Amplifying action-context greater: Image segmentation-guided intraoperative active bleeding detection

SeulGi Hong, SeungBum Hong, Junyoung Jang, Keunyoung Kim, Woo Jin Hyung and Min-Kook Choi

VisionAI, hutom, South Korea

ARTICLE HISTORY
Compiled September 13, 2022

ABSTRACT
Intraoperative active bleeding (iAB) is a representative adverse event, and in most surgical operations, it delays the operation time and damages organs, affecting the patient's outcome. The iAB detection model can be used for image-guided surgery and as a major statistical index in predicting patient outcomes after surgery. However, detecting iAB is difficult due to the similarity between active and non-active bleeding or active bleeding in a small area. Using the spatial and temporal characteristics of the iAB area within frames simultaneously can overcome this. We propose a novel training method that can adequately fuse image segmentation and temporal action localization models for effective iAB detection. The proposed active bleeding detection model has the following supervision process: First, annotate temporal localization information for active bleeding that is relatively easy to annotate. Next, in the active bleeding section, where temporal localization is annotated, spatial localization information for selected frames is annotated and used as auxiliary information for active bleeding detection. We constructed a cross-validation set of 40 robotic subtotal gastrectomies and verified the ability of an active bleeding model guided by image segmentation information to bring improvements to the active bleeding recognition task. In addition, we applied performance evaluation for outcome analysis by measuring errors in iAB duration and counting in surgical videos for each algorithm. [1]

KEYWORDS
active bleeding detection; temporal action localization; semantic segmentation

1. Introduction

Since adverse intraoperative events (iAEs) delay surgery time and harm the patient (Bohnen et al. 2016; Bonrath et al. 2013), it is essential to prevent adverse events in advance. About 23% of iAEs accompany active bleeding, which is detrimental to the patient during the hemostasis process and delays the operative time. (Zegers et al. 2011). Detecting intraoperative Active Bleeding (iAB) in surgery has two key implications. The first is to provide surgeons with a computer-assisted surgery (CAS) environment that can increase the efficiency of surgery by detecting the bleeding point of iAB early in minimally invasive surgery. It requires highly accurate real-time recognition accompanied by localization of active bleeding points. The second is to be a surgical index for predicting the prognosis of patients after surgery and evaluating

[1]Please refer to https://sghong977.github.io/bleeding/ for supplementary materials.

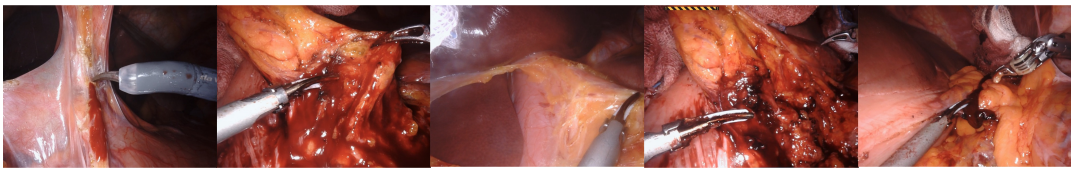


Figure 1. Taxonomy of intraoperative bleedings.

surgery. Here, the surgical index is related to a statistical index that can affect the surgical outcome. Assuming accurate detection of iAB can be achieved in a surgical video recorded during surgery, it can be utilized as a major surgical index related to the patient's outcome and recoveries, such as bleeding time, frequency, and amount of bleeding (Numata et al. 2021; Aoyama et al. 2020). At the same time, these indices can be actively used for postoperative surgical evaluation and review.

Although detecting intraoperative Active Bleeding (iAB) is important, it demands heavy cost and time of experts. The automated system of detecting iAB is essential for this reason, however, there are two main reasons why automated iAB detection has been difficult so far. First, the color and texture of iAB are similar to those of other organs or non-active bleeding. Unlike medical videos such as diagnostic endoscopy, the surgical videos of active bleeding need to be defined more precisely, as bleeding requires hemostasis during surgery or for use as a surgical index. Figure 1 shows the types of bleeding that can be found in the surgical video. Active bleeding refers to bleeding that is actively flowing by the time of surgical intervention (Bohnen et al. 2016; Bonrath et al. 2013). Blood-stained tissue, inner fascia bleeding, and stagnant blood in a specific area should be excluded from active bleeding. Next, if the model is eager to recognize iAB appearing in the surgical video, it must be able to encode the spatiotemporal features of active bleeding properly. Actions in the general video appear throughout its duration, but iAB in surgical video occurs locally in the spatiotemporal domain, resulting in a severe class imbalance. Effective spatiotemporal feature encoding should be involved in the training and inference process on the iAB to mitigate this.

