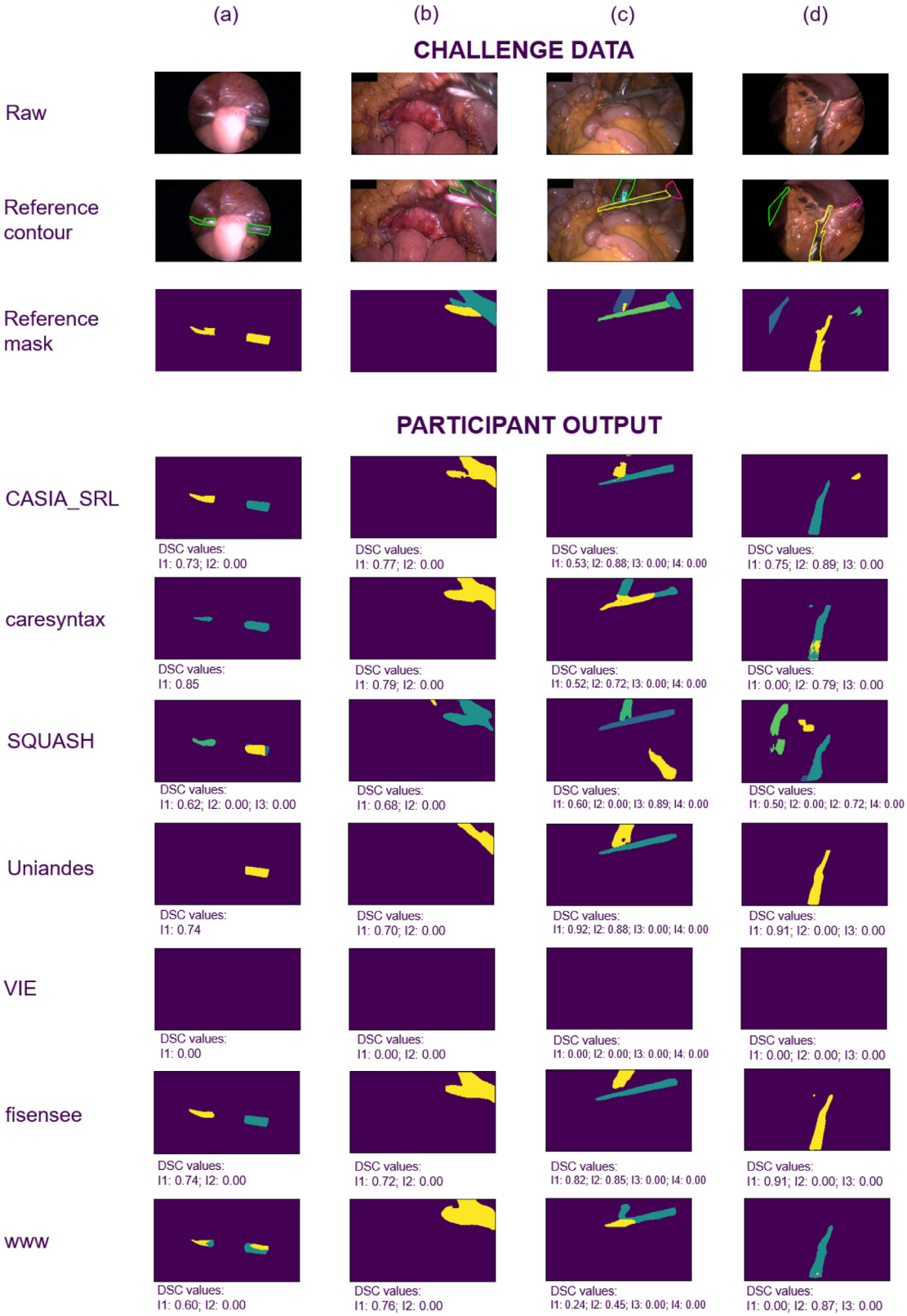


Fig. 10. Test cases with low corresponding algorithm performances. Each row shows the raw frame, the reference contours and mask as well as the algorithm output of the participating teams of the multi-instance segmentation (MIS) task for one representative frame. The columns represent image frames with (a) an instrument overlain by tissue, (b) motion, (c) multiple instruments, (d) underexposure and multiple instruments, respectively. For participants, the Dice Similarity Coefficient (DSC) values are provided per instrument instance (I_i).

fidence scores from the algorithms, we decided to update the ranking with the F1-score.

4.1.3. Challenge data

In general, we observed many inconsistencies in the initial data annotation, which is why we introduced a structured multi-stage annotation process involving medical experts and following a pre-defined annotation protocol (see Appendix B). We recommend challenge organizers to generate such a protocol from the outset of their challenge.

It should be noted that three different surgical procedures were used for the challenge, yet, these three procedures are all colorectal surgeries that share similarities. A rectal resection incorporates parts of a sigmoid resection, for example. It is possible that performance drops will be more radical when analyzing a wider variety of procedures such as biliopancreatic or upper gastrointestinal surgeries.

In the future, we will also prevent the potential side effects which resulting from pre-processing. The fact that we downsampled our video images may have harmed performance. However, due to the fact that (1) all participants had the same starting conditions, (2) the applied CNNs methods had to fit to GPUs and (3) all participants reduced the resolution further, we think that these effects are only minor.

4.2. Challenge outcome

4.2.1. Methods

The variability of all of the methods, submitted for the binary segmentation was vast and ranged from 2D U-Net versions (TernausNet, multi scale U-Net) to different implementations of the Mask R-CNN with a ResNet backbone to the latest DeepLabV3 network architecture. For the multi-instance detection and multi-instance segmentation tasks, however, the range of the underlying architecture was much narrower, with multiple Mask R-CNN variations and one combination of a U-Net, a classical approach and the principal component analysis (see Table 3).

The most successful participating team (*haoyun*) in the binary segmentation task implemented a DeepLabV3+ architecture which gave them the top rank in three out of the four rankings for the binary segmentation task. A relatively simple approach based on the combination of a U-Net with a connected component analysis by the *fisensee* team turned out to be a strong baseline and won accuracy rankings in both the binary segmentation task and the *DSC* accuracy ranking for the multi-instance segmentation task. It was, however, less successful in terms of robustness.

An increasingly relevant problem in reporting challenge results is the fact that it is often hard to understand which specific design choice for a certain algorithm make this algorithm better than the competing methods (Maier-Hein et al., 2018). Based on our challenge analysis, we hypothesize that data augmentation and the specifics of the training process are the key to a winning result. In other words, we believe that focusing on one architecture and performing a broad hyperparameter search in combination with an extensive data augmentation technique and a well-thought-out training procedure will create more benefit than testing many different network architectures without optimizing the training process. This is in line with recent findings in the field of radiological data science (Isensee et al., 2018).

4.2.2. Results

The key insights have already been summarized at the beginning of the discussion. Methods that tackle the multi-instance segmentation performed worse compared to the binary segmentation task. In fact, when multiple instrument instances were visible in one image, the algorithm performance decreased dramatically from over 0.8 for one instance to less than 0.6 for more than three instances (see Fig. 7). This is also reflected in Fig. 2 (c) and (d), which show clusters in the boxplots at specific metric values. These clusters correspond to the performance with respect to different numbers of instrument instances. For a single instrument, metric values are high, for multiple instruments the metric values are grouped around lower values. We thus conclude that detection of multiple instances remains an unsolved problem.

