Medical Image Analysis 74 (2021) 102224

Contents lists available at ScienceDirect

Medical Image Analysis

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

Real-time medical phase recognition using long-term video understanding and progress gate method

Yanyi Zhang [a,*], Ivan Marsic [a], Randall S. Burd [b]

[a] Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854, USA
[b] Division of Trauma and Burn Surgery, Children's National Medical Center, Washington, DC 20010, USA

ARTICLE INFO

Article history:
Received 9 September 2020
Revised 31 August 2021
Accepted 2 September 2021
Available online 3 September 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Phase recognition
Trauma resuscitation
Deep learning
Video understanding
Reduced long-term operation
Process gate

ABSTRACT

We introduce a real-time system for recognizing five phases of the trauma resuscitation process, the initial management of injured patients in the emergency department. We used depth videos as input to preserve the privacy of the patients and providers. The depth videos were recorded using a Kinect-v2 mounted on the sidewall of the room. Our dataset consisted of 183 depth videos of trauma resuscitations. The model was trained on 150 cases with more than 30 minutes each and tested on the remaining 33 cases. We introduced a reduced long-term operation (RLO) method for extracting features from long segments of video and combined it with the regular model having short-term information only. The model with RLO outperformed the regular short-term model by 5% using the accuracy score. We also introduced a progress gate (PG) method to distinguish visually similar phases using video progress. The final system achieved 91% accuracy and significantly outperformed previous systems for phase recognition in this setting.

1. Introduction

Trauma is the leading cause of mortality in children and young adults (Kaplan (2002)). The initial resuscitation of injured patients is critical for identifying and managing life-threatening injuries. Despite the use of a standardized protocol, errors remain frequent during this initial evaluation (Rodziewicz and Hipskind (2020); Wolf and Hughes (2008)). Computerized decision support has been proposed as a method for reducing errors in this setting (Jia et al. (2016); Reis et al. (2017); Castaneda et al. (2015)). Trauma resuscitation is divided into phases based on the prioritization of activities within each phase. The pre-arrival phase is focused on preparation for the patient, the primary survey for identifying and managing life-threatening injuries, the secondary survey phase for identifying additional injuries that need management, and the post-secondary phases for initiating additional injury management. Although some activities are shared between phases, the types and order of many activities differ between phases. Identification of phases aids in the determination of errors in the type and order of activities. Decision support in this domain should reflect the priorities of each phase. Knowledge of the current phases aids in the prioritization of required activities based on the underlying goals in each. The duration of each phase varies across resuscitation, preventing the use of fixed time points to separate the phases. An automatic phase recognition system is needed to address this challenge. Our real-time phase recognition system uses depth video as input (Fig. 1). Building on modeling experience with deep learning in the field of computer vision, video classification has been rapidly developed for activity recognition. Compared to activity recognition, recognizing phases in medical settings has three challenges. First, the system needs to manage privacy considerations because RGB videos reveal patient and providers' faces. Second, the system training needs to rely on small datasets because of the increased time requirement of annotating and limited access to videos of patient care compared to general activities. Third, the system needs to rely on long-term context because the phases are defined by the occurrence of multiple and often overlapping activities that may occur in several phases. Models using short-term input (individual frames or subsequent frames) will make erroneous predictions because the same or visually similar activities may occur in different phases.

* Corresponding author.
E-mail address: yz593@scarletmail.rutgers.edu (Y. Zhang).

https://doi.org/10.1016/j.media.2021.102224
1361-8415/

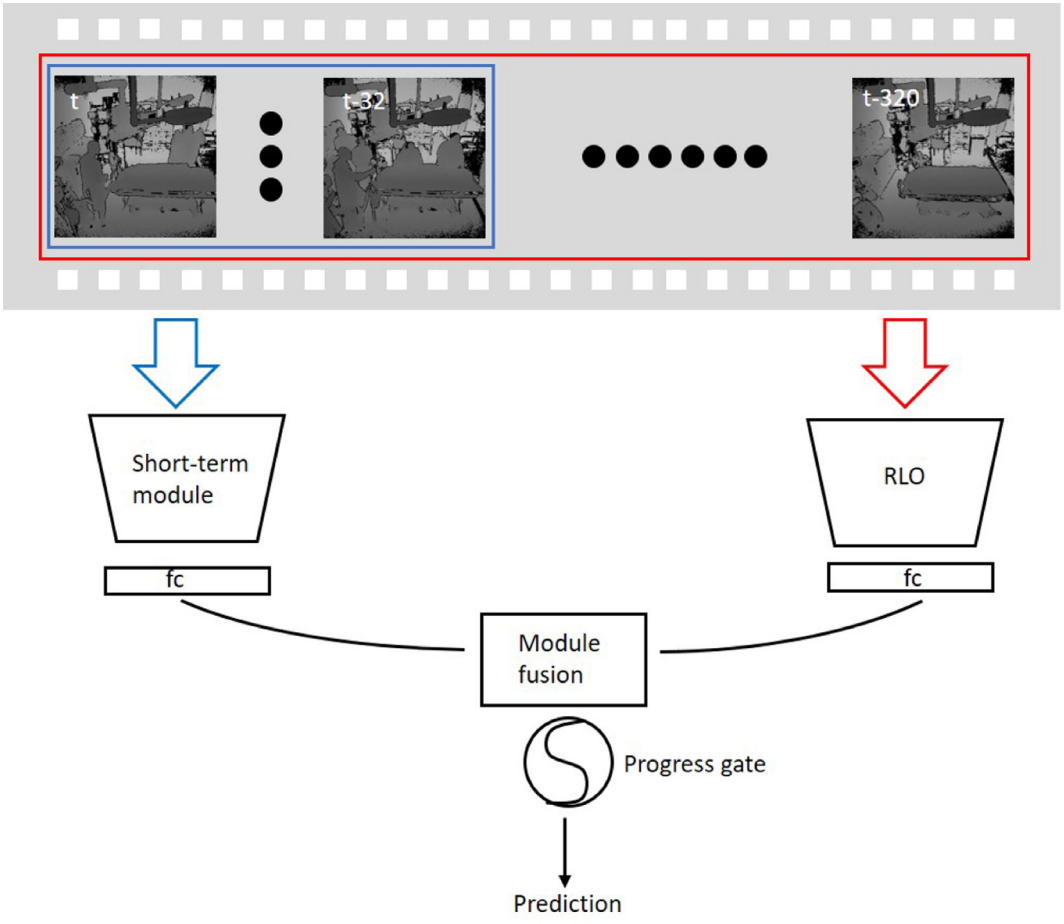

Fig. 1. Overview of our phase recognition system. The system takes depth videos as input, and the phase is predicted using 32-second-frame inputs to the short-term module and 320-second-frame inputs to RLO. The outputs of the short-term and long-term modules are fused for making the final prediction.

To address the first challenge, we used depth instead of RGB videos. The depth videos contain gray-scale images, with pixel values denoting the distances between the objects and the camera. These gray-scale images do not contain recognizable facial textures but include the contour of the people and objects relevant to activity performance. To address the challenge of limited data, we used transfer learning, pre-training the model using public large activity datasets (Kay et al. (2017)). We then fine-tuned the model weights using our smaller dataset. This approach could be used because activity and phase recognition rely on similar low-level features, such as the presence of people and objects and the occurrence of associated gestures. We evaluated our system using two model structures, an inflated 3D ConvNet (i3D) (Carreira and Zisserman (2017)) and a nonlocal neural network (NL) (Wang et al. (2018)). These two model structures have achieved state-of-the-art performance on the Kinetics-400 dataset for activity recognition(Kay et al. (2017)), making them an appropriate starting point. Previous research introduced a 2-stage CNN-RNN structure, where CNN pre-computes features followed by a RNN, which learns temporal dependencies among features (Al Hajj et al. (2018)). Their method is not end-to-end trainable which causes an error propagating problem, and RNN-based networks are slow and have information loss problems when given long-range inputs. (Vaswani et al. (2017)). To enable the model learn long-range video contexts better, we introduced a reduced long-term operation (RLO) method (Fig. 1, right) that uses frame inputs from a long time sequence (320 seconds). Some phases have similar visual features that persist for a relatively long time, making it difficult for the model to distinguish them using RLO. For example, the pre-arrival (before patient arrival) and the patient departure (after patient leave) phases have similar visual characteristics. People can easily distinguish these phases using the time since the start of the process. Previous research proposed a phase-inference network (RSDNet) to predict the surgical progresses (Twinanda et al. (2018)). We introduced a progress gate (PG) method (Fig. 1, bottom), which is using the estimated process progress for phase recognition unlike the RSDNet which used the phase information to predict the process progress.

