Comparative validation of surgical phase recognition, instrument keypoint estimation, and instrument instance segmentation in endoscopy: Results of the PhaKIR 2024 challenge

Tobias Rueckert [a,b,c,*], David Rauber [a], Raphaela Maerkl [a], Leonard Klausmann [a,d,e], Suemeyye R. Yildiran [a,d,e], Max Gutbrod [a], Danilo Weber Nunes [a], Alvaro Fernandez Moreno [f,g], Imanol Luengo [f], Danail Stoyanov [f,g], Nicolas Toussaint [f], Enki Cho [h], Hyeon Bae Kim [h], Oh Sung Choo [h], Ka Young Kim [h], Seong Tae Kim [h], Gonçalo Arantes [i], Kehan Song [j], Jianjun Zhu [j], Junchen Xiong [j], Tingyi Lin [j], Shunsuke Kikuchi [k], Hiroki Matsuzaki [k], Atsushi Kouno [k], João Renato Ribeiro Manesco [l], João Paulo Papa [l], Tae-Min Choi [m], Tae Kyeong Jeong [m], Juyoun Park [m], Oluwatosin Alabi [n], Meng Wei [n], Tom Vercauteren [n], Runzhi Wu [o], Mengya Xu [o], An Wang [o], Long Bai [o], Hongliang Ren [o], Amine Yamlahi [p], Jakob Hennighausen [p], Lena Maier-Hein [p], Satoshi Kondo [q], Satoshi Kasai [r], Kousuke Hirasawa [s], Shu Yang [t], Yihui Wang [t], Hao Chen [t,u,v], Santiago Rodríguez [w], Nicolás Aparicio [w], Leonardo Manrique [w], Juan Camilo Lyons [w], Olivia Hosie [w], Nicolás Ayobi [w], Pablo Arbeláez [w], Yiping Li [x], Yasmina Al Khalil [x], Sahar Nasirihaghighi [x,y], Stefanie Speidel [z,aa], Daniel Rueckert [ab,ac], Hubertus Feussner [c,ad], Dirk Wilhelm [c,ad], Christoph Palm [a,d,e]

a Regensburg Medical Image Computing (ReMIC), OTH Regensburg, Regensburg, Germany
b AKTORmed Robotic Surgery, Neutraubling, Germany
c Research Group MITI, TUM University Hospital, School of Medicine and Health, Technical University of Munich, Munich, Germany
d Regensburg Center of Biomedical Engineering (RCBE), OTH Regensburg and Regensburg University, Regensburg, Germany
e Regensburg Center of Health Sciences and Technology (RCHST), OTH Regensburg, Regensburg, Germany
f AI Centre of Excellence, Medtronic Ltd., Watford, UK
g Engineering Sciences, University College London, London, UK
h Augmented Intelligence Lab, Kyung Hee University, Seoul, South Korea
i University of Minho, Braga, Portugal
j Hanglok Tech, Zhuhai City, China
k Jmees Inc., Kashiwa City, Japan
l School of Sciences, São Paulo State University (UNESP), Bauru, Brazil
m KIST HARILAB, Center for Humanoid Research, Artificial Intelligence and Robot Institute, Korea Institute of Science and Technology (KIST), Seoul, South Korea
n King's College London, London, UK
o The Chinese University of Hong Kong, Hong Kong, China
p Division of Intelligent Medical Systems, German Cancer Research Center (DKFZ), Heidelberg, Germany
q Muroran Institute of Technology, Hokkaido, Japan
r Niigata University of Health and Welfare, Niigata, Japan
s Konica Minolta, Inc., Osaka, Japan
t Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
u Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
v HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen, China
w Center for Research and Formation in Artificial Intelligence (CinfonIA), Los Andes University, Bogota, Colombia

* Corresponding author.
   E-mail address: tobias.rueckert@re-mic.de (T. Rueckert).

[x] *Department of Biomedical Engineering, Medical Image Analysis Group, Eindhoven University of Technology, Eindhoven, Netherlands*
[y] *Institute of Information Technology (ITEC), Klagenfurt University, Klagenfurt, Austria*
[z] *Center for Tactile Internet with Human-in-the-loop (CeTI), TU Dresden, Dresden, Germany*
[aa] *Department of Translational Surgical Oncology, National Center for Tumor Diseases (NCT/UCC), Partner Site Dresden, Dresden, Germany*
[ab] *Chair for AI in Healthcare and Medicine, Technical University of Munich (TUM) and TUM University Hospital, Munich, Germany*
[ac] *Biomedical Image Analysis Group, Department of Computing, Imperial College London, London, UK*
[ad] *Department of Surgery, TUM University Hospital, School of Medicine and Health, Technical University of Munich, Munich, Germany*

ARTICLE INFO

ABSTRACT

Reliable recognition and localization of surgical instruments in endoscopic video recordings are foundational for a wide range of applications in computer- and robot-assisted minimally invasive surgery (RAMIS), including surgical training, skill assessment, and autonomous assistance. However, robust performance under real-world conditions remains a significant challenge. Incorporating surgical context – such as the current procedural phase – has emerged as a promising strategy to improve robustness and interpretability.

To address these challenges, we organized the Surgical Procedure Phase, Keypoint, and Instrument Recognition (PhaKIR) sub-challenge as part of the Endoscopic Vision (EndoVis) challenge at MICCAI 2024. We introduced a novel, multi-center dataset comprising thirteen full-length laparoscopic cholecystectomy videos collected from three distinct medical institutions, with unified annotations for three interrelated tasks: surgical phase recognition, instrument keypoint estimation, and instrument instance segmentation. Unlike existing datasets, ours enables joint investigation of instrument localization and procedural context within the same data while supporting the integration of temporal information across entire procedures.

We report results and findings in accordance with the BIAS guidelines for biomedical image analysis challenges. The PhaKIR sub-challenge advances the field by providing a unique benchmark for developing temporally aware, context-driven methods in RAMIS and offers a high-quality resource to support future research in surgical scene understanding.

1. Introduction

In recent years, significant progress has been made in the field of computer- and robot-assisted minimally invasive surgery (RAMIS) with the aim of supporting the surgical team during laparoscopic interventions with innovative assistance systems (Dagnino and Kundrat, 2024). The foundation for the development of such applications often relies on the determination of the current phase of a procedure, the segmentation of the surgical instruments in endoscopic images, and the localization of certain keypoints and the associated estimation of the instruments' pose. Knowing the current intervention phase offers further possibilities, such as improving safety in the operating room (OR) by providing early context-sensitive warnings (Katić et al., 2013), by optimizing the OR management (Tanzi et al., 2020), or by providing accurate procedure time predictions (Aksamentov et al., 2017), e. g., for anesthesia planning. Segmentation of surgical instruments provides a variety of application areas, such as surgical navigation systems (Ahmed et al., 2024), surgical skill assessment (Lam et al., 2022), or autonomous endoscope guidance (Gruijthuijsen et al., 2022). Based on the localization of certain keypoints that describe the pose of an instrument, the distance to risk structures can be determined (Bernhardt et al., 2017) and the assessment of surgical skills of the surgeon can further be automated (Law et al., 2017). Purely image-based processing of endoscopic data is of particular importance for all three tasks, as no changes need to be made to the surgical environment and the procedure workflow, allowing a simple integration into existing clinical practice (Rueckert et al., 2024).

In the past, the development of methods was often based on the collection and annotation of application-specific data. In order to avoid the disadvantages of this procedure, such as the lack of reproducibility or the inability to compare the results between the methods, the concept of *challenges* was introduced. The most well-known platform for conducting such challenges in the area of medical image processing is the annual conference of the Medical Image Computing and Computer Assisted Intervention (MICCAI) society, which enables the organization of a large number of challenges for various medical applications. One of them is the Endoscopic Vision Challenge (EndoVis), which has been held annually since 2015. It coordinates a series of sub-challenges, all focused exclusively on endoscopic image data. Within the domain of laparoscopic phase recognition, three dedicated sub-challenges have been

conducted to date (Speidel et al., 2017; Bodenstedt et al., 2021; Wagner et al., 2023). For the task of surgical instrument segmentation, eight sub-challenges have been performed (Bodenstedt et al., 2018; Allan et al., 2019, 2020; Roß et al., 2021; Bodenstedt et al., 2021; Psychogyios et al., 2024; Malpani and Glock, 2023; Bodenstedt et al., 2023). As for surgical instrument keypoint estimation, one dedicated sub-challenge has been held so far (Bodenstedt et al., 2018).

However, the datasets provided by these sub-challenges do not cover all the necessary aspects required for the development of real-world applications. For the phase recognition task, only two out of three papers present a dataset that originates from multiple centers (Bodenstedt et al., 2021; Wagner et al., 2023). In the context of the instrument segmentation task, several challenges do not differentiate between distinct instrument classes (Bodenstedt et al., 2018; Allan et al., 2020; Roß et al., 2021; Bodenstedt et al., 2021; Malpani and Glock, 2023) or between individual instances within the same class (Bodenstedt et al., 2018; Allan et al., 2019, 2020; Bodenstedt et al., 2021; Psychogyios et al., 2024; Malpani and Glock, 2023). This lack of distinction hampers the ability to accurately identify overlapping instruments and complicates the task of tracking the tools over time, which relies on the consistent identification of each instrument across successive frames. The recordings of some sub-challenges operate on porcine tissue (Allan et al., 2019, 2020), which significantly simplifies the recognition of surgical instruments compared to human tissue (Allan et al., 2020), or do not use any biological tissue at all (Malpani and Glock, 2023). Some datasets only provide single frames (Allan et al., 2019, 2020; Bodenstedt et al., 2021, 2023; Malpani and Glock, 2023) or short clips (Roß et al., 2021) instead of full video sequences, precluding the use of temporal information for the segmentation task. Information on relevant instrument keypoints is available in only one sub-challenge (Bodenstedt et al., 2018). However, for in-vivo recordings involving manual instruments, only the coordinates of a single keypoint and the instrument orientation are provided. Meanwhile, sequences featuring robotic instruments are based on ex-vivo data from animal tissue, representing highly simplified movements that do not account for real-world surgical complexities. Additionally, none of the sub-challenge datasets for the segmentation of surgical instruments or the keypoint estimation task include multi-center data, thereby limiting the ability to evaluate the participant methods across different clinical environments and surgical tools. An overview regarding previous related
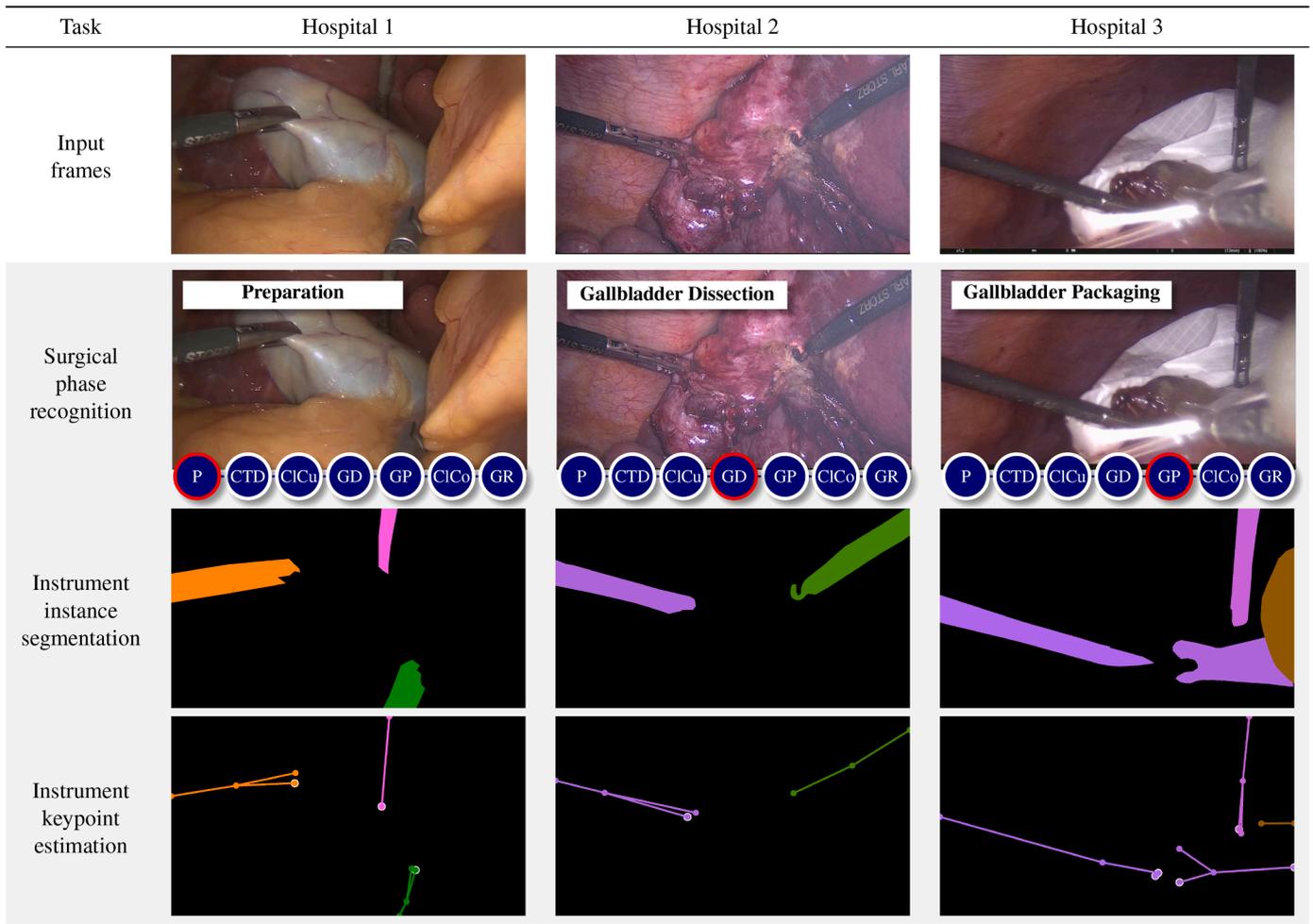
Task	Hospital 1	Hospital 2	Hospital 3

Fig. 1. Visualization of the PhaKIR tasks and annotations for each of the three medical centers. For the phase recognition task, the Phases Preparation (P), Calot Triangle Dissection (CTD), Clipping and Cutting (ClCu), Gallbladder Dissection (GD), Gallbladder Packaging (GP), Cleaning and Coagulation (ClCo), and Gallbladder Retraction (GR) are shown. For the instrument instance segmentation task, the color-encoded masks are presented. For the instrument keypoint estimation task, a visualization of the keypoint coordinates is depicted, including hidden keypoints surrounded by white.

EndoVis sub-challenges together with the conducted tasks, the number of participating medical centers, and key properties of the challenges and datasets is given in Table 1.

To address these challenges, we provided a dataset comprising full endoscopic video sequences of thirteen human cholecystectomies from three medical centers, enabling the incorporation of temporal information in method development as well as the consideration of real-world conditions. The annotations span three tasks: determining the surgical phase of an intervention, performing pixel-precise instance segmentation of surgical instruments, and localizing specific keypoints of these instruments. This setup supports both the recognition of surgical instruments and the understanding of surgical context in one dataset. An overview of the tasks and the associated annotations is shown in Fig. 1.

Our paper is structured according to the transparent reporting of biomedical image analysis challenges (BIAS) guidelines published by Maier-Hein et al. (2020). As our challenge consists of three tasks, we refrain from using an overarching methods section and instead present the Sections 2–5 separately to ensure a clear structure.

The appendix contains information regarding the challenge organizers (see Appendix A), the data description and labeling instructions (see Appendix B), the submission instructions for the participants (see Appendix C), as well as the challenge design document (see Appendix D).

2. Challenge organization

As part of the MICCAI 2024 conference, which took place in Marrakech, Morocco, from October 6 to 10, we conducted our sub-challenge entitled PhaKIR: Surgical Procedure **Pha**se, **K**eypoint, and **I**nstrument **R**ecognition as one of a total of eight EndoVis sub-challenges. The number of registered participants and submissions for each individual sub-challenge is shown in Fig. 2.

Three parties were involved in the challenge organization (see Appendix A). The research group for Minimally Invasive Interdisciplinary Therapeutic Intervention (MITI)[1] at the Technical University of Munich (TUM) obtained the source data from the three different medical centers and created the annotations (see Appendix B). The Regensburg Medical Image Computing (ReMIC)[2] laboratory at the Ostbayerische Technische Hochschule Regensburg (OTH Regensburg) carried out the technical implementation, which included the pre-processing and the provision of the data, the registration and administration of the participants, the submission process, and the evaluation and presentation of the results. The company AKTORmed - Robotic Surgery[3] provided

---

Table 1

Overview of related previous EndoVis sub-challenges including the conducted tasks Phase Recognition (PR), Instrument Segmentation (IS), and Keypoint Estimation (KE), the number of participating medical centers (#Med. centers), and key properties of the challenges and datasets, organized according to the task sequence presented in Section 3.

Challenge	Task			#Med.	Challenge / dataset properties
	PR	IS	KE	centers	
Speidel et al. (2017)	✓			1	single-centered, single-task, methods no longer considered state of the art
Wagner et al. (2023)	✓			3	single-task, strong focus on one leading, more traditional architecture
Bodenstedt et al. (2021)	✓	✓		3	no instrument classification, no instance segmentation, single frames
Allan et al. (2019)		✓		1	single-centered, single-task, porcine tissue, no instance segmentation, single frames
Allan et al. (2020)		✓		1	single-centered, single-task, porcine tissue, no instrument classification, no instance segmentation, single frames
Roß et al. (2021)		✓		1	single-centered, single-task, no instrument classification, short clips
Malpani and Glock (2023)		✓		-	single-centered, single-task, synthetic data, single frames
Bodenstedt et al. (2023)		✓		1	single-centered, single-task, single frames
Psychogyios et al. (2024)		✓		1	single-centered, single-task, no instance segmentation
Bodenstedt et al. (2018)		✓	✓	1	single-centered, no instrument classification, no instance segmentation
PhaKIR (2025)	✓	✓	✓	3	multi-centered, multi-task, instrument classification, instance segmentation, real-world surgeries, full sequences

the prize money of €3,000 which was distributed equally across the three tasks. The sub-challenge is designed as a one-time event with a fixed submission deadline, after which no submissions are possible. As soon as the joint sub-challenge paper is published, the training dataset will be publicly available. The sub-challenge platform can be accessed at https://phakir.re-mic.de/ and was hosted entirely on infrastructure set up and managed by the ReMIC laboratory (Klausmann et al., 2025), i.e., no established platforms such as Synapse or Grand Challenge were used.

With respect to the participation policies, only fully automated methods were permitted. Participants were restricted to using only the training data provided by the challenge organizers and publicly available datasets. In addition, the use of networks pre-trained on publicly available datasets was allowed. While members of the organizing institutions were permitted to participate, they were not eligible to receive any awards. The results of all teams were presented at the MICCAI 2024 EndoVis challenge event on October 10, 2024 and were subsequently published on the sub-challenge website. All results and further analysis from participating teams are included in this joint challenge publication, with all contributing members of each team listed as co-authors. Authors of submitted methods were not permitted to publish any results prior to the release of this joint challenge paper.

Information on the submission process (see Appendix C) was provided to the participants both directly on the website[4] and in a separate GitHub repository[5]. While the instructions on the website are more concerned with organizational aspects, the repository explanations provide more technical details, such as the structure of the input and desired output data of the algorithms, information on the Docker submission, on the submission process in general, and on the metrics used. A Dockerfile, together with sample data and an exemplary network model, was also made available to the participants via this GitHub repository. The submissions were provided as Docker containers, which were manually reviewed by the organizers to verify both technical functionality and compliance with the required output format. Submissions with missing results for test cases were not permitted. Participants were notified of the status of their submission by email and were allowed to update their models as often as needed until the final deadline.

The sub-challenge was structured into four distinct phases: registration, training, submission, and evaluation. The most important dates corresponding to these phases are shown in Fig. 3. Key milestones included the opening of the registration on April 29, the start of the training phase on May 27, the beginning of the submission phase on September 8, the submission deadline on September 24, and the publication

of the results at the EndoVis event at MICCAI 2024 and online on the website on October 10.

Regarding the dataset, only anonymized data collected in an ethically approved research project approved by the local ethics committee of the TUM (approval code 337/21 S-EB) was used. During the registration phase, each participant was required to sign the EndoVis rules document, which specifies that the provided data may only be used for the purposes of the challenge, must not be shared with third parties and cannot be used for commercial purposes. Use of the data for individual publications by participants is permitted only after the release of this joint challenge publication, with notification provided via email and through the official website. The dataset is published under a Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA)[6] license, which means that it can be used for non-commercial purposes once the sub-challenge has been conducted and this paper has been released. Regarding code availability, participants were provided with Docker container templates, including functional example code and example outputs. While sharing the code for their submissions was not mandatory, participants were encouraged to do so. Apart from the company AKTORmed, no additional sponsors or supporters were involved in the challenge. Access to the test data was strictly limited: only a predefined group within the MITI research group responsible for acquiring the raw data and annotations, and a designated team at the ReMIC laboratory, tasked with data provisioning and evaluation, were granted access.

3. Mission of the challenge

The objective of our PhaKIR sub-challenge was to make a contribution in the field of RAMIS by providing a high-quality, comprehensive real-world dataset of minimally invasively performed cholecystectomies, along with the benchmarking of various algorithms on this dataset. We defined the target cohort as all individuals around the world with gallbladder-related problems that require cholecystectomy. For this purpose, the challenge cohort consisted of adult humans of different ages and genders who had undergone cholecystectomy at three medical centres in Germany. The data collected consists of endoscopic video recordings of the abdominal cavity during interventions, without any additional clinical patient data. The challenge comprised the following three tasks, in which participants were allowed to take part either individually or as part of a team and could choose to participate in one or multiple tasks. Each task was evaluated individually.

- Surgical phase recognition: Classification of the surgical phases of a cholecystectomy as accurately as possible.

---

[4] https://phakir.re-mic.de/participation/
[5] https://github.com/schnadoslin/PhaKIR_Submission_Template/

[6] https://creativecommons.org/licenses/by-nc-sa/4.0/

Fig. 2. Number of participants that registered and submitted for each of the eight individual EndoVis-2024 sub-challenges, sorted in ascending order based on the number of registrations.

- Instrument instance segmentation: Segmentation of the contours of the surgical instruments as precisely as possible through pixel-accurate predictions and distinguishing different instrument classes and instances of the same instrument class.
- Instrument keypoint estimation: Location of certain keypoints on the surgical instruments as precisely as possible through pixel-accurate coordinates, and distinguishing different instrument classes and instances of the same instrument class, considering that the number of keypoints depends on the type of the individual instrument.

