Intraoperative Adverse Event Detection in Laparoscopic Surgery: Stabilized Multi-Stage Temporal Convolutional Network with Focal-Uncertainty Loss

Haiqi Wei CELINE.WEI@MAIL.UTORONTO.CA
University of Toronto, International Centre for Surgical Safety, Surgical Safety Technologies

Frank Rudzicz FRANK@SPOCLAB.COM
University of Toronto, Vector Institute, International Centre for Surgical Safety, Surgical Safety Technologies

David Fleet FLEET@CS.TORONTO.EDU
University of Toronto, Vector Institute

Teodor Grantcharov T.GRANTCHAROV@SURGICALSAFETY.COM
University of Toronto, International Centre for Surgical Safety, Surgical Safety Technologies, University Health Network

Babak Taati TAATI@CS.TORONTO.EDU
University of Toronto, Toronto Rehabilitation Institute, University Health Network

Abstract

Intraoperative adverse events (iAEs) increase rates of postoperative mortality and morbidity. Identifying iAEs is important to quality assurance and postoperative care, but requires expertise, is time consuming, and expensive. Automated or partially-automated techniques are, therefore, desirable. Previous work showed that conventional image processing has not worked well with real-world laparoscopic videos. We present a novel modular deep learning system that can partially automate the process of iAE screening using videos of laparoscopic procedures. The system consists of a stabilizer to reduce camera motion, a spatiotemporal feature extractor, and a multi-stage temporal convolutional neural network to detect adverse events. We apply a novel focal-uncertainty smoothing loss to handle class imbalance and to address multi-task uncertainty. The system is evaluated using 5-fold cross-validation on a large (228 hours) dataset of laparoscopic videos, and we perform ablation studies to investigate the effects of stabilization and focal-uncertainty loss. Our system achieves an AUROC of 0.952, an average precision (AP) of 0.626 in thermal injury detection, and an AUROC of 0.823 and an AP of 0.336 in bleeding detection. Our novel modular deep learning system outperforms conventional deep learning baselines. The model can be used as a screening tool to search for high risk events and to provide feedback for operation quality improvements and postoperative care. Source code available on GitHub: https://github.com/ICSSresearch/IAE-video.

1. Introduction

Intraoperative adverse events (iAEs) lead to substantial increases in mortality, morbidity, and length of postoperative stay (Bohnen et al., 2017). Approximately 70% of iAEs involve medical errors (Gawande et al., 2003), more than half of which are preventable (Gawande

et al., 1999; Thomas et al., 1999). Indeed, iAEs caused by medical errors contribute to the deaths of between 98,000 and 400,000 people each year (Makary and Daniel, 2016). With 15 million laparoscopic procedures annually (BIS Research, 2018), millions are at risk. Currently, surgeons must identify iAEs and report them *post hoc* in patient records, relying on the imperfect memory of the staff. We advocate for recording laparoscopic procedures and for automatically analyzing them to address these problems.

Intraoperative adverse events account for 48% of adverse events (Leape et al., 1991), 23% of which involve bleeding (Zegers et al., 2011). In our analysis, thermal injury caused by cautery accounts for a high proportion of remaining causes. Our aim is, therefore, to detect both bleeding and thermal injuries. We formulate iAE detection as a temporal action localization (TAL) problem and we identify the start- and end-time of all iAEs.

Previous studies (Jo et al., 2016; Garcia-Martinez et al., 2017; Okamoto et al., 2019) used handcrafted features, such as colour descriptors, to detect bleeding events in laparoscopic videos. Deep learning techniques have not been used for iAE detection in laparoscopic videos, but a deep learning model – namely an InceptionV3 (Szegedy et al., 2016) convolutional neural network (CNN) – has successfully been used for lesion detection in wireless capsule endoscopy (WCE), outperforming traditional handcrafted feature methods (Jia and Meng, 2016; Li et al., 2017). Both of these approaches generate frame-by-frame predictions without considering temporal dependencies. Incorporating temporal dependencies is expected to generate smoother outputs and also improve prediction accuracy. For example, a bleeding event may leave blood residue on tissues that frame-by-frame analysis will have difficulty differentiating from active bleeding.

Architectures that incorporate temporal dependencies in their decision making – e.g., recurrent neural networks (RNNs), temporal convolutional networks (TCNs), or 3D convolutional neural networks (3DCNNs) – have produced state-of-the-art results in TAL tasks in other domains but, as of yet, have not been used or adapted for iAE detection in laparoscopic videos. It was, however, unclear how such techniques will perform on long unstable intraoperative videos, as they have primarily been used for TAL in significantly shorter and more stable videos. For instance, in the THUMOS'15 dataset, which is one of the largest datasets in TAL, the average action length is 4.6 seconds, and actions on average take ∼20% of video lengths (Idrees et al., 2017). By contrast, in laparoscopic videos, each iAE could last up to a few minutes in a procedure, and the procedures are on average 1.5 hours long, so iAEs on average occupy less than 3.5% of a procedure. This, in turn, results in a machine learning training set with a significant label imbalance. Similarly, many of the action samples in most TAL datasets are filmed on a fixed camera, for instance, in the Breakfast dataset (Kuehne et al., 2014), or sports broadcast videos in the THUMOS'15 dataset that are filmed on professional stabilizing equipment. By contrast, laparoscopic videos are filmed on hand-held laparoscopes and are highly unstable. The laparoscopic camera movements, in most cases, are larger than the expansion of blood or burn mark on tissues.

In this work, to move beyond the aforementioned limitations, we propose to stabilize the long unstable videos and to use a novel focal-uncertainty loss to address the highly unbalanced class labels. We use a 3DCNN model (I3D (Carreira and Zisserman, 2017)) as a feature extractor and propose a method to learn a fine-grained feature representation of iAE classes by learning spatiotemporal features while preserving and sharing 2D features of more granular sub-classes. We feed the feature vector of a full procedure into a multi-stage TCN

model (MS-TCN) (Farha and Gall, 2019) for temporal segmentation. We conduct ablation studies to show the significance of stabilization and focal-uncertainty loss. Finally, we compare our system to multiple baselines: InceptionV3, I3D, and MS-TCN. Our proposed system outperforms these baselines, and establishes the state-of-the-art for iAE detection. To date, we are the first to detect thermal injury events, to use stabilization in feature learning, and we increase MS-TCN performance using a focal-uncertainty loss.

Generalizable Insights about Machine Learning in the Context of Healthcare

Detecting iAEs in laparoscopic procedures is an action recognition and localization problem with unique characteristics. While action recognition – and particularly human activity recognition – is a widely studied problem, iAE detection requires the localization of short event instances in very long videos filmed on unstable cameras, making it substantially different from typical activity recognition benchmarks. Our work addresses this hereto unstudied problem by:

- integrating video stabilization with 3DCNN feature extraction and multi-stage temporal convolutional neural network to achieve best performance;
- investigating the effect of stabilization and various loss function terms (multi-task loss, focal loss, an uncertainty loss, smoothing loss) to gain insight about models and conditions under which state-of-the-art performance can be achieved on highly unbalanced data and given the unique characteristics of the problem and its differences with standard benchmarks; and
- analyzing the readability and comprehensibility of the visual detection results for clinical use.

2. Related Work

iAE detection in laparoscopic surgery. Most studies in iAE detection in laparoscopic videos used statistical parameters of colour descriptors (Jo et al., 2016; Garcia-Martinez et al., 2017; Okamoto et al., 2019). These methods are not adaptive to realistic variability, such as changes of light source, and they depend on various hyper-parameters that can be difficult to determine. While existing literature in iAE detection in laparoscopic surgery is sparse, more studies have been conducted in a closely related area, i.e., lesion detection in wireless capsule endoscopy.

Wireless capsule endoscopy lesion detection. Among various TAL tasks in different clinical imaging modalities, bleeding detection in WCE is the closest to iAE detection in laparoscopic videos. While earlier methods in this domain used handcrafted features (Fu et al., 2014; Ghosh et al., 2014; Usman et al., 2016), two new studies recently explored using CNNs and achieved significantly improved performance (Jia and Meng, 2016; Li et al., 2017). The success of deep learning in WCE motivates applying CNNs to iAE detection in laparoscopic videos. There are, however, some important differences between WCE and laparoscopic videos. For example, unlike WCE, it is normal to see blood or blood residue during a laparoscopic surgery. Therefore, while bleeding detection in WCE can be translated to a simple blood detection, the task is significantly more complicated in laparoscopic videos

and involves differentiating between active bleeding vs. the presence of blood residue. This necessitates the use of temporal models that can detect the expansion and the flow of a blood pool. The same concept applies in thermal injury detection and distinguishing between the presence of a burn mark vs. active injury.

