Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition

Saurav Sharma[1] · Chinedu Innocent Nwoye[1] · Didier Mutter[2,3] · Nicolas Padoy[1,2]

Abstract

Purpose   One of the recent advances in surgical AI is the recognition of surgical activities as triplets of ⟨instrument, verb, target⟩. Albeit providing detailed information for computer-assisted intervention, current triplet recognition approaches rely only on single-frame features. Exploiting the temporal cues from earlier frames would improve the recognition of surgical action triplets from videos.

Methods   In this paper, we propose *Rendezvous in Time (RiT)*—a deep learning model that extends the state-of-the-art model, Rendezvous, with temporal modeling. Focusing more on the verbs, our RiT explores the connectedness of current and past frames to learn temporal attention-based features for enhanced triplet recognition.

Results   We validate our proposal on the challenging surgical triplet dataset, CholecT45, demonstrating an improved recognition of the verb and triplet along with other interactions involving the verb such as ⟨instrument, verb⟩. Qualitative results show that the RiT produces smoother predictions for most triplet instances than the state-of-the-arts.

Conclusion   We present a novel attention-based approach that leverages the temporal fusion of video frames to model the evolution of surgical actions and exploit their benefits for surgical triplet recognition.

Keywords   Surgical triplet recognition · Laparoscopic surgery · Temporal modeling · Action triplet · Attention model

Introduction

Laparoscopy, despite its many advantages, added some level of complexity to surgical procedures [1]. This brings the need to support surgeon's efforts with highly adaptive AI systems that can learn from streams of digital data [2] available in the operating room (OR) to provide intraoperative context-awareness [1]. Bulk of the existing research focuses on surgical phase recognition [3] and instrument spatial and presence detection [4]. The phase information is coarse-grained, concealing fine details of the activities taking place [5], whereas the instrument bounding box location might not necessarily contain information about interaction with the anatomy. Even the finer action recognition [6] that summarizes scene activity by a single verb is devoid of instrument and tissue details.

Taking everything into consideration, surgical action triplet is proposed [7, 8] to formalize and recognize surgical activities as triplets of ⟨*instrument, verb, target*⟩ enabling detailed and holistic surgical activity understanding for more helpful AI assistance in the OR. At this milestone, current methods are yet to explore the freely available temporal data in surgical videos. Tripnet [8], the first deep learning model on this task, uses a multitask learning framework to capture instances of instrument, verb, and target from a single image frame and associate them to form a triplet using a 3D interaction function. The state-of-the-art method, Rendezvous (RDV) [9], improves on the performance using a hybrid attention mechanism, yet on single frames. This could explain why the RDV performs low on some triplets such as ⟨*irrigator, aspirate, fluid*⟩, ⟨*grasper, retract, omentum*⟩, ⟨*bipolar, coagulate, cystic-pedicle*⟩, which are heavily dif-

✉ Saurav Sharma
  ssharma@unistra.fr

  Chinedu Innocent Nwoye
  nwoye@unistra.fr

  Didier Mutter
  didier.mutter@ihu-strasbourg.eu

  Nicolas Padoy
  npadoy@unistra.fr

[1] ICube, University of Strasbourg, CNRS, Strasbourg, France

[2] IHU Strasbourg, Strasbourg, France

[3] University Hospital of Strasbourg, Strasbourg, France

ferentiated by temporal dynamics of their instruments and targets. Temporal modeling has been employed in many other surgical workflow analyses, e.g., phase recognition [10, 11], tool tracking [12], to improve performance leveraging inherent motion characteristics in video-based data.

Motivated by the need for temporal modeling in fine-grained analysis, we propose *RiT*, which stands for *Rendezvous in Time*, that extends the RDV with temporal modeling. We focus our temporal modeling on the *verb* component of the triplet following our strong assumption that the verb is the most temporal-dependent component of the triplet. As an example, the triplet ⟨clipper, clip, cystic-duct⟩ would require more than one frame to discern the *clip* activity, whereas the *clipper* and *cystic-duct* terms can be obvious from a single frame. Predicting the verbs from a single frame, as has been the norm in existing models, would be misleading and unreliable, especially in the case of the instrument's occlusion and other procedural noise. Therefore, the RiT is designed to study the effects of integrating the discriminative temporal context of the verb features into the triplet recognition task. We specifically ask—*How can we model the verb features across time with higher precision without sacrificing the triplet recognition accuracy*? To answer this research question, RiT employs a novel temporal attention module (TAM) on the verb processing layer of the RDV to aggregate verb features from current and past frames in a late fusion setup. This allows for the learning of local motion representations of instrument-target interactions, which are pivotal for activity understanding. We further evaluate RiT on the challenging CholecT45 dataset [9] and show improvements on verb and triplet recognition as well as on other sub-level associations.