A central goal across all challenge tasks was the development of algorithms robust to real-world surgical conditions. These include sudden smoke development, image blur, motion blur due to rapid instrument movement, occlusions from tissue and organ overlap, and reduced visibility caused by significant bleeding. Another key objective was to leverage the temporal information inherent in the provided video sequences. By moving beyond single-frame, participants were encouraged to improve both the accuracy and the robustness of their approaches in the face of these challenging conditions.

4. Challenge dataset

4.1. Data collection

A total of $n = 13$ video sequences ranging between 23 and 60 minutes in duration and showing laparoscopic cholecystectomies were recorded with different endoscope systems from three surgical centers in Germany, namely the TUM University Hospital Rechts der Isar[7] ($n = 9$), the Heidelberg University Hospital[8] ($n = 2$), and the Weilheim-Schongau Hospital[9] ($n = 2$). For the sake of readability, we will refer to these three hospitals as hospital 1, hospital 2, and hospital 3 in the remainder of this article. Heidelberg recordings have already been published as part of the Surgical Workflow and Skill Analysis Challenge (HeiChole Benchmark, Wagner et al., 2023) and comply with the video files entitled `HeiChole2.mp4` and `HeiChole3.mp4`. All sequences were recorded with monocular laparoscopic cameras, a resolution of 1920×1080 pixels, and a frame rate of 25 frames per second (FPS). A video is recorded from the moment the camera is first inserted into the patient's abdomen and ends when the laparoscope is removed from the abdominal cavity for the last time.

---

[7] https://www.mri.tum.de/
[8] https://www.klinikum.uni-heidelberg.de/
[9] https://www.meinkrankenhaus2030.de/

4.2. Training and test case definition

The training and test cases each corresponded to a single video frame of a cholecystectomy recording and were structured as individual, independent samples. Neither video frames nor labels were being provided for the test cases. The participants were provided with the full-length sequences of the interventions. Care was taken to ensure that full sequences were used either for training or for testing and that no subdivision was made within a video. From hospital 1, $n = 6$ videos were randomly selected for training and the remaining $n = 3$ videos were used for testing. The sequences from the other two surgical centers were divided equally into training and test data. This approach was intended to ensure that each center was represented in the test data, thus requiring a high degree of generalizability of the participants' methods. The recordings were named consistently, with `Video_01`, `Video_02`, …, `Video_09` originated from hospital 1, the sequences `Video_11` and `Video_12` were taken from hospital 2, and the recordings `Video_13` and `Video_14` were recorded at hospital 3. It is important to note that the file entitled `Video_10` was initially considered, but was excluded before the start of the challenge due to the significantly different instrumentation used, which varied greatly from that of the other videos. The test dataset therefore consisted of the sequences `Video_06`, `Video_08`, `Video_09`, `Video_12`, `Video_14`. To ensure the anonymity of the patient and the surgical staff, all frames recorded outside the abdominal cavity (e.g., to clean the laparoscope) were removed from the videos. For each video sequence, the participants were provided with a file containing the frame number after which such outside abdomen frames were extracted. For the phase recognition task, every individual frame was annotated, whereas for the instrument instance segmentation and the instrument keypoint estimation tasks, annotations were provided for one frame per second. The participants were therefore provided with 485,875 frames for training, with the same number of labels for the phase recognition task, as well as with instance segmentation and keypoint annotations for 19,435 frames. The hidden test data contained 323,025 images, with an identical number of phase recognition labels and segmentation and keypoint labels for 12,921 frames.

4.3. Data annotation

The data was annotated by a team of four medical professionals, with the head of the group being an expert surgeon with many years of experience and the remaining members being students with a medical background. The generation of the annotations followed specific rules
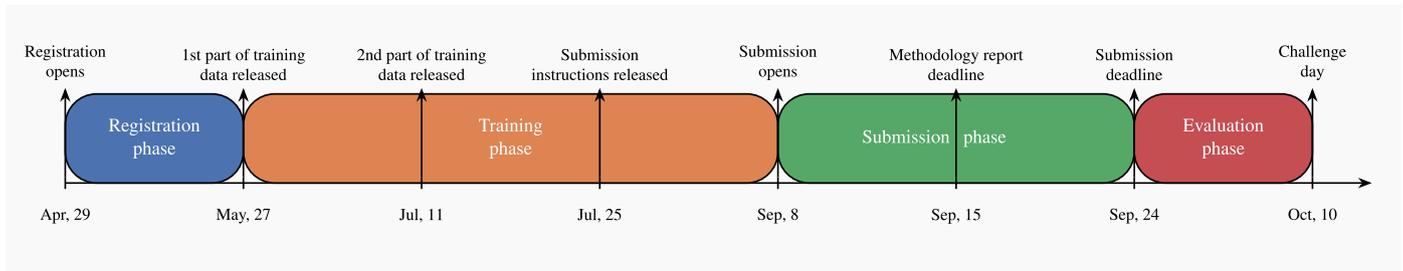
Fig. 3. Timeline of our PhaKIR sub-challenge highlighting key milestones and dates. All dates refer to the year 2024.

which were defined and documented at the beginning of the process (see Appendix B). A visualization of the annotations for each task, as well as for the three medical centers that provided the data, is shown in Fig. 1. To ensure high annotation quality and to avoid potential sources of error such as inaccurate timestamps or incorrectly classified instruments, a three-stage process was employed for each video. This process involved the initial creation of the annotations, followed by two independent review stages conducted by different annotators. The following sections describe the ground truth generation process in detail for each task.

4.3.1. Task 1: surgical phase recognition

The ground truth data for the surgical phase recognition task was generated by setting timestamps during the visual inspection of the video recordings. Each frame between a defined start and end timestamp was labeled with the corresponding surgical phase, based on the actions performed in that segment. A total of seven distinct surgical phases was used, following the classification scheme introduced in the Cholec80 dataset by Twinanda et al. (2017). Additionally, an eighth category – an undefined phase – was introduced to label transitional frames between two phases. Although the video recordings from Wagner et al. (2023), captured at the Heidelberg University, already included phase annotations, we re-annotated these sequences to ensure consistency with our annotation guidelines and alignment with the data from the other medical centers. Examples of the surgical phase annotations are presented in the second row of Fig. 1.

4.3.2. Task 2: instrument instance segmentation

For the creation of ground truth data for the instrument instance segmentation task, the Computer Vision Annotation Tool (CVAT)[10] was used. The surgical instruments were defined by manually drawn contours, where each contour defined the area of an instrument, from which segmentation masks were subsequently generated. Following the approach by Roß et al. (2021) in the ROBUST-MIS 2019 challenge, this method ensured that each pixel in an image was assigned to exactly one instrument or to the background, enabling pixel-precise multi-instance segmentation. Annotators had access to entire video sequences – including past and future frames – for both the training and the test data, allowing for more accurate segmentation. However, only visible portions of the instruments were annotated. No assumptions were made regarding occluded or inferred parts. A total of 19 different instrument classes was annotated. Within each frame, individual instruments were distinguished by color-encoded masks, where the red (R) and green (G) channels encoded the instrument class, and the blue (B) channel indicated the instance of the object within that class. For example, two instruments of the same class in a single frame shared identical R and G values, while differing in B values. To ensure consistency in manual segmentation, a set of annotation rules and visual examples was provided and strictly followed by all annotators. A visualization of the color-coded instance segmentation masks for the three contributing medical centers is presented in the third row of Fig. 1.

---

[10] https://www.cvat.ai/

4.3.3. Task 3: instrument keypoint estimation

For the instrument keypoint estimation task, annotations were also created using the CVAT tool, which allowed annotators to specify the coordinates of the relevant keypoints. Depending on the type of the surgical instrument, between two and four keypoints were annotated per tool to cover the relevant locations. The annotated keypoints included:

- the endpoint, marking where the instrument intersects the image boundary,
- the shaft-tip transition, indicating the junction between the shaft and the tip, and
- the instrument tip, which may consist of two separate keypoints for opening instruments.

For each keypoint, its visibility status was documented – categorized as visible, hidden or outside the image (and therefore unannotated). As in the instrument instance segmentation task, annotators had access to entire video sequences for both training and test data, allowing them to consult preceding and following frames to achieve accurate localization. In contrast to segmentation, the annotation of keypoints did involve inference based on temporal continuity – i.e., estimated instrument motion was considered to refine keypoint placement, particularly when visibility was limited. A visualization of the annotated keypoints and joints between the instrument parts is depicted in the last row of Fig. 1, following the same color-scheme used for the instrument instance segmentation masks. Both visible and invisible keypoints displayed, with hidden keypoints indicated by a white circle around their respective coordinates.

After publication of this challenge paper and the corresponding dataset paper (Rueckert et al., 2025b), the training dataset will be publicly available (Rueckert et al., 2025a).

5. Assessment method

The following outlines the procedure for evaluation of the performance of the participants' methods. First, the aggregation of metric values is explained in Section 5.1, then the ranking strategy is described in Section 5.2, and finally, the selected metrics are presented in detail in Section 5.3.

5.1. Metric aggregation

For the selection of evaluation metrics, we considered the work of Reinke et al. (2024) and followed the recommendations of the Metrics Reloaded Framework proposed by Maier-Hein et al. (2024). Metrics were calculated per surgical phase and instrument class for each annotated frame in a video. As illustrated in Fig. 4, the distribution of instrument classes is highly imbalanced, with some instruments appearing significantly more frequently than others. To ensure a fair evaluation that does not favor common classes or phases, we first averaged the metric values across all frames of a video for each phase and instrument class individually. These class-specific values were then averaged with equal weighting to compute a single overall score per video. To aggregate across the test set, video-level scores were again averaged with
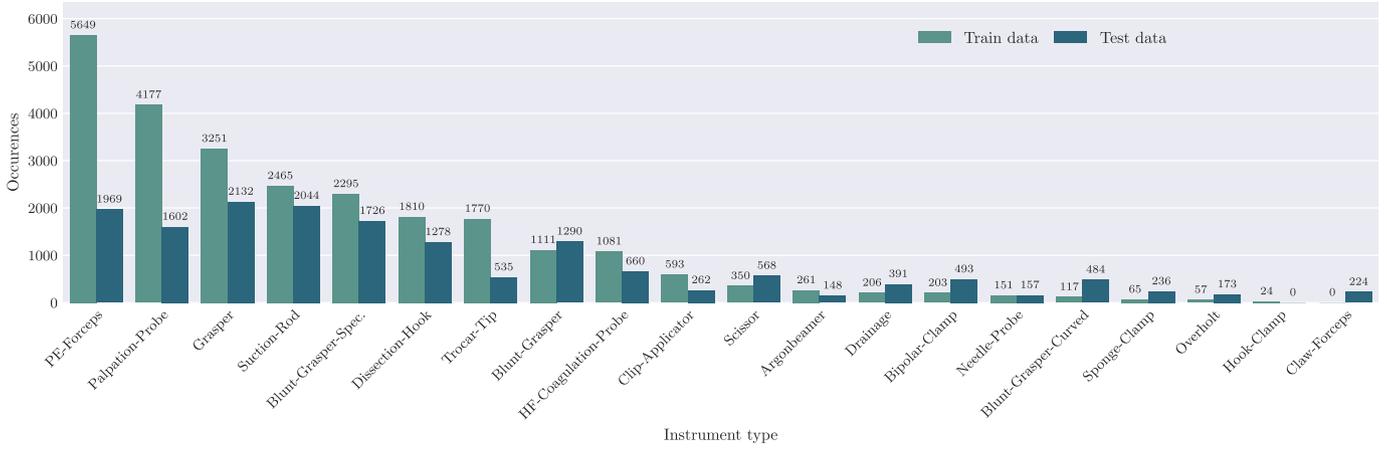
Fig. 4. Number of occurrences for each instrument type, specified for training and test data and sorted in descending order according to the number of occurrences in the training data.

equal weighting, ensuring that all videos contributed equally to the final evaluation – regardless of their lengths.

To improve robustness and stability, we applied bootstrapping (Efron, 1992) with 10,000 iterations. For each run, a number of samples equal to the number of annotated frames in the test set was drawn randomly with replacement. The resulting metric values were averaged across all iterations to produce a final score per metric and participant, forming the ranking basis.

5.2. Ranking strategy

Since the number and type of evaluation metrics varied across tasks ($n = 2$ for surgical phase recognition, $n = 3$ for instrument instance segmentation, $n = 1$ for instrument keypoint estimation), direct aggregation was not feasible due to differences in scale and units. To address this, we calculated a ranking score for each individual metric. These scores were then averaged with equal weighting to obtain an overall ranking value per participating group. The final challenge ranking was derived by comparing these aggregated ranking values across participating groups.

5.3. Metric description

In the following, the employed metrics are described in detail for the tasks surgical phase recognition (Section 5.3.1), instrument instance segmentation (Section 5.3.2), and instrument keypoint estimation (Section 5.3.3).

5.3.1. Task 1: surgical phase recognition

For the task of surgical phase recognition, the F1-score (Van Rijsbergen, 1979; Chinchor, 1991) was used as the per-class performance metric, while Balanced Accuracy (BA, Tharwat, 2021) served as the metric for evaluating the overall multi-class performance. The F1-score, defined as:

$$\text{F1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{1}$$

represents the harmonic mean of precision and recall, making it particularly suitable for the surgical phase recognition task as it accounts for both false positives and false negatives. Similarly, the BA, defined as:

$$\text{BA} = \frac{\text{sensitivity} + \text{specificity}}{2}, \tag{2}$$

corresponds to the unweighted average of sensitivity and specificity, ensuring that both shorter and longer phases contribute equally to the metric's value. It is important to note that for frames labeled as undefined in

the ground truth – typically corresponding to phase transitions – participants were free to handle them as they deemed appropriate. However, these frames were excluded from the evaluation and were not considered in the calculation of the performance metrics for the test data.

5.3.2. Task 2: instrument instance segmentation

For the instrument instance segmentation task, three metrics were employed. For localization, the Dice Similarity Coefficient (DSC, Dice, 1945) was used as the multi-instance multi-class overlap metric, the mean Average Precision (mAP$_{\text{IoU}}$, Lin et al., 2014; Everingham et al., 2015) was applied as the multi-threshold metric, and the 95% Hausdorff-Distance (95% HD, Huttenlocher et al., 1993) served as the boundary-based metric. The DSC measures the overlap between two binary segmentation masks A and B, and is computed as twice the area of overlap between A and B divided by the sum of the areas of the two masks. Formally, for two sets A and B, the DSC is given by:

$$\text{DSC}(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|}, \tag{3}$$

where $|A|$ and $|B|$ denote the cardinalities of A and B, respectively. The DSC ranges from 0 (no overlap) to 1 (perfect overlap), and is commonly used as a measure of segmentation accuracy in medical imaging research and challenge design (Taha and Hanbury, 2015; Müller et al., 2022; Roß et al., 2021).

The mean Average Precision (mAP) is computed following the COCO evaluation protocol (Lin et al., 2014) and the procedure described by Roß et al. (2021). It is defined as the mean of the Average Precision (AP) values across all C object classes and all Intersection-over-Union (IoU) thresholds $T = \{0.50, 0.55, \dots, 0.95\}$:

$$\text{mAP}_{\text{IoU}} = \frac{1}{|C|} \sum_{c=1}^{C} \left(\frac{1}{10} \sum_{t \in [0.5:0.05:0.95]} \left(\int_0^1 p_{c,t}(r) \, dr \right) \right). \tag{4}$$

For each class $c \in C$ and IoU threshold $t \in T$, the Average Precision $\text{AP}_{c,t}$ is defined by the area under the precision-recall curve:

$$\text{AP}_{c,t} = \int_0^1 p_{c,t}(r) \, dr. \tag{5}$$

Here, $p_{c,t}(r)$ denotes the precision at a given recall level r. The final mAP score is obtained by averaging $\text{AP}_{c,t}$ across all classes and IoU thresholds. Since participants' algorithms produced only binary segmentation masks without class probabilities, instance matching was based on the IoU between predicted instances M_1 and ground truth instances M_2, as defined by:

$$\text{IoU}(M_1, M_2) = \frac{|M_1 \cap M_2|}{|M_1 \cup M_2|}, \tag{6}$$

which was originally proposed by Jaccard (1912). The 95% Hausdorff Distance (HD) quantifies the boundary agreement between two point sets A and B, typically representing the contours of predicted and ground truth segmentations. It is defined as:

$$95\% \text{ HD}(A, B) = \max \left\{ \text{quant}_{95} \left(\min_{b \in B} d(a, b) \right), \right.$$
$$\left. \text{quant}_{95} \left(\min_{a \in A} d(b, a) \right) \right\}. \tag{7}$$

This metric computes the maximum of the 95th percentile of the shortest distances from each point in one set to its closest point in the other. The point-to-point distance $d(a, b)$ is measured using the Euclidean norm: $d(a, b) = \|a - b\|_2$. By using the 95th percentile instead of the absolute maximum, this variant of the Hausdorff Distance is less sensitive to outliers and offers a more robust measure of segmentation accuracy. To assign predicted segmentations to ground truth instances, we applied the Hungarian Maximum Matching Algorithm (Kuhn, 2010) based on the IoU between all possible prediction-ground truth pairs. Note that one instrument class present in the test data did not appear in the training set. This class was therefore excluded from the evaluation to ensure fairness and metric consistency.

5.3.3. Task 3: instrument keypoint estimation

The instrument keypoint estimation task was evaluated using the mAP_{OKS} metric based on the Object Keypoint Similarity (OKS) following the COCO evaluation protocol (Consortium, 2025). This approach is conceptually similar to the one used for instance segmentation, with the key difference being the substitution of the IoU with the OKS as the matching criterion between ground truth and predicted instances. The matching strategy follows the procedure described in Section 5.3.2. The OKS is defined as:

$$\text{OKS} = \frac{\sum_i [\exp(-d_i^2 / 2s^2 k_i^2) \cdot \delta(v_i > 0)]}{\sum_i [\delta(v_i > 0)]}, \tag{8}$$

where d_i represents the Euclidean distance between the i-th predicted and ground truth keypoint, v_i denotes the visibility flag of the ground truth keypoints, s is the object scale, and k_i is a constant controlling the falloff of the Gaussian distribution for each keypoint. The function $\delta(v_i > 0)$ is a conditional operator that checks whether the visibility flag of the ground truth keypoint is greater than zero. The calculation of the OKS involves passing d_i through an unnormalized Gaussian distribution with standard deviation sk_i, with only visible keypoints being considered; hidden keypoints, where $v_i = 0$, are excluded from the calculation. In cases where predicted and ground truth instances contain different numbers of keypoints:

- If the predicted instance has more keypoints, only those corresponding to the ground truth class are considered.
- If fewer keypoints are predicted than present in the ground truth, placeholder coordinates $(0, 0)$ are appended for the missing ones.

For instruments with two tip keypoints, the optimal one-to-one correspondence is determined using the Hungarian Algorithm. If the prediction includes more tip keypoints than expected, only the first predicted tip keypoint is considered for evaluation.

6. Participating teams and methods

This section provides a detailed overview of the participating teams and their methodological contributions across the three challenge tasks. Table 2 summarizes the task-wise participation. A total of 14 unique teams took part in the challenge. Of these, seven teams submitted solutions for the surgical phase recognition task, nine for instrument instance segmentation, and two for instrument keypoint estimation. Notably, five teams participated in more than one task, demonstrating interdisciplinary interest and expertise. Within each task, the teams are

listed in ascending alphabetical order. The descriptions of their approaches are based on the methodological reports submitted by the participants and reflect the diversity of strategies employed to tackle the challenge problems.

6.1. Task 1: Surgical phase recognition

This section provides a detailed overview of the methods submitted for the surgical phase recognition task. Table 3 complements this discussion by summarizing key methodological components for each approach. Specifically, it outlines the model architectures, strategies for temporal modeling, and whether external data sources were utilized. Additionally, the table includes relevant training configurations such as loss functions, optimizers, learning rate schedules, initial learning rates, weight decay parameters, number of training epochs, and batch sizes.

6.1.1. *Augi* - *Kyung Hee University, South Korea*
Architecture The Augi team employed the LoViT architecture (Liu et al., 2025) as a feature extractor to capture rich spatial representations.
Temporal context To model temporal dependencies, they used AS-Mamba (Chen et al., 2024), a modified version of ASFormer (Yi et al., 2021) originally designed for temporal action segmentation. ASMamba replaces the attention mechanism in ASFormer with the state-space model (SSM) operation from Vision Mamba (Zhu et al., 2024), improving efficiency of temporal interaction learning. While ASMamba typically leverages bidirectional Vision Mamba blocks to predict the phase of the current frame, the authors modified the model to operate in a unidirectional, causal manner to ensure causal inference. Additionally, they extended the loss function with a penalty term to discourage implausible phase transitions and encourage sequences that align with typical surgical workflows.
Dataset split An eight-fold cross-validation was conducted, selecting the median-performing video for validation.
Augmentations No data augmentation was applied. Input frames were resized to 224×224 and normalized using ImageNet statistics (Russakovsky et al., 2015) before feature extraction.

6.1.2. *Hanglok* - *Hanglok Tech, China*
Architecture The Hanglok team based their approach on the TranSVnet architecture (Gao et al., 2021), which follows a two-stage design. In the first stage, spatial embeddings are extracted from each video frame using a ResNet-50 backbone (He et al., 2016).
Temporal context In the second stage, these embeddings are passed to a temporal convolutional network (TeCNO, Czempiel et al., 2020), which combines the spatial features of the current frame with temporal features from preceding frames. A subsequent transformer layer produces the final classification, where the spatial embeddings of the current frame serve as queries, and the temporal embeddings from preceding frames serve as keys and values.
Dataset split For evaluation, the authors used a fixed dataset split, allocating six videos for training and one video each for validation and testing.
Augmentations Input frames were resized to 250×250 pixels. During training, data augmentation consisted of random horizontal flipping and rotation within a range of $[-5°, +5°]$. Normalization was applied to both training and validation sets using channel-wise means of $[0.417, 0.261, 0.258]$ and standard deviations of $[0.219, 0.198, 0.193]$ for the RGB channels.

6.1.3. *Jmees_inc.* - *Jmees Inc., Japan*
Architecture The Jmees_inc. team employed the EVA-02 architecture (Fang et al., 2024) using default hyperparameter settings.
Temporal context Temporal information was incorporated via an Exponential Moving Average (EMA) update of the model weights, combined with predefined post-processing rules. These rules stipulated that for frames with an ID greater than 30,000, the "Preparation" phase was

Table 2
The available tasks of the PhaKIR challenge, surgical phase recognition, instrument instance segmentation, and instrument keypoint estimation, along with the participation of individual teams in each task.