We formulate the detection of iAB as a temporal action localization (TAL) problem, similar to the study in (Wei et al. 2021). We applied the following approach to overcome the two challenges in iAB detection:

- The temporal and spatial locations are utilized simultaneously as supervision information for effective spatiotemporal feature encoding of iAB.
- A semantic segmentation network is pre-trained on the relatively small dataset to reduce the annotation burden of spatial location; it generates pseudo-labels while training the detection model for iAB.
- The operation is evaluated from the TAL perspective, and the evaluation is meaningful as a surgical index in the entire operation video. The iAB error of the duration and count is evaluated in the post-operation analysis.

We propose a model that AMplifies Action-context Greater: the Image segmentation-guided active bleeding detection model (AMAGI) to encode these spatiotemporal feature information simultaneously. Utilizing the AMAGI model, we achieved high accuracy in iAB detection compared to other SOTA action recognition models or semantic segmentation models in the surgical video, along with the lowest error rate even in the surgical index.

2. Related Work

Intraoperative active bleeding (iAB) detection. Bleeding detection in medical videos was performed first in diagnostic endoscopy or wireless capsule endoscopy (WCE) videos. Bleeding detection in endoscopic videos utilized early handcrafted features (Fu et al. 2014; Usman et al. 2016), and relatively recent studies using deep learning have been introduced (Li et al. 2019). Although there is a technical similarity in the bleeding recognition in the video, the detection difficulty of iAB is relatively high in the classification of the target's active bleeding and the complexity of the environment in the video. Bleeding detection in minimally invasive surgery has been relatively recent. For iAB detection, similar to endoscopic video, an approach using color features was first introduced (Jo et al. 2016; Garcia-Martinez et al. 2017; Okamoto et al. 2019). (Jo et al. 2016) proposed a method to utilize the distribution of color values in a specific color space, and (Garcia-Martinez et al. 2017; Okamoto et al. 2019) suggested a machine learning-based classifier for color features. Recently, a CNN-based object detector-based bleeding recognizer has also been proposed (Hua et al. 2022). However, it performed iAB detection in a limited number of frames and did not evaluate from the perspective of the entire surgical video.

More recently, in terms of CAS and surgical video analysis, iAB detection has been introduced in combination with modern computer vision techniques (Wei et al. 2021; Rabbani 2022). In (Wei et al. 2021), the detection of iAB was introduced in the context of iAE and active bleeding detection in the surgical video was performed using a Multi-Stage Temporal Convolutional Network (MS-TCN) (Li et al. 2020) based on the extracted I3D features (Carreira and Zisserman 2017). It attempts to alleviate the class imbalance problem by utilizing the focal loss (Lin et al. 2020) during training, but the evaluation performance was not high enough throughout the surgical video. In (Rabbani 2022), a Space-Time Memory (STM) (Oh et al. 2019) network was utilized to formulate iAB detection as a video object segmentation problem. Synthetic data was used for training to overcome the limitation of requiring a large amount of spatial information for annotation data due to the characteristics of video object segmentation. Also, an adversarial domain adaptation technique was applied to effectively train on synthetic data. However, training and evaluation were performed from a dataset consisting of limited frames, but the full surgical video was not evaluated.

Spatiotemporal feature fusion. For active bleeding recognition, a model should deeply appreciate the spatiotemporal context. Unlike standard multi-modal computer vision architectures, we only use video frames without any other modality, and delicate feature processing for understanding spatiotemporal context is crucial in our case. In the previous work, BTSNet (Hong and Choi 2021) introduced architecture to understand input video clips by spatiotemporal feature fusion. The temporal-spatial pathway (TSP) block of BTSNet contains candidate features from various spatial and temporal scales and calculates the softmax attention among them. Then these attentions are used for re-weighting given input features, and the features are aggregated. Similar to BTSNet in terms of feature engineering, we also define several candidate features for the spatiotemporal surgical information wanted to be obtained and re-weight candidates. The difference between our fusion idea and BTSNet is that in our method, the bleed/non-bleed attentions are obtained from the segmentation branch. These region attentions with slow and fast temporal scale information are aggregated for the following convolutional operations. Finally, we define four candidate features.