Although the described winning methods produce median *MI_DSC* results above 0.9 (see Fig. 6), most of them could not outperform the expert baseline in the multi-instance segmentation task, especially if more than one instrument was present in the image frames. In fact, only the teams *fisensee* (binary segmentation) and *Uniandes* (multi-instance segmentation) produced similar performances to the human annotator in stage 2 of the challenge. It should be noted that for pragmatic reasons, the additional labeling was performed only on a subset of images and with only one additional medical expert. The discrepancy in performance between algorithms and experts may differ based on the data and the annotator.

Generally, the expert accuracy is independent of the number of visible instances, while the performance of the algorithm drops with an increasing number. However, to our surprise, the expert also achieved comparatively low values in the robustness rankings (aggregated values of 0.43 or 0.47 for $n = 1$ and $n = 2$ instruments). We found this mainly to be caused by missing or wrong instrument instances (see Fig. E.1). However, where the expert did detect an instance, the segmentation quality of this instance is almost always good (*MI_DSC*= 0.9 is on the 10th percentile and *MI_DSC*= 0.95 on the 37th percentile), which is not the case for the algorithms as shown in see Fig. E.1 (Team *Uniandes* with *MI_DSC*=0.9 on the 14th percentile and *MI_DSC*= 0.95 on the 48th percentile; Team *fisensee* with *MI_DSC*= 0.9 on the 14th percentile and *MI_DSC*= 0.95 on the 37th percentile).

By analyzing the worst 100 cases across all of the methods, we found that all methods generally had issues with small, transparent or fast moving instruments. In addition, instruments close to other instruments or the image border, as well as partially hidden or crossing instruments were difficult to detect and segment (see Fig. 9 and 10). We also observed that classic challenges (Bodenstedt et al., 2018) such as reflections, blood, different illumination conditions did not pose any great problems. Images acquired when the lens of the endoscope was inside of a trocar were not particularly difficult to process.

It should be noted that only three of the ten methods incorporated the temporal video information provided with the frames to be annotated. One method used the video information to predict the likelihood of instrument presence in a multi-task setting while two approaches used the videos to calculate the optical flow. However, based on the team reports and on the challenge results, none of the teams where able taking a benefit from using the video data, neither for the binary segmentation task, nor for the multi-instance detection/segmentation tasks. Given the way in which medical and technical experts annotated the data, this is surprising, and we speculate that much of the potential of temporal context remains to be discovered.

Finally, it should be noted that an evaluation of the inference time of methods was not included in this paper because a respective metric had not been announced to the challenge participants. Although we assume that the participating teams had not optimized their methods for performance, we performed a preliminary analysis of the docker submissions to approximate computation times. This yielded runtimes between 0.07 and 7.3 seconds per image frame (mean: 1.09 seconds per image frame). Given the need for real-time inference, we recommend using a runtime-based metric in future challenges of this kind.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments and conflicts of interest

This challenge has been funded by the Surgical Oncology Program of the National Center for Tumor Diseases (NCT) Heidelberg and the project "OP4.1", funded by the German Federal Ministry of Economic Affairs and Energy (grant number BMWI 01MT17001C). It was further supported by UNDERSTAND.AI[18], NVIDIA GmbH[19] and Digital Surgery[20]. The challenge was further supported by the Helmholtz Association under the joint research school HIDSS4Health (Helmholtz Information and Data Science School for Health). Furthermore, the authors wish to thank Tim Adler, Janek Gröhl, Alexander Seitel and Minu Dietlinde Tizabi for proofreading the paper.

L.M.-H., T.R., A.R., S.B. and S.S. worked with device manufacturer Karl Storz GmbH & Co. KG in the joint research project "OP4.1", funded by the German Federal Ministry of Economic Affairs and Energy (grant number BMWI 01MT17001C). M.W., H.G.K. and P.P.M. worked with device manufacturer Karl Storz GmbH & Co. KG in the joint research project "InnOPlan", funded by the German Federal Ministry of Economic Affairs and Energy (grant number BMWI 01MD15002E).

G.-B.B., Z.-G.H., Z.-L.N., Y.-J.Z. and H.-B.C. were supported by the National Key Research and Development Program of China (Grant 2017YFB1302704).

D.G., G.W. and L.W. were funded by National Natural Science Fundation of China (81771921, 61901084).

P.H., M.A.R. and D.J. were funded by Research Council of Norway projects number 263,248 (Privaton).

S.K. and K.S. were funded by the FWF Austrian Science Fund under grant P 32010-N38.

All challenge organizers and some members of their institute had access to training and test cases and were therefore not eligible for awards.

Appendix A. Challenge organization

The "Robust Medical Instrument Segmentation Challenge 2019 (ROBUST-MIS 2019)" was organized as a sub-challenge of the Endoscopic Vision Challenge 2019 at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in Shenzhen, China. It was organized by T. Roß, A. Reinke, M. Wagner, H. Kenngott, B. Müller, A. Kopp-Schneider and L. Maier-Hein. See Section A.1 for detailed description. The challenge was intended as a one-time event with a fixed submission deadline. The platforms grand-challenge.org (Roß et al., 2019a) and synapse.org (Roß et al., 2019b) served as websites for the challenge. Synapse served as data providing platform which was further used to upload the challenge participants' submissions.

The participation policies for the challenge allowed only fully automatic algorithms to be submitted. Although it was possible to use publicly available data released outside the field of medicine to train the methods or to tune hyperparameters, it was forbidden to use any medical data, besides the training data offered by the challenge. For members of the organizers' departments it was possible to participate in the challenge but they were not eligible for awards and their participation would have been highlighted in the leaderboards. The challenge was funded by the company Digital Surgery with a total monetary award of 10,000€. As the challenge comprised 9 rankings in total (see Section 2.3.2), each winning team was awarded 1,000€ and each runner-up team 125€. Moreover, the top three performing methods for each ranking were announced publicly. The remaining teams could decide whether or not their identity was revealed. One team decided not to be mentioned in the rankings. Finally, for this publication, each participating team could nominate members of their team as co-authors. The method description submitted by the authors was used in the publication (see Section 3.1). Personal data of the authors include their names, affiliations and contact addresses. References used in the method description were published as well. Participating teams are allowed to publish their results separately with explicit permission from the challenge organizers once this paper has been accepted for publication.

The submission instructions for the participating methods are published on the Synapse website and consist of a detailed description of the submission of docker containers which were used to evaluate the results. The complete submission instructions are provided in Appendix C. Algorithms were only evaluated on the test data set, so no leaderboard was published before the final result submission. The initial training data set was released on 1st July 2019, the final training data set on 5th August 2019. Participants could register for the challenge until 14th September 2019. The docker submission took place between 15th September and 28th September 2019. There where two deadlines, the 21th September for participants, whose methods would require more than 3h of runtime and the 28th September for participants, whose dockers needs less than 3h runtime. Participating teams had to submit a method description in addition to the docker containers.