1.1. Related work

Medical Phase Recognition: Medical phases often define the progress of a medical event. For surgical procedures, several phases can be defined, including preparation, execution, and termination phases. For protocols such as Advanced Cardiovascular Life Support (ACLS) and Advanced Trauma Life Support (ATLS), phases can be defined based on the choice and priority of management and treatment activities (Kortbeek et al. (2008)). Phase prediction in medical settings can be used for several purposes, including targeting recommendations based on the current phase (context-aware), comparison of performance between individuals and teams, and estimating process duration for workflow tracking and improvement (Li et al. (2016); Twinanda et al. (2016a); Bardram et al. (2011)). Previous work in phase recognition has achieved good results using body-worn sensors (Ahmadi et al. (2008); Meißner et al. (2014)). In a medical setting, wearable sensors may require the active participation of providers or may interfere with the performance of medical

tasks, potentially limiting the usability of this approach. Computer vision has advantages over wearable sensors by relying on data from fixed cameras without interfering with the conduct of the medical event. Video images are a rich source of information about phases and may enhance performance in a context in which wearable sensors are impractical. Deep convolution networks have been used to recognize surgical phases using laparoscopic and ocular videos (Twinanda et al. (2016a); Yengera et al. (2018); Loukas (2018); Zisimopoulos et al. (2018); Chen et al. (2018); Loukas (2018)). These studies showed that video-based systems work well on phase recognition without requiring wearable sensors that may interfere with work. Surgical phase recognition has used videos focused on specific regions around medical tools (Twinanda et al. (2016a); Zisimopoulos et al. (2018); Chen et al. (2018)). For example, the cholecystectomy dataset (Chen et al. (2018); Twinanda et al. (2016a)) contains videos from a laparoscopic view, while the CATARACTS dataset (Zisimopoulos et al. (2018)) includes only video of the orbital region during cataract surgery. In contrast to this previous work, phase recognition in a team-based medical setting requires video that covers the entire scene for recognition of activities relevant to each phase. A real-time state identification system in operating rooms has been proposed using RGB videos (Bhatia et al. (2007)). Because these scenes include the patient and the individuals providing medical care, the use of RGB videos has privacy concerns that needed to be addressed. Several strategies have been used to manage concerns with RGB videos including the use of using extremely low-resolution images to anonymize faces (Dai et al. (2015); Ryoo et al. (2018); Ren et al. (2018)). An alternative approach for ensuring that images do not allow individual detection is the use of depth videos that include gray-scale images. This representation makes it difficult to identify individuals but may be sufficient for recognizing activities and phases (Li et al. (2016, 2017b)).

Medical Workflow Analysis using Depth Videos: Depth videos contain gray-scale frames that represent the distance between the camera and objects in the scene. Previous research on monitoring hand hygiene, human pose, and patient mobilization activities in Intensive Care Unit (ICU) used depth videos instead of RGB due to the privacy concerns in ICU (Srivastav et al. (2019); Yeung et al. (2019); Reiter et al. (2016); Yeung et al. (2016)). Other research used RGBD videos that rely on distance information in depth images to improve system performance on surgical phase recognition and activity recognition in operating rooms (Twinanda et al. (2015, 2016b)).

Video Understanding: In many settings, activities and phases are continuous rather than fixed point events. Recognition of these components of human work benefits from analysis of spatio-temporal features available in videos. Detection of dynamic components of work differs from standard image recognition that only requires spatial features from a single image. Several model structures are available for extracting spatio-temporal features for activity recognition in videos. The Two-stream and CNN-LSTM network structures have been used for large-scale video classification and activity recognition by extracting temporal associations between subsequent frames (Wang et al. (2016); Simonyan and Zisserman (2014); Feichtenhofer et al. (2016); Li et al. (2017a); Mutegeki and Han (2020)). Recent works applied 3D Convolution structures for video understanding, supported by 3D ConvNets being end-to-end trainable (unlike Two-stream networks) and allowing parallel computing (unlike CNN-LSTM networks) (Carreira and Zisserman (2017); Tran et al. (2015)). The Slow-Fast, and the Channel-separated networks were proposed for reducing computational complexity for training 3D ConvNets (Feichtenhofer et al. (2019); Tran et al. (2019)). The X3D expands hyper-parameters of 3D Convolution architectures for building efficient video recognition networks (Feichtenhofer (2020)). The non-local neural network also has been used to obtain long-range associations between distant pixels by including nonlocal blocks into the 3D ConvNets. This type of structure has achieved better performance for activity recognition than the i3D network (Wang et al. (2018)). Although these approaches perform well for recognizing activities, additional spatio-temporal information from longer video context is needed for recognizing phases rather than any single activity. An additional challenge is that some activities may be performed in different phases, limiting the use of short-range spatio-temporal features.

1.2. Contributions

We introduce a real-time phase recognition system that can be used to provide contextual information that supports a context-aware recommend system for trauma resuscitation. This system is privacy-preserving and extends previous preliminary works (Li et al. (2016, 2017b)) as follows:

- We applied recent video understanding methods that extract spatio-temporal features from consecutive frames instead of spatial-only features from static images for recognizing phases. Our system significantly outperformed our previous systems (Li et al. (2016, 2017b)).
- We introduced a RLO strategy that increased the performance by extracting long-term spatio-temporal features for phase recognition.
- We introduced a PG method that allows the model to distinguish visually similar phases using estimated video progress as an additional input.
- We collected depth videos and created their corresponding ground truth for more trauma resuscitation cases (183 cases vs. 60 cases Li et al. (2017b)). The system evaluated on larger testing set is more convincing.

This paper is organized as follows. Section 2 describes our phase recognition system. Section 3 presents data collection and the implementation details. Section 4 and Section 5 shows the experiment results. Section 6 discuss the model visualization results, and Section 7 concludes the paper.

2. Methodology

2.1. Method overview

We represented a video input as $(T, W, H, 1)$, that includes T consecutive frames, each with three dimensions: width (W), height (H), and the color channel. We applied these inputs to recognize medical phases in three stages. We first trained short-term spatio-temporal models that take 32-second depth frames as input and each phase as output (Fig. 2, up). We next applied a novel reduced long-term operation (RLO) method to learn long-range contexts from the video (Fig. 2, bottom). This method takes long-range history frames (320 seconds) as input for tuning the long-term module branch. We then fused the predictions between using short-term and long-term spatio-temporal features to generate the final phase predictions. Finally, we applied the progress gate (PG) after the fused predictions to help the model distinguish visually similar phases using estimated video progresses (Fig. 2, middle left).

2.2. Short-term module

The short-term module takes 32-second consecutive depth frames as the input and extracts spatio-temporal features for phase recognition. We evaluated the short-term module using two spatio-temporal network structures, the inflated 3D ConvNet (i3D) and the nonlocal neural network.

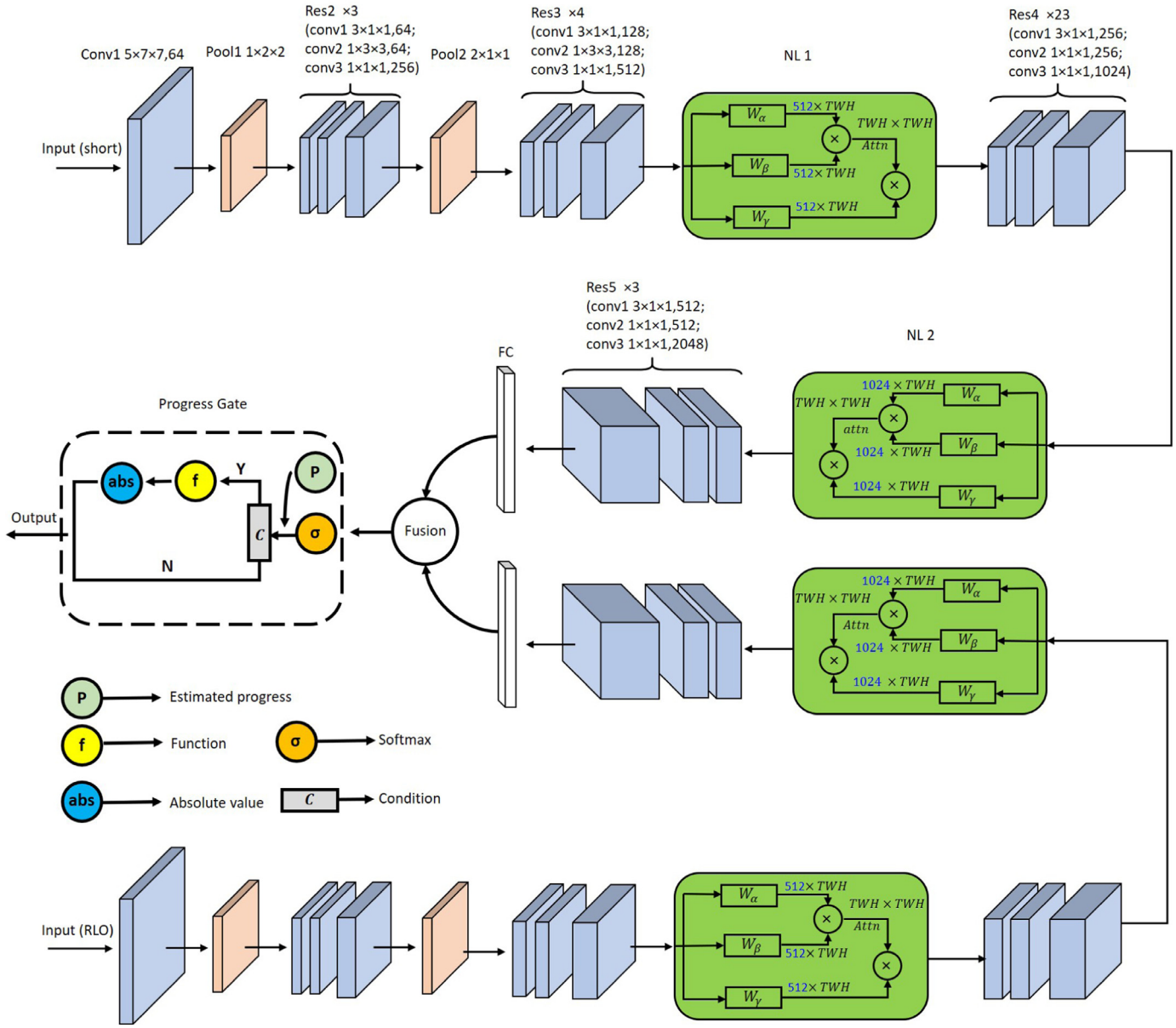

Fig. 2. Network structure of our phase recognition system. Shown are the convolution kernel sizes for each network stage (Conv1-Res5) and the dimension transformation of the features in the nonlocal blocks. The W_α, W_β, and W_γ are the parameters of nonlocal blocks in equation 2 and 3. The "f" and "C" in the progress gate module (dash-lined block on the left of the middle row) are the function and the condition introduced in Section 2.5..