Temporal action localization. iAE detection in laparoscopic video is analogous to TAL, i.e. identifying the start- and end-time of an event in an untrimmed video. One approach for TAL is to generate and classify action proposals, and regress the boundaries (Xu et al., 2019; Zhao et al., 2020). The robustness of this method depend on the selected initial proposal size, the number of anchors, and the anchor size. Another approach is segment-level classification such as via and RNN, 3DCNN, or a TCN (Montes et al., 2016; Lea et al., 2017; Farha and Gall, 2019; Carreira and Zisserman, 2017). This approach learns frame-level or segment-level features and feed them to an action classifier. Our work is inspired by the success of MS-TCN (Farha and Gall, 2019) in using full temporal features of a complete video. In contrast to the short actions filmed on a fixed camera that the original MS-TCN used (the Breakfast dataset (Kuehne et al., 2014)), our videos are significantly longer and filmed by a moving laparoscope.

3. Cohort

We collected 130 real laparoscopic procedure videos, a total of 228 hours of video data in 1280 x 720 resolution at 30 frames per second (fps), from the years 2016 to 2018. This is the largest dataset of its type. It consists of 24,637,433 annotated frames, and 12,529 unique bleeding and thermal events. Each event is labelled with start- and end-time. The data are collected in the operating room of St. Michael's Hospital in Toronto Canada, whose Research Ethics Board approved this study (REB# 16-243). We use the OR Black Box® (Surgical Safety Technologies™, Toronto, Canada) to collect the data. The data are collected in the operating room of a hospital and the Institutional Review Board approved the study. Identifiable information such as patient information and audio was excluded during data storage and processing. The data is reviewed and annotated by three trained surgeons frame-by-frame with labels blood, bleeding, burn, and thermal injury.

Bleeding leaves residual blood, stained tissues, and blood clots in laparoscopic procedures. These can resemble an active bleeding event if the detector only detects blood pixels. Similarly, an active thermal injury event can resemble a burn scar if temporal dependencies are not considered. As such, we consider four distinct types of event, namely,

- **blood**: when blood is visible, including bleeding, blood residue, and blood clots,
- **burn**: when tissue discolouration (e.g., a burn mark) is visible,
- **bleeding**: when blood is actively flowing, and
- **thermal injury**: when a thermal device is actively damaging tissues.

Labels can appear simultaneously in the same frame. The bleeding and the thermal injury labels subsume the blood and the burn labels respectively. That is, if bleeding is present, blood is also present, but not vice versa. A frame without labels indicates no adverse

event. Assigning these labels to video frames is modelled as a multi-task classification problem. Fig. 1A shows examples of these events.

Table 1: Distribution of events, frames, and windows in each task: a highly unbalanced dataset. The number of windows indicates the number of 10-second (50 frames) sliding windows.

Task	# Frames	% Frames	# Events	% Events
Blood	13,940,462	51.8%	N/A	N/A
Bleeding	203,809	0.8%	1,450	11.6%
Burn	1,723,831	6.4%	N/A	N/A
Thermal Injury	938,583	3.5%	11,079	88.4%
Background	10,118,163	37.6%	N/A	N/A

Table 1 shows the distribution of each label. Blood labels dominate the distribution, accounting for 51.8% of all frames, whereas bleeding, burn, and thermal injuries combined for less than 11%. Among individual active events, thermal injury events are the most common (88.4%).

Rather than train separate classifiers for each event type, we train one system for all event types, thereby learning fine-grained shared properties and distinctive features of each event. For example, an active flow of bright red blood provides an indication of an active bleeding event, and a spread of yellow-white burn marks is a visual indication of an active thermal injury event. The model can also explicitly learn dependencies between events.

4. Methods

Our proposed iAE detection system is a 4-stage stabilized temporal convolutional network ("4-stage Stab-TCN"). This system is evaluated on the largest laparoscopic iAE dataset, which has 130 laparoscopic procedures.The system identifies bleeding and thermal injury event predictions in laparoscopic videos based on extracted features on stabilized frames.

We use a bundled camera path estimation algorithm (Liu et al., 2013) to stabilize camera motion in laparoscopic views. We extract 1024 latent features using I3D, a 3D CNN, on a sliding window of 50 frames. As a result, the input size to the next module (i.e., MS-TCN) is $N \times 1024$, where N is the number of sliding windows in each video. The MS-TCN module determines if an event happens in a given window, and produces N predictions for each video.

In what follows we provide details about the model architectures, loss functions, experimental setup, and evaluation metrics.

4.1. The 4-stage Stab-TCN

Our 4-stage Stab-TCN consists of three modules: a stabilizer, a spatiotemporal feature extractor, and a multi-task event classifier. Fig. 1B shows the workflow of these modules, and Fig. 1C shows a typical case with the ground truth provided by human experts, compared with our model's predictions. The vertical axis represents a detection of a bleeding

(in blue), and/or a thermal injury event (in orange), and the length of the bars are different only to showcase event overlap. Additional event plots of different cases are in Appendix A.

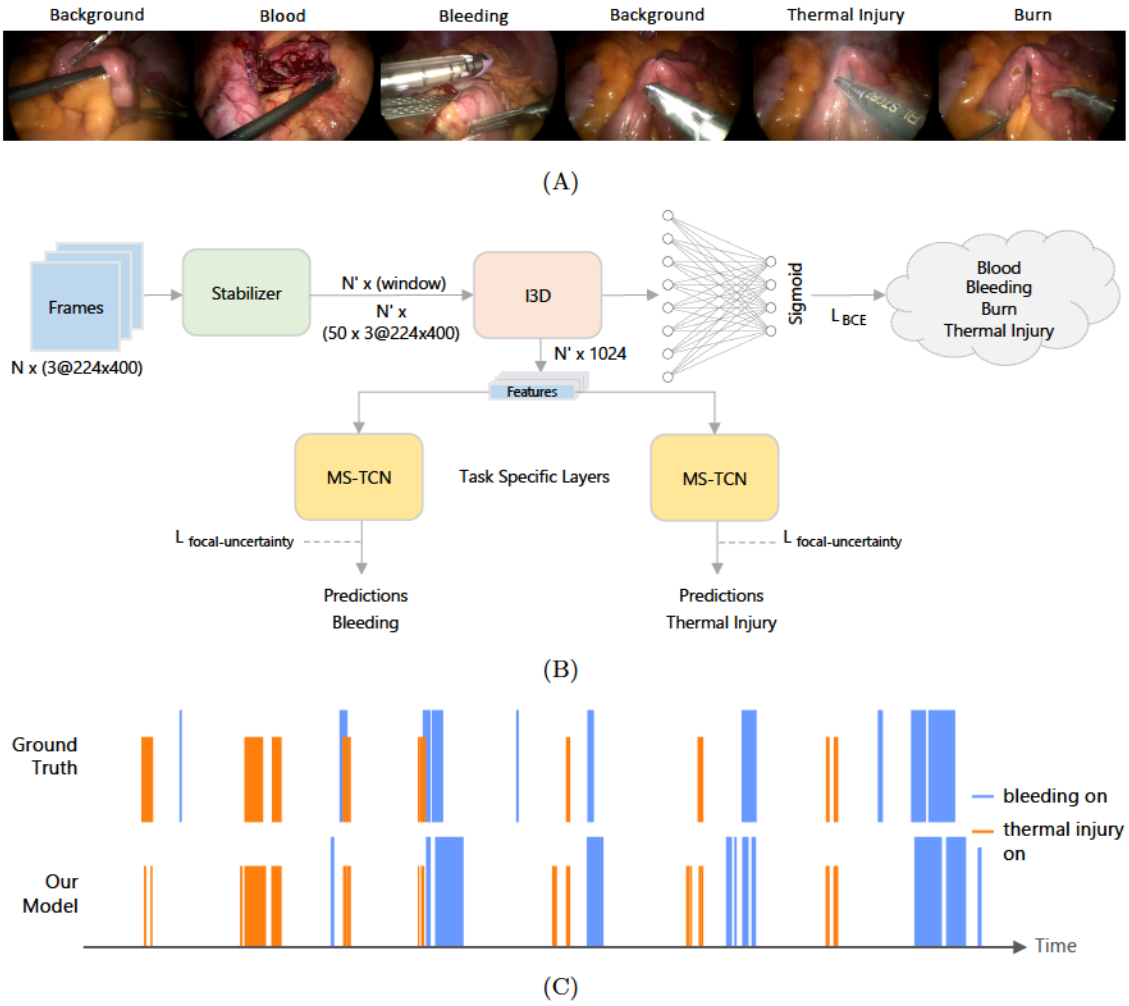


Figure 1: System overview: Our system takes as input a video containing frames such as those shown in A that share strong similarities among events. It consists of three modules shown in B, and detects iAEs accurately and precisely as shown in C, in which each of the vertical bars represents a detection of bleeding and/or thermal injury.

In the first module, we stabilize a window of consecutive frames to reduce camera motion. The camera in laparoscopy is controlled by an operating surgeon. The stabilizer, based on bundled-camera path stabilization (Liu et al., 2013), reduces jitter, and smooths camera paths so that features can be accumulated over several frames. The algorithm warps images to align each frame's camera view via a homography, and it does not require training a

model. The stabilizer is applied to every 10-second window at 5 fps with a stride of 1 second.

The second module, a 3D convolutional neural network based on InceptionV1 (I3D) (Carreira and Zisserman, 2017), extracts a vector of 1024 latent features from each sliding window. Each sliding window is a 10-second stabilized clip (i.e., 50 stabilized frames). The spatiotemporal feature vector is therefore a representation of its corresponding frames. We minimize a binary cross-entropy loss function to train the multi-task classifier.