The data sets of the challenge were fully anonymized (see Section 2.2) and could therefore be used without any ethics approval (Recital26, 2016). By registering in the challenge, each team agreed (1) to use the data provided only in the scope of the challenge and (2) to neither pass it on to a third party nor use it for any publication or for commercial use. The data will be made publicly available for non-commercial use.

The evaluation code for the challenge was made publicly available (Roß and Reinke, 2019) and participants were encouraged to release their methods in open source.

A1. Author contributions

All authors read the paper and agreed to publish it.

- T. Roß and A. Reinke organized the challenge, performed the evaluation and statistical analyses and wrote the manuscript
- P.M. Full, H. Hempe, D. Mindroc-Filimon, P. Scholz, T.N. Tran and P. Bruno reviewed and labeled the challenge data set
- M. Wagner, H. Kenngott, B.P. Müller-Stich organized the challenge and performed the medical expert review of the challenge data set
- M. Apitz performed the medical expert review of the challenge data set
- K. Kirtac, J. Lindström Bolmgrem, M. Stenzel, I. Twick and E. Hosgor participated in the challenge as team *caresyntax* in all three tasks
- Z.-L. Ni, H.-B. Chen, Y.-J. Zhou, G.-B. Bian and Z.-G. Hou participated in the challenge as team *CASIA_SRL* in the binary and multi-instance segmentation tasks
- D. Jha, M.A. Riegler and P. Halvorsen participated in the challenge as team *Djh* in the binary segmentation task
- F. Isensee and K. Maier-Hein participated in the challenge as team *fisensee* in all three tasks

[18] https://understand.ai.
[19] https://www.nvidia.com.
[20] https://digitalsurgery.co.

- L. Wang, D. Guo and G. Wang participated in the challenge as team *haoyun* in the binary segmentation task
- S. Leger, S. Bodenstedt and S. Speidel participated in the challenge as team *NCT* in the binary segmentation task
- S. Kletz and K. Schoeffmann participated in the challenge as team *SQUASH* in all three tasks
- L. Bravo, C. González and P. Arbeláez participated in the challenge as team *Uniandes* in all three tasks
- R. Shi, Z. Li, T. Jiang participated in the challenge as team *VIE* in all three tasks
- J. Wang, Y. Zhang, Y. Jin, L. Zhu, L. Wang and P.-A. Heng participated in the challenge as team *www* in all three tasks
- A. Kopp-Schneider and M. Wiesenfarth performed statistical analyses
- L. Maier-Hein organized the challenge, wrote the manuscript and supervised the project

Appendix B. Annotation instructions

B1. Terminology

- Matter: Anything that has mass, takes up space and can be clearly identified.
- Examples: tissue, surgical tools, blood
- Counterexamples: reflections, digital overlays, movement artifacts, smoke

Medical instrument to be detected and segmented: Elongated rigid object introduced into the patient and manipulated directly from outside the patient.

- Examples: grasper, scalpel, (transparent) trocar, clip applicator, hooks, stapling device, suction
- Counterexamples: non-rigid tubes, bandage, compress, needle (not directly manipulated from outside but manipulated with an instrument), coagulation sponges, metal clips

B2. Tasks

Participating teams may enter competitions related to the following tasks:

Binary segmentation:

- Input: 250 consecutive frames (10sec) of a laparoscopic video with the last frame containing at least one medical instrument.
- Output: A binary image, in which "0" indicates the absence of a medical instrument and a number ">0" represents the presence of a medical instrument.

Multi-instance detection and segmentation:

- Input: 250 consecutive frames (10sec) of a laparoscopic video with the last frame containing at least one medical instrument.
- Output: An image, in which "0" indicates the absence of a medical instrument and numbers "1", "2",... represent different instances of medical instruments.

For all three tasks, the entire corresponding video of the surgery is provided along with the training data as context information. In the test phase, only the test image along with the preceding 250 frames is provided. See Supplementary file S1.

Appendix C. Submission instructions

The following section provides the instruction document that challenge participants obtained. See Supplementary file S3.

Appendix D. Rankings for all stages

The ranking schemes described in Section 2.3.2 were also computed for stages 1 and 2. To compare the performance of participating teams across stages, stacked frequency plots of the observed ranks, separated by the algorithms, for each ranking of the binary and multi-instance segmentation tasks are displayed in Fig. D.1 to D.8. Observed ranks across bootstrap samples are presented over the three stages the stages. The metric values for the multi-instance detection task are displayed in Table D.1

Table D.1
Results over all stages for the multi-instance detection task.

Team identifier	mAP		
	Stage 1	Stage 2	Stage 3
Uniandes	1.000	0.833	1.000
VIE	0.750	0.778	0.978
caresyntax	0.944	0.833	0.972
SQUASH	0.967	1.000	0.966
fisensee	1.000	1.000	0.964
www	0.900	0.833	0.944

Table D.2
Results over all stages for the multi-instance detection task as reported during the challenge event. Those values have to be interpreted with care due to an implementation error in the validation.

Team identifier	F1-score		
	Stage 1	Stage 2	Stage 3
Uniandes	0.94	0.93	0.91
www	0.92	0.90	0.90
caresyntax	0.92	0.91	0.89
SQUASH	0.90	0.86	0.86
fisensee	0.89	0.89	0.86
VIE	0.84	0.82	0.82

.