The contributions in this work are twofold: (1) we propose a novel temporal attention mechanism (TAM) targeting the verb component of triplet utilizing current and past frames; (2) we justify our choice of temporal model by providing extensive ablation studies for different configurations of TAM.

Related works

Surgical workflow analysis

Activity recognition is one of the numerous branches of surgical workflow analysis that has deepened in research over time: from coarse-grained phase recognition [13–15] to fine-grained gesture [16], step [17], and action [6] recognition. The finer actions describe surgical activities leaving out details about the instruments and tissues involved in the interactions [5].

Surgical action triplets

Surgical triplets in a cholecystectomy procedure present a way to perform fine-grained analysis involving all concerned components by modeling activity as a triplet ⟨*instrument, verb, target*⟩. A 40-video surgical dataset, known as CholecT40 [8], provides triplet-labeled annotations for research, which is extended to CholecT50 [9] with additional 10 videos. A related dataset, SARAS-ESAD [18] that formalized activity as only ⟨*verb, anatomy*⟩, provides bounding box annotations for the surgical instruments on 4 videos of radical prostatectomy procedures. QDNet [19] maintains the complete triplet formalism and provides bounding boxes for both the instruments and targets for cataract surgery.

In terms of method, Tripnet [8] utilizes an instrument-centric multitask model built on a weakly supervised instrument localization module, a class activation-based verb and target recognition, and a learnable interaction matrix to capture triplets. A subsequent approach, RDV [9], leverages a transformer-inspired attention method. More methods are presented in the 2021 CholecTriplet challenge [20] for triplet recognition.

Temporal modeling

Modeling temporal dynamics of surgical activities across multiple frames has become a promising way of improving model performance. Temporal convolution network is applied in TeCNO [10] in a stage-wise manner to recognize phases. Trans-SVNet [21] utilizes a transformer to aggregate both spatial and temporal embeddings to predict phase labels. ConvLSTM [12] and MTRCNet-CL [11] apply recurrent models for instrument motion tracking and phase recognition. Notwithstanding, temporal modeling is yet to be fully explored on the task of surgical action triplet recognition. Few exceptions in the CholecTriplet challenge [20] added well-known temporal units like LSTM, GRU, ConvLSTM, etc., on the triplet model predictions. Our proposed method innovates a temporal attention module that aggregates verb features from past frames into the current frame for surgical action triplet understanding.