2.2.1. Inflated 3D ConvNet

Inflated 3D ConvNet (i3D) (Carreira and Zisserman (2017)) is a spatio-temporal structure that extends successful 2D image recognition models (Inception v1) into 3D ConvNets with an additional temporal dimension. 3D ConvNet learns spatio-temporal features from the video input as:

$$\begin{aligned} X_f(k,j,i) &= Conv3D(X(k+t, j+h, i+w), \theta) \\ &= \sum_t Conv2D(X(k+t, j+h, i+w), \theta_t) \end{aligned} \quad (1)$$

where X is the input spatio-temporal feature descriptors, $X_f \in \mathbb{R}^{TWH \times F}$ is the output feature map of the 3D ConvNet, $X_f(k,i,j)$ is a feature point in the 3D feature space, θ denotes the parameters of the 3D convolution, T is the number of consecutive frames in each input, and F is the number of channels in the feature map X_f. Our i3D network is extended from the ResNet-101 (He et al. (2016)), which is the 2D image recognition network that achieved the first place on the ImageNet challenge (Deng et al. (2009)). Table 1 shows the detail network structure and parameters of the i3D network that we used. The network includes five stages ($Conv_1$, and $Res_2 - Res_5$), Res_n denotes the bottleneck block including three 3D convolution layers. The i3D also benefits from loading the pre-trained 2D convolution parameters that have already learned spatial features on image classification datasets and duplicating the 2D convolution kernels T times for generating 3D convolution kernels. Learning spatio-temporal features by fine-tuning the well-learned spatial features converges faster than training from scratch by randomly initializing the 3D convolution parameters.

2.2.2. Nonlocal neural network

The attention mechanism was introduced for capturing long-term dependencies within sequential inputs, which is commonly used in nature language processing systems, such as text classification, and machine translation (Vaswani et al. (2017);

Table 1
The detail structure and parameters of the i3D network that we are using.

Stage	Details	Output Size
$Conv_1$	$5 \times 7 \times 7$, 64, stride 1, 2, 2	$32 \times 112 \times 112 \times 64$
$Maxpool_1$	$2 \times 3 \times 3$, stride 2, 2, 2	$32 \times 56 \times 56 \times 64$
Res_2	$\begin{pmatrix} 3\times1\times1, 64 \\ 1\times3\times3, 64 \\ 1\times1\times1, 256 \end{pmatrix}\times3$	$16 \times 56 \times 56 \times 256$
$Maxpool_2$	$2 \times 3 \times 3$, stride 2, 2, 2	$8 \times 28 \times 28 \times 256$
Res_3	$\begin{pmatrix} 3\times1\times1, 128 \\ 1\times3\times3, 128 \\ 1\times1\times1, 512 \end{pmatrix}\times4$	$8 \times 28 \times 28 \times 512$
Res_4	$\begin{pmatrix} 3\times1\times1, 256 \\ 1\times3\times3, 256 \\ 1\times1\times1, 1024 \end{pmatrix}\times23$	$8 \times 14 \times 14 \times 1024$
Res_5	$\begin{pmatrix} 3\times1\times1, 512 \\ 1\times3\times3, 512 \\ 1\times1\times1, 2048 \end{pmatrix}\times3$	$8 \times 7 \times 7 \times 2048$

Shen et al. (2018); Bahdanau et al. (2014)). The nonlocal neural network (Wang et al. (2018)) extends i3D by inserting nonlocal blocks between the stages in the i3D network that learns long-term spatio-temporal features from the feature maps extracted by 3D convolution by generating spatio-temporal attentions as:

$$Attn = softmax(X_f^T W_\alpha^T W_\beta X_f) \tag{2}$$

$$X_{nl} = Attn(W_\gamma X_f) + X_f \tag{3}$$

where $X_{nl} \in \mathbb{R}^{TWH \times F}$ was the output after applying the nonlocal block, $Attn \in \mathbb{R}^{TWH \times TWH}$ was the spatio-temporal attention that represents the association between pairs of positions in X_f, W_α, W_β, and W_γ are the parameters of the linear functions, and $+X_f$ denotes the residual operation between X_f and the output after applying $Attn$ on X_f. Nonlocal neural network learns long-rang spatio-temporal features using $Attn$. The attention $Attn$ was generated using batch matrix multiplication between two linear projections of the input X_f ($W_\alpha X_f$ and $W_\beta X_f$) that captures the association between two points in x, regardless of their distance.

The two networks (i3D and nonlocal neural network) are pretrained on Kinetics-400, a large-scale video set for activity recognition (Kay et al. (2017)). Pre-training the network using general large-scale datasets achieves better performance than training the network only using the available limited domain-specific data (Carreira and Zisserman (2017); Wang et al. (2018)). Although Kinetics-400 is an activity recognition dataset that is somewhat different from phase recognition, these two phenomena (activity and phase) share similar low-level features such as edges, objects contours and personal motions. To predict the phases, we then applied a fully-connected layer that takes the extracted spatio-temporal features as input.

2.3. Reduced long-term operation

3D convolution extracts spatio-temporal features from videos and the long-range dependencies in the feature maps can be captured using nonlocal blocks. This information, however, is constrained by the input duration (32 seconds). The short-term video inputs are sufficient for activity recognition because most activities are performed within seconds. Phase recognition requires longer-duration video contexts. Multiple activities may be performed during a phase, and the same activity may occur in different phases. For example, during trauma resuscitation, the blood pressure measurement may be performed in both the primary survey phase and the secondary survey phase. In this case, short-term inputs that contain features for this activity may be labeled as different phases (primary survey and secondary survey), which would confuse the model. A straightforward solution to this problem is to enlarge the input duration. This approach, however, increases the complexity of training and evaluating the model. We introduced a reduced long-term operation (RLO) method that enables the model to learn features from long-range video contexts without increasing the model complexity.

The input to our reduced long-term operation (RLO) method were the video frames over the last 320-seconds before the current time. We did not use the frames after the current time to enable the model to generate online predictions. To reduce the model complexity when using longer video inputs, we increased the down-sample rate of the inputs as:

$$x_{long} = \{x_\alpha, x_{2\alpha}, \ldots, x_{T\alpha}\} \tag{4}$$

where x_{long} denotes the long-range frame inputs of the RLO, α is the down-sample rate, and T is the frame number ($\alpha = 10$, and $T = 32$). An additional fully-connected layer provided phase predictions that takes the long-term spatio-temporal features. The 320-second inputs in RLO are frames constructed by the current 32-second frames and the preceding 288-second frames (from history). The historical frames help the model eliminate implausible predictions. For example, a prediction of the secondary survey phase cannot be made based on the inputs that having historical frames that occur before the primary survey phase. The model using RLO achieved accuracy that was 5% higher than using short-term module only.

2.4. Module fusion

The next step of our system was to fuse the outputs from the short-term and long-term modules. The long-term module provides more accurate predictions because of the long-range inputs. It will not produce phase predictions during the first 320 seconds until it observes a sufficient past interval. We used the short-term module to provide phase predictions during these 320 seconds and fused the short-term and long-term modules for the predictions of the remaining time. We used the output-level fusion to aggregate the outputs of the long-term and short-term modules as:

$$y_{fuse} = y_{short} + y_{long} \tag{5}$$

$$\hat{y} = softmax(y_{fuse}) \tag{6}$$

where y_{short} and y_{long} are in R^5, and they denote the outputs of the short-term and long-term modules, respectively. $\hat{y}$ denotes the model output by applying softmax function over y_{fuse}. We also evaluated the potential use of multi-modal fusion strategies (e.g., by concatenating or using nonlocal gates (Wu et al. (2019)) to merge the features outputted by the long-term and short-term modules. We did not adopt these multi-modal fusion strategies in our system because our evaluation showed that their use imposed higher computation cost without a performance increase.