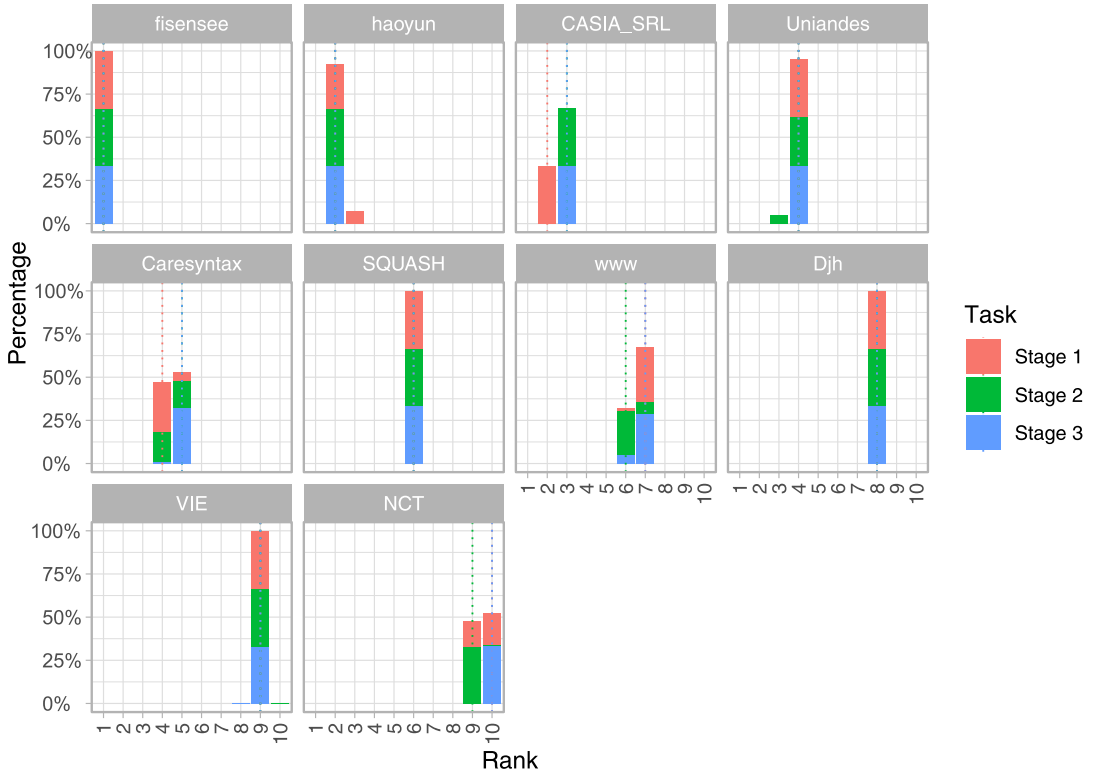


Fig. D.1. Stacked frequency plot for stages 1 to 3 with the Dice Similarity Coefficient (*DSC*) accuracy ranking of the binary segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

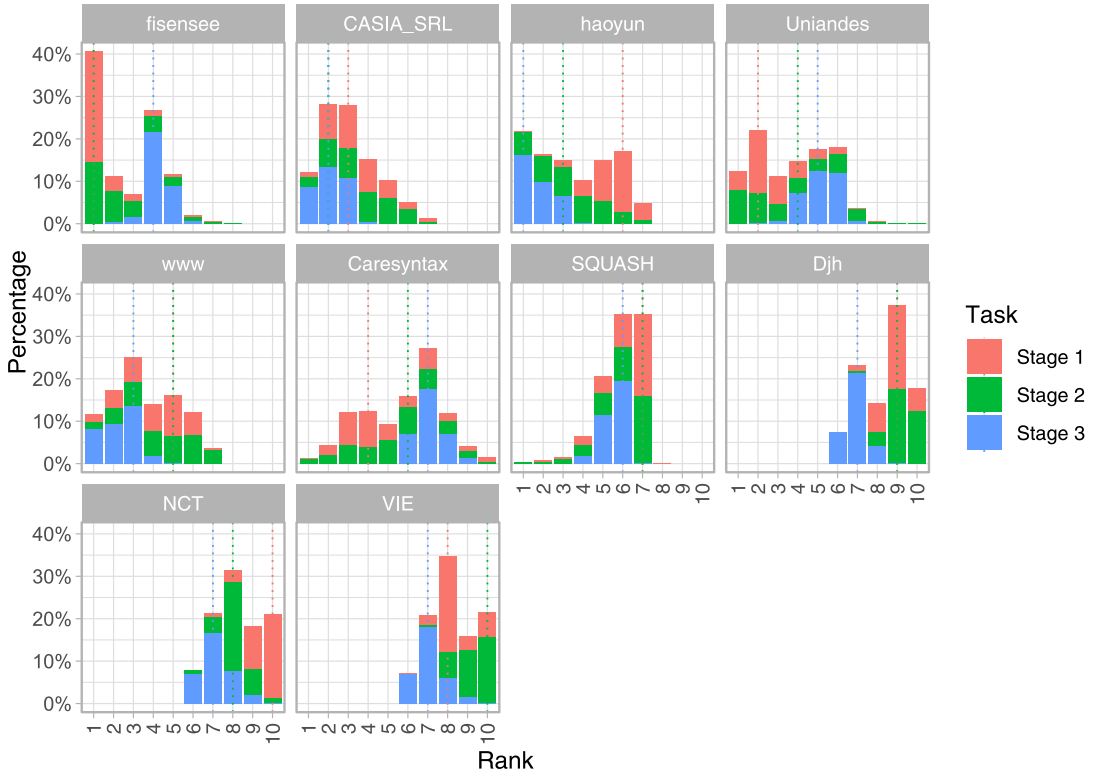


Fig. D.2. Stacked frequency plot for stages 1 to 3 with the Dice Similarity Coefficient (*DSC*) robustness ranking of the binary segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

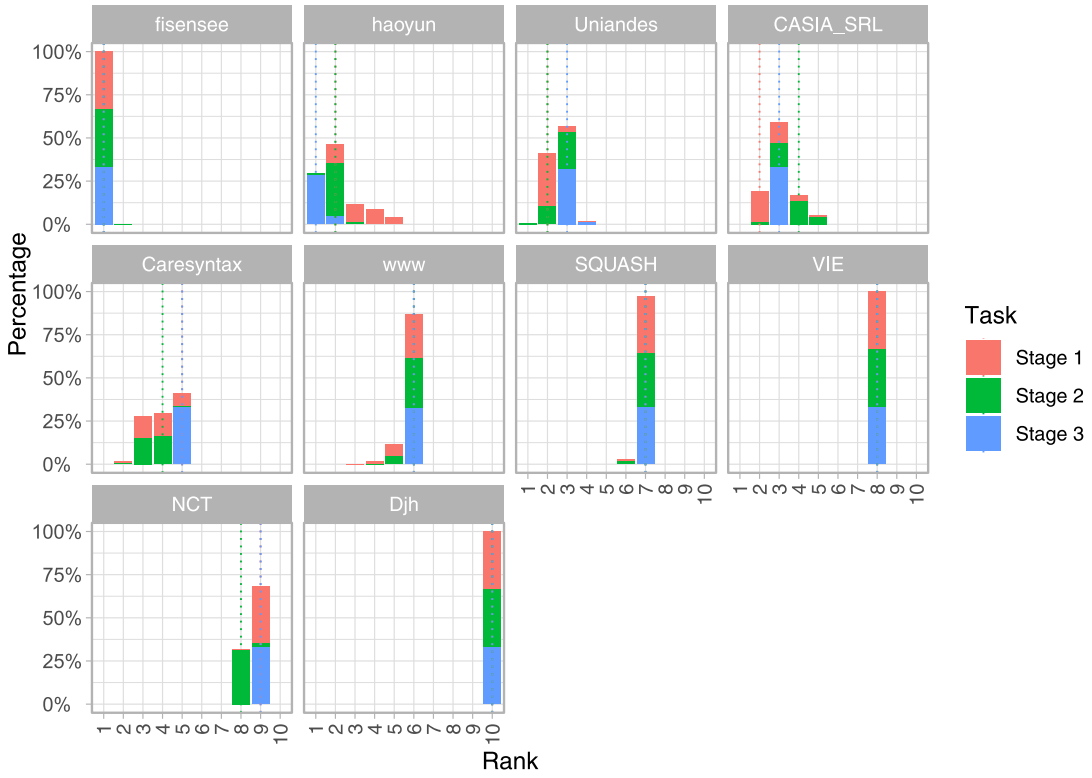


Fig. D.3. Stacked frequency plot for stages 1 to 3 with the Normalized Surface Distance (*NSD*) accuracy ranking of the binary segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