Methodology

Base architecture

Our proposed method is based on RDV [9] framework, which is an encoder-decoder style of architecture where the encoder learns individual components—*instrument*(**i**), *verb*(**v**), *target*(**t**), and the decoder models the interactions between components to predict *triplet*(**ivt**). The encoder takes an ImageNet pre-trained ResNet-18 [22] to extract visual features.
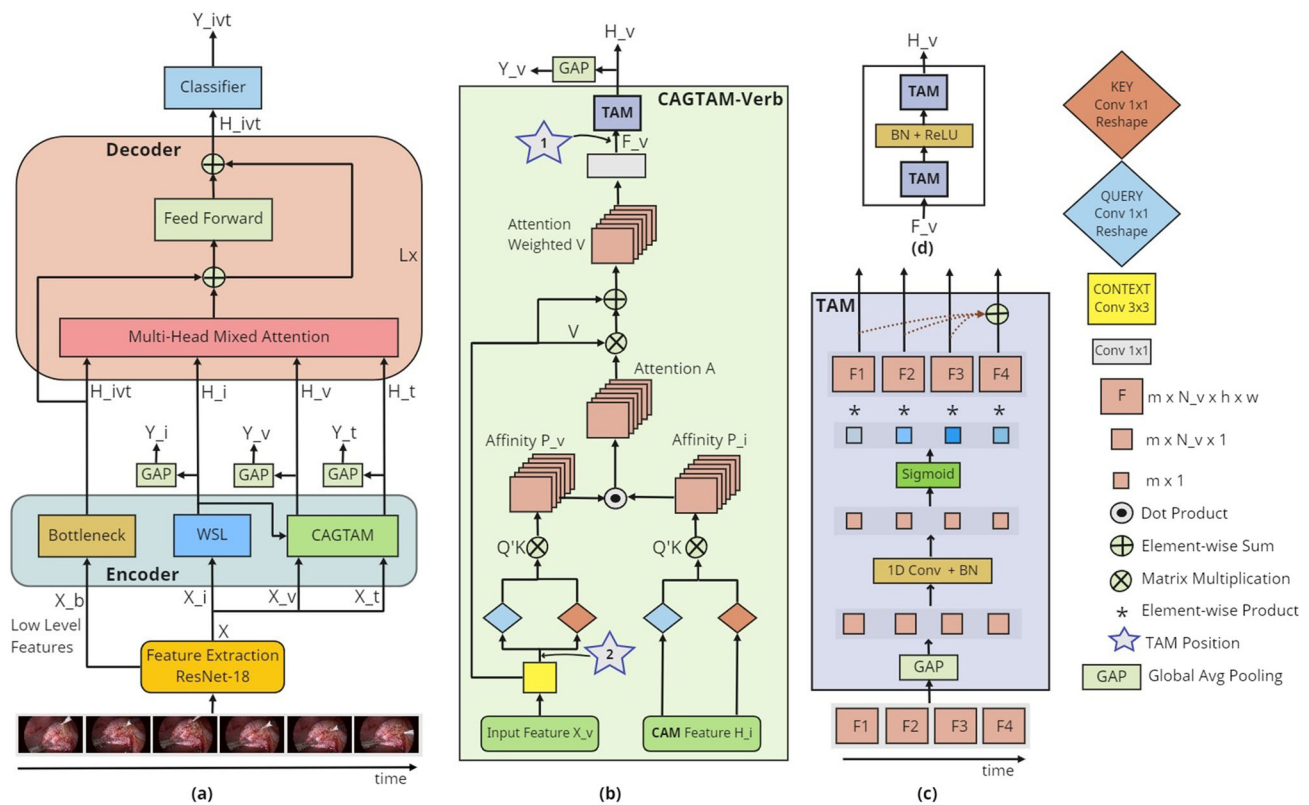
Fig. 1 Model illustration: **a** Rendezvous in Time (RiT) architecture with CAGTAM module extending RDV [9], **b** CAGTAM module layout to process verb features with proposed temporal attention module (TAM), **c** TAM in detail for the temporal fusion of verb features from current and past frames weighted by attention scores following Eq. (1). **d** Configuration of two TAM layers stacked together. *Note m* and N_v refer to the video clip size and number of verb classes, respectively

Acting on these features, a weakly supervised localization (WSL) module with two-layer convolution and a global pooling operation models the instrument and generates class activation maps (CAM). These instrument CAMs are input to the class activation guided attention mechanism (CAGAM) module as contextual features to learn verb and target features with the help of attention. Additionally, the encoder uses a bottleneck layer to extract low-level scene features **H_ivt**. As an input, the decoder then takes the scene features and activation maps from triplet components (**H_i**, **H_v**, **H_t**) and applies layers of multi-head attention to capture their interrelations and generates subtle triplet features **Y_ivt**. To induce multitask training, activation maps **H_i**, **H_v**, **H_t** are average pooled to generate corresponding logits **Y_i**, **Y_v**, **Y_t**. RDV, modeled at the frame level, does not consider the temporal cues from the video data, which can improve verb performance and enhance instrument interaction with the verb component of the triplet.

To address the above limitation, we present our proposed framework, *Rendezvous in Time (RiT)*, an RDV-based neural network that adds a new temporal model on the verb component of triplet to capture relations across video frame

features. RiT, illustrated in Fig. 1a, introduces the class activation guided **Temporal** attention mechanism (CAGTAM) module, adding the long-range capability to the CAGAM module. RiT enhances surgical action triplet recognition by unifying the verb features from past frames to build holistic verb features for the current video frame.

