# CholecTriplet2022: Show me a tool and tell me the triplet - an endoscopic vision challenge for surgical action triplet detection

Chinedu Innocent Nwoye, Tong Yu, Saurav Sharma, Aditya Murali, Deepak Alapatt, Armine Vardazaryan, Kun Yuan, Jonas Hajek, Wolfgang Reiter, Amine Yamlahi, et al.

# CholecTriplet2022: Show me a tool and tell me the triplet

## *- an endoscopic vision challenge for surgical action triplet detection -*

Chinedu Innocent Nwoye [a,*], Tong Yu[a], Saurav Sharma[a], Aditya Murali[a], Deepak Alapatt[a], Armine Vardazaryan[a,u], Kun Yuan[a,k], Jonas Hajek[b], Wolfgang Reiter[b], Amine Yamlahi[c], Finn-Henri Smidt[c], Xiaoyang Zou[e], Guoyan Zheng[e], Bruno Oliveira[f,g,h], Helena R. Torres[f,g,h], Satoshi Kondo[i], Satoshi Kasai[j], Felix Holm[k], Ege Özsoy[k], Shuangchun Gui[l], Han Li[l], Sista Raviteja[m], Rachana Sathish[m], Pranav Poudel[p], Binod Bhattarai[n,o], Ziheng Wang[r], Guo Rui[r], Melanie Schellenberg[c,d], João L. Vilaça[f], Tobias Czempiel[k], Zhenkun Wang[l], Debdoot Sheet[m], Shrawan Kumar Thapa[q], Max Berniker[r], Patrick Godau[c,d], Pedro Morais[f], Sudarshan Regmi[q], Thuy Nuong Tran[c], Jaime Fonseca[h], Jan-Hinrich Nölke[c,d], Estevão Lima[g], Eduard Vazquez[p], Lena Maier-Hein[c], Nassir Navab[k], Pietro Mascagni[s], Barbara Seeliger[a,t,u], Cristians Gonzalez[t,u], Didier Mutter[t,u], Nicolas Padoy[a,u]

[a] *ICube, University of Strasbourg, CNRS, France*
[b] *Riwolink GmbH, Germany*
[c] *Division of Intelligent Medical Systems (IMSY), German Cancer Research Center (DKFZ), Heidelberg, Germany*
[d] *National Center for Tumor Diseases (NCT), Heidelberg, Germany*
[e] *Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, China*
[f] *2Ai School of Technology, IPCA, Barcelos, Portugal*
[g] *Life and Health Science Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal*
[h] *Algoritimi Center, School of Engineering, University of Minho, Guimeraes, Portugal*
[i] *Muroran Institute of Technology, Japan*
[j] *Niigata University of Health and Welfare, Japan*
[k] *Technical University Munich, Germany*
[l] *Southern University of Science and Technology, China*
[m] *Indian Institute of Technology Kharagpur, India*
[n] *University College London, UK*
[o] *University of Aberdeen, UK*
[p] *Redev Technology Ltd, UK*
[q] *Nepal Applied Mathematics and Informatics Institute for research (NAAMII), Nepal*
[r] *Intuitive Surgical, USA*
[s] *Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy*
[t] *University Hospital of Strasbourg, France*
[u] *IHU Strasbourg, France*

ABSTRACT

Formalizing surgical activities as triplets of the used instruments, actions performed, and target anatomies is becoming a gold standard approach for surgical activity modeling. The benefit is that this formalization helps to obtain a more detailed understanding of tool-tissue interaction which can be used to develop better Artificial Intelligence assistance for image-guided surgery. Earlier efforts and the CholecTriplet challenge introduced in 2021 have put together techniques aimed at recognizing these triplets from surgical footage. Estimating also the spatial locations of the triplets would offer a more precise intraoperative context-aware decision support for computer-assisted intervention. This paper presents the *CholecTriplet2022* challenge, which extends surgical action triplet modeling from recognition to detection. It includes weakly-supervised bounding box localization of every visible surgical instrument (or tool), as the key actors, and the modeling of each tool-activity in the form of ⟨*instrument, verb, target*⟩ triplet. The paper describes a baseline method and 10 new deep learning algorithms presented at the challenge to solve the task. It also provides thorough methodological comparisons of the methods, an in-depth analysis of the obtained results, their significance, and useful insights for future research directions and applications in surgery.

KEYWORDS: *Action detection, tool localization, fine-grained activity recognition, surgical action triplet, weak supervision, CholecT50, computer-assisted surgery.*

## 1. Introduction

Real-time video analysis will become an essential part of surgical intervention (Mascagni et al., 2022). It can offer context-aware intraoperative decision support to surgeons using AI computational models capable of extracting knowledge from live video data with reliability, accuracy, and speed (Vercauteren et al., 2019; Nwoye, 2021). The goal is to provide new intraoperative assistance thanks to the complementary analysis of the surgical workflow in terms of activity recognition (Twinanda et al., 2016; Lecuyer et al., 2020), tool (or instrument) detection and tracking (Jin et al., 2018a; Al Hajj et al., 2018; Nwoye et al., 2019), tissue/tumor segmentation and classification (Luengo et al., 2021; Wang et al., 2022; Maqbool et al., 2020), etc. Despite the tremendous progress made in the research community, activity modeling, such as phase recognition (Czempiel et al., 2020; Gao et al., 2021), is still coarse-grained and does not contain the details needed for highly adaptive AI assistance in the operating room (OR). The fine-grained counterparts, such as action recognition (Khatibi and Dezyani, 2020; Wagner et al., 2021), leave out the details about the anatomy operated on.

The quest for a better and more comprehensive activity modeling led to the formalization of surgical activities as triplets (Nwoye, 2021; Katic et al., 2014, 2015) taking into account the details of the operating instruments, manipulated tissues/targets, and verbs of action describing the interaction. Earlier work on surgical action triplet recognition (Nwoye et al., 2020, 2022b; Xi et al., 2022; Li et al., 2022; Cheng et al., 2022) including our previous CholecTriplet2021 challenge (Nwoye et al., 2022a) concentrate primarily on the presence detection of these triplets from laparoscopic videos. Localizing also the spatial regions of action triplets across video frames would advance research on tool-tissue interaction understanding and help develop assistance systems delivering valuable AI feedback and automated warnings.

This paper presents **CholecTriplet2022**[1], an endoscopic vision challenge organized at MICCAI 2022 for the detection of surgical action triplets in laparoscopic videos. This is an extension of the previous edition of the international competition on triplet presence detection at MICCAI 2021 with **spatial localization** of the instruments performing the actions (illustrated in Fig 1). The challenge was organized under the aegis of the Endoscopic Vision (EndoVis) grand-challenge (Zia et al., 2022) and presented at MICCAI 2022 in Singapore. The international challenge provided a scientific platform for the development of deep learning solutions for surgical action triplet detection in the OR. During the contest, registered participants were granted privileged access to a part of the CholecT50 dataset (Nwoye et al., 2022b) used as training data. The provided data labels are based on binary presence only, motivating the contestants to innovate by proposing alternative approaches, such as weak supervision, for modeling the localization aspect of the task. Meanwhile, some spatially annotated mini-samples of the dataset are provided for checks and validation. Useful code repositories, a validation system, a discussion forum, a submission server, support, and other helpful resources were also provided. A total of 11 teams participated in the challenge and presented several promising technologies using AI to replicate a detailed understanding of tool-tissue interactions that experts in minimally invasive surgery have acquired through many years of training.

In addition to these contributions brought by the event itself, the post-challenge report offers contributions on its own. Here, we study and present in this paper, a summary of the challenge activities, a theoretical description of all presented methods, an in-depth methodological analysis, and a method comparison. We benchmark the developed algorithms against their rivals and the baseline using the same criteria, and report the obtained results which are in the range of 18.8-35.0% for triplet recognition, 0.3-41.9% on instrument localization, and 0.08-4.49% on triplet detection using average precision metrics at a threshold of 0.5 IoU, which is very strict for weakly supervised methods. We also illustrate, using a rich selection of qualitative results, the behavior of proposed methods under various visual and procedural challenges associated with surgical video datasets. We highlight some strategic findings on the suitability of the observed computer vision techniques for surgical workflow activity modeling and discuss the significance and relevance of the presented results to the advancement of AI in surgery. We conclude with a survey polling the participants on their experiences and how to standardize and improve future events.

The rest of the paper is organized as follows: a review of related works in the next section helps to position our work in the research domain. This is followed by a summary of the challenge setup and activities. A detailed description, analysis, and comparison of the methods with the adapted evaluation protocols are presented in the subsequent sections. Afterwards, the findings are analyzed and discussed in terms of their benefits and limitations. And the study concludes by discussing the future potential of the work done.

## 2. Related work

The CholecTriplet challenge relates to several research topics, for which we present the relevant literature in this section.

### *2.1. Activity recognition*

Recognizing human activities from videos is a central task in computer vision, approached in different ways depending on the visual context - natural or medical images. Early methods from general computer vision employed architectures based only on 2D Convolutional neural networks (CNN): Karpathy et al. (2014) explored them in various pooling configurations, while Simonyan and Zisserman (2014) used a pair of 2D CNN to simultaneously exploit RGB data and optical flow. More advanced forms of temporal modeling for activity recognition followed, starting with the long-term recurrent convolutional networks (LRCN) from Donahue et al. (2015): this architecture incorporated a long short time memory (LSTM) recurrent neural network as a model for temporal dependencies between

*Corresponding author
[1] https://cholectriplet2022.grand-challenge.org

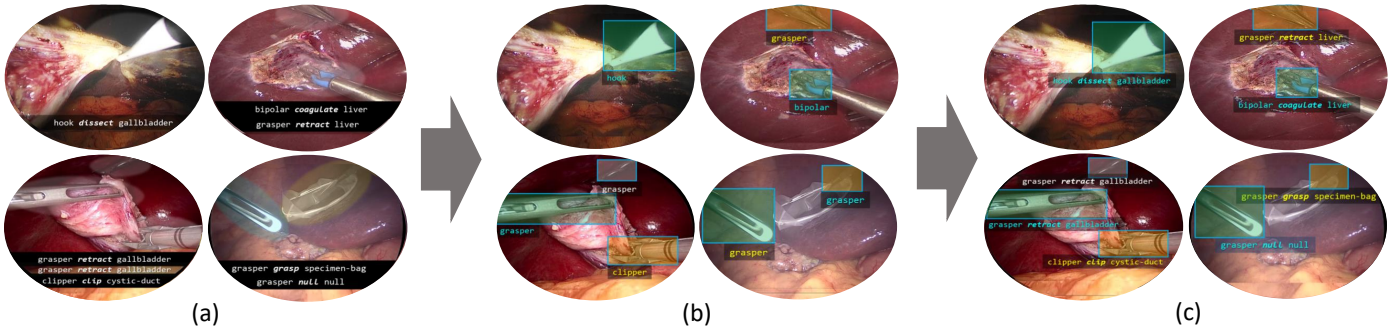


**Fig. 1. Illustration of the 3 sub-tasks of the CholecTriplet2022 challenge on the CholecT50 dataset: (a) Triplet recognition: triplet binary labels, (b) Instrument localization: instrument binary + instrument spatial labels, and (c) Triplet detection: triplet binary + instrument spatial labels. Represented surgical action triplets are illustrated for different time points during laparoscopic cholecystectomy.**

frames. 3D CNN (Tran et al., 2015; Carreira and Zisserman, 2017; Xie et al., 2018; Feichtenhofer et al., 2019) take a different approach by using spatio-temporal convolutions. Recent developments led to the video Transformers (Bertasius et al., 2021a; Liu et al., 2022b), built around spatio-temporal attention mechanisms.

In surgical computer vision, the most common activity recognition task is phase recognition. Several models were proposed for this task: Twinanda et al. (2016) used the same concept as the LRCN (Donahue et al., 2015), combining a ResNet-152 model with an LSTM on the Cholec80 dataset. Variations of the same architecture were employed in subsequent works on the same dataset (Funke et al., 2018; Yu et al., 2019; Jin et al., 2018b, 2021; Gao et al., 2021). Other types of temporal models were explored, such as temporal convolutional networks (TCN) by Czempiel et al. (2020), as well as Ramesh et al. (2021), who, in addition to phase, carried out the finer-grained task of step recognition. Most recently, the Transformer models (Czempiel et al., 2021) were explored for surgical activity recognition.

Overall, surgical computer vision has mostly focused on coarse-grained, long-range activity recognition tasks, leading to models that are different from those found in natural computer vision.

### *2.2. Action triplet: from recognition to detection*

Beyond coarse-grained activity recognition, activities that involve more than one component provide an exciting way to understand the complexity of an activity. Fine-grained activities are commonly denoted by a triplet of ⟨*subject, verb, object*⟩ where subject denotes the actor, verb the type of activity, and object the end target of the activity. To this end, Chao et al. (2015) analyzed a diverse set of human-object interaction instances. Mallya and Lazebnik (2016) first extracted human and object region of interest features from a CNN and applied Multi-Instance Learning (MIL) technique for activity prediction. Chao et al. (2018), on the other hand, modeled individual and pairwise associations of detected objects along with graph-based label correlation in a multi-stream architecture. Gkioxari et al. (2018) utilized FasterRCNN to detect humans and objects in a multi-task learning setup. Qi et al. (2018) adopted a graph structure with detected humans and objects as nodes in an iterative message-passing method. Zou et al. (2021) and Tamura et al. (2021) employed transformer-like architecture in an end-to-end learning approach with the help of matching loss and learnable queries. Zhang et al. (2022) proposed a two-stage unary and pairwise token-based transformer to analyze human-object interaction.

Action triplets in the context of surgical computer vision were introduced in Katic et al. (2014, 2015) as a piece of complementary information from surgical triplets to help with the main task of surgical phase recognition. Action triplet recognition as a primary task was first introduced in Nwoye et al. (2020) where they designed unary branches to predict triplet components: instrument, verb, and target, and a learnable 3D interaction space to model an association between the components. Later, Nwoye et al. (2022b) introduced an improved transformer-style model with an attention-based mechanism to further enhance triplet components association towards triplet recognition. Xu et al. (2021) employed a transformer model along with adversarial learning to generate captions, akin to triplets, depicting semantic relationships between components involved in a surgical scene. Lin et al. (2022) assigned instrument and target bounding boxes to triplet information and utilized a spatio-temporal graph for instrument-target interaction detection in cataract surgery.

### *2.3. Datasets: from recognition to detection*

Action datasets offer a large choice of tasks to be learned and performed by algorithms, with variations in granularity as well as the nature of the proposed task. From general computer vision, one early example of action classification datasets is PASCAL VOC (Everingham et al., 2005), with 10 action classes and static images only. To better capture actions, video datasets were released: UCF-101 (Soomro et al., 2012), followed by Charades (Sigurdsson et al., 2016) and Kinetics (Carreira and Zisserman, 2017). Another solution for capturing more information on actions is to add spatial detection: V-COCO (Gupta and Malik, 2015) extended the original COCO dataset (Lin et al., 2014) with bounding boxes around interacting elements. HICO-DET (Chao et al., 2015), centered around human-object interaction, offered similar annotations. Video datasets with spatial action detection, such as UCFSports (Soomro and Zamir, 2014) or AVA (Gu et al., 2018) exist as well.