Besides, high-quality spatial and temporal information is required to improve the recognition model of active bleeding. Primarily, temporal context is essential to distinguish active bleeding from a visually similar scene, for example, blood-stained tissues in Figure 1. Video action recognition, one of the representative computer vision tasks, has been actively developed to understand the video context (Tran et al. 2015; Feichtenhofer et al. 2019; Lin et al. 2019; Feichtenhofer 2020). MS-TCN (Li et al. 2020) focused on temporal dependencies and suggested a dilated layer that combines small and large receptive fields, and the previous work (Wei et al. 2021) used this model for the active bleeding task. In our case, we embedded the commonly used SlowFast (Feichtenhofer et al. 2019) as a video backbone into our fusion model. Since SlowFast is capable of extracting slow and fast context in each pathway, these features are best suited for our fusion model to be manipulated.

Furthermore, we attained spatial context from the semantic segmentation architecture. Semantic segmentation is a pixel-wise classification task for generating label masks (Chen et al. 2018; Xiao et al. 2018; Yuan et al. 2020). In this work (Hong et al. 2020), the semantic segmentation benchmark for cholecystectomy is released for computer-assisted surgery mechanisms. In our case, we also utilize our spatial localization dataset to understand active bleeding events in surgery. We trained the widely-used segmentation model OCR (Yuan et al. 2020) to generate possibly bleeding regions. In summary, we exploit existing well-designed spatial and temporal context backbones and devise a fusion-based architecture for active bleeding recognition. Our model is end-to-end except for using the pre-trained segmentation branch for inference.

3. Methods

In this section, we will explain our model that amplifies action-context greater: the image segmentation-guided active bleeding localization model (AMAGI). We define the active bleeding task as having a temporal localization. Temporal localization is one of the video action recognition tasks that aim to detect activities in the video stream by predicting the beginning and end timestamps. In the active bleeding scenario, our framework should predict the start and end frames of active bleeding in the entire surgery recording. We verify our method in the temporal localization task since our dataset marks entire bleeding events for given surgery videos. For example, the annotation can evaluate when the bleeding occurred and how long the bleeding lasted.

Our approach is based on a fusion idea that exploits different specialized networks for recognition to solve the given active bleeding task. In Figure 3, our active bleeding model fuses the information extracted from two branches: video-based action recognition and frame-based semantic segmentation. Our model exquisitely grabs the surgical context by taking advantage of both temporal and spatial features. Furthermore, the model output is post-processed in the evaluation protocol for better extraction of the surgical analysis indexes. These surgical indexes are provided to clients to advise how much the patient loses blood during surgery. The target surgery type of our framework is robotic subtotal gastrectomy.

Our elaborately designed active bleeding framework can well extract the surgical analysis index for clinical use. We will describe the temporal context model in 3.1, the spatial context model in 3.2, the proposed fusion-based bleeding recognition model in 3.3, and the post-processing and surgical analysis index in 3.4.

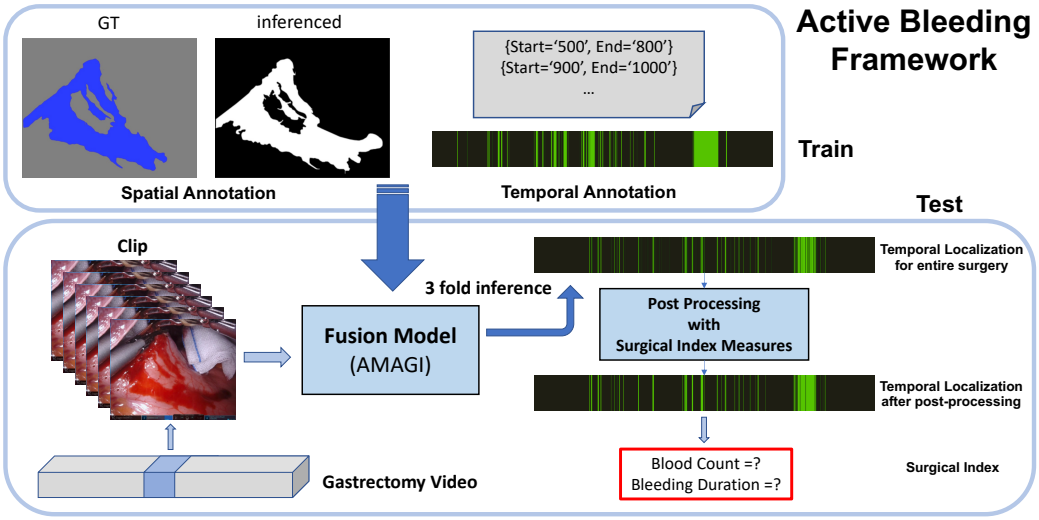


Figure 2. Our active bleeding framework for the entire surgery has a long sequence of more than hours.

