



HAL
open science

Rendezvous: Attention Mechanisms for the Recognition of Surgical Action Triplets in Endoscopic Videos

Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Nicolas Padoy

► **To cite this version:**

Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, et al.. Rendezvous: Attention Mechanisms for the Recognition of Surgical Action Triplets in Endoscopic Videos. *Medical Image Analysis*, 2022, 78, pp.102433. 10.1016/j.media.2022.102433 . hal-03765190

HAL Id: hal-03765190

<https://hal.science/hal-03765190v1>

Submitted on 22 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

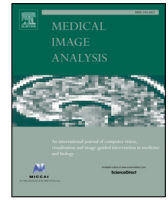


Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License



Contents lists available at ScienceDirect

Medical Image Analysis

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

Rendezvous: attention mechanisms for the recognition of surgical action triplets in endoscopic videos

Chinedu Innocent Nwoye^{a,*}, Tong Yu^a, Cristians Gonzalez^{b,c}, Barbara Seeliger^{b,c}, Pietro Mascagni^{a,d}, Didier Mutter^{b,c}, Jacques Marescaux^e, Nicolas Padoy^{a,b}

^aICube, University of Strasbourg, CNRS, France

^bIHU Strasbourg, France

^cUniversity Hospital of Strasbourg, France

^dFondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy

^eIRCAD France

ARTICLE INFO

Article history:

Received -

Received in final form -

Accepted -

Available online -

Communicated by -

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Surgical workflow analysis, Tool-tissue interaction, CholecT50, Attention, Transformer, Deep learning, Laparoscopic surgery

ABSTRACT

Out of all existing frameworks for surgical workflow analysis in endoscopic videos, action triplet recognition stands out as the only one aiming to provide truly fine-grained and comprehensive information on surgical activities. This information, presented as $\langle instrument, verb, target \rangle$ combinations, is highly challenging to be accurately identified. Triplet components can be difficult to recognize individually; in this task, it requires not only performing recognition simultaneously for all three triplet components, but also correctly establishing the data association between them. To achieve this task, we introduce our new model, the *Rendezvous* (RDV), which recognizes triplets directly from surgical videos by leveraging attention at two different levels. We first introduce a new form of spatial attention to capture individual action triplet components in a scene; called *Class Activation Guided Attention Mechanism* (CAGAM). This technique focuses on the recognition of verbs and targets using activations resulting from instruments. To solve the association problem, our RDV model adds a new form of semantic attention inspired by Transformer networks; called *Multi-Head of Mixed Attention* (MHMA). This technique uses several cross and self attentions to effectively capture relationships between instruments, verbs, and targets. We also introduce *CholecT50* - a dataset of 50 endoscopic videos in which every frame has been annotated with labels from 100 triplet classes. Our proposed RDV model significantly improves the triplet prediction mAP by over 9% compared to the state-of-the-art methods on this dataset.

© 2022 Elsevier B. V. All rights reserved.

1. Introduction

Laparoscopic cholecystectomy, as one of the most commonly performed surgical procedures in the world (Shaffer, 2006; Majumder et al., 2020), has become the gold standard approach

(Pucher et al., 2018) over its open surgery counterpart. As a minimally invasive procedure, it significantly alleviates some of the preoperative, intraoperative, and postoperative burden: the patient generally experiences decreased odds of nosocomial infection, less pain, less bleeding, and faster recovery times (Velanovich, 2000). Yet, this success comes at a price for the surgeon, who now has to deal with increased technical difficulty coming from the indirect vision and laparoscopic instruments (Ballantyne, 2002), especially during complex cases

*Corresponding author: Tel.: +33 (0) 3 904 13 535;
e-mail: nwoye@unistra.fr (Chinedu Innocent Nwoye),
npadoy@unistra.fr (Nicolas Padoy)

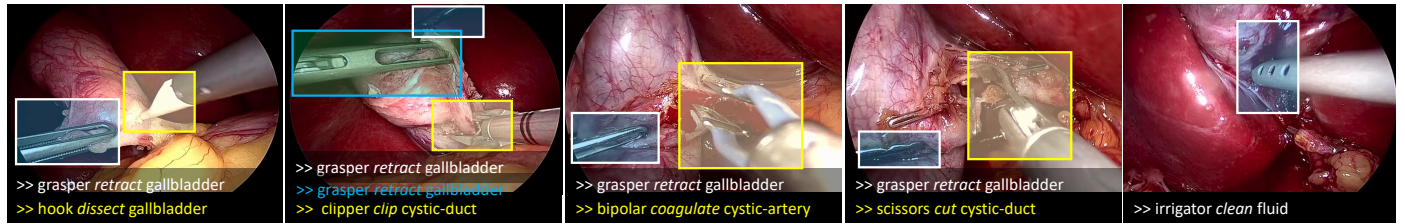


Fig. 1: Some examples of action triplets from CholecT50 dataset. The localization is not part of the dataset, but a representation of the weakly-supervised output of our recognition model.

(Felli *et al.*, 2019). The elevated complexity of laparoscopy is one of the motivations driving the development of context-aware support systems for surgery (Maier-Hein *et al.*, 2017); i.e. systems capable of assisting surgeons, for example via automated warnings (Vercauteren *et al.*, 2019), based on their dynamic perception and understanding of the surgical scene and workflow.

Developing this understanding is the focus of surgical workflow analysis methods: *given a scene from surgery, what is happening in it?* The finer-grained the answer becomes, the more value it gains in terms of clinical utility: for instance according to Mascagni *et al.* (2021), an automated surgical safety system would benefit from the ability to identify individual actions such as a *clipper* applying a *clip* to the *cystic-artery* or other blood vessels.

Methods from the literature have so far only given incomplete answers, with our previous work as the only exception (Nwoye *et al.*, 2020). The main task studied by the community, surgical phase recognition (Ahmadi *et al.*, 2006; Lo *et al.*, 2003a), only describes scenes at a very coarse level. As an example the *clipping and cutting* phase (Twinanda *et al.*, 2017) in cholecystectomy contains a multitude of important actions: *graspers* holding anatomical landmarks, a *clipper* applying several clips, laparoscopic *scissors* cutting the *cystic-duct* and so on. The phase information on its own does not, by any means, provide an accurate picture of the activities taking place. Even finer-grained workflow divisions such as steps (Ramesh *et al.*, 2021) are composed of multiple individual actions. Limited attempts were made in other works (Khatibi and Dezyani, 2020; Rupprecht *et al.*, 2016) as well as in MICCAI 2019’s Endoscopic Vision challenge¹ to capture those actions focusing on key verbs such as *dissect*, *cut*, or *coagulate*. This type of framework, however, overlooks interactions with the anatomy.

To obtain a comprehensive account of a surgical scene, simultaneous recognition of the instruments, verbs, target anatomy and the relationships between them needs to be achieved. This goes beyond the conventional action recognition used in the EndoVis 2019 sub-challenge to a deeper understanding of visual semantics that depicts the complex relationships between instruments and tissues. Katić *et al.* (2014) proposed the *surgical action triplet* as the most detailed and expressive formalism for surgical workflow analysis. However, while Katić *et al.* (2014, 2015) leveraged triplet formulation provided by manual annotation to better recognize surgical phases, no

attempt outside of our previous work (Nwoye *et al.*, 2020) has been made to directly recognize those triplets from surgical images or videos.

Nonetheless, the difficulty of this recognition task, for all its utility, is not to be overlooked. First, action triplets are **instrument-centric**: meaning that visibility alone does not determine the consideration of an anatomy as a part of a triplet, but also their involvement in an interaction carried out by an instrument. For instance, the *liver*, which is visible most of the time in laparoscopic cholecystectomy, is labeled a target only when being acted upon by an instrument. Similarly, a verb is defined by the instrument’s action, and so, without an instrument, there cannot be a verb. Furthermore, the level of **spatial reasoning** involved is highly challenging: given an instrument, its role (verb) with respect to a given target can imperceptibly change: $\langle \textit{grasper}, \textit{retract}, \textit{gallbladder} \rangle$, $\langle \textit{grasper}, \textit{grasp}, \textit{gallbladder} \rangle$, $\langle \textit{grasper}, \textit{dissect}, \textit{gallbladder} \rangle$ are tough to distinguish even for experienced surgeons and require careful observation of the area surrounding the tooltip. The applications of the surgical instruments vary according to the surgeon’s intention for use. Multiplicity and **semantic reasoning** are the other major challenges. Overlaps are found between different instruments used for the same action (or verb), e.g. *dissection* performed by *bipolar*, *grasper*, *hook*, *irrigator*, and *scissors* (see Table 2). Similarly, when operating on an organ or structure, multiple instruments can interact with the target. Thus, a target can be simultaneously involved in multiple distinct actions, e.g. $\langle \textit{grasper}, \textit{retract}, \textit{cystic-duct} \rangle$ and $\langle \textit{hook}, \textit{dissect}, \textit{cystic-duct} \rangle$ happening at the same time. Since multiple triplets can occur in one frame, associating the matching components of the triplets is akin to solving a complex tripartite data association problem between the entities. The model from our previous work, as a first attempt to tackle the triplet recognition problem, addresses those challenges in a limited manner, without explicit spatial focus or an advanced enough model of the instrument - verb - target relationships.

In this paper, we extend Nwoye *et al.* (2020), our conference paper published in MICCAI 2020. Our extension is both in data and in methods. On the data contribution, we introduce the *CholecT50* dataset, which is a quantitative and qualitative expansion of the *CholecT40* dataset (Nwoye *et al.*, 2020). The dataset consists of 50 videos of cholecystectomy annotated with 161K instances from 100 triplet classes. Examples of such action triplets include: $\langle \textit{grasper}, \textit{retract}, \textit{gallbladder} \rangle$, $\langle \textit{hook}, \textit{dissect}, \textit{cystic-plate} \rangle$, $\langle \textit{clipper}, \textit{clip}, \textit{cystic-artery} \rangle$, $\langle \textit{scissors}, \textit{cut}, \textit{cystic-duct} \rangle$, $\langle \textit{irrigator}, \textit{aspirate}, \textit{fluid} \rangle$, etc, as also shown in

¹<https://endovissub-workflowandskill.grand-challenge.org/>

Fig 1.

Method-wise, we develop a new recognition model called *Rendezvous* (RDV), which is a transformer-inspired neural network for surgical action triplet recognition. RDV improves on the existing *Tripnet* model proposed in Nwoye et al. (2020) by leveraging the attention mechanism to detect the various components of the triplets and learn their association. Our first effort at exploiting attention mechanisms in this task led to the development of a *Class Activation Guided Attention Mechanism* (CAGAM) to better detect the verb and target components of the triplet, which are instrument-centric. The CAGAM is achieved by redesigning the saliency-guided attention mechanism in (Ji et al., 2019; Yao and Gong, 2020) to utilize a more adequate and easier to learn class activation map (CAM). While our approach is similar in the attention guiding principle, it differs in three respects: (a) our attention network is guided by the instrument's activations which are learnable in the same network, using a global pooling layer without relying on a *third-party* saliency generation network, (b) our attention guide implements a combination of position and channel attention for the target and verb detection tasks respectively, (c) we employ cross-attention from the instrument domain to the other task domains (i.e.: *verb and target*) as opposed to self-attention in Yao and Gong (2020). Meanwhile, the CAGAM is an improvement on the class activation guide (CAG) module introduced in Nwoye et al. (2020) which is simply a concatenation of the model's intermediary features with the instrument's activation features. As an ablation experiment, we show the improved performance of our previous *Tripnet* model, only upgraded with the CAGAM. This upgraded model is called *Attention Tripnet*.

The proposed RDV uses the CAGAM unit as part of its encoder and a *multiple heads of mixed attention* (MHMA) decoder to learn the action triplets. The MHMA in RDV is inspired by the Transformer model (Vaswani et al., 2017) used in Natural Language Processing (NLP), particularly, its multi-head attention mechanism. Unlike in the NLP Transformer, which implements a multi-head of self-attention, we design a novel multi-head attention module that is a mixture of self- and cross-attention suitable for the action triplet recognition task. As opposed to Transformers in NLP which use multi-head attention temporally (i.e. focusing on a sequence of words in a sentence), and the Vision Transformer (Dosovitskiy et al., 2020) which uses it spatially (by forming its sequence from different patches of an image), our RDV model takes a different approach by employing it semantically, with redesigned multi-head modeling attention across the discriminating features of various components that are interacting to form action triplets. With this method, we outperform the state-of-the-art models significantly on triplet recognition. We plan to release our source code along with the evaluation script on our public github^{2,3,4} upon acceptance of this paper. The dataset will also become public on the CAMMA team's website⁵.

In summary, the contributions of this work are as follows:

1. We present a comprehensive study on surgical action triplet recognition directly from videos.
2. We propose a Class Activation Guided Attention Mechanism (CAGAM) for detecting the target and verb components of the triplets conditioned on the instrument's appearance cue.
3. We propose a Multi-Head of Mixed Attention (MHMA) by modeling self- and cross-attention on semantic sequences of class-wise representations to learn the interaction between the instrument, verb, and target in a surgical scene.
4. We develop *Rendezvous* (RDV): a transformer-inspired neural network model that utilizes CAGAM and MHMA for surgical triplet recognition in laparoscopic videos.
5. We present a large endoscopic action triplet dataset, *CholecT50*, for this task.
6. We analyze the surgical relevance of our methods and results, setting the stage for clinical translation and future research.