2.5. Progress gate

During trauma resuscitation, several phases may have similar visual appearance over long intervals, making it difficult for the model to distinguish them even with RLO. For example, in our dataset, the pre-arrival and the patient departure phases look similar in some cases, but people can distinguish them based on the current progress of the video as additional information. We therefore applied the progress gate (PG) after the fused predictions by using estimated progress for additional input as:

$$\hat{y}_p^{'} = \begin{cases} \frac{1}{5}(1 - \hat{y}_p) & \text{, if } C \\ \hat{y}_p & \text{, otherwise} \end{cases} \tag{7}$$

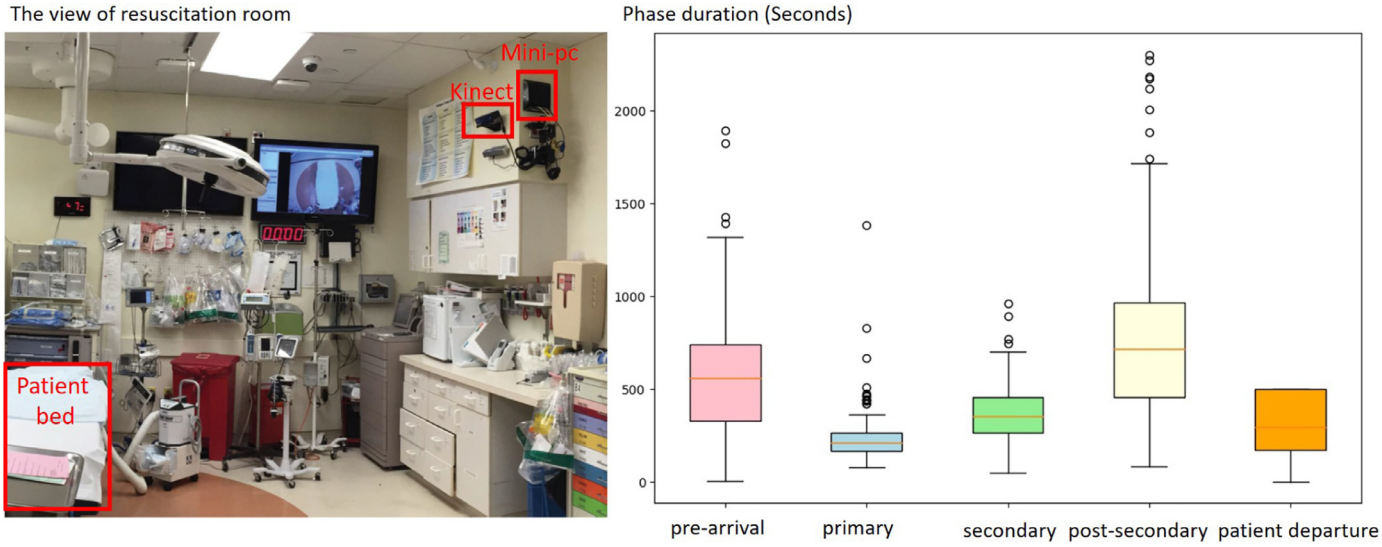

Fig. 3. The view of trauma resuscitation room (left) and the duration boxplot (right) for the five phases in our dataset (in seconds). The patient departure does not have an upper whisker (UW) because we truncated the videos 500 seconds after the patient left the room.

where $\hat{y} \in \mathbb{R}^5$ is the model prediction, $p \in 0, \ldots, 4$ is the element of $\hat{y}$ that denotes the p^{th} phase, and C is the condition which is represented as:

$$C : argmax(\hat{y}) = p \text{ and } (\lambda < \lambda_{p\_min} \text{ or } \lambda > \lambda_{p\_max}) \quad (8)$$

where λ denotes the estimated progress of the current video by dividing the current time played with the average duration of the videos in our training set. We used the estimated progress instead of the progress relative to the total length of the current video to be able to use this system for real-time phase prediction. The λ_{p_min} and λ_{p_max} are the lowest and highest estimated progress values for the p^{th} phase across all the cases in the training set. We multiplied $1 - \hat{y}_p$ by $\frac{1}{5}$ (5 is the number of phases) in Eq. 7 to ensure that the phase p will not be selected as the prediction when the condition C is satisfied. In some cases, $1 - \hat{y}_p$ will still make $argmax(\hat{y}_p) = p$ (e.g., $\hat{y} = [0.5, 0.2, 0.1, 0.05, 0.05]$, and both $argmax(\hat{y})$ and $argmax(1 - \hat{y}_p)$ are equal 0). Multiplying $1 - \hat{y}_p$ by $\frac{1}{5}$ will make at least one other phase have a larger prediction score than the phase p. Note that multiplying by a smaller number than $\frac{1}{5}$ or re-setting $\hat{y}_p$ to 0 would have the same result. $\frac{1}{5}$ is a boundary case when one phase is assigned the maximum possible score $\hat{y} = [0.5, 0.1, 0.1, 0.1, 0.1]$. Then, $\frac{1}{5}(1 - \hat{y}_0) = 0.1$ and $max(\hat{y}_{1:4}) = 0.1$.

3. Data collection and implementation details

3.1. Data collection

We evaluated our system using videos of trauma resuscitations conducted at a level 1 trauma center. This research was approved by the hospital's Institutional Review Board (IRB).We installed a Microsoft Kinect V2 for capturing depth videos and connected it to a local computer for controlling the recording and storing videos (Fig. 3 left). We mounted the Kinect on the sidewall of the room at a position 2.5 m above the ground and tilted it downwards at 20°. We applied the build-in skeleton detection function from the Kinect API on our system to detect the number of persons in the view. The system is triggered and begins recording after the Kinect detects more than two people in view for at least one minute. This triggering function is required to decrease data storage needs and avoid the need to manually start recording, a task that can easily be forgotten in this type of setting. After the camera is triggered, the system stores a depth frame (Fig. 4 right) every second and stops after the Kinect detects that no person is present in the room for more than one minute. We collected depth videos for 183 trauma resuscitation cases, using 150 cases for training and 33 cases (20%) for testing. We segmented the videos into 32-second clips (32 consecutive frames) using a 16-step sliding window (overlapped by 16 seconds). Table 2 shows the number of clips for different phases in both training and testing set. Ground truth labelling was performed by manual reviewing RGB videos (without using audio), based on predetermined definitions of the process phases (RGB videos were not available for model training). Each video was annotated independently by the three providers, and any conflicting annotations were resolved by consensus. When there is a phase transition in an input clip, the system assign the clip to the phase that dominates the clip (having longer duration).

Table 2
Number clips for different phases in training and testing sets. Each clip contains 32-second consecutive frames.

Phase Name	Num Clips (train)	Clip Num (test)
Pre-arrival	2150	490
Primary Survey	1077	205
Secondary Survey	1712	375
Post-Secondary Survey	3697	1095
Patient Departure	1973	454

In contrast to other medical processes, trauma resuscitation is a highly structured process that is taught as part of the Advanced Trauma Life Support (ATLS) protocol (Kortbeek et al. (2008)). Although rare deviations may occur because of unusual patient conditions or provider error, this phase structure is consistently observed during trauma resuscitation. This consistency makes this domain ideal for phase-based decision support. The system was designed to recognize five phases of the trauma resuscitation process: pre-arrival, primary survey, secondary survey, post-secondary survey, and patient departure (Table 3). The pre-arrival phase occurs in the time between a notification that an injured patient will be arriving and the arrival of the patient in the room. During this phase, a multidisciplinary team of up to 15 individuals assembles and begins preparing equipment needed for evaluating and treating the patient. The endpoint of this phase is defined as when the patient is moved from the prehospital gurney to the hospital

Fig. 4. The RGB (left) and depth (right) view of trauma resuscitation room from the Kinect.

Table 3
Description of five medical phases that our system recognize.

Pre-arrival	Start: When first personnel member enters the room End: When the patient enters the room
Primary Survey	Start: When first primary survey or primary survey related task (i.e. warm sheet placement) begins End: When examining provider performs first secondary survey task.
Secondary Survey	Start: When examining provider performs first secondary survey task. End: When the last secondary survey task is performed in the normal progression of the examination. Secondary survey tasks and secondary survey adjunct tasks that are completed after the secondary survey has been conducted should not be used as the end time. (i.e. if the examining provider completes the secondary survey and returns to re-evaluate an injury minutes later, the second occurrence should not be used as the end time.
Post-Secondary Survey	Start: When the last secondary survey task is performed in the normal progression of the examination. Secondary survey tasks and secondary survey adjunct tasks that are completed after the secondary survey has been conducted should not be used as the end time. (i.e. if the examining provider completes the secondary survey and returns to re-evaluate an injury minutes later, the second occurrence should not be used as the end time. End: When only the patient's head remains visible in the "foot view" video frame (rest of body already through the doorway) *if pt dies exit is time of death and label exit with "death" attribute)
Patient Departure	When only the patient's head remains visible in the "foot view" video frame (rest of body already through the doorway) *if pt dies exit is time of death and label exit with "death" attribute)

bed. The primary survey phase then begins. The primary survey includes a series of activities that are performed for identifying and immediately managing potentially life-threatening conditions. The activities within these phases follow five steps: acronym-named as A through E which stands for airway assessment and management (A), evaluation of adequacy of ventilation/breathing (B), assessment of circulatory status and perfusion (C), assessment of neurological status/disability (D), and the complete exposure of the patient for visual inspection of injuries (E). These five steps occur in this order in most resuscitations unless patient requirements require omission or delay of a step until later in the resuscitation. The secondary survey follows the primary survey. This phase is a head to toe physical examination focused on identifying additional injuries not found in the primary survey. The post-secondary phase begins at the completion of this assessment. The patient departure phase begins when the patient leaves the room, a period when the members of the team may remain to clean and prepare the room for another patient. The phases of trauma resuscitation are sequential (Fig. 5). Although overlap occurs between some activities in each phase, phase order is preserved across resuscitation. The duration of the five phases vary through different cases (Fig. 3, and Fig. 5). In some cases, the Kinect built-in function wrongly detected some background objects as people and made the system to keep recording after the patient departed. This type of event caused the label unbalance issue because the patient departure phase was extremely long in these cases. We solved this problem by truncating the videos 500 seconds after the patient left the room.