3.1. Temporal Context Model

First, we use a video-based backbone model to extract the temporal context. We considered two video backbones, SlowFast (Feichtenhofer et al. 2019) and Multi-Stage TCN (Li et al. 2020; Wei et al. 2021). SlowFast (Feichtenhofer et al. 2019) is a widely used 3D CNN-based video action recognition network. It has slow and fast pathways responsible for extracting contextual information at different tempos. We choose SlowFast as the video backbone of our fusion model, AMAGI. Besides, the existing active bleeding framework (Wei et al. 2021) uses Multi-Stage TCN (Li et al. 2020) to see the temporal context. But MS-TCN is not used for the proposed feature fusion architecture due to its lower performance.

To summarize, we executed three experiments related to temporal context. First, the proposed fusion model AMAGI contains a video branch. We use the features of SlowFast for our fusion architecture. Second, we also report the active bleeding performance of the vanilla SlowFast (Feichtenhofer et al. 2019) network to verify the effectiveness of our proposed fusion algorithm. Third, we show the performance of the MS-TCN network to compare with previous work (Wei et al. 2021).

3.2. Spatial Context Model

Although the video-based recognition model comprehends video sequences well, high-level intelligence is necessary to understand active bleeding in surgery videos. To this end, we adopt the semantic segmentation backbone model to extract the spatial context. Semantic segmentation is a well-developed computer vision task that generates semantic label masks by pixel-level classification. The model should consider the fine details and general context for successful segmentation. Accordingly, existing semantic segmentation models specialize in picking out detailed spatial features. Even though the definition of semantic segmentation does not match our goal, since we define our active bleeding framework as a temporal localization task, we use the segmentation model as a spatial context provider. In other words, a semantic segmentation network is embedded in the proposed fusion model as a detailed spatial feature extractor that

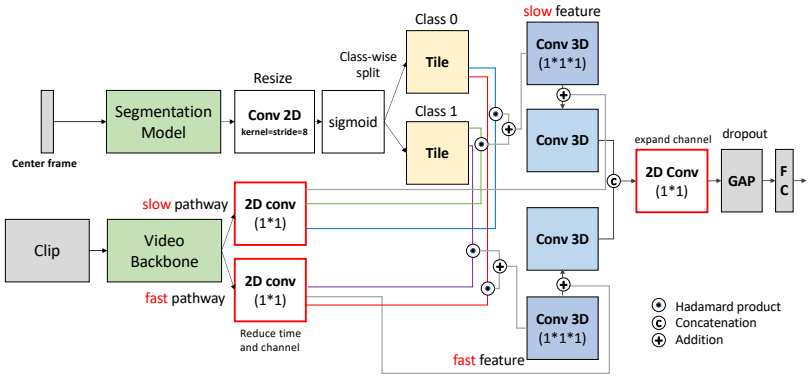


Figure 3. Our active bleeding recognition model AMAGI which is composed of semantic segmentation, action recognition and feature fusion part.

indicates *where the active bleeding occurred in a scene.* In a given frame, we extract features from the last layer composed of two soft mask information, active bleeding, and background. These spatial features are used in the fusion part of the network, which is described in the next section.

Spatial features are informative; however, mask annotations on surgical scene semantic segmentation are highly demanding. We pretrain the segmentation model on a relatively small dataset and freeze the segmentation branch while training the fusion model. Then the segmentation model only uses inference to generate spatial features, and we utilize these features as a pseudo label. For implementation, we use one of the representative semantic segmentation models, HRNet+OCR (Yuan et al. 2020).

3.3. Fusion Model

In this subsection, we will explain our suggesting fusion network which utilizes the previously mentioned temporal and spatial context specialized branches. As we previously mentioned, the temporal context model SlowFast (Feichtenhofer et al. 2019) contains slow and fast pathway features, and the spatial context model OCR is able to highlight the area where the bleeding is likely to occur. In our fusion model, the spatial feature is used for emphasizing each slow and fast features. As shown in Figure 3, we will describe in the next paragraph how the fusion layers combine spatiotemporal information.