Action recognition in the surgical domain mainly comes in

the form of surgical phase classification, as proposed by the Cholec80 dataset (Twinanda et al., 2016) dividing cholecystectomy into 7 phases. The CATARACTS dataset (Al Hajj et al., 2019) contained similar phase annotations for cataract surgery. Finer-grained descriptions of activities were featured in the Bypass40 dataset (Ramesh et al., 2021) via surgical steps. Additional information can also come in the form of spatial detections: Vardazaryan et al. (2018); Nwoye et al. (2019) used an extension of Cholec80 with instrument bounding boxes on the test set, for the evaluation of weakly supervised object detectors; however, those only located actions in an indirect manner. Actual action localization was offered by the SARAS-ESAD dataset (Bawa et al., 2021; Lin et al., 2022), with bounding boxes pointing to action verbs being performed.

Overall, action datasets have evolved towards more detailed and complex tasks, gradually adding more information in the form of interaction labels or bounding boxes. Our work continues in this direction by proposing a weakly supervised action localization task on the CholecT50 dataset (Nwoye et al., 2022b).

### 2.4. Benchmark challenge: from recognition to detection

International challenges have become a de facto standard for benchmarking image analysis algorithms (Wiesenfarth et al., 2021). They have a significant impact on the research community, oftentimes, encouraging a surge in new research directions owing to their characteristic creation of shared datasets, as well as incentives for best-performing algorithms. And more significantly, they offer a more reliable result benchmarking by withholding their test data from the public domain and comparing rival algorithms using the same criteria. The PASCAL VOC challenge (Everingham et al., 2005) and the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015) are among the earliest image analysis challenges featuring deep learning methods albeit focusing on simple classification. Subsequent challenges advance the recognition task to object detection (Lin et al., 2014; Everingham et al., 2010), single and multi-object tracking (Dendorfer et al., 2020; Chen et al., 2021; Kristan et al., 2016), segmentation (Cordts et al., 2016; Voigtlaender et al., 2019), etc., spearheading the development of state-of-the-art models. Repeating some challenges over the years comes with either or both increasing the size of datasets and task difficulty.

The first MICCAI grand challenge organized in 2007 in Brisbane brought similar benefits to biomedical image analysis. Notable in the surgical data science domain is the M2CAI 2016 challenge[2] featuring both surgical workflow analysis and tool presence detection for laparoscopic cholecystectomy leading to the creation of widely used datasets in the field: m2cai16-workflow (Stauder et al., 2016; Twinanda et al., 2016), m2cai16-tool and Cholec80 (Twinanda et al., 2016). Other datasets such as BraTS (Menze et al., 2014), CATARACTS (Al Hajj et al., 2019), ROBUST-MIS (Roß et al., 2021), SurgVisDom (Zia et al., 2021), MISAW (Huaulmé et al., 2021), HeiChole (Wagner et al., 2021), SARAS-ESAD (Bawa et al., 2021), Robotic Instrument Segmentation (Allan et al., 2019), CaDIS (Grammatikopoulou et al., 2019; Luengo et al., 2021), etc, are all products of biomedical challenges.

In 2021, we introduce the first endoscopic vision challenge focusing on the recognition of surgical activities in the form of triplets (Nwoye et al., 2022a). A rerun of this challenge in the current edition is with two notable advances: (1) the increment of the task difficulty from presence to spatial detection, which also increases the task usefulness, and (2) the addition of spatial bounding box labels to challenge evaluation data thereby enriching the dataset.

## 3. Challenge description

CholecTriplet is an endoscopic vision challenge that orchestrates research, development, and evaluation of AI methods for the automatic analysis of surgical video activities syntactically as ⟨*instrument, verb, target*⟩ triplets; also known as surgical action triplet (Nwoye et al., 2020). Its inception in 2021 under the code-name of *CholecTriplet2021* schemes to the presence detection (or simply, *recognition*) of action triplets directly from surgical videos. The current edition, *CholecTriplet2022*, organized by an 8-person committee from the Research Group CAMMA, advances the previous edition to also include the bounding box *localization* of regions of action triplet's likelihood in laparoscopic videos. We describe in detail the organization and execution of the CholecTriplet2022 challenge in this section.

### 3.1. Task

The primary goal is to develop machine learning methods for the *detection* of surgical action triplets directly from surgical videos. The term "detection" in this context translates to joint recognition and localization tasks. This added to the complexity of surgical action triplet modeling coupled with the required simultaneous identification of the correct components (*vis-à-vis*: instruments, verbs, targets), and resolving their association in the case of multi-instance triplets per frame.

The task is categorized into 3 sub-tasks as shown in Fig 1:

1. **Triplet recognition**: identification of the correct triplets, including their components, in every video frame,
2. **Spatial box localization**: estimation of the bounding box location of the principal actor (the instrument's tip) in every recognized triplet.
3. **Box-triplet association**: pairing of every localized instrument's bounding boxes to their corresponding triplets.

During the challenge, the three sub-tasks are jointly treated as a single task: a submission, comprising either a single model or collaborating multiple models, must produce the three outputs in a single docker run to be considered complete.

### 3.2. Challenge design

The CholecTriplet 2022 challenge is designed following the BIAS Reporting Guideline (Maier-Hein et al., 2020) for enhanced quality and transparency of biomedical research. The

[2] http://camma.u-strasbg.fr/m2cai2016/

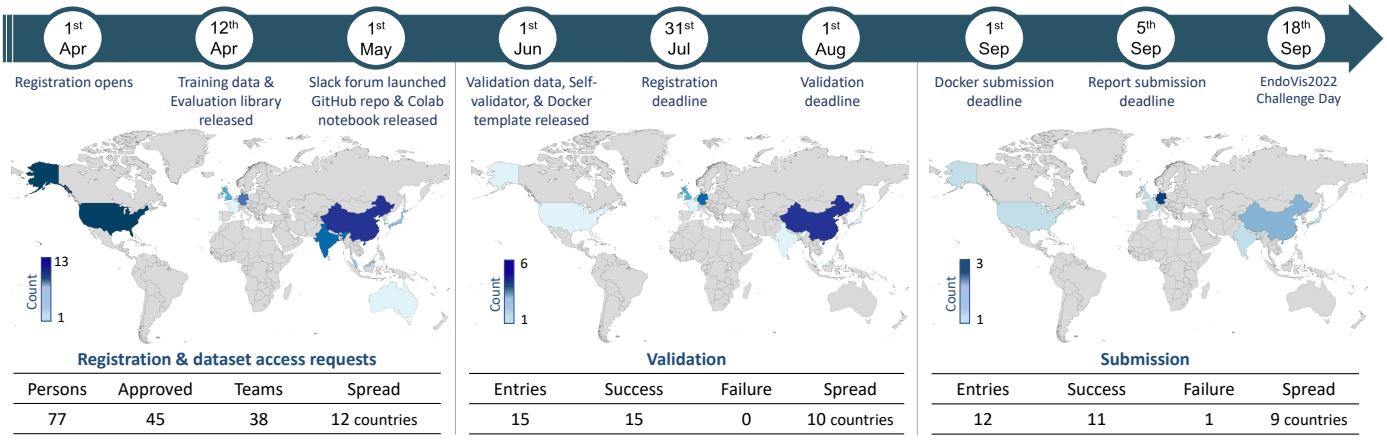


**Fig. 2. CholecTriplet2022 challenge timeline of activities and participation statistics.**

structured design is submitted as part of the Endoscopic Vision (EndoVis) grand challenge (Zia et al., 2022) in December 2021. The proposal was approved after two rounds of MICCAI review followed by a call for participation circulated online and offline.

The challenge was officially launched on April 1, 2022, and run through a 6-months window as shown in Fig. 2. This period is characterized by several activities such as the release of training data, customized metrics library, slack communication channel, a snippet of "getting started" code, GitHub repositories, etc., to guide and support the participants' method development. The participating teams develop their novel methods, fine-tune a state-of-the-art method, or improve on existing solutions during this period. The challenge timeline also involves a validation phase, harnessed by the use of a self-validation system, validation data samples, a Docker template, and guidelines provided to facilitate method submission. The whole process is concluded with the presentation of the method, results, and award winners at MICCAI 2022 conference in Singapore on Sept. 18, 2022.

### *3.3. Dataset and spatial annotation*

The challenge experiments are conducted on CholecT50 (Nwoye et al., 2022b), the largest endoscopic video dataset for surgical action triplet recognition. The dataset consists of 50 video footages of laparoscopic cholecystectomy that has been annotated with 100 distinct categories of surgical action triplets. At 1 frame per second (fps), a total of 100.9K frames of the dataset has been annotated with ~151$K$ triplet instances formed from 6 instrument, 10 verb, and 15 target categories. We follow the official data splits (Nwoye and Padoy, 2022) for challenge purposes (*aka* CholecT50-challenge version) which ensures that the test set is drawn from only the 5 videos that are not in the public domain for fairness in the competition. The rest of the 45 videos, *aka* CholecT45, are released to the participants for training their models. The CholecT45 provides only binary presence labels for the triplets, its individual components (instruments, verbs, targets), and the phase labels. Without spatial labels in the training data, the challenge allows for the modeling of the instrument's localization by weak supervision, thus, alleviating the cost and tedious annotation effort. While pretraining/transfer learning on third-party spatially annotated datasets is allowed, the downstream task involves learning the instrument localization and triplet-box association from imperfect annotations.

**Table 1. CholecT50 dataset configuration used in the challenge**

| | Training | Validation | Testing |
|---|---|---|---|
| # full videos | 45 | - | 5 |
| # short clips | - | 5 | - |
| # frames | 90.5K | 1.1K | 10.4K |
| # triplet classes | 100 | 100 | 100 |
| # instrument classes | 7 | 7 | 7 |
| # verb classes | 10 | 10 | 10 |
| # target classes | 15 | 15 | 15 |
| # phase classes | 7 | 7 | 7 |
| # triplet labels | 137.9K | 1.3K | 13.0K |
| # instrument labels | 137.9K | 1.3K | 13.0K |
| # verb labels | 137.9K | 1.3K | 13.0K |
| # target labels | 137.9K | 1.3K | 13.0K |
| # phase labels | 90.5K | 1.1K | 10.4K |
| # bounding boxes | - | 1.3K | 13.0K |

Training set = CholecT45. Validation set = subset(CholecT45 ∩ m2cai16-tool-location).

For evaluation, we annotate the test set with bounding boxes over the instruments using the IHU MoSaiC video annotation tool for surgical datasets. Conventionally, an instrument can be divided into two sub-parts: *effector*, which is the instrument tip responsible for primary action on anatomy, and *shaft* which connects the instrument tip with the instrument handle on its other end. As an annotation protocol, we consider only the effector part of the instrument and draw bounding boxes for instrument tips: (*cold*) grasper, bipolar (*grasper*), (*monopolar*) hook, (*monopolar*) scissors, clipper (short for *clip applier*), and irrigator (short for *suction/irrigation device*). Further, we match the bounding box with the corresponding triplet label and the annotations are stored in JSON format that includes both triplet binary presence labels as well as bounding box details. To give participants an insight into the testing labels, a mock-up validation set is generated. This involves 5 short video clips with triplet binary presence labels, instrument's bounding box labels, and box-triplet matching labels. The validation set's spatial annotation is outsourced from the overlapping videos of the m2cai16-tool-location dataset (Jin et al., 2018a) and merged

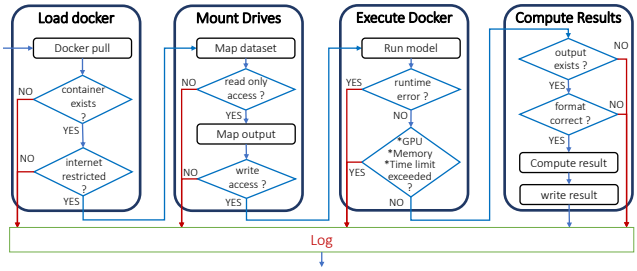


**Fig. 3. Flowchart of the self-validation system used in the challenge**

with the CholecT45 binary labels. The statistics of the entire dataset as used in the challenge are provided in Table 1.

To access the training and validation datasets, participants first register on the challenge website and sign a non-disclosure contract on the usage of the dataset. Afterward, participants are provided with a download link to the online repository containing the dataset. The test set will be made public at the discretion of the organizing lab.

### *3.4. Validation system*

In the validation step, participants were tasked with testing their docker containers using a validation script, to ensure that they properly ingested input data as well as adhered to the proper specifications for their model outputs. This validation script, implemented in Python, ensured that the participants' models respected the specified compute constraints, loaded and output data properly, and ran within the specified evaluation time and resource limits. The pipeline of the validation system is illustrated in Figure 3.

To aid participants, we split the validation phase into two parts: a local phase and a host phase. During the local validation phase, we gave the participants access to a properly formatted validation dataset which they could use as a reference to develop their models and docker container. Then, during the host testing phase, the participants were tasked with using the provided validation script to evaluate their finalized docker containers. This validator outputted a success report with performance on each test; participants were required to submit a successful validation report to the challenge submission portal before submitting their final docker container. The whole self-validation concept is to mimic the actual evaluation server but on the participant's personal computer (PC). The validation phase was highly effective, as 91.7% of the finally submitted docker containers ran directly during the submission phase as can be seen in Fig. 2; this is in contrast to the 2021 edition of the challenge, wherein 105 docker image submissions were required to successfully validate 24 teams. We attribute this difference to the inclusion of the local validation phase, which accelerated docker container development, as well as having provided the validation script directly to participants, which increased evaluation transparency.

### *3.5. Submission protocol*

Method submission is based on a Docker image uploaded via a dedicated challenge DockerHub. Private repositories are created for the teams with access granted by tokens valid until 5th September 2022. Smooth submission is enabled with a guideline published on the challenge GitHub page providing a Docker template with basic libraries, such as ivtmetrics, scikit-image, etc., required to run the model and evaluate their outputs. The template follows the same format used in the validation phase allowing participants to locally validate the Docker containers before pushing them to the submission portal. Submission update is possible only before the deadline. Along with the method Docker, a success self-validation log file of the final Docker, a draft summary report, and a PowerPoint/video presentation of the proposed method are required to complete a team submission. Afterward, all uploaded Docker containers are evaluated on the private test data (CholecT5), and results are collated for task benchmarking and team ranking. Only a single (last) submission is evaluated for each team.

### *3.6. Participation statistics*

During the call for participants, a total of 32 teams registered to participate in the challenge. Of these teams, a total of 11 teams progressed beyond the validation phase; as detailed in Fig 2, these teams were drawn from 9 different countries and 3 continents. This final participation was just over half of the CholecTriplet 2021 challenge (Nwoye et al., 2022a); we also observed a drop in the relative ratio between the final participants and total registrants (45.5% in 2021; 34.4% in 2022). We identified 2 likely reasons for the drop in each of these statistics: (1) the added difficulty of surgical triplet localization, which could have deterred registered teams from progressing to the validation phase, and (2) the use of almost the same dataset as in the 2021 edition, which may have reduced the perceived novelty of the challenge. The participating teams and their method are presented in Table 2.

### *3.7. Awards*

There are several IHU-sponsored monetary and certificate awards for the winners and runners-up of each task category. Additionally, there is an NVIDIA-sponsored GPU award for the winner of the main task (category 3). These awards are targeted at motivating and accelerating the research of the awardees even further.

## 4. Methodology

A summary of each participating team's proposed method is provided in Table 2; in the following subsections, we elaborate on these summaries, providing detailed descriptions of each method and a comparative analysis.