3.2. Implementation details

We implemented our model using the Pytorch framework. We set the length of input video clips as 32 consecutive frames to match the input size of the pre-trained networks and expand to 320 frames for the long-term branch (RLO). We added a batch normalization after every convolutional layer to speed up the model convergence (Ioffe and Szegedy (2015)). A ReLU was used as the activation function. Adam (Kingma and Ba (2014)) was used as the optimizer with the initialized learning rate of 1e-4, and 1e-8 as the weights decay. We set the batch size to 12 (constrained by the GPU memory size) and trained the model for 14k iterations. The model was trained using three RTX 2080 ti and required about one day to converge. To avoid overfitting, we applied the scale-jittering method in range of [256, 320] to augment the frames in spatial (Feichtenhofer et al. (2019)). We also applied Dropout (Srivastava et al. (2014)) after the fully-connected layers to avoid overfitting.

4. Experimental results

4.1. Experimental results overview

Fig. 6 shows the confusion matrices of our system for prediction five phases. Based on the confusion matrices (Fig. 6), our system performed best on the pre-arrival and the patient departure phases. During the pre-arrival, fewer than three people are typically in the trauma room, and no patient is on the bed. These features are visually recognizable. During the patient departure phase, the patient's bed has been wheeled out of the room, a feature providing a strong visual cue. When the patient bed stayed in the room after the resuscitation, but the patient has left, these two phases were sometimes confused. Prediction of the post-secondary phase (Fig. 6, row 4) was slightly worse because of the confusion between the secondary and post-secondary phases. During the post-secondary survey phase, the patient is still on the bed

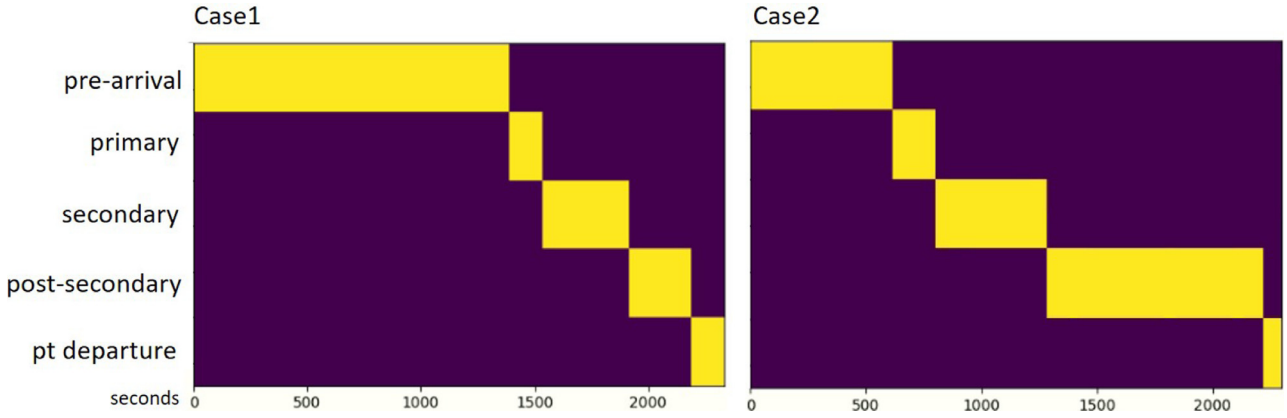

Fig. 5. The workflow of the two cases. The phases' duration varies between different cases.

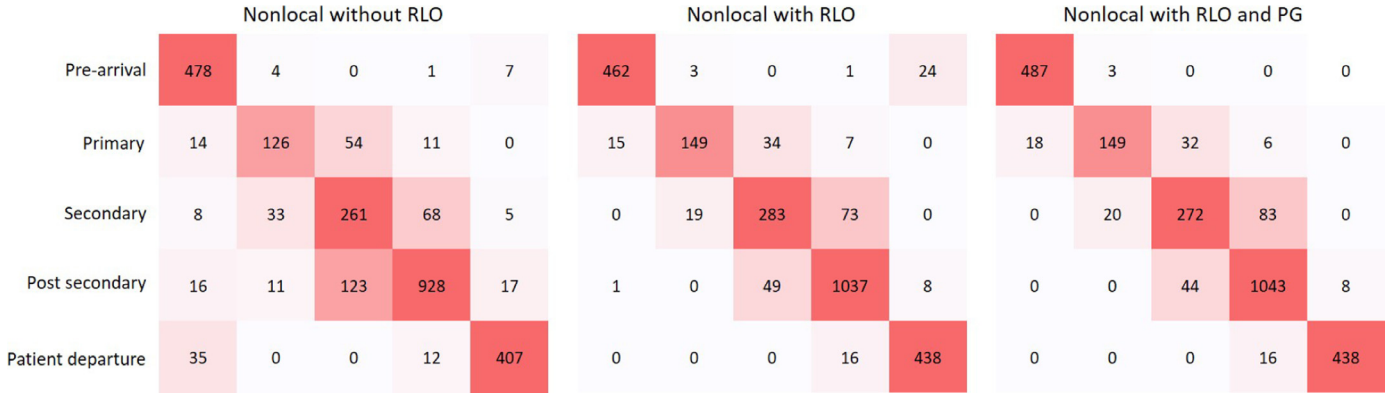

Fig. 6. Confusion matrices for phase recognition using nonlocal network. The values in the confusion matrices denote the number of input clips across the 33 testing cases. The left diagram is the confusion matrices using nonlocal network without RLO and PG, the middle diagram is the confusion matrices using nonlocal network with RLO only, and the right diagram is the confusion matrices with both RLO and PG.

and only a few providers remain in the room. The lowest performance was achieved on the primary and secondary survey phases. These two phases are difficult to distinguish based on depth video because detailed visual textures are not available that help a human reviewer using RGB videos. Human reviewers presented with depth video had the most difficulty identifying the transition between these two phases. Using RLO (Fig. 6 middle) significantly increased the detection of the primary survey because this phase is relatively short. This short duration caused the long-range inputs (320-second frames) for the primary survey to partially include views from the pre-arrival phase and helped the model distinguish the primary survey from the secondary survey. The model using PG (Fig. 6 right) eliminated the incorrect predictions between the pre-arrival and the patient departure phases because the estimated progress of the video helped distinguish these two phases.

4.2. Ablation study

We performed ablation experiments on phase recognition for comparing the performance using different network structures and hyper-parameters.

Network structures: We evaluated our model using three different network structures (ResNet2D-101, i3D, and Nonlocal) that were introduced for image and video recognition (He et al. (2016); Carreira and Zisserman (2017); Wang et al. (2018)). The ResNet2D-101 achieved the worst performance because the model recognized phases using single-frames as input without considering the context between the consecutive frames (Table 4a). The nonlocal network slightly outperformed the i3D network because of the long-range spatio-temporal associations captured by the nonlocal blocks. The nonlocal network with RLO and PG (Table 4a, last row) significantly outperformed the nonlocal network without (Table 4a, second last row) because the RLO helps the model to learn spatio-temporal features from a longer video context for phase recognition rather than from short-term inputs (320 vs. 32 seconds). The PG also helps the model to distinguish visually similar phases using video progress.

RLO input length: We also evaluated our model using RLO with different input lengths (Table 4b). The model using 320-second-frame inputs for the RLO achieved the best performance. Inputs from longer video contexts contained more information but had a lower temporal resolution that lacked the continuity of the videos. The model achieved the best performance with 320-second input duration and a decreased performance with inputs longer than 320 seconds (Table 4b, last row).

4.3. Comparison with previous systems

We compared our system with two previous systems for phase recognition during trauma resuscitation, both that used depth videos as input (Li et al. (2016, 2017b)). We evaluated our system on both the smaller video set (Table 5, 50/10 train/test) that the previous systems used and our larger video set (Table 5, 150/33 train/test). Our system outperformed these systems (Table 5) because by the use of spatio-temporal network structures and our proposed methods (RLO and PG) for capturing features from long-range video contexts and including estimated video progress as additional input (Table 5). One system (Table 5, first row) applied a spatial-only network structure previously used for single image recognition, which did not include the temporal asso-

Table 4
Ablation experiments on phase recognition. We show the accuracy, F1, precision and recall scores by using different network structures and hyper-parameters.

Network	Acc.	F1.	Prec.	Rec.
ResNet2D-101	74.9	74.8	75.3	75.0
I3D	82.1	81.4	81.8	82.9
Nonlocal	83.7	83.1	82.6	84.0
Nonlocal + RLO	89.3	89.1	89.1	89.3
Nonlocal + PG	84.6	84.3	84.4	85.7
Nonlocal + RLO + PG	**90.8**	**90.6**	**90.6**	**90.8**
Network structure: performance on phase recognition when using different network structures.				
RLO Inputs (seconds)	Acc.	F1.	Prec.	Rec.
64	83.9	83.5	84.0	83.8
160	84.0	83.9	84.6	84.0
320	**89.3**	**89.1**	**89.1**	**89.3**
640	89.0	88.7	88.7	89.0
RLO inputs: performance on phase recognition when using different input length for the RLO method.				

Table 5
Experimental results and comparison with previous work. The evaluation results are accuracy, F1-score, precision and recall in percentages. The column "Data set" denotes the number of cases that were used for training and testing.

Method	Data set	Online	Acc.	F1.	Prec.	Rec.
CNN Frame-wise (Li et al. (2016))	50/10	Yes	67.5	-	-	-
CNN Frame-wise + constrain (Li et al. (2016))	50/10	Yes	80.0	70.0	72.0	76.0
CNN-LSTM + GMM (Li et al. (2017b))	50/10	No	86.0	72.0	69.0	67.0
Nonlocal	50/10	Yes	87.2	87.1	87.5	86.9
Nonlocal + RLO + PG	50/10	Yes	92.1	90.9	91.7	91.1
Nonlocal	150/33	Yes	83.7	83.1	82.6	84.0
Nonlocal + RLO	150/33	Yes	89.3	89.1	89.1	89.3
Nonlocal + RLO + PG	150/33	**Yes**	90.8	90.6	90.6	90.8
Nonlocal + RLO + PG (filtered)	150/33	No	**91.2**	**90.9**	**91.1**	**91.4**

Table 6
Phase independent evaluation: we compared our system with other previous systems using independent F1-scores of each phase.