Given input clip $x = x_0 ... x_{t-1}$, we extract spatial and temporal context with each backbone. We notate slow and fast pathway features from the SlowFast as *slow* and *fast* which contain the temporal context of active bleeding in robotic subtotal gastrectomy. The spatial context features seg_{out} are extracted from the center frame of the clip by using the pre-trained semantic segmentation model. The pre-trained segmentation model provides the soft label to temporal context features which suggest where should be focused for recognizing active bleeding. We resize and normalize the segmentation output seg_{out} through 2D convolution with the kernel size and stride as 8 following sigmoid as activation function. The feature is split along the channel dimension to attain the class-wise spatial information: each class is responsible for the background and active bleeding area.

$$m_0, m_1 = Split(\sigma(Conv2D(seg_{out}))) \tag{1}$$

Before fusing spatial and temporal context, a linear transform is applied to temporal context features and notate them as $slow'$ and $fast'$. We use 3d convolution with 1 spatial kernel, and time step and output channel dimension of convolution are properly controlled by model hyperparameters $channel\_reduction\_rate$ and $time\_reduction\_rate$. The class-wise spatial features in equation (1) are tiled to overlap to temporal context features $slow'$ and $fast'$.

$$m'_0, m'_1 = Tile(m_0), Tile(m_1) \tag{2}$$

There are several steps to fuse the spatial and temporal context information. First, we calculate features in class-wise and tempo-wise ways. We compute every 4 possible combinations (a slow or fast, and active bleeding area or not) by Hadamard product between temporal and spatial features. These features carefully observe spatiotemporal information from the background and active bleeding regions separately. Second, we combine divided features with the same tempo. For example, we add the background-highlighted fast pathway feature to the active bleeding highlighted one. Then, the class-aggregated feature of the fast pathway passes through the 3D convolution with a 1*1*1 kernel and the same output channel dimension. The same procedure is applied to process slow embedding. To obtain $embed_{fast}$ and $embed_{slow}$, we have the residual connection on each pathway feature.

$$embed_{fast} = Conv3D((fast' \odot m'_0) + (fast' \odot m'_1)) + fast' \tag{3}$$

Then $embed_{fast}$ and $embed_{slow}$ are concatenated to get the fusion feature embedding $embed_{fuse}$. Before combining, we flatten the temporal dimension of features to be the same size, by 3D convolution with the 1*1 spatial kernel, calculated temporal kernel and output channel size.

$$embed_{fuse} = Concatenate(Conv3D(embed_{slow}), Conv3D(embed_{fast})) \tag{4}$$

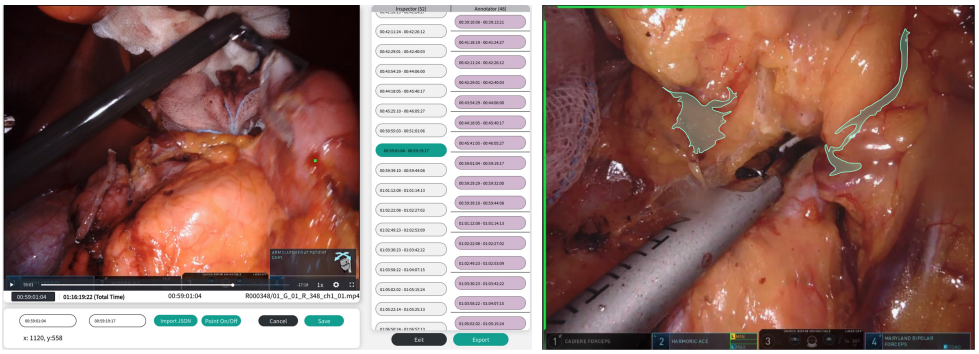
Before flattening the spatial dimension, the feature $embed_{fuse}$ is processed with 1*1 2D convolution. Then we use global average pooling to make the feature flat. Last, dropout and fully connected layers are applied to feature for the final decision.

$$out = FC(Dropout(GAP(Conv2D(embed_{fuse})))) \tag{5}$$

Output logits indicate whether the center frame of the input clip contains active bleeding or not. To summarize, our active bleeding recognition model AMAGI has three components: spatial, temporal context backbones, and the feature fusion part. The fusion layer manipulates semantics of slow and fast, bleed/non-bleed region as we described in this paragraph, and helps make a decision well. The entire network is trained in an end-to-end manner except for the semantic segmentation branch.