The final module is a multi-stage temporal convolutional network (MS-TCN) (Farha and Gall, 2019) that takes the extracted features of a whole video as input. As bleeding and thermal injuries are not causally related, we use the MS-TCN as a task-specific layer and evaluate bleeding, and thermal injury detection separately.

4.2. Stabilization

The bundled camera path algorithm divides a frame into several meshes. At each mesh, the algorithm estimates localized camera motion by computing local homography for each mesh. This homography is constrained by a shape-preserving term so that the final warp of the image is not strongly distorted (Liu et al., 2013). The amount of shape regularization is controlled by a factor α. In the original implementation, α is an adaptive threshold that is determined empirically by computing the fitting error using a range of α from 0.3 to 3. This process is too computationally costly in our context. Instead, we use $\alpha = 3$ to ensure smooth camera paths between cells because occlusion and camera depth variation happen very often in the procedure. We use ORB features (Rublee et al., 2011) for model estimation instead of SURF features.

Once camera paths are estimated, we optimize the camera paths so that the camera motion is smooth. We use a discontinuity-preserving term ('smoothing term') to preserve motion discontinuity (G_m). This prevents cropping for a quickly panning scene, and is originally computed as the sum of transition in camera position. After a few experiments, we find that computing the Gaussian distribution ($\sigma^2 = 800$) of the sum produces smoother stabilized frames, and more consistent frames without the need for large cropping. This discontinuity-preserving term is controlled by an adaptive parameter λ. Cropping and distortion ratios, as a measure of the stabilization quality, are computed at each λ until the ratios are both below the thresholds. Instead of using this adaptive method, we empirically select $\lambda = 3$. This speeds up our stabilization process by a factor of 3.

We divide each frame into an 8 x 8 mesh grid for stabilization. During camera optimization, we consider all 50 frames used in feature extraction in the smoothing term.

4.3. Loss Function

We use a combination of a multi-task loss, a focal loss term (Lin et al., 2017), an uncertainty loss term (Cipolla et al., 2018), and a smoothing loss (Farha and Gall, 2019). The multi-task loss is a sigmoid binary cross entropy, and the smoothing loss is a truncated mean-square error (T-MSE) between log-probabilities of the current frame and the previous frame. The smoothing loss is only used in TCN training where the full video is used as an input.

For the multi-task loss, we use a binary cross entropy loss

$$\mathcal{L}_{BCE_{cn}} = y_{t_{cn}} \log(y_{p_{cn}}) \; + \; (1 - y_{t_{cn}}) \log(1 - y_{p_{cn}}), \tag{1}$$

$$\mathcal{L}_{BCE} = \frac{1}{CN}\sum_{c}\sum_{n}\mathcal{L}_{BCE_{cn}}, \tag{2}$$

where $y_{p_{cn}}$ is prediction probability and $y_{t_{cn}}$ is the true label of label c and sample n (1 for positive, and 0 for negative).

The smoothing loss is in the form of its original implementation (Farha and Gall, 2019)

$$\mathcal{L}_{T-MSE} = \frac{1}{CN}\sum_{c}\sum_{n}\max(\tau,\ |\log y_{p_{cn}} - \log y_{p_{cn-1}}|), \tag{3}$$

where τ is a hyper-parameter. The normal loss in the experiment is a weighted sum of the two losses

$$\mathcal{L}_{normal} = \mathcal{L}_{BCE}\ +\ \lambda\ \mathcal{L}_{T-MSE}, \tag{4}$$

where λ is a hyper-parameter. Focal loss scales $\mathcal{L}_{BCE}$ for each sample so that the model will focus on learning harder examples.

$$\mathcal{L}_{f} = \frac{1}{CN}\sum_{c}\sum_{n}(1-p_{cn})^{\gamma}\mathcal{L}_{BCE_{cn}} \tag{5}$$

where

$$p_{cn} = \begin{cases} y_{p_{cn}} & \text{when } y_{t_{cn}} = 1 \\ 1 - y_{p_{cn}} & \text{when } y_{t_{cn}} = 0 \end{cases} \tag{6}$$

is the prediction confidence, and $\mathcal{L}_{BCE_{cn}}$ is binary cross-entropy loss of label c at sample n.

Task-dependent uncertainty depends on each task's representation or measurement units. In adverse event detection, each event has different representation and measurement scales. In order to model this task-dependent uncertainty that captures uncertainty embedded in each task, we used the multi-task likelihoods, and scale it by a factor $\frac{1}{\sigma^2}$, where σ can be interpreted as the temperature term in a Gibbs distribution (Cipolla et al., 2018). The derivation of such model is done on classification and regression tasks (Cipolla et al., 2018). For our problem, we derive the multi-task loss function based on negative log-likelihood of sigmoid activation $\mathcal{L} = -\log\left(Sigmoid(f^w(x)\right)$. For a single task, we have

$$-\log(p(y|f^w(x),\sigma)) \tag{7}$$

$$= -\log\left(Sigmoid(\frac{1}{\sigma^2}f^w(x))\right) \tag{8}$$

$$= \log\left(1+\exp\left(-\frac{1}{\sigma^2}f^2(x)\right)\right) \tag{9}$$

$$= \frac{1}{\sigma^2}\mathcal{L} + \log\left(\frac{1+\exp\left(-\frac{1}{\sigma^2}f^w(x)\right)}{\left(1+\exp\left(-f^w(x)\right)\right)^{\frac{1}{\sigma^2}}}\right) \tag{10}$$

$$\approx \frac{1}{\sigma^2}\mathcal{L} + \log\left(\sigma\right) \tag{11}$$

In (10), we let $\left(1+\exp(-f^w(x)\right)^{\frac{1}{\sigma^2}} = \frac{1}{\sigma^2}(1+\exp(-f^w(x)))$. To generalize to multi-task tasks, each label c has a scalar σ_c, and is embedded in the loss function

$$\mathcal{L}_{u} = \left(\frac{1}{CN}\sum_{c}\sum_{n}\frac{1}{\sigma_c^2}\mathcal{L}_{BCE_{cn}} + \log\sigma_c\right). \tag{12}$$

Focal and uncertainty loss functions are used with the smoothing loss. We take a weighted sum of the loss, over all samples and labels, plus the smoothing loss

$$\mathcal{L}_{focal} = \mathcal{L}_f + \lambda \, \mathcal{L}_{T-MSE} \tag{13}$$

$$\mathcal{L}_{uncertainty} = \mathcal{L}_u + \lambda \, \mathcal{L}_{T-MSE} \tag{14}$$

$$\mathcal{L}_{focal-uncertainty} = \mathcal{L}_f + \mathcal{L}_u + \lambda \, \mathcal{L}_{T-MSE}, \tag{15}$$

where λ is a constant. In all experiments, we use $\gamma = 2$, $\lambda = 0.15$, and $\tau = 16$.

5. Results on Real Data

5.1. Experimental Setup

We train all models with stochastic gradient descent with an initial learning rate of 0.001, after which the learning rate decays by a factor of 0.95 after each epoch. We use 5-fold cross-validation to select the best epoch and threshold, which are then used for testing. All models use a sigmoid function as their final activation function as opposed to softmax function since the classes are not mutually exclusive. We resize frames from 1280 x 720 to 256 x 256, and apply data augmentation such as random flip and random contrast adjustment for all training.

We pre-train the I3D network on ImageNet (Russakovsky et al., 2015) and Kinetics (Kay et al., 2017) with a mini-batch size of 8 windows. Each sample windows of duration 10 seconds at 5 fps, with a stride of 1 frame. To balance the training data, we randomly sample windows by combined labels. The data has 12 possible combinations of labels, or 'classes'. For example, class *[1 1 1 0]* has positive blood, positive bleeding, positive burn, and no thermal injury. Impossible combinations are discarded (e.g., a frame with no blood, but positive bleeding is not possible). We over-sample classes with fewer frames using random over-sampling, randomly selecting samples for augmentation, and under-sampling remaining classes so all classes have the same number of frames (150,000 frames). We show the data distribution by class in Appendix B. We take the label of the last frame as the label for the window.

SS-TCN and MS-TCN are trained with a batch size of 1 with no pre-trained weights. Sampling does not balance the dataset because a training sample is a full procedure with a varying numbers of events. We add focal and uncertainty loss terms to address the effects of dta imbalance, improving model performance. SS-TCN has one layer of TCN with 10 layers of dilated residuals, whereas MS-TCN has four layers of TCN with 10 layers of dilated residuals. Each layer has 64 filters with a filter width of 3.

As no previous work used deep learning in detecting iAEs in laparoscopic videos, we evaluate a baseline model – InceptionV3 (Szegedy et al., 2016), a popular 2D convolutional neural network for image classification. We fine-tune InceptionV3 with pre-trained weights on the ImageNet dataset (Russakovsky et al., 2015). Although images in ImageNet are very different from laparoscopic images, the large data volume and diversity of ImageNet helps the model learn general features. We use a mini-batch of 64 samples, randomly sampling classes from the data so different classes are balanced in each mini-batch. Training converges at 100 epochs.