Task	alvaro	Augi	goncalo	Hanglok	Jmees_inc.	recogna	kist_harilab	floor9	Ryze	sds-hd	sk	Smartlab_hkust	Uniandes24	Yipingli
Phase Recognition		✓		✓	✓				✓			✓	✓	✓
Instance Segmentation		✓	✓	✓	✓	✓	✓	✓				✓	✓	
Keypoint Estimation	✓									✓				

Table 3
Overview of participating teams and their methodological choices for the surgical phase recognition task. The table summarizes the architectures used, incorporation of temporal information, utilization of external data, loss functions, optimizers, schedulers, Learning Rates (LR), Weight Decays (WD), number of training epochs, and Batch Sizes (BS). Frequently used abbreviations and commonly referenced elements include: Stochastic Gradient Descent (SGD, Kiefer and Wolfowitz, 1952), Adaptive Moment Estimation (Adam, Kingma and Ba, 2015), Exponential Moving Average (EMA, Klinker, 2011), Long Short-Term Memory (LSTM, Hochreiter and Schmidhuber, 1997), Cross-Entropy (CE, Good, 1956), weighted cross-entropy (WCE, Ho and Wookey, 2020), Cosine Annealing (CA, Loshchilov and Hutter, 2017), Reduce LR on Plateau (RLROP, Al-Kababji et al., 2022), and ImageNet (Russakovsky et al., 2015).

Team	Basic architecture	Video data used?	Additional data	Loss functions	Optimizer / Scheduler	LR / WD	Epochs / BS
Augi	LoViT (Liu et al., 2025)	Yes, unidirectional ASMamba (Chen et al., 2024)	LoViT pre-trained on Cholec80 (Nwoye et al., 2023)	WCE, phase transition penalties	Adam / RLROP	1e-3 (LoViT), 5e-4 (ASMamba) / No	10 / 1
Hanglok	TranSVnet (Gao et al., 2021)	Yes, TeCNO (Czempiel et al., 2020)	ResNet-50 backbone pre-trained on Cholec80 (Nwoye et al., 2023)	CE	Adam / RLROP	1e-4 (stage 1), 5e-4 (stage 2) / 5e-4	50 (stage 1), 150 (stage 2) / 100
Jmees_inc.	EVA-02 (Fang et al., 2024)	Yes, EMA + post-processing rules	Pre-trained on ImageNet	CE	SGD / CA	1e-5 / 2e-5	5 / 32
Ryze	ResNet-50 (He et al., 2016)	Yes, LSTM layer	Pre-trained on ImageNet	CE	N/A / RLROP	5e-4 / 5e-4	25 / 150 (train), 320 (validation)
Smartlab_hkust	Surgformer (Yang et al., 2024)	Yes, HTA (Yang et al., 2024)	Weights of TimeSFormer (Bertasius et al., 2021) pre-trained on Kinetics-400 dataset (Kay et al., 2017)	CE	AdamW / CA	5e-4 / 5e-2	50 / 32
Uniandes24	MuST (Pérez et al., 2024)	Yes, MTFE + TCM (Pérez et al., 2024)	MViT (Fan et al., 2021) backbone pre-trained on Kinetics-400 (Kay et al., 2017) and Cholec80 (Twinanda et al., 2017)	CE	AdamW / CA	1e-4 (MViT, MuST), 1e-5 (TCM) / 1e-4	20 (MViT pre-training, TCM), 5 (MuST finetuning) 24 (MViT pre-training, MuST finetuning), 256 (TCM)
Yipingli	CAFormer (Yu et al., 2023)	Yes, LSTM layer	Yes, pre-trained weights from SurgeNet (Jaspers et al., 2024)	CE	AdamW / Not used	1e-4 / 1e-2	200 / 1 sequence of 128 frames

disallowed, while for frames with an ID below 2,000, this phase was enforced. For prediction, if the model's confidence score exceeded 0.9, the class with the highest adjusted probability was selected as the final output. Otherwise, a weighted sum of class probabilities over the previous 1000 frames was calculated, with linearly increasing weights ranging from 1 to 1,000.

Dataset split To ensure label diversity, two videos from different hospitals were selected for validation. For the final submission, the model was trained on the full dataset.

Augmentations Input images were resized to 448×448 pixels. Data augmentation during training included random horizontal flipping, random resizing with a scale factor sampled from the range [0.08, 1.0], random aspect ratio adjustments in the range [0.75, 1.33], and color jittering with brightness, contrast, saturation, and hue variations each sampled from the range [−0.4, +0.4].

6.1.4. *Ryze - University of Hong Kong, China*

Architecture The Ryze team employed a ResNet-50 architecture (He et al., 2016) for feature extraction, followed by a fully connected layer for surgical phase classification.

Temporal context To model temporal dependencies, a Long Short-Term Memory (LSTM, Hochreiter and Schmidhuber, 1997) layer was added

after the feature extractor, configured with an input size of 2048, a hidden size of 512, and a sequence length of ten frames.

Dataset split For dataset partitioning, a five-fold cross-validation strategy was used.

Augmentations During training, input images were resized to 250×250 pixels; for validation, they were resized to 224×224 pixels. Data augmentation during training included random cropping to 224×224 pixels, random horizontal flipping, random rotation in the range [−5°, +5°], and color jittering. The color jittering parameters allowed for brightness, contrast, and saturation variations within [−0.1, +0.1] and hue adjustments within [−0.05, +0.05]. Normalization was performed using dataset-specific mean and standard deviation values.

6.1.5. *Smartlab_hkust - The Hong Kong University of Science and Technology, China*

Architecture The Smartlab_hkust team employed the Surgformer architecture (Yang et al., 2024) to process untrimmed surgical videos by constructing frame volumes from temporally sampled sequences.

Temporal context Frames were extracted at predefined intervals and divided into non-overlapping patches, which were then transformed into spatial-temporal tokens. These tokens were passed through a series of stacked transformer blocks, each incorporating Hierarchical Temporal

Attention (HTA), Spatial Attention (SA), and a Multilayer Perceptron (MLP). The SA mechanism aggregates pixel features at identical temporal positions across the spatial dimension, effectively capturing spatial representations within individual frames. In contrast, the HTA mechanism aggregates features at identical spatial locations over the temporal dimension, leveraging multiple temporal windows at varying resolutions to enhance temporal dependency modeling.

Dataset split The dataset was partitioned into a fixed split, allocating six videos for training and two for validation.

Augmentations During training, the RandAugment augmentation strategy (Cubuk et al., 2020) was applied, producing input images with a resolution of 224×224 pixels. The configuration used was rand-m7-n4-mstd0.5-inc1, which indicates that four random transformations (n4) were applied to each image, each with a base magnitude of 7 (m7), a magnitude standard deviation of 0.5 (mstd0.5) for variability, and inclusion of the identity operation (inc1) to probabilistically retain the original image. To further enhance data diversity, random erasing (Zhong et al., 2020) was employed, with a probability of 0.25. Additionally, mixup (Zhang et al., 2018) was incorporated as a regularization strategy.

6.1.6. *Uniandes24 - Los Andes University, Colombia*

Architecture The Uniandes24 team utilized the Multi-Scale Transformers for Surgical Phase Recognition (MuST) architecture (Pérez et al., 2024), which follows a two-stage pipeline.

Temporal context In the first stage, a Multi-Term Frame Encoder (MTFE) is employed to adapt to diverse surgical contexts by capturing dependencies across different temporal scales. This module constructs temporal windows of varying lengths around the target frame to generate embeddings that encode both short- and mid-term temporal information. The second stage consists of a Temporal Consistency Module (TCM), which enhances long-term temporal understanding by modeling relationships among the frame embeddings generated by the MTFE within an extended temporal window. The overall approach builds a temporal multi-scale pyramid, where lower levels provide fine-grained temporal detail and higher offer broader but sparser temporal context. The method employs Multiscale Vision Transformers (MViT, Fan et al., 2021) as the video backbone, leveraging its Multi-Head Pooling Attention mechanism to progressively reduce spatio-temporal resolution while increasing feature dimensionality.

This generates multi-scale spatio-temporal features at each pyramid level, consisting of a class token and patch embeddings. At each level, the features are processed through a Multi-Temporal Attention Module, which combines cross-attention with residual multi-temporal self-attention units, thereby enriching the representation with information from other temporal scales. The class tokens from all levels are then extracted and aggregated to form a comprehensive multi-term embedding of the keyframe, which is further refined through the TCM for the final phase classification. To enhance model predictions, the authors applied a post-processing technique that smooths output probabilities using a sliding spanning two frames. This smoothing considers both the predicted class probabilities and their associated class labels.

Dataset split A fixed dataset split was used, assigning six videos for training and two for validation. During training, all images were resized to a resolution of 250×250 pixels.

Augmentations Data augmentation included random cropping to 224×224 pixels, random horizontal flipping, and AlexNet-style PCA color augmentation using the same eigenvalues and eigenvectors as in MViT. For validation, images were resized to 224×224 pixels. Normalization for both training and validation followed the standard ImageNet values (Russakovsky et al., 2015).

6.1.7. *Yipingli - Eindhoven University of Technology, Netherlands*

Architecture The Yipingli team developed their approach using a backbone composed of a pre-trained CAFormer model (Yu et al., 2023) to extract spatial features from input images.

Temporal context This feature extractor was coupled with a temporal head incorporating an LSTM network (Hochreiter and Schmidhuber, 1997) configured with a sequence length of 128 frames to capture temporal dependencies. The LSTM module propagates hidden states across batches to maintain temporal continuity and enable a more comprehensive understanding of the video content. During fine-tuning, only the final block of the CAFormer backbone was unfrozen, allowing effective end-to-end supervised training while preserving the rich feature representations learned during pre-training.

Dataset split A cross-validation strategy was employed, with the dataset partitioned into six videos for training and two for validation.

Augmentations To enhance model robustness, several data augmentation techniques were applied, including random shifting and scaling with a factor in the range $[-0.05, +0.05]$, random rotation between $[-15°, +15°]$, color shifts applied individually to each RGB channel within the range $[-15, +15]$, and random brightness and contrast adjustments with factors in the range $[-0.2, +0.2]$. All images were resized to 224×224 pixels and normalized using the standard ImageNet values (Russakovsky et al., 2015).

6.2. Task 2: instrument instance segmentation

This section provides a detailed overview of the methods submitted by the participating teams for the instrument instance segmentation task. Table 4 is structured analogously to Table 3 and summarizes the key methodological components of each approach, including the architectures employed, temporal modeling strategies, data utilization, and training configurations. The table serves as a concise reference for comparing the core characteristics and design decisions of the submitted solutions.

6.2.1. *Augi - Kyung Hee University, South Korea*

Architecture The Augi team employed the Mask2Former architecture (Cheng et al., 2022) in combination with a Presence-aware Instrumental Segmentation (PINS) module, specifically designed to refine segmentation predictions in frames devoid of surgical instruments. In the first stage, feature extraction is performed at multiple scales using a Swin-Large (Swin-L) backbone (Liu et al., 2021), allowing the model to accurately segment both small and large surgical tools within complex surgical scenes. Leveraging the transformer-based architecture of Mask2Former, the model generated multiple object queries per frame, each corresponding to a detected instance of a surgical instrument. To reduce false-positive predictions, the authors introduced the PINS module as a second stage of their approach. This module, build upon a Swin-Base (Swin-B) backbone (Liu et al., 2021), performs binary classification to determine whether any surgical instrument is present in a given frame. If Mask2Former produces a segmentation mask for a frame classified by PINS as instrument-free, the PINS module suppresses the erroneous mask, thereby enhancing the overall segmentation quality. During training, deep supervision (Lee et al., 2015) was employed to guide learning across the network's multiple stages.

Temporal context No temporal information from the video sequences was utilized.

Dataset split A fixed dataset split was used, with seven videos designated for training and one for validation. For Mask2Former training, only frames containing at least one instrument were included, while the PINS module was trained on all available frames.

Augmentations Input images for Mask2Former were resized to 1024×1024 pixels with random scaling applied using a factor in the range $[-0.9, 1.0]$. For the PINS module, augmentation techniques included random horizontal and vertical flipping, random rotation within $[-30°, +30°]$, random resized cropping with scale range $[0.8, 1.0]$, color jittering (brightness, contrast, saturation, and hue variation between $[0.5, 1.5]$), and the addition of Gaussian noise with a standard deviation of $\sigma = 0.05$. All training and validation images were resized to 224×224 pixels.

Table 4

Overview regarding teams and methods participated in the task of surgical instrument instance segmentation. Shown are the used architectures, the usage of the temporal information, the utilization of external data, the employed loss functions, optimizers, schedulers, Learning Rates (LR), Weight Decays (WD), Epochs, and Batch Sizes (BS). Common terms are abbreviated as follows, together with their citations: Stochastic Gradient Descent (SGD, Kiefer and Wolfowitz, 1952), Adaptive Moment Estimation (Adam, Kingma and Ba, 2015), Cross-Entropy (CE, Good, 1956), step decay (SD, Johnson et al., 2023), Cosine Annealing (CA, Loshchilov and Hutter, 2017), Flat Cosine Annealing (FCA, Lyu et al., 2022), Reduce LR on Plateau (RLROP, Al-Kababji et al., 2022), and ImageNet (Russakovsky et al., 2015).

Team	Basic architecture	Video data used?	Additional data	Loss functions	Optimizer / Scheduler	LR / WD	Epochs / BS
Augi	Mask2Former (Cheng et al., 2022), PINS module	No	Swin-L (Liu et al., 2021) backbone pre-trained on ImageNet	Binary CE + dice loss (segmentation), focal loss (classification)	AdamW / SD	1e-4 (Mask2Former, PINS) / 5e-2 (Mask2Former)	4775 (Mask2Former), 20 (PINS) / 8 (Mask2Former), 16 (PINS)
goncalo	Mask R-CNN (He et al., 2017)	No	Microsoft COCO (Lin et al., 2014)	Binary CE (mask head, objectness), CE (classification)	Adam / SD	1e-3 / 1e-2	30 / 8
Hanglok	SAM (Kirillov et al., 2023), DETR (Carion et al., 2020)	No	SAM pre-trained ViT-H weights (Kirillov et al., 2023), DETR R50 weights (Carion et al., 2020), DETR backbone pre-trained on ImageNet	Dice loss + NTXentLoss (He et al., 2020; Chen et al., 2020) (SAM), CE + IoU loss (DETR)	Adam (SAM), AdamW (DETR) / SD (DETR)	1e-3 (SAM), 1e-4 (DETR), 1e-5 (DETR backbone) / 1e-4 (SAM), 1e-4 (DETR, DETR backbone)	300 (SAM), 200 (DETR, DETR backbone) / 32 (SAM), 10 (DETR, DETR backbone)
Jmees_inc.	Mask2Former (Cheng et al., 2022), SAM 2 (Ravi et al., 2024)	No	No external data, pseudo labels generated from unlabeled PhaKIR video frames	CE (classification), dice loss + CE (segmentation)	AdamW / CA	1e-4 / 5e-2	10 / 4
kist_harilab	TP-SIS (Zhou et al., 2023), BERT (Devlin et al., 2019), Swin transformer (Liu et al., 2021)	No	Pre-trained TP-SIS based on CLIP, EV-17, EV-18 (Radford et al., 2021; Allan et al., 2019, 2020), BERT pre-trained on BooksCorpus (Zhu et al., 2015) and English wikipedia, Swin backbone pre-trained on ImageNet	CE	AdamW / SD	3e-5 / 1e-2	50 / 32
floor9	RTMDet-Ins-l (Lyu et al., 2022), SAM 2 (Ravi et al., 2024)	Yes, match instances along frames	RTMDet-Ins-l pre-trained on CholeInstanceSeg (Alabi et al., 2024)	Dice-loss	AdamW / FCA	4e-3 / 5e-2	300 / 16
recogna	U-Net (Ronneberger et al., 2015), SAM (Kirillov et al., 2023)	No	U-Net trained from scratch, SAM pre-trained on SA-1B (Kirillov et al., 2023)	CE	Adam / RLROP	1e-3 / 1e-4	100 / 20
sk	U-Net (Ronneberger et al., 2015)	No	ConvNext-Base (Liu et al., 2022) backbone pre-trained on ImageNet	Dice focal loss, Hausdorff distance loss	AdamW / CA	1e-4 / 1e-5	50 / 24
Uniandes24	MATIS (Ayobi et al., 2023)	Yes, TAPIS (Ayobi et al., 2024), SAM2 (Ravi et al., 2024)	Pre-trained ViT on ImageNet, Mask2Former on Microsoft COCO (Lin et al., 2014), MViT on Kinetics-400 (Kay et al., 2017), SAM2 on SA-V (Ravi et al., 2024)	CE	AdamW (ViT, Mask2Former), Adam (MViT, Transformer) / SD (Mask2Former), CA (MViT, Transformer)	1e-5 (ViT, MViT, Transformer), 1e-4 (Mask2Former) / 1e-4 (MViT, Transformer)	30 / 24 (Mask2Former), 64 (MViT, Transformer)

6.2.2. goncalo - *University of Minho, Portugal*

Architecture The goncalo team developed their method based on the Mask R-CNN architecture (He et al., 2017), employing a ResNet-50 backbone (He et al., 2016). To leverage pre-trained weights effectively, the first three layers of the backbone were frozen during training, while only the subsequent layers were fine-tuned.

Temporal context Temporal information of the video sequences was not incorporated into the approach.

Dataset split A fixed dataset split was used, with six videos allocated for training and two for validation.

Augmentations No data augmentation techniques were applied during either training or validation phases. All input images were resized to a uniform resolution of 512×512 pixels.

6.2.3. Hanglok - *Hanglok Tech, China*

Architecture The Hanglok team employed the Segment Anything Model (SAM, Kirillov et al., 2023), utilizing the ViT-H variant of the Vision Transformer (ViT, Dosovitskiy et al., 2021) as the backbone. The segmentation prompt of their model was derived from the outputs of the DEtection TRansformer (DETR, Carion et al. (2020)), which itself was based on a Swin-Transformer (Liu et al., 2021) backbone. The SAM model comprises three core components: an image encoder, a prototype-based class prompt encoder, and a mask decoder. The image encoder extracts embeddings from the input image, which are subsequently processed by the prompt encoder. Following the methodology of Surgical-SAM (Yue et al., 2024), the authors replaced the original SAM prompt encoder with a class prompt encoder to enhance performance. The class

prompt encoder calculates the spatial similarity between the image embedding and a class embedding bank, activating class-specific regions within the image. The class predicted by the DETR model is then used to generate prompt embeddings for the mask decoder. To improve class discrimination and facilitate more accurate class prompting, contrastive learning was applied to the class prompt encoder.

Temporal context Temporal information from the video sequences was not utilized.

Dataset split For dataset partitioning, the authors adopted a class-based splitting strategy rather than a video-based one, assigning individual frames to training and validation sets in a 9:1 ratio.

Augmentations During training, data augmentation techniques included horizontal and vertical flipping, random rotation in the range [−30°, +30°], and random batch sampling.

6.2.4. *Jmees_inc.* - *Jmees Inc., Japan*

Architecture The Jmees_inc. team utilized the Mask2Former architecture (Cheng et al., 2022) with a Swin-Transformer-Base (Liu et al., 2021) backbone for the instance segmentation task. Pseudo labels were generated using the Segment Anything Model 2 (SAM 2, Ravi et al., 2024) and subsequently employed as supervision targets for training the Mask2Former model. To convert instance segmentation outputs into semantic segmentation masks, a post-processing step was applied, in which each predicted instance was binarized using a threshold of 0.5.

Temporal context The authors experimented with incorporation of temporal information by considering the frequency of instances across multiple frames for determining the instrument class. However, this approach was found to degrade validation performance and was therefore excluded from the final method.

Dataset split A fixed dataset split was used, allocating six videos for training and the remaining two for validation, with careful attention to ensuring all instrument classes were present in the validation set. For the final submission, the model was trained on the entire dataset.

Augmentations During training, several data augmentation techniques were applied, including random horizontal flipping, large-scale jittering (Ghiasi et al., 2021) with a resize factor randomly sampled from the range [0.1, 2.0], target resolution of 750 × 1333 pixels, and random cropping to 750 × 1333, allowing for negative crops. For validation, no data augmentation was performed beyond resizing to the fixed resolution of 750 × 1333 pixels.

6.2.5. *kist_harilab* - *Korea Institute of Science and Technology, South Korea*

Architecture The kist_harilab team employed a text-promptable surgical instrument segmentation model (TP-SIS, Zhou et al., 2023), which leverages text embeddings generated by Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019). For image encoding, a Swin Transformer (Liu et al., 2021) was used as the backbone. The model operates based on text prompts constructed from instrument class names and additional prompts derived from object location information, specifically the centroid of each object. To mitigate potential bias from only positive prompts (e.g., prompts for objects present in the image), the authors also generated negative prompts for instruments not present in the frame. This strategy enhanced the model's ability to distinguish object existence more reliably. To ensure training stability, the number of positive and negative prompts was balanced.

Temporal context The temporal context from the video sequences was not utilized.

Dataset split Additionally, the team did not define a separate validation dataset, opting to use the entire available dataset for model training.

Augmentations All input images were resized to a resolution of 480 × 480 pixels, and the network outputs were upsampled to match the original image resolutions for the final segmentation mask generation. No data augmentation techniques were applied during training.

6.2.6. *floor9* - *King's College London, UK*

Architecture The floor9 team utilized a model from the Real-Time Models for Object Detection (RTMDet) framework, specifically the RTMDet-Ins-l architecture (Lyu et al., 2022), which was initialized with weights pre-trained on the CholecInstanceSeg dataset (Alabi et al., 2024) – a large-scale tool instance segmentation benchmark specifically curated for cholecystectomy procedures. This model was chosen for its optimal balance between high performance and efficient training time compared to other architectures in its category. As part of the post-processing pipeline, the authors applied non-maximum suppression across classes to ensure that, for instruments with multiple predicted instances across different classes, only the instance with the highest confidence score was retained.

Temporal context In a subsequent step, temporal information was incorporated by comparing the predicted class labels of each instance in the keyframe with those of neighbouring frames. To facilitate instance matching across frames, the SAM 2 model (Ravi et al., 2024) was employed to propagate labels from adjacent predictions. The instrument class in the keyframe was updated only when sufficient supporting evidence from surrounding frames was present.

Dataset split The authors did not report any specific details regarding a training-validation dataset split.

Augmentations For training, the default augmentation strategies from the RTMDet framework were employed, including Cached Mosaic, Cached MixUp, and large scale jittering (Lyu et al., 2022; Bochkovskiy et al., 2020; Zhang et al., 2018; Ghiasi et al., 2021). However, the parameter ranges used for these transformations were not specified.

6.2.7. *recogna* - *São Paulo State University, Brazil*

Architecture The recogna team developed their method based on a two-stage coarse-to-fine procedure. In the first stage, a U-Net architecture (Ronneberger et al., 2015) was employed for object prediction and initial coarse segmentation. The second stage involved segmentation refinement using the Segment Anything Model (SAM, Kirillov et al., 2023), where bounding boxes extracted from the U-Net outputs were used as prompt inputs, enabling more precise segmentations during the refinement phase.

Temporal context Temporal information from the video sequences was not utilized.

Dataset split A fixed data split was applied, with 80% of the images assigned to training and 20% to validation. The images for both training and validation were randomly selected.