3.4. Post-processing and Surgical Analysis Index

The ultimate goal of our active bleeding framework is to provide AI-predicted surgical measures that represent implicit information about bleeding events during the entire surgery. To this end, we have two more procedures in our framework: post-processing and surgical index measuring.



(a) Our annotation tool for TAL annotation (b) CVAT for spatial localization annotation

Figure 4. TAL and spatial localization labeling examples for iAB annotation.

In the post-processing step, we use traditional noise removal algorithms in the computer vision area. The proposed active bleeding model predicts the entire surgical video in the evaluation stage. However, the prediction results tend to be noisy due to the limited input size of the clip-by-clip prediction model. We introduce the noise removal steps based on classical computer vision operations to alleviate the issue. We apply a median filter, opening and closing morphological operations sequentially with kernel sizes of 11, 30, and 10 each. The kernel size means the window size for each operation to be applied on per-frame predicted outputs.

In addition, we use two surgical analysis indexes to extract statistics from the entire surgery video. The bleeding count indicates the number of consecutive bleeding events in our temporal localization output. Bleeding duration is the time that active bleeding occurs in entire surgery videos. Post-processing should be applied before measuring surgical indices. Figure 2 shows the surgical analysis index and how the inference output changes before and after post-processing.

4. Experiments

We experimented with various settings to verify our active bleeding framework. First, we show a performance comparison of models that are in spatial context only, temporal context models, and our fusion model AMAGI in Table 1. We applied the same post-processing for all models. The results indicate our proposed model consistently outperforms state-of-the-art existing algorithms. Moreover, we analyze the results with binary classification metrics and a surgical analysis index. We also provide a detailed analysis of model robustness with the area under the curves (AUC) of the ROC and precision-recall curves. Lastly, we have Grad-CAM visualizations to prove that our model is successfully trained to concentrate on areas where bleeding is likely to occur.

Dataset. We generated a dataset using 40 robotic subtotal gastrectomies for gastric cancer to train AMAGI models for iAB detection. The baseline, initiation, and end points for iAB were defined in discussion with three surgeons. In our data generation process, the spatial and temporal bleeding localization datasets are all defined and annotated by surgical experts. Each surgical video was then annotated with the start and end points of iAB by each of the trained annotators. With two annotations, one specialist confirms the TAL label of the final merged iAB. For spatial annotation on iAB, one frame is selected randomly within one iAB TAL event. In iAB's spatial

Table 1. Performance of Active bleeding frameworks. **First Table.** We report after post-processing 3-fold case averaged scores with standard deviations. All models are trained on R-40 temporal localization dataset. **Second Table.** We also report a confusion matrix.

R-40	Evaluation Metrics of Binary Classification				surgical analysis index	
Models	Accuracy	Precision	Recall	F1 score	Count Err.[a]	Duration Err.[b]
HRNet+OCR	32.04 ± 5.58	15.17 ± 0.23	$\mathbf{95.22} \pm 3.63$	25.96 ± 0.20	28.1 ± 11.92	87.88 ± 6.54
MS-TCN	36.68 ± 49.35	6.57 ± 5.76	66.00 ± 57.17	11.77 ± 10.29	16.0 ± 3.0	28.34 ± 20.36
SlowFast-50	77.14 ± 10.76	32.37 ± 10.17	51.37 ± 18.96	36.60 ± 3.92	26.8 ± 19.77	17.00 ± 23.98
SlowFast-152	76.26 ± 3.59	29.77 ± 4.15	60.42 ± 5.51	$\mathbf{39.09} \pm 2.50$	40.8 ± 13.50	18.11 ± 7.88
AMAGI-50	82.67 ± 0.92	35.05 ± 1.09	39.96 ± 9.74	36.06 ± 4.29	$\mathbf{12.3} \pm 1.37$	5.83 ± 0.68
AMAGI-152	$\mathbf{84.99} \pm 1.05$	$\mathbf{40.18} \pm 4.25$	31.29 ± 3.89	34.04 ± 1.48	13.2 ± 1.60	$\mathbf{5.09} \pm 1.94$

[a] Blood count error indicates the 3-fold averaged MAE (Mean Absolute Error) counts.
[b] Blood duration error is the 3-fold averaged MAE (Mean Absolute Error) minutes.