Class activation guided temporal attention mechanism (CAGTAM)

Temporal modeling of verb features is necessary to recognize verb classes such as *clip*, *cut*, and *aspirate* that rely heavily on the temporal dynamics of activity. We begin with the CAGAM module from RDV [9] and adapt it to take video clips as input. We then add our novel temporal attention module (TAM) layer at position 1 denoted by a star in Fig. 1b, to fuse past and current frame features. We name this setup consisting of CAGAM with multi-frame clip support and our proposed TAM layer as CAGTAM.

In detail, TAM as shown in Fig. 1c, takes as input a verb feature tensor $(F\_v)_t \in \mathcal{R}^{b \times m \times c \times h \times w}$ at time step t, where c denotes total number of verb classes and m the

Table 1 Triplet recognition AP (%) on CholecT45 dataset using official cross-validation splits

Method	#Params (M)	Component detection			Triplet association		
		AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
Tripnet	14.95	**89.9 ± 1.0**	59.9 ± 0.9	37.4 ± 1.5	31.8 ± 4.1	27.1 ± 2.8	24.4 ± 4.7
RDV	16.61	89.3 ± 2.1	62.0 ± 1.3	40.0 ± 1.4	34.0 ± 3.3	30.8 ± 2.1	29.4 ± 2.8
RiT	17.12	88.6 ± 2.6	**64.0 ± 2.5**	**43.4 ± 1.4**	**38.3 ± 3.5**	**36.9 ± 1.0**	**29.7 ± 2.6**

Bold numbers denote best performance for each metric

Table 2 Per-class verb recognition AP (%) on CholecT45 dataset using official cross-validation splits

Classes	Tripnet	RDV	RiT
Grasp	**70.5 ± 5.8**	69.8 ± 3.7	69.6 ± 3.9
Retract	**90.5 ± 5.4**	89.7 ± 7.2	89.3 ± 9.0
Dissect	93.0 ± 2.8	**93.2 ± 3.9**	92.6 ± 2.2
Coagulate	67.2 ± 6.1	**68.7 ± 5.5**	68.5 ± 6.1
Clip	85.4 ± 6.4	85.5 ± 3.7	**87.3 ± 5.3**
Cut	70.5 ± 9.1	72.0 ± 4.8	**74.9 ± 10.9**
Aspirate	60.7 ± 9.2	57.8 ± 9.9	**64.9 ± 7.8**
Irrigate	**29.6 ± 8.2**	25.7 ± 5.8	22.7 ± 9.0
Pack	32.1 ± 9.9	31.2 ± 9.9	**43.2 ± 16.2**
Null-verb	23.0 ± 2.4	24.0 ± 4.1	**26.4 ± 4.8**
Mean	59.9 ± 0.9	62.0 ± 1.3	**64.0 ± 2.5**

Bold numbers denote best performance for each metric

video clip size. Afterward, TAM applies global average-pooling to get feature $(F_-v)_t \in \mathcal{R}^{b \times m \times c}$. Then, a sequence of *1D-convolution-batch-normalization-sigmoid nonlinearity* generates the probability score w_i, where $i \in [t-m+1 \ldots t]$. Following Eq. (1), TAM then fuses verb features from past and current frames, weighted by w_i to generate refined verb features $(H_-v)_t$ at the time step t.

$$(H_-v)_t = \sum_{i=t-m+1}^{t} w_i \odot (F_-v[:, i, :, :, :]). \tag{1}$$

This enables TAM to model contextual dependencies across past frames while also moderating the noisy or less discriminative features that could generate due to the naive addition of verb features from past frames. Furthermore, due to the utilization of temporal context, TAM learns better instrument-verb semantics which otherwise would be ambiguous in a single-frame context.

Verb feature fusion

The position of TAM plays a vital role in deciding when to fuse frame features and directly impacts verb recognition capability. Inspired from [23], we explore two strategies to aggregate past and current frame features—early and late fusion marked as position 1 and 2, respectively, in Fig. 1b.