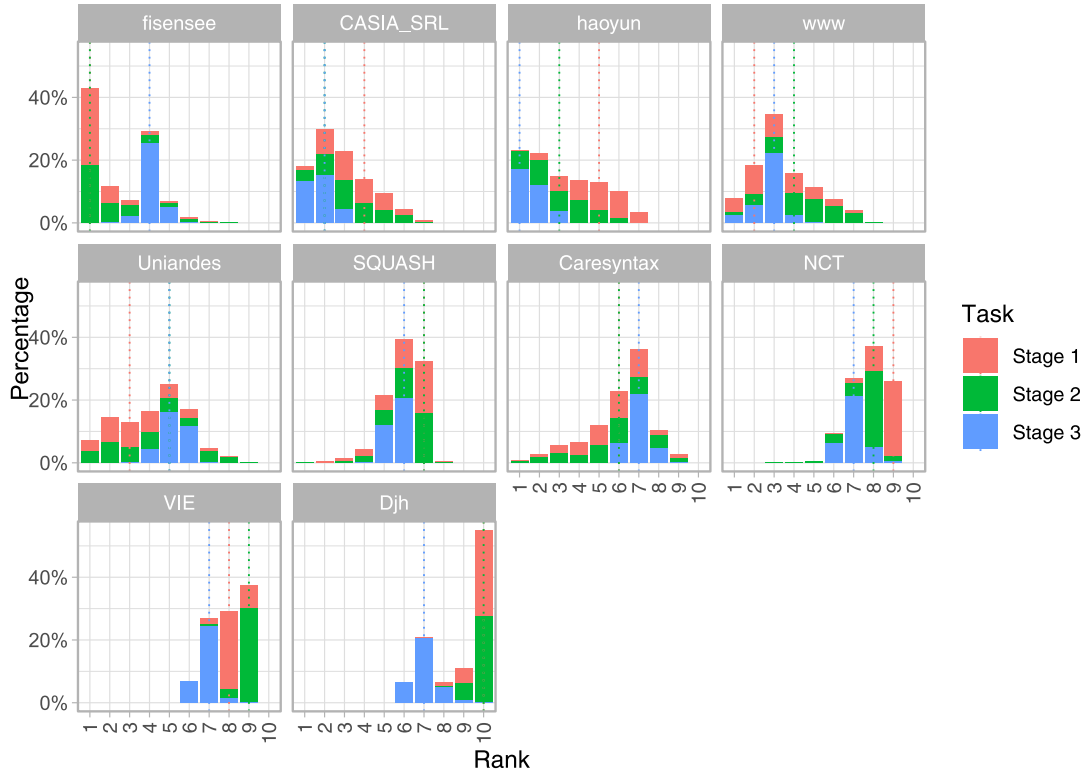


Fig. D.4. Stacked frequency plot for stages 1 to 3 with the Normalized Surface Distance (*NSD*) robustness ranking of the binary segmentation task.

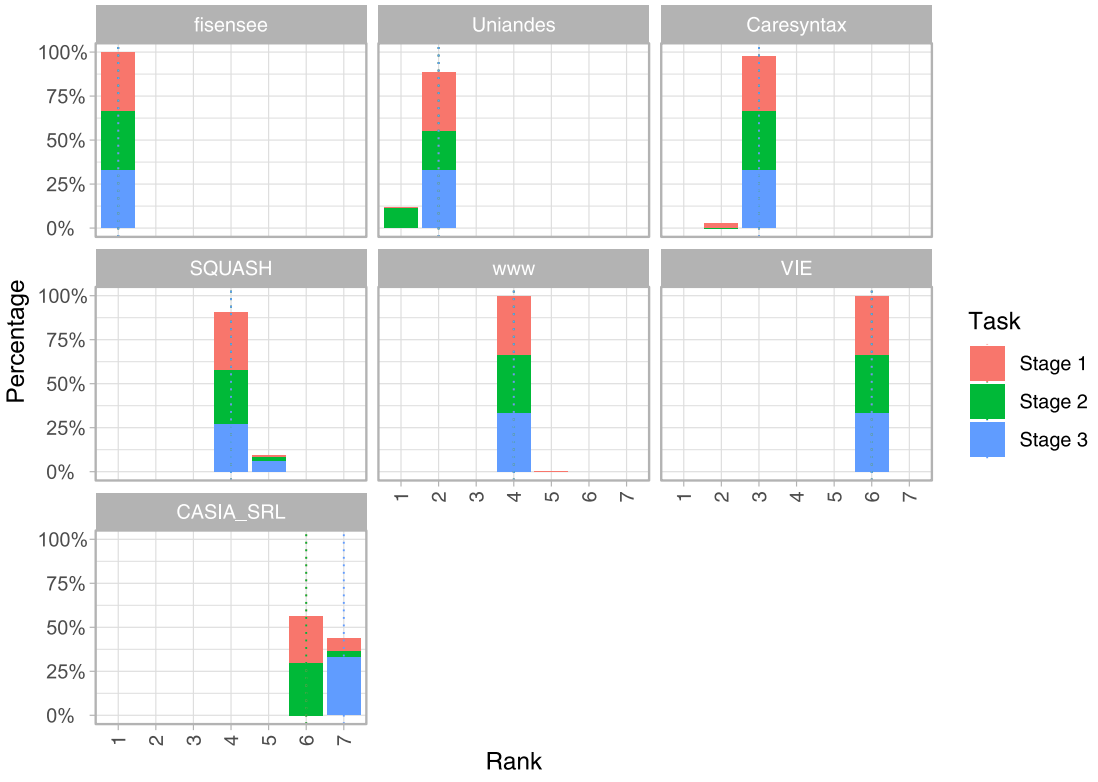


Fig. D.5. Stacked frequency plot for stages 1 to 3 with (multi-instance) Dice Similarity Coefficient (*(MI_)DSC*) accuracy ranking of the multi-instance segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