Confusion Matrix	TP	FP	FN	TN
SlowFast-50	166,448	423,576	148,593	1,673,875
AMAGI-50	126,315	231,495	188,726	1,865,956
AMAGI-152	99,016	146,860	216,025	1,950,591

annotation on selected frames, one annotator creates polygons, and one specialist finalizes them to confirm polygon information. For the TAL annotation of iAB, a self-implemented video annotation tool and CVAT (Sekachev et al. 2020) was used to annotate spatial polygon information. Figure 4 shows an example of TAL annotation and spatial localization annotation for iAB.

In our experiments, we use these two types of annotations to train the active bleeding recognition model: temporal localization and spatial localization. The temporal localization dataset is a set of start and end points of active bleeding in the entire surgery recordings. The spatial localization dataset is composed of pairs of images and binary masks. The zero-one labels in the binary mask indicate whether it is an active bleeding pixel or not. These large-scaled annotations are all executed by clinical specialists. In Table 2, we divide 40 cases into 30 and 10 for each train and test set of the 3-fold cross-validation set. We have spatial localization annotations on several frames according to the cross-validation cases. In the case of temporal localization annotations, we generated clips with lengths of 240 and less than 50 for each non-bleeding and bleeding clip for training. We evaluate the models as a temporal localization task in 10 fps of all 10 test cases for each cross-validation set.

Evaluation Metrics. We have two types of evaluation metrics. We report the performance of binary classification metrics generally used for machine learning tasks. We sum up the frame-by-frame prediction and count true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). These can produce accuracy, precision, recall, and an f1 score. We also provide ROC and precision-recall curves with AUC on various thresholds for detailed explanations. Additionally, we show surgical analysis indexes that are straightforward to carry the bleeding information for clinical use. All evaluation metrics are calculated after post-processing since the process is one of the major components of our framework. This process is crucial to extracting the surgical analysis index, although post-processing slightly decreases the binary classification metrics.

Model Types and Experimental Settings. We have three different types of model for comparison: semantic segmentation only, action recognition only, and fusion-based model AMAGI. First, semantic segmentation only setting use the spatial localization dataset to train the model. For semantic segmentation, we use OCR (Yuan et al. 2020) algorithm with HRNet backbone (Wang et al. 2019) and follow the of-

Table 2. The size of 3-fold cross validation set. Our dataset of spatial localization and temporal localization.

Dataset	train_split1	train_split2	train_split3	test_split1	test_split2	test_split3
Cross-Validation Surgery Cases	30	30	30	10	10	10
Spatial Localization (images)	1315	1388	1439	532	459	408
Temporal Localization (clips)	51846	51718	51711	all video frames in 10 fps		

Table 3. Performance of pretrained semantic segmentation in spatial localization dataset. We report best epoch for validation set.

IoU	split1	split2	split3	3-fold avg.
Background	98.70	98.90	99.00	98.87 ± 0.15
Bleeding	41.51	46.92	49.16	45.86 ± 3.93
Mean	70.10	72.91	74.07	$\mathbf{72.36} \pm 2.04$

ficial github implementation. Our configuration for OCR is the batch size 32, SGD optimizer, learning rate 0.01, weight decay 5e-4, 300 epochs, and resize image into 512×512. We choose the best model at validation epoch about mean IoU. Second, action recognition only setting use the temporal localization dataset for training. We experimented SlowFast (Feichtenhofer et al. 2019) and MS-TCN (Li et al. 2020) for action recognition. Third, the fusion-based model takes both spatial and temporal localization annotations. We implemented the SlowFast (Feichtenhofer et al. 2019) and fusion model on mmaction2 (Contributors 2020) environment of open-mmlab. We pre-trained the segmentation model with spatial localization, then port this code to mmaction2 (Contributors 2020) to extract the spatial context features as pseudo labels. We use exactly the same learned parameters with segmentation only setting, and these parameters are not updated during training the fusion model. We only use temporal localization annotation when training the fusion model, which is the same procedure as SlowFast setting. Our configurations for SlowFast and AMAGI are clip length 8, frame interval 3, speed ratio 2, channel ratio 8, 2 GPUs, batch size 64 per GPU, optimizer SGD with learning rate 0.01, momentum 0.9, weight decay 4e-5, cosine annealing scheduler with linear warm up, 100 epochs, choose last epoch for testing, and resize image into 224×224. We loaded the pretrained SlowFast parameters in Kinetics-400. For the fusion part, time reduction rate and channel reduction rates are set to 2.