Augmentations During training, the authors employed several data augmentation techniques, including horizontal and vertical flipping, random rotation within the range of [−30°, +30°], and random color jittering with individual adjustment factors for hue, saturation, brightness, and contrast, all sampled from the interval [−0.1, +0.1]. All training and validation images were resized to a resolution of 256 × 256 pixels.

6.2.8. *sk* - *Muroran Institute of Technology, Japan*

Architecture The sk team utilized a U-Net architecture (Ronneberger et al., 2015) with a ConvNext-Base (Liu et al., 2022) backbone for feature extraction. The encoder was initialized with pre-trained weights from the ImageNet dataset (Russakovsky et al., 2015), while the decoder weights were initialized randomly.

Temporal context The authors did not incorporate any temporal information into their method.

Dataset split A fixed data split for training and validation was applied. Specifically, three complete video sequences, along with the first halves of two additional videos, were used for training, while two complete videos and the second halves of the aforementioned videos were assigned to the validation set.

Augmentations During training, random horizontal and vertical flipping were applied as data augmentation techniques, and all input images were resized to a resolution of 960 × 544.

Table 5

Overview of participating teams and their methods for the surgical instrument keypoint estimation task. The table summarizes the architectures employed, the use of temporal information, the utilization of external data, as well as key training parameters including loss functions, optimizers, schedulers, Learning Rates (LR), Weight Decays (WD), epochs, and Batch Sizes (BS). Common abbreviations and frequently used terms together with their associated references are as follows: Stochastic Gradient Descent (SGD, Kiefer and Wolfowitz, 1952), Cross-Entropy (CE, Good, 1956), Cosine Annealing (CA, Loshchilov and Hutter, 2017), One Cycling LR (OCLR, Smith and Topin, 2019).

Team	Basic architecture	Video data used?	Addi-tional data	Loss functions	Optimizer / Scheduler	LR / WD	Epochs / BS
alvaro	YOLOv8 (yolo-v8-x variant) (Varghese and Sambath, 2024)	No	No	CE (classification), complete IoU loss (bounding box regression), distributed focal loss (bounding box identification), OKS loss (keypoint localization)	SGD / OCLR with CA	1e-2 / 5e-4	300 / 32
sds-hd	YOLOv8 (yolo-v8-x-pose-p6 variant) (Varghese and Sambath, 2024)	No	No	CE (classification, keypoint locations and visibility), complete IoU loss (bounding box regression), distributed focal loss (bounding box identification)	SGD / No	1e-2 / 0.0	500 / 104

6.2.9. *Uniandes24* - Los Andes University, Colombia

Architecture The Uniandes24 team proposed a fully supervised approach based on a two-stage mask classification pipeline using Masked-Attention Transformers for Surgical Instrument Segmentation (MATIS, Ayobi et al., 2023), extended to enhance temporal consistency in mask predictions. The core method relies on an adapted version of Mask2Former (Cheng et al., 2022), which predicts region candidates for each surgical instrument within a frame and classifies them into specific instrument types.

Temporal context To improve classification consistency across frames, the team integrated a Temporal Consistency Module (TCM) inspired by the Transformers for Actions, Phases, Steps, and Instrument Segmentation (TAPIS) model (Ayobi et al., 2024), employing the MViT (Fan et al., 2021) backbone for video analysis. This module aggregates spatio-temporal features over time to provide essential contextual information for classification.

To maintain temporal coherence in instance predictions, SAM2 (Ravi et al., 2024) was employed to track individual instrument instances across frames, generating a temporal mask sequence for each instance. These tracked instances were then associated with semantic features from Mask2Former's predictions using the IoU metric. Each semantic feature window was processed independently by a transformer encoder, capturing local temporal dependencies for each instance.

Further improvements in cross frame mask classification were achieved by incorporating a cross-attention module, again inspired by TAPIS (Ayobi et al., 2024) and leveraging MViT (Fan et al., 2021) to aggregate global spatio-temporal features. These were fused with the local features from the transformer encoder, providing rich contextual information to support accurate classification. To reduce false positives in frames without instruments, a ViT (Lin et al., 2014) model was trained for binary classification to distinguish background from non-background frames. If a frame was classified as background, the mask threshold for all classes was increased to 0.9, thereby reducing the likelihood of erroneous predictions.

Dataset split The dataset was split into six videos for training and two for validation.

Augmentations For Mask2Former, input frames were resized to 750×1333 pixels, and random horizontal flipping was used as data augmentation. For the MViT and transformer pipeline, training images were resized to 250×250, followed by a random crop to 224×224. Additional augmentations included horizontal flipping, color jittering, and AlexNet-style PCA color augmentation (using the same eigenvalues and eigenvectors as in MViT), followed by ImageNet normalization (Russakovsky et al., 2015) values. During validation, images were resized to 224×224 pixels, with only the normalization step applied.

6.3. Task 3: instrument keypoint estimation

The following sections outline the methodological approaches employed by the participants in the instrument keypoint estimation task. Table 5 follows the same structure as the Tables 3 and 4 providing a concise overview of each team's approach. It summarizes the key characteristics of the submitted methods, including the architectures used, the incorporation of temporal information, the utilization of external data, as well as essential training parameters.

6.3.1. *alvaro* - Medtronic, UK, and University College London, UK

Architecture The alvaro team addressed the task of surgical instrument keypoint localization by incorporating uncertainty estimation into a state-of-the-art object detection framework. Specifically, they utilized the YOLOv8 architecture, selecting the yolo-v8-x variant (Varghese and Sambath, 2024) and extended it to jointly predict keypoint localizations, bounding box coordinates, and class labels. Their objective was to leverage recent developments in deterministic uncertainty quantification methods. particularly the Deep Deterministic Uncertainty (DDU) approach (Mukhoti et al., 2021), building upon the foundation established by Deterministic Uncertainty Methods (DUM, Postels et al., 2022). To realize this, the authors introduced spectral normalization layers (Miyato et al., 2018) into each convolutional block of the YOLOv8 model. The rationale behind this modification is that spectral normalization enforces Lipschitz continuity, effectively constraining both the upper and lower Lipschitz bounds. This regularization is intended to prevent the functional collapse and enhance training stability.

Temporal context The authors did not incorporate any temporal information from the video sequences.

Dataset split The dataset was partitioned into 65% for training, 10% for validation, and 25% for testing, ensuring that each instrument class was represented in all subsets.

Augmentations During training, a range of data augmentation techniques were applied. Initially, images were resized to 640×500 while preserving the aspect ratio through padding, followed by horizontal and vertical flipping. Further augmentations included brightness reduction, contrast enhancement, and various blurring transformations – such as motion blur, median blur, and random fog generation. The specific parameter ranges for these transformations were not reported. Following augmentation, images were resized again to the final resolution of 640×384.

6.3.2. *sds-hd* - German Cancer Research Center, Germany

Architecture The sds-hd team utilized the yolov8x-pose-p6 variant of the YOLOv8 architecture (Varghese and Sambath, 2024). Given that

Table 6

Results and rankings for each individual team and metric regarding the surgical phase recognition task, together with the averaged and overall rank for each team. For each metric, the mean value over the 10,000 bootstrap runs (Mean$_{BT}$), the 95% Confidence Interval (95% CI), and the metric-specific rank are shown, and it is indicated whether a higher or lower value is desired.

Team	F1-score (in %) ↑			BA (in %) ↑			Avg. rank	Global rank
	Mean$_{BT}$	95% CI	Rank	Mean$_{BT}$	95% CI	Rank		
Uniandes24	69.12	[68.90, 69.35]	1	84.20	[84.09, 84.31]	1	1.0	**1st**
Jmees_inc.	65.46	[65.23, 65.70]	2	82.20	[82.07, 82.33]	2	2.0	**2nd**
Yipingli	58.01	[57.80, 58.23]	3	79.02	[78.90, 79.15]	4	3.5	**3rd**
Smartlab_hkust	57.34	[57.13, 57.55]	4	80.31	[80.21, 80.42]	3	3.5	**3rd**
Ryze	48.32	[48.09, 48.55]	5	73.78	[73.63, 73.93]	5	5.0	**5th**
Hanglok	28.83	[28.62, 29.03]	6	62.21	[62.06, 62.35]	6	6.0	**6th**
Augi	17.46	[17.34, 17.58]	7	57.97	[57.86, 58.08]	7	7.0	**7th**

the provided annotations featured a variable number of keypoints per object depending on the instrument class, a dedicated padding strategy was implemented to standardize the keypoint count. This ensured comparability with YOLOv8-pose, which requires a consistent number of keypoints across all object classes. After pre-processing, the standardized annotations and corresponding images were fed into the YOLOv8 model. During post-processing, the predictions were filtered to retain only the relevant keypoints for each instrument, while padded keypoints were excluded from the final output.

Temporal context Temporal context of the video sequences was not employed.

Dataset split For model development, a leave-one-out cross-validation strategy was applied across the eight available videos, with each video iteratively excluded for validation. For the final submission, the model was trained on all eight videos to fully leverage the available data.

Augmentations A variety of data augmentation techniques were employed during training. These included color-based augmentations such as random hue shifts within the range [−0.015, +0.015], saturation adjustments in the range [−0.7, +0.7], and value modifications in the range [−0.4, +0.4], aimed at addressing variations in lighting and color conditions. Geometric augmentations included random translation in the range [−0.1, +0.1], random scaling within [−0.5, +0.5], and both vertical and horizontal flipping. In addition, the mosaic augmentation technique (Kaur et al., 2021) was employed with a factor of 1.0. This augmentation was disabled after the 50th epoch to allow for more refined training in the later stages of model convergence.

7. Results

The following sections present the evaluation results, beginning with an overview of the two statistical analysis methods used (Section 7.1). This is followed by the detailed results for the three challenge tasks: surgical phase recognition (Section 7.2), instrument instance segmentation (Section 7.3), and instrument keypoint estimation (Section 7.4). For each task, we first report the overall performance results (Tables 6, 8, 12), followed by a statistical analysis of ranking stability (Figs. 5, 9, 14) and task-specific supplementary analyses. For the phase recognition task, we include a team-specific confusion matrix visualization (Fig. 6). For the segmentation and keypoint estimation tasks, we present qualitative results (Tables 9, 10, 13). In addition, the instrument segmentation task includes a comparison of the methodological approaches used by the participating teams (Fig. 10). Finally, for all three tasks, we analyze team performance at the level of individual test videos, enabling a more fine-grained assessment of robustness and consistency (Figs. 7, 8, 11, 12, 13, 15; Tables 7, 11, 14).

7.1. Statistical analysis

This section introduces two statistical analysis methods used in the subsequent chapters: ranking stability plots (Figs. 5, 9, and 14) and kernel density estimation plots (KDE, Silverman, 2018, Figs. 7, 8, 11, 12, 13, 15).

Ranking stability plots illustrate the robustness of metric-based rankings by repeating the evaluation across multiple bootstrap iterations. In these plots, the size of each color-coded blob at position (A$_i$, rank j) reflects the relative frequency with which algorithm A$_i$ achieved rank j across b = 10,000 bootstrap samples. The median rank of each algorithm is indicated by a black cross, and the corresponding 95% bootstrap interval, ranging from the 2.5th to the 97.5th percentile, is shown as a black line. This visualization technique is adapted from the blob plot concept in Wiesenfarth et al. (2021).

Kernel density estimation with a Gaussian kernel is a non-parametric technique used to estimate the probability density function of a continuous variable from sample data. Each KDE is constructed by centering a Gaussian kernel at each data point x_i, and summing the contributions to form a smooth approximation of the underlying distribution. The KDE is defined as:

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{(x - x_i)^2}{2h^2}\right),\tag{9}$$

where x is the evaluation point and x_i are the observed data points and h is the bandwidth parameter controlling the smoothness of the estimate. High-density regions of the bootstrap distribution appear as sharp peaks in the KDE, while low-density regions produce flatter shapes. To select an appropriate bandwidth h, we apply Scott's Rule (Scott, 2015), which balances the sample size and data dimensionality:

$$h = n^{-\frac{1}{d+4}},\tag{10}$$

where n = 10,000 denotes the number of bootstrap samples and d = 1 represents the dimensionality of the data.

KDE plots are valuable for visualizing bootstrap results because they provide an intuitive and smooth representation of an algorithm's performance across repeated evaluation runs. Rather than focusing solely on single summary statistics such as the mean or median, KDE plots reveal the full distribution of performance. This makes it easier to compare methods, understand variability, and assess consistency in a more comprehensive and interpretable way.

7.2. Task 1: surgical phase recognition

The results for the surgical phase recognition task are summarized in Table 6. The team Uniandes24 achieved the highest scores in both metrics (69.12% F1-score, 84.20% BA), securing first place in the overall ranking and thereby winning the task. The individual rankings across both metrics are largely consistent, with the exception of a swap between third and fourth place: Yipingli ranked higher in terms of
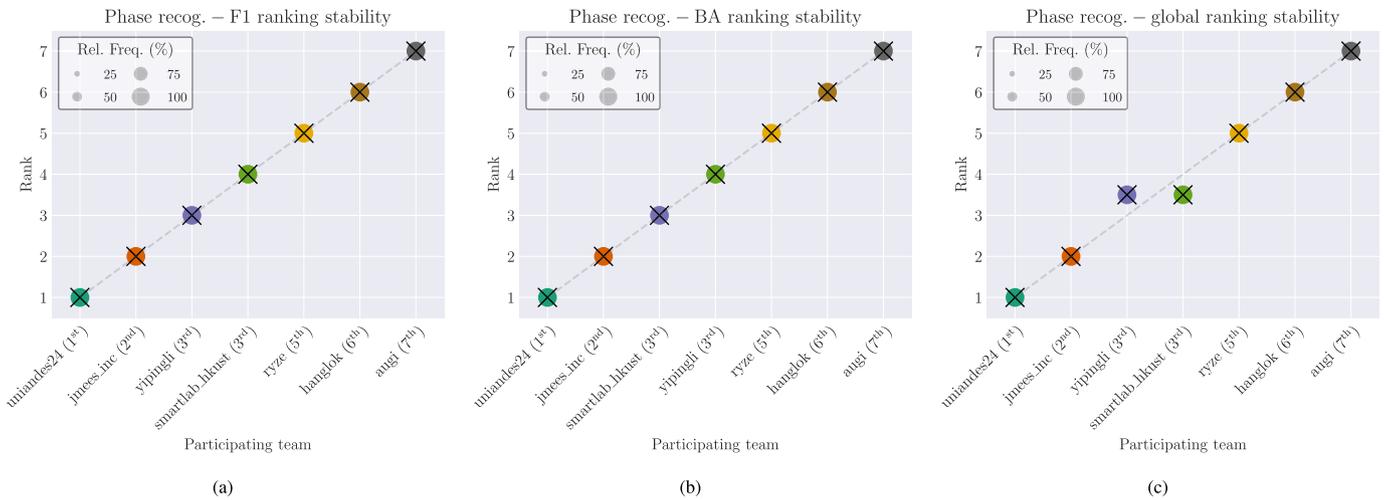
Fig. 5. Visualization of ranking stability based on bootstrapping for the surgical phase recognition task for the metrics F1-score (a), BA (a), and Global (c). For each metric, the individual ranking stability is shown and the global rank is indicated after each team name.

Table 7
Results of the surgical phase recognition task for each team and each individual video in the test dataset, together with the mean metric values for each video.

Team	F1-score (in %) ↑					BA (in %) ↑				
	V_06	V_08	V_09	V_12	V_14	V_06	V_08	V_09	V_12	V_14
Uniandes24	81.30	72.59	71.58	78.49	41.66	90.50	83.84	85.64	89.59	71.43
Jmees_inc.	66.76	64.08	86.24	64.42	45.81	82.19	81.01	94.23	81.19	72.38
Yipingli	68.53	49.91	79.32	48.77	43.54	82.58	75.07	92.07	75.89	69.50
Smartlab_hkust	76.67	46.57	83.43	44.70	35.34	88.34	76.82	91.42	73.91	71.08
Ryze	57.15	43.64	56.33	43.33	41.14	75.90	71.13	80.74	70.77	70.35
Hanglok	34.19	22.20	45.56	20.80	21.39	64.05	61.27	70.11	55.42	60.20
Augi	12.99	10.05	19.46	18.35	26.44	53.85	56.97	54.35	63.30	61.39
mean	**56.80**	**44.15**	**63.13**	**45.55**	**36.47**	**76.77**	**72.30**	**81.22**	**72.87**	**68.05**

F1-score, while Smartlab_hkust outperformed in BA. Since both teams attained the same average rank of 3.5, they shared third place in the final ranking.

The ranking stability, assessed via bootstrapping, is visualized in Fig. 5. Fig. 5(a) and (b) depict the ranking stability for the F1-score and the BA metric, respectively, while Fig. 5(c) presents the stability of the overall ranking based on the average of both metrics. The results demonstrate high robustness across all rankings and confirm the stability of Yipingli's third place in F1-score and Smartlab_hkust's third place in BA.

Fig. 6 presents the confusion matrices for all participating teams, illustrating the relationship between the actual and predicted surgical phases and highlighting the proportion of correctly and incorrectly classified categories. Among all phases, the Preparation Phase (Pr) achieved the highest average classification accuracy, with 74.05% across all participants. In contrast, the Clipping and Cutting phase (ClCu) proved to be the most challenging, with the lowest average accuracy of 29.89%.

The average classification accuracies for the remaining phases were as follows:

- Gallbladder Retraction (GR): 71.00%
- Gallbladder Packaging (GP): 57.89%
- Gallbladder Dissection (GD): 53.28%
- Cleaning and Coagulation (ClCo): 45.85%
- Calot Triangle Dissection (CTD): 42.71%.

A closer inspection of misclassification patterns reveals consistent trends:

- The Pr phase was most frequently confused with GR (12.35%) and CTD (10.07%).
- The CTD phase was most often misclassified as GD (20.05%) and GP (12.43%).
- The ClCu phase was commonly mistaken for CTD (27.46%) and GD (20.92%).
- The GD phase was misclassified as ClCo (17.72%) and GP (13.65%).
- The GP phase was most frequently confused with GR (18.22%) and ClCo (9.57%).
- The ClCo phase was often mistaken for GD (19.52%) and GP (12.77%).
- The GR phase was most commonly misclassified as GP (13.44%) and Pr (7.81%).

Figs. 7 and 8 display the KDE plots for the F1-score and BA, respectively, derived from bootstrap evaluation across individual test videos. In the KDE plot for the F1-score (Fig. 7), video nine yielded the highest metric values for four out of seven teams, indicating it was the most accurately predicted video overall. In contrast, video 14 produced the lowest F1-scores for five out of seven teams, making it the most challenging sequence in the dataset. A similar pattern emerges in the KDE plot for BA (Fig. 8): video nine achieved the highest scores for five teams, while video 14 registered the lowest values for five teams.

Table 7 presents quantitative results summarizing the mean metric values for each individual video in the test dataset. These numerical results align well with the KDE visualizations shown in Figs. 7 and 8, reinforcing the observed performance trends. Among all test sequences, video nine was the most successfully recognized across teams, achieving an average F1-score of 63.13% and BA of 81.22%. In contrast, video 14
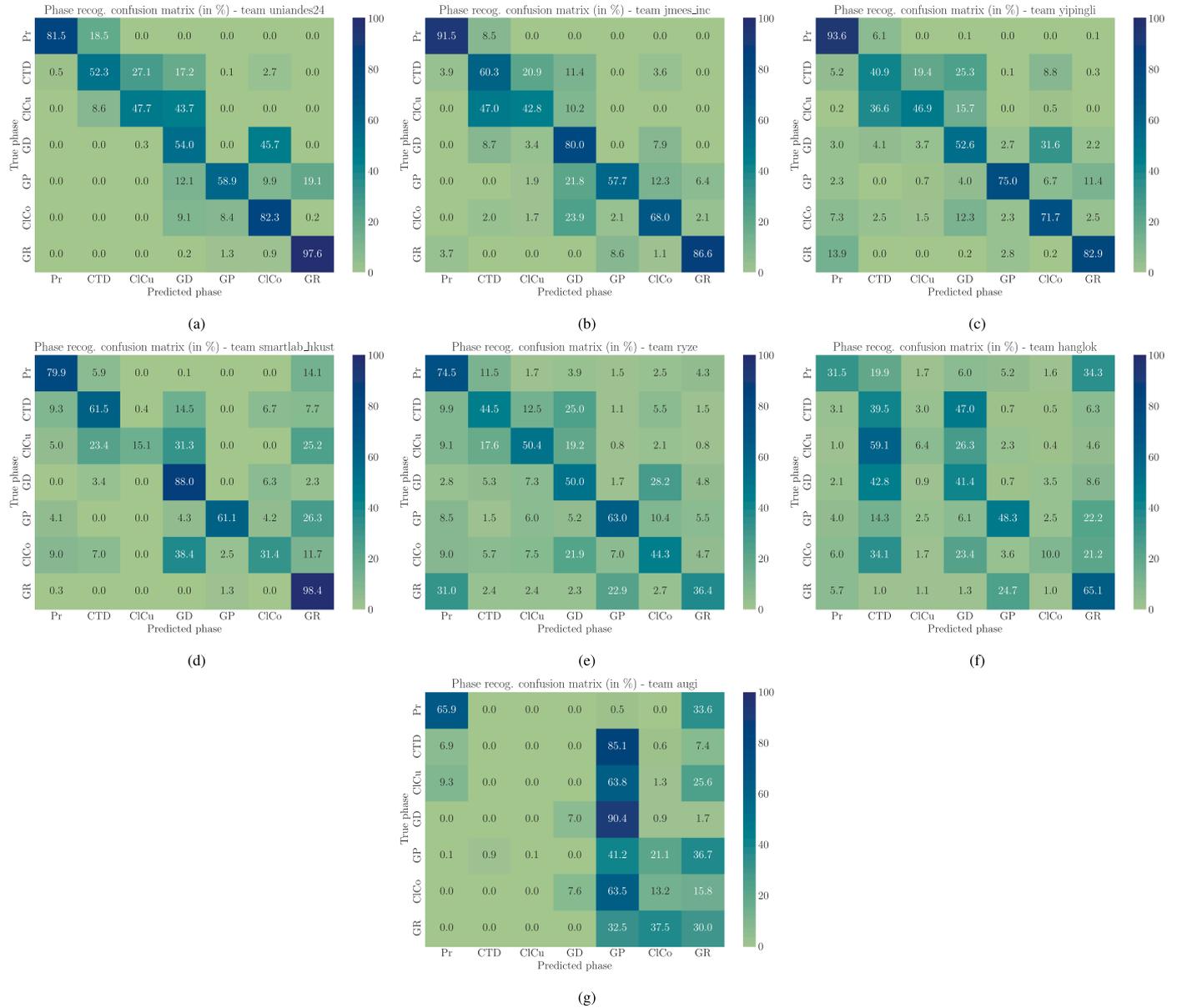
Fig. 6. Confusion matrices for the surgical phase recognition task across all test videos, given in row-wise relative values, for teams Uniandes24 (a) Jmees_inc. (b), Yipingli (c), Smartlab_hkust (d), Ryze (e), Hanglok (f), Augi (g), sorted according to the results and rankings shown in Table 6. For clearer visualization, the phase names are abbreviated as follows: Preparation (Pr), Calot Triangle Dissection (CTD), Clipping and Cutting (ClCu), Gallbladder Dissection (GD), Gallbladder Packaging (GP), Cleaning and Coagulation (ClCo), and Gallbladder Retraction (GR).