2. Related Work

2.1. Surgical Workflow Analysis

The paradigm shift brought by Artificial Intelligence (AI) across several fields has seen the application of deep learning techniques for the recognition of surgical workflow activities to provide assisted interventions in the operating room (OR). However, compared to other fields such as natural Vision, NLP, Commerce, etc., there has been a delay in introducing large-scale data science to interventional medicine. This is partly due to the unavailability of large annotated dataset (Maier-Hein et al., 2017) and the particular need for precision in medicine. Some research focuses on detecting elements such as instruments/tools used during surgery (Al Hajj et al., 2018; Garcia-Peraza-Herrera et al., 2017; Nwoye et al., 2019; Sznitman et al., 2014; Vardazaryan et al., 2018; Voros et al., 2007), while others model the sequential workflow by recognizing surgical phases either from endoscopic videos (Blum et al., 2010; Dergachyova et al., 2016; Funke et al., 2018; Lo et al., 2003a; Twinanda et al., 2017; Yu et al., 2018; Zisimopoulos et al., 2018) or from ceiling-mounted cameras (Chakraborty et al., 2013; Twinanda et al., 2015). Some works go deeper in the level of granularity, recognizing the steps within each surgical phase (Charriere et al., 2014; Lecuyer et al., 2020; Ramesh et al., 2021), while others learn phase transitions (Sahu et al., 2020). Another work (Lo et al., 2003b) investigated the four major events in minimally invasive surgery (MIS) and categorized these events into their main actions; namely, *idle*, *retraction*, *cauterization*, and *suturing*. From the perspective of robotic surgery, similar research focused more on gesture recognition from kinematic data (DiPietro et al., 2016, 2019), and robotized surgeries (Kitaguchi et al., 2019; Zia et al., 2018), system events (Malpani et al., 2016), and the recognition of other events, such as the presence of smoke or bleeding (Loukas and Georgiou, 2015). These surgical events are explored for the recognition of surgeon's deviation from standard processes in laparoscopic videos (Hualmé et al., 2020).

²<https://github.com/CAMMA-public/tripnet>

³<https://github.com/CAMMA-public/rendezvous>

⁴<https://github.com/CAMMA-public/attention-tripnet>

⁵<http://camma.u-strasbg.fr/datasets>

Aside the coarse-grained activities, some works (Khatibi and Dezyani, 2020; Rupprecht *et al.*, 2016) explored fine-grained action in laparoscopic videos, however, the recognition task is limited to verb classification. Within the EndoVis challenge at MICCAI 2019, Wagner *et al.* (2021) introduced a similar action recognition task for only four prevalent verbs in surgery (*cut*, *grasp*, *hold*, and *clip*), however, this does not consider the target anatomy or the instrument performing the action. In the SARAS-ESAD challenge organized within MIDL 2020, the proposed action labels encompass 21 classes (Bawa *et al.*, 2021). The EASD challenge dataset is an effort to capture more details in surgical action recognition. While this dataset provides spatial labels for action detection, just like some human-object interaction (HOI) datasets, it formalizes action labels as verb-anatomy relationship such as *clippingTissue*, *pullingTissue*, *cuttingTissue*, etc., and thus, do not take into account the instrument performing the actions. Although humans are not categorized in the general vision HOI problem, it is imperative to recognize the surgical instruments by their categories as they play semantically different roles; their categories are informative in distinguishing the surgical phases. Recognizing surgical actions as single verbs is also being explored in other closely related procedures such as gynecologic laparoscopy (Khatibi and Dezyani, 2020; Kletz *et al.*, 2017; Petscharnig *et al.*, 2018).

For a more detailed workflow analysis, Nwoye *et al.* (2020) proposed to recognize surgical actions at a fine-grained level directly from laparoscopic cholecystectomy videos, modeling them as triplets of the used instrument, its role (verb), and its underlying target anatomy. Such fine-grained activity recognition gives a detailed understanding of the image contents in laparoscopic videos.

2.2. Surgical Action Triplet Recognition

In the existing surgical ontology, an action is described as a triplet of the used instrument, a verb representing the action performed, and the anatomy acted upon (Neumuth *et al.*, 2006; Katić *et al.*, 2014). Earlier works such as Katić *et al.* (2014, 2015) used triplet annotation information to improve surgical phase recognition. Recently, Nwoye *et al.* (2020) introduced CholecT40, an endoscopic video dataset annotated with action triplets. Tripnet (Nwoye *et al.*, 2020) is the first deep learning model designed to recognize action triplets directly from surgical videos. The model relies on a class activation guide (CAG) module to detect the verb and target in triplets, leveraging instrument appearance cues. It models the final triplet association by projecting the detected components to a 3D interaction space (3Dis) to learn their association while maintaining a triplet structure. In this paper, we improve on the verb and target detections using an attention mechanism. The triplet dataset (Nwoye *et al.*, 2020) is also expanded and refined.

With fine-grained action recognition now gaining momentum, a recent work in robotic surgery (Xu *et al.*, 2021) extended two robotic surgery datasets, MICCAI's robotic scene segmentation challenge (Allan *et al.*, 2020) and Transoral Robotic Surgery (TORS), with 11 and 5 semantic relationship labels respectively. They in turn proposed a cross-domain method for the two datasets generating surgical captions that are comparable to action triplets.

Detecting multi-object interaction in natural images/videos is widely explored by the research on human-object interaction (HOI) (Hu *et al.*, 2013; Mallya and Lazebnik, 2016) where activities are formulated as triplets of (*human*, *verb*, *object*) (Chao *et al.*, 2015). Detecting or recognizing HOI is enabled by triplet datasets with spatial annotations (e.g. HICO-DET (Chao *et al.*, 2018), VCOCO (Lin *et al.*, 2014)) or simply binary presence labels (e.g. HICO (Chao *et al.*, 2015)). CNN models with simple (Mallya and Lazebnik, 2016) or multi-stream architectures (Chao *et al.*, 2018) are widely used to model human and object detections as well as resolving spatial relationships between them. Considering the often large number of possible combinations, Shen *et al.* (2018) proposed a zero-shot method to predict unseen verb-object pairs at test time.

2.3. Attention Mechanism

Since the advent of the attention mechanism (Bahdanau *et al.*, 2014), many deep learning models have exploited it in various forms: from self (Vaswani *et al.*, 2017) to cross (Mohla *et al.*, 2020), and from spatial (Fu *et al.*, 2019) to temporal (Sankaran *et al.*, 2016). Methods relying on attention mechanisms (Wang *et al.*, 2019; Kolesnikov *et al.*, 2019) are proposed to focus the HOI detection networks only on crucial human and object context features. An action-guided attention mining loss (Lin *et al.*) has also been used in HOI recognition tasks; however, all these attention models rely on expensive *spatial* annotations.

Recently, Ji *et al.* (2019) proposed a form of attention that rely on saliency features without requiring additional supervision. While Ji *et al.* (2019) used a combination of spatial and textual attention modules to capture fine-grained image-sentence correlations, another work by Yao and Gong (2020) utilized image saliency to guide an attention network for weakly supervised object segmentation. In medical imaging, Attention U-Net (Oktay *et al.*, 2018) is used to focus on target structures for pancreas segmentation.

Action triplets are instrument-centric: the instrument is the verb's subject, and a visible anatomical part is only considered a target if an instrument operates on it; therefore learning the verb and target are conditioned on the instrument's presence and position. Nwoye *et al.* (2020) addressed this with an activation guide layer named CAG, where the verb and target features are each attuned to instrument activation maps. Even for HOI detection, which is human-centric, human appearance cues are leveraged to predict action-specific densities over target object locations, albeit fully supervised on human bounding boxes (Gkioxari *et al.*, 2018). Ulutan *et al.* (2020) opined that attention modeling is superior to feature concatenation in terms of spatial reasoning. We improve on the CAG principle with a class activation-guided attention mechanism (CAGAM) achieved by redesigning the saliency-guided attention mechanism in (Ji *et al.*, 2019; Yao and Gong, 2020) to utilize a more adequate and easier to learn class activation map (CAM). Our implementation combines both channel and position attention mechanisms for the verb and target detections respectively.

The Transformer model (Vaswani *et al.*, 2017) introduced in NLP shows that attention can be expanded to capture long-range dependencies without recurrence. The Vision Transformer (Dosovitskiy *et al.*, 2020) explored this technique for

image understanding with encouraging performance. Another Transformer with end-to-end self-attention (Zou et al., 2021; Kim et al., 2021) modeled long-range attentions for both HOI components detection and their interaction association. In surgical data science, transformers have been explored for surgical instrument classification (Kondo, 2020) and recently for phase recognition (Gao et al., 2021; Czempiel et al., 2021). Similarly, we propose *Rendezvous* (RDV), a transformer-inspired method, for online surgical action triplet recognition. The novelty of RDV is found in the powerful way it incorporates self- and cross-attentions in its multi-head layers to decode the interactions between the detected instruments and tissues in a laparoscopic procedure.

The Transformer, as used in Natural Language Processing (Vaswani et al., 2017), learns attention maps over a *temporal sequence*, considering a sentence to be a sequence of words. In computer vision, many works have tried to replicate this by modeling input sequence over temporal frames (Girdhar et al., 2019). A single image, however, can be as informative as a complete sentence. Even the Vision Transformer (Dosovitskiy et al., 2020) shows that an image is equivalent to 16×16 words, modeled as a *sequence of patches* from a single image. Many works have followed similar approaches in image understanding (Dosovitskiy et al., 2020), object detection (Carion et al., 2020), segmentation (Chen et al., 2021; Valanarasu et al., 2021), captioning (Liu et al., 2021; Sundaramoorthy et al., 2021), and activity recognition (Bertasius et al., 2021; Gavriljuk et al., 2020). Alternatively, the hybrid architecture of the Vision Transformer (Dosovitskiy et al., 2020) shows that, aside from the raw image, the input sequence can also be obtained from CNN features. It also shows that a patch can have a 1×1 spatial size which is akin to using an image with no explicit sequence modeling. We propose another hybrid approach to obtain the appropriate feature, one that can preserve the spatial and class-wise relationships of the interacting triplet components in a surgical image frame, a *semantic sequence* in this regard. While our attention input features are extracted from a CNN as done in the hybrid Vision Transformer, our attention is design to leverage the CNN’s features in a manner that helps the attention network benefit from the learned class representations. This means we preserve the spatial relationship in the grid of features without breaking it up into patches. This provides further insight into the decision-making of attention networks.

3. CholecT50: Cholecystectomy Action Triplet Dataset

CholecT50 is a dataset of endoscopic videos of laparoscopic cholecystectomy surgery introduced to enable research on fine-grained action recognition in laparoscopic surgery. It is annotated with triplet information in the form of $\langle instrument, verb, target \rangle$. The dataset is a collection of 50 videos consisting of 45 videos from the Cholec80 dataset (Twinanda et al., 2017) and 5 videos from an in-house dataset of the same surgical procedure. It is an extension of CholecT40 (Nwoye et al., 2020) with 10 additional videos and standardized classes.

The cholecystectomy recordings were annotated by two surgeons using the software *Surgery Workflow Toolbox-Annotate*

Table 1: Statistics of the triplet’s component labels in the dataset

instrument		Verb		Target	
Label	Count	Label	Count	Label	Count
bipolar	6697	aspirate	3122	abd-wall/cavity	847
clipper	3379	clip	3070	adhesion	228
grasper	90969	coagulate	5202	blood-vessel	416
hook	52820	cut	1897	cystic-artery	5035
irrigator	5005	dissect	49247	cystic-duct	11883
scissors	2135	grasp	15931	cystic-pedicle	299
		irrigate	572	cystic-plate	4920
		null-verb	10841	fluid	3122
		pack	328	gallbladder	87808
		retract	70795	gut	719
				liver	17521
				null-target	10841
				omentum	9220
				peritoneum	1227
				specimen-bag	6919

from the B-com institute⁶. Annotators set the beginning and end on a timeline for each identified action, then assigned to the corresponding *instrument*, *verb* and *target* class labels. An action ends when the corresponding instrument exits the frame, or if the verb or target changes. Out-of-frame actions are not reported, and video frames that are recorded outside the patient’s body are zeroed out.