We use an ablation study on different combinations of each term of the loss functions to study their influence on model performance. The different losses are 1) the normal loss $\mathcal{L}_{normal}$, 2) the combined focal and T-MSE loss $\mathcal{L}_{focal}$, 3) the combined uncertainty and T-MSE loss $\mathcal{L}_{uncertainty}$, 4) the combined focal, uncertainty and T-MSE loss $\mathcal{L}_{focal-uncertainty}$, and 5) the combined focal and uncertainty loss $\mathcal{L}_{focal-uncertainty-notmse}$.

We ran another ablation study on I3D and MS-TCN with and without stabilization to study the effect of stabilization. The threshold is chosen at maximum segmental recall.

Finally, we analyze all model performance including InceptionV3, I3D, a single-stage TCN with stabilization (1-stage Stab-TCN), and a 4-stage Stab-TCN with the four labels (i.e., 4-label 4-stage Stab-TCN) to study the combined effect of loss function, stabilization, and other parameters. The 4-stage Stab-TCN, and the 1-stage Stab-TCN are trained on bleeding and thermal injury separately.

5.2. Evaluation Metrics

Frame-wise metrics consider correctly detected events in each frame as true positives. Event-wise metrics count individual (multi-frame) events as true positives if the intersection over union (IOU) of the true and detected frames is larger than a selected threshold.

Frame-wise metrics are commonly used to evaluate detector performance. In our analysis, we use area under ROC curves (AUROC), average precision (AP), precision, and recall. However, these metrics do not represent model performance when a majority class dominates the number of frames. For example, when a model detects a very long event correctly and misses many events of short duration, its frame-wise performance could be high, even though the event-wise performance may be weak.

Therefore, we adapt event-wise metrics—segmental AP at IOUs of k =0.1, and 0.25 (AP@k), segmental precision, and segmental recall (Richard and Gall, 2016). If there are multiple correct detections for a true event, only the first is marked as a true positive. This way this metric encourages a continuous true positive segment, and penalizes over-segmentation error by marking the rest of the multiple detections as false positives (Richard and Gall, 2016), even if its IOU is above threshold.

The SS-TCN study finds segmental F1 more robust than segmental AP (Lea et al., 2017) in solving the multi-class action recognition problem. On the contrary, our result shows it is more robust to use segmental AP in the multi-task problem.

Additionally, to show the significance of our comparison, we plotted metrics in bar graphs with confidence intervals (CIs) marked in red lines with the measurements above each bar. Model name with its loss function name is used in the legend, where *focal-uncertainty-notmse* is a loss combining focal and uncertainty loss without truncated mean-square error, and *focal-uncertainty* is the combined focal and uncertainty loss with truncated mean-square error. We used the Holm-Bonferroni method to test all comparisons. The blue numbers on the bars indicates the significance group the proportion is in. When the results show significance between multiple proportions, we ranked them starting from 1. If they are in the same group, they will have the same rank. If a comparison shows no significance to any other group, it will be grouped into group 0.

5.3. Results

The 4-stage Stab-TCN significantly outperforms other models. When compared to InceptionV3, our system increases AUROC from 0.740 to 0.803 in bleeding detection, and from 0.829 to 0.930 in thermal injury detection; our system increases the average precision (AP) from 0.237 to 0.356, and 0.375 to 0.560 in bleeding and thermal injury detection, respectively. We discuss use cases such as using 4-stage Stab-TCN as a screening tool in this section. We provide demo videos of stabilization, and detection in the supporting materials (see Appendix C for more details).

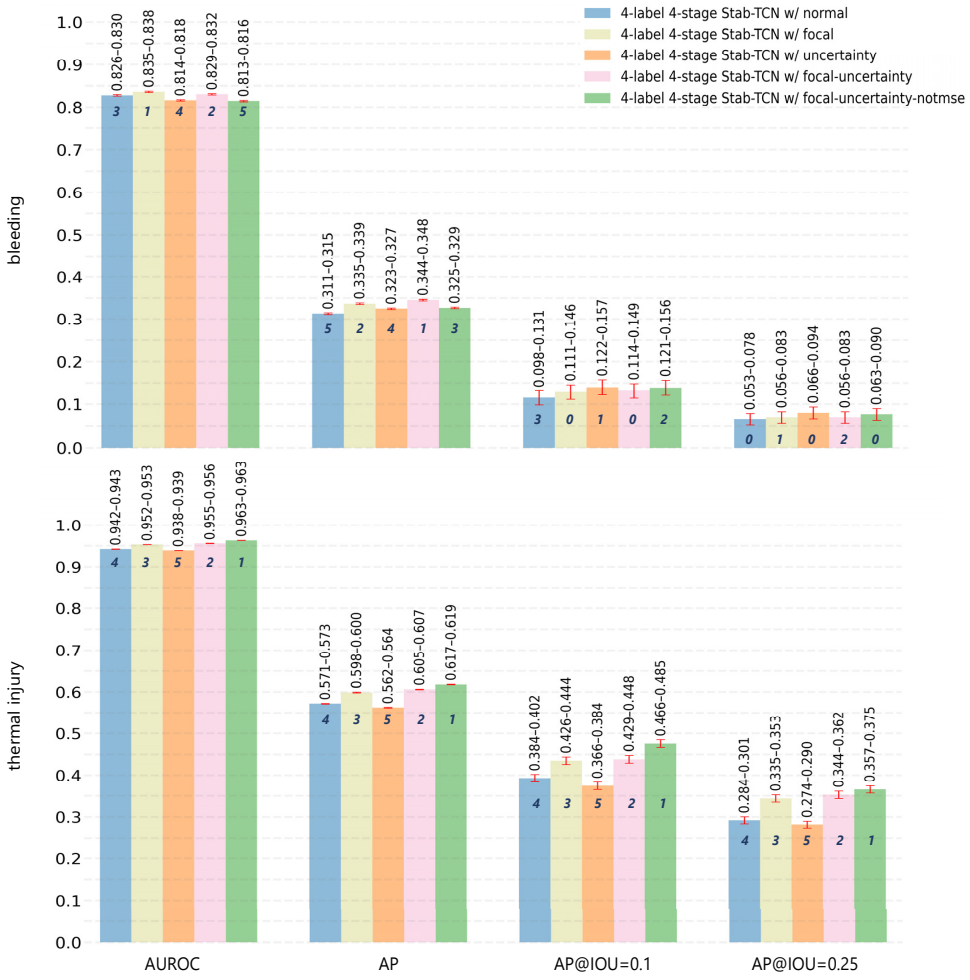


Figure 2: Comparison of different loss function effects on Stab-TCN: Loss functions *focal-uncertainty-notmse* and *focal-uncertainty* improves detection of shorter events. The vertical axis shows the frame-wise AUROC and AP, and segmental AP at IOU of 0.1, and 0.25 of each model