In early fusion, TAM processes abstract ResNet-18 features to generate verb features without contextualizing them with instrument CAMs. In late fusion, on the other hand, TAM processes instrument contextualized verb features and exploits useful information present in verb features from past frames to enrich current frame verb features.

Video slicing input pipeline

To train RiT, we create video clips from m frames following Eq. (2):

$$\text{clip}_t = ([f_{t-m+1}..f_t], y_t) \quad \forall t \in \{0, 1, \ldots, N\}, \tag{2}$$

where N is the total number of frames in a video and y_t is the ground truth label for the tth frame. To update verb feature at $(F_-v)_t$, verb features from $[(F_-v)_{t-m+1}..(F_-v)_{t-1}]$ are utilized and during training, the loss is computed only at time step t. This clip sampling strategy helps to train RiT keeping multiple clips from different videos in a batch and improve the learning process by shuffling across clips and videos. Moreover, the input pipeline is causal as the model learns to predict label y_t from clip_t with no access to future frames.

Experiments

Dataset and evaluation metrics

We follow the official protocol for data splits and metrics [24] to conduct our experiments on the multi-label CholecT45 triplet dataset that consists of 6 instrument, 10 verb, 15 target, and 100 triplet classes. The video frames, sampled at a 1fps frame rate, are annotated with binary presence labels of instrument, verb, target, and triplet. We use the official five-fold cross-validation splits for model evaluation and RDV splits for ablation studies. Also, the logits for instrument, verb, and target components are extracted from main triplet logits as suggested in [24]. We use video-specific average precision (AP) to evaluate model performance.

Implementation details

Following [24], we resize the video clip frames to 256×448 resolution and apply data augmentation strategies such as

(a) triplet < grasper, retract, gallbladder >

(b) triplet < scissors, cut, cystic_duct >

(c) triplet < clipper, clip, cystic_artery >

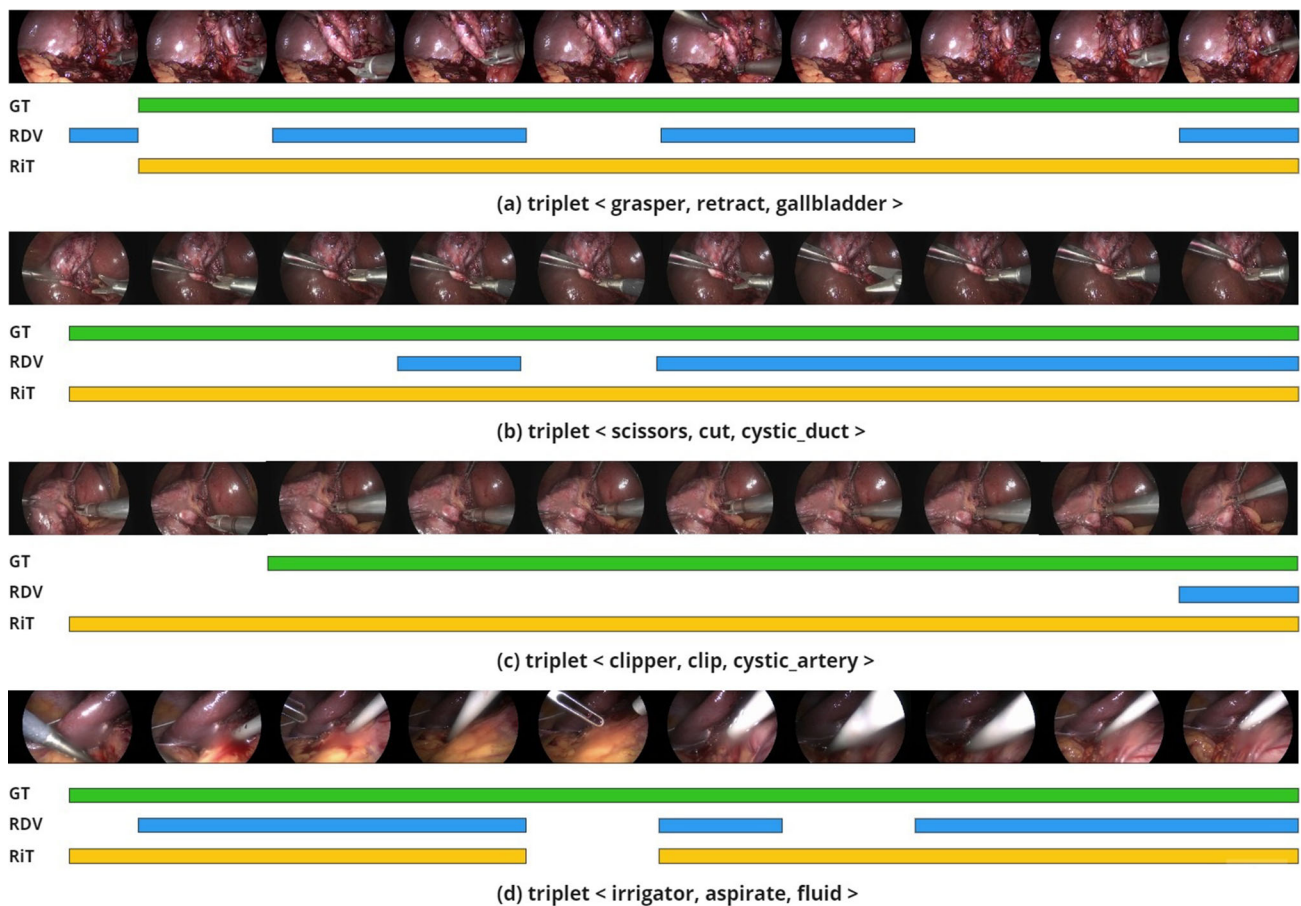(d) triplet < irrigator, aspirate, fluid >