We then define classes for the triplet. Due to the number of instruments, verbs, and targets available, the theoretical number of all possible triplet configurations (900) is prohibitively high. Even limiting those configurations to the approximately 300 observed in the dataset has little clinical relevance due to the presence of many spurious classes. To have a reasonable number of classes with maximum clinical utility, a team of clinical experts selected the top relevant labels for the triplet dataset. This is achieved in two steps. In the first instance, class grouping (\cup) is carried out to super-class triplets that are semantically the same. Some examples of triplets grouped include:

- $\langle irrigator, aspirate, bile \rangle \cup \langle irrigator, aspirate, fluid \rangle \cup \langle irrigator, aspirate, blood \rangle \rightarrow \langle irrigator, aspirate, fluid \rangle$
- $\langle grasper, pack, gallbladder \rangle \cup \langle grasper, store, gallbladder \rangle \rightarrow \langle grasper, pack, gallbladder \rangle$
- $\langle grasper, retract, gut \rangle \cup \langle grasper, retract, duodenum \rangle \cup \langle grasper, retract, colon \rangle \rightarrow \langle grasper, retract, gut \rangle$
- $\langle bipolar, coagulate, liver \rangle \cup \langle bipolar-grasper, coagulate, liver \rangle \cup \langle bipolar, coagulate, liver-bed \rangle \rightarrow \langle bipolar, coagulate, liver \rangle$
- $\langle grasper, grasp, gallbladder-fundus \rangle \cup \langle grasper, grasp, gallbladder-neck \rangle \cup \langle grasper, grasp, gallbladder \rangle \cup \langle grasper, grasp, gallbladder-body \rangle \rightarrow \langle grasper, grasp, gallbladder \rangle$

In addition to class grouping, surgical relevance rating and label mediation of the annotated data are carried out by three clinicians. For the rating, the clinicians assigned a score from a range of [1-5] to each triplet composition based on their possibility and usefulness in the considered procedure. Their average scores, as well as the triplet’s number of occurrences, is used to order the triplet classes, after which the top relevant classes are

⁶<https://b-com.com/>

Table 2: Dataset statistics showing the number of occurrences of the triplets

Name	Count	Name	Count	Name	Count
bipolar,coagulate,abdominal-wall/cavity	434	grasper,grasp,cystic-artery	76	hook,dissect,gallbladder	29292
bipolar,coagulate,blood-vessel	251	grasper,grasp,cystic-duct	560	hook,dissect,omentum	3649
bipolar,coagulate,cystic-artery	68	grasper,grasp,cystic-pedicle	26	hook,dissect,peritoneum	337
bipolar,coagulate,cystic-duct	56	grasper,grasp,cystic-plate	163	hook,null-verb,null-target	4397
bipolar,coagulate,cystic-pedicle	77	grasper,grasp,gallbladder	7381	hook,retract,gallbladder	479
bipolar,coagulate,cystic-plate	410	grasper,grasp,gut	33	hook,retract,liver	179
bipolar,coagulate,gallbladder	343	grasper,grasp,liver	83	irrigator,aspirate,fluid	3122
bipolar,coagulate,liver	2595	grasper,grasp,omentum	207	irrigator,dissect,cystic-duct	41
bipolar,coagulate,omentum	262	grasper,grasp,peritoneum	380	irrigator,dissect,cystic-pedicle	89
bipolar,coagulate,peritoneum	73	grasper,grasp,specimen-bag	6834	irrigator,dissect,cystic-plate	10
bipolar,dissect,adhesion	73	grasper,null-verb,null-target	4759	irrigator,dissect,gallbladder	29
bipolar,dissect,cystic-artery	187	grasper,pack,gallbladder	328	irrigator,dissect,omentum	100
bipolar,dissect,cystic-duct	183	grasper,retract,cystic-duct	469	irrigator,irrigate,abdominal-wall/cavity	413
bipolar,dissect,cystic-plate	54	grasper,retract,cystic-pedicle	41	irrigator,irrigate,cystic-pedicle	29
bipolar,dissect,gallbladder	353	grasper,retract,cystic-plate	1205	irrigator,irrigate,liver	130
bipolar,dissect,omentum	176	grasper,retract,gallbladder	48628	irrigator,null-verb,null-target	573
bipolar,grasp,cystic-plate	8	grasper,retract,gut	686	irrigator,retract,gallbladder	30
bipolar,grasp,liver	95	grasper,retract,liver	13646	irrigator,retract,liver	350
bipolar,grasp,specimen-bag	85	grasper,retract,omentum	4422	irrigator,retract,omentum	89
bipolar,null-verb,null-target	632	grasper,retract,peritoneum	289	scissors,coagulate,omentum	17
bipolar,retract,cystic-duct	8	hook,coagulate,blood-vessel	57	scissors,cut,adhesion	155
bipolar,retract,cystic-pedicle	9	hook,coagulate,cystic-artery	10	scissors,cut,blood-vessel	21
bipolar,retract,gallbladder	32	hook,coagulate,cystic-duct	41	scissors,cut,cystic-artery	613
bipolar,retract,liver	164	hook,coagulate,cystic-pedicle	15	scissors,cut,cystic-duct	808
bipolar,retract,omentum	69	hook,coagulate,cystic-plate	9	scissors,cut,cystic-plate	20
clipper,clip,blood-vessel	51	hook,coagulate,gallbladder	217	scissors,cut,liver	90
clipper,clip,cystic-artery	1097	hook,coagulate,liver	189	scissors,cut,omentum	27
clipper,clip,cystic-duct	1856	hook,coagulate,omentum	78	scissors,cut,peritoneum	56
clipper,clip,cystic-pedicle	13	hook,cut,blood-vessel	15	scissors,dissect,cystic-plate	12
clipper,clip,cystic-plate	53	hook,cut,peritoneum	92	scissors,dissect,gallbladder	52
clipper,null-verb,null-target	309	hook,dissect,blood-vessel	21	scissors,dissect,omentum	93
grasper,dissect,cystic-plate	78	hook,dissect,cystic-artery	2984	scissors,null-verb,null-target	171
grasper,dissect,gallbladder	644	hook,dissect,cystic-duct	7861		
grasper,dissect,omentum	31	hook,dissect,cystic-plate	2898	Total	161005

Table 3: Statistics of the dataset split

Data split	Videos	Frames	Labels
Training	35	72815	113884
Validation	5	6797	10267
Testing	10	21251	36854
Total	50	100863	161005

selected. Moreso, the third clinician performed label mediation in the case of label disagreement.

The final dataset comprises 100 triplet classes that follow the format of $\langle instrument, verb, target \rangle$. The triplets are composed from 6 instruments, 10 verbs and 15 target classes, presented with their instance counts in Table 1. We present the CholecT50 dataset triplet labels including their number of occurrences in Table 2. We also present the co-occurrence statistics for $\langle instrument, target \rangle$ and $\langle instrument, verb \rangle$ pairs within triplets in the supplementary material.

For our experiment, we down-sampled the videos to 1fps yielding 100.86K frames annotated with 161K triplet instances. The video dataset is split into training, validation, and testing sets as in Table 3. The videos in the dataset splits are distributed

in the same ratio to include annotations from each surgeon.

4. Methodology

Action triplet recognition is a complex and challenging task, since it requires: (1) simultaneously solving three multi-label classification problems, and (2) performing associations while accounting for multiple triplet instances. In this work, we propose two methods that tackle each aspect of these tasks.

We address the first point with the *class activation guided attention mechanism* or CAGAM, which explicitly uses tool type and location information to highlight discriminative features for verbs and targets respectively. We demonstrate its utility by replacing our previous Tripnet (Nwoye et al., 2020) model’s class activation guide (CAG) with CAGAM, resulting in a preliminary model which we call *Attention Tripnet*.

The second point is addressed by the *multi-head of mixed attention* (MHMA), as an advanced model of semantic attention for triplet association, and a successor to the previous state-of-the-art Tripnet’s more primitive 3D interaction space (3Dis). The MHMA resolves the triplet’s components association using multiple heads of self and cross attention mechanisms.

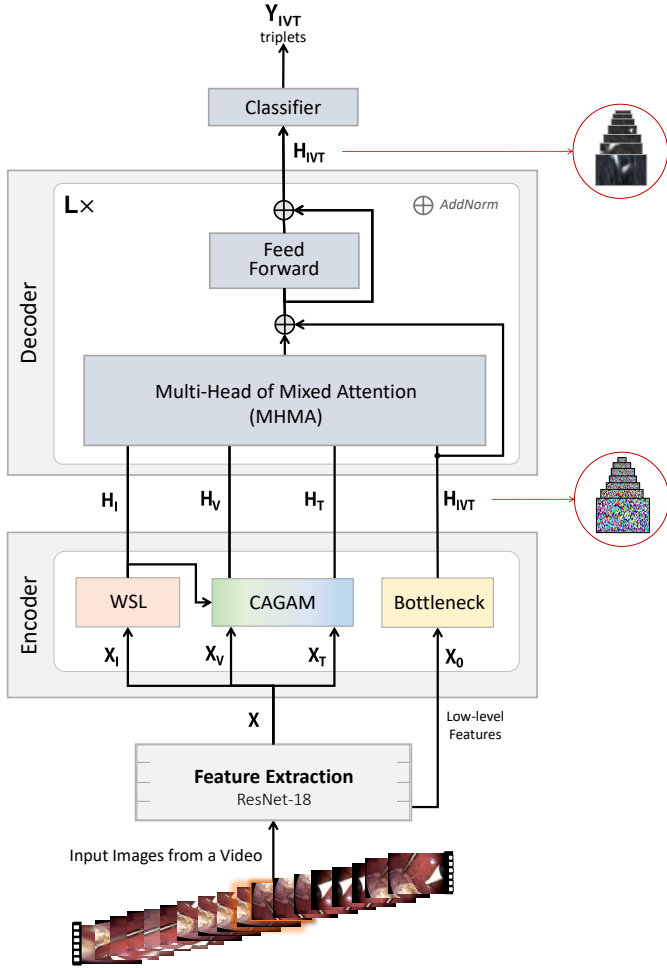


Fig. 2: Architecture of the *Rendezvous: meeting of attentions*, a transformer-inspired model with channel and position spatial attention for triplet components detection and a multi-head of self and cross semantic attention for action triplet recognition.

Our final model is called the *Rendezvous* (RDV): a transformer-inspired neural network for surgical action triplet recognition. The RDV combines the CAGAM in its encoder with the MHMA in its Transformer-inspired decoder for enhanced triplet component detection and association respectively. This model provides the highest performance on action triplet recognition.

The proposed RDV network is conceptually divided into four segments: feature extraction backbone, encoder, decoder, and classifier as shown in Fig. 2.

4.1. Feature Extraction

We model the visual feature extraction using the ResNet-18 base model. Our choice is motivated by the excellent performance of residual networks in visual object classification tasks. To facilitate more precise localization, the strides of the last two blocks of the ResNet are lowered to one pixel providing higher output resolution.

The Resnet-18 base model takes an RGB image frame from a video as input and extracts its visual features $\mathbf{X} \in \mathbb{R}^{32 \times 56 \times 512}$. The extracted feature is triplicated into $(\mathbf{X}_I, \mathbf{X}_V, \mathbf{X}_T)$ for multi-

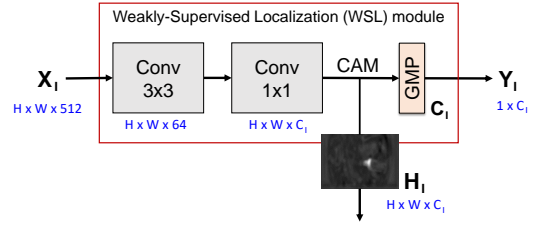


Fig. 3: Weakly supervised localization (WSL) layer for instrument detection. Feature dimension (height $H = 32$, width $W = 56$, depth (class size) $C_I = 6$).

task learning of the instrument, verb, and target components of the triplets respectively.

4.2. Components Encoding

The encoder is responsible for detecting the various components of the triplets, while the decoder resolves the relationships between them. The encoder is composed of the weakly-supervised localization (WSL) module for instrument detection, class activation guided attention mechanism (CAGAM) module for verb and target recognition, and a bottleneck layer collecting unfiltered low-level features from Resnet-18's lower layer.

4.2.1. Weakly Supervised Localization (WSL)

While this work primarily focuses on *recognizing* surgical action triplets, *localizing* actions -similarly to HOI tasks- is an interesting addition. We therefore go beyond simply detecting the presence of surgical instruments by locating their position, which represents the region of interaction. In the absence of spatial annotations we achieve this with weak supervision.

As shown in Fig. 3, the WSL module consists of a 3×3 convolution layer (Conv) of 64 channels, then followed by a 1×1 Conv of $C_I = 6$ channels for instrument localization in form of class activation maps (CAM).

Specifically, the WSL module takes \mathbf{X}_I from the feature extraction layer as input and returns the instruments' CAM, marked as (\mathbf{H}_I) , from its last Conv layer. **The output CAM (\mathbf{H}_I) are trained for localization via their Global Maximum Pooled (GMP) values \mathbf{Y}_I representing instrument class-wise presence probabilities similar to Vardazaryan et al. (2018).**

The discriminative CAM features (\mathbf{H}_I) alongside these remaining extracted features $(\mathbf{X}_V, \mathbf{X}_T)$ are passed to the CAGAM for verb and target detection.

4.2.2. Class Activation Guided Attention Mechanism (CAGAM)

Surgical action triplets are instrument-centric. Detecting the correct verbs and target anatomies is very challenging, because the visibility as well as the subtly involvement of a tool and anatomy in an action have to be taken into consideration.

A limited effort is made in our previous method, Tripnet (Nwoye et al., 2020), to handle this using a CAG module conditioning the detection of verbs and targets on the instruments activations, via concatenated features. Since attention modeling is found to be superior to feature concatenation (Ulutun et al., 2020), we explore several types of attention and propose a new form of spatial attention, named CAGAM.