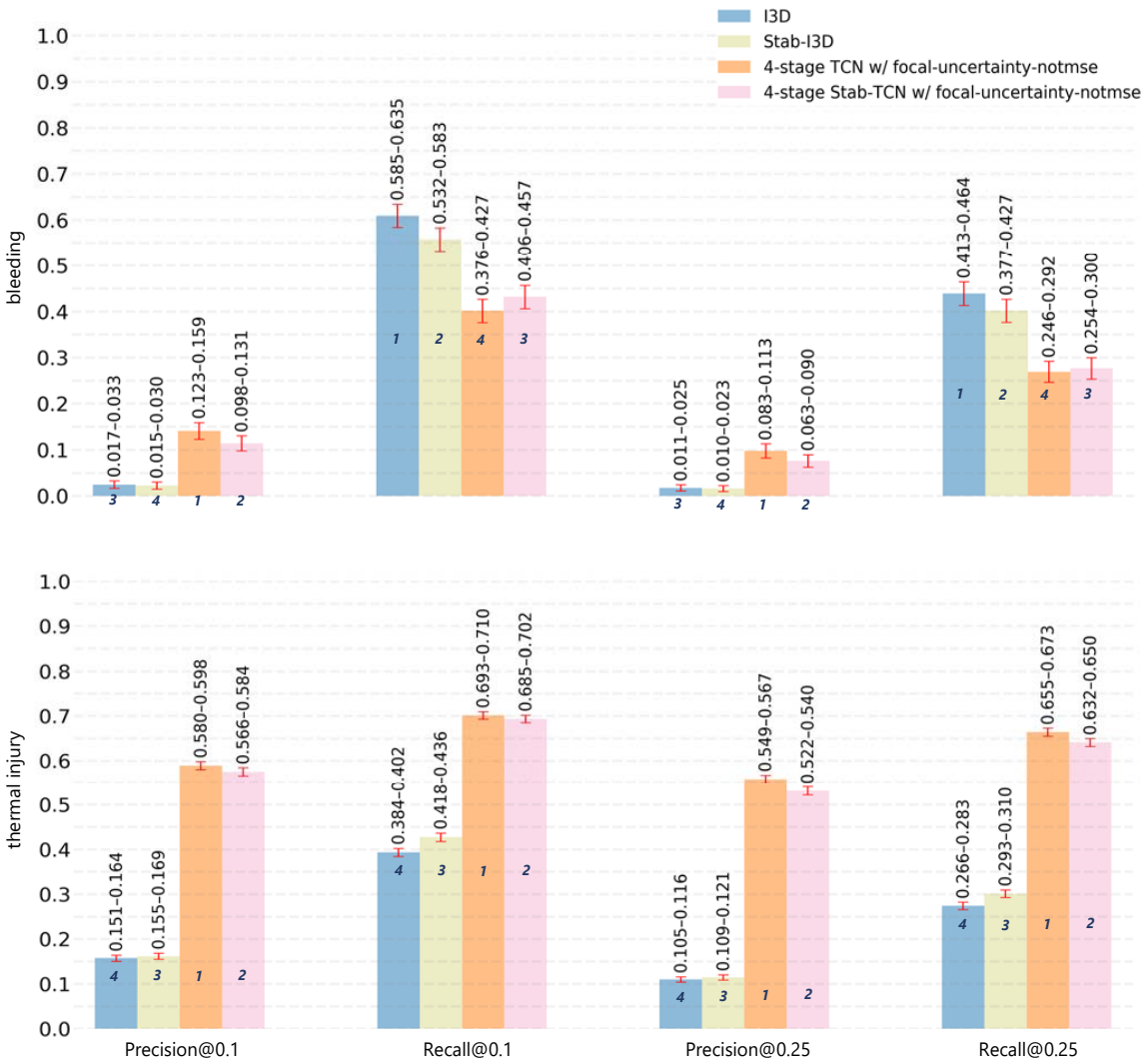


Figure 3: Effect of stabilization on segmental results: Stabilization improves model performance in segmental recall in bleeding detection of 4-stage TCN. The vertical axis shows the segmental precision, and recall of each model at IOU of 0.1, and 0.25.

5.4. Effect of Focal and Uncertainty Loss

The effect of the losses on 4-stage Stab-TCN is shown in Fig. 2. The ranking of losses performance in thermal injury detection is $\mathcal{L}_{focal-uncertainty-notmse} > \mathcal{L}_{focal-uncertainty} > \mathcal{L}_{focal} > \mathcal{L}_{normal} > \mathcal{L}_{uncertainty}$. There is no clear trend in bleeding detection.

For thermal injury detection, the model trained with $\mathcal{L}_{focal-uncertainty-notmse}$ loss perform significantly better than those with the other losses in all metrics, shown in the bottom figure of Fig. 2. The differences in frame-wise metric values between the best loss and the worst loss are 0.025 in AUROC (lower bound CI: 0.938–0.963) and 0.055 in AP (lower bound CI: 0.562–0.617), the $\mathcal{L}_{focal-uncertainty-notmse}$ loss improves AP@0.1 and AP@0.25

of the second best loss, $\mathcal{L}_{focal-uncertainty}$, by 8.62% (0.429–0.466) and 3.78% (lower bound CI: 0.344–0.357) respectively.

For bleeding detection, shown in the top figure of Fig. 2, the effect of $\mathcal{L}_{focal-uncertainty-notmse}$ loss in event-wise metrics are not as evident as in thermal injury detection. Its AP@0.1, and its AP@0.25 performance are 0.001, and 0.003 lower than the highest ones in the two metrics respectively.

Due to large gains in performance in thermal injury detection using the $\mathcal{L}_{focal-uncertainty-notmse}$ loss, we ignore1 the minor decrease in performance in bleeding detection, and choose the $\mathcal{L}_{focal-uncertainty-notmse}$ loss for further analysis.

5.5. Effect of Stabilization

Stabilization has very different effects on different models. Fig. 3 illustrates a comparison of model performance with and without stabilization.

Stabilization increases I3D precision in thermal injury detection by 2.65% and 3.81% in precision@0.1 and precision@0.25 respectively. Its recall@0.1, and recall@0.25 increase by 8.85% and 10.15% to 0.384 and 0.266 respectively. Despite of a significant improvement in thermal injury detection, improvements in bleeding detection by I3D are not significant.

By contrast, the 4-stage Stab-TCN has better segmental recall than the 4-stage TCN in bleeding detection. Its segmental recall increases from 0.376 to 0.406 for recall@0.1, and from 0.246 to 0.254 for recall@0.25.

5.6. Comparison to Baseline Models

The 4-stage Stab-TCN systems outperform the InceptionV3, and the Stab-I3D in AUROC, AP, and AP@k. The comparison results are shown in Fig. 4, and the precision-recall curves of all cross-validation results are in Appendix D.

The 4-label 4-stage Stab-TCN has similar bleeding detection performance as the 4-stage Stab-TCN at the range around 0.121–0.156 in AP@0.1. However, the 4-label 4-stage Stab-TCN has significantly lower AP@k in thermal injury detection at around 0.466–0.485 comparing to 0.531–0.550 for 4-stage Stab-TCN in AP@0.1. This suggests that the focal-uncertainty loss does not fully address class imbalance, and we need task-specific classifier. The 4-stage Stab-TCN, which is a task-specific classifier trained only on one class, has the best overall performance among the four models.

Comparing to the 1-stage Stab-TCN, the 4-stage Stab-TCN's event-wise performance are 2 to 3 times higher in bleeding detection, despite they share similar performance in frame-wise metrics. In thermal injury, 1-stage Stab-TCN has better performance that improves 4-stage Stab-TCN's AP@0.1 from 0.531–0.550 to 0.566–0.584, which is an increase of 6.60%. Although 1-stage Stab-TCN provide decent improvement in thermal injury detection, its event-wise performance in bleeding detection is the worst among all models.

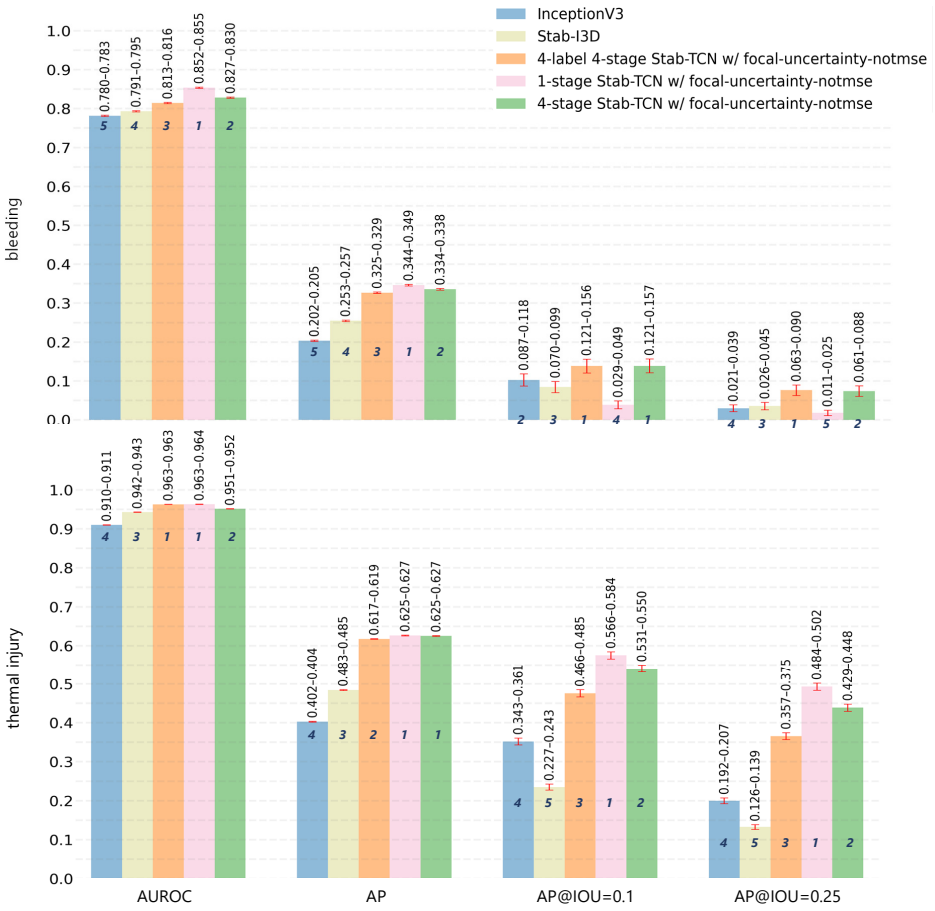


Figure 4: Baseline comparison: Stab-TCN models outperforms InceptionV3 and I3D with more than 10% difference in most metrics, and 4-stage Stab-TCN has the best overall performance. The vertical axis shows the frame-wise AUROC and AP, and segmental AP at IOU of 0.1, and 0.25 of each model.

To further study the performance of each model for high recall and low precision, we evaluate model performance using frame-wise precision, recall, and specificity, and event-wise precision, and recall at thresholds chosen at the best validation segmental recall. Using the selected thresholds, we 'event plot' the detection of a randomly selected full procedure in Fig. 5. We use two types of lines with different lengths and colours indicating multiple tasks in one frame. The longer blue line indicates a bleeding event occurs at the time, and the shorter orange one indicates the occurrence of a thermal injury event.