Table 8
Results and rankings for each individual team and metric regarding the instrument instance segmentation task, together with the averaged and overall rank for each team. For each metric, the mean value over the 10,000 bootstrap runs ($Mean_{BT}$), the 95% Confidence Interval (95% CI), and the metric-specific rank are shown, and it is indicated whether a higher or lower value is desired.

Team	DSC (in %) ↑			mAP_{IoU} (in %) ↑			95% HD ↓			Avg. rank	Global rank
	$Mean_{BT}$	95% CI	Rank	$Mean_{BT}$	95% CI	Rank	$Mean_{BT}$	95% CI	Rank		
Jmees_inc.	35.48	[34.12, 36.96]	1	36.05	[34.51, 37.70]	1	75.48	[68.98, 82.41]	3	1.67	**1st**
Uniandes24	33.10	[31.43, 35.20]	2	33.39	[31.52, 35.66]	3	71.47	[56.32, 85.03]	1	2.00	**2nd**
Augi	32.95	[31.61, 34.56]	3	33.53	[31.99, 35.31]	2	73.10	[62.15, 82.24]	2	2.33	**3rd**
kist_harilab	26.72	[26.01, 27.54]	4	26.40	[25.63, 27.29]	4	163.91	[145.75, 182.09]	6	4.67	**4th**
floor9	21.64	[20.47, 23.07]	6	20.63	[19.36, 22.16]	6	82.91	[76.99, 89.24]	4	5.33	**5th**
sk	24.24	[23.59, 24.90]	5	23.73	[22.96, 24.50]	5	243.56	[222.71, 259.75]	7	5.67	**6th**
goncalo	18.57	[17.60, 19.73]	8	18.34	[17.19, 19.62]	8	155.12	[142.39, 166.73]	5	7.00	**7th**
Hanglok	18.92	[17.80, 20.15]	7	18.84	[17.45, 20.33]	7	246.13	[232.84, 259.82]	8	7.33	**8th**
recogna	13.46	[12.73, 14.24]	9	12.39	[11.57, 13.26]	9	300.12	[286.60, 312.53]	9	9.00	**9th**

Phase recognition − F1 bootstrap density
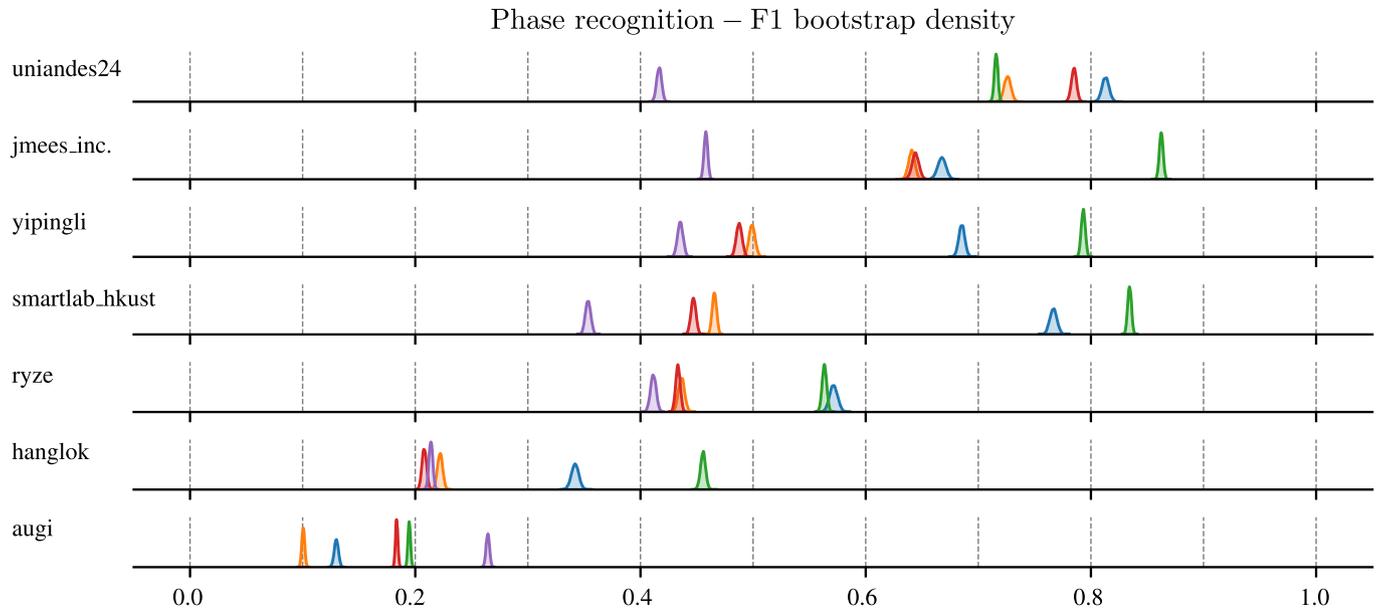
Fig. 7. KDE plot based on the bootstrap results for the participating teams of the surgical phase recognition task with regard to the F1-score values for the test set videos 6 ■, 8 ■, 9 ■, 12 ■, and 14 ■.

Phase recognition − BA bootstrap density
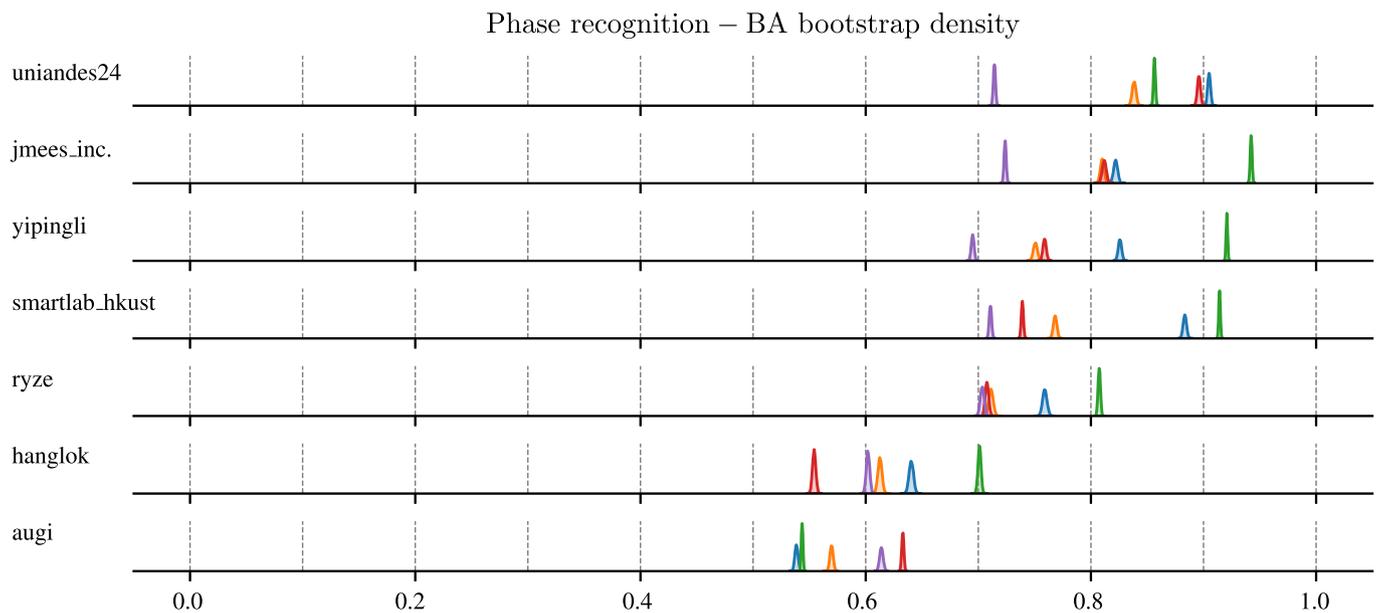
Fig. 8. KDE plot based on the bootstrap results for the participating teams of the surgical phase recognition task with regard to the BA values for the test set videos 6 ■, 8 ■, 9 ■, 12 ■, and 14 ■.

proved most challenging, with the lowest average scores (36.47% F1-score, 68.05% BA). The second-highest performance was observed for video six (56.80% F1-score, 76.77% BA), followed by video 12 (45.55% F1-score, 72.87% BA), and video eight (44.15% F1-score, 72.30% BA).

7.3. Task 2: instrument instance segmentation

Table 8 summarizes the detailed results of the instrument instance segmentation task. The team Jmees_inc. achieved the highest scores in two out of three evaluation metrics (35.48% DSC and 36.05% mAP_{IoU}) and ranked third in the 95% HD metric (75.48%), securing the best overall performance in the task. A largely consistent ranking pattern emerges between the DSC and mAP_{IoU} metrics, with only minor rank fluctuations caused by closely aligned performance values among com-

peting teams. In contrast, the ranking based on the 95% HD metric diverges considerably from those observed in the other two metrics. Notably, only the teams recogna and Augi exhibit stable rankings that align with their performance in the mAP_{IoU} metric, highlighting differences in boundary-level versus region-based segmentation performance.

Fig. 9 depicts the ranking stability of each participating team across the three evaluation metrics DSC (9a), mAP_{IoU} (9b), 95% HD (9c), as well as the global ranking based on aggregated metrics (9d). The results reveal that team rankings are considerably more stable for both the DSC and mAP_{IoU} metrics, indicating consistent performance across bootstrap iterations. In contrast, the rankings for the 95% HD metric exhibit greater variability, with a higher incidence of outliers and rank fluctuations, reflecting the increased sensitivity of this metric to boundary-level discrepancies. In the global ranking plot (Fig. 9d), a clear ordering
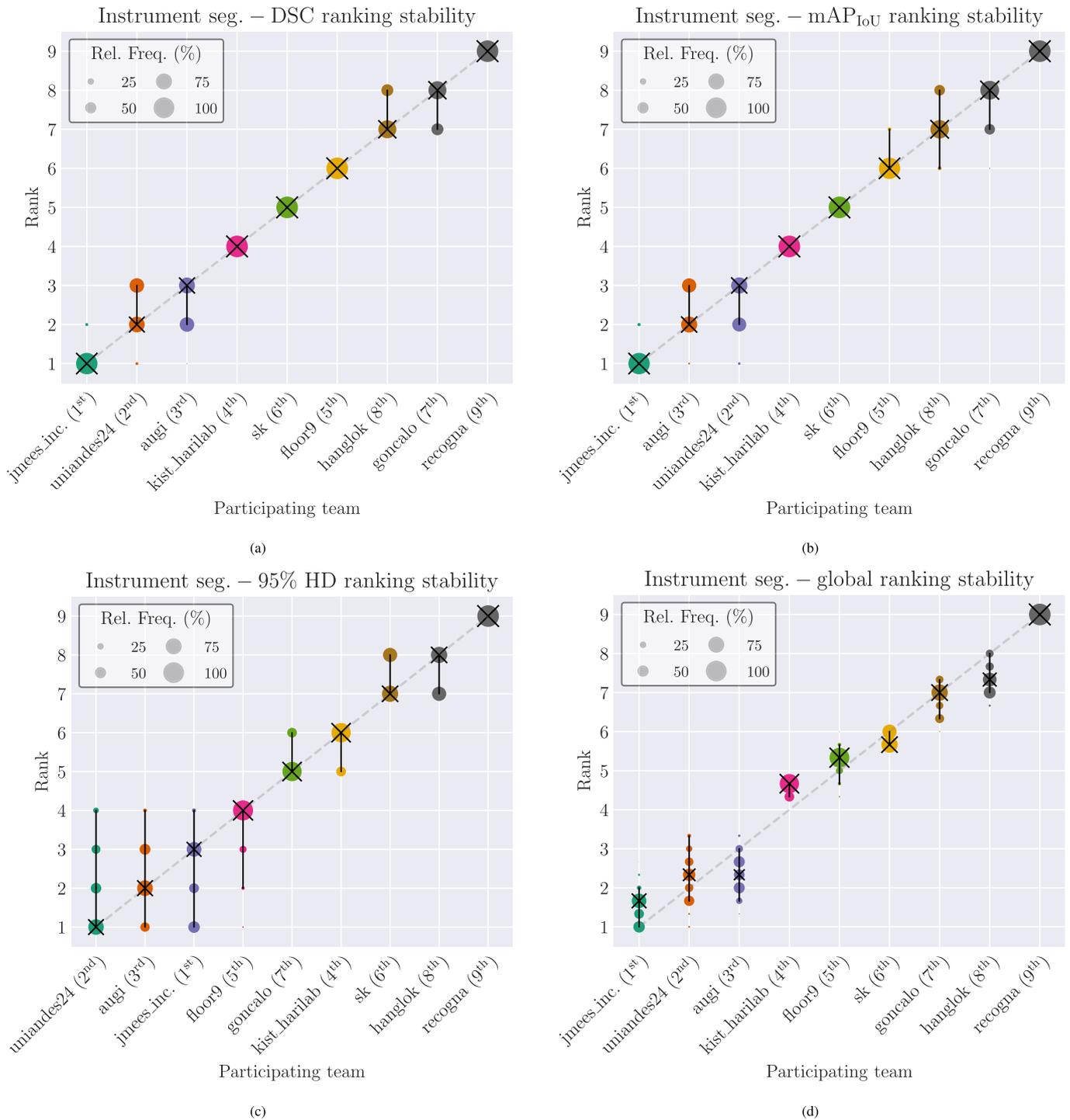
Fig. 9. Visualization of ranking stability based on bootstrapping for the instrument instance segmentation task for the metrics DSC (a), mAP$_{IoU}$ (b), 95% HD (c), and Global (d). For each metric, the individual ranking stability is shown and the global rank is indicated after each team name.

of teams is observed. Notably, the teams, Uniandes24 and Augi, which ranked second and third overall, demonstrate nearly indistinguishable performance.

Qualitative results are presented in Tables 9 and 10, showcasing eight representative test samples along with their respective ground truth segmentations. For each sample, the predicted outputs of the participating teams are visualized, and the corresponding metric values are reported. The dataset distribution is as follows: Samples one, four, five, and eight originate from hospital one, samples two and six from hospital two, and samples three and seven from hospital three,

with hospital identifiers listed in Fig. 1. Instruments that appeared frequently in the training data (e.g., pe-forceps, grasper, palpation-probe, clip-applicator, HFcoag-probe) were generally segmented with higher accuracy. In contrast, instruments with limited representation (e.g., scissor, blunt-grasper, blunt-grasper-spec., bipolar-clamp) were more prone to misclassification or missed detection. Additionally, a notable limitation across many submissions was the lack of true instance segmentation. Several teams instead applied semantic segmentation, failing to distinguish between multiple instances of the same instrument class. This issue is particularly evident in sample three, where
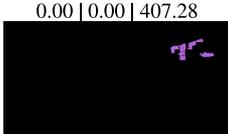
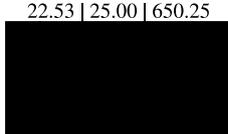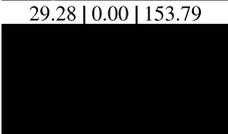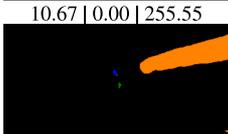**Table 9**

Visualization of example participant predictions, together with quantitative results for DSC (in %) ↑, mAP$_{IoU}$ (in %) ↑, and 95% HD ↓ below each image. An assignment of the color codings to the respective instrument categories is provided in Section 4.1 in B.

	Challenge data			
Team	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 4**
Input				
Ground truth				
	Participant predictions			
jmees_inc.	97.30 | 100.00 | 7.06	49.01 | 50.00 | 19.00	65.06 | 50.00 | 530.66	95.87 | 100.00 | 8.37
uniandes24	97.17 | 100.00 | 10.38	96.51 | 100.00 | 11.61	64.90 | 50.00 | 529.53	96.27 | 100.00 | 31.58
augi	97.76 | 100.00 | 9.64	96.50 | 100.00 | 14.52	64.97 | 50.00 | 377.58	96.61 | 100.00 | 31.37
kist_harilab	97.44 | 100.00 | 7.31	0.00 | 0.00 | 1045.36	49.07 | 50.00 | 10.00	92.49 | 100.00 | 13.82
floor9	96.70 | 100.00 | 10.91	0.00 | 0.00 | 208.81	96.65 | 100.00 | 12.52	94.03 | 100.00 | 33.84
sk	96.34 | 100.00 | 30.88	11.08 | 0.00 | 109.27	47.95 | 33.33 | 225.46	94.53 | 100.00 | 33.28
goncalo	96.12 | 100.00 | 35.79	0.00 | 0.00 | 211.69	15.01 | 0.00 | 806.70	26.12 | 28.57 | 35.67
hanglok	64.23 | 66.67 | 11.11	0.00 | 0.00 | nan	0.00 | 0.00 | nan	77.41 | 66.67 | 152.42
recogna	39.93 | 33.33 | 582.06	0.00 | 0.00 | 573.50	0.00 | 0.00 | 570.86	61.69 | 66.67 | 43.59

Table 10

Visualization of example participant predictions, together with quantitative results for DSC (in %) ↑, mAP$_{IoU}$ (in %) ↑, and 95% HD ↓ below each image. An assignment of the color codings to the respective instrument categories is provided in Section 4.1 in B.

|	Challenge data			
Team	Sample 5	Sample 6	Sample 7	Sample 8
Input				
Ground truth				
Participant predictions				
jmees_inc.	73.10 | 80.00 | 25.02	26.12 | 0.00 | 789.72	96.81 | 100.00 | 12.39	94.36 | 100.00 | 24.06
uniandes24	73.62 | 80.00 | 16.34	96.77 | 100.00 | 8.50	97.09 | 100.00 | 17.66	94.02 | 100.00 | 31.20
augi	92.12 | 100.00 | 18.09	32.28 | 33.33 | 9.33	97.50 | 100.00 | 9.66	88.97 | 100.00 | 95.51
kist_harilab	69.38 | 75.00 | 18.36	91.57 | 100.00 | 34.18	0.00 | 0.00 | nan	30.72 | 33.33 | 30.53
floor9	67.61 | 75.00 | 111.06	23.46 | 25.00 | 33.02	32.47 | 33.33 | 611.79	89.11 | 100.00 | 60.73
sk	69.32 | 75.00 | 28.73	32.49 | 33.33 | 25.06	15.14 | 0.00 | 335.29	46.39 | 66.67 | 430.51
goncalo	70.45 | 80.00 | 28.82	23.45 | 25.00 | 70.35	0.00 | 0.00 | 407.28	22.53 | 25.00 | 650.25
hanglok	83.37 | 100.00 | 165.37	29.28 | 0.00 | 153.79	10.67 | 0.00 | 255.55	0.00 | 0.00 | 1386.63
recogna	52.65 | 60.00 | 268.49	0.00 | 0.00 | nan	0.00 | 0.00 | 496.23	13.22 | 0.00 | 649.44
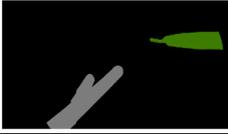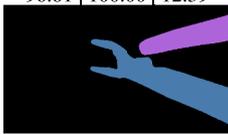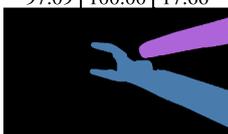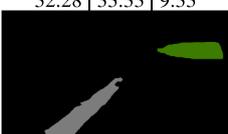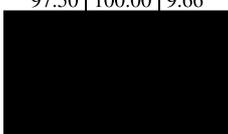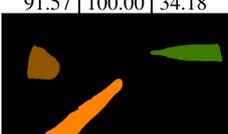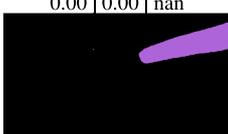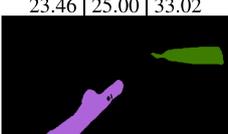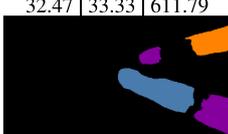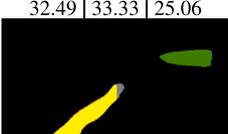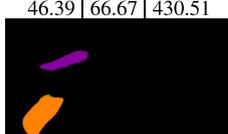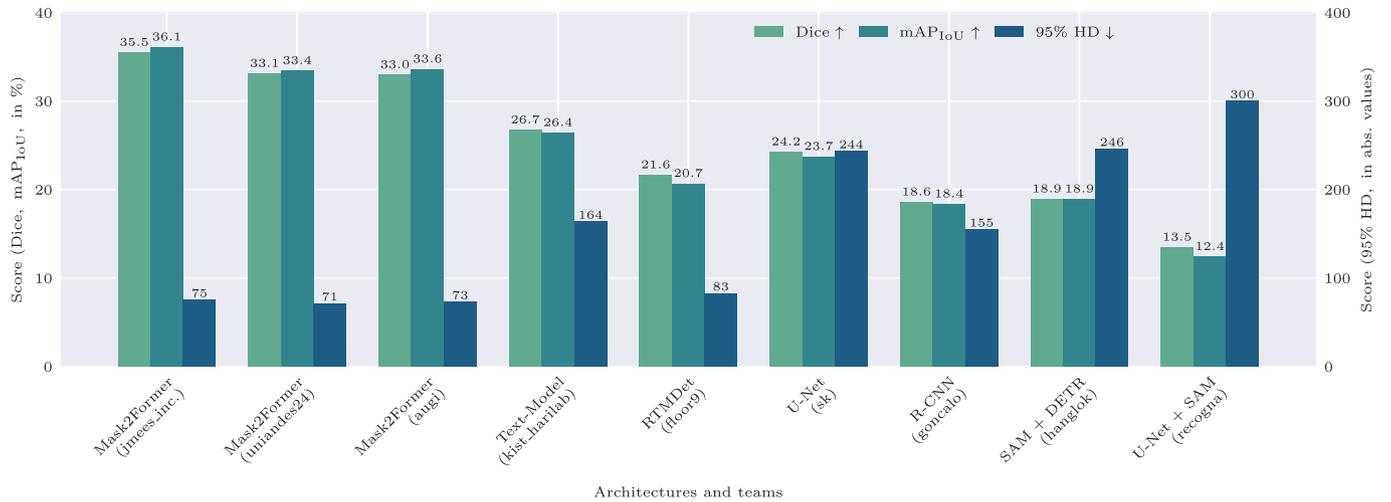
Fig. 10. Submitted methods together with the corresponding participants and results with regard to the instrument instance segmentation task. If a methodology consists of several parts, the primary approach is listed first.
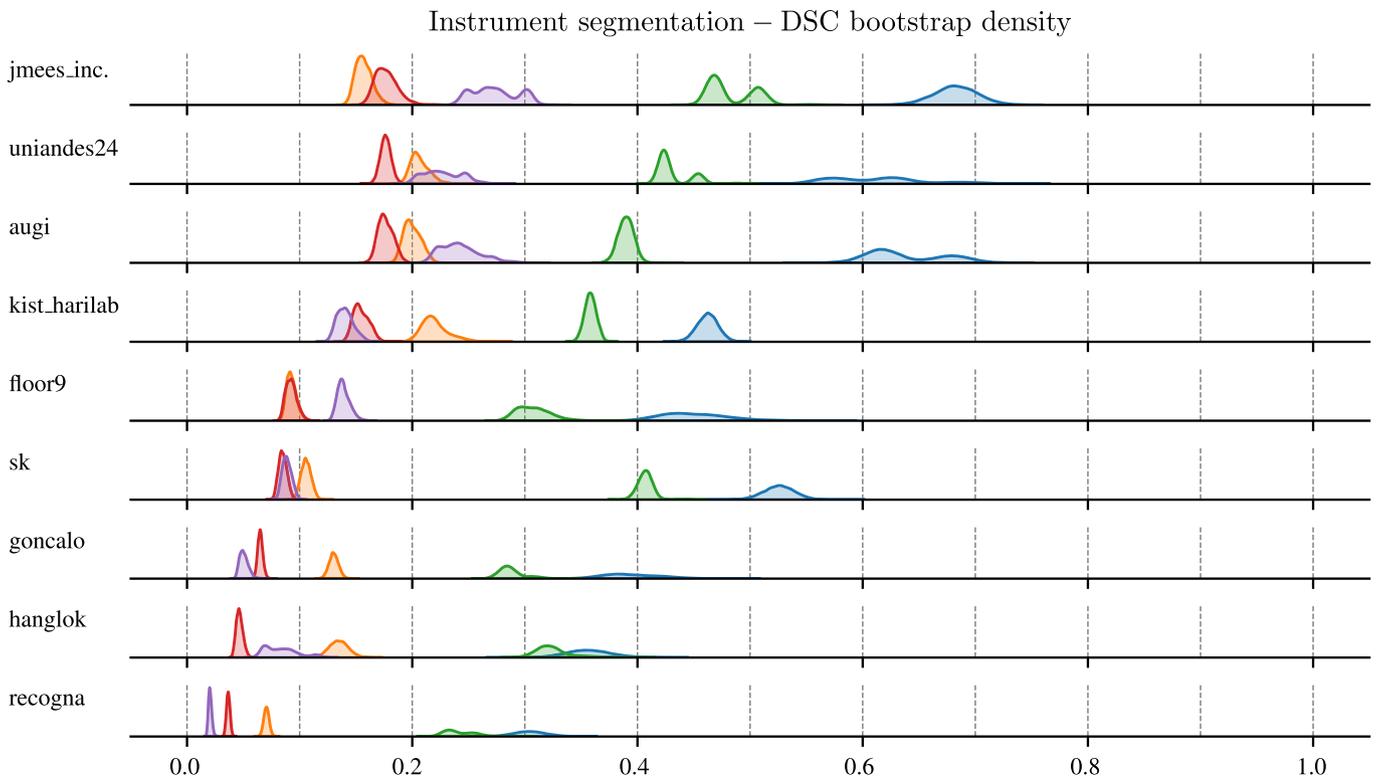