Fig. 2 *Qualitative results*: Rendezvous in Time (RiT) (*in orange*) compared with RDV (*in blue*) and ground truth (GT) (*in green*) on random images from test videos. For verbs such as retract, cut, aspirate, and clip, RiT tries to achieve temporally smooth predictions compared to RDV as seen for triplets. Empty spaces are missed triplet predictions. (Best viewed in color)

horizontal flipping and brightness/contrast shift. We remove scaling and vertical flipping augmentation as these hurt the triplet recognition performance. We use weighted binary cross-entropy loss for the multi-label classification task following:

$$L = \sum_{c=1}^{C} \frac{-1}{N} \left(W_c y_c \log(\sigma(\overline{y}_c)) + (1 - y_c) \log(1 - \sigma(\overline{y}_c)) \right), \tag{3}$$

where C refers to total number of classes, y_c and $\overline{y}_c$, respectively, denote correct and predicted labels, σ is the sigmoid activation function, and W_c is the weight applied for class balancing following [12]. We apply loss L on instrument, verb, target, and triplet logits to create the total loss L_{total}:

$$L_{\text{total}} = L_{\text{Instrument}} + L_{\text{Verb}} + L_{\text{Target}} + L_{\text{Triplet}}. \tag{4}$$

Using our presented input pipeline in section "Video Slicing Input Pipeline", we feed to RiT a batch of random video

Table 3 Triplet recognition AP (%) in early (EF) vs late (LF) fusion

Method	Component detection			Triplet association		
	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
RiT (EF)	87.3	59.3	42.1	34.9	32.5	25.1
RiT (LF)	**89.5**	**63.0**	**45.2**	**38.7**	**37.5**	**30.8**
RiT (EF+LF)	85.8	60.3	42.6	37.1	32.8	27.8

Bold numbers denote best performance for each metric

clips, $\mathcal{V} \in \mathcal{R}^{b \times 3 \times m \times h \times w}$, where m is the video clip size, $h = 256$ is the image height, and $w = 448$ is the image width. Given the memory constraint during model training, we consider batches of 16 video clips. The model is trained end-to-end using stochastic gradient descent with $1e^{-6}$ weight decay and a combination of linear and exponential learning rate schedules. The linear learning rate schedule helps to simulate model warmup where it first learns each triplet component and gradually learns triplet information. RiT, implemented in PyTorch, is trained for 50 epochs on Nvidia V100 and RTX6000 GPUs. We tune model hyperparameters on the 5 validation videos.