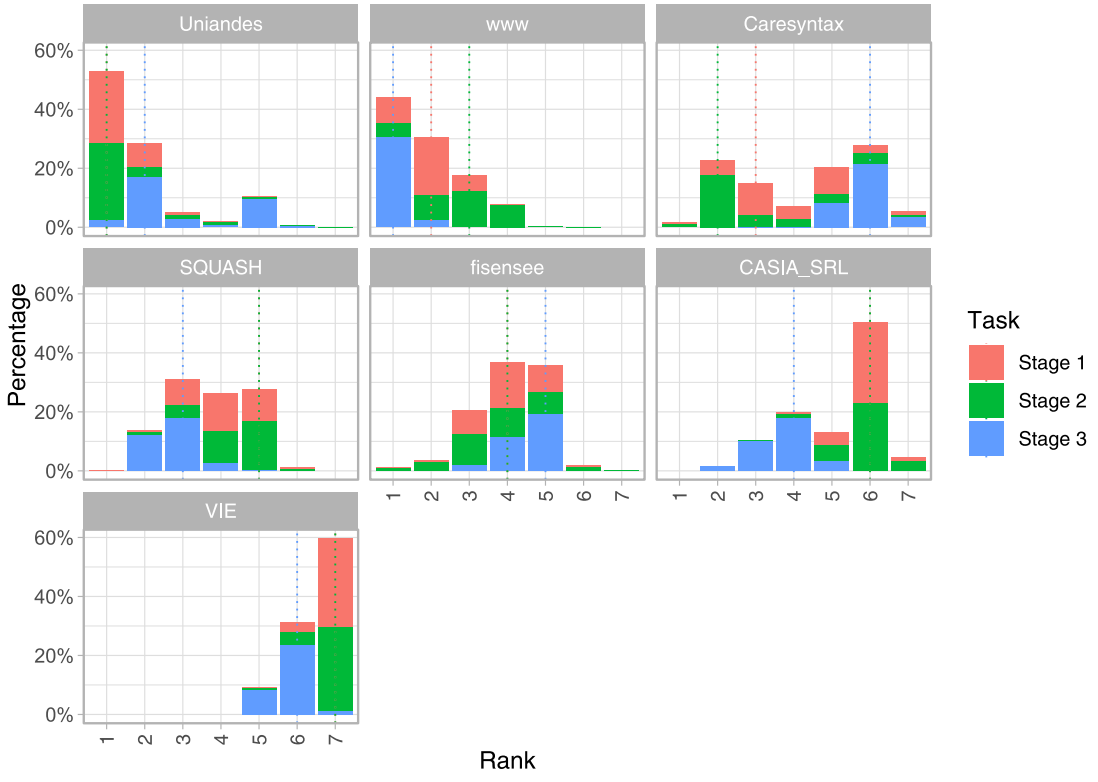


Fig. D.6. Stacked frequency plot for stages 1 to 3 with the (multi-instance) Dice Similarity Coefficient (*(MI_)DSC*) robustness ranking of the multi-instance segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

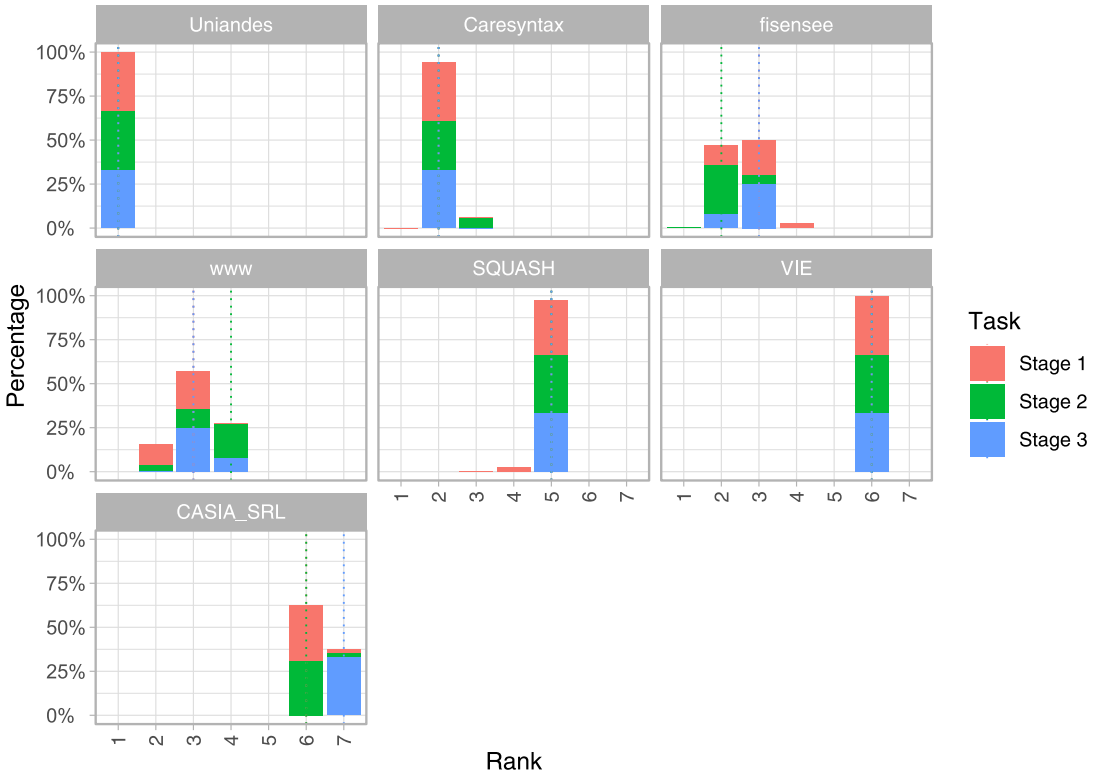


Fig. D.7. Stacked frequency plot for stages 1 to 3 with the (multi-instance) Normalized Surface Distance (*(MI_)NSD*) accuracy ranking of the multi-instance segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

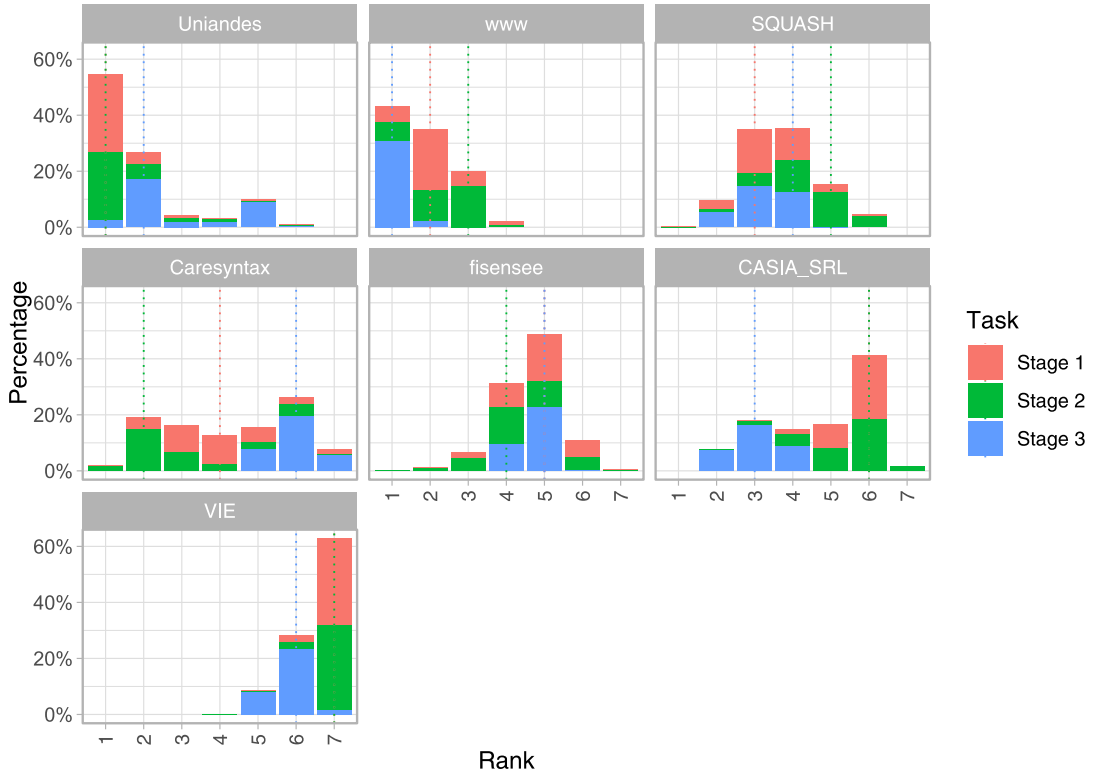


Fig. D.8. Stacked frequency plot for stages 1 to 3 with the (multi-instance) Normalized Surface Distance (*(MI_)NSD*) robustness ranking of the multi-instance segmentation task. The plots were generated using the package challengeR (Wiesenfarth et al., 2019b; 2019a).