Fig. 11. KDE plot based on the bootstrap results for the participating teams of the instrument instance segmentation task with regard to the DSC values for the test set videos 6 ■, 8 ■, 9 ■, 12 ■, and 14 ■.

multiple algorithms generated reasonable segmentation masks for the blunt-grasper-spec. class but did not differentiate between separate instruments positioned at the top of the frame. This shortcoming negatively impacted performance metrics that rely on accurate instance separation.

Fig. 10 provides a comparative visualization of team performances across the three evaluation metrics, accompanied by an overview of the primary method employed by each team. In cases where a method comprised multiple components, only the principal approach is listed. A clear trend emerges: the top three performing teams all implemented transformer-based architectures, specifically the Mask2Former

model. In contrast, teams that relied on convolutional neural networks, such as U-Net and R-CNN, demonstrated notably lower performance across all metrics. Furthermore, teams utilizing Mask2Former achieved consistently strong results across all evaluation criteria, highlighting the architecture's robustness and adaptability for the instrument instance segmentation task. In comparison, alternative methods showed a discernible performance gap, suggesting the growing advantage of transformer-based approaches in medical image segmentation. Figs. 11, 12, and 13 display the KDE plots for the three evaluation metrics: DSC, mAP_{IoU}, and 95% HD. Across all participating teams, video six consistently yielded the highest performance for both DSC and mAP_{IoU}
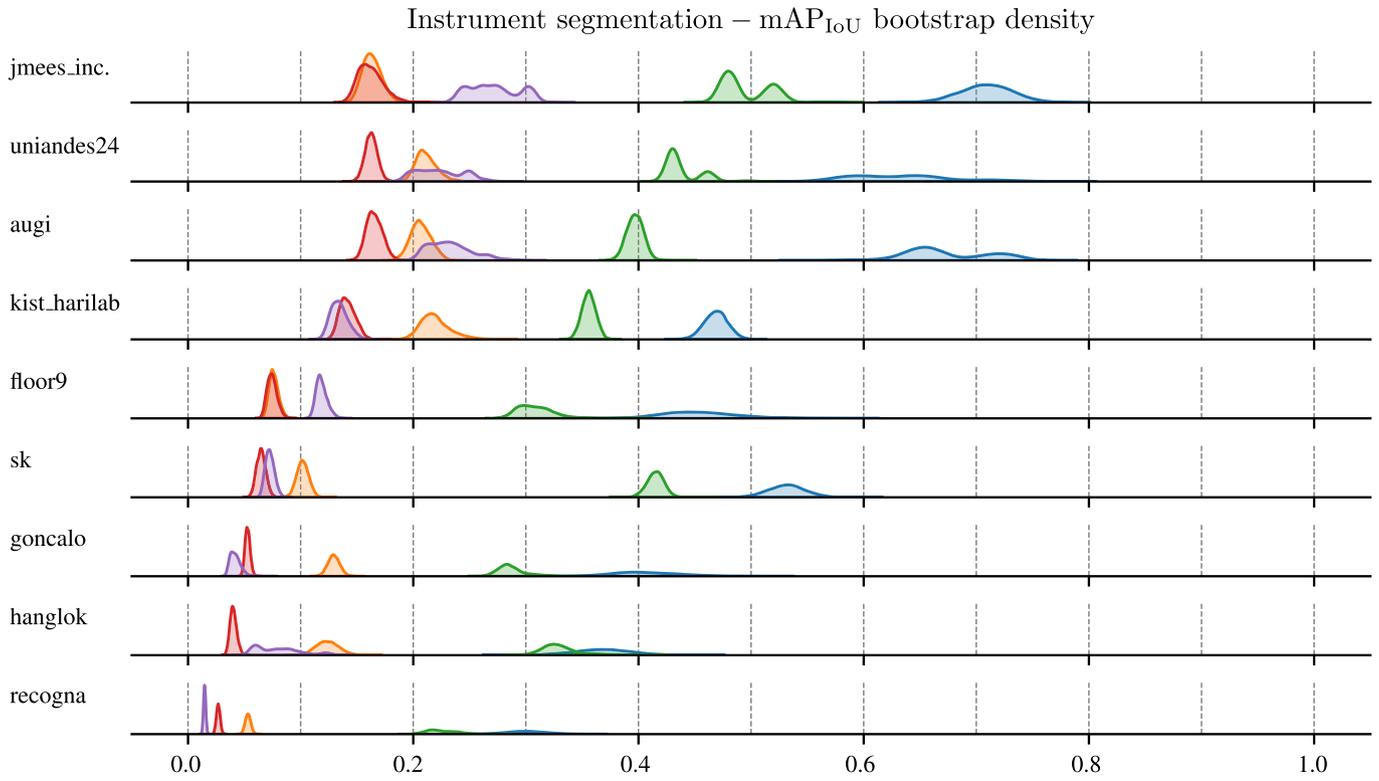
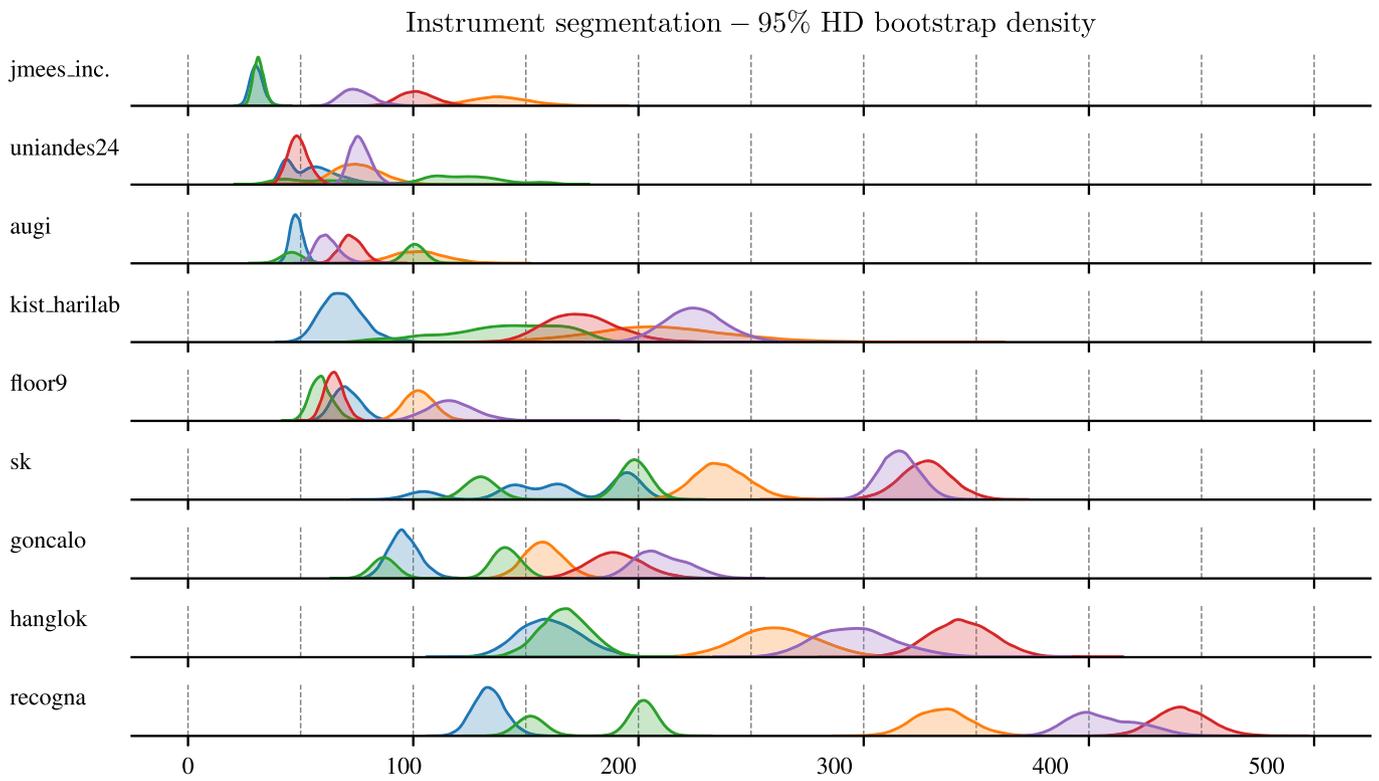Instrument segmentation − mAP$_{\text{IoU}}$ bootstrap density

Fig. 12. KDE plot based on the bootstrap results for the participating teams of the instrument instance segmentation task with regard to the mAP$_{\text{IoU}}$ values for the test set videos 6 ■, 8 ■, 9 ■, 12 ■, and 14 ■ .

Instrument segmentation − 95% HD bootstrap density

Fig. 13. KDE plot based on the bootstrap results for the participating teams of the instrument instance segmentation task with regard to the 95% HD values for the test set videos 6 ■, 8 ■, 9 ■, 12 ■, and 14 ■ .

Table 11

Results of the instrument instance segmentation task for each team and each individual video in the test dataset, together with the mean metric values for each video.

Team	DSC (in %) ↑					mAP$_{IoU}$ (in %) ↑					95% HD ↓				
	V_06	V_08	V_09	V_12	V_14	V_06	V_08	V_09	V_12	V_14	V_06	V_08	V_09	V_12	V_14
Jmees_inc.	68.26	15.70	48.35	17.70	27.40	70.90	16.34	49.57	16.23	27.21	30.50	138.61	31.38	101.83	75.09
Uniandes24	61.24	20.68	43.17	17.65	22.74	63.26	21.17	43.88	16.27	22.38	55.79	75.71	100.37	49.02	76.49
Augi	63.85	20.02	38.98	17.64	24.24	67.42	20.68	39.70	16.55	23.30	48.19	102.31	80.46	72.57	61.97
kist_harilab	46.23	22.03	35.82	15.47	14.04	46.88	21.97	35.59	14.16	13.41	68.48	209.79	141.25	174.87	225.17
floor9	45.01	9.28	30.73	9.26	13.90	45.43	7.63	30.77	7.47	11.87	70.59	102.64	59.18	64.93	117.20
sk	52.57	10.57	40.70	8.51	8.88	53.15	10.18	41.54	6.49	7.27	164.34	237.17	173.00	327.48	315.82
goncalo	39.48	13.04	28.76	6.51	5.06	40.61	12.94	28.65	5.29	4.20	95.91	158.02	121.35	190.05	210.28
Hanglok	35.64	13.48	32.45	4.67	8.36	36.81	12.40	32.99	4.03	7.99	160.83	262.00	167.43	343.76	296.62
recogna	30.43	7.05	24.09	3.66	2.05	30.03	5.32	22.45	2.69	1.48	133.71	334.87	183.70	441.37	406.95
mean	**49.19**	**14.65**	**35.89**	**11.23**	**14.07**	**50.50**	**14.29**	**36.13**	**9.91**	**13.23**	**92.04**	**180.12**	**117.57**	**196.21**	**198.40**

Table 12

Results and rankings for each individual team and metric regarding the surgical instrument keypoint estimation task, together with the averaged and overall rank for each team. For the mAP$_{OKS}$ metric, the mean value over the 10,000 bootstrap runs (Mean$_{BT}$), the 95% Confidence Interval (95% CI), and the metric-specific rank are shown, and it is indicated whether a higher or lower value is desired.

Team	mAP$_{OKS}$ (in %) ↑			Averaged	Global
	Mean$_{BT}$	95% CI	Rank	rank	rank
sds-hd	30.82	[29.17, 32.34]	1	1.0	**1st**
alvaro	15.29	[13.75, 16.75]	2	2.0	**2nd**

metrics, followed by video nine. In contrast, the remaining three videos exhibit substantially lower scores, indicating greater segmentation difficulty. While the KDE plot for the 95% HD metric reveals a generally similar trend, it is characterized by larger fluctuations and a less clearly defined ranking pattern. The broader Gaussian distributions observed for this metric – across all videos and methods – highlight its greater sensitivity to boundary discrepancies and higher variability compared to the DSC and mAP$_{IoU}$.

Table 11 summarizes the quantitative performance per video.

- Video six achieved the best results overall (49.19% DSC, 50.50% mAP$_{IoU}$, 92.04% 95% HD),
- followed by video nine (35.89% DSC, 36.13% mAP$_{IoU}$, 117.57% 95% HD),
- and video eight (14.65% DSC, 14.29% mAP$_{IoU}$, 180.12% 95% HD).

Performance for videos 12 and 14 was more mixed:

- video 14 yielded slightly better scores for DSC (14.07%) and mAP$_{IoU}$ (13.23%),
- whereas video 12 performed marginally better with respect to the 95% HD metric (196.21%).

These findings underscore the video-specific variability in segmentation difficulty and highlight the importance of evaluating performance across diverse test scenarios.

7.4. Task 3: instrument keypoint estimation

Table 12 presents the results for the instrument keypoint estimation task. Team sds-hd achieved the highest performance with a mAP$_{OKS}$ score of 30.82%, clearly outperforming team alvaro, which reached 15.29% mAP$_{OKS}$. As only two teams participated, the average rank equaled the global rank, directly determining the final ranking outcome.

The ranking stability analysis shown in Fig. 14 confirmed the robustness of these results. Team sds-hd consistently ranked first in all 10,000 bootstrap iterations, demonstrating highly stable performance across the
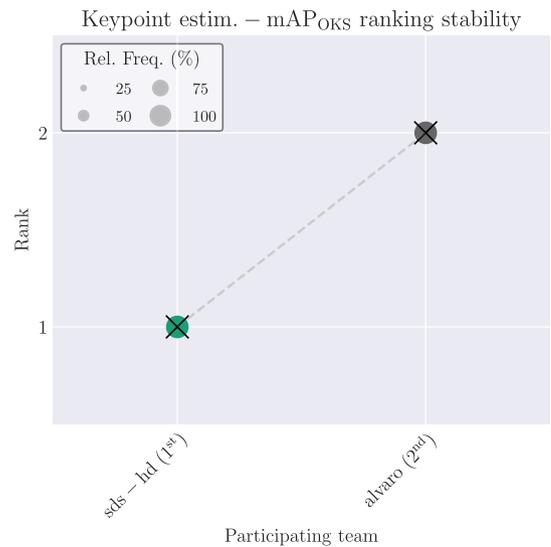

Fig. 14. Visualization of ranking stability based on bootstrapping for the instrument keypoint estimation task for the mAP$_{OKS}$ metric. The individual ranking stability for the mAP$_{OKS}$ metric is shown and the global rank is indicated after each team name.

test dataset. Since this task was evaluated using a single evaluation metric, the global ranking plot shown corresponds directly to this mAP$_{OKS}$ metric, and therefore only the ranking stability of mAP$_{OKS}$ is visualized.

Table 13 presents a selection of qualitative results for the instrument keypoint estimation task, using the same image samples as in Tables 9 and 10. Given the limited number of participating teams, all results are consolidated into a single table. The upper half displays the predicted keypoints and corresponding mAP$_{OKS}$ scores for the first four samples, while the lower half shows results for the remaining four. In several cases – specifically samples two, six, seven, and eight – the algorithms clearly struggled with accurate instrument classification, highlighting the inherent difficulty of the task. Consistent with the findings from the instance segmentation task, instruments with higher representation in the training set were generally estimated more accurately than those with fewer examples.

The KDE plot in Fig. 15 provides further insight into performance variability. It shows that both submitted algorithms most accurately predicted videos six and nine, whereas the remaining three videos exhibited substantially lower performance. The aggregated Gaussian distributions for team sds-hd appear slightly broader, indicating a greater spread of metric values and reflecting more variability in performance across different video samples compared to team alvaro.

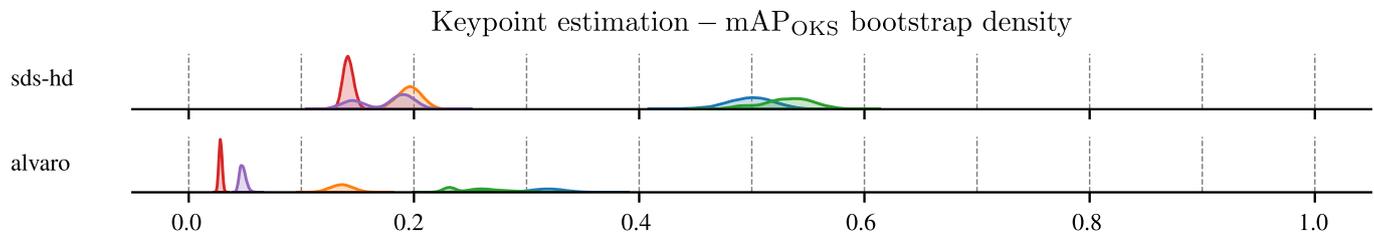Table 14 summarizes the average mAP$_{OKS}$ scores for each test video:

Fig. 15. KDE plot based on the bootstrap results for the participating teams of the instrument keypoint estimation task with regard to the mAP_{OKS} values for the test set videos 6 .

- Video six achieved the highest average score (39.78%),
- closely followed by video nine (39.28%),
- Video eight saw a notable drop (16.58%),
- while videos 14 (11.13%) and 12 (8.49%) yielded the lowest overall results.

These outcomes highlight the challenges of keypoint estimation, particularly for underrepresented instruments and more complex video sequences.

8. Discussion

In the following, we reflect on the key findings of our PhaKIR challenge and discuss potential limitations. We begin by discussing the organizational structure of the challenge (Section 8.1), and examining the characteristics of the provided dataset (Section 8.2). We then outline the evaluation methodology used to assess the submitted approaches (Section 8.3), and analyze the performance of the submitted methods for each individual task (Sections 8.4–8.6). We then derive overarching insights across all tasks (Section 8.7), before concluding with perspectives on future research directions in this domain (Section 8.8).

8.1. Challenge organization

As illustrated in Fig. 2, the PhaKIR challenge attracted the highest number of registrations and submissions among all EndoVis-2024 sub-challenges, highlighting both the high relevance of the research questions and the strong interest from the scientific community. Notably, PhaKIR was hosted on a bespoke platform developed according to the principles outlined in the "DIY Challenge Blueprint" publication (Klausmann et al., 2025). Self-hosting provided several advantages that likely contributed to the high level of engagement. The platform was fully customizable to the specific needs of the challenge, enabling streamlined submission workflows, tailored evaluation pipelines, and flexible task definitions. Participants were continuously kept informed via email updates and news posts on the challenge website, reflecting an active communication strategy by the organizers. This close interaction reduced participation barriers and further supported the recruitment of a large and diverse participant cohort. In addition, the large dataset size and the high quality of annotations offered a uniquely rich benchmarking opportunity, enhancing the attractiveness of the challenge. This combination of platform usability, dataset quality, task diversity, and proactive participant engagement distinguished PhaKIR from challenges hosted on standard portals such as Synapse or Grand Challenge. It should be noted, however, that PhaKIR comprises three distinct tasks, whereas most other sub-challenges focused on a single task. This difference should be considered when interpreting participation metrics.

8.2. Challenge dataset

Our PhaKIR challenge introduced a novel dataset specifically designed to overcome key limitations of existing datasets and previous challenges, as outlined in Chapter 1. By collecting surgical videos from

three different medical centers, we ensured a high degree of variability in both surgical phases and instrument usage – variability that not only exists between institutions but also within a single hospital or procedure type. To maintain consistency with established benchmarks, we adopted the widely used surgical phase definitions from the Cholec80 dataset (Twinanda et al., 2017) for the surgical phase recognition task. Only real clinical interventions were included, with no restrictions on instrument types or preselection of video material. Providing full-length video sequences of entire procedures allowed participants to leverage temporal context, a valuable yet underutilized aspect in surgical AI research (Rueckert et al., 2024). To ensure high-quality annotations, we followed a rigorous three-step process involving two independent validation stages, all guided by a detailed annotation protocol. This approach resulted in consistent and reliable labels across the dataset.