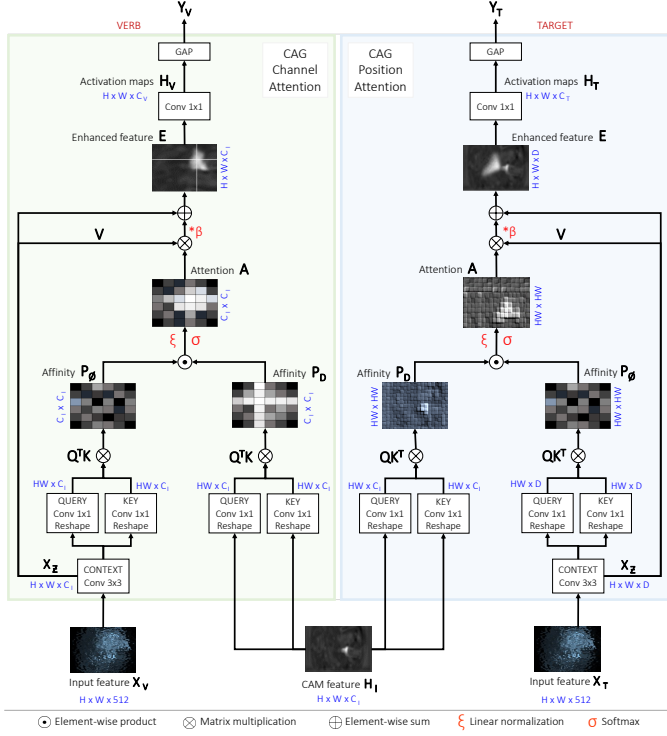


Fig. 4: Class Activation Guided Attention Mechanism (CAGAM): uses the attention learned from the instrument’s CAM to highlight the verb class (left) and the anatomy in contact with the instrument (right). Feature dimension (height $H = 32$, width $W = 56$, depth $D = 64$, instrument’s class size $C_I = 6$, verb’s class size $C_V = 10$, target’s class size $C_T = 15$).

According to Vaswani *et al.* (2017), an attention function can be described as matching a query (\mathbf{Q}) and a set of key-value (\mathbf{K}, \mathbf{V}) pairs to form an output. The output is computed as a weighted sum of the values ($w\mathbf{V}$), where the weight ($w = \mathbf{Q}\mathbf{K}^T$) is computed by an affinity score function of the query with the corresponding key. CAGAM is a new form of spatial attention mechanism that propagates attention from known to unknown context features, thereby enhancing the unknown context for relevant pattern discovery. It is an adaptation of the saliency-guided attention mechanism in (Ji *et al.*, 2019; Yao and Gong, 2020) to utilize a more adequate and easier to learn class activation map (CAM) suitable for action triplet recognition. It is used, in this case, to discover the verbs and targets that are involved in tool-tissue interactions leveraging the instrument’s contextual dependencies, by propagating attention from the discriminative \mathbf{H}_I to the non-discriminative \mathbf{X}_V and \mathbf{X}_T features. The CAM (\mathbf{H}_I) serves as the known context features in this regard, since they are already discriminated class-wise and localized for the instruments.

As shown in Fig. 4, we model the CAGAM to enhance the verb’s and target’s unfiltered features by element-wise addition of an *enhancement*: this enhancement is a computed spatial attention \mathbf{A} from the instrument affinity maps (\mathbf{P}_D) as well as the component affinity maps (\mathbf{P}_0) themselves. The \mathbf{P}_D are termed discriminative because they originate from the instrument CAM features, whereas \mathbf{P}_0 are termed non-discriminative because they are formed from the unfiltered component features.

We observe that verbs and targets behave differently with regards to their instrument; that is, verbs are mostly affected by the instrument’s type, while targets tend to be determined by instrument’s position. This distinction is a key factor in the choices of attention mechanism in the CAGAM which indeed combines **channel attention** for verb detection (Fig. 4: left) and **position attention** for target detection (Fig. 4: right). Both types of spatial attention mechanisms are similar, except for the dimensions used, and therefore the nature of the information attended to. The channel attention is captured in the $C_I \times C_I$ channel dimensions, informed by instrument type, whereas the position attention is captured in the $HW \times HW$ spatial dimensions, informed by instrument location. This choice is well validated in ablation studies shown further (Table 4).

CAG channel attention for verbs

As illustrated in Fig. 4 (left), verb features are first remapped to $\mathbf{X}_Z \in \mathbb{R}^{H \times W \times C_I}$, which we call the *context features*. Following two separate 1×1 Conv and reshaping, mapping it to a query \mathbf{Q} and a key \mathbf{K} of size $HW \times C_I$, the *non-discriminative affinity map* $\mathbf{P}_0 \in \mathbb{R}^{C_I \times C_I}$ is obtained via matrix multiplication of the transposed \mathbf{Q} by \mathbf{K} as illustrated in Equation 1:

$$\mathbf{P}_0 = \mathbf{Q}^T \mathbf{K}. \quad (1)$$

Applying a similar process to the CAM results in the *discriminative affinity map*: $\mathbf{P}_D \in \mathbb{R}^{C_I \times C_I}$. As done in Yao and Gong (2020), an element-wise product of the two affinity maps, scaled by a factor ξ and passed through red *softmax* (σ) gives the attention \mathbf{A} :

$$\mathbf{A} = \sigma \left(\frac{\mathbf{P}_D \mathbf{P}_0}{\xi} \right). \quad (2)$$

Meanwhile, we obtain the value features $\mathbf{V} \in \mathbb{R}^{HW \times C_I}$ by reshaping the verb context \mathbf{X}_Z to $\mathbb{R}^{HW \times C_I}$. Next, we obtain an enhancement by matrix multiplication of \mathbf{A} by \mathbf{V} , weighted by a learnable temperature β . This enhancement is reshaped to $\mathbb{R}^{H \times W \times C_I}$ and added back to \mathbf{X}_Z to produce the enhanced features, \mathbf{E} .

$$\mathbf{E} = \beta(\mathbf{V}\mathbf{A}) + \mathbf{X}_Z. \quad (3)$$

The features \mathbf{E} are transformed into per-verb activation maps $\mathbf{H}_V \in \mathbb{R}^{H \times W \times C_V}$ via a 1×1 Conv. Finally, verb logits $\mathbf{Y}_V \in \mathbb{R}^{1 \times C_V}$ are obtained by global average pooling of \mathbf{H}_V , where $C_V = 10$ is the number of verb classes.

CAG position attention for targets

As illustrated in Fig. 4 (right), obtaining the \mathbf{Q} , \mathbf{K} , and \mathbf{V} terms for the CAG position attention is similar to the CAG channel attention mechanism. However to obtain an instrument location-aware attention, we multiply \mathbf{Q} by \mathbf{K}^T (instead of \mathbf{Q}^T by \mathbf{K} as done for verbs in Equation 1) producing affinity maps ($\mathbf{P}_D, \mathbf{P}_0$) and a subsequent attention map \mathbf{A} of the desired size $HW \times HW$, informed by instrument position rather than instrument type.

Furthermore, we obtain enhanced target features (\mathbf{E}), which we also feed to a 1×1 Conv of $C_T = 15$ channels to obtain the per-target activation maps $\mathbf{H}_T \in \mathbb{R}^{H \times W \times C_T}$. Using a global pooling on \mathbf{H}_T , we then obtain the target logits $\mathbf{Y}_T \in \mathbb{R}^{1 \times C_T}$.

To ensure that the $\mathbf{H}_I, \mathbf{H}_V$ and \mathbf{H}_T class maps properly capture their corresponding components, we train their global pooled logits ($\mathbf{Y}_I, \mathbf{Y}_V, \mathbf{Y}_T$) as auxiliary classification tasks.

4.2.3. Bottleneck Layer

In addition to these refined, component-specific features ($\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T$), a global context feature is also necessary for modeling their contextual relationship; which is why we also draw a unfiltered low-level feature \mathbf{X}_0 from the first block of ResNet and feed it to the bottleneck layer that consists of $3 \times 3 \times 256$ and $1 \times 1 \times C$ convolution layers, where $C = 100$ is the number of triplet classes. This gives the global context feature for triplets \mathbf{H}_{IVT} , with channels matched to the triplet classes.

The unfiltered triplets feature \mathbf{H}_{IVT} as well as the triplet component's class maps ($\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T$) are fed to the decoder layer for decoding the triplet association.

4.3. Interaction Decoding

We describe here the modeling of the triplet components' relationship in the RDV decoder. An existing approach attempts to model every triplet possibility from the outer product combination of the three components using a 3D feature space (Nwoye et al., 2020). This design models more than the required triplets, including irrelevant and impossible combinations, making the module hard to train. Hence, we follow a Transformer-like architecture (Vaswani et al., 2017; Dosovitskiy et al., 2020; Chen et al., 2021) leveraging long-range attention to efficiently model the required relationships. To take into consideration the constituting components of the triplets (Nwoye et al., 2020), we utilize the semantic features of each component, captured in their class maps (H_I, H_V, H_T). Unlike the Vision Transformer (Dosovitskiy et al., 2020), however, we do not break class maps into patches. As shown by ablation results in the supplementary material, the patch sequence degrades representations, especially information on instruments that is important for locating actions.

Hence, we model the RDV attention decoder on the semantic sequence of learnt class-wise representations. From $\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T$ and the global triplet feature \mathbf{H}_{IVT} , RDV decodes all the self- and cross-interactions between the triplet's global context feature and the three features corresponding to individual components, using scaled dot-product attention (Vaswani et al., 2017) without using recurrence. In addition to self-attention, cross-attention adds the capability to better model the relationships with components participating in the action triplet. This is important when resolving interactions: for instance, an anatomical part can appear in the frame without being a target, often making the interaction with the instrument ambiguous.

To understand the attention decoder used in this work, we explain the **decoding-by-attention** concept below:

1. Firstly, attention decoding is described as a search process whereby a query (\mathbf{Q}), that is issued by a user (*sink* or *receiver*), is used to retrieve data from a repository (*source*). Normally, \mathbf{Q} is a user's abridged description of the requested data also known as *search terms*.

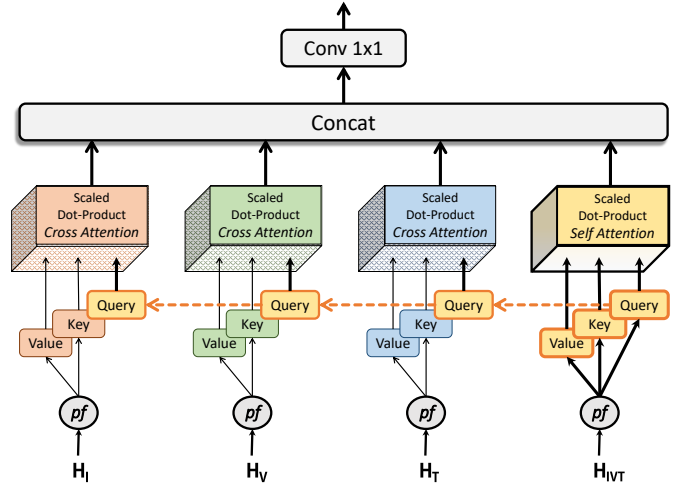


Fig. 5: Architecture of the multi-head of mixed attention (MHMA): showing the feature projection into Q, K and V, and subsequent multiple heads of self and cross attentions using scale-dot product attention mechanism.

2. The source context consists of a key-value (\mathbf{K}, \mathbf{V}) pair where \mathbf{V} is a collection of several data points or *records* and \mathbf{K} is the mean descriptor for each record also known as *keywords*.
3. To retrieve the requested data, the issued \mathbf{Q} is matched with the available \mathbf{K} s to create an *affinity* (\mathbf{P}), also known as the *attention weight*.
4. The \mathbf{P} , when matched with \mathbf{V} , creates an *attention map* (\mathbf{A}) which helps retrieve the most appropriate data to the sink.

We implement a transformer-inspired decoder that is composed of a stack of $L = 8$ identical layers as shown in Fig. 2. Each layer receives the triplet features \mathbf{H}_{IVT} and the encoded class maps ($\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T$) as inputs which are processed successively by its two internal modules: MHMA and feed-forward, to produce refined triplet features, \mathbf{H}_{IVT} . The output of each module is followed by a residual connection and a layer normalization (AddNorm) as it is done in other multi-head attention networks. The entire cycle repeats, with a more refined \mathbf{H}_{IVT} output, until the L^{th} layer.