Appendix E. Results for stage 2 including expert baseline

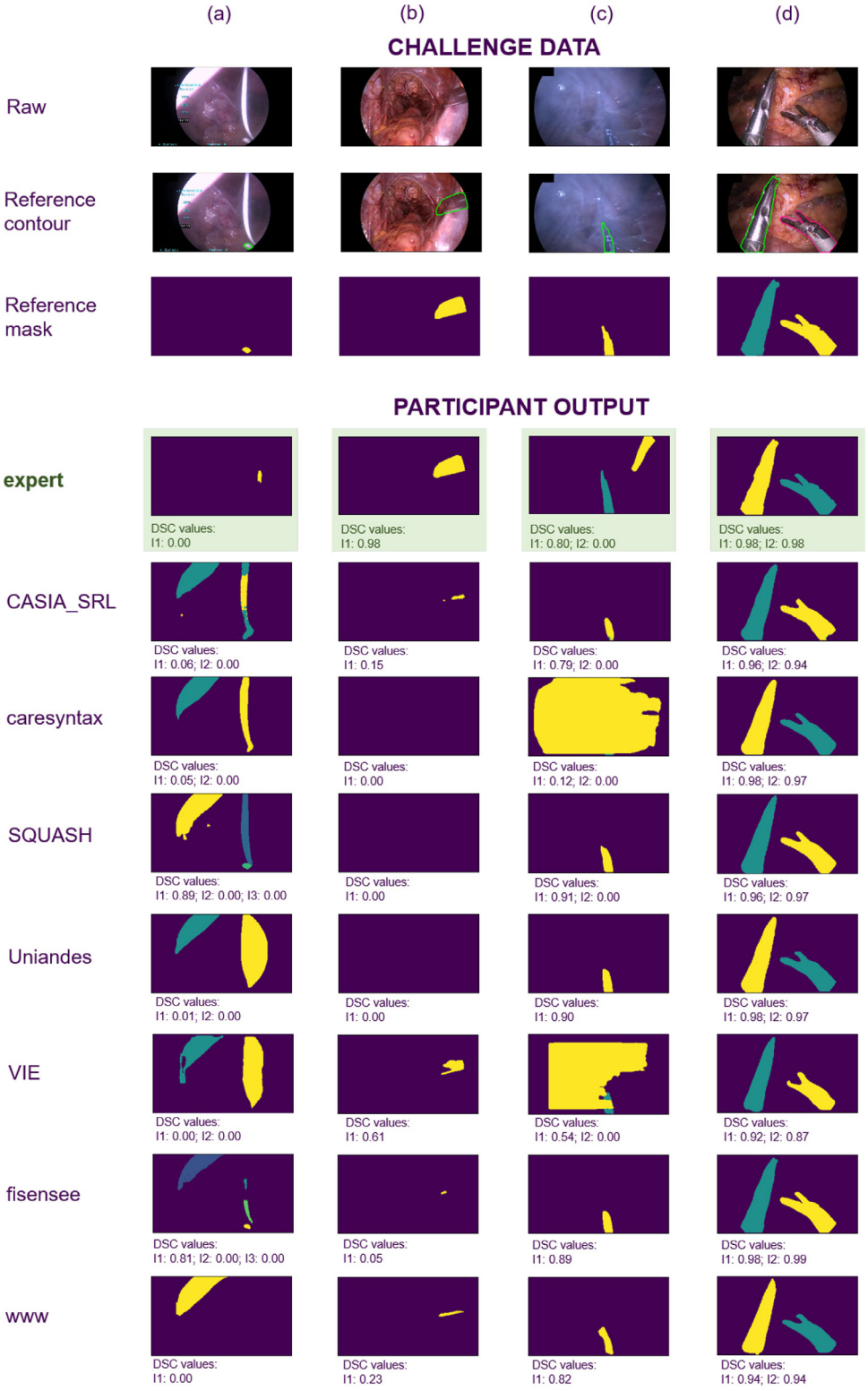


Fig. E.1. Example frames from stage 2 with corresponding participant and expert performances. Each row shows the raw frame, the reference contours and mask as well as the (algorithm) output of the participating teams/expert of the multi-instance segmentation (MIS) task for one representative frame. The columns represent image frames with (a) low expert, low algorithm performances, (b) high expert, low algorithm performances, (c) low expert, high performances, (d) high expert, high algorithm performances, respectively. For participants, the Dice Similarity Coefficient (DSC) values are provided per instrument instance (I_i).

Appendix F. Challenge design document

See Supplementary file S2..

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2020.101920.

References

Allan, Max, Kondo, Satoshi, Bodenstedt, Sebastian, Leger, Stefan, Kadkhodamohammadi, Rahim, Luengo, Imanol, Fuentes, Felix, Flouty, Evangello, Mohammed, Ahmed, Pedersen, Marius, et al., 2020. 2018 Robotic Scene Segmentation Challenge. arXiv:2001.11190.

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.-H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al., 2019. 2017 Robotic instrument segmentation challenge. arXiv:1902.06426.

Amini Khoiy, K., Mirbagheri, A., Farahmand, F., 2016. Automatic tracking of laparoscopic instruments for autonomous control of a cameraman robot. Minimally Invasive Therapy & Allied Technologies 25 (3), 121–128.

Armstrong, T.G., Moffat, A., Webber, W., Zobel, J., 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 601–610.

Bianchi, F., Masaracchia, A., Shojaei Barjuei, E., Menciassi, A., Arezzo, A., Koulaouzidis, A., Stoyanov, D., Dario, P., Ciuti, G., 2019. Localization strategies for robotic endoscopic capsules: a review. Expert Rev. Med. Devices 16 (5), 381–403.

Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., Müller-Stich, B., Ourselin, S., Pakhomov, D., et al., 2018. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. arXiv:1805.02475.

Burström, G., Nachabe, R., Persson, O., Edström, E., Terander, A.E., 2019. Augmented and virtual reality instrument tracking for minimally invasive spine surgery: a feasibility and accuracy study. Spine 44 (15), 1097–1104.

Cardoso, M.J., 2018. Medical segmentation decathlon. http://medicaldecathlon.com/. Accessed: 2019-10-29.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818.

Clevert, D.-A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv:1511.07289.

De Paolis, L.T., De Luca, V., 2019. Augmented visualization with depth perception cues to improve the surgeon's performance in minimally invasive surgery. Medical & biological engineering & computing 57 (5), 995–1013.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. Nat. Med. 25 (1), 24–29.

Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: aretrospective. Int. J. Comput. Vis. 111 (1), 98–136.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88 (2), 303–338.

Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.-A., 2019. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. 14 (9), 1611–1617.

García-Peraza-Herrera, L.C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., Ourselin, S., 2016. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In: International Workshop on Computer-Assisted and Robotic Endoscopy. Springer, pp. 84–95.

Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J., 2019. Ce-net: context encoder network for 2d medical image segmentation. IEEE Trans. Med. Imaging 38 (10), 2281–2292.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Iglovikov, V., Shvets, A., 2018. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv:1801.05746.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2462–2470.

Ioannidis, J.P., 2005. Why most published research findings are false. PLos med 2 (8), e124.

Isensee, F., Maier-Hein, K.H., 2020. Or-unet: an optimized robust residual u-net for instrument segmentation in endoscopic images. arXiv:2004.12668.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. Nnu-net: self-adapting framework for u-net-based medical image segmentation. arXiv:1809.10486.

Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., DeLange, T., Halvorsen, P., Johansen, H.D., 2019. Resunet++: An advanced architecture for medical image segmentation. In: Proceedings of the IEEE International Symposium on Multimedia (ISM). IEEE, pp. 225–2255.

Kiefer, J., Wolfowitz, J., et al., 1952. Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics 23 (3), 462–466.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv:1412.6980.

Kuhn, H.W., 1955. The hungarian method for the assignment problem. Naval research logistics quarterly 2 (1–2), 83–97.

Kurmann, Thomas, Neila, Pablo Marquez, Du, Xiaofei, Fua, Pascal, Stoyanov, Danail, Wolf, Sebastian, Sznitman, Raphael, 2017. Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. International Conference on Medical Image Computing and Computer-Assisted Intervention 505–513.

Laina, I., Rieke, N., Rupprecht, C., Vizcaíno, J.P., Eslami, A., Tombari, F., Navab, N., 2017. Concurrent segmentation and localization for tracking of surgical instruments. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 664–672.

Law, H., Ghani, K., Deng, J., 2017. Surgeon technical skill assessment using computer vision based analysis. In: Machine learning for healthcare conference, pp. 88–99.

Lin, S., Qin, F., Bly, R.A., Moe, K.S., Hannaford, B., 2019. Automatic sinus surgery skill assessment based on instrument segmentation and tracking in endoscopic video. In: International Workshop on Multiscale Multimodal Medical Imaging. Springer, pp. 93–100.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European conference on computer vision. Springer, pp. 740–755.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. 9 (1), 5217.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbuary, A., Jannin, P., Müller, H., Onogur, S., et al., 2019. Bias: transparent reporting of biomedical image analysis challenges. arXiv:1910.04071.

Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al., 2017. Surgical data science for next-generation interventions. Nat. Biomed. Eng. 1 (9), 691–696.

Nguyen, X.A., Ljuhar, D., Pacilli, M., Nataraja, R.M., Chauhan, S., 2019. Surgical skill levels: classification and analysis using deep neural network model and motion signals. Comput. Methods Programs Biomed. 177, 1–8.

Ni, Z.-L., Bian, G.-B., Wang, G.-A., Zhou, X.-H., Hou, Z.-G., Xie, X.-L., Li, Z., Wang, Y.-H., 2020. Barnet: bilinear attention network with adaptive receptive field for surgical instrument segmentation. arXiv:2001.07093.

Ni, Z.-L., Bian, G.-B., Xie, X.-L., Hou, Z.-G., Zhou, X.-H., Zhou, Y.-J., 2019. Rasnet: Segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 5735–5738.

Nikolov, S., Blackwell, S., Mendes, R., De Fauw, J., Meyer, C., Hughes, C., Askham, H., Romera-Paredes, B., Karthikesalingam, A., Chu, C., et al., 2018. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv:1809.04430.

Pakhomov, Daniil, Premachandran, Vittal, Allan, Max, Azizian, Mahdi, Navab, Nassir, 2019. Deep residual learning for instrument segmentation in robotic surgery. International Workshop on Machine Learning in Medical Imaging 566–573.

Panch, T., Mattie, H., Celi, L.A., 2019. The 'inconvenient truth' about ai in healthcare. Npj Digital Medicine 2 (1), 1–3.

Qayyum, A., Qadir, J., Bilal, M., Al-Fuqaha, A., 2020. Secure and robust machine learning for healthcare: asurvey. arXiv:2001.08103.

Recital26, 2016. General data protection regulation of the european union. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679_d1e1374-1-1. Accessed: 2019-10-29.

Reinke, A., Eisenmann, M., Onogur, S., Stankovic, M., Scholz, P., Full, P.M., Bogunovic, H., Landman, B.A., Maier, O., Menze, B., et al., 2018. How to exploit weaknesses in biomedical challenge design and organization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 388–395.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Roß, T., Reinke, A., 2019. Robustmis2019. https://phabricator.mitk.org/source/rmis2019/. Accessed: 2019-10-29.

Roß, T., Reinke, A., Maier-Hein, L., 2019a. Robust medical instrument segmentation (ROBUST-MIS) challenge (grand-challenge.org). https://robustmis2019.grand-challenge.org/. Accessed: 2019-10-29.

Roß, T., Reinke, A., Maier-Hein, L., 2019b. Robust medical instrument segmentation (ROBUST-MIS) challenge (synapse.org). https://www.synapse.org/_!Synapse:syn18779624/wiki/. Accessed: 2019-10-29.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., Schmidt, L., 2020. Evaluating machine accuracy on imagenet. In: International Conference on Machine Learning (ICML).
Shapiro, L.G., 1996. Connected Component Labeling and Adjacency Graph Construction. In: Machine Intelligence and Pattern Recognition, 19. Elsevier, pp. 1–30.
Siddaiah-Subramanya, M., Tiang, K.W., Nyandowe, M., 2017. A new era of minimally invasive surgery: progress and development of major technical innovations in general surgery over the last decade. The Surgery Journal 3 (04), e163–e166.
Su, Y.-H., Huang, K., Hannaford, B., 2018. Real-time vision-based surgical tool segmentation with robot kinematics prior. In: 2018 International Symposium on Medical Robotics (ISMR). IEEE, pp. 1–6.
Wang, R., Zhang, M., Meng, X., Geng, Z., Wang, F.-Y., 2017. 3-D tracking for augmented reality using combined region and dense cues in endoscopic surgery. IEEE J. Biomed. Health Inform. 22 (5), 1540–1551.
Wiesenfarth, M., Reinke, A., Landman, B.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2019a. challengeR: Methods and open-source toolkit for analyzing and visualizing challenge results. https://github.com/wiesenfa/challengeR. Accessed: 2019-10-29.
Wiesenfarth, M., Reinke, A., Landman, B.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2019. Methods and open-source toolkit for analyzing and visualizing challenge results. arXiv:1910.05121.
Zhang, J., Gao, X., 2020. Object extraction via deep learning-based marker-free tracking framework of surgical instruments for laparoscope-holder robots. Int. J. Comput. Assist. Radiol. Surg. 15 (8), 1335–1345.
Zhao, Z., Chen, Z., Voros, S., Cheng, X., 2019. Real-time tracking of surgical instruments based on spatio-temporal context and deep learning. Computer Assisted Surgery 24 (sup1), 20–29.
Zhou, S.K., Rueckert, D., Fichtinger, G., 2019. Handbook of medical image computing and computer assisted intervention. Academic Press.