Nevertheless, several limitations remain. While the dataset contains a substantial number of annotated frames and complete procedural videos, it is based on only 13 patient cases, which may limit the variety of surgical scenarios. Furthermore, it focuses exclusively on minimally invasive cholecystectomies – a specific procedure type that restricts broader applicability to other laparoscopic surgeries. All videos were recorded under uniform conditions (25 fps, 1920×1080 resolution), which may not reflect the diversity of imaging systems used in clinical practice and could impact the generalizability of results. To strengthen future evaluations, it will be important to address these aspects through more extensive and diverse data collection, as well as systematic robustness analyses (Roß et al., 2021).
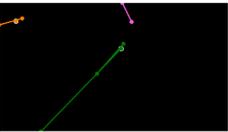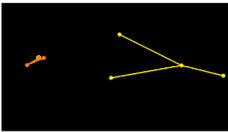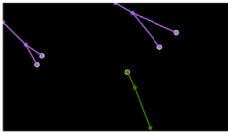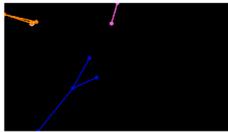
8.3. Assessment method

For the selection of evaluation metrics, we followed the recommendations of the Metrics Reloaded Framework by Maier-Hein et al. (2024). Accordingly, we employed the F1-score and BA for the surgical phase recognition task, DSC, mAP_{IoU}, and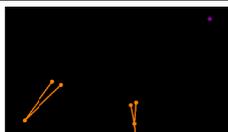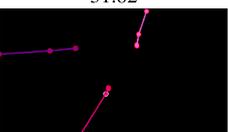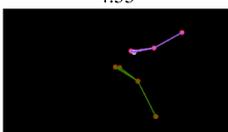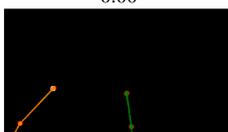 95% HD for the instrument instance segmentation task, and mAP_{OKS} for the instrument keypoint estimation task. Metric values were first computed individually for each phase or instrument class within each frame. These values were then aggregated across the sequence in two steps: first, we averaged across all phases or classes within a sequence using equal weights; second, we averaged across all sequences, again applying uniform weighting. This approach ensures that all categories contribute equally to the final metric score, preventing an overrepresentation of frequent phases or classes. An alternative approach – averaging per frame before aggregating - was considered but ultimately not adopted, as it would have been more susceptible to distortions caused by class imbalances. As shown in Fig. 4, the frequency of certain instruments varies widely. Our aggregation strategy was deliberately chosen to minimize the impact of such skewed distributions and ensure fairer comparisons.

To further enhance the robustness of our results, we applied a bootstrapping approach with 10,000 iterations for all evaluations. By averaging the metric-specific ranks instead of the raw metric values, the global ranking reflects consistency across multiple performance dimensions. This approach favors balanced methods and prevents single-metric dominance in the final ranking.

Table 13

Visualization of example participant predictions, together with quantitative results for mAP$_{OKS}$ (in %) ↑ below each image. White-bordered dots indicate predicted hidden keypoints, red-bordered circles indicate keypoiknts for which coordinates were predicted but marked as not present. According to the metric description in Section 5.3.3, for metric calculation, only the visibility flag of the GT keypoints is taken into account. An assignment of the color codings to the respective instrument categories is provided in Section 4.1 in B.

8.4. Task 1: surgical phase recognition

The results presented in Section 7.2 reveal notable variability in the performance of the submitted methods, regardless of the evaluation metric used. This is evident from the quantitative differences between teams, with an average Mean$_{BT}$ difference of 18.89 percentage points for the F1-score, and 5.93 percentage points for the BA. Ranking stability, as illustrated in Fig. 5, further underscores the spread in performance across submissions.

A methodical comparison reveals that the architectures employed are highly individual, suggesting that there is currently no consensus within the community on a common model for this task. It can further

Table 14

Results of the instrument keypoint estimation task for each team and each individual video in the test dataset, together with the mean metric values for each video.

Team	mAP_{OKS} (in %) ↑				
	V_06	V_08	V_09	V_12	V_14
sds-hd	49.79	19.58	53.08	14.17	17.46
alvaro	29.76	13.58	25.49	2.82	4.80
mean	**39.78**	**16.58**	**39.28**	**8.49**	**11.13**

Task 1: Lessons learned

Key design choices:

- Usage of various individual architectures
- Wide spread in performance across submissions
- Highly specialized methods outperform others
- Transformer-based models represent the majority
- Advanced strategies for temporal context processing are greatly beneficial

Failure cases:

- High number of misclassifications between phases
- Poor generalization ability between hospitals due to varying quantities of training data

Future research:

- Exploration on reasons for superior performance of specialized models
- Investigation for improved temporal context usage
- Enhancing generalization capabilities between medical centers

---

Task 2: Lessons learned

Key design choices:

- Top three methods based on Mask2Former
- Pseudo-label generation as key factor
- Superior results from transformer-based networks
- Usage of temporal information not decisive
- Individual pre-training strategies and datasets

Failure cases:

- Instrument instances rarely differentiated
- Underrepresented instruments often misclassified
- Poor generalization ability between hospitals due to varying quantities of training data

Future research:

- Investigation for improved temporal context usage
- Better identification of instrument instances
- Improved classification of underrepresented tools
- Enhancing generalization capabilities between medical centers

---

be stated that the highly specialized phase recognition architecture of the winning team Uniandes24 is superior to the general classification models used by the following ranked teams. This result indicates that the development of task-specific methods has a significant impact on performance and is particularly influential in conjunction with the size and quality of the available dataset. Furthermore, it can be assumed that the advanced use of temporal information in the task-specific architecture of the Uniandes24 team also contributes to the better results. Compared to the more straightforward approaches of the Jmees_inc. and Yipingli teams to incorporate temporal context, it is evident that Uniandes24's two-stage strategy for capturing temporal context across different time periods improves the results.

Our findings align with those of Wagner et al. (2023), especially concerning the challenges of accurately identifying the ClCu phase, which once again proved to be the most difficult for participating teams. Moreover, consistent with Wagner et al. (2023)'s earlier challenge, all methods submitted to PhaKIR incorporated temporal information, lending further support to the hypothesis that temporal modeling offers clear advantages over frame-wise processing in the context of surgical phase recognition. A methodological comparison between both challenges reveals an important shift: whereas the earlier study exclusively featured convolutional neural network approaches, all but one team (Ryze) in the PhaKIR challenge employed transformer-based architectures. This highlights an emerging consensus in the field, reflecting the growing preference for transformer-based models in surgical workflow analysis.

8.5. Task 2: instrument instance segmentation

The analysis of results in Section 7.3 highlights that the top three teams – Jmees_inc., Uniandes24, and Augi – achieved similarly strong performances, while a noticeable gap separates them from the remaining six teams. Rankings based on the DSC and mAP_{IoU} metrics were substantially more stable than those derived from the 95% HD, which showed greater variability.

As detailed in Table 4 and visualized in Fig. 10, the submitted methods demonstrated considerable architectural diversity. Nevertheless, all three top-performing teams based their solutions on Mask2Former, underscoring its current dominance for instance segmentation in surgical scenes. A comparison of the methodological approaches suggests that the semi-supervised learning strategy leveraging pseudo-label generation was likely the decisive factor behind the success of the Jmees_inc. team. Furthermore, it can be observed that the advanced use of temporal context implemented by team Uniandes24 did not confer a clear advantage over the two competitors who did not incorporate any temporal information.

Despite this, the qualitative evaluation – particularly sample 3 in Table 9 – revealed a key limitation: even the highest-ranking methods failed to distinguish between multiple instances of the same instrument class (of the blunt-grasper-spec.), underscoring the persistent challenge of true instance-level segmentation in cluttered surgical environments. Interestingly, only two of the nine teams incorporated temporal information into their approach, implying that most participants regarded temporal modeling as less beneficial for this task. One team even reported removing a temporal component after observing decreased performance during ablation testing. This suggests that the advantages of temporal integration may be highly task-specific in the context of surgical instrument segmentation.

8.6. Task 3: instrument keypoint estimation

The results of the instrument keypoint estimation task, as outlined in Section 7.4, reveal that only two teams participated. Nonetheless, a substantial performance gap was observed, as reflected by the mAP_{OKS} metric. A methodological comparison of the two approaches indicates that the introduction of uncertainty estimation techniques by team alvaro was not a decisive factor in securing victory. Although both teams

---

Task 3: Lessons learned

Key design choices:

- YOLOv8 architecture represents current standard
- Human pose estimation networks may not be suitable for instrument recognition
- Uncertainty estimation techniques could not improve performance
- Substantial differences in results due to deviations in training procedure

Failure cases:

- Underrepresented instruments often misclassified
- Coordinate accuracy varies across tool types
- Poor generalization ability between hospitals due to varying quantities of training data

Future research:

- Investigation for improved temporal context usage
- Improved classification of underrepresented tools
- More accurate predictions of keypoint coordinates
- Enhancing generalization capabilities between medical centers

---

built their approaches on the same underlying architecture, differences in training procedures contributed significantly to the observed performance discrepancy. Despite these efforts, the overall mAP$_{OKS}$ scores remained low for both teams. This can likely be attributed to the inherent limitations of pose estimation networks, which are typically optimized for human pose recognition involving a single class and a fixed number of keypoints. In contrast, the PhaKIR dataset includes multiple instrument classes, each associated with a class-specific number of keypoints. This added considerable complexity, as accurate classification became a critical prerequisite for precise keypoint localization.

A closer look at Table 13 reveals two central limitations of the submitted methods. Firstly, analogous to the instrument instance segmentation task, the instrument classes are often confused, particularly in the case of infrequently occurring tools. Secondly, the coordinates of the predicted keypoints are often rather inaccurate, which can also be seen especially in the case of rarer instruments.

8.7. Overall findings

An analysis of team participation per task reveals a clear trend: significantly more teams submitted methods for the surgical phase recognition ($n = 7$) and instrument instance segmentation ($n = 9$) tasks compared to the instrument keypoint estimation task ($n = 2$). This imbalance suggests that the former two tasks currently attract greater attention within the research community. However, it is important to note that the keypoint estimation task addresses clinically meaningful challenges, such as the localization of occluded instrument parts, not covered by the other tasks and highly relevant for computer-assisted surgery.

In terms of temporal modeling, all teams in the surgical phase recognition task incorporated temporal context, whereas only two out of nine teams did so for the instrument instance segmentation task, and none for the instrument keypoint estimation task. These findings indicate that temporal context integration is currently more established and accepted in phase recognition, while its potential in the other areas remains underexplored.

A further trend emerges in the architecture choices: in the two tasks where transformer-based models were used, the top-performing teams

consistently employed transformer-based approaches, suggesting a shift in preference towards such architectures for surgical video analysis.

Finally, cross-hospital comparison of test video results reveals a performance bias: algorithms performed best on videos from hospital 1 ($n = 3$), which was also more prominently represented in the training data. In contrast, performance on videos from hospital 2 ($n = 1$) and hospital 3 ($n = 1$) was consistently lower. These results indicate that the predominance of data from hospital 1 within the training set has a significant impact on test performance, enabling the models to achieve improved adaptability to this hospital's data distribution. This outcome also highlights a key limitation of all submissions – the challenge of generalizability across unseen clinical environments remains unsolved.

8.8. Future research directions

Future research across the three task domains presented in this challenge should place a strong emphasis on improving the generalizability of algorithmic approaches. None of the current methods demonstrated satisfactory performance when confronted with data from unfamiliar clinical environments – a critical shortcoming that must be addressed to enable real-world applicability. Tackling this challenge will require the development of more comprehensive and diverse datasets that include surgical procedures from a broad range of medical institutions and cover different types of surgeries. Such datasets should be explicitly designed to support generalization across centers, equipment, and surgical styles.

Additionally, future work could investigate cross-hospital generalization by training models on one or two centers and testing on a completely unseen third. Such experiments would help identify failure modes and provide insights into the robustness of methods under real-world distribution shifts.

Beyond dataset diversity, another key direction is the design of model architectures and algorithms that explicitly promote generalization. Architectural choices, regularization strategies, and training paradigms can all influence a model's ability to transfer across institutions, equipment, and surgical styles. For instance, approaches that incorporate domain-invariant features, multi-scale representations, or advanced attention mechanisms may improve robustness to distribution shifts. Investigating such strategies in combination with diverse training data could therefore provide a powerful way to enhance real-world applicability of the developed models.

In particular, the accurate detection and localization of instrument keypoints holds significant promise for real-world clinical applications, such as robotic assistance or intraoperative guidance, but remains underexplored. Advancing keypoint estimation methods tailored to the specific characteristics of surgical tools and contexts is therefore a key direction for future research.

Moreover, temporal modeling has proven effective in the surgical phase recognition task, but is still rarely employed in the other two domains. Exploring how temporal context can be harnessed in instance segmentation and keypoint estimation may lead to substantial performance gains and should be a focus of subsequent work.

Unfortunately, we did not have access to manually annotated metadata that would allow us to identify video sequences that are particularly challenging for state-of-the-art methods through an extended structural analysis described by Roß et al. (2023). Conducting such a systematic analysis also represents a promising direction for future research, as it could provide valuable insights into the underlying conditions that lead to failure cases and inform the development of more robust methods.

9. Conclusion

To overcome the limitations of existing datasets and earlier challenges, we organized the PhaKIR challenge – covering surgical phase recognition, keypoint estimation, and instrument instance segmentation – as part of the EndoVis event at MICCAI 2024. For this purpose, we introduced a novel multi-center dataset featuring complete real-world la-

paroscopic cholecystectomy sequences from 13 patients, annotated consistently for all three tasks. Using this dataset, we systematically benchmarked the performance of submitted algorithms and conducted an in-depth evaluation across a range of metrics. Our findings reveal that surgical phase recognition and instrument instance segmentation currently receive greater attention in the research community than the more complex and underexplored task of instrument keypoint estimation.

Temporal modeling was applied in all phase recognition approaches, confirming its value in this domain. In contrast, temporal context was rarely exploited in keypoint estimation or instrument instance segmentation methods, highlighting untapped potential in these areas. A key challenge identified across all tasks is generalizability: none of the submitted approaches demonstrated robust performance across surgical videos from different medical centers. We believe that the dataset and challenge presented here make a valuable contribution to the field of computer- and robot-assisted minimally invasive surgery. We hope the insights gained will stimulate further research, foster innovation, and support the development of more reliable and transferable AI methods in surgical contexts.

CRediT authorship contribution statement

Tobias Rueckert: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Formal analysis, Data curation, Conceptualization; **David Rauber:** Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization; **Raphaela Maerkl:** Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization; **Leonard Klausmann:** Validation, Software, Resources, Methodology, Conceptualization; **Suemeyye R. Yildiran:** Software, Methodology, Conceptualization; **Max Gutbrod:** Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization; **Danilo Weber Nunes:** Software, Methodology, Conceptualization; **Alvaro Fernandez Moreno:** Writing – review & editing, Software; **Imanol Luengo:** Writing – review & editing, Software; **Danail Stoyanov:** Writing – review & editing, Software; **Nicolas Toussaint:** Writing – review & editing, Software; **Enki Cho:** Writing – review & editing, Software; **Hyeon Bae Kim:** Writing – review & editing, Software; **Oh Sung Choo:** Writing – review & editing, Software; **Ka Young Kim:** Writing – review & editing, Software; **Seong Tae Kim:** Writing – review & editing, Software; **Gonçalo Arantes:** Writing – review & editing, Software; **Kehan Song:** Writing – review & editing, Software; **Jianjun Zhu:** Writing – review & editing, Software; **Junchen Xiong:** Writing – review & editing, Software; **Tingyi Lin:** Writing – review & editing, Software; **Shunsuke Kikuchi:** Writing – review & editing, Software; **Hiroki Matsuzaki:** Writing – review & editing, Software; **Atsushi Kouno:** Writing – review & editing, Software; **João Renato Ribeiro Manesco:** Writing – review & editing, Software; **João Paulo Papa:** Writing – review & editing, Software; **Tae-Min Choi:** Writing – review & editing, Software; **Tae Kyeong Jeong:** Writing – review & editing, Software; **Juyoun Park:** Writing – review & editing, Software; **Oluwatosin Alabi:** Writing – review & editing, Software; **Meng Wei:** Writing – review & editing, Software; **Tom Vercauteren:** Writing – review & editing, Software; **Runzhi Wu:** Writing – review & editing, Software; **Mengya Xu:** Writing – review & editing, Software; **An Wang:** Writing – review & editing, Software; **Long Bai:** Writing – review & editing, Software; **Hongliang Ren:** Writing – review & editing, Software; **Amine Yamlahi:** Writing – review & editing, Software; **Jakob Hennighausen:** Writing – review & editing, Software; **Lena Maier-Hein:** Writing – review & editing, Software; **Satoshi Kondo:** Writing – review & editing, Software; **Satoshi Kasai:** Writing – review & editing, Software; **Kousuke Hirasawa:** Writing – review & editing, Software; **Shu Yang:** Writing – review & editing, Software; **Yihui Wang:** Writing – review & editing, Software; **Hao Chen:** Writing – review & editing, Software; **Santiago Rodríguez:** Writing – review & editing, Software; **Nicolás Aparicio:** Writing – review & editing, Software; **Leonardo Manrique:** Writing – review & editing, Software; **Juan Camilo Lyons:** Writing – review & editing, Software; **Olivia Hosie:** Writing – review & editing, Software; **Nicolás Ayobi:** Writing – review & editing, Software; **Pablo Arbeláez:** Writing – review & editing, Software; **Yiping Li:** Writing – review & editing, Software; **Yasmina Al Khalil:** Writing – review & editing, Software; **Sahar Nasirihaghighi:** Writing – review & editing, Software; **Stefanie Speidel:** Writing – review & editing, Methodology, Conceptualization; **Daniel Rueckert:** Writing – review & editing, Supervision, Methodology, Conceptualization; **Hubertus Feussner:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Data curation, Conceptualization; **Dirk Wilhelm:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Data curation, Conceptualization; **Christoph Palm:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Data availability statement

The dataset was published under a Creative Commons AttributionNonCommercial-ShareAlike (CC BY-NC-SA) license on Zenodo (https://doi.org/10.5281/zenodo.15740620), which means that it is publicly available for non-commercial usage. Should you wish to use or refer to this dataset, you must cite this paper and the associated dataset paper (Rueckert et al., 2025b). The licensing of new creations must use the exact same terms as in the current version of the dataset.

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.

Acknowledgements

During the preparation of this work the authors used ChatGPT (GPT-4o, OpenAI) for language refinement to improve the clarity and readability of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix A. Challenge organization

Three parties were involved in the challenge organization. The research group for Minimally Invasive Interdisciplinary Therapeutic Intervention (MITI) (https://web.med.tum.de/en/miti/) at the Technical University of Munich (TUM) obtained the source data from the three different medical centers and created the annotations. The Regensburg Medical Image Computing (ReMIC) (https://re-mic.de/) laboratory at the Ostbayerische Technische Hochschule Regensburg (OTH Regensburg) carried out the technical implementation, which included the preprocessing and the provision of the data, the registration and administration of the participants, the submission process, and the evaluation and presentation of the results. The company AKTORmed - Robotic Surgery (https://aktormed.info/) provided the prize money of €3,000 which was distributed equally across the three tasks.

Appendix B. Data description and labeling instructions

See Supplementary file S1.

Appendix C. Submission instructions

Information on the submission process was provided to the participants both directly on the website (https://phakir.re-mic.de/participation/) and in a separate GitHub repository (https://github.com/schnadoslin/PhaKIR_Submission_Template/).

Appendix D. Challenge design document

See Supplementary file S2.

Appendix E. Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/10.1016/j.media.2026.103945.

References

Ahmed, F.A., Yousef, M., Ahmed, M.A., Ali, H.O., Mahboob, A., Ali, H., Shah, Z., Aboumarzouk, O., Al Ansari, A., Balakrishnan, S., 2024. Deep learning for surgical instrument recognition and segmentation in robotic-assisted surgeries: a systematic review. Artif. Intell. Rev. 58 (1), 1. https://doi.org/10.1007/s10462-024-10979-w

Aksamentov, I., Twinanda, A.P., Mutter, D., Marescaux, J., Padoy, N., 2017. Deep neural networks predict remaining surgery duration from cholecystectomy videos. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2017 - 20th International Conference, Quebec City, QC, Canada, September 11–13, 2017, Proceedings, Part II. Springer, pp. 586–593. https://doi.org/10.1007/978-3-319-66185-8_66

Al-Kababji, A., Bensaali, F., Dakua, S.P., 2022. Scheduling techniques for liver segmentation: reducelronplateau vs onecyclelr. In: International Conference on Intelligent Systems and Pattern Recognition. Springer, pp. 204–212. https://doi.org/10.1007/978-3-031-08277-1_17

Alabi, O., Toe, K. K.Z., Zhou, Z., Budd, C., Raison, N., Shi, M., Vercauteren, T., 2024. Cholecinstanceseg: a tool instance segmentation dataset for laparoscopic surgery. ArXiv preprint abs/2406.16039. https://doi.org/10.1038/s41597-025-05163-w

Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes-Hurtado, F., Flouty, E., Mohammed, A.K., Pedersen, M., Kori, A., Varghese, A., Krishnamurthi, G., Rauber, D., Mendel, R., Palm, C., Bano, S., Saibro, G., Shih, C., Chiang, H., Zhuang, J., Yang, J., Iglovikov, V., Dobrenkii, A., Reddiboina, M., Reddy, A., Liu, X., Gao, C., Unberath, M., Azizian, M., Stoyanov, D., Maier-Hein, L., Speidel, S., 2020. 2018 robotic scene segmentation challenge. ArXiv preprint abs/2001.11190. https://doi.org/10.48550/arXiv.2001.11190

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., García-Peraza, L.C., Li, W., Iglovikov, V., Luo, H., Yang, J., Stoyanov, D., Maier-Hein, L., Speidel, S., Azizian, M., 2019. 2017 robotic instrument segmentation challenge. https://doi.org/10.48550/arXiv.1902.06426

Ayobi, N., Pérez-Rondón, A., Rodríguez, S., Arbeláez, P., 2023. Matis: masked-attention transformers for surgical instrument segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1–5. https://doi.org/10.1109/ISBI53787.2023.10230819

Ayobi, N., Rodríguez, S., Pérez, A., Hernández, I., Aparicio, N., Dessevres, E., Peña, S., Santander, J., Caicedo, J.I., Fernández, N., Arbeláez, P., 2024. Pixel-wise recognition for holistic surgical scene understanding. https://doi.org/10.48550/arXiv.2401.11174

Bernhardt, S., Nicolau, S.A., Soler, L., Doignon, C., 2017. The status of augmented reality in laparoscopic surgery as of 2016. Medical Image Anal. 37, 66–90. https://doi.org/10.1016/J.MEDIA.2017.01.007

Bertasius, G., Wang, H., Torresani, L., 2021. Is space-time attention all you need for video understanding? In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021. PMLR, pp. 813–824. https://doi.org/10.48550/arXiv.2102.05095

Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M., 2020. Yolov4: optimal speed and accuracy of object detection. https://doi.org/10.48550/arXiv.2004.10934

Bodenstedt, S., Allan, M., Agustinos, A., Du, X., García-Peraza-Herrera, L.C., Kenngott, H., Kurmann, T., Müller-Stich, B.P., Ourselin, S., Pakhomov, D., Sznitman, R., Teichmann, M., Thoma, M., Vercauteren, T., Voros, S., Wagner, M., Wochner, P., Maier-Hein, L., Stoyanov, D., Speidel, S., 2018. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. https://doi.org/10.48550/arXiv.1805.02475

Bodenstedt, S., Jenke, A., Speidel, S., Martin Wagner, M.D., Tabibian, A., 2023. Sims: surgical instrument multi-domain segmentation challenge. https://www.synapse.org/Synapse:syn47193563/wiki/620035 [Accessed: 2025-07-16].