4.3.1. Multi-head of Mixed Attention (MHMA)

The multi-head attention combines both self- and cross-attentions, encouraging high-level learning of triplets from the interacting components as shown in Fig. 5. It starts with a projection function, pf , which generates a set of value \mathbf{V} , key \mathbf{K} , and/or query \mathbf{Q} for each context feature ($\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T, \mathbf{H}_{IVT}$). In the implementation as shown in Equation 4, the pf function generates vectors of $\mathbf{Q} \in \mathbb{R}^{1 \times C}$ and $\mathbf{K} \in \mathbb{R}^{1 \times C_z}$ that represent the abridged mean descriptors of the contexts by leveraging the global average pooling (GAP) operation. Here, $C = 100$ for triplet, whereas $C_z = [6, 10, 15, 100]$ for either instrument, verb, target, or triplet classes, respectively. We map each descriptor to a feature embedding layer where we mask (dropout $\lambda = 0.3$) parts of \mathbf{Q} to avoid repeating the same query in the L alternating layers. Using the pf function, we also obtain the $\mathbf{V} \in \mathbb{R}^{H \times W \times C_z}$ by a convolution operation on the feature context

and reshape to $\mathbb{R}^{HW \times Cz}$. Hence, the extracted \mathbf{Q} , \mathbf{K} , and \mathbf{V} features follow the aforementioned decoding-by-attention concept (items 1 & 2). The pf function generates each \mathbf{K} and \mathbf{Q} using FC layers as done in (Vaswani et al., 2017; Dosovitskiy et al., 2020), and generates the \mathbf{V} using convolution layers as done in (Fu et al., 2019; Wang et al., 2018; Huang et al., 2019).

$$pf(H) = \begin{cases} Q: & FC(DROPOUT(GAP(H))), \\ K: & FC(GAP(H)), \\ V: & CONV(H). \end{cases} \quad (4)$$

Next, we build 4 attention heads for the instrument, verb, target, and triplet attention features. In the existing Transformer and Transformer-based models, each of the heads learns a self-attention. Self-attention helps a model understand the underlying meaning and patterns within its own feature representation. This is needed for initial scene understanding. However, when each feature representation (such as a class-map) has been discriminated to attend to only one component in an image scene, understanding their underlying relationship requires a cross-attention across the component features. In a cross-attention mechanism, the attention built from one context (the *source*) is used to highlight features in another context (the *sink*) as done in Mohla et al. (2020). While the self-attention mechanism computes the focal representation on the same triplet features, cross attentions learn the triplet representations by drawing attention from the individual components: namely instrument, verb, and target. This models how the features of each component affect the triplet composition, by propagating the affinities from their respective context features to the required triplet features.

To utilize both self and cross attentions, we model the source context from the encoded class-map features ($\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T$) representing the triplet components and the sink context from the triplet features (\mathbf{H}_{IVT}). Of course, the source context remains the same as the sink in the self-attention mechanism. This means we generate the corresponding \mathbf{K} s and \mathbf{V} s from both the source and sink contexts, but generate the \mathbf{Q} only from the sink context using the projection function, pf , as shown in Fig. 5. With \mathbf{Q} coming from the triplet features, we actually focus the image understanding on the actions of interest by pointing the cross-attention heads at the component's discriminative features ($\mathbf{H}_I, \mathbf{H}_V, \mathbf{H}_T$) in a manner that helps the attention network benefit from the learnt class representations. This also respects the aforementioned decoding-by-attention concept. We then learn a scaled dot product attention of the \mathbf{Q} on the (\mathbf{K}, \mathbf{V}) pair for each attention head as shown in Fig. 6. Specifically, we derive the scaled dot product attention using the widely used attention formula (Vaswani et al., 2017) in Equation 5:

$$\mathbf{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V} \cdot \sigma \left(\frac{\mathbf{K}\mathbf{Q}^T}{\sqrt{d_{\mathbf{K}}}} \right), \quad (5)$$

where σ is a softmax activation function, $\sqrt{d_{\mathbf{K}}}$ is a scaling factor, and $d_{\mathbf{K}}$ is the dimension of \mathbf{K} after linear transformation. The cross attention is implemented on the instrument, verb, and target attention heads, whereas self-attention is implemented on

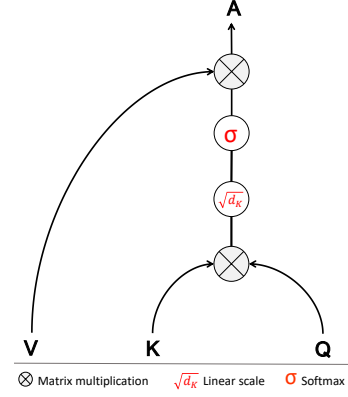


Fig. 6: Structure of scale dot-product attention mechanisms: in self-attention, the $(\mathbf{K}, \mathbf{V}, \mathbf{Q})$ triple comes from one feature context, whereas in cross-attention, the (\mathbf{K}, \mathbf{V}) pair comes from the source feature context while \mathbf{Q} comes from the sink feature context.

the triplet attention head. While each attention head simultaneously concentrates on its own features of interest, the multi-head module combines heads $\mathbf{A}_{1..N}$ to jointly capture the triplet features as in Equation 6:

$$\mathbf{A}_{1..N} = \mathbf{W} \left(\parallel_{i=1}^N \mathbf{A}_i \right), \quad (6)$$

where \parallel is a concatenation operator for $N = 4$ attention heads. \mathbf{A}_1 is the triplet self-attention, $\mathbf{A}_{2..N}$ are the triplet cross attentions with the interacting components. \mathbf{W} is the matrix of convolution weights. This packed convolution scheme merges the information from all attention heads while preserving its spatial structure.

4.3.2. Feed-forward

The output of the multi-head attention is further refined by a feed-forward layer which is a stack of 2 convolutions with an AddNorm. The output is a refined \mathbf{H}_{IVT} with each channel attending to each triplet class.

4.4. Triplet Classification

The RDV model terminates with a linear classifier for the final classification of the triplets. In this layer, we apply a global pooling operation on the \mathbf{H}_{IVT} from the L^{th} layer of the RDV decoder, followed by an FC-layer (with $C = 100$ neurons) for the triplet classification. The output logits (\mathbf{Y}_{IVT}) are trained jointly end-to-end with the auxiliary logits from the encoder.

4.5. Attention Tripnet

In our previous work (Nwoye et al., 2020), Tripnet relies on two modules: (1) the class activation guide (CAG), which leverages instrument activations to detect verbs and targets via concatenated features, and (2) the 3D interaction space (3Dis), where features corresponding to the three components are projected in an attempt to resolve their interactions.

As an ablation model, we extend this to *Attention Tripnet* by only replacing the CAG in Tripnet (Nwoye et al., 2020) with CAGAM as shown in Fig. 7. This validates the contribution of attention modeling for verb and target detections using Attention Tripnet.

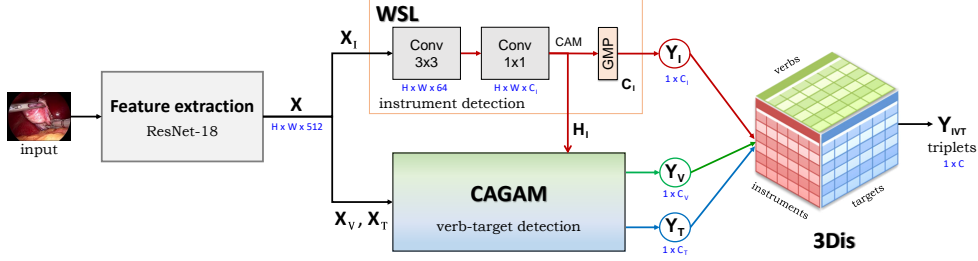


Fig. 7: Architecture of the Attention Tripnet showing the base (feature extraction backbone), neck (instrument detection branch and CAGAM module), and head (3D interaction space). *Feature dimension values* ($H = 32$, $W = 56$, $C_I = 6$, $C_V = 10$, $C_T = 15$, $C = 100$)

5. Experiments

5.1. Data Setup

Due to variability in the video dataset, frame resolution varies from 480×854 to 1080×1920 . We unified these spatial dimensions by resizing them to 256×448 . We also employed random scaling $[0.5, 1.5]$ and brightness/contrast shift ($\delta = 0.2$) as data augmentation for training. The models are trained on 35 videos, validated on 5 videos, and tested on 10 videos according to the data split in Table 3. To obtain specific labels for the component tasks, we design a mapping function, which extracts per-component labels from the triplet labels; those are three vectors of binary presence labels with length $N = [6, 10, 15]$ per frame, where $n \in N$ is the class size for each triplet's component trained as auxiliary task. For a high-performance data loading pipeline, we store our training data as serialized TFRecords binaries.

5.2. Training and Loss Functions

Since classifying each triplet component, namely instrument, verb, and target, is a multi-label classification problem, we employ weighted sigmoid cross-entropy losses: L_I , L_V , and L_T respectively. The weighted cross-entropy with logits is as follows:

$$L = \sum_{c=1}^C \frac{-1}{N} (W_c y_c \log(\sigma(\hat{y}_c)) + (1 - y_c) \log(1 - \sigma(\hat{y}_c))), \quad (7)$$

where y_c and \hat{y}_c are respectively the ground truth and predicted labels for class c , σ is the sigmoid function, and W_c is a weight for class balancing. The three component detection tasks are jointly learned in a multi-task manner following the uncertainty loss procedure given in Kendall *et al.* (2018) that uses learnable parameters w_I , w_V , w_T to automatically balance the tasks training as follows:

$$L_{comp} = \frac{1}{3} \left(\frac{1}{e^{w_I}} L_I + \frac{1}{e^{w_V}} L_V + \frac{1}{e^{w_T}} L_T + w_I + w_V + w_T \right) \quad (8)$$

This is only used for the auxiliary tasks captured by multi-task learning.

The triplet association loss L_{assoc} is also modeled as a sigmoid cross-entropy. To jointly learn the complete tasks end-to-end, we define the total loss (L_{total}) using the equation:

$$L_{total} = L_{comp} + \rho L_{assoc} + \lambda L_2, \quad (9)$$

where ρ is a warm-up parameter that allows the network to focus solely on learning the individual components' information within the first 18 epochs. $\lambda = 1e^{-5}$ is a regularization weight decay for the L_2 normalization loss.

5.3. Hyper-parameters

The feature extraction backbone is pretrained on ImageNet. All the models are trained using Stochastic Gradient Descent with Momentum ($\mu = 0.95$) as optimizer. We maintain a step-wise learning rate ($\eta = 0.001$) policy, decayed by $\delta = 0.1$ after every 50 epochs. The models are trained in batches of size 8 for 200 epochs. The final model weights are selected based on their validation loss saturation. All the hyper-parameters are tuned on the validation set (5 videos) with up to 74 grid search experiments.

5.4. Hardware and Schedule

Our networks are implemented using TensorFlow and trained on GeForce GTX 1080 Ti, Tesla P40, RTX6000, and V100 GPUs. Full training takes approximately 118-180 hours on a single GTX 1080 Ti. Total storage space consumption for the model, input data, output weights, and summaries is under 10GB. Parameter counts for the MTL baseline, Tripnet, Attention Tripnet, and 8-layer RDV models reach 14.94M, 14.95M, 11.81M, 16.61M respectively.

5.5. Inference and Evaluation Protocols

Model outputs are probability scores that can be thresholded to indicate class presence or absence. We statistically evaluate the model's performance at recognizing surgical actions as a triplet using three metrics:

1. **Component average precision:** This measures the average precision (AP) of detecting the correct components of the triplet, as the area under the precision-recall curve per class. Using this, we measure the AP for instrument (AP_I), verb (AP_V), and target (AP_T) detections. To use these metrics for the naive models or for any model that predicts only the triplet labels \mathbf{Y}_{IVT} , we decompose their predictions into the constituting components ($\mathbf{Y}_I, \mathbf{Y}_V, \mathbf{Y}_T$) following Equation 10:

$$\begin{aligned} Y_I &= [\text{MAX}(Y_{IVT}|I = i) \quad \forall i \in \{0, 1, \dots, C_I\}], \\ Y_V &= [\text{MAX}(Y_{IVT}|V = v) \quad \forall v \in \{0, 1, \dots, C_V\}], \\ Y_T &= [\text{MAX}(Y_{IVT}|T = t) \quad \forall t \in \{0, 1, \dots, C_T\}], \end{aligned} \quad (10)$$

where C_1 , C_2 and C_3 are the class sizes for the instrument, verb, and target components respectively. This directly translates to obtaining the probability of a given component class as the maximum probability value among all triplet labels having the same component class label in a given frame. For instance, the predicted probability of a *grasper* instrument in a frame is the maximum probability of all triplet labels having *grasper* as their instrument component label. The ground truth for these components is also derived in the same manner.

2. **Triplet average precision:** This measures the AP of recognizing the tool-tissue interactions by observing elements of the triplet in conjunction. Using the same metrics as Nwoye et al. (2020), we measure the APs for the instrument-verb (AP_{IV}), instrument-target (AP_{IT}), and instrument-verb-target (AP_{IVT}). During the AP computation, a prediction is registered as correct if all of the components of interest are correctly identified (e.g. instrument and verb for AP_{IV}). The main metric in this study is AP_{IVT} , which evaluates the recognition of the complete triplets.
3. **Top-N recognition performance:** Due to high similarities between triplets, we also measure the ability of a model to predict the exact triplets within its top N confidence scores. For every given test sample x_i , a model made an error if the correct label y_i does not appear in its top N confident predictions \hat{y}_i for that sample. Using this setup, we measure the top-5, top-10, and top-20 accuracies for the triplet prediction. We also show the top 10 predicted triplet class labels and their AP scores for a more insightful analysis of the model’s performance.

Video-specific AP scores are computed per category, across all frames of a given video. Averaging those APs over all videos gives us the mean AP (mAP), serving as our main metric.

6. Results and Discussion

In this section, we rigorously validate individual components of the Attention Tripnet and Rendezvous (RDV) through careful ablation studies. We then provide a comparative analysis with baseline and state-of-the-art (SOTA) methods to show our methods’ superiority.