Network	Data set	Online	Pre-arrival	Primary	Secondary	Post-secondary	Pt-departure
CNN Frame-wise + constrain (Li et al. (2016))	50/10	Yes	80.0	49.0	43.0	76.0	77.0
CNN-LSTM + GMM Li et al. (2017b)	50/10	No	61.0	56.0	72.0	43.0	94.9
Nonlocal + RLO	50/10	Yes	95.2	**82.7**	82.2	94.6	90.3
Nonlocal + RLO + PG	50/10	Yes	**98.8**	82.1	**82.9**	**94.6**	**97.2**
Nonlocal + RLO	150/33	Yes	95.5	**79.3**	**76.4**	93.1	94.8
Nonlocal + RLO + PG	150/33	Yes	**97.9**	79.0	75.2	**93.1**	**97.3**

ciations between consecutive frames (Li et al. (2016)). This system (Li et al. (2016)) applied a constrained softmax to eliminate the illegal predictions from the model output. We tried to apply this constrained softmax method in our system, but it worsened performance. This decrease may have occurred because the constrained softmax depends on the predictions of the preceding models, which will increase the error rate when these models made incorrect predictions. The second system (Table 5, third row) estimated progress using depth videos as input and then predicted phases using the generated progress (Li et al. (2017b)). The errors propagated from the progress estimation step may have resulted in incorrect phase prediction. In addition, the second system used a filtering algorithm to smoothen the generated progress and enhance the performance of phase prediction (Li et al. (2017b)). This method can only be applied offline because progress can only be generated from consideration of performance of the entire case. We also evaluated our model by applying average filtering method. Application of this method only increased accuracy by about 1% (Table 5, last row).

We also compared our current system with our previous systems using independent F1-scores of the five phases. Based on the F1-scores in Table 6 (rows 3 and 5), our current system significantly outperformed our two previous systems on Pre-arrival, Primary, Secondary, and Post-secondary (31.5% on average, Table 6, rows 3 and 5) on the same dataset (Table 6, 50/10) because of the use of RLO and PG methods that we introduced. These four phases are more important for detecting human errors during the resuscitation (especially the Primary and Secondary Surveys) compared to the Pt-Departure phase after the patient has left. We also evaluated our current system on our current video set (Table 6, 150/33), which is significantly larger than the video set we used in the past. Based on the evaluation matrices in Table 6, our current system significantly outperformed our previous systems (Li et al. (2017b, 2016)) on both video sets.

5. Experiment results on Endotube and Cholec80

To show the generalizability of our approach, we evaluate our system on the EndoTube dataset and, the Cholc80 dataset for surgical phase recognition (Lea et al. (2016)).

EndoTube: The EndoTube dataset contains 25 videos captured from full cholecystectomy procedures performed at 19 different hospitals in nine countries. The average video length is 11.4 minutes in the range of 4 to 27 minutes. The procedures were manually labeled into seven different phases: trocar placement, preparation, clip/cut, dissection, retrieval, hemostasis, and drainage/finish. We applied 5-fold cross-validation on EndoTube that using 20 videos for training and the remaining five videos for testing, as was done in this previous study (Lea et al. (2016)).

Table 7
Experimental results on the EndoTube dataset. The evaluation results are accuracy, f1-score, precision and recall in percentage.

Method	with tool	Acc.	F1.	Prec.	Rec.
Spatial CNN + tool (Lea et al. (2016))	Yes	63.7	-	-	-
ST-CNN + tool (Lea et al. (2016))	Yes	62.4	-	-	-
Nonlocal	No	70.9	71.3	75.5	70.9
Nonlocal + RLO	No	73.5	73.6	75.9	73.5
Nonlocal + RLO + PG	No	75.1	75.9	76.8	75.1

Table 8
Experimental results on the Cholec80 dataset. The evaluation results are accuracy, f1-score, precision and recall in percentage.

Method	Acc.	F1.	Prec.	Rec.
Phase-LSTM Twinanda et al. (2016a)	79.68	-	72.8	73.45
Endo-LSTM Twinanda (2017)	80.8	-	76.8	72.1
MTRCNet Jin et al. (2020)	82.8	-	76.1	78.0
ResNet-LSTM Jin et al. (2017)	86.6	-	80.5	79.9
TeCNO Czempiel et al. (2020)	88.6	-	81.6	85.2
Nonlocal	87.1	87.0	88.2	87.1
Nonlocal + RLO	90.5	90.4	91.5	90.6
Nonlocal + RLO + PG	**91.2**	**91.0**	**91.6**	**91.1**

Cholec80: The Cholec80 dataset contains 80 videos of cholecystectomy surgeries performed by 13 surgeons (Twinanda et al. (2016a)). The videos were captured at 25 fps and labeled into seven phases: preparation, calot triangle dissection, clipping cutting, gallbladder dissection, gallbladder packaging, cleaning coagulation, and gallbladder retraction. We used 40 videos for training, 8 videos for validation and the remaining 40 videos for testing as was done in previous research (Twinanda et al. (2016a); Czempiel et al. (2020)). These datasets also included surgical tool labels, as additional information for phase recognition. We only used videos as input for our experiments.

Based on the accuracy score in Table 7, and Table 8, our method significantly outperformed previous state-of-the-art approaches on both EndoTube, and Cholec80 datasets (75.1% vs.63.7% on EndoTube, and 91.2 vs. 88.6 on Cholec80), even without using the instrument labels as additional information (Lea et al. (2016); Czempiel et al. (2020)). The previous system separately extracted spatial features from individual frames and then represented the temporal associations from consecutive frames using temporal convolution and LSTM, which poorly represented the motions in the consecutive frames. Our method learns spatio-temporal features using 3D convolution filters and nonlocal blocks. Our proposed RLO extracts long-term spatio-temporal features from the video and also benefits from the pre-trained weights using large-scale activity recognition datasets. (Kay et al. (2017)). We set the downsample rate as 5 ($\alpha = 5$) based on the experiment result. The model increased the accuracy score by around 3% when using RLO to extract long-term video context across the video. The model with PG also had about a 1% accuracy score enhancement compared to the model without PG. The evaluation on using EndoTube data supports that our system generalizes across different processes. These findings also highlight that the proposed RLO and PG methods can improve the model performance on the phase recognition tasks for processes other than trauma resuscitation.

6. Discussion

6.1. Phase recognition consistency

We visualized the phase predictions and their corresponding ground truth in three resuscitation cases (Fig. 7). We compared

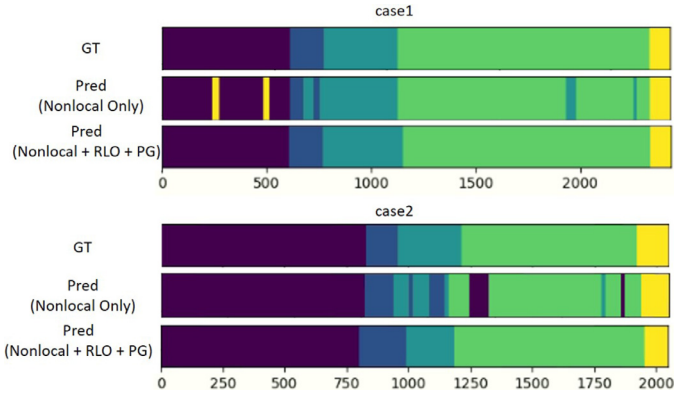

Fig. 7. Phase recognition consistency: we visualized the system predictions and corresponding ground truth in three cases.

the predictions between the system with and without the introduced RLO and PG methods. Based on the visualizations in Fig. 7, the system with RLO and PG can provide more consistent predictions with very few incorrect fragments (Fig. 7, case3). The model having limited incorrected fragments is caused by the use of RLO and PG methods that capture long-term information, and eliminate implausible predictions.

6.2. Temporal modeling

We compared our RLO with other temporal modeling methods by evaluating them on both our Trauma dataset and Cholec80. Based on the evaluation matrices in Table 9, the Nonlocal and TCN-based networks outperformed traditional temporal modeling structures such as CNN-HMM, and CNN-LSTM that are unable to model long distance temporal associations. The HMM, and LSTM-based networks cannot build correlations between long distance frames. The performance did not increase when using LSTM and TCN on top of the Nonlocal network. The spatio-temporal features in short-term inputs (e.g., 32-second in Trauma dataset) have already been well captured by the Nonlocal network. Our proposed RLO method improved the system performance based on the Nonlocal network by including long-term inputs (320-second) while reducing the requirements on memory and computation resources. The TCN-based method reported in (Czempiel et al. (2020)) performed slightly better than our implementation of the TCN-based network (88.6 vs. 87.4) on Cholec80. The difference in performance might be caused by using different training settings.