### *4.1. Analytical description of the conceptual frameworks*

#### *4.1.1. Rendezvous-Det (RDV-Det)*

This is proposed by the challenge organizers as a proof-of-concept and as a baseline to the competing methods. The RDV-det, shown in Fig. 4, is a detection version of the Rendezvous (Nwoye et al., 2022b) model popular for surgical action triplet recognition. Notably, the Transformer-inspired RDV is conceptually made up of 3 major components: (1) a weakly supervised

**Table 2. A cross-view of the methods presented by the teams at MICCAI EndoVis CholecTriplet 2022 challenge.**

| | Network | Method | Team | Affiliation(s) |
|---|---|---|---|---|
| 1 | AtomTKD | A multiple atomic tasks knowledge distillation framework for triplet recognition and detection | SHUANGCHUN | Southern University of Science and Technology, China |
| 2 | DATUM | Detection of action triplets using multi-graph networks | KLIV-IITKGP | Indian Institute of Technology Kharagpur, India |
| 3 | Distilled-Swin-YOLO | Self-distilled swin transformer ensemble | SDS-HD | German Cancer Research Center (DKFZ), Germany |
| 4 | DualMFFNet | Weakly-supervised surgical action triplet detection with dual multiplicative feature fusion networks | SK | Muroran Institute of Technology, Japan<br>Niigata University of Health and Welfare, Japan |
| 5 | EndoSurgTRD | Multi-task spatial-temporal triplet recognition with weakly-supervised tool detection | INTUITIVE-CORTEX-ML | Intuitive Surgical, USA |
| 6 | IF-Net | Instrument first: multi-instance instrument detection and instance-wise triplet classification | CAMP | Technical University Munich, Germany |
| 7 | MTTT | Multi-task triplet transformer for end-to-end surgical action triplet recognition | CITI | Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, China |
| 8 | ResNet-CAM-YOLOv5 | Combining ResNet class activation mapping with YOLOv5 for instrument detection and triplet recognition | WINTEGRAL | Riwolink GmbH, Germany |
| 9 | RDV-Det | Rendezvous-det: A baseline extension of the rendezvous network for surgical action triplet detection | CAMMA | Universite de Strasbourg, France |
| 10 | SurgNet | Surgical triplet recognition and detection using an ensemble of multi-task recurrent convolutional neural networks | 2AI-ICVS | Applied Artificial Intelligence Laboratory, Portugal |
| 11 | URN-Net | Mullti-task transformer with learnable orthogonal queries for triplet classification | URN | University College London, UK |

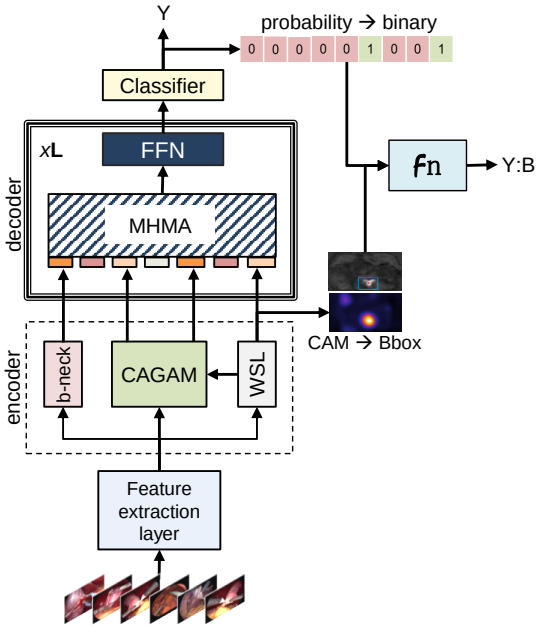


**Fig. 4. Architecture and data flow of the RDV-det model. It features an attention mechanism, multi-task learning, class activation modeling, weak supervision, and linear assignment.**

localization (WSL) layer for estimating the positions of surgical instruments in video frames, (2) a class activation guided attention module (CAGAM) for learning the verbs and targets leveraging an attention built from instruments activations, and (3) a multi-head of mixed attention (MHMA) for learning to resolve the instrument-verb-target relationships forming triplets. The MHMA is terminated by a linear classifier producing the triplet log probability scores, **log Y**. Without loss of generality, we ignore other modules of the RDV such as: the feature extraction base model, bottleneck layer, projection layer, etc., that are not of paramount interest in the discussion of RDV-det.

To advance this model for detection, the instruments are first localized via the WSL layer. Here, the interest is in the last 6-channel convolution layer designed to learn the location of each of the six distinct instrument categories in the CholecT50 dataset in the form of class activation maps (CAM). As a post-process, we extract bounding box coordinates, **B**, for every positive activation in the CAMs. Non-maximum suppression is applied to remove noisy labels. Afterward, triplet binary presence scores, **Y**, are obtained by thresholding the class probability vector of **log Y**. Finally, a heuristic-based data assignment module, **fn**, is employed to associate every extracted bounding box, $\mathbf{b} \in \mathbf{B}$, to the corresponding triplet instance $\mathbf{y} \in \mathbf{Y}$ within a given frame:

$$y : b \Leftarrow \mathbf{fn}(y_i, b_j, \phi) \quad \forall\, i = 1..|\mathbf{Y}|,\ j = 1..|\mathbf{B}|, \tag{1}$$

leveraging the instrument categories and triplet's detection confidence scores as the matching features, $\phi$, where $|\mathbf{Y}|$ and $|\mathbf{B}|$ are the cardinalities of Y and B respectively.

### *4.1.2. AtomTKD*

AtomTKD is a teacher-student approach to triplet recognition in which the student model, S-TRNet, is trained to directly predict triplets while being additionally supervised by a teacher model, T-AtomNet, trained to predict each individual (atomic) triplet component (instrument, verb, and target) in a multi-task fashion. Both S-TRNet and T-AtomNet consist of an ImageNet-pretrained ResNet18 backbone followed by a four-stage decoder consisting of a Transformer and convolutional layers (Dai et al., 2022); the respective loss functions are applied to the outputs of each stage, resulting in a hierarchical model structure. For the student model S-TRNet, the architecture is modified to use causal dilated convolutions, greatly reducing the number of model parameters. Both S-TRNet and T-AtomNet are trained with binary cross-entropy loss functions: on triplets for S-TRNet, and separately on instrument, verb, and

target for T-AtomNet. Meanwhile, for the teacher-driven supervision signal, the predicted triplet logits from S-TRNet are filtered to obtain separate instrument, verb, and target prediction logits (e.g. aggregating triplet logits across all verbs and targets yields instrument logits), which are supervised by the respective logits predicted by T-AtomNet using MSE loss functions.

For triplet localization, AtomTKD processes a sequence of frames with a ResNet50 backbone followed by a transformer encoder-decoder (Zou et al., 2021) to produce spatio-temporal embeddings; these embeddings are then processed with a three-layer perceptron to predict triplets as human-object interaction instances (instrument representing human, target representing object). This output is weakly supervised based on the ground-truth triplet labels, and at test-time, the bounding box predictions are extracted from the output.

#### *4.1.3. DualMFFNet*

DualMFFNet makes use of multiplicative feature fusion networks (MFF-Net) (Wei et al., 2021), which include a mechanism to aggregate features from different levels of the backbone, to predict each triplet component. For instrument and target prediction, two separate MFF-Nets with a ResNet50 backbone are trained, yielding instrument and target feature maps. To model tool-tissue interaction, the instrument and target feature maps are concatenated and processed with a 1D convolution, generating a verb feature map. Each feature map is passed through a GAP layer and a sigmoid activation, yielding instrument, verb, and target probabilities. Then, for triplet association, these probabilities are multiplied to generate triplet probabilities, which are then thresholded to obtain final triplet predictions.

For triplet localization, class activation maps are obtained from the instrument feature map using the corresponding instruments of the predicted triplets; the activation map is then thresholded to obtain the instrument bounding box.

#### *4.1.4. DATUM*

DATUM relies on a multi-head, multi-graph neural network (MGNet) as shown in Fig. 5. The MGNet is composed of several connected semantic attention modules (SAMs), dynamic graph neural networks (DGCNs) (Ye et al., 2020), and one classification head per triplet component plus a triplet head; as such, it is inherently multi-task. Features extracted from a ResNet-18 backbone (trained by self-supervised DINO approach (Caron et al., 2021)) are fed to the MGNet where they are processed by individual SAMs for instrument, verb, and target. The resulting class activation maps (CAMs) are then forwarded to three DGCNs with edges determined by co-occurrence relations. Features from these three DGCNs are subsequently pooled, concatenated and fed to the triplet DGCN, controlled this time by component-to-triplet mappings. Classification heads are positioned after each of the DGCNs to yield the final class probabilities. In this approach, each frame is processed with two passes through the MGNet: a first pass to obtain CAMs, and a second one feeding these maps back into the MGNet to refine the predictions. A mean square error loss ensures consistency between the two passes, in addition to cross-entropy losses for classification.

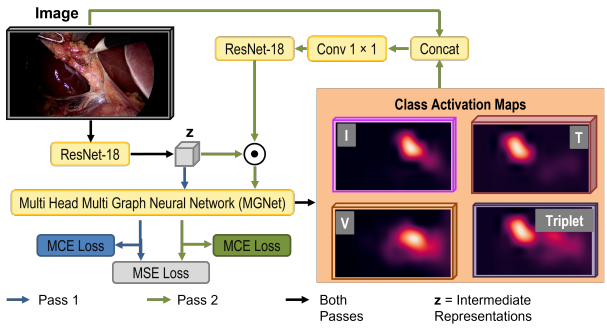


**Fig. 5. Architecture & data flow for the DATUM method featuring graphical modeling, multi-task learning, attention mechanism, class activation maps, and weak supervision.**

Triplet localization is obtained during inference from a series of post-processing steps: from the top 3 triplet predictions for a given frame, the corresponding instrument and triplet class activation maps are normalized and binarized using a heuristically determined threshold and then multiplied together. A bounding box is then fit to the resulting blobs.

#### *4.1.5. EndoSurgTRD*

EndoSurgTRD combines two models: a single-frame ResNet50 model for instrument classification and a spatio-temporal TimeSformer model (Bertasius et al., 2021b) for modeling the verbs and the targets. The instrument recognition/localization follows the Rendezvous' (Nwoye et al., 2022b) WSL approach albeit with a ResNet-50 backbone for feature extraction. The verb and target components are modeled using a modified TimeSformer which adapts the frame-based ViT (Dosovitskiy et al., 2020) to videos such that the inputs are processed as a sequence of patches extracted from the individual frames in the multi-image pathway. This modification extends the ViT's self-attention from its traditional image space to also include time 3D volume. Each of the triplet's component modules is terminated by a classification head. Their three output logits are in turn concatenated and passed to a final FC layer that predicts triplet logits. The complete model is trained end-to-end using the binary cross-entropy loss.

The localization part is achieved with an ad-hoc function that extracts bounding box coordinates for every positive class activation from the ResNet50 model. Bounding boxes with low class-wise probability scores and small bounding boxes are ignored in the process. This model, likely due to Docker error, produces the same output box coordinates for all input image, and hence excluded from the localization results.

#### *4.1.6. Distilled-Swin-YOLO*

Distilled-Swin-YOLO is a method that makes use of soft-labels for the recognition task and pseudo-labels for the localization task to tackle label uncertainty and overconfidence, which can be especially harmful when using ranking-based metrics like mAP. For triplet recognition, as depicted in Fig. 6, Distilled-Swin-YOLO uses an ensemble of three Swin transformers: a Swin-B model, a Swin-L model, and another Swin-B

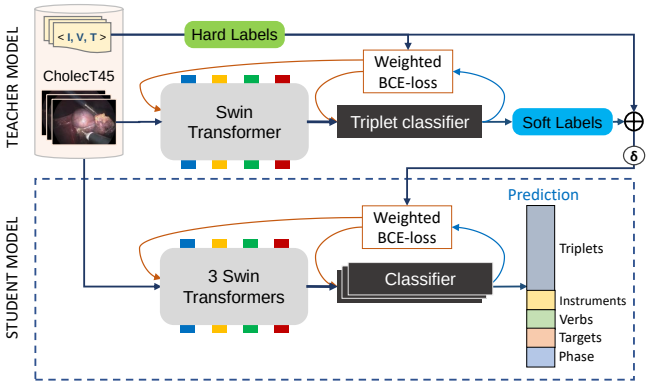


**Fig. 6. Architecture & data flow for the recognition part of Distilled-Swin-YOLO featuring knowledge distillation, transformer, transfer learning, and ensemble modeling.**

model trained in a multi-task fashion for instrument, verb, target, and phase recognitions. To train these models, a Swin-B teacher model is first trained and used to produce softened versions of the original labels, which are in turn used to train each of the 3 final models. Distilled-Swin-YOLO additionally uses a weighted binary cross-entropy loss during training to handle class imbalance.

For the triplet localization task, Distilled-Swin-YOLO uses a teacher-student approach to train an instrument detector, the output of which is heuristically combined with the triplet recognition output at test time. Initially, a YOLOv5 teacher model is pretrained on three external robotic instrument segmentation datasets (Allan et al., 2020, 2019; Bodenstedt et al., 2015) for which bounding box annotations are automatically generated from the original segmentation masks. This teacher model is then used to iteratively generate pseudo-labels on the challenge training set, which are in turn used to train a student model (also a YOLOv5 detector). For the pseudo-label generation, the teacher predictions are filtered based on the prediction confidence and following the ground truth instrument annotations.

#### *4.1.7. InstrumentFirst-Net (IF-Net)*

IF-Net decomposes the problem of action triplet detection into two sequential tasks: (1) instrument detection and (2) triplet classification, taking advantage of the simple constraint that each triplet is bound to a single instrument. The instrument detection model consists of a ResNet50 model for instrument detection, from which class activation maps are extracted and used to regress instrument bounding boxes following a percentile-based thresholding approach. In the case of multiple peaks in a single instrument's activation map, the activation map is split into multiple parts and multiple bounding boxes are regressed.

For triplet classification, IF-Net uses a modified ResNet50 model that takes as input the original image concatenated with an instrument class activation map and outputs a probability distribution of triplets. This probability distribution is then filtered based on the class of the predicted instrument (predicted probabilities for triplets with a different instrument are set to 0). The triplet classification model is run separately for each instrument detected by the first stage of IF-Net; thus, in the case of no detected instruments, the model is not run. IF-Net, therefore, models triplet classification as a multi-class problem rather than a multi-class multi-label problem, training the second stage to predict a single triplet. To obtain the final localized triplet result, the triplet predictions for each detected instrument are aggregated and associated with the previously computed localization results.

#### *4.1.8. Multi-Task Triplet Transformer (MTTT)*

MTTT is an end-to-end transformer trained in a multi-task fashion. It consists of a SwinV2-B (Liu et al., 2022a) feature extractor followed by 2 masked multi-head attention layers (Vaswani et al., 2017) to incorporate temporal context. These features are then passed to 4 different prediction branches (similar to SurgNet): instrument detection, verb detection, target detection, and phase detection. The output logits of each of these branches are then concatenated along with the original features and passed to a triplet prediction branch. Each branch consists of a fully connected layer and an activation function (sigmoid for instrument, verb, target, triplet; or softmax for phase).

For triplet localization, a CNN is applied to a high-resolution version of the input image. Following Nwoye et al. (2019), a modified ResNet-50 (He et al., 2016) (stride in layer 4 is set as one) is employed for feature extraction, followed by two convolutional Long Short-Term Memory (ConvLSTM) (Shi et al., 2015) layers for incorporating spatio-temporal context. Finally, the triplet localization is realized by class activation mapping (Shen et al., 2016).