6.1. Quantitative Results

6.1.1. Ablation study on the encoder’s attention type

Table 4: Ablation study on the task-attention suitability

Guided detection	AP_V	AP_T
None (as in MTL baseline)	48.4	28.2
CAM (as in Tripnet’s CAG)	51.3	32.1
CAM + Channel attention	59.0	31.5
CAM + Position attention	51.2	35.1
CAM + Dual ¹ attention	61.1	40.2

¹ Dual = (channel + position) attentions

We begin with an ablation study for the choice of the attention type in the CAGAM module. We compare the module with

a baseline model (MTL) (Nwoye et al., 2020), which implements a multi-task learning of instruments, verbs, and targets in separate branches with no attention (None), and show that attention guidance helps better detect the components in general (Table 4). We also justify the distinct attention types for verbs and targets. Firstly, the channel attention is used for both verb and target detections (row 3), and the position attention is used for both verb and target detections (row 4), before they are combined (Dual attention) in the last row. Channel attention is better suited for verbs than targets, with +10.6% vs +3.5% improvement respectively. Position attention behaves the opposite: +2.8% vs +6.9%. Matching verbs with channel attention and targets with position attention gives the most balanced and highest improvement: +12.4% verbs, +12.0% targets. We, therefore, retain this choice in the proposed models.

6.1.2. Ablation Study on Decoder’s Attention Type

Table 5: Ablation study on the attention type in the multi-head decoder

Model	Layer size	AP_{IV}	AP_{IT}	AP_{IVT}
Single Self	6	29.8	23.3	18.8
Multiple Self	6	35.7	32.8	26.1
Self + Cross (RDV)	6	39.4	36.9	29.9

One of the novel contributions of this work is its hybrid multi-head attention mechanism for resolving tool-tissue interactions, combining self- and cross-attention. This is a substantial innovation over transformers found in sequence modeling, which instead rely on multi-heads of self-attention only. Our choice of multi-head attention is justified in the following ablation study presented in Table 5.

Our first ablation model in this regards (*Single Self*) uses a multi-head attention with the input feature coming from the high-level features (\mathbf{X}) of ResNet-18 to compute a successive scale dot-product attention over 8 decoder layers as in RDV. It can be observed that using a multi-head of self-attention coming from a single source (triplet features) yields insufficient results for action triplet recognition.

The *Multiple Self* ablation model, as a ”self-attention only” version of the RDV, uses self-attention in all four contexts: instrument, verb, target, and triplet. The RDV clearly performs the best in terms of association, justifying our use of cross-attention.

Table 6: A scalability study on the multi-head layer size: showing the mean average precision (mAP) for varying triplet associations, number of learning parameters (Params) in millions (M), and inference time (i-Time) in frame per seconds (FPS) on GTX 1080 Ti GPU.

Layer size	mAP_{IV} (%) \uparrow	mAP_{IT} (%) \uparrow	mAP_{IVT} (%) \uparrow	Params (M) \downarrow	i-Time (FPS) \uparrow
1	35.8	30.7	24.6	12.6	54.2
2	36.0	41.1	27.0	13.1	47.9
4	38.5	32.9	27.3	14.3	39.2
8	39.4	36.9	29.9	16.6	28.1

Table 7: Performance summary of the proposed models compared to state-of-the-art and baseline models

Method		Component detection			Triplet association		
		AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
Naive Baseline	CNN	57.7	39.2	28.3	21.7	18.0	13.6
	TCN	48.9	29.4	21.4	17.7	15.5	12.4
	MTL	84.5	48.4	28.2	26.6	21.2	17.6
SOTA	Tripnet (Nwoye et al., 2020)	92.1	54.5	33.2	29.7	26.4	20.0
Ours	Attention Tripnet	92.0	60.2	38.5	31.1	29.8	23.4
	Rendezvous	92.0	60.7	38.3	39.4	36.9	29.9

6.1.3. Scalability Study on Multi-Head Layer Size

We carried out a scalability study to observe the performance of the RDV when increasing the number of multi-head layers while keeping track of the number of parameters and GPU requirements. These results presented in Table 6 show that the proposed model improves when scaled up, at the cost of increased computational requirements. To balance performance and resource usage, we choose $L = 8$ as default settings in all our experiments. An 8-layer RDV with > 25 FPS processing speed can be used in real-time for OR assistance.

More ablation studies on the sequence modeling of the class-wise features, use of auxiliary classification loss, etc., are provided in the supplementary material.

6.1.4. Component Detection and Association mAP

For ease of reference, we summarize the overall performance of the experimented models on the considered metrics for both triplet component detection and the recognition of their interactions in Table 7. As a baseline, we design a CNN model that models the triplet recognition as a simple classification of 100 distinct labels without taking any special reference to the constituting components. The performance of this baseline model shows that it is not sufficient to naively classify the triplet IDs without considering the triplet components. Even a temporal refinement of the naive CNN model outputs using a (TCN) (Lea et al., 2016) is still sub-optimal. Multi-task learning (MTL) of the triplet components helps the model gain some performance, but still scores low on triplet association. The MTL outperforms the TCN here likely because a temporal refinement would not matter much if a Naive CNN does not capture significant representative features for triplet recognition. The Tripnet model proposed in (Nwoye et al., 2020) leverages the CAG to improve the MTL in the triplet components detection. It also improves the interaction recognition AP_{IVT} by 2.4% using the 3Dis.

The Attention Tripnet uses the CAGAM to further improve the Tripnet’s verb detection by 5.7% and target detection by 5.3%. The Attention Tripnet is on par for instrument detection AP; this is likely due to the instrument detection being already saturated. The overall performance does increase, with indeed a 3.4% improvement for triplet recognition. The RDV, on the other hand, uses a multi-head attention decoder to further improve the association performance (+9.7% on instrument-verb, +10.5% on instrument-target). It improves the overall final triplet recognition by 9.9% mAP_{IVT} compared to the SOTA,

tripling the improvement from the Attention Tripnet. A breakdown of per-class detection of the triplet components and their association performance is presented in the supplementary material.

6.1.5. Top-N Triplet Recognition Performance

Table 8: Top N Accuracy of the triplet predictions

Method		Top-5	Top-10	Top-20
Naive Baseline	CNN	67.0	80.0	90.2
	TCN	54.5	69.4	84.3
	MTL	70.2	80.2	89.5
SOTA	Tripnet	70.5	81.9	91.4
Ours	Attention Tripnet	75.3	86.0	93.8
	Rendezvous	76.3	88.7	95.9

In our multi-class problem with 100 action triplet classes, getting a comprehensive view of a model’s strength is difficult. Here we focus on the top N predictions. As shown in Table 8, when considering the model’s top 20 predictions, the model records an AP of $\approx 95\%$. The model’s confidence however decreases when considering more top predictions, suggesting how closely related most of the triplet classes could be.

6.1.6. Surgical Relevance of the Top Detected Triplets

The result of the top 10 correctly detected triplets for the experimented models, presented in Table 9, reveals the individual strengths of the experimented models in recognizing the tool-tissue interaction. All triplets predicted in the top results are clinically sensible, with none of the more unexpected instrument-verb or instrument-target pairings.

Of importance, triplets with high surgical relevance in cholecystectomy procedure, i.e., $\langle \text{clipper, clips, cystic duct or artery} \rangle$ and $\langle \text{scissors, cut, cystic duct or artery} \rangle$, which are critical for safety monitoring, are better detected by the RDV than the SOTA. The proposed models learn to detect rare but clinically important uses of surgical instruments in their top 10 correctly predicted labels. This holds true for ambiguous instruments, like the *irrigator* that is mostly used to aspirate or irrigate but can as well be used to dissect in rare cases ($\langle \text{irrigator, dissect, cystic-pedicle} \rangle$). Another detected rare case include $\langle \text{bipolar,}$

Table 9: Top-10 predicted Triplets (AP_{IVT} for Instrument-Verb-Target Interactions).

Tripnet (SOTA)		Attention Tripnet		Rendezvous	
Tripnet	AP	Tripnet	AP	Tripnet	AP
grasper,retract,gallbladder	77.30	grasper,grasp,specimen-bag	82.34	grasper,retract,gallbladder	85.34
grasper,grasp,specimen-bag	76.50	grasper,retract,gallbladder	78.41	grasper,grasp,specimen-bag	81.75
bipolar,coagulate,liver	67.39	bipolar,coagulate,liver	68.85	hook,dissect,gallbladder	75.90
hook,dissect,gallbladder	57.54	irrigator,dissect,cystic-pedicle	66.21	grasper,retract,liver	66.70
irrigator,aspirate,fluid	57.51	hook,dissect,gallbladder	63.22	bipolar,coagulate,liver	63.12
grasper,retract,liver	54.25	grasper,retract,liver	58.06	clipper,clip,cystic-duct	59.68
clipper,clip,cystic-artery	47.44	grasper,grasp,cystic-pedicle	55.35	bipolar,coagulate,blood-vessel	57.18
scissors,cut,cystic-duct	42.57	scissors,cut,cystic-artery	48.44	scissors,cut,cystic-artery	53.84
scissors,cut,cystic-artery	40.37	irrigator,aspirate,fluid	47.11	irrigator,aspirate,fluid	51.95
clipper,clip,cystic-duct	39.62	bipolar,coagulate,abdominal-wall-cavity	46.07	clipper,clip,cystic-artery	51.52
mean	56.05		61.41		64.70

coagulate, blood-vessel). This suggests that the models effectively learned the surgical semantics of instrument usage even with small examples of peculiar classes.

The triplet $\langle \textit{grasper}, \textit{grasp}, \textit{specimen-bag} \rangle$ always appears in the top 2 even though its prevalence (6K) is not particularly high, compared to triplets such as $\langle \textit{hook}, \textit{dissect}, \textit{gallbladder} \rangle$ (29K), $\langle \textit{grasper}, \textit{retract}, \textit{liver} \rangle$ (13K), etc. This may be due to its consistent appearance in the workflow, usually towards the end; another factor could be the discernability of the *bag*.

For every triplet in the top-10 predictions of both the SOTA and the proposed models, the performance is usually higher in the proposed models. Remarkably, the entire top 10 for the RDV is recognized at an AP above 50%. Compared to SOTA, the proposed models show improvements in the more complex task of detecting complete triplets and instrument-target while showing comparable performance for the visually simpler instrument-verb detection task (see supplementary material and statistical analysis in Table 10).

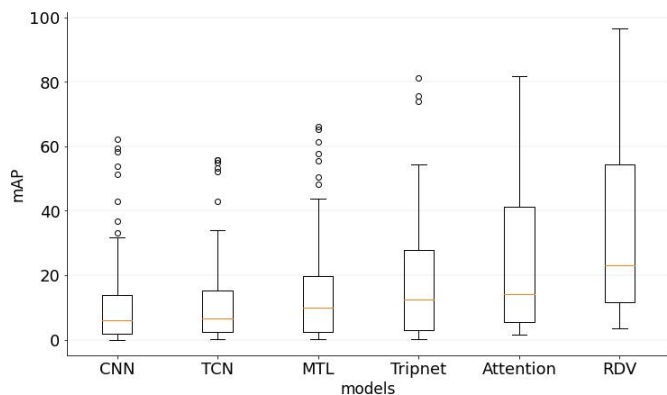


Fig. 8: Distribution of the model AP for the 100 triplet class predictions.

In addition to the top 10, we also present the full extent of the model’s performance on all 100 classes using the AP box plots in Fig. 8, showing upper and lower performance bounds for each model as well the spread around the mean. The rectangular box represents the middle 50% of the score for each model also known as *interquartile range*. As can be seen from Fig. 8, the proposed models maintain higher median and upper-

quartile performance than the baselines. They also maintain higher upper-whiskers showing the extent of their performance distribution above the interquartile range.