The event plot of Stab-TCN (Fig. 5) has the best readability and usability with fewer incorrect transitions, and more correctly and precisely detected events comparing to I3D and InceptionV3.

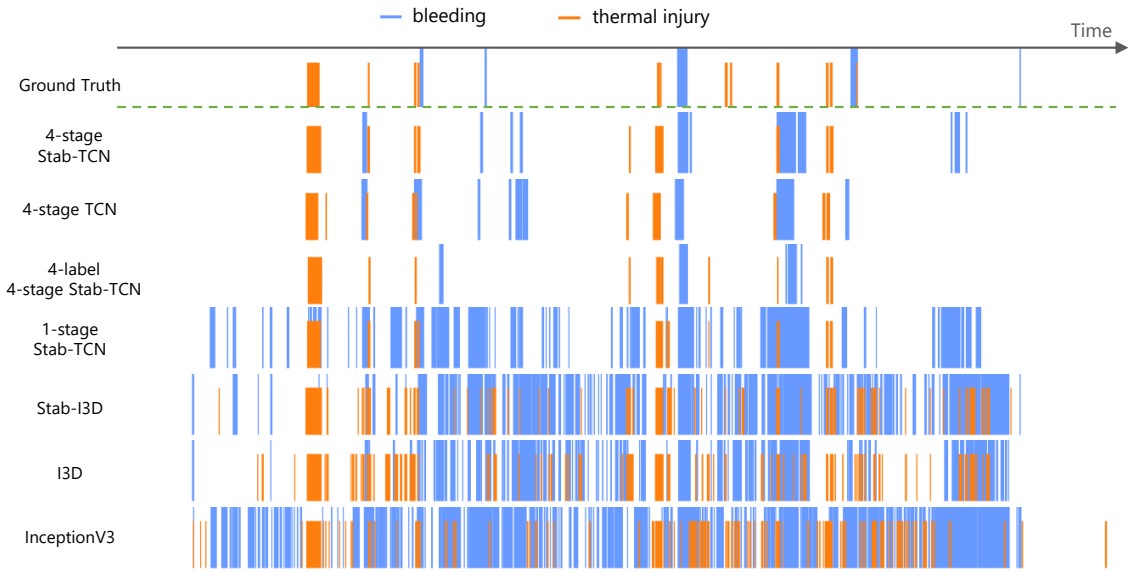


Figure 5: Event plots: the results of the 4-stage TCN with or without stabilization has less false positives, and are more readable and useful for a clinical to identify region of interests in a long procedure.

5.7. Use Case Analysis

Due to trade-offs between recall and precision, we computed the 4-stage Stab-TCN performance in the high-recall and low-precision scenario, and in the high-precision and low-recall scenario, as shown in Table 2. In the former scenario, our system has a segmental recall@0.1 of 0.694 and a segmental precision@0.1 of 0.575 for thermal injury detection, and 0.432 and 0.114 for bleeding detection.

Table 2: Performance of 4-stage Stab-TCN with $\mathcal{L}_{focal-uncertainty-notmse}$ loss in different use case scenarios ([a]"R." as recall, [b]"P." as precision.).

	bleeding		thermal injury	
	High R.[a] Low P.[b]	High P. Low R.	High R. Low P.	High P. Low R.
Precision @0.1	0.114	0.324	0.575	0.702
Recall @0.1	0.432	0.223	0.694	0.558
Precision @0.25	0.076	0.302	0.531	0.657
Recall @0.25	0.277	0.113	0.641	0.522

6. Discussion

In an ablation study of loss functions, our approach of integrating focal and uncertainty losses outperforms other losses. The combined loss assigns more weights to samples that are harder to learn, as well as learn task uncertainty in a separate term to force the model weight to be more precise. Therefore, the model trained on the focal and uncertainty loss better addresses the unbalanced dataset and multi-label problem.

We also found that, contrary to Farha and Gall (2019), removing the T-MSE loss from $\mathcal{L}_{focal-uncertainty}$ improves model performance in thermal injury. The T-MSE loss acts as a smoothing filter that corrects over-segmentation errors by forcing the model to minimize generated transitions between events. Although it reduces short false positives in event transition, it also removes short positive event segments. Unlike previously studied action recognition tasks, which produce a long series of actions, our TAL task has long empty gaps between the short instances (e.g., the orange events in Fig. 1C). Using the T-MSE loss eliminates short instances along with the false positives. As the MS-TCN model corrects over-segmentation errors over each stage, the T-MSE over-corrects the errors and lowers performance. Therefore, the model with $\mathcal{L}_{focal-uncertainty-notmse}$ performs better than the one with $\mathcal{L}_{focal-uncertainty}$.

In an ablation study of stabilization algorithms, we found that the trade-off of using stabilization in both of I3D and MS-TCN is that the increase in thermal injury performance will lead to a decrease in bleeding detection performance. We found the same behaviour in recall, and precision — stabilization improves recall, and leads to decrease in precision. This is because small distortions and cropping are inevitable in bundled-camera path estimation, and it becomes more visible in the longer bleeding event. Due to this trade-off, stabilization is more useful in detection of events that have more motion, and when we prefer higher recall to precision.

The model comparison shows MS-TCN based model has higher segmental precision and recall than I3D or InceptionV3 model. It extracts longer temporal features, as it performs multiple dilated convolution in the time domain of a full procedure. In accordance with Farha and Gall (2019), the 4-stage Stab-TCN has better performance than single-stage Stab-TCN because multi-stage model corrects more over-segmentation errors over multiple stages of temporal convolution. For our suggested clinical use case as a screening tool, the detection of 4-stage Stab-TCN shown in Fig. 5 has better readability with less false positives.

Fig. 5 also shows strongly relying on frame-wise metrics, such as AUROC and AP only, will result in a wrong model selection, such as 1-stage Stab-TCN. Due to its low precision, 1-stage Stab-TCN would complicate efficiently identifying probable iAEs, especially bleeding events. For temporal action localization problem, segmental metrics are more reliable in model selection, and performance evaluation.

Clinically, using our system as a screening tool, with segmental precision of 0.575 (thermal injury) and 0.114 (bleeding), human annotators would only need to filter out approximately 40% of identified false-positive thermal injury events, and approximately 90% of identified false-positive bleeding events. Note that this system only highlights an average of 13.45 thermal injury events and an average of 13.74 bleeding events per 1.5-hour procedure. Reviewing these events (in less than 10 minutes) to remove false positives is less

laborious than reviewing an entire video (average 1.5 hour per video). In the high-precision and low-recall scenario, the system would miss many true events, but each detection has a high probability to be a true positive. Our system has a segmental precision@0.1 of 0.702 and a segmental recall@0.1 of 0.558 for thermal injury detection, and 0.324 and 0.223 for bleeding detection. With some assistance from human experts to filter out false positives, this system can be used as a data mining tool to collect an intraoperative adverse event database.

Our proposed system integrates stabilization with multi-stage temporal convolutional neural network with a focal-uncertainty loss. This system addresses problems including highly unbalanced dataset, multi-label detection task, and temporal localization of short and rare event in long videos. This unique temporal action localization problem is more common in clinical event detection than the well-studied human activity recognition problem. Our analysis shows the effect of each module contribution to the state-of-the-art performance. This method can be applied to other event detection in long surgical videos.

Limitations Our system recognizes adverse events of all levels including less serious adverse events. Clinicians that are only interested in highly severe adverse events would need to filter them out. This system does not support real time processing as stabilization is time consuming. For a video with a resolution of 1280 x 720, our stabilizer reaches a speed of 7 FPS. Using resolution of 224 x 400 increases the speed from 7 FPS to 11 FPS. However, the speed of stabilization is still a bottleneck for real time processing.

References

BIS Research. Global laparoscopy and endoscopy devices market: Focus on surgical procedures (cholecystectomy and hysterectomy) and product types (arthroscopes, neuroendoscopes, cystoscope, and bronchoscopes) - analysis and forecast, 2018-2025. Technical report, BIS Research, 2018. URL https://www.researchandmarkets.com/research/626pqp/global?w=12.

Jordan D. Bohnen, Michael N. Mavros, Elie P. Ramly, Yuchiao Chang, D. Dante Yeh, Jarone Lee, Marc De Moya, David R. King, Peter J. Fagenholz, Kathryn Butler, George C. Velmahos, and Haytham M.A. Kaafarani. Intraoperative adverse events in abdominal surgery: What happens in the operating room does not stay in the operating room. In *Annals of Surgery*, volume 265, pages 1119–1125. Lippincott Williams and Wilkins, jun 2017. doi: 10.1097/SLA.0000000000001906. URL http://journals.lww.com/00000658-201706000-00015.

J. Carreira and A. Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, July 2017. doi: 10.1109/CVPR.2017.502.

R. Cipolla, Y. Gal, and A. Kendall. Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, June 2018. doi: 10.1109/CVPR.2018.00781.

Yazan Abu Farha and Juergen Gall. MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL http://arxiv.org/abs/1903.01945.