6.3. Runtime efficiency

We evaluated the runtime efficiency of our system to show that the system is able to provide real-time phase predictions. Table 10 shows the latency of our complete system (including RLO and PG) using both i3D and Nonlocal as backbones and running on multiple processors. Based on the latency in Table 10, even running on a CPU, our model required less than 2 seconds to provide a prediction for a 32-second input (plus 320-seconds history for RLO). The

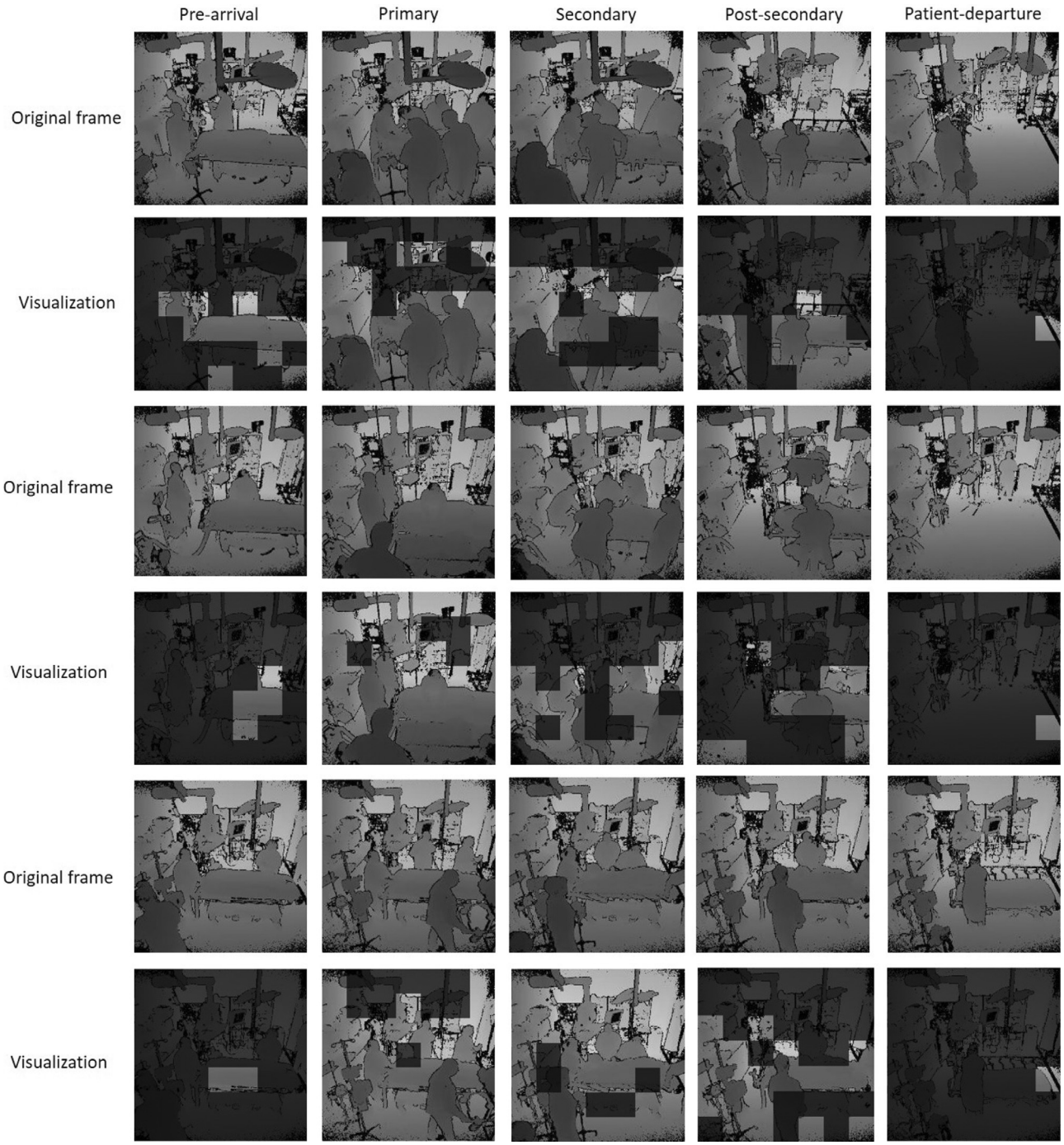

Fig. 8. Feature visualizations for the five phases and their corresponding depth frames. We overlapped the feature maps on the original frames and used the 0.5 threshold for the values for better visualization.

runtime efficiency show in Table 10 demonstrated that our system was able to provide real-time phase predictions because the time that the model uses for providing a prediction is significantly smaller than the sliding window when extracting video clips (16 seconds).

6.4. System transferability

The transferability of any vision-based system is partially dependent on the camera view used. The Kinect in our setting was mounted on the wall, a location that is unobtrusive and easy to maintain. Transfer to another setting that uses different camera views may require tuning. We have obtained a domain-specific dataset that will speed this tuning process when other camera views are used. Equipment may vary in different emergency room settings. Our system relies on environmental features common to other resuscitation settings, including the position of the patient bed and the location of providers performing specific activities. For example, airway activities are performed during the primary survey phase by individuals at the head of the bed. Transferability will require fine tuning the model in other settings that having different background features. Image segmentation models that masking out the unrelated backgrounds may also help to improve the performance for system transferability (He et al. (2017))).

Table 9
Discussion of temporal modules. We compared our method (Nonlocal + RLO) with other temporal modules on both our Trauma Resuscitation dataset and Cholec80.

Method	dataset	Acc.	F1.	Prec.	Rec.
CNN-HMM	Trauma Resuscitation	75.1	74.9	79.2	74.4
CNN-LSTM	Trauma Resuscitation	78.0	79.1	82.4	78.0
Nonlocal	Trauma Resuscitation	83.7	83.1	82.6	84.0
Nonlocal + LSTM	Trauma Resuscitation	82.9	81.1	82.3	83.7
Nonlocal + TCN	Trauma Resuscitation	83.5	83.6	84.2	83.9
Ours (Nonlocal + RLO)	Trauma Resuscitation	**90.8**	**90.6**	**90.6**	**90.8**
EndoNet (CNN-HMM) Twinanda et al. (2016a)	Cholec80	75.2	-	70.0	66.0
ResNet-LSTM (CNN-LSTM) Jin et al. (2017)	Cholec80	86.6	-	80.5	79.9
TCN Czempiel et al. (2020)	Cholec80	88.6	-	81.6	85.2
Nonlocal	Cholec80	87.1	87.0	88.2	87.1
Nonlocal + LSTM	Cholec80	87.1	86.5	86.9	86.8
Nonlocal + TCN	Cholec80	87.4	86.9	87.7	86.8
Ours (Nonlocal + RLO)	Cholec80	**90.5**	**90.4**	**91.5**	**90.6**

Table 10
Runtime efficiency of our system. We evaluated the latency of our system using different backbones and processors.

Method	Input	Processors	Latency
Ours (i3D)	32 + 320 (s)	RTX-2080 ti	0.07 (s)
Ours (Nonlocal)	32 + 320 (s)	RTX-2080 ti	0.14 (s)
Ours (i3D)	32 + 320 (s)	GTX-1080 ti	0.10 (s)
Ours (Nonlocal)	32 + 320 (s)	GTX-1080 ti	0.20 (s)
Ours (i3D)	32 + 320 (s)	i7-6850k (CPU)	1.11 (s)
Ours (Nonlocal)	32 + 320 (s)	i7-6850k (CPU)	1.75 (s)

6.5. Model visualization

To evaluate our hypotheses about the reasons for differences in phase prediction (Section 5.1), we visualized feature maps obtained from the intermediate output of the model and their corresponding depth inputs for different phases (Fig. 8). We overlapped the feature maps on their corresponding depth frames and used a threshold value of 0.5 to generate clearer visualizations. Based on these visualizations (Fig. 8), the feature map has high values for the region of patient bed during the pre-arrival phase (Fig. 8, left) and focuses on the floor during the patient departure phase (Fig. 8, right). The feature map during the post-secondary survey phase (Fig. 8, second last) focused on the patient bed and the few providers around the patient bed. Finally, during the primary and secondary survey, a large area on the feature map (around the patient bed, Fig. 8, second and third) was highlighted reflecting the complexity of the environment in these phases. The model appeared to focus on multiple regions that have features for phase recognition. These visualizations showed that the model focused on regions likely to distinguish different phases and learned representative features for phase recognition.

6.6. Limitation and future work

We have built the system using depth video to ensure that our system is privacy preserving. Our results show that the performance is lower in recognizing the primary and secondary survey phases, but with a relatively high F1-score ($>$ 82%). Human annotators have used RGB videos ground truth coding because unique activities need to be detected to distinguish these two phases. Our next step will be to implement a system that uses enriched texture features from RGB videos preserves privacy-sensitive regions from frames using generating adversarial networks (Goodfellow (2016); Ronneberger et al. (2015); Mirza and Osindero (2014)). RGB/depth video may not be adequate for distinguishing the primary and secondary survey in some cases. For example, secondary survey activities may be performed in parallel with primary survey activities or the primary survey may be interrupted by performance of secondary survey activities before returning to primary survey activities. Although uncommon, these variations will be managed in our future work using a multi-label phase prediction network that provides concurrent phase predictions. Modeling phase-wise correlations will improve multi-label phase prediction in this framework (Sun et al. (2010); Huang et al. (2017)).

Our system segments the trauma resuscitation cases into phases and reduces the challenge of detecting and localizing process errors by setting the focus on a phase of interest. Phase recognition can also be used to improve activity recognition. Because some activities occur uniquely or more (or less) frequently in certain phases, the initial step of phase recognition can provide this needed context. The single camera system may miss some activities because of view occlusion when providers are crowded around the patient bed. Additional cameras may improve this performance even more but at a cost of reducing the transferability of our system. Building a system for recognizing activities using multiple RGB cameras without privacy violation and reducing transferability will be our future work.