Table 4 Triplet recognition AP (%) on the effect of video clip size

Method	Component detection			Triplet association		
	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
4	88.5	61.9	41.9	33.2	37.3	27.9
6	**89.5**	**63.0**	**45.2**	**38.7**	**37.5**	**30.8**
8	86.5	59.0	42.0	32.8	36.2	28.7

Bold numbers denote best performance for each metric

Results

We present our results in Table 1 for the official fivefold cross-validation splits where we use RiT with 2 TAM layers and video clip of six frames. Inside CAGTAM, the 2 TAM layers are stacked together with BatchNorm and ReLU as shown in Fig. 1d. Thanks to the temporal context exploited by the TAM, RiT improves upon state-of-the-art RDV on the verb and triplet recognition further enhancing instrument-verb association. Our novel TAM improves the recognition of five verb classes—*clip*, *cut*, *aspirate*, *pack*, and *null-verb* while the performance of other verb classes is similar to that of state-of-the-art. Interestingly, we also observe gains in target recognition and instrument-target association. We believe this is due to the multitask nature of the training where temporal modeling on verbs impacts target understanding. We further report the classwise performance of verb categories in Table 2. Qualitative results comparing RiT with RDV are shown in Fig. 2. Note that for baseline results, we consider the PyTorch reproduced version of Tripnet and RDV [24].

Ablations

Effect of position of TAM

We investigate the effect of the position of the TAM on triplet recognition performance. As shown in Fig. 1b, we add 1 TAM layer on the verb component at position 1 and 2 for late and early fusion, respectively. Results in Table 3 show that late fusion performs better than early fusion due to the verb features contextualized with instrument class activation map adding more discriminative power. Moreover, with both early and late fusion, the performance of the triplet component recognition and their interaction association drops. We

keep late fusion (position 1) setup of TAM for our proposed method.

Effect of clip size m on triplet recognition

Ablation study on the clip size m helps in understanding the impact of the duration of temporal context on triplet recognition. Table 4 shows the triplet recognition for different m values where performance diminishes beyond $m = 6$. This is because most action triplets are repetitive with short bursts of action for which fusing features from more than the optimal number of frames introduces noise. We keep $m = 6$ as the default for our proposed method.

TAM vs other temporal models

We also explore other temporal models such as a ConvLSTM, ConvGRU to highlight the impact of TAM. ConvLSTM has been shown to be effective in the temporal modeling of features, particularly for surgical instrument tracking [12]. We choose convolutional recurrent unit over standard LSTM/GRU as it maintains structural locality without flattening the features. We put 1 ConvLSTM layer with 512-dim features at position 1 like TAM as shown in Fig. 1b. We repeat the same with ConvGRU. Additionally, we also modify 1 non-local block [25] to operate on the spatial verb feature maps. Results in Table 5 demonstrates the power of TAM in modulating the contribution of verb features from past frames over other temporal models.

We provide additional ablation results on TAM in the supplementary material. We also evaluate our model offline in CholecTriplet (MICCAI 2022) Challenge (https://cholectriplet2022.grand-challenge.org/results) where we would rank 4th in the triplet recognition leaderboard. This is encouraging as the high-ranked methods employ bounding box annotations to model triplet recognition, whereas RiT relies only on binary presence labels of the triplet and its components. (Detailed ranking is in the supplementary material).

Conclusion

This work introduces a deep learning method for improved surgical action triplet recognition leveraging temporal infor-

Table 5 Triplet recognition AP (%) for TAM vs other temporal models

Method	Component detection			Triplet association		
	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
RiT (with Non-Local Block)	87.1	59.7	41.4	36.0	34.7	29.2
RiT (with ConvLSTM)	87.8	59.5	41.1	37.1	32.8	26.6
RiT (with ConvGRU)	**89.7**	61.7	43.0	38.2	34.6	29.3
RiT (with TAM)	89.5	**63.0**	**45.2**	**38.7**	**37.5**	**30.8**

Bold numbers denote best performance for each metric

mation in video clips. We design a temporal attention module (TAM) to modulate verb features from past frames. The learned features are in turn utilized by our final model, RDV in Time (RiT), which extends the frame-based RDV network with attention-based temporal modeling to predict more accurate triplet labels. Our results on the challenging CholecT45 dataset demonstrate that learning temporally coherent features while focusing attention on the most relevant locations, is crucial to reap benefits from refined verb features. This leads to capturing instrument–tissue interactions across video frames more accurately.