#### *4.1.9. ResNet-CAM-YOLOv5*

ResNet-CAM-YOLOv5 combines class activation mapping with supervised pre-training on external datasets to obtain accurate instrument localization, then employs a multi-task approach for triplet recognition. Specifically, for triplet recognition, a multi-task ResNet50 model is trained on phase recognition, instrument classification, verb classification, and target classification, as in several other methods. In addition, rather than including a module to associate the predicted instrument, verb, and target into a predicted triplet, it includes another branch for direct triplet classification.

For triplet localization, a YOLOv5 instrument detector is first pre-trained on instrument bounding boxes from the CholecSeg8k (Hong et al., 2020) and HeiCo (Maier-Hein et al., 2021) datasets; then, to obtain a model that localizes instrument tips rather than the entire instrument, the YOLOv5 model is fine-tuned using the LapChole (Stauder et al., 2016) and Endovis Instrument (Bodenstedt et al., 2015) datasets. During inference, to predict the instrument class for each detection, the YOLO-detected bounding boxes are matched with instrument class activation maps ResNet using a threshold-based decision tree, which considers the scores of the recognized instrument classes, the scores of detected bounding boxes, and the mean of the class activation in the region of the detected bounding box. Meanwhile, for triplet classification, the highest-scoring triplet (obtained from the triplet classification branch) that includes the previously predicted instrument class is predicted as the triplet.

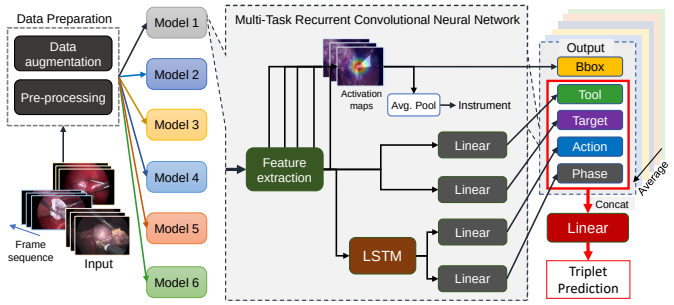


**Fig. 7. Architecture & data flow for the SurgNet method featuring temporal modeling, ensemble modeling, multi-task learning, and weak supervision.**

#### *4.1.10. SurgNet*

SurgNet is an ensemble of six deep models for surgical action triplet prediction. As shown in Fig. 7, each of the models consists of a feature extraction layer, built on one of the popular base architectures (ResNet50, ResNest50, ResNest101, SENet, EfficientNet-B0, EfficientNetB4), and a 5-heads multi-task layer for predicting five intermediary tasks (instrument presence, instrument bounding box location, targets, actions or verbs, phases). The instrument and target prediction heads are each modeled using a fully connected layer and a sigmoid activation function to enable their multi-label classifications. The phase and action prediction heads are each modeled using an LSTM unit, followed by an activation function (softmax for multi-class phases and sigmoid for multi-label actions) to compute their class probabilities. The instrument localization head builds a cascade of intermediary features from various stages of the backbone to capture the instruments' location in the form of heatmaps. The spatial dimension of the heatmaps is scaled to the input image size and their channels are mapped to the number of instrument classes. The output heatmaps are transformed to class presence probability scores using average pooling operation and supervised weakly on instruments' binary presence labels.

Each model in the ensemble is individually trained following a multi-task pipeline. An ensemble prediction per intermediary task is obtained by averaging each intermediary prediction across the six ensemble models. The ensembled instrument presence, targets, verbs, and phases predictions are concatenated to generate a feature vector with a size of 39 and remapped to the 100 triplet class probabilities using a fully connected layer. Meanwhile, the ensembled instrument bounding box location outputs are thresholded with a value of 0.95, after which the smallest enclosing box can be computed for each instrument.

#### *4.1.11. URN-Net*

URN-Net uses a ResNet-50 backbone and a Transformer decoder with learnable queries to recognize triplets. The ResNet50 extracts feature fed as input to the Transformer decoder from which it derives key and value features. The transformer decoder outputs processed queries that are used to directly predict each triplet component (instrument, verb, target). In addition, to prevent redundancies in the queries, an additional loss function is imposed on the final processed queries to ensure that they are mutually orthogonal. Lastly, to encourage temporal consistency, a temporal consistency loss is imposed during training by minimizing the cosine distance between the predicted triplet logits of two frames in a window of 10 frames.

URN-Net does not include a triplet localization component and is hence omitted from the localization results.

### *4.2. Theoretical comparative analysis*

We analyze and compare the computer vision technologies in the presented methods under 9 categories as follows:

#### *4.2.1. Multi-task learning (MTL)*

To learn relevant triplet component features, multi-task learning of the instruments, verbs, and targets is a common inclusion observed among the presented methods. This helps models to learn refined component-specific features before aggregating them for triplet prediction or to learn triplet features from shared features that are trained to be component-aware. MTL approach was highly popular among contestants: SurgNet, MTTT, EndoSurgTRD, DATUM, DualMFFNet, ResNet-CAM-YOLOv5, and AtomTKD associate features from triplet components to perform triplet recognition. URN-Net and Distilled-Swin-YOLO, on the other hand, utilize a shared feature extractor where they create separate branches for both triplet and its components. Moreover, due to correlations between surgical phases and actions, SurgNet, MTTT, Distilled-Swin-YOLO, and ResNet-CAM-YOLOv5 incorporate phase recognition along triplet components as an additional task.

#### *4.2.2. Temporal modeling*

The verb component in a surgical triplet captures the type of action performed by the instrument on the target. Although single frames are sometimes sufficient to identify actions, it is worth exploring how temporal models can help with more ambiguous triplets by using the context from past frames. Two frequent flavors of temporal models are featured in this challenge: Transformers and LSTMs, with variations on the triplet component of interest for the temporal head. SurgNet applies an LSTM only on phase and verb branches to learn temporal features; this choice is motivated by the interplay between phase, verb, and instrument motion patterns. MTTT employs ConvLSTM layers to refine bounding box locations from the noisy instrument class activation map (CAM). EndoSurgTRD adds a Transformer to model verb and target components separately from single-frame instrument features. Given the importance of past frame features to disambiguate triplet classes, for inference, URN-Net adds a discount factor for every past frame to weight their contribution for triplet recognition.

#### *4.2.3. Attention methods*

Attention-based methods are currently the state of the art on triplet recognition as shown by RDV model (Nwoye et al., 2022b). These methods learn attention maps that score the importance of a region to achieve high performance on a given task and serve as a powerful tool to explain model predictions. Aside from the RDV-Det, 6 methods utilize attention mechanisms,

with the Transformer being the primary model choice. Atom-TKD and URN-Net apply the Transformer for the recognition of triplets and their components whereas MTTT and Distilled-Swin-YOLO additionally use the Transformer to learn phase features. Moreover, the choice of where to place the Transformer varies across methods and depends on the type of triplet component, as EndoSurgTRD applies the Transformer on only the verb and target components. As a non-Transformer attention method, DATUM applies a semantic attention module to extract relevant regions of interest from scene features. Interestingly, for localization, none of the methods uses an attention mechanism.

#### *4.2.4. Knowledge distillation (KD)*

Distillation is the process of carrying over knowledge between a pair of models, with the source of knowledge commonly referred to as the "teacher" and the recipient as the "student". While oftentimes distillation serves as a model compression technique by having a lightweight model copy a heavier one, its purpose in the two entries utilizing it herein was to implement a form of soft label learning. Atom-TKD performed distillation between two models: T-AtomNet as the teacher, focusing on separate triplet components, and S-TRNet as the student, with more focus on the triplet as a whole. Distilled-Swin-YOLO used self-distillation: three models were involved in this process, one for each triplet component, each one acting both as the teacher and the student. In both cases, KD is performed through logits, via an MSE (Atom-TKD) or cross-entropy (Distilled-Swin-YOLO) loss, to handle uncertainty in predictions. KD is also used by Distilled-Swin-YOLO to train a YOLOv5 for instrument localization.

#### *4.2.5. Activation modeling*

Class activation map (CAM) is an effective way of obtaining the localization of scene objects using only image-level labels. CAMs highlight the discriminative regions in the image that can contribute to object localization. Furthermore, Nwoye et al. (2020, 2022b) demonstrate a useful property of CAMs: high activation regions correspond to the area around the instrument tip. This style of localization is helpful when instrument bounding box annotations are time-consuming and expensive. RDV-Det, SurgNet, IF-Net, EndoSurgTRD, and DualMFFNet integrate weakly supervised methods based on CAMs for localizing the instruments present in the surgical scene. Surgical triplets usually feature instrument movements over a certain time range, which MTTT exploits by utilizing ConvLSTM models on the CAM output to modulate features across time. Moreso, DualMFFNet combines instrument and triplet CAMs multiplicatively to locate triplets.

#### *4.2.6. Ensemble methods*

Methods that ensemble model predictions provide a comprehensive way to factor knowledge from multiple models, and limit noise in predictions. SurgNet and Distilled-Swin-YOLO employ ensemble with variations on the style of architecture to model triplet components. SurgNet creates an ensemble of large models such as ResNet50 and ResNest50/10 as well as parameter efficient models EfficientNetB0/B4 and SENet and averages the predictions for both the triplet recognition and localization tasks. Distilled-Swin-YOLO, on the other hand, utilizes both base and large variants of Swin-Transformer to train solely on triplet and an additional base variant of Swin-Transformer to train in a multi-task setting, involving triplet components.

#### *4.2.7. Transfer learning*

Transfer Learning is a de facto approach to utilize models trained on one task such as object detection for another related task such as instrument/triplet detection. Distilled-Swin-YOLO and ResNet-CAM-YOLOv5 train YOLOv5, a state-of-the-art object detector, on external laparoscopic datasets as well as on the challenge training data for triplet localization. This allows the model to exploit YOLOv5's capabilities to better detect instruments. A rudimentary form of transfer learning, known as pre-training, involves prior training of a model on another dataset, usually a larger one, to mitigate overfitting. Rather than actual model training, pre-training is most approached indirectly via full or partial weights initialization; especially for the backbone parameters. 72.7% models presented in this challenge are pretrained on ImageNet; Cholec80 and other endoscopic datasets make up the rest. Only one model is without pretraining.

#### *4.2.8. End-to-end vs separate training strategy*

Models employed for complex tasks such as the ones proposed in this challenge tend to include several parts, each part performing a different function. For such models, choosing a training sequence can be an issue: end-to-end training, with all parts trained simultaneously, or separate training. Note that in many situations, end-to-end training can be experimentally inconvenient, if not completely infeasible, especially with long videos such as those offered by CholecT50. Three entries have opted for end-to-end learning: MTTT (CNN + ConvLSTM - Vision Transformer + attention), DualMFFNet (2 CNN), and URN-Net (CNN + Transformer). MTTT in particular is a spatio-temporal combination, limiting the temporal range for end-to-end learning: in this case, clips of 5 frames were used.

#### *4.2.9. Levels of model supervision*

Alternatives to full supervision are highly sought after in surgical data science due to the heavy cost of manually labeled surgical data. In that spirit, no labels at all were provided for the localization task of this challenge, which left no possibility for true full supervision of challenge entrants. To solve this, weak supervision based on Class Activation Maps (CAM) was an overwhelmingly popular choice. The activation in CAMs manifests as a blob covering the entity to detect; further manipulations of the CAM, e.g. fitting a bounding box to the blob, enable spatial detection at no annotation cost since CAMs are a byproduct of training for binary presence detection. We observed that 7 methods used this type of CAM-based weak supervision: SurgNet, IF-Net, MTTT, EndoSurgTRD, DATUM, DualMFFNet, and RDV-Det. Two methods, however, opted for more conventional object detectors: Distilled-Swin-YOLO and ResNet-CAM-YOLOv5 both rely on YOLOv5, prefatorily trained using full spatial supervision from external datasets.