6.1.7. Statistical Significance Analysis

Table 10: The p -values obtained in Wilcoxon signed-rank test of the proposed methods using the SOTA model (Tripnet) as the alternative method. (Lower p -value is preferred)

Tasks		Proposed methods	
		Attention Tripnet	Rendezvous
Component Detection	AP_I	$p \approx 0.327$	$p \approx 0.374$
	AP_V	$p \ll 0.001$	$p \approx 0.003$
	AP_T	$p \ll 0.001$	$p \ll 0.001$
Triplet Association	AP_{IV}	$p \approx 0.018$	$p < 0.001$
	AP_{IT}	$p \approx 0.010$	$p \approx 0.005$
	AP_{IVT}	$p < 0.005$	$p \ll 0.001$

We also measure the statistical significance of the proposed model performance using the SOTA model as the alternative method. Using the Wilcoxon signed-rank test, we sample $N = 30$ random batches of 100 consecutive frames instead of 30 random frames to simulate the evaluation on video clips. The null hypothesis (H_0) states that the difference between the proposed method and the alternative method follows a symmetric distribution around zero. We perform the significance analysis for each task, and based on the obtained p -values, presented in Table 10, we draw the following conclusions:

1. Both proposed models do not significantly improve the instrument detection sub-task. Their p -values fall short of the standard 0.05. This is mainly because the instrument detection performance is already saturated in the alternative method; there is no new modeling in the proposed methods targeting their improvement. Being a two-tailed test, the p -value also shows that the SOTA does not outperform the proposed models on instrument detection.
2. The guided attention mechanism is very useful in improving the verb and target detections in both the Attention

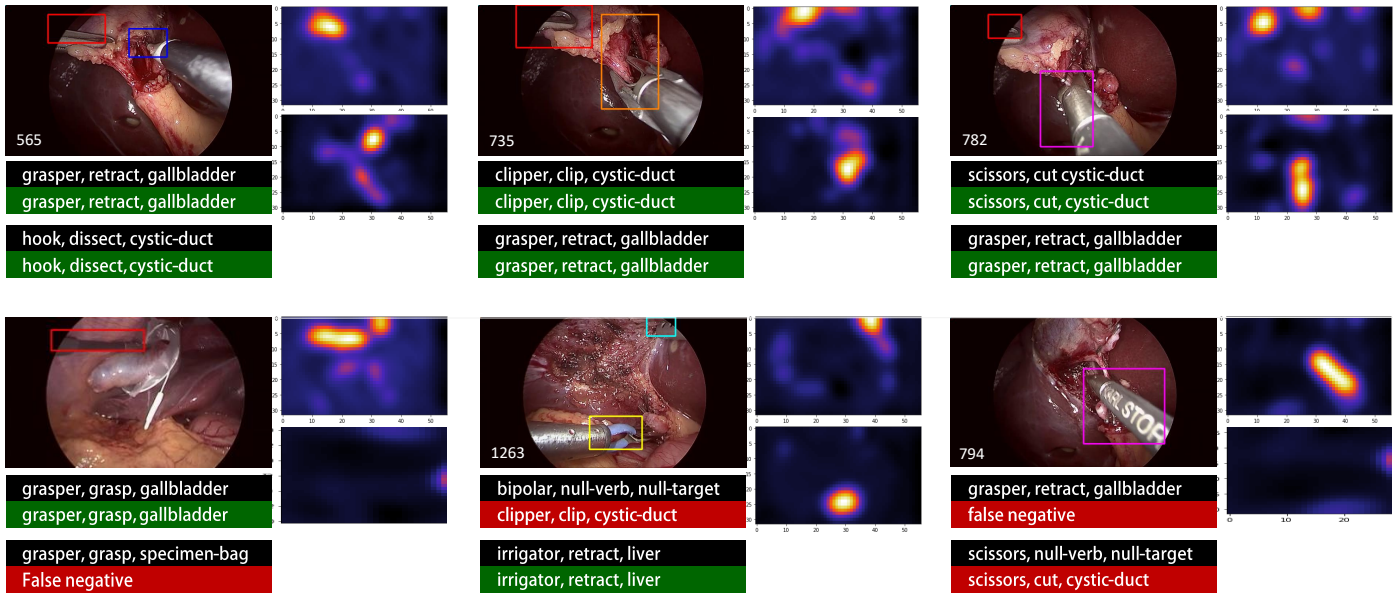


Fig. 9: Qualitative results showing the triplet predictions and the heatmaps for the triplet detection. Localization bounding boxes are obtained from the WSL module of the proposed RDV model. Predicted and ground truth triplets are displayed below each image: black = ground truth, green = correct prediction, red = incorrect prediction. A missed triplet is marked as false negative and a false detection is marked as false positive (Best viewed in color).

Tripnet and RDV models. Their contributions are significant enough to even beat a more narrow 0.01 significant level.

- Our contributions are also significant in improving the recognition of the tool-tissue interaction, with Attention Tripnet's improvement on AP_{IVT} relevant at a 0.005 significance level. Our best method (RDV) is more significant, with a p -value far below 0.001.

In summary, we reject the null hypothesis H_0 at a confidence level of 5%, concluding that there is a significant difference between the proposed models and the alternative method.

6.2. Qualitative Results

6.2.1. Triplet Recognition with Weak Localization

The predicted class labels are obtained by applying a 0.5 threshold on the output probabilities of the proposed RDV model. We present those predicted labels in Fig. 9, alongside the localization of the regions of action obtained from the weakly supervised learning (WSL) module of the network. This localization, depicted by bounding boxes overlaid on the image, shows the focus of the model when it makes a prediction, thereby providing insight into its rationale. Those results are solid arguments in favor of the model's ability for spatial reasoning when recognizing surgical actions. This suggests that the model can be further exploited for action triplet detection and segmentation. We also provide a short *video* of this qualitative performance in the supplementary material (also accessible via: https://youtu.be/d_yHdJtCa98).

6.2.2. Qualitative Analysis of Top 5 Predicted Triplets

We also examine the top 5 prediction confidence of the proposed models compared to baselines on random frames (Fig. 10). Fully correct predictions are signaled by the color green,

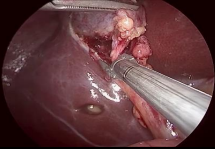
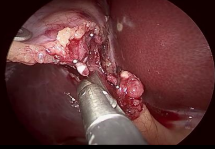

while red indicates errors on all three components. Other colors indicate partially correct predictions. RDV and Attention Tripnet outperform the baseline (MTL) and SOTA (Tripnet) each time, with the two actions correctly recognized each time within their top 5 predictions. Moreover, other actions in their top 5 have relevant components, showing these models' understanding of surgical actions by clustering triplets related to the performed actions. More qualitative results on this are provided in the supplementary material.


6.2.3. Attention Map Visualization

To understand the benefit of the CAGAM's attention mechanism, we visualize its attention maps in Fig. 11. For each input image, we randomly selected a few points ($marked\ i \in [1, 2, 3, 4]$) in the images and reveal the corresponding attention maps for the tool-tissue interaction captured in the CAGAM's position attention map. We observe that the attention module could capture semantic similarity and full-image dependencies, which change based on the contribution of the selected pixel to the action understanding. This shows that the model learns attention maps that contextualize every pixel in the image feature in relation to the action performed. For instance in the top image: point 2, a pixel location on the instrument - *grasper*, creates an attention map that highlights both the instrument and its target - *gallbladder*. Indeed, the attention guidance introduced in this model helps to highlight the triplet's interest regions while suppressing the rest. This effect is shown further in the supplementary video.


7. Conclusion

We have presented methods featuring new forms of attention mechanisms that surpass the state-of-the-art for surgical actions triplet (*instrument, verb, target*) recognition. We first proposed


Groundtruth	MTL Baseline	Tripnet	Attention Tripnet	Rendezvous
	clipper, clip, cystic-duct	grasper, retract, gallbladder	clipper, clip, cystic-duct	clipper, clip, cystic-duct
	grasper, retract, gallbladder	clipper, clip, cystic-duct	grasper, retract, gallbladder	grasper, retract, gallbladder
	clipper, clip, blood-vessel	clipper, clip, cystic-artery	clipper, clip, blood-vessel	clipper, clip, cystic-artery
	hook, retract, gallbladder	grasper, grasp, gallbladder	grasper, grasp, cystic-plate	clipper, clip, blood-vessel
clipper, clip, cystic-duct	bipolar, retract, liver	clipper, clip, blood-vessel	clipper, clip, cystic-plate	clipper, clip, cystic-plate
grasper, retract, gallbladder				
	grasper, retract, gallbladder	grasper, retract, gallbladder	scissors, cut, blood-vessels	grasper, retract, gallbladder
	hook, retract, gallbladder	grasper, grasp, liver	scissors, cut, cystic-artery	scissors, cut, cystic-duct
	grasper, retract, omentum	scissors, null-verb, null-target	grasper, retract, gallbladder	scissors, cut, cystic-artery
	scissors, cut, cystic-duct	scissors, cut, cystic-duct	scissors, cut, cystic-duct	scissors, null-verb, null-target
grasper, retract, gallbladder	grasper, null-verb, null-target	grasper, grasp, gallbladder	grasper, grasp, cystic-plate	grasper, grasp, gallbladder
Scissors, cut, cystic-duct				
	grasper, dissect, cystic-plate	grasper, grasp, specimen-bag	grasper, grasp, specimen-bag	grasper, grasp, specimen-bag
	grasper, null, null	grasper, null-verb, null-target	grasper, null-verb, null-target	grasper, grasp, gallbladder
	hook, dissect, gallbladder	grasper, grasp, gut	grasper, grasp, gallbladder	grasper, null-verb, null-target
	grasper, retract, gallbladder	grasper, retract, liver	grasper, retract, cystic-plate	grasper, retract, liver
grasper, grasp, specimen-bag	grasper, grasp, specimen-bag	grasper, grasp, liver	grasper, grasp, liver	grasper, pack, gallbladder
grasper, grasp, gallbladder				




Correct complete triplet (IVT)




Correct pair (IT) instrument-target



Correct pair (IV) instrument-verb



Correct instrument (I)



Incorrect prediction

Fig. 10: Qualitative results showing the top-5 triplet predictions for the best performing baseline, SOTA, and the proposed models (Best viewed in colour).

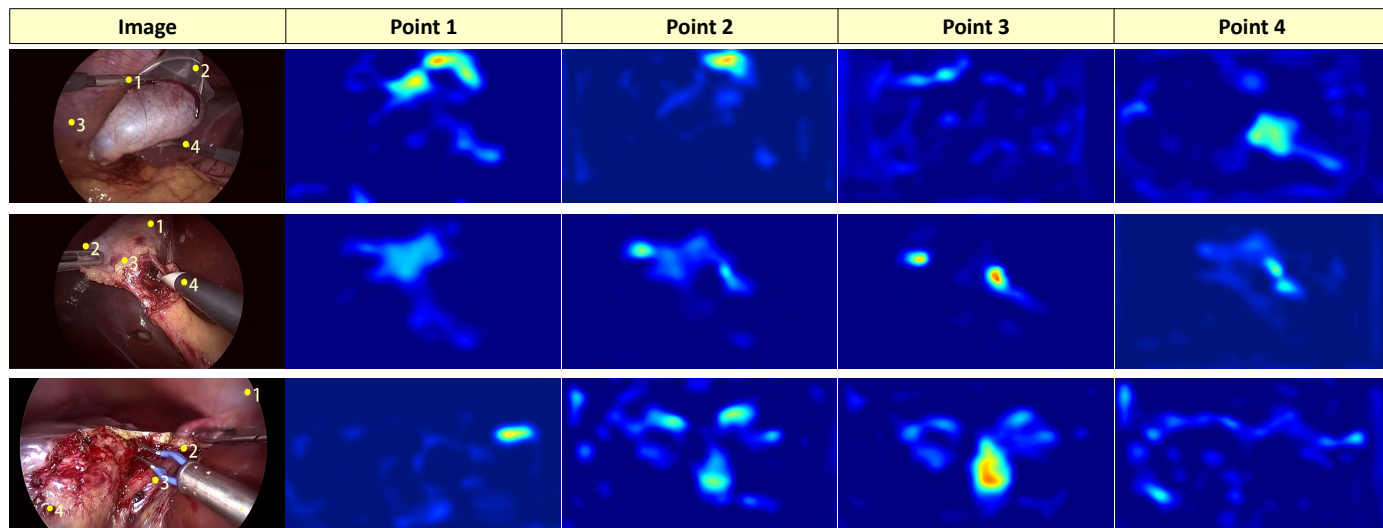


Fig. 11: Attention maps in the CAGAM module on the CholecT50 test set. The left column is the input image, the subsequent columns are the attention maps captured by the different points as marked in the input image. The attention map shows the focus on the target (best seen in color).

a novel approach for attention intended for verbs and targets, using instrument class activation maps. We have also introduced a novel hybrid attention mechanism that resolves component association in triplets by leveraging multiple heads of both self and cross attentions on the component features.

We have rigorously validated our performance claims on *CholecT50*, a new large-scale endoscopic video dataset also contributed in this work. We also discussed the benefits of the

proposed methods in terms of clinical significance. Qualitative results suggest possible extensions to different tasks, including automated surgical report generation and spatial action segmentation.

While these initial results are encouraging, many challenges remain. One is the scalability on unseen triplets which may likely be tackled by zero-, one- or few-shot learning. Our results on rare triplets already hint at promising prospects for this

approach. Inference speed is another challenge: increasing the number of layers generally drives up the performance, but is computationally very costly. Implementing a more lightweight Rendezvous would help alleviate some of these costs.

One limitation of this work is that target localization using the same weakly-supervised technique as for instruments is not yet achieved. This is likely due to the target's visibility not being the sole indicator for a positive binary label, both in the ground truth annotations and the model predictions. **We also observed that it is not possible to recognize multiple instances of the same triplet, e.g., two (*grasper*, *grasp*, *gallbladder*). This is due to the nature of the binary presence annotation, which does not provide an instance count for each unique triplet class. Only actions performed with the *grasper* instrument can have multi-instance occurrence in this dataset. Nonetheless, this does not affect recognition but is considered a limitation in future work on triplet localization, where multiple instances would need to be detected differently.**

With high-profile potential applications such as safety monitoring, skill evaluation, and objective reporting, our surgical action triplet method, as well as the release of our dataset for the 2021 Endoscopic Vision challenge, bring considerable value to the field of surgical activity understanding.

Future work will consider temporal modeling as some of the action verbs could be better recognized by the temporal dynamics of the tool-tissue interaction.