7. Conclusion

We introduced a real-time medical phase recognition system during trauma resuscitation. The system is privacy-preserving and achieved more than 90% accuracy score, which outperformed the previous systems using depth videos as input for phase recognition during trauma resuscitation. We also evaluated our system on the EndoTube dataset, outperforming results using a previous system supporting the generalizability of our approach. We introduced novel methods (RLO and PG) for learning spatio-temporal features from long-range video contexts. These methods include estimation of the video progresses to enhance the accuracy of phase prediction. The system's accuracy in distinguishing the primary-survey and secondary-survey phases was affected by the limited texture information in the depth videos. To apply this system within an activity recognition system, we are implementing an RGB-based phase recognition system that manages privacy considerations.

Declaration of Competing Interest

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

CRediT authorship contribution statement

Yanyi Zhang: Conceptualization, Methodology, Formal analysis, Writing – review & editing. **Ivan Marsic:** Supervision, Writing –

review & editing. **Randall S. Burd:** Supervision, Writing – review & editing.

Acknowledgments

We would like to thank the trauma experts from Trauma and Burn Surgery, Children's National Medical Center for their work on data collection and processing. This research has been funded under NIH/NLM grant 2R01LM011834-05 and NSF grants IIS-1763827, IIS-1763355, and IIS-1763509.

References

Ahmadi, S., Padoy, N., Heining, S., Feussner, H., Daumer, M., Navab, N., 2008. Introducing wearable accelerometers in the surgery room for activity detection. Computer-und Roboter-Assistierte Chirurgie (CURAC).
Al Hajj, H., Lamard, M., Conze, P.-H., Cochener, B., Quellec, G., 2018. Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. Med Image Anal 47, 203–218.
Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
Bardram, J.E., Doryab, A., Jensen, R.M., Lange, P.M., Nielsen, K.L., Petersen, S.T., 2011. Phase recognition during surgical procedures using embedded and body-worn sensors. In: 2011 IEEE international conference on pervasive computing and communications (PerCom). IEEE, pp. 45–53.
Bhatia, B., Oates, T., Xiao, Y., Hu, P., 2007. Real-time identification of operating room state from video. In: AAAI, Vol. 2, pp. 1761–1766.
Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308.
Castaneda, C., Nalley, K., Mannion, C., Bhattacharyya, P., Blake, P., Pecora, A., Goy, A., Suh, K.S., 2015. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J Clin Bioinforma 5 (1), 4.
Chen, W., Feng, J., Lu, J., Zhou, J., 2018. Endo3D: online workflow analysis for endoscopic surgeries based on 3d CNN and LSTM. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. Springer, pp. 97–107.
Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N., 2020. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 343–352.
Dai, J., Wu, J., Saghafi, B., Konrad, J., Ishwar, P., 2015. Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 68–76.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee, pp. 248–255.
Feichtenhofer, C., 2020. X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 203–213.
Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. Slowfast networks for video recognition. In: Proceedings of the IEEE international conference on computer vision, pp. 6202–6211.
Feichtenhofer, C., Pinz, A., Zisserman, A., 2016. Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1933–1941.
Goodfellow, I., 2016. Nips 2016 tutorial: generative adversarial networks. arXiv preprint arXiv:1701.00160.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
Huang, J., Li, G., Wang, S., Xue, Z., Huang, Q., 2017. Multi-label classification by exploiting local positive and negative pairwise label correlation. Neurocomputing 257, 164–174.
Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.
Jia, P., Zhang, L., Chen, J., Zhao, P., Zhang, M., 2016. The effects of clinical decision support systems on medication safety: an overview. PLoS ONE 11 (12), e0167683.
Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.-W., Heng, P.-A., 2017. Sv-rcnet: workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans Med Imaging 37 (5), 1114–1126.
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.
Kaplan, L.J., 2002. Trauma resuscitation. In: Gullo, A. (Ed.), Anaesthesia, Pain, Intensive Care and Emergency Medicine — A.P.I.C.E.. Springer Milan, Milano, pp. 107–123.
Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al., 2017. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.
Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Kortbeek, J.B., Al Turki, S.A., Ali, J., Antoine, J.A., Bouillon, B., Brasel, K., Brenneman, F., Brink, P.R., Brohi, K., Burris, D., et al., 2008. Advanced trauma life support, the evidence for change. Journal of Trauma and Acute Care Surgery 64 (6), 1638–1650.
Lea, C., Choi, J.H., Reiter, A., Hager, G., 2016. Surgical phase recognition: from instrumented ORs to hospitals around the world. In: Medical image computing and computer-assisted intervention M2CAIMICCAI workshop, pp. 45–54.
Li, X., Zhang, Y., Li, M., Chen, S., Austin, F.R., Marsic, I., Burd, R.S., 2016. Online process phase detection using multimodal deep learning. In: 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, pp. 1–7.
Li, X., Zhang, Y., Zhang, J., Chen, S., Marsic, I., Farneth, R.A., Burd, R.S., 2017. Concurrent activity recognition with multimodal cnn-lstm structure. arXiv preprint arXiv:1702.01638.
Li, X., Zhang, Y., Zhang, J., Zhou, M., Chen, S., Gu, Y., Chen, Y., Marsic, I., Farneth, R.A., Burd, R.S., 2017. Progress estimation and phase detection for sequential processes. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 1 (3), 1–20.
Loukas, C., 2018. Surgical phase recognition of short video shots based on temporal modeling of deep features. arXiv preprint arXiv:1807.07853.
Meißner, C., Meixensberger, J., Pretschner, A., Neumuth, T., 2014. Sensor-based surgical activity recognition in unconstrained environments. Minimally Invasive Therapy & Allied Technologies 23 (4), 198–205.
Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.
Mutegeki, R., Han, D.S., 2020. A CNN-LSTM approach to human activity recognition. In: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, pp. 362–366.
Reis, W.C., Bonetti, A.F., Bottacin, W.E., Reis Jr, A.S., Souza, T.T., Pontarolo, R., Correr, C.J., Fernandez-Llimos, F., 2017. Impact on process results of clinical decision support systems (cdsss) applied to medication use: overview of systematic reviews. Pharmacy Practice (Granada) 15 (4).
Reiter, A., Ma, A., Rawat, N., Shrock, C., Saria, S., 2016. Process monitoring in the intensive care unit: Assessing patient mobility through activity analysis with a non-invasive mobility sensor. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 482–490.
Ren, Z., Jae Lee, Y., Ryoo, M.S., 2018. Learning to anonymize faces for privacy preserving action detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 620–636.
Rodziewicz, T.L., Hipskind, J.E., 2020. Medical error prevention. StatPearls [Internet]. StatPearls Publishing.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.
Ryoo, M.S., Kim, K., Yang, H.J., 2018. Extreme low resolution activity recognition with multi-siamese embedding learning. In: Thirty-Second AAAI Conference on Artificial Intelligence.
Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., Zhang, C., 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In: Thirty-Second AAAI Conference on Artificial Intelligence.
Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp. 568–576.
Srivastav, V., Gangi, A., Padoy, N., 2019. Human pose estimation on privacy-preserving low-resolution depth images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 583–591.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15 (1), 1929–1958.
Sun, L., Ji, S., Ye, J., 2010. Canonical correlation analysis for multilabel classification: a least-squares formulation, extensions, and analysis. IEEE Trans Pattern Anal Mach Intell 33 (1), 194–200.
Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp. 4489–4497.
Tran, D., Wang, H., Torresani, L., Feiszli, M., 2019. Video classification with channel-separated convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5552–5561.
Twinanda, A.P., 2017. Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos. Strasbourg.
Twinanda, A.P., Alkan, E.O., Gangi, A., de Mathelin, M., Padoy, N., 2015. Data–driven spatio-temporal rgbd feature encoding for action recognition in operating rooms. Int J Comput Assist Radiol Surg 10 (6), 737–747.
Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N., 2016. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging 36 (1), 86–97.
Twinanda, A.P., Winata, P., Gangi, A., Mathelin, M., Padoy, N., 2016. Multi-stream deep architecture for surgical phase recognition on multi-view rgbd videos. In: Proc. M2CAI Workshop MICCAI, pp. 1–8.
Twinanda, A.P., Yengera, G., Mutter, D., Marescaux, J., Padoy, N., 2018. Rsdnet: learn-

ing to predict remaining surgery duration from laparoscopic videos without manual annotations. IEEE Trans Med Imaging 38 (4), 1069–1078.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008.
Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L., 2016. Temporal segment networks: towards good practices for deep action recognition. In: European Conference on Computer Vision. Springer, pp. 20–36.
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.
Wolf, Z.R., Hughes, R.G., 2008. Error Reporting and Disclosure. Patient Safety and Quality: An Evidence-based Handbook for Nurses. Agency for Healthcare Research and Quality (US).
Wu, C.-Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R., 2019. Long-term feature banks for detailed video understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 284–293.
Yengera, G., Mutter, D., Marescaux, J., Padoy, N., 2018. Less is more: surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks. arXiv preprint arXiv:1805.08569.
Yeung, S., Alahi, A., Haque, A., Peng, B., Luo, Z., Singh, A., Platchek, T., Milstein, A., Li, F.-F., 2016. Vision-based hand hygiene monitoring in hospitals.. AMIA.
Yeung, S., Rinaldo, F., Jopling, J., Liu, B., Mehra, R., Downing, N.L., Guo, M., Bianconi, G.M., Alahi, A., Lee, J., et al., 2019. A computer vision system for deep learning-based detection of patient mobilization activities in the icu. NPJ digital medicine 2 (1), 1–5.
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. Springer, pp. 265–272.