**Table 3. Implementation details across entrants.**

| Team | RDV-Det (*baseline*) | AtomTKD | DATUM | Distilled-Swin-YOLO | DualMFFNet | EndoSurgTRD | IF-Net | MTTT | ResNet CAM YOLOv5 | SurgNet | URN-Net |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Architecture | Rendezvous Transformer | CNN-TCN-Transformer | Dynamic GCN | Vision Transformer | CNN + Multiplicative Feature Fusion (MFF-Net) | Two-stream network (single-image & multi-image) | CNN + CNN | *Classification*: Vision Transformer *Localization*: CNN-ConvLSTM | CNN + object detector | Multi-task Recurrent Convolutional Network | CNN + Transformer |
| Backbone(s) | ResNet18 | ResNet18 ResNet50 | ResNet18 | Swin-Transformer-Base Swin-Transformer-Large | ResNet50 | ResNet50 TimeSFormer | ResNet50 | *Classification*: SwinV2-B *Localization*: ResNet-50 | ResNet50 YOLOv5 | EfficientNetB0 EfficientNetB4 SENet Resnet50 Resnest50 Resnest101 | ResNet50 |
| Multi-task | Instrument, Verb, Target, Triplet | Instrument, Verb, Target, Triplet | Instrument, Verb, Target, Triplet | Instrument, Verb, Target, Phase, Triplet | Instrument, Verb, Target, Triplet | Instrument, Verb, Target, Triplet | Instrument, Verb, Target, Triplet | Instrument, Verb, Target, Phase, Triplet, Instrument location | Instrument, Verb, Target, Phase, Triplet | Instrument, Verb, Target, Phase, Triplet | N/A |
| Temporal component | N/A | Conv-Transformer | N/A | N/A | N/A | TimeSFormer | N/A | *Classification*: Transformer *Localization*: ConvLSTM | N/A | LSTM | N/A |
| Attention | Transformer Multi-head of mixed attention | Transformer Multi-head of self-attention | Spatial attention | Transformer | Weakly supervised class activation map | Spatio-temporal attention | N/A | *Classification*: Self-attention | N/A | N/A | Transformer multi-head self-attention |
| Model size (# params) | 17.1M | 12.54M | *Training:* 30.75M. *Inference:* 19.57M | 88M | 28.3M | 145M | 2 x ~23M | *Classification*:103.8M *Localization*: 77.9M | *ResNet50*: 24M *YOLOv5*: 140M | *Backbones*: 382M | 32M |
| Output components | Probability vectors for instruments, verbs, targets, triplets. Class activation maps | Probability vectors for Instruments, verbs, targets, triplets | Probability vectors for instruments, verbs, targets, triplets, phases. Class activation maps | Probability vector for triplets | Probability vectors for instruments, verbs, targets, triplets Class activation maps | Probability vectors for instruments, verbs, targets, triplets. Class activation maps | Probability vectors for instruments, verbs, targets, triplets. Class activation maps | Probability vectors for instruments, verbs, targets, triplets. Class activation maps | Probability vector for triplets. Class activation maps. Bounding boxes | Probability vectors for verbs, instruments, targets, triplets. Class activation maps | Probability vectors for instruments, verbs, targets, triplets. Query instances |
| Data preprocessing | Normalization, 256x448 image resize | None | Normalization, 240x427 image resize | 224x224 image resize | Normalization, 416x240 image resize | Normalization, 256x448 image resize | Normalization, 256x448 image resize | Black-margin cropping | Normalization, Image resize - *ResNet50*: 480x854 *YOLOv5* : 640x640 | Normalization, 250×250 image resize | None |
| Data augmentation | Random flipping, color jitter | Random flip, rotation, adjust sharpness, color jitter, auto contrast | Random contrast, brightness, rotation (± 10°), sun flare, blur, scale (± 0.2) | Rotation, flip, hue-saturation-value | Horizontal flip, scale, rotate, color jitter, shift, blur, cutout | Random horizon flip, vertical flip, color jitter, gaussian blur | PyTorch trivial augment wide with vertical flip, horizontal flip | Random flipping, rotation, color jitter | Rotation, scale, color jitter | Random 224x224 crop, horizontal flip; color jitter; 10° rotation | Vertical flip, horizontal flip; contract; 90° rotation |
| Output Post-processing | Bounding box extraction from CAM activations | None | Normalizing, thresholding and bounding box generation from triplet and instrument CAMs. | Downscale bounding box by 1%. Select triplets based on the instrument detection output | Bounding box extraction from CAM activations | Bounding box extraction from CAM activations | Triplet filtering by tool prediction. Percentile-based thresholding of CAMs to extract bounding boxes | None | Analytical mapping of instrument & triplet to instrument bounding box via CAM | Temporal smoothing: smooth variation according to previous predictions | Temporal smoothing |
| Levels of supervision | Full supervision for classification tasks, weak supervision for localization task | Full supervision for classification tasks, weak supervision for localization task | Self supervision for weight initialization, full supervision for classification tasks, weak supervision for localization task | Full supervision for classification tasks, Cross supervision from external data for localization task | Full supervision for classification tasks, weak supervision for localization task | Full supervision for classification tasks, weak supervision for localization task | Full supervision for classification tasks, weak supervision for localization task | Full supervision for classification tasks, weak supervision for localization task | Full supervision for classification tasks, Cross supervision from external data for localization task | Full supervision for classification tasks, weak supervision for localization task | Self and Full supervision for classification tasks |
| Loss function | Weighted Binary Cross Entropy (BCE) | BCE | Modified CE | Weighted BCE | CE | BCE | Categorical CE, BCE | Weighted BCE | BCE, Complete Intersection over Union (CIoU) | Categorical CE, BCE | BCE, CE, Cosine embedding loss |
| Optimizer | SGD (with momentum 0.9) | SGD | Adam | Adam | Adam | Adam (momentum=0.9) | AdamW | AdamW | Adam, SGD (momentum 0.937) | SGD (momentum 0.9) | AdamW |
| Learning rate | 1e-3 | 1e-2 | 1e-6 (backbone) 5e-5 (other) | 2e-4 | 1e-5 | 1e-2 | 1e-3 | 1e-5 | *YOLOv5:* 1e-2 *Others:* 1e-4 | 5e-5 | 1e-5(backbone) 1e-4 (other) |
| Learning rate schedule | Linear (warm up 8) + Exponential | Linear + Exponential | Constant | Cosine annealing | Cosine annealing | Cosine annealing, multi-step learner, linear (warm up 5) | Constant | Cosine scheduling | One cycle (1e-4) | Step : 5 step size, 0.1 gamma | Linear |
| Weight decay | L2 norm (0.01) | N/A | N/A | L2 penalty (1e-6) | L2 norm (1e-5) | N/A | L2 norm (1e-2) | L2 norm (0.01) | L2 norm (YOLOv5: 5e-4, *Others:* 1e-4) | L2 norm (5e-4) | L2 norm (1e-4) |
| Model weight initialization / pre-training | ImageNet | ImageNet | SSL on training data using DINO | ImageNet | ImageNet | *ResNet50*: ImageNet *TimeSformer*: *google/vit-base-patch16-224* | ImageNet | ImageNet, Cholec80 | ImageNet | ImageNet, Cholec80 | ImageNet |
| Use external data | None | None | None | Cholec80, Robotic Instrument Segmentation, Robotic Scene Segmentation, EndoVis Instrument | None | None | None | Cholec80 | COCO, LapChole, CholecSeg8k, HeiCo, Endovis Instrument | Cholec80 | None |
| Epochs | 120 | 200 (backbone) 2000(T-AtomNet) 1000 (S-TRNet) | 25 | 30-40 | 60 | 30 | 100 (max) | 30 | *YOLOv5:* 46 *Others:* 16 | 15 | 25 |
| Batch size | 32 | 64 (backbone) 256 (T-AtomNet) 1 (S-TRNet) | 64 | 64 | 384 | 4 | 32 | 12 | *YOLOv5:* 16 *Others:* 8 | 400 | 128 |
| GPU resources | 1x NVIDIA P4 | 1x NVIDIA Tesla V100 32GB | 2x NVIDIA Tesla V100 SXM2 16GB HBM2 NVLink | 2x NVIDIA RTX 3090 + 1x NVIDIA V100 32GB | 3x NVIDIA RTX 3090 24GB | 3x NVIDIA GeForce RTX 2080 | 1x NVIDIA A40 | 1x NVIDIA TITAN RTX | 1 x NVIDIA Tesla V100 32GB | 4 A100 NVIDIA | 1x V100 NVIDIA |
| Training time (Hours) | 36 | N/A | 8 | 100 | 6 | 7 | 26 | 20 | *ResNet50:* 14 *YOLOv5*: 20 | 80 | 11 |

## 5. Experiment and evaluation setup

### 5.1. Implementation details

The implementation details for the reproducibility of all submissions are summarized in Table 3.

### 5.2. Evaluation protocol

All presented models are trained on the publicly released part of the CholecT50 dataset otherwise known as CholecT45. Evaluations are conducted on the hidden 5 video test set. At the evaluation, submitted Docker containers are executed on the challenge test set producing a JSON file per video containing the outputs for the three sub-tasks. Triplet recognition evaluation is based on the $Y_{IVT}$ probability scores for valid 94 out of 100 triplet classes ignoring the last 6 labels with at least one NULL component in the triplet composition, as done in the previous challenge (Nwoye et al., 2022a). Model's direct outputs for any component of the triplet are not considered, instead, the predictions for instruments, verbs, and targets are filtered from $Y_{IVT}$ using Nwoye et al. (2022b)'s disentanglement function for uniformity across all architectural designs. Detection evaluation for instruments and triplets is based on the predicted bounding box coordinates and class identities. For each metric, we use video-specific averaging to obtain the mean score: per-category metric scores are averaged across videos, before computing the mean across categories. Averaging ignores unrepresented categories. All the performance scores are computed using `ivtmetrics` (Nwoye and Padoy, 2022)[3], which is a dedicated metrics library for surgical action triplet evaluation.

### 5.3. Evaluation metrics

The performance of the challenge models is assessed using the average precision (AP) metric. The AP metric is computed from precision (p) and recall (r) scores which are estimated from the true positive (TP), false positive (FP), and false negative (FN) scores as follows:

$$p = \frac{TP}{TP + FP}, \qquad r = \frac{TP}{TP + FN} \tag{2}$$

The AP metric is used for both the recognition and detection task categories albeit computed differently. For the classification, the TP, FP, and FN are computed by thresholding the predicted probability scores at a range of thresholds and the AP is computed as the area under the precision-recall curve. Following the established protocol Nwoye et al. (2020), we measure the AP of the triplet components recognition: $AP_I$, $AP_V$, $AP_T$, and of the triplet associations: $AP_{IV}$, $AP_{IT}$, $AP_{IVT}$. The $AP_{IVT}$ which evaluates the complete instrument-verb-target combination remains the main metric for this category.

For the detection, the TP, FP, and FN are computed by measuring the level of overlap between the groundtruth and predicted bounding boxes. A detection is assigned a TP if the degree of overlap exceeds a certain threshold and the ID (i.e. instrument ID for category 2 or triplet ID for category 3) is correct. The corresponding AP is computed as a weighted mean of the precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight:

$$AP = \int_0^1 p(r)dr. \tag{3}$$

We also utilize the triplet association scores (TAS) introduced in Nwoye and Padoy (2022) to assess the capacity of the models at associating the bounding boxes to the correct triplet IDs. The TAS metric, presented in Table 4 offers an in-depth analysis of the task 3 category by evaluating the detection from different perspectives.

We equally analyze the quality of the model predictions by computing the top$K$ accuracy where $K \in [5, 10, 15, 20]$ and the average over top$K$@[5:20] at intervals of 5 steps as done in Nwoye et al. (2022a).

**Table 4. TAS: Triplet association metrics**

| Metric | Name | Description |
|---|---|---|
| LM | Localize & Match | Percentage of tools localized at a given IoU and matched with correct triplet IDs. |
| *p*LM | Partial Localize & Match | Percentage of tools localized below the a given IoU but matched with correct triplet IDs. |
| IDS | Identity Switch | Percentage of tools localized at a given IoU but IDs are swapped within the frame. |
| IDM | Identity Missed | Percentage of tools localized at a given IoU but IDs are missed entirely. |
| MIL | Missed Localization | Percentage of correctly recognized triplet IDs without localization. |
| RFP | Remaining False Positives | Percentage of false alarms after other TAS metrics have been considered. |
| RFN | Remaining False Negatives | Percentage of missed predictions after other TAS metrics have been considered. |

## 6. Results and discussion

### 6.1. Triplet recognition

We present a brief overview of the AP scores for both triplet component recognition and component association in Table 5. Expectedly, the instrument is the most accurately recognized component with all the teams scoring over 70% AP and half the participating teams scoring above 80%. For the verb component, performance ranges between ~42 – 52% AP with a mean of 47.3 ± 5.4%. As with the previous edition of the challenge (Nwoye et al., 2022a), accurate recognition of the target being acted upon remains an open problem. Here, we see a relatively wide range of performances between ~26 – 46% with a majority of methods recognizing the target with an AP of < 40%. However, we also do note that there appears to be a strong correlation between recognition of the overall triplet and recognition of the target (R=0.93). Further, the fact that the triplet recognition performance here appears to be more strongly linked with the target performance than the instrument (R=0.74) or verb (R=0.84) performance may be indicative of a bottleneck that target recognition brings and should inform future research. Finally, we note that the top 3 methods on the triplet recognition task also attempt to learn phase features, potentially demonstrating how coarser-grained information could

[3] https://pypi.org/project/ivtmetrics

**Table 5. Performance summary of the presented methods for the three tasks.**

| Model | Triplet recognition | | | | Instrument localization ($\theta$ = 0.5) | | Triplet detection ($\theta$ = 0.5) | |
|---|---|---|---|---|---|---|---|---|
| | $AP_I$ | $AP_V$ | $AP_T$ | $AP_{IVT}$ | $AP_I$ | $AR_I$ | $AP_{IVT}$ | $AR_{IVT}$ |
| ResNet-CAM-YOLOv5 | 80.3 | 50.3 | 38.4 | 29.0 | **41.9** | **49.3** | **4.49** | **7.87** |
| Distilled-Swin-YOLO | 83.8 | 52.0 | **45.9** | **35.0** | 17.3 | 30.4 | 2.74 | 6.16 |
| MTTT | 84.0 | 49.5 | 40.3 | 34.5 | 11.0 | 21.1 | 1.47 | 3.65 |
| DualMFFNet | 64.9 | 40.4 | 29.9 | 20.0 | 04.6 | 06.6 | 0.36 | 0.73 |
| RDV-Det ‡ | 78.2 | 46.6 | 35.9 | 29.0 | 03.0 | 07.6 | 0.24 | 0.79 |
| IF-Net | 72.0 | 42.3 | 30.4 | 23.6 | 00.7 | 03.6 | 0.22 | 0.92 |
| AtomTKD | **85.7** | 52.3 | 39.2 | 27.2 | 00.9 | 02.4 | 0.15 | 0.32 |
| SurgNet | 78.8 | **55.7** | 40.3 | 34.4 | 10.8 | 19.6 | 0.13 | 0.39 |
| DATUM | 66.2 | 39.7 | 32.3 | 23.1 | 00.3 | 03.2 | 0.08 | 0.48 |
| URN-Net | 82.4 | 49.3 | 38.1 | 27.3 | - - | - - | - - | - - |
| EndoSurgTRD | 73.6 | 42.1 | 26.3 | 18.8 | - - | - - | - - | - - |
| Average ± standard deviation (stdev) | 77.3 ± 7.2 | 47.3 ± 5.4 | 36.1 ± 5.8 | 27.4 ± 5.7 | 10.1 ± 12.5 | 16.0 ± 15.0 | 1.1 ± 1.4 | 2.4 ± 2.7 |

**bold** = best score. underlined = second best. ‡ organizers' baselines.

inform fine-grained tasks such as triplet recognition. Operative phases are defined through the different steps/actions surgeons perform to complete a procedure. For each phase, defined by a specific procedural task, there is a characteristic set of triplets used. The correlation could ensures that the likelihood of detecting a specific triplet will increase in associated phases and thus, conditions the model on a subset of probable triplets rather than on a whole using phase features. Additionally, leveraging phase information could also help in discriminating closely related triplets with imperceptible change in roles which are usually tough to distinguish visually, e.g. ⟨*grasper, retract, gallbladder*⟩ occurring at the *calot triangle dissection* phase and ⟨*grasper, pack, gallbladder*⟩ during *gallbladder packaging*.

In terms of association for the different components, we see a comparable performance in associating the instrument with each of the two other components. Overall triplet recognition performance maxes out at 35.0% by the Distilled-Swin-YOLO model.

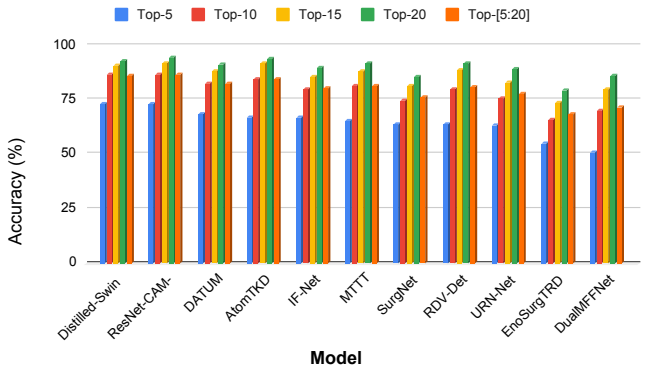


**Fig. 8. Action triplet recognition top-K accuracy.**

Further, using Top K accuracy, in Fig. 8, we also measure the ability of models to predict triplets within their K highest-scoring outputs. This allows us to consider the difficulty of the problem introduced by the high semantic overlap between certain classes (eg. sharing multiple components). Unsurprisingly, we see that the Distilled-Swin-YOLO model also comes out on top with top 5 and top 10 accuracies of 73.5 and 87.9, respectively. The gains in the performance are for two reasons - first, the choice of model, a Swin-Transformer that generates meaningful hierarchical features, and second, the use of an ensemble that suppresses noise in the predictions. Interestingly, the ResNet-CAM-YOLOv5 model, which places fourth for triplet recognition comes in joint-highest in the top 5 and top {5:20} performances, respectively. The higher topK accuracy compared to the AP scores show that a reduced number of triplet classes might theoretically be easier for a deep learning model to train. This would imply super-classifying related classes, resulting in more data to train each class. It would also increase inter-class variability and help to balance the loss optimization strategy affected by intra-class conflicts. However, a larger number of triplet categories increases the granularity and becomes more beneficial and clinically relevant.