Y. Fu, W. Zhang, M. Mandal, and M. Q. . Meng. Computer-Aided Bleeding Detection in WCE Video. *IEEE Journal of Biomedical and Health Informatics*, 18(2):636–642, March 2014. ISSN 2168-2194. doi: 10.1109/JBHI.2013.2257819.

Alvaro Garcia-Martinez, Jose María Vicente-Samper, and José María Sabater-Navarro. Automatic detection of surgical haemorrhage using computer vision. *Artificial Intelligence in Medicine*, 78:55 – 60, 2017. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2017.06.002. URL http://www.sciencedirect.com/science/article/pii/S0933365716305590.

Atul A. Gawande, Eric J. Thomas, Michael J. Zinner, and Troyen A. Brennan. The incidence and nature of surgical adverse events in Colorado and Utah in 1992. *Surgery*, 126(1):66–75, Jul 1999. ISSN 0039-6060. doi: 10.1067/msy.1999.98664. URL https://doi.org/10.1067/msy.1999.98664.

Atul A. Gawande, Michael J. Zinner, David M. Studdert, and Troyen A. Brennan. Analysis of errors reported by surgeons at three teaching hospitals. *Surgery*, 133(6):614–621, Jun 2003. ISSN 0039-6060. doi: 10.1067/msy.2003.169. URL https://doi.org/10.1067/msy.2003.169.

T. Ghosh, S. K. Bashar, S. A. Fattah, C. Shahnaz, and K. A. Wahid. A feature extraction scheme from region of interest of wireless capsule endoscopy images for automatic bleeding detection. In *2014 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 256–260, Dec 2014. doi: 10.1109/ISSPIT.2014.7300597.

Haroon Idrees, Amir R. Zamir, Yu Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. ISSN 1090235X. doi: 10.1016/j.cviu.2016.10.018. URL www.elsevier.com/locate/cviu.

X. Jia and M. Q. . Meng. A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 639–642, Aug 2016.

K. Jo, B. Choi, S. Choi, Y. Moon, and J. Choi. Automatic detection of hemorrhage and surgical instrument in laparoscopic surgery image. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1260–1263, Aug 2016. doi: 10.1109/EMBC.2016.7590935.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. *ArXiv*, abs/1705.06950, 2017.

Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 780–787. IEEE Computer Society, sep 2014. ISBN 9781479951178. doi: 10.1109/CVPR.2014.105.

C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal Convolutional Networks for Action Segmentation and Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1003–1012, July 2017. doi: 10.1109/CVPR.2017.113.

Lucian L. Leape, Ann G. Lawthers, A. Russell Localio, Benjamin A. Barnes, Liesi Hebert, Joseph P. Newhouse, Howard Hiatt, Nan Laird, Troyen A. Brennan, Howard Hiatt, Troyen A. Brennan, Joseph P. Newhouse, and Paul C. Weiler. The nature of adverse events in hospitalized patients: Results of the harvard medical practice study II. *New England Journal of Medicine*, 324(6):377–384, feb 1991. ISSN 15334406. doi: 10.1056/NEJM199102073240605. URL http://www.nejm.org/doi/abs/10.1056/NEJM199102073240605.

X. Li, H. Zhang, X. Zhang, H. Liu, and G. Xie. Exploring transfer learning for gastrointestinal bleeding detection on small-size imbalanced endoscopy images. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1994–1997, July 2017. doi: 10.1109/EMBC.2017.8037242.

T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, Oct 2017. doi: 10.1109/ICCV.2017.324.

Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. Bundled Camera Paths for Video Stabilization. *ACM Trans. Graph.*, 32(4):78:1–78:10, July 2013. ISSN 0730-0301. doi: 10.1145/2461912.2461995. URL http://doi.acm.org/10.1145/2461912.2461995.

Martin A Makary and Michael Daniel. Medical error—the third leading cause of death in the US. *BMJ*, 353, 2016. doi: 10.1136/bmj.i2139. URL https://www.bmj.com/content/353/bmj.i2139.

Alberto Montes, Amaia Salvador, Santiago Pascual, and Xavier Giro-i Nieto. Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks. In *1st NIPS Workshop on Large Scale Computer Vision Systems*, December 2016.

Takayuki Okamoto, Takashi Ohnishi, Hiroshi Kawahira, Olga Dergachyava, Pierre Jannin, and Hideaki Haneishi. Real-time identification of blood regions for hemostasis support in laparoscopic surgery. *Signal, Image and Video Processing*, 13(2):405–412, Mar 2019. ISSN 1863-1711. doi: 10.1007/s11760-018-1369-7. URL https://doi.org/10.1007/s11760-018-1369-7.

A. Richard and J. Gall. Temporal Action Detection Using a Statistical Language Model. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3131–3140, June 2016. doi: 10.1109/CVPR.2016.341.

E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, pages 2564–2571, Nov 2011. doi: 10.1109/ICCV.2011.6126544.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015. ISSN 1573-1405. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, June 2016. doi: 10.1109/CVPR.2016.308.

Eric J. Thomas, David M. Studdert, Joseph P. Newhouse, Brett I. W. Zbar, K. Mason Howard, Elliott J. Williams, and Troyen A. Brennan. Costs of Medical Injuries in Utah and Colorado. *Inquiry*, 36(3):255–264, 1999. ISSN 00469580, 19457243. URL http://www.jstor.org/stable/29772835.

Muhammad Arslan Usman, G.B. Satrya, Muhammad Rehan Usman, and Soo Young Shin. Detection of small colon bleeding in wireless capsule endoscopy videos. *Computerized Medical Imaging and Graphics*, 54:16 – 26, 2016. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2016.09.005. URL http://www.sciencedirect.com/science/article/pii/S0895611116300970.

Huijuan Xu, Abir Das, and Kate Saenko. Two-Stream Region Convolutional 3D Network for Temporal Activity Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2319–2332, mar 2019. ISSN 19393539. doi: 10.1109/TPAMI.2019.2921539. URL http://arxiv.org/abs/1703.07814.

Marieke Zegers, Martine C de Bruijne, Bertus de Keizer, Hanneke Merten, Peter P Groenewegen, Gerrit van der Wal, and Cordula Wagner. The incidence, root-causes, and outcomes of adverse events in surgical units: Implication for potential prevention strategies. *Patient Safety in Surgery*, 5(1):13, 2011. ISSN 17549493. doi: 10.1186/1754-9493-5-13. URL http://pssjournal.biomedcentral.com/articles/10.1186/1754-9493-5-13.

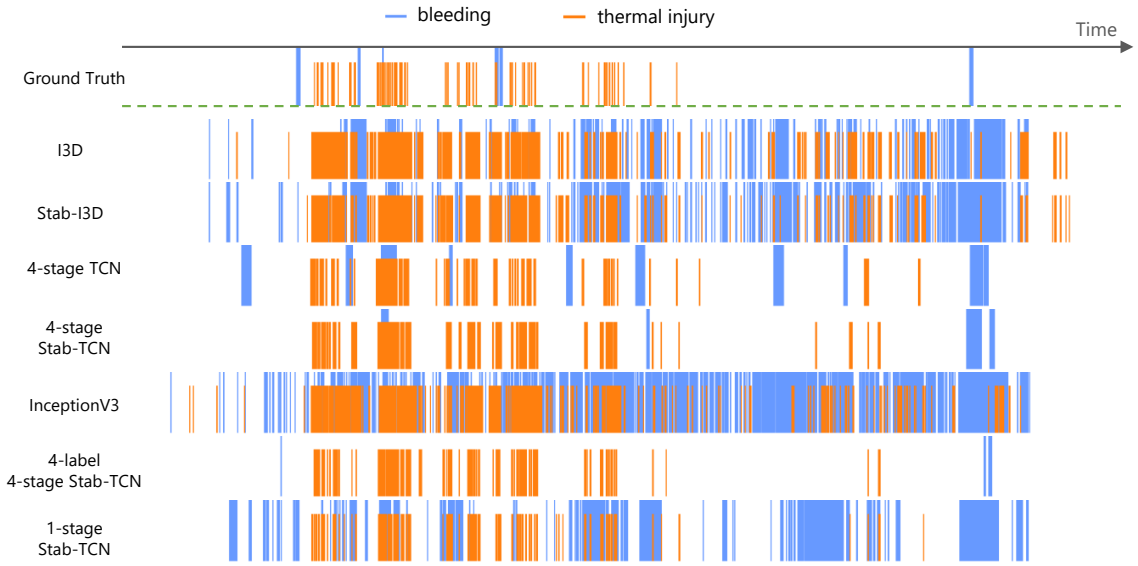
Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal Action Detection with Structured Segment Networks. *International Journal of Computer Vision*, 128(1):74–95, apr 2020. ISSN 15731405. doi: 10.1007/s11263-019-01211-2. URL http://arxiv.org/abs/1704.06228.