Code Availability The source code is publicly available at https://github.com/CAMMA-public/rendezvous-in-time.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent This manuscript does not contain any patient data.

References

1. Vercauteren T, Unberath M, Padoy N, Navab N (2020) CAI4CAI: the rise of contextual artificial intelligence in computer-assisted interventions. Proc IEEE 108(1):198–214
2. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, Eisenmann M et al (2017) Surgical data science for next-generation interventions. Nat Biomed Eng 1(9):691–696
3. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N (2017) Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE TMI 36(1):86–97
4. Jin A, Yeung S, Jopling J, Krause J, Azagury D, Milstein A, Fei-Fei L (2018) Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: WACV, pp 691–699
5. Nwoye C (2021) Deep learning methods for the detection and recognition of surgical tools and activities in laparoscopic videos. PhD thesis, Université de Strasbourg
6. Wagner M, Müller-Stich B-P, Kisilenko A, Tran D, Heger P, Mündermann L, Lubotsky DM, 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:2109.14956
7. Katić D, Wekerle A-L, Gärtner F, Kenngott H, Müller-Stich BP, Dillmann R, Speidel S (2014) Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance. In: IPCAI, pp 158–167
8. Nwoye CI, 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: MICCAI, pp 364–374
9. Nwoye CI, Yu T, Gonzalez C, Seeliger B, Mascagni P, Mutter D, Marescaux J, Padoy N (2022) Rendezvous: attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Med Image Anal 78:102433
10. Czempiel T, Paschali M, Keicher M, Simson W, Feussner H, Kim ST, Navab N (2020) Tecno: surgical phase recognition with multi-stage temporal convolutional networks. In: MICCAI, pp 343–352
11. Jin Y, Li H, Dou Q, Chen H, Qin J, Fu C-W, Heng P-A (2020) Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Med Image Anal 59:101572
12. Nwoye CI, Mutter D, Marescaux J, Padoy N (2019) Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. IJCARS 14(6):1059–1067
13. Dergachyova O, Bouget D, Huaulmé A, Morandi X, Jannin P (2016) Automatic data-driven real-time segmentation and recognition of surgical workflow. In: IJCARS, pp 1081–1089
14. Funke I, Jenke A, Mees ST, Weitz J, Speidel S, Bodenstedt S (2018) Temporal coherence-based self-supervised learning for laparoscopic workflow analysis. LNCS 11041:85–93
15. 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. In: MICCAI, pp 593–603
16. DiPietro R, Ahmidi N, Malpani A, Waldram M, Lee GI, Lee MR, Vedula SS, Hager GD (2019) Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks. In: IJCARS, pp 2005–2020
17. 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. In: IJCARS, pp 1111–1119
18. Bawa VS, Singh G, KapingA F, Skarga-Bandurova I et al (2021) The SARAS endoscopic surgeon action detection (ESAD) dataset: challenges and methods. arXiv:2104.03178
19. Lin W, Hu Y, Hao L, Zhou D, Yang M, Fu H, Chui C, Liu J (2022) Instrument-tissue interaction quintuple detection in surgery videos. In: MICCAI, pp 399–409
20. Nwoye CI, Alapatt D, Yu T, Vardazaryan A, Xia F, Zhao Z, Xia T, Jia F, Yang Y, Wang H et al (2022) Cholectriplet2021: a benchmark challenge for surgical action triplet recognition. arXiv:2204.04746
21. Jin Y, Long Y, Gao X, Stoyanov D, Dou Q, Heng P-A (2022) Trans-svnet: hybrid embedding aggregation transformer for surgical workflow analysis. In: IJCARS, pp 1–10
22. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR, pp 770–778
23. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: CVPR, pp 1725–1732
24. Nwoye CI, Padoy N (2022) Data splits and metrics for method benchmarking on surgical action triplet datasets. arXiv:2204.05235
25. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: CVPR, pp 7794–7803