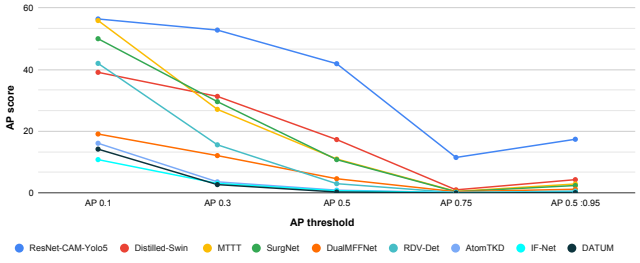


**Fig. 9. Surgical instrument localization mAP**

## 6.2. *Instrument localization*

Fig. 9 shows the instrument localization performance for each competing method, evaluated using the AP metric at various IoU thresholds. Immediately, we observe a large contrast in performance between the top-performing ResNet-CAM-YOLOv5 and the rest of the models; this can be attributed to the fact that the ResNet-CAM-YOLOv5 includes components that are pretrained on the CholecSeg8k, HeiCo, M2CAI, and EndoVisSub datasets for instrument localization, in contrast to the other methods, which rely purely on weak supervision signals. This is further substantiated when studying the $AP_{0.1}$ results, where a couple of other methods (MTTT, SurgNet), are competitive with the ResNet-CAM-YOLOv5; this indicates that these models are able to achieve a rough localization of the instruments, but lack precision and are therefore harshly penalized by increasing the IoU threshold for the AP computation.

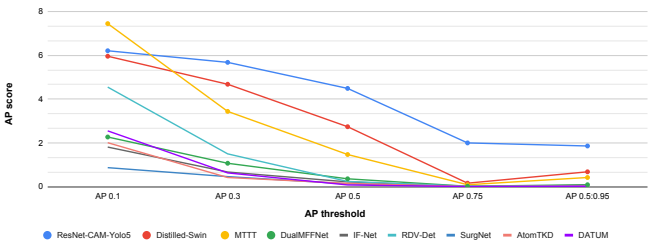


**Fig. 10. Surgical action triplet detection mAP**

### *6.3. Triplet detection (box-triplet matching)*

The triplet detection assessment evaluates joint performance on triplet presence recognition and instrument localization. It is also affected by the correct association of the triplet IDs to their corresponding instrument bounding boxes. The detection AP is presented in Fig. 10. Across different IoU, the models' ranking is stable, however, the performances decrease with a stricter overlapping threshold. It is observed that correct instrument localization plays a crucial role in this metric for the top 3 models (ResNet-CAM-YOLOv5, Distilled-Swin-YOLO, MITT) when compared with Fig. 9, though without discounting the associating part which could account for the altering of the model ranking on this metric after the top 3.

**Table 6. Bbox-triplet association performance**

| Model | LM ↑ | *p*LM | IDS ↓ | IDM ↓ | MIL ↓ | RFP ↓ | RFN ↓ |
|---|---|---|---|---|---|---|---|
| ResNet-CAM-YOLOv5 | **23.9** | 8.2 | 0.9 | 0.1 | **0.1** | **3.3** | 63.5 |
| Distilled-Swin-YOLO | 12.0 | 11.3 | 0.3 | **0.0** | 0.3 | 33.0 | 43.1 |
| MTTT | 8.6 | 25.4 | 0.1 | 0.1 | 0.4 | 11.6 | 53.8 |
| RDV-Det ‡ | 3.3 | 29.0 | **0.0** | **0.0** | 1.1 | 5.4 | 61.1 |
| DualMFFNet | 3.0 | 17.4 | 0.1 | **0.0** | 1.2 | 3.5 | 74.8 |
| IF-Net | 2.0 | 14.9 | 0.1 | 0.7 | 1.5 | 35.9 | 45.1 |
| SurgNet | 2.0 | 8.5 | **0.0** | **0.0** | 1.6 | 18.9 | 69.0 |
| DATUM | 0.4 | 22.8 | **0.0** | **0.0** | 1.3 | 45.1 | **30.5** |
| AtomTKD | 0.1 | 14.9 | **0.0** | **0.0** | 3.1 | 28.4 | 53.6 |

**bold** = best score. underlined = second best. ‡ organizers' baselines.

We, hence, quantify the association performance at varying localization conditions as presented in Table 6 using the triplet association score (TAS) metrics described in Section 5.3. We observe that while the association scores when keeping only the well-localized bounding boxes (LM) are low, they become higher when considering also the partially localized boxes (pLM).

The TAS metrics also show that identity switches (IDS) and identity misses (IDM) are minimal for all the methods. This argument is further strengthened by low MIL scores showing that the number of correctly recognized triplets that are not localized at any overlapping degree is negligible. Hence, the bottleneck lies in obtaining substantial localization overlap by a weakly-supervised method. The remaining FP is relatively low for 3 models (ResNet-CAM-YOLOv5, RDV-Det, DualMFFNet) whereas the remaining FN is high for all the methods. Both false negatives and false positives account for ~65% of the triplet association strength of the models. The FN/FP cases are frequently observed among semantically similar triplets such as ⟨*grasper, grasp, gallbladder*⟩ in place of ⟨*grasper, retract, gallbladder*⟩, ⟨*scissors, cut, cystic-artery*⟩ in place of ⟨*scissors, cut, blood-vessel*⟩. Discriminating such triplets could be a potential area of improvement for future works.

### *6.4. Qualitative results*

We also analyze a rich set of qualitative results from the models on different surgical situations, vis-à-vis, instrument usage patterns (Fig. 11) and visual challenges (Fig. 12). In both figures, varying surgical conditions are presented in each column, the groundtruths are on the first rows whereas predictions from 3 randomly selected models are on the other rows. Instrument localization is represented with bounding boxes of different colors for different triplet classes.

It is observed that more common triplets (Fig. 11a) in a clear scene are better detected than rarer triplets (Fig. 11b) that might occur only once and during a short moment in a procedure. Predicting the correct category is oftentimes an issue with rare triplets. Another observation is that a miss is usually the case when only a tiny portion of the instrument is visible (Fig. 11c), notwithstanding, the Distilled-Swin-YOLO model appears robust to this case. A crowded scene (Fig. 11d) presents a combination of missed prediction (e.g. ResNet-CAM-YOLOv5) and overlapping localization of multiple instruments as one (e.g. SurgNet). Even with correct localization, the triplet identities are oftentimes switched or mismatched (e.g. Distilled-Swin-YOLO). This could become an issue in certain surgical procedures, e.g. gastric bypass, requiring up to 6 trocar ports or when more than 2 instruments are visible in each frame, such as during retraction for suturing or knot-tying, stitching, etc. Occlusion occurs when instruments seem to overlap in a 2D view, the instrument closest to the camera (partially) obstructs the view of the instrument at the back. Under this situation (Fig. 11e), the occluded instrument is either localized as part of the occluding instrument (e.g. MTTT, ResNet-CAM-YOLOv5) or missed entirely (e.g. DATUM).

We observe quite interesting behaviors of the models under challenging visual conditions (Fig. 12). Bleeding in surgical procedures can cause blood stains on the instruments or camera lens. Under this condition (Fig. 12a), model localizing capacity is challenged: most models would localize just the non-stained part of the instruments (e.g. IF-Net, RDV-Det). The view (or image quality) can also be impaired by rapid motion of camera and instruments, or lack of focus adjustment (Fig. 12b). This leads to imprecise localization which could be attributed to the blurred instrument boundaries (e.g. DualMFFNet). In the cleaning and coagulation phase, instruments, such as the irrigator, are immersed in fluid/water (Fig. 12c). We observed that this leads to high false negatives for all the considered models. Specular reflection from endoscopic light can change the appearance of the surgical scene as seen in Fig. 12d. These light reflections outshine the surgical scene and thus affects the models in both the localization of obvious instruments like grasper and hook (e.g. AtomTKD, DATUM, MTTT) and their triplet classification. Localizing the same instrument with multiple bounding boxes is also observed (e.g. MTTT). Sometimes, when there is smoke generated through use of coagulation instruments, the effect ranges from minimal to adverse depending on the thickness of the smoke. In Fig. 12e, it can be observed

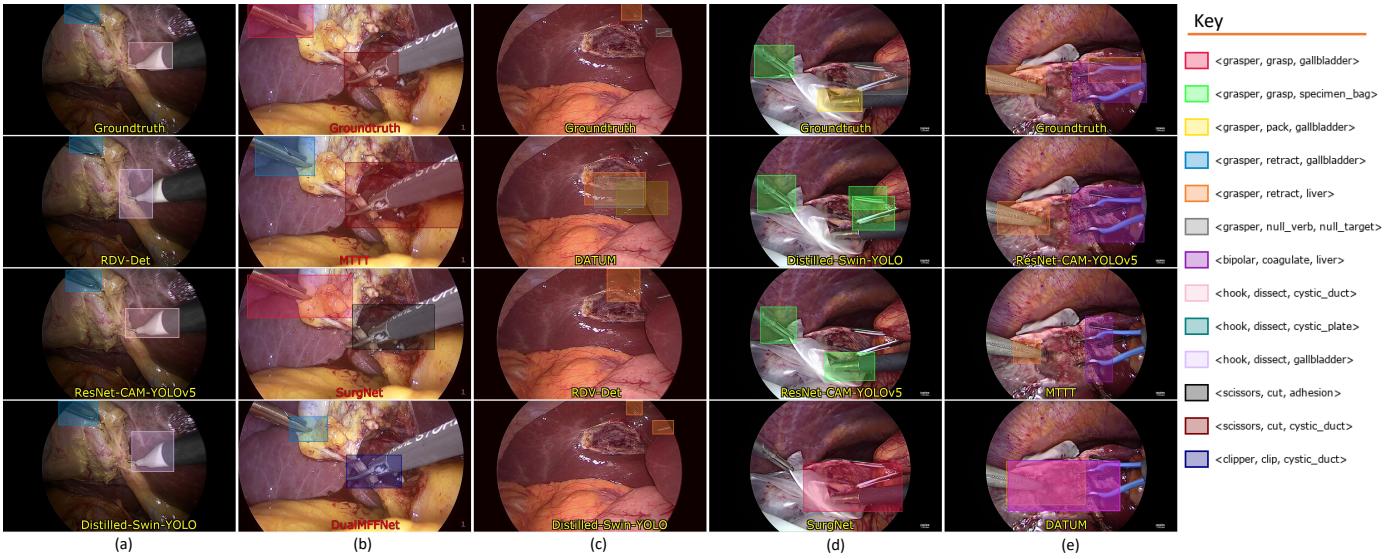


**Fig. 11. Qualitative results visualizing triplet detection under different instrument usage conditions, grouped by columns:** (a) regular case occurring too often, (b) rare case occurring only once in a procedure, (c) nearly invisible instrument, (d) crowded scene, and (e) partial occlusion. Ground truths are shown at the top rows while predictions from three randomly selected models are shown in other rows. Localization is illustrated using bounding boxes with color coding (Key at the right) matching the triplet labels.

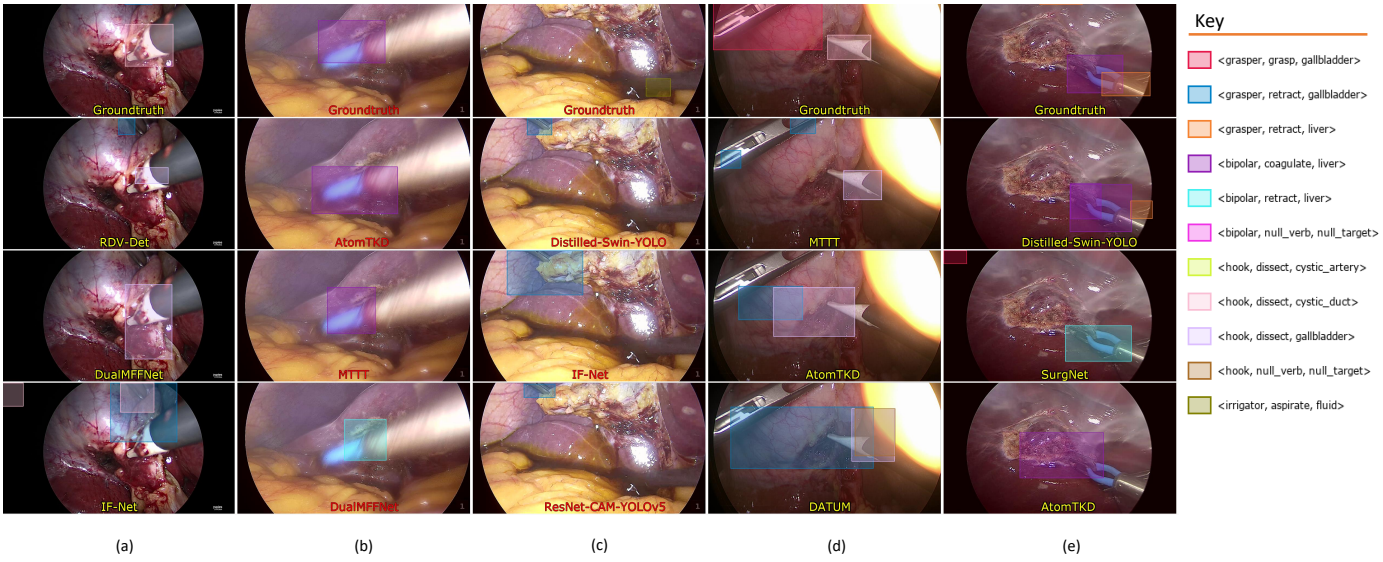


**Fig. 12. Qualitative results visualizing triplet detection under different visual challenges (surgical noise), grouped by columns:** (a) blood stain, (b) rapid motion blurring, (c) immersion in fluid, (d) specular light reflection, and (e) presence of smoke. Ground truths are shown at the top rows while predictions from three randomly selected models are shown in other rows. Localization is illustrated using bounding boxes with color coding (Key at the right) matching the triplet labels.

that the grasper held by the assistant surgeon is missed by all models as it is heavily overlaid by smoke. The light smoke on bipolar has minimal effects on the models though it affects the instrument localization for the AtomTKD and the triplet classification for the SurgNet.

In general, the instruments are mostly predicted correctly. Most of the missed predictions arise from incorrect verb and/or target predictions. Notwithstanding, the major factor contributing to the low AP scores is insufficient overlap between the localized instruments and their corresponding groundtruths leading to prohibitively high false negatives as shown in the last column of Table 6. This is also confirmed in Fig. 10 where the AP performance decreases with increasing the overlapping threshold. Triplet misclassification and their identity mismatch contribute to further decrease of the detection AP score.