Appendix A. Generalization to Different Cases

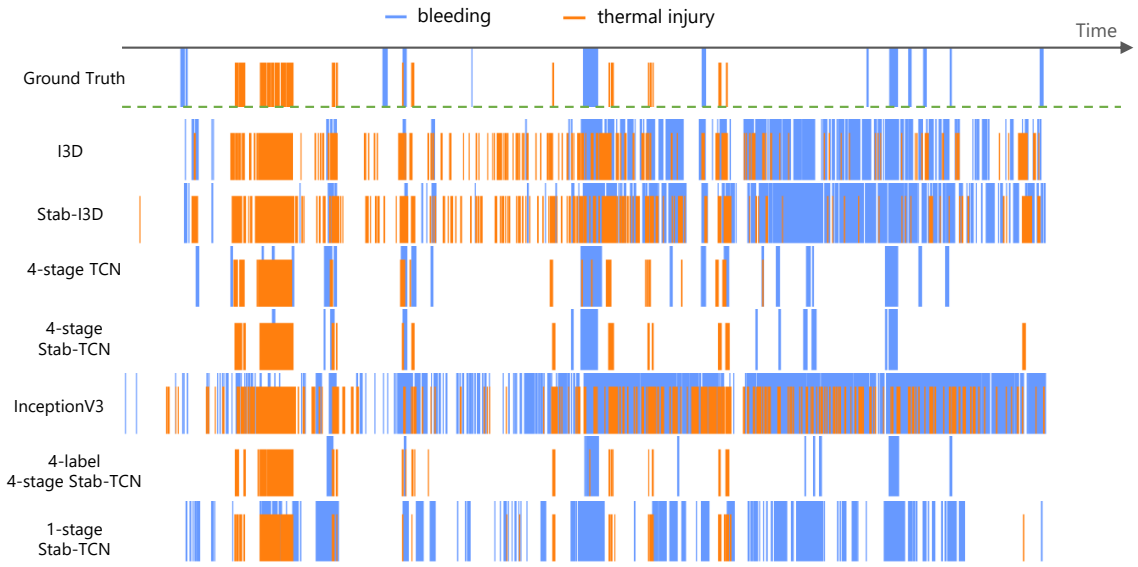
Fig. 6 shows the quality of our system in comparison to the others in generalizing to different cases.

Appendix B. Data Distribution by Classes

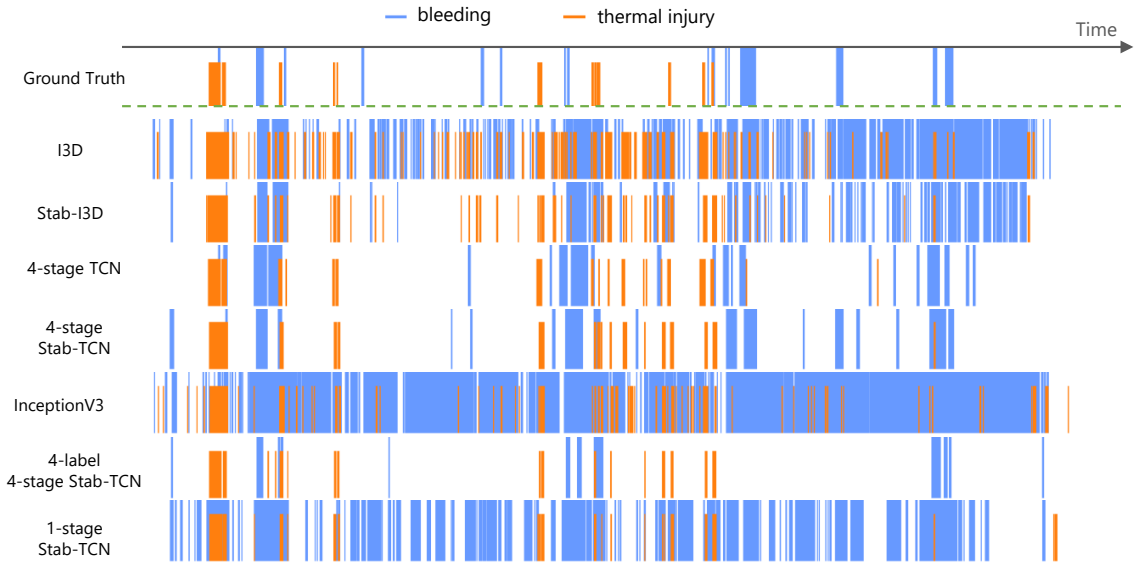
The data distribution by task is highly unbalanced. This is true for distribution of each class. Table. 3 shows a highly unbalanced distribution of each class.



(A)



(B)



(C)

Figure 6: Event plots of three cases where the horizontal axis represents time. Each of the vertical bars represents a detection of bleeding and/or thermal injury. The lines are in different lengths and colours indicating multiple tasks in one frame. The longer blue line represents bleeding, and the shorter orange one represents thermal injury.

Table 3: Distribution of frames in each class shows the dataset is highly unbalanced. Each class represents a unique combination of tasks. For example, *1 1 0 0* represents a class that has blood, active bleeding, but not burn and not active thermal injury.

Class	Blood	Bleeding	Burn	Thermal Injury	# Frames
0 0 0 0	Off	Off	Off	Off	10,118,163
0 0 0 1	Off	Off	Off	***On***	277,806
0 0 1 0	Off	Off	***On***	Off	272,719
0 0 1 1	Off	Off	***On***	***On***	28,283
1 0 0 0	***On***	Off	Off	Off	10,941,450
1 0 0 1	***On***	Off	Off	***On***	445,355
1 0 1 0	***On***	Off	***On***	Off	1,203,348
1 0 1 1	***On***	Off	***On***	***On***	127,454
1 1 0 0	***On***	***On***	Off	Off	1,080,771
1 1 0 1	***On***	***On***	Off	***On***	50,057
1 1 1 0	***On***	***On***	***On***	Off	82,399
1 1 1 1	***On***	***On***	***On***	***On***	9,628
				Total	**24,637,433**

Appendix C. Demo Videos

Demo videos of stabilization, and intra-operative adverse event detection are provided in supporting files. The detection demo includes bleeding detection with zooming, and with occlusion, as well as thermal injury detection with and without smoke in the scene. Before and after the events, we leave some period with no-event to demonstrates the detection of true negatives. You can download the demo videos here https://drive.google.com/drive/folders/1ypiXuk9vaAiM7eMc1ilEL3_23x-A2iij?usp=sharing.

Appendix D. Precision-Recall Curves

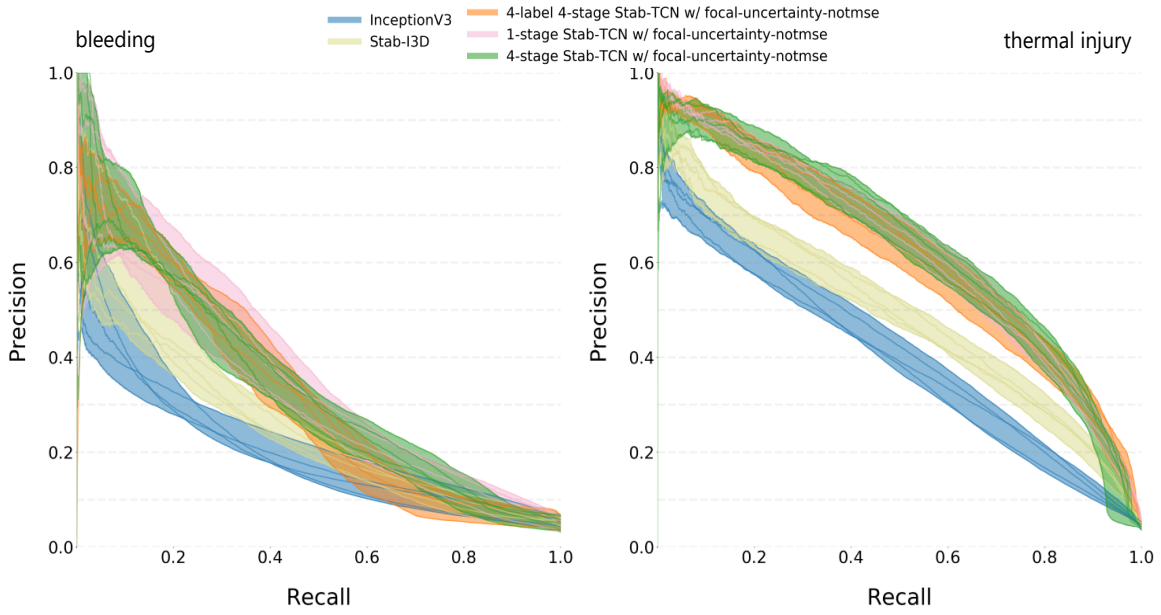


Figure 7: Precision-recall curves of each task. Model name with its loss function name is used in the legend, where *focal-uncertainty-notmse* is a loss combining focal and uncertainty loss without truncated mean-square error. Each model has 5 corresponding lines from 5-fold cross-validation. The region with colour is an indication of the range of the precision-recall curves.

We plot a precision-recall curve of each model on cross-validation results at each fold. Each model has 5 curves that forms an area in between shown in Fig. 7. Model with stabilization and temporal convolutional neural networks are on top of the curves of InceptionV3, and Stab-I3D, and the area between the curves are narrower than those of InceptionV3, and Stab-I3D. This provides empirical evidences of the performance improvements using stabilization and temporal convolutional network in comparison to 2D or 3D convolutional neural network.