Acknowledgements

This work was supported by French state funds managed within the Investissements d'Avenir program by BPI France (project CONDOR) and by the ANR under references ANR-11-LABX-0004 (Labex CAMI), ANR-16-CE33-0009 (DeepSurg), ANR-10-IAHU-02 (IHU Strasbourg) and ANR-20-CHIA-0029-01 (National AI Chair AI4ORSafety). It was granted access to HPC resources of Unistra Mesocentre and GENCI-IDRIS (Grant 2021-AD011011638R1). The authors also thank the IHU and IRCAD research teams for their help with the data annotation during the CONDOR project.

References

- Ahmadi, S.A., Sielhorst, T., Stauder, R., Horn, M., Feussner, H., Navab, N., 2006. Recovery of surgical workflow without explicit models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 420–428.
- Al Hajj, H., Lamard, M., Conze, P.H., Cochener, B., Quelled, G., 2018. Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. *Medical Image Analysis* 47, 203–218.
- 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.
- Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
- Ballantyne, G.H., 2002. The pitfalls of laparoscopic surgery: challenges for robotics and telerobotic surgery. *Surgical Laparoscopy Endoscopy & Percutaneous Techniques* 12, 1–5.
- Bawa, V.S., Singh, G., KapingA, F., Skarga-Bandurova, I., Oleari, E., Leporini, A., Landolfo, C., Zhao, P., Xiang, X., Luo, G., et al., 2021. The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods. arXiv preprint arXiv:2104.03178.
- Bertasius, G., Wang, H., Torresani, L., 2021. Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095.
- Blum, T., Feußner, H., Navab, N., 2010. Modeling and segmentation of surgical workflow from laparoscopic video, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 400–407.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer. pp. 213–229.
- Chakraborty, I., Elgammal, A., Burd, R.S., 2013. Video based activity recognition in trauma resuscitation, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8.
- Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J., 2018. Learning to detect human-object interactions, in: 2018 IEEE winter conference on applications of computer vision (wacv), IEEE. 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: Proceedings of the IEEE International Conference on Computer Vision, pp. 1017–1025.
- Charriere, K., Quelled, G., Lamard, M., Coatrieux, G., Cochener, B., Cazuguel, G., 2014. Automated surgical step recognition in normalized cataract surgery videos, in: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE. pp. 4647–4650.
- Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transnet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.
- Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N., 2021. Opera: Attention-regularized transformers for surgical phase recognition. arXiv preprint arXiv:2103.03873.
- Dergachyova, O., Bouget, D., Huauilmé, A., Morandi, X., Jannin, P., 2016. Automatic data-driven real-time segmentation and recognition of surgical workflow. *International journal of computer assisted radiology and surgery* 11, 1081–1089.
- DiPietro, R., Ahmidi, N., Malpani, A., Waldram, M., Lee, G.I., Lee, M.R., Vedula, S.S., Hager, G.D., 2019. Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks. *International journal of computer assisted radiology and surgery* 14, 2005–2020.
- DiPietro, R., Lea, C., Malpani, A., Ahmidi, N., Vedula, S.S., Lee, G.I., Lee, M.R., Hager, G.D., 2016. Recognizing surgical activities with recurrent neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 551–558.
- 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.
- Felli, E., Mascagni, P., Wakabayashi, T., Mutter, D., Marescaux, J., Pessaux, P., 2019. Feasibility and value of the critical view of safety in difficult cholecystectomies. *Annals of surgery* 269, e41.
- Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154.
- 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, pp. 85–93.
- Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A., 2021. Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. arXiv preprint arXiv:2103.09712.
- García-Peraza-Herrera, L.C., Li, W., Fidon, L., Grijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., et al., 2017. Toolnet: holistically-nested real-time segmentation of robotic surgical tools, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 5717–5722.
- Gavrilyuk, K., Sanford, R., Javan, M., Snoek, C.G., 2020. Actor-transformers for group activity recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 839–848.
- Girdhar, R., Carreira, J., Doersch, C., Zisserman, A., 2019. Video action transformer network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 244–253.
- Gkioxari, G., Girshick, R., Dollár, P., He, K., 2018. Detecting and recognizing human-object interactions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8359–8367.
- Hu, J.F., Zheng, W.S., Lai, J., Gong, S., Xiang, T., 2013. Recognising human-

- object interaction via exemplar based modelling, in: Proceedings of the IEEE international conference on computer vision, pp. 3144–3151.
- Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W., 2019. Cc-net: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 603–612.
- Huaultm, A., Jannin, P., Reche, F., Faucheron, J.L., Moreau-Gaudry, A., Voros, S., 2020. Offline identification of surgical deviations in laparoscopic resection. *Artificial Intelligence in Medicine* 104, 101837.
- Ji, Z., Wang, H., Han, J., Pang, Y., 2019. Saliency-guided attention network for image-sentence matching, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5754–5763.
- Katić, D., Julliard, C., Wekerle, A.L., 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. *International journal of computer assisted radiology and surgery* 10, 1427–1434.
- Katić, D., Wekerle, A.L., 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: *International Conference on Information Processing in Computer-Assisted Interventions*, Springer. pp. 158–167.
- Kendall, A., Gal, Y., Cipolla, R., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7482–7491.
- Khatibi, T., Dezyani, P., 2020. Proposing novel methods for gynecologic surgical action recognition on laparoscopic videos. *Multimedia Tools and Applications* 79, 30111–30133.
- Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J., 2021. Hotr: End-to-end human-object interaction detection with transformers. *arXiv preprint arXiv:2104.13682*.
- Kitaguchi, D., Takeshita, N., Matsuzaki, H., Takano, H., Owada, Y., Enomoto, T., Oda, T., Miura, H., Yamanashi, T., Watanabe, M., et al., 2019. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surgical Endoscopy*, 1–8.
- Kletz, S., Schoeffmann, K., Münzer, B., Primus, M.J., Husslein, H., 2017. Surgical action retrieval for assisting video review of laparoscopic skills, in: Proceedings of the 2017 ACM Workshop on Multimedia-based Educational and Knowledge Technologies for Personalized and Social Online Training, pp. 11–19.
- Kolesnikov, A., Kuznetsova, A., Lampert, C., Ferrari, V., 2019. Detecting visual relationships using box attention, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0.
- Kondo, S., 2020. Lapformer: surgical tool detection in laparoscopic surgical video using transformer architecture. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 1–6.
- Lea, C., Vidal, R., Reiter, A., Hager, G.D., 2016. Temporal convolutional networks: A unified approach to action segmentation, in: *European Conference on Computer Vision*, Springer. pp. 47–54.
- Lecuyer, G., Ragot, M., Martin, N., Launay, L., Jannin, P., 2020. Assisted phase and step annotation for surgical videos. *International journal of computer assisted radiology and surgery*, 1–8.
- Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: *European conference on computer vision*, Springer. pp. 740–755.
- Lin, X., Zou, Q., Xu, X., . Action-guided attention mining and relation reasoning network for human-object interaction detection .
- Liu, W., Chen, S., Guo, L., Zhu, X., Liu, J., 2021. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804* .
- Lo, B.P., Darzi, A., Yang, G.Z., 2003a. Episode classification for the analysis of tissue/instrument interaction with multiple visual cues, in: *Int. conference on medical image computing and computer-assisted intervention*, pp. 230–237.
- Lo, B.P., Darzi, A., Yang, G.Z., 2003b. Episode classification for the analysis of tissue/instrument interaction with multiple visual cues, in: *Int. conference on medical image computing and computer-assisted intervention*, pp. 230–237.
- Loukas, C., Georgiou, E., 2015. Smoke detection in endoscopic surgery videos: a first step towards retrieval of semantic events. *The International Journal of Medical Robotics and Computer Assisted Surgery* 11, 80–94.
- Maier-Hein, L., Vedula, S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al., 2017. Surgical data science: Enabling next-generation surgery. *Nature Biomedical Engineering* 1, 691–696.
- Majumder, A., Altieri, M.S., Brunt, L.M., 2020. How do i do it: laparoscopic cholecystectomy. *Annals of Laparoscopic and Endoscopic Surgery* 5, 15.
- Mallya, A., Lazebnik, S., 2016. Learning models for actions and person-object interactions with transfer to question answering, in: *European Conference on Computer Vision*, Springer. pp. 414–428.
- Malpani, A., Lea, C., Chen, C.C.G., Hager, G.D., 2016. System events: readily accessible features for surgical phase detection. *International journal of computer assisted radiology and surgery* 11, 1201–1209.
- Mascagni, P., Vardazaryan, A., Alapatt, D., Urade, T., Emre, T., Fiorillo, C., Pessaux, P., Mutter, D., Marescaux, J., Costamagna, G., et al., 2021. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Annals of Surgery* .
- Mohla, S., Pande, S., Banerjee, B., Chaudhuri, S., 2020. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 92–93.
- Neumuth, T., Strauß, G., Meixensberger, J., Lemke, H.U., Burgert, O., 2006. Acquisition of process descriptions from surgical interventions, in: *International Conference on Database and Expert Systems Applications*, pp. 602–611.
- 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*, Springer, Cham. pp. 364–374.
- Nwoye, C.I., Mutter, D., Marescaux, J., Padoy, N., 2019. Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos. *International journal of computer assisted radiology and surgery* 14, 1059–1067.
- Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* .
- Petschmann, S., Schöffmann, K., Benois-Pineau, J., Chaabouni, S., Keckstein, J., 2018. Early and late fusion of temporal information for classification of surgical actions in laparoscopic gynecology, in: *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE. pp. 369–374.
- Pucher, P.H., Brunt, L.M., Davies, N., Linsk, A., Munshi, A., Rodriguez, H.A., Fingerhut, A., Fanelli, R.D., Asbun, H., Aggarwal, R., 2018. Outcome trends and safety measures after 30 years of laparoscopic cholecystectomy: a systematic review and pooled data analysis. *Surgical endoscopy* 32, 2175–2183.
- 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. *arXiv preprint arXiv:2102.12218* .
- Rupprecht, C., Lea, C., Tombari, F., Navab, N., Hager, G.D., 2016. Sensor substitution for video-based action recognition, in: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE. pp. 5230–5237.
- Sahu, M., Szengel, A., Mukhopadhyay, A., Zachow, S., 2020. Surgical phase recognition by learning phase transitions. *Current Directions in Biomedical Engineering* 6.
- Sankaran, B., Mi, H., Al-Onaizan, Y., Ittycheriah, A., 2016. Temporal attention model for neural machine translation. *arXiv preprint arXiv:1608.02927* .
- Shaffer, E.A., 2006. Epidemiology of gallbladder stone disease. *Best practice & research Clinical gastroenterology* 20, 981–996.
- Shen, L., Yeung, S., Hoffman, J., Mori, G., Fei-Fei, L., 2018. Scaling human-object interaction recognition through zero-shot learning, in: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE. pp. 1568–1576.
- Sundaramoorthy, C., Kelvin, L.Z., Sarin, M., Gupta, S., 2021. End-to-end attention-based image captioning. *arXiv preprint arXiv:2104.14721* .
- Sznitman, R., Becker, C., Fua, P., 2014. Fast part-based classification for instrument detection in minimally invasive surgery, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 692–699.
- Twinanda, A.P., Alkan, E.O., Gangi, A., de Mathelin, M., Padoy, N., 2015. Data-driven spatio-temporal rgb-d feature encoding for action recognition in

- operating rooms. *Int. journal of computer assisted radiology and surgery* 10, 737–747.
- Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N., 2017. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging* 36, 86–97.
- Ulutun, O., Iftekhhar, A., Manjunath, B.S., 2020. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13617–13626.
- Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M., 2021. Medical transformer: Gated axial-attention for medical image segmentation. *arXiv preprint arXiv:2102.10662*.
- Vardazaryan, A., Mutter, D., Marescaux, J., Padoy, N., 2018. Weakly-supervised learning for tool localization in laparoscopic videos. *arXiv preprint arXiv:1806.05573*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.
- Velanovich, V., 2000. Laparoscopic vs open surgery. *Surgical endoscopy* 14, 16–21.
- 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.
- Voros, S., Long, J.A., Cinquin, P., 2007. Automatic detection of instruments in laparoscopic images: A first step towards high-level command of robotic endoscopic holders. *The International Journal of Robotics Research* 26, 1173–1190.
- Wagner, M., Müller-Stich, B.P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., et al., 2021. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *arXiv preprint arXiv:2109.14956*.
- Wang, T., Anwer, R.M., Khan, M.H., Khan, F.S., Pang, Y., Shao, L., Laaksonen, J., 2019. Deep contextual attention for human-object interaction detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5694–5702.
- Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803.
- Xu, M., Islam, M., Lim, C.M., Ren, H., 2021. Learning domain adaptation with model calibration for surgical report generation in robotic surgery. *arXiv preprint arXiv:2103.17120*.
- Yao, Q., Gong, X., 2020. Saliency guided self-attention network for weakly and semi-supervised semantic segmentation. *IEEE Access* 8, 14413–14423.
- Yu, T., Mutter, D., Marescaux, J., Padoy, N., 2018. Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition. *arXiv preprint arXiv:1812.00033*.
- Zia, A., Hung, A., Essa, I., Jarc, A., 2018. Surgical activity recognition in robot-assisted radical prostatectomy using deep learning, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 273–280.
- Zisimopoulos, O., Flouty, E., Luengo, I., Giataganas, P., Nehme, J., Chow, A., Stoyanov, D., 2018. Deepphase: surgical phase recognition in cataracts videos, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 265–272.
- 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. *arXiv preprint arXiv:2103.04503*.