### *6.5. Comparison with the CholecTriplet2021 challenge*

While the CholecTriplet2022 challenge is an evolution of the 2021 edition (Nwoye et al., 2022a), we note here several differentiating factors in the organization, participation, and methodology. The primary distinction between the two editions is the scope of each challenge. The first edition aimed at benchmarking, analyzing, and improving methods for surgical triplet

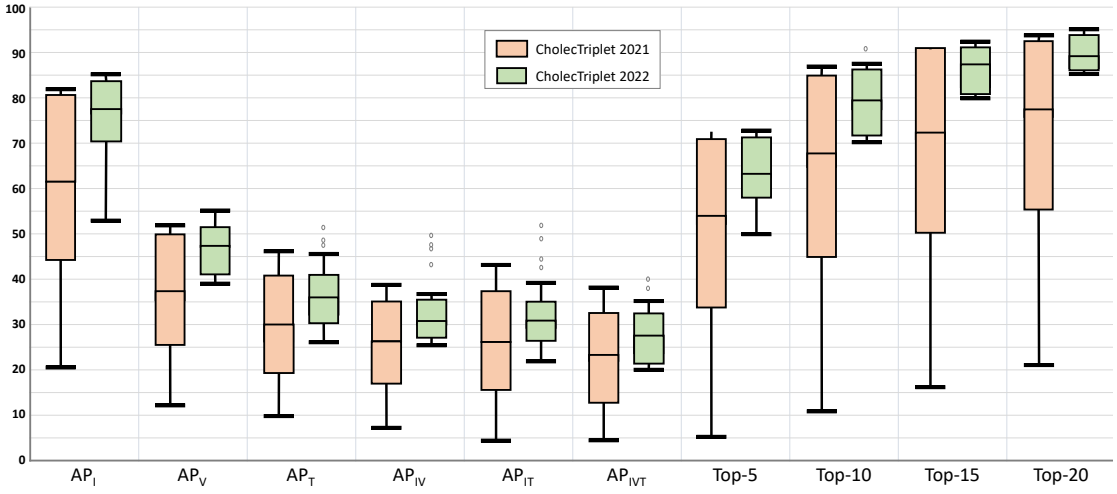


**Fig. 13. Box plot comparison of the quantitative results (Category 1) with the previous edition of the challenge (CholecTriplet 2021).**

recognition. As a natural progression, the current edition extends these tasks to be able to localize the operating instrument corresponding to each detected triplet (i.e. from triplet recognition to triplet detection) without spatial labels for training. To evaluate these methods, the 5 test videos of the CholecT50 dataset were augmented to include bounding box labels around the instrument tips in addition to the binary triplet labels available for all 50 videos.

Further, we note that several aspects of the validation and submission process were also refined. Regarding validation, the CholecTriplet2022 challenge utilized a self-validation approach with strict requirements on input/output formats and computational resources. Compared to the CholecTriplet2021 challenge, this enabled a more transparent validation process, and faster debugging, and helped focus organizational communication toward only significant issues. The submission portal was also moved to DockerHub to allow native submission of docker files without conversion or compression. The evaluation protocol itself was modified to allow structured outputs as a JavaScript Object Notation (JSON file) rather than a text file for quicker processing. The metrics were also appropriately modified to be able to effectively assess the localization of instruments and triplets in addition to the recognition metrics used in the previous edition. To support these additional organizational efforts, the committee was expanded from 4 members to 8.

In terms of the outcome, the 2022 edition offers a wide range of methods: 6 out of 11 models have extremely distinctive methodologies. Interestingly, methods focusing on knowledge distillation are proposed to supplement weakly-supervised learning of spatial localization from different datasets. There is also a significant reduction in the gap between the top and bottom placed models in the leaderboard as illustrated by box plots in Fig 13: the smaller standard deviation across all metrics illustrates an improved and better understanding of the task modeling compared to the 2021 edition. The final triplet recognition mAP is not outperformed in the latest edition of the challenge. This could be attributed to the increased complexity of the task and the challenge of optimizing more task heads at various levels of supervision (full and weak). The top-K results, meanwhile, generally improved.

### *6.6. Limitations*

Notwithstanding the milestone, the challenge on surgical action triplet detection faces some limitations. Firstly, as with most challenges, given that the participants are not constrained in terms of modeling, the comparative analysis may be taken as indications rather than facts, especially since aside from the strength of the modeling techniques, lots of other factors such as data preprocessing, augmentation styles, model size, hyperparameter tuning, training computational power, weight initialization, etc., can influence model performances.

Another limitation is the choice of supervision. The localization task of the challenge is defined by weak supervision by default. However, weak supervision is approached differently: in this case, some teams train directly on weaker binary presence labels while others train a teacher model on external data which in turn generates pseudo-spatial labels for model fine-tuning. Both approaches are acceptable standards for weak supervision, however, their advantages differ thus affecting their comparability. The best we could do in terms of constraints is to withhold the spatial labels of the benchmark dataset.

Looking at the obtained results, there is still room for improvement, especially on the localization task. Also, localization, which is explored on the instruments in this challenge, is yet to be extended to the surgical targets. The task of weakly supervised target anatomy localization remains an open problem as generating spatial annotation is quite expensive, and importantly, the target anatomy is not solely determined by its visibility, which further complicates the problem. Furthermore, there are fewer datasets in the community for differentiating the anatomy/organ structures causing a delay in research focusing on recognition of anatomical targets: there is also no known pre-training source. This contributes to the imbalance and suggests why the target recognition is poor compared to the instruments.

Finally, how the traditional AP metrics translates a model capacity at triplet recognition and detection is yet to be investigated in full. In the CholecT45 and CholecT50 datasets, there is

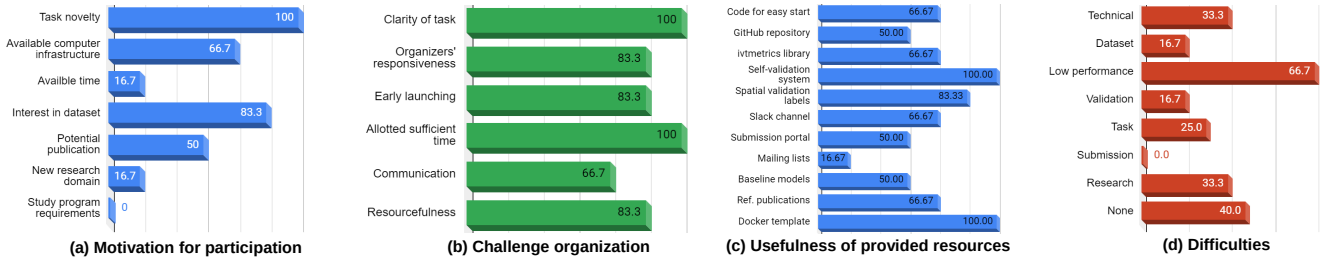


**Fig. 14. Summary of a survey conducted on CholecTriplet2022 at the end of the challenge.** Each item is independently rated on a scale of 1-5. Average values expressed in percentage. Similar items are grouped in one plot for analysis. Values of grouped items are not normalized to a fixed number.

a level of similarity between certain distinct triplets. As an example, the triplets ⟨*grasper, grasp, gallbladder*⟩ and ⟨*grasper, retract, gallbladder*⟩ are different but at certain stage of the procedure could be used interchangeably without misrepresentation. At some other point, an action, e.g. *dissect*, can happen at an unclear boundary such that the two adjacent anatomies, e.g. *gallbladder-neck* versus *cystic-duct*, can be interchangeably considered as the target without loss of generality. Taking everything into consideration, it may be fairer to design AP metrics that would consider the proximity of the predicted triplets to their interchangeable ground truth counterparts. This would to an extent address the low performance scores of deep learning models on triplet recognition and detection.

## 7. Challenge survey

We conducted an anonymized survey on the organization of the challenge. The questionnaires were circulated online to the participants including those unable to proceed to the submission stage. The respondents were allowed to rate on a scale of 1-5 (later converted to percentile in the analysis) their satisfaction on each of the parameters of the challenge, and offer recommendations where they feel so. We collated the responses into four groups, presented in Fig. 14, and summarized as follow:

(a) *Motivation for participation:* The novelty of the task of triplet detection as well as the interest in the CholecT50 dataset top the reasons for many entrants into the contest. Participation is mildly influenced by the availability of computer infrastructure and the opportunity for a joint publication of the challenge findings.

(b) *Challenge organization:* The organization and coordination of the challenge receive a very high rating, especially on the clarity of the task instructions, responsiveness to participants' queries, and provision of helpful resources. Participants agree that the challenge started early enough with sufficient time allotted for its completion.

(c) *Usefulness of provided resources:* The Docker template, self-validation system, and spatial-labeled dataset were rated to be the most useful. Providing snippets of code for easy starting, metrics library, reference papers, and GitHub repositories are also of great help. The creation of group and support mailing lists is rated lowest and this could be attributed to the alternative use of a dedicated Slack channel for daily communication.

(d) *Difficulties:* The participants identify the low performance of their models as a major factor discouraging them from proceeding beyond the validation phase. This can be attributed to the task difficulty which is also reported in the survey. Technical difficulties mostly arise from the building of Docker containers. It is observed that dataset usage and validation process pose less issues.

These observations can be valuable in improving future computer vision competitions.

## 8. Conclusion

With CholecTriplet2022, we presented a study in the form of an international contest focusing on the development of deep learning models for the detection of fine-grained surgical activities in laparoscopic videos. The task is subdivided into three categories of (1) recognizing surgical actions as triplets ⟨*instrument, verb, target*⟩, (2) localizing the instruments performing the actions, and (3) pairing the localized instruments with their correct action triplets. The research and experiments are conducted on the CholecT50 dataset which has been annotated with triplet labels. The localization sub-task is handled by weak supervision. The challenge ran through a 5-month window recording a total of 11 teams' participation showcasing 11 new deep learning methods to solve the problems.

The presented models employed several computer vision techniques such as state-of-the-art CNN architectures, multi-task learning, temporal modeling, graphical modeling, knowledge distillation, attention/transformer, transfer learning, model ensembling, end-to-end training, weak supervision, etc. While performances close to the baseline are recorded on the localization task, higher and promising results are obtained on the triplet recognition task. Our in-depth analysis of the results, quantitatively and qualitatively, reveals the behaviors of presented methods under various surgical conditions and visual challenges prevalent in surgical video data. With this meta-analysis, we provide insights for the advancement of the present and similar studies in the field.

In summary, this study contributes, in no small amount, to the gradual advancement of research on tool-tissue interaction understanding, which would be helpful for the development of future generation operating rooms that include AI assistance. Following a similar setup, the design and experiments contained in this paper can be extended to other surgical procedures.

## Acknowledgment

The organizers would like to thank the IHU and IRCAD research teams for their help with the initial data annotation during the CONDOR project. We also thank Stefanie Speidel, Lena Maier-Hein, Danail Stoyanov, and the entire EndoVis 2022 organizing committee for providing the platform for this challenge.

## Funding

This work was supported by French state funds managed within the Plan Investissements d'Avenir by the ANR under references: National AI Chair AI4ORSafety [ANR-20-CHIA-0029-01], Labex CAMI [ANR-11-LABX-0004], DeepSurg [ANR-16-CE33-0009], IHU Strasbourg [ANR-10-IAHU-02] and by BPI France under references: project CONDOR, project 5G-OR.

Software validation and evaluation were performed with servers managed by CAMMA at University of Strasbourg and IHU Strasbourg, as well as HPC resources from Unistra Mésocentre, and GENCI-IDRIS [Grant 2021-AD011011638R1, 2021-AD011011638R2].

Awards for the challenge winners were sponsored by IHU Strasbourg, NVIDIA, and Medtronic Ltd.

Participating teams would like to acknowledge the following funding: **CITI**: Shanghai Municipal Science and Technology Commission [20511105205]. **SDS-HD**: Twinning Grant [DKFZ+RBCT]; the Surgical Oncology Program of the National Center for Tumor Diseases (NCT) Heidelberg, by the German Federal Ministry of Health under the reference number 2520DAT0P1 as part of the pAItient project, and by HELMHOLTZ IMAGING, a platform of the Helmholtz Information & Data Science Incubator. European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (NEURAL SPICING; grant agreement No. [101002198]) and Surgical Oncology Program of the National Center for Tumor Diseases (NCT) Heidelberg. **2AI-ICVS**: Fundação para a Ciência e a Tecnologia (FCT), Portugal and the European Social Fund, European Union, for funding support through the "Programa Operacional Capital Humano" (POCH) in the scope of the Ph.D. grants [SFRH/BD/136721/2018, SFRH/BD/136670/2018]. Grants [NORTE-01-0145-FEDER-000045, NORTE-01-0145-FEDER-000059], supported by Northern Portugal Regional Operational Programme (NORTE 2020), under the Portugal 2020 Partnership Agreement, through the European Regional Development Fund (FEDER). Also funded by national funds, through the FCT and FCT/MCTES in the scope of the project [UIDB/05549/2020, UIDP/05549/2020]. **SHUANGCHUN**: Guangdong Climbing Plan under Grant [pdjh2023c21602]. **CAMP**: partially supported by Carl Zeiss AG.

## CRediT authorship contribution statement

**C.I. Nwoye**: Conceptualization, Data Curation, Methodology, Software, Investigation, Validation, Formal Analysis, Visualization, Data Analysis & Interpretation, Writing - Original Draft, Writing - Review & Editing, Resources, Challenge Organization. **T. Yu**: Conceptualization, Data Curation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing, Resources, Challenge Organization. **S. Sharma**: Conceptualization, Data Curation, Software, Formal Analysis, Validation, Visualization, Writing - Original Draft, Resources, Challenge Organization. **A. Murali**: Conceptualization, Data Curation, Formal Analysis, Software, Writing - Original Draft, Resources, Challenge Organization. **D. Alapatt**: Conceptualization, Data Curation, Formal Analysis, Writing - Original Draft, Challenge Organization. **A. Vardazaryan**: Conceptualization, Data Curation, Investigation, Writing - Review & Editing, Resources, Challenge Organization. **K. Yuan**: Formal Analysis, Visualization, Writing - Original Draft, Challenge Organization. **J. Hajek, W. Reiter, A. Yamlahi, F. Smidt, X. Zou, G. Zheng, B. Oliveira, H, Torres, S. Kondo, S. Kasai, F. Holm, E. Özsoy, S. Gui, H. Li, S. Raviteja, R. Sathish, P. Poudel, B. Bhattarai, Z. Wang, G. Rui**: Methodology, Software, Investigation, Visualization, Writing - Review & Editing. **M. Schellenberg, J. Vilaça, T. Czempiel, Z. Wang, D. Sheet, S.K. Thapa, M. Berniker, P. Godau, P. Morais, S. Regmi, T. Tran, J. Fonseca, J.H. Nölke, E. Lima, E. Vazquez, L. Maier-Hein, N. Navab**: Team Supervision, Writing - Review & Editing. **B. Seeliger, C. Gonzalez, P. Mascagni** : Data Curation, Writing - Review & Editing. **D. Mutter** : Data Curation, Writing - Review & Editing, Resources, Supervision. **N. Padoy**: Conceptualization, Project Administration, Project Supervision, Funding Acquisition, Resources, Writing - Review & Editing, Challenge Organization.

## Declaration of competing interest

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

## Data availability

The CholecT50 dataset and the validation data used in the challenge have been made available to the public and accessible via https://github.com/CAMMA-public/cholect50. The test set spatial labels will be released publicly. The baseline model code will be released as well. Participants can release their code on their own volition. All released code would be linked to the central GitHub repository for the challenge: https://github.com/CAMMA-public/cholectriplet2022.

## References

Al Hajj, H., Lamard, M., Conze, P.H., Cochener, B., Quellec, G., 2018. Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. Medical Image Analysis 47, 203–218.

Al Hajj, H., Lamard, M., Conze, P.H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., et al., 2019. Cataracts: Challenge on automatic tool annotation for cataract surgery. Medical Image Analysis 52, 24–41.

Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al., 2020. 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 .

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

Bawa, V.S., Singh, G., Kaping'a, F., Skarga-Bandurova, I., Oleari, E., Leporini, A., Landolfo, C., Zhao, P., Xiang, X., Luo, G., Wang, K., Li, L., Wang, B., Zhao, S., Li, L., Stabile, A., Setti, F., Muradore, R., Cuzzolin, F., 2021. The SARAS endoscopic surgeon action detection (ESAD) dataset: Challenges and methods. CoRR abs/2104.03178.