Bodenstedt, S., Speidel, S., Wagner, M., Chen, J., Kisilenko, A., Müller-Stich, B.P., Maier-Hein, L., 2021. Heichole surgical workflow analysis and full scene segmentation (heisurf). https://www.synapse.org/Synapse:syn25101790/wiki/608802 [Accessed: 2025-07-16].

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: Computer Vision - ECCV 2020. Springer, pp. 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

Chen, G., Huang, Y., Xu, J., Pei, B., Chen, Z., Li, Z., Wang, J., Li, K., Lu, T., Wang, L., 2024. Video mamba suite: state space model as a versatile alternative for video understanding. https://doi.org/10.48550/arXiv.2403.09626

Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E., 2020. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020. PMLR, pp. 1597–1607. https://doi.org/10.48550/arXiv.2002.05709

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022. IEEE, pp. 1280–1289. https://doi.org/10.1109/CVPR52688.2022.00135

Chinchor, N., 1991. MUC-3 evaluation metrics. In: Third Message Understanding Conference (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21–23, 1991. Association for Computational Linguistics.

Consortium, T.C., 2025. COCO - common objects in context - keypoint evaluation. https://cocodataset.org/#keypoints-eval. Accessed: 2025-07-16.

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q., 2020. Randaugment: practical automated data augmentation with a reduced search space. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 3008–3017. https://doi.org/10.1109/CVPRW50498.2020.00359

Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N., 2020. TeCNO: surgical phase recognition with multi-stage temporal convolutional networks. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III. Springer, pp. 343–352. https://doi.org/10.1007/978-3-030-59716-0_33

Dagnino, G., Kundrat, D., 2024. Robot-assistive minimally invasive surgery: trends and future directions. https://doi.org/10.1007/s41315-024-00341-2

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302. https://doi.org/10.2307/1932409

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021. OpenReview.net. https://doi.org/10.48550/arXiv.2010.11929

Efron, B., 1992. Bootstrap methods: another look at the jackknife. In: Breakthroughs in Statistics: Methodology and Distribution. Springer, pp. 569–593. https://doi.org/10.1214/aos/1176344552

Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C. K.I., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vision 111, 98–136. https://doi.org/10.1007/s11263-014-0733-5

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C., 2021. Multiscale vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. IEEE, pp. 6804–6815. https://doi.org/10.1109/ICCV48922.2021.00675

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y., 2024. Eva-02: a visual representation for neon genesis. Image Vision Comput. 149, 105171. https://doi.org/10.1016/j.imavis.2024.105171

Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P., 2021. Trans-SVNet: accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2021. Springer, pp. 593–603. https://doi.org/10.1007/978-3-030-87202-1_57

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T., Cubuk, E.D., Le, Q.V., Zoph, B., 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. Computer Vision Foundation / IEEE, pp. 2918–2928. https://doi.org/10.1109/CVPR46437.2021.00294

Good, I.J., 1956. Some terminology and notation in information theory. Proc. IEEE-Part C Monograph. 103 (3), 200–204. https://doi.org/10.1049/pi-c.1956.0024

Gruijthuijsen, C., Garcia-Peraza-Herrera, L.C., Borghesan, G., Reynaerts, D., Deprest, J., Ourselin, S., Vercauteren, T., Vander Poorten, E., 2022. Robotic endoscope control via autonomous instrument tracking. Front. Rob. AI 9, 832208. https://doi.org/10.3389/frobt.2022.832208

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B., 2020. Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2020 IEEE, pp. 9726–9735. https://doi.org/10.1109/CVPR42600.2020.00975

He, K., Gkioxari, G., Dollár, P., Girshick, R.B., 2017. Mask R-CNN. In: IEEE International Conference on Computer Vision, ICCV 2017. IEEE Computer Society, pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.322

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016. IEEE Computer Society, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

Ho, Y., Wookey, S., 2020. The real-world-weight cross-entropy loss function: modeling the costs of mislabeling. https://doi.org/10.48550/arXiv.2001.00570

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15 (9), 850–863. https://doi.org/10.1109/34.232073

Jaccard, P., 1912. The distribution of the flora in the alpine zone. New Phytol. 11 (2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jaspers, T. J.M., de Jong, R. L. P.D., Khalil, Y.A., Zeelenberg, T., Kusters, C. H.J., Li, Y., van Jaarsveld, R.C., Bakker, F. H.A., Ruurda, J.P., Brinkman, W.M., de With, P. H.N., van der Sommen, F., 2024. Exploring the effect of dataset diversity in self-supervised

learning for surgical computer vision. In: Data Engineering in Medical Imaging - Second MICCAI Workshop, DEMI 2024. Springer, pp. 43–53. https://doi.org/10.1007/978-3-031-73748-0_5

Johnson, O.V., XinYing, C., Johnson, O.E., Khaw, K.W., Lee, M.H., 2023. Learning rate schedules and optimizers, a game changer for deep neural networks. In: International Conference of Reliable Information and Communication Technology. Springer, pp. 327–340. https://doi.org/10.1007/978-3-031-59711-4_28

Katić, D., Wekerle, A.-L., Görtler, J., Spengler, P., Bodenstedt, S., Röhl, S., Suwelack, S., Kenngott, H.G., Wagner, M., Müller-Stich, B.P., Dillmann, R., Speidel, S., 2013. Context-aware augmented reality in laparoscopic surgery. Comput. Med. Imaging Graph. 37 (2), 174–182. https://doi.org/10.1016/j.compmedimag.2013.03.003

Kaur, P., Khehra, B.S., Mavi, E.B.S., 2021. Data augmentation for object detection: a review. In: 2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS). IEEE, pp. 537–543. https://doi.org/10.1109/MWSCAS47672.2021.9531849

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A., 2017. The kinetics human action video dataset. https://doi.org/10.48550/arXiv.1705.06950

Kiefer, J., Wolfowitz, J., 1952. Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. , 462–466. https://doi.org/10.1214/aoms/1177729392

Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015. ICLR. https://doi.org/10.48550/arXiv.1412.6980

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B., 2023. Segment anything. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023. IEEE, pp. 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371

Klausmann, L., Rueckert, T., Rauber, D., Maerkl, R., Yildiran, S.R., Gutbrod, M., Palm, C., 2025. DIY challenge blueprint: from organization to technical realization in biomedical image analysis. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2025. Springer, pp. 85–95. https://doi.org/10.1007/978-3-032-05141-7_9

Klinker, F., 2011. Exponential moving average versus moving exponential average. Math. Semesterber. 58, 97–107. https://doi.org/10.1007/s00591-010-0080-8

Kuhn, H.W., 2010. The hungarian method for the assignment problem. In: 50 Years of Integer Programming 1958–2008 - from the Early Years to the State-of-the-Art. Springer, pp. 29–47. https://doi.org/10.1007/978-3-540-68279-0_2

Lam, K., Chen, J., Wang, Z., Iqbal, F.M., Darzi, A., Lo, B., Purkayastha, S., Kinross, J.M., 2022. Machine learning for technical skill assessment in surgery: a systematic review. npj Dig. Med. 5 (1), 24. https://doi.org/10.1038/s41746-022-00566-0

Law, H., Ghani, K., Deng, J., 2017. Surgeon technical skill assessment using computer vision based analysis. In: Machine Learning for Healthcare Conference. Vol. 68. PMLR, pp. 88–99.

Lee, C., Xie, S., Gallagher, P.W., Zhang, Z., Tu, Z., 2015. Deeply-supervised nets. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015. JMLR.org. https://doi.org/10.48550/arXiv.1409.5185

Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. In: Computer Vision - ECCV 2014. Springer, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, Y., Boels, M., García-Peraza-Herrera, L.C., Vercauteren, T., Dasgupta, P., Granados, A., Ourselin, S., 2025. Lovit: long video transformer for surgical phase recognition. Medical Image Anal. 99, 103366. https://doi.org/10.1016/J.MEDIA.2024.103366

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. IEEE, pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022. IEEE, pp. 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167

Loshchilov, I., Hutter, F., 2017. SGDR: stochastic gradient descent with warm restarts. In: 5th International Conference on Learning Representations, ICLR 2017. OpenReview.net. https://doi.org/10.48550/arXiv.1608.03983

Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., Chen, K., 2022. Rtmdet: an empirical study of designing real-time object detectors. https://doi.org/10.48550/arXiv.2212.07784

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M.A., Wiesenfarth, M., Kavur, A.E., Sudre, C.H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, A., Rädsch, T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M.B., Cardoso, M.J., Cheplygina, V., Cimini, B.A., Collins, G.S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Haase, R., Hashimoto, D.A., Hoffman, M.M., Huisman, M., Jannin, P., Kahn, C.E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B.A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A.L., Mattson, P., Meijering, E., Menze, B., Moons, K. G.M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C.I., Shetty, S., van Smeden, M., Summers, R.M., Taha, A.A., Tiulpin, A., Tsaftaris, S.A., Van Calster, B., Varoquaux, G., Jäger, P.F., 2024. Metrics reloaded: recommendations for image analysis validation. Nat. Methods 21 (2), 195–212. https://doi.org/10.1038/s41592-023-02151-z

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., 2020. BIAS: transparent reporting of biomedical image analysis challenges. Medical Image Analysis 66, 101796. https://doi.org/10.1016/J.MEDIA.2020.101796

Malpani, A., Glock, K.M., 2023. Syn-ISS: synthetic data for instrument segmentation in surgery. https://www.synapse.org/Synapse:syn50908388/wiki/620516 [Accessed: 2025-07-16].

Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., 2018. Spectral normalization for generative adversarial networks. In: 6th International Conference on Learning Representations, ICLR 2018. OpenReview.net. https://doi.org/10.48550/arXiv.1802.05957

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H.S., Gal, Y., 2021. Deep deterministic uncertainty: a simple baseline. https://doi.org/10.48550/arXiv.2102.11582

Müller, D., Soto-Rey, I., Kramer, F., 2022. Towards a guideline for evaluation metrics in medical image segmentation. BMC Res. Notes 15 (1), 210. https://doi.org/10.1186/s13104-022-06096-y

Nwoye, C.I., Yu, T., Sharma, S., Murali, A., Alapatt, D., Vardazaryan, A., Yuan, K., Hajek, J., Reiter, W., Yamlahi, A., Smidt, F., Zou, X., Zheng, G., Oliveira, B., Torres, H.R., Kondo, S., Kasai, S., Holm, F., Özsoy, E., Gui, S., Li, H., Raviteja, S., Sathish, R., Poudel, P., Bhattarai, B., Wang, Z., Rui, G., Schellenberg, M., Vilaça, J.L., Czempiel, T., Wang, Z., Sheet, D., Thapa, S.K., Berniker, M., Godau, P., Morais, P., Regmi, S., Tran, T.N., Fonseca, J.C., Nölke, J., Lima, E., Vazquez, E., Maier-Hein, L., Navab, N., Mascagni, P., Seeliger, B., Gonzalez, C., Mutter, D., Padoy, N., 2023. Cholectriplet2022: show me a tool and tell me the triplet - an endoscopic vision challenge for surgical action triplet detection. Medical Image Anal. 89, 102888. https://doi.org/10.1016/J.MEDIA.2023.102888

Pérez, A., Rodríguez, S., Ayobi, N., Aparicio, N., Dessevres, E., Arbeláez, P., 2024. MuST: multi-scale transformers for surgical phase recognition. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2024. Springer, pp. 422–432. https://doi.org/10.1007/978-3-031-72089-5_40

Postels, J., Segù, M., Sun, T., Sieber, L.D., Gool, L.V., Yu, F., Tombari, F., 2022. On the practicality of deterministic epistemic uncertainty. In: International Conference on Machine Learning, ICML 2022. PMLR, pp. 17870–17909. https://doi.org/10.48550/arXiv.2107.00649

Psychogyios, D., Colleoni, E., van Amsterdam, B., Li, C., Huang, S., Li, Y., Jia, F., Zou, B., Wang, G., Liu, Y., Boels, M., Huo, J., Sparks, R., Dasgupta, P., Granados, A., Ourselin, S., Xu, M., Wang, A., Wu, Y., Bai, L., Ren, H., Yamada, A., Harai, Y., Ishikawa, Y., Hayashi, K., Simoens, J., DeBacker, P., Cisternino, F., Furnari, G., Mottrie, A., Ferraguti, F., Kondo, S., Kasai, S., Hirasawa, K., Kim, S., Lee, S.H., Lee, K.E., Kong, H., Fu, K., Li, C., An, S., Krell, S., Bodenstedt, S., Ayobi, N., Pérez, A., Rodríguez, S., Puentes, J., Arbeláez, P., Mohareri, O., Stoyanov, D., 2024. SAR-RARP50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. https://doi.org/10.48550/arXiv.2401.00496

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021. PMLR, pp. 8748–8763. https://doi.org/10.48550/arXiv.2103.00020

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C., 2024. Sam 2: segment anything in images and videos. https://doi.org/10.48550/arXiv.2408.00714

Reinke, A., Tizabi, M.D., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Kavur, A.E., Rädsch, T., Sudre, C.H., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Buettner, F., Cardoso, M.J., Cheplygina, V., Chen, J., Christodoulou, E., Cimini, B.A., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Glocker, B., Godau, P., Hashimoto, D.A., Hoffman, M.M., Huisman, M., Isensee, F., Jannin, P., Kahn, C.E., Kainmueller, D., Kainz, B., Karargyris, A., Kleesiek, J., Kofler, F., Kooi, T., Kopp-Schneider, A., Kozubek, M., Kreshuk, A., Kurc, T., Landman, B.A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A.L., Meijering, E., Menze, B., Moons, K. G.M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rafelski, S.M., Rajpoot, N., Reyes, M., Riegler, M.A., Rieke, N., Saez-Rodriguez, J., Sánchez, C.I., Shetty, S., Summers, R.M., Taha, A.A., Tiulpin, A., Tsaftaris, S.A., Van Calster, B., Varoquaux, G., Yaniv, Z.R., Jäger, P.F., Maier-Hein, L., 2024. Understanding metric-related pitfalls in image analysis validation. Nat. Methods 21 (2), 182–194. https://doi.org/10.1038/s41592-023-02150-0

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Springer, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Roß, T., Bruno, P., Reinke, A., Wiesenfarth, M., Koeppel, L., Full, P.M., Pekdemir, B., Godau, P., Trofimova, D., Isensee, F., Adler, T.J., Tran, T.N., Moccia, S., Calimeri, F., Müller-Stich, B.P., Kopp-Schneider, A., Maier-Hein, L., 2023. Beyond rankings: learning (more) from algorithm validation. Med. Image Anal. 86, 102765. https://doi.org/10.1016/j.media.2023.102765

Roß, T., Bruno, P., Reinke, A., Wiesenfarth, M., Koeppel, L., Full, P.M., Pekdemir, B., Godau, P., Trofimova, D., Isensee, F., Moccia, S., Calimeri, F., Müller-Stich, B.P., Kopp-Schneider, A., Maier-Hein, L., 2021. How can we learn (more) from challenges? a statistical approach to driving future algorithm development. https://doi.org/10.48550/arXiv.2106.09302

Roß, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Mîndroc-Filimon, D., Scholz, P., Tran, T.N., Bruno, P., Arbeláez, P., Bian, G., Bodenstedt, S., Bolmgren, J.L., Sánchez, L.B., Chen, H., González, C., Guo, D., Halvorsen, P., Heng, P., Hosgor, E., Hou, Z., Isensee, F., Jha, D., Jiang, T., Jin, Y., Kirtaç, K., Kletz, S., Leger, S., Li, Z., Maier-Hein, K.H., Ni, Z., Riegler, M.A., Schoeffmann, K., Shi, R., Speidel, S., Stenzel, M., Twick, I., Wang, G., Wang, J., Wang, L., Wang, L., Zhang, Y., Zhou, Y., Zhu, L., Wiesenfarth, M., Kopp-Schneider, A., Müller-Stich, B.P., Maier-Hein, L., 2021. Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge. Medical Image Anal. 70, 101920. https://doi.org/10.1016/J.MEDIA.2020.101920

Rueckert, T., Maerkl, R., Rauber, D., Klausmann, L., Gutbrod, M., Rueckert, D., Feussner, H., Wilhelm, D., Palm, C., 2025a. PhaKIR dataset - surgical procedure phase, keypoint, and instrument recognition. https://doi.org/10.5281/zenodo.15740620

Rueckert, T., Maerkl, R., Rauber, D., Klausmann, L., Gutbrod, M., Rueckert, D., Feussner, H., Wilhelm, D., Palm, C., 2025b. Video dataset for surgical phase, keypoint, and

instrument recognition in laparoscopic surgery (phaKIR). https://doi.org/10.48550/arXiv.2511.06549

Rueckert, T., Rueckert, D., Palm, C., 2024. Methods and datasets for segmentation of minimally invasive surgical instruments in endoscopic images and videos: a review of the state of the art. Comput. Biol. Med. 169, 107929. https://doi.org/10.1016/J.COMPBIOMED.2024.107929

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vision 115, 211–252. https://doi.org/10.48550/arXiv.1409.0575

Scott, D.W., 2015. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons. https://doi.org/10.1002/9780470316849

Silverman, B.W., 2018. Density estimation for statistics and data analysis. Routledge. https://doi.org/10.1002/bimj.4710300745

Smith, L.N., Topin, N., 2019. Super-convergence: very fast training of neural networks using large learning rates. In: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. Vol. 11006. SPIE, pp. 369–386. https://doi.org/10.1117/12.2520589

Speidel, S., Bodenstedt, S., Kenngott, H., Wagner, M., Müller-Stich, B.P., Maier-Hein, L., 2017. Surgical workflow analysis in the sensorOR. https://endovissub2017-workflow.grand-challenge.org/ [Accessed: 2025-07-16].

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC Med. Imaging 15, 1–28. https://doi.org/10.1186/s12880-015-0068-x

Tanzi, L., Piazzolla, P., Vezzetti, E., 2020. Intraoperative surgery room management: a deep learning perspective. Int. J. Med. Rob. Comput. Assisted Surg. 16 (5), 1–12. https://doi.org/10.1002/rcs.2136

Tharwat, A., 2021. Classification assessment methods. Appl. Comput. Inf. 17 (1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N., 2017. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Medical Imaging 36 (1), 86–97. https://doi.org/10.1109/TMI.2016.2593957

Van Rijsbergen, C., 1979. Information retrieval: theory and practice. In: Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems. IBM/University of Newcastle, pp. 1–14.

Varghese, R., Sambath, M., 2024. Yolov8: a novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). IEEE, pp. 1–6. https://doi.org/10.1109/ADICS58448.2024.10533619

Wagner, M., Müller-Stich, B.P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., Reinke, A., Reid, C., Yu, T., Vardazaryan, A., Nwoye, C.I., Padoy, N., Liu, X., Lee, E., Disch, C., Meine, H., Xia, T., Jia, F., Kondo, S., Reiter, W., Jin, Y., Long, Y., Jiang, M., Dou, Q., Heng, P., Twick, I., Kirtaç, K., Hosgor, E., Bolmgren, J.L., Stenzel, M., von Siemens, B., Zhao, L., Ge, Z., Sun, H., Xie, D., Guo, M., Liu, D., Kenngott, H.G., Nickel, F., von Frankenberg, M., Mathis-Ullrich, F., Kopp-Schneider, A., Maier-Hein, L., Speidel, S., Bodenstedt, S., 2023. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Med. Image Anal. 86, 102770. https://doi.org/10.1016/J.MEDIA.2023.102770

Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci. Rep. 11 (1), 2369. https://doi.org/10.1038/s41598-021-82017-6

Yang, S., Luo, L., Wang, Q., Chen, H., 2024. Surgformer: surgical transformer with hierarchical temporal attention for surgical phase recognition. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2024. Springer, pp. 606–616. https://doi.org/10.1007/978-3-031-72089-5_57

Yi, F., Wen, H., Jiang, T., 2021. Asformer: transformer for action segmentation. In: 32nd British Machine Vision Conference 2021, BMVC 2021. BMVA Press, p. 236. https://doi.org/10.48550/arXiv.2110.08568

Yu, W., Si, C., Zhou, P., Luo, M., Zhou, Y., Feng, J., Yan, S., Wang, X., 2023. Metaformer baselines for vision. IEEE Trans. Pattern Anal. Mach. Intell. https://doi.org/10.1109/TPAMI.2023.3329173

Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z., 2024. SurgicalSAM: efficient class promptable surgical instrument segmentation. In: Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014. AAAI Press, pp. 6890–6898. https://doi.org/10.1609/AAAI.V38I7.28514

Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D., 2018. Mixup: beyond empirical risk minimization. In: 6th International Conference on Learning Representations, ICLR 2018. OpenReview.net. https://doi.org/10.48550/arXiv.1710.09412

Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2020. Random erasing data augmentation. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020. AAAI Press, pp. 13001–13008. https://doi.org/10.48550/arXiv.1708.04896

Zhou, Z., Alabi, O., Wei, M., Vercauteren, T., Shi, M., 2023. Text promptable surgical instrument segmentation with vision-language models. In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023. NeurIPS. https://doi.org/10.48550/arXiv.2306.09244

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X., 2024. Vision mamba: efficient visual representation learning with bidirectional state space model. In: Forty-first International Conference on Machine Learning, ICML 2024. OpenReview.net. https://doi.org/10.48550/arXiv.2401.09417

Zhu, Y., Kiros, R., Zemel, R.S., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015. IEEE Computer Society, pp. 19–27. https://doi.org/10.1109/ICCV.2015.11