Bertasius, G., Wang, H., Torresani, L., 2021a. Is space-time attention all you need for video understanding?, in: Meila, M., Zhang, T. (Eds.), ICML 2021, PMLR. pp. 813–824.

Bertasius, G., Wang, H., Torresani, L., 2021b. Is space-time attention all you need for video understanding?, in: International Conference on Machine Learning ICML, p. 4.

Bodenstedt, S., Speidel, S., Allan, M., Stoyanov, D., Maier-Hein, L., Kenngott, H., Wagner, M., 2015. EndoVis'15 Instrument Segmentation and Tracking Challenge. https://endovissub-instrument.grand-challenge.org/. [Online; accessed 19-December-2022].

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers, in: International Conference on Computer Vision ICCV, IEEE/CVF. pp. 9650–9660.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE Computer Society. pp. 4724–4733.

Chao, Y., Liu, Y., Liu, X., Zeng, H., Deng, J., 2018. Learning to detect human-object interactions, in: Winter Conference on Applications of Computer Vision WACV, IEEE Computer Society. pp. 381–389.

Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J., 2015. Hico: A benchmark for recognizing human-object interactions in images, in: International Conference on Computer Vision, IEEE. pp. 1017–1025.

Chen, G., Wang, W., He, Z., Wang, L., Yuan, Y., Zhang, D., Zhang, J., Zhu, P., Van Gool, L., Han, J., et al., 2021. Visdrone-mot2021: The vision meets drone multiple object tracking challenge results, in: International Conference on Computer Vision, IEEE/CVF. pp. 2839–2846.

Cheng, Y., Liu, L., Wang, S., Jin, Y., Schönlieb, C.B., Aviles-Rivero, A.I., 2022. Why deep surgical models fail?: Revisiting surgical action triplet recognition through the lens of robustness. arXiv arXiv:2209.08647.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding, in: Conference on Computer Vision and Pattern Recognition CVPR, pp. 3213–3223.

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, Springer. pp. 343–352.

Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N., 2021. Opera: Attention-regularized transformers for surgical phase recognition, in: Medical Image Computing and Computer Assisted Intervention MICCAI, Springer. pp. 604–614.

Dai, R., Das, S., Kahatapitiya, K., Ryoo, M.S., Bremond, F., 2022. Ms-tct: Multi-scale temporal convtransformer for action detection, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE. pp. 20041–20051.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L., 2020. Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 .

Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., Saenko, K., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE Computer Society. pp. 2625–2634.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88, 303–338.

Everingham, M., Zisserman, A., Williams, C.K., Gool, L.V., Allan, M., Bishop, C.M., Chapelle, O., Dalal, N., Deselaers, T., Dorkó, G., et al., 2005. The 2005 pascal visual object classes challenge, in: Machine Learning Challenges Workshop, Springer. pp. 117–176.

Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. Slowfast networks for video recognition, in: International Conference on Computer Vision ICCV, IEEE/CVF. pp. 6201–6210.

Funke, I., Jenke, A., Mees, S.T., Weitz, J., Speidel, S., Bodenstedt, S., 2018. Temporal coherence-based self-supervised learning for laparoscopic workflow analysis, in: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, - and - Skin Image Analysis, Springer. pp. 85–93.

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: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention MICCAI, Springer. pp. 593–603.

Gkioxari, G., Girshick, R.B., Dollár, P., He, K., 2018. Detecting and recognizing human-object interactions, in: Conference on Computer Vision and Pattern Recognition CVPR, Computer Vision Foundation / IEEE Computer Society. pp. 8359–8367.

Grammatikopoulou, M., Flouty, E., Kadkhodamohammadi, A., Quellec, G., Chow, A., Nehme, J., Luengo, I., Stoyanov, D., 2019. Cadis: Cataract dataset for image segmentation. arXiv preprint arXiv:1906.11586 .

Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J., 2018. AVA: A video dataset of spatio-temporally localized atomic visual actions, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE Computer Society. pp. 6047–6056.

Gupta, S., Malik, J., 2015. Visual semantic role labeling. arXiv preprint arXiv:1505.04474 .

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE. pp. 770–778.

Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S., 2020. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 .

Huaulmé, A., Sarikaya, D., Le Mut, K., Despinoy, F., Long, Y., Dou, Q., Chng, C.B., Lin, W., Kondo, S., Bravo-Sánchez, L., et al., 2021. Micro-surgical anastomose workflow recognition challenge report. Computer Methods and Programs in Biomedicine 212, 106452.

Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A., Fei-Fei, L., 2018a. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks, in: Winter conference on applications of computer vision (WACV), IEEE. pp. 691–699.

Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C., Heng, P., 2018b. Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network. Transactions on Medical Imaging 37, 1114–1126.

Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.A., 2021. Temporal memory relation network for workflow recognition from surgical video. Transactions on Medical Imaging 40, 1911–1923.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE Computer Society. pp. 1725–1732.

Katic, D., Julliard, C., Wekerle, A., Kenngott, H., Müller-Stich, B.P., Dillmann, R., Speidel, S., Jannin, P., Gibaud, B., 2015. Lapontospm: an ontology for laparoscopic surgeries and its application to surgical phase recognition. Int. J. Comput. Assist. Radiol. Surg. 10, 1427–1434.

Katic, D., Wekerle, A., Gärtner, F., Kenngott, H., Müller-Stich, B.P., Dillmann, R., Speidel, S., 2014. Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance, in: Information Processing in Computer-Assisted Interventions IPCAI, Springer. pp. 158–167.

Khatibi, T., Dezyani, P., 2020. Proposing novel methods for gynecologic surgical action recognition on laparoscopic videos. Multim. Ton ols Appl. 79, 30111–30133.

Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Čehovin, L., 2016. A novel performance evaluation methodology for single-target trackers. Transactions on Pattern Analysis and Machine Intelligence 38, 2137–2155.

Lecuyer, G., Ragot, M., Martin, N., Launay, L., Jannin, P., 2020. Assisted phase and step annotation for surgical videos. Int. J. Comput. Assist. Radiol. Surg. 15, 673–680.

Li, L., Li, X., Ding, S., Fang, Z., Xu, M., Ren, H., Yang, S., 2022. Sirnet: Fine-

grained surgical interaction recognition. Robotics and Automation Letters 7, 4212–4219.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European Conference on Computer Vision ECCV. Springer, pp. 740–755.

Lin, W., Hu, Y., Hao, L., Zhou, D., Yang, M., Fu, H., Chui, C., Liu, J., 2022. Instrument-tissue interaction quintuple detection in surgery videos, in: International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI, Springer. pp. 399–409.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al., 2022a. Swin transformer v2: Scaling up capacity and resolution, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE. pp. 12009–12019.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H., 2022b. Video swin transformer, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE Computer Society. pp. 3192–3201.

Luengo, I., Grammatikopoulou, M., Mohammadi, R., Walsh, C., Nwoye, C.I., Alapatt, D., Padoy, N., Ni, Z.L., Fan, C.C., Bian, G.B., et al., 2021. 2020 cataracts semantic segmentation challenge. arXiv preprint arXiv:2110.10965 .

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., et al., 2020. Bias: Transparent reporting of biomedical image analysis challenges. Medical Image Analysis 66, 101796.

Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P.M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., et al., 2021. Heidelberg colorectal data set for surgical data science in the sensor operating room. Scientific data 8, 1–11.

Mallya, A., Lazebnik, S., 2016. Learning models for actions and person-object interactions with transfer to question answering, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), European Conference on Computer Vision ECCV, Springer. pp. 414–428.

Maqbool, S., Riaz, A., Sajid, H., Hasan, O., 2020. m2caiseg: Semantic segmentation of laparoscopic images using convolutional neural networks. arXiv preprint arXiv:2008.10134 .

Mascagni, P., Alapatt, D., Sestini, L., Altieri, M.S., Madani, A., Watanabe, Y., Alseidi, A., Redan, J.A., Alfieri, S., Costamagna, G., et al., 2022. Computer vision in surgery: from potential to clinical value. npj Digital Medicine 5, 163.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (brats). Transactions on medical imaging 34, 1993–2024.

Nwoye, C., 2021. Deep Learning Methods for the Detection and Recognition of Surgical Tools and Activities in Laparoscopic Videos. Ph.D. thesis. Université de Strasbourg. URL: http://icube-publis.unistra.fr/8-Nwoy21.

Nwoye, C.I., Alapatt, D., Yu, T., Vardazaryan, A., Xia, F., Zhao, Z., Xia, T., Jia, F., Yang, Y., Wang, H., et al., 2022a. Cholectriplet2021: A benchmark challenge for surgical action triplet recognition. arXiv preprint arXiv:2204.04746 .

Nwoye, C.I., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N., 2020. Recognition of instrument-tissue interactions in endoscopic videos via action triplets, in: International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI, Springer. pp. 364–374.

Nwoye, C.I., Mutter, D., Marescaux, J., Padoy, N., 2019. Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos. Int. J. Comput. Assist. Radiol. Surg. 14, 1059–1067.

Nwoye, C.I., Padoy, N., 2022. Data splits and metrics for method benchmarking on surgical action triplet datasets. arXiv preprint arXiv:2204.05235 .

Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N., 2022b. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Medical Image Analysis 78, 102433 .

Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S., 2018. Learning human-object interactions by graph parsing neural networks, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), European Conference on Computer Vision ECCV, Springer. pp. 407–423.

Ramesh, S., Dall'Alba, D., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Fiorini, P., Padoy, N., 2021. Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. Int. J. Comput. Assist. Radiol. Surg. , 1 – 9.

Roß, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., Bruno, P., Arbeláez, P., Bian, G.B., Bodenstedt, S., Bolmgren, J.L., Bravo-Sánchez, L., Chen, H.B., González, C., Guo, D., Halvorsen, P., Heng, P.A., Hosgor, E., Hou, Z.G., Isensee, F., Jha, D., Jiang, T., Jin, Y., Kirtac, K., Kletz, S., Leger, S., Li, Z., Maier-Hein, K.H., Ni, Z.L., 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.J., 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 Analysis 70, 101920.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. International journal of computer vision 115, 211–252.

Shen, W., Zhao, K., Jiang, Y., Wang, Y., Zhang, Z., Bai, X., 2016. Object skeleton extraction in natural images by fusing scale-associated deep side outputs, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE. pp. 222–230.

Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c., 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems 28.

Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A., 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), European Conference on Computer Vision ECCV, Springer. pp. 510–526.

Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Conference on Neural Information Processing Systems Nurips, pp. 568–576.

Soomro, K., Zamir, A., 2014. Action recognition in realistic sports videos. Advances in Computer Vision and Pattern Recognition 71, 181–208.

Soomro, K., Zamir, A.R., Shah, M., 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. ArXiv abs/1212.0402.

Stauder, R., Ostler, D., Kranzfelder, M., Koller, S., Feußner, H., Navab, N., 2016. The tum lapchole dataset for the m2cai 2016 workflow challenge. arXiv preprint arXiv:1610.09278 .

Tamura, M., Ohashi, H., Yoshinaga, T., 2021. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE Computer Society. pp. 10410–10419.

Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks, in: International Conference on Computer Vision ICCV, IEEE Computer Society. pp. 4489–4497.

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N., 2016. Endonet: a deep architecture for recognition tasks on laparoscopic videos. Transactions on medical imaging 36, 86–97.

Vardazaryan, A., Mutter, D., Marescaux, J., Padoy, N., 2018. Weakly-supervised learning for tool localization in laparoscopic videos, in: Intravascular Imaging and Computer Assisted Stenting - and - Large-Scale Annotation of Biomedical Data and Expert Label Synthesis MICCAI LABELS, Springer. pp. 169–179.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems Nurips 30.

Vercauteren, T., Unberath, M., Padoy, N., Navab, N., 2019. Cai4cai: the rise of contextual artificial intelligence in computer-assisted interventions. Proceedings of the IEEE 108, 198–214.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B., 2019. Mots: Multi-object tracking and segmentation, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE Computer Society. pp. 7942–7951.

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., Yu, T., Vardazaryan, A., Nwoye, C.I., Padoy, N., Liu, X., Lee, E.J., Disch, C., Meine, H., Xia, T., Jia, F., Kondo, S., Reiter, W., Jin, Y., Long, Y., Jiang, M., Dou, Q., Heng, P.A., Twick, I., Kirtac, K., Hosgor, E., Bolmgren, J.L., Stenzel, M., von Siemens, B., Kenngott, H.G., Nickel, F., von Frankenberg, M., Mathis-Ullrich, F., Maier-Hein, L., Speidel, S., Bodenstedt, S., 2021. Comparative validation of machine learning algo-

rithms for surgical workflow and skill analysis with the heichole benchmark. arXiv:2109.14956.

Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y., 2022. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy, in: International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI, Springer. pp. 486–496.

Wei, J., Wang, Q., Li, Z., Wang, S., Zhou, S.K., Cui, S., 2021. Shallow feature matters for weakly supervised object localization, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE. pp. 5993–6001.

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. Scientific reports 11, 1–15.

Xi, N., Meng, J., Yuan, J., 2022. Forest graph convolutional network for surgical action triplet recognition in endoscopic videos. Transactions on Circuits and Systems for Video Technology .

Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K., 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), European Conference on Computer Vision ECCV, Springer. pp. 318–335.

Xu, M., Islam, M., Lim, C.M., Ren, H., 2021. Learning domain adaptation with model calibration for surgical report generation in robotic surgery, in: International Conference on Robotics and Automation ICRA, IEEE. pp. 12350–12356.

Ye, J., He, J., Peng, X., Wu, W., Qiao, Y., 2020. Attention-driven dynamic graph convolutional network for multi-label image recognition, in: European conference on computer vision ECCV, Springer. pp. 649–665.

Yu, T., Mutter, D., Marescaux, J., Padoy, N., 2019. Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition. Information Processing in Computer-Assisted Interventions IPCAI arXiv:1812.00033.

Zhang, F.Z., Campbell, D., Gould, S., 2022. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE Computer Society. pp. 20104–20112.

Zia, A., Bhattacharyya, K., Liu, X., Wang, Z., Kondo, S., Colleoni, E., van Amsterdam, B., Hussain, R., Hussain, R., Maier-Hein, L., et al., 2021. Surgical visual domain adaptation: results from the miccai 2020 surgvisdom challenge. arXiv preprint arXiv:2102.13644 .

Zia, A., Liu, X., Bhattacharyya, K., Wang, Z., Berniker, M., Jarc, A., Nwoye, C., Alapatt, D., Murali, A., Sharma, S., Vardazaryan, A., Padoy, N., Amsterdam, B.V., Psychogyios, D., Colleoni, E., Rau, A., Bano, S., Jin, Y., Cartucho, J., Giannarou, S., Ali, S., Jin, Y., López, Y.E., Buc, E., Roy, B.L., Teoule, P., Reissfelder, C., Bailey, A., Soonawalla, Z., Gordon-Weeks, A., Silva, M., Bartoli, A., Roß, T., Reinke, A., Bodenstedt, S., Stoyanov, D., Maier-Hein, L., Speidel, S., 2022. Endoscopic vision challenge 2022. doi:10.5281/zenodo.6390403.

Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., et al., 2021. End-to-end human object interaction detection with hoi transformer, in: Conference on Computer Vision and Pattern Recognition CVPR, CVF / IEEE. pp. 11825–11834.