In the case of MS-TCN, we use the official MS-TCN Github for implementation. We use mean squared error and cross-entropy losses to train 100 epochs. 2D CNN features are extracted from ResNet. Although the model is designed to see the entire video, it was hard to converge the loss due to the class imbalance and long sequence. So we split the input video into 10 segments and use 1 video per iteration, and the batch size was set to 10. We use 2 stages and 3 layers since it performs better than the larger one.

4.1. Overall Performance

In Table 1, we report the overall performance of the active bleeding task. We report case-averaged 3-fold cross-validation results to normalize the length of each surgery case. First, our fusion model AMAGI consistently outperforms those of other models. AMAGIs have the best accuracy and precision compared to other models, including the previous active bleeding work (Wei et al. 2021). The best recall and F1 models, OCR and SlowFast, have a low score in respect of accuracy and precision. We focused

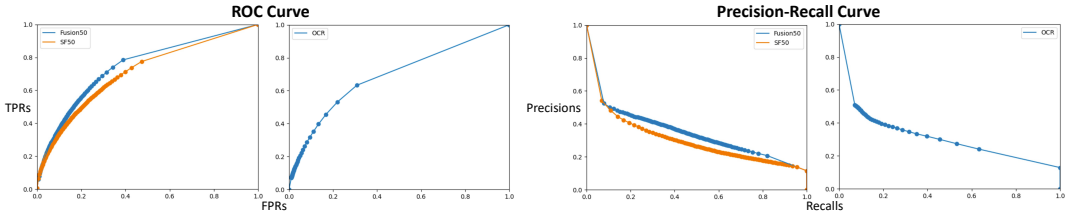


Figure 5. First Row. There are ROC and precision-recall curves of fusion model AMAGI and temporal context model SlowFast. We give 0.01 as threshold steps to draw the graph. **Second Row.** There are ROC and precision-recall curves of spatial context model OCR. We use 2000 for the pixel threshold steps to draw the graph. Both two curves are important to observe the trade-off of the active bleeding prediction.

Table 4. Performance comparison respect to ROC AUC and Precision-Recall AUC. We averaged the performance of validation cases in 3-fold set.

	ROC AUC		Precision-Recall AUC	
threshold step	SlowFast-50	AMAGI-50	SlowFast-50	AMAGI-50
0.05	0.69398	**0.71937**	**0.34852**	0.34651
0.01	0.70745	**0.74363**	0.33728	**0.33933**
0.001	0.71767	**0.75998**	0.33265	**0.33658**
threshold step	HRNet+OCR		HRNet+OCR	
2000	0.68839		0.31409	
4000	0.66917		0.31455	

on precision among the binary classification metrics because we aim to measure the amount of active bleeding accurately. Figure 6 shows the importance of precision. At this point, the fusion model AMAGIs generally achieve better than other settings. Our models also beat others on the surgical analysis index with small standard deviations. Besides, we find out that the MS-TCN is hard to converge in our case since the sequence is too long and a small portion of the bleeding periods are negligible. Then the model tends to be biased toward one class to converge the loss function. We also experimented on different model depths $\{50, 152\}$, and the results in Table 1 confirmed that our model consistently defeats others. In the second table of Table 1, we report the confusion matrix. The result indicates that a ratio of true and false positives is improved using AMAGI. Also, true negative is relatively enhanced much considering the trade-offs. Since it is difficult to judge due to the trade-off in binary classification measures, a more detailed analysis will be provided in the next subsection in relation to model robustness and bleeding thresholds.

Additionally, we present the performance of pre-trained segmentation used by our fusion model AMAGI in Table 3. Segmentation performance must be good, which can be evaluated with our limited number of spatial localization datasets to give a better guide. As a result, the mean IoU of 3-fold cross-validation is about 72. This segmentation model is the same as HRNet+OCR in Table 1, and we loaded these pre-trained model parameters into the fusion model AMAGI. When these pre-trained model parameters were utilized in the fusion model, the fusion models showed excellent performance in Table 1.

4.2. Model Robustness

Although our fusion model AMAGI achieves great performance, we confirmed that there is a trade-off between binary classification metrics in the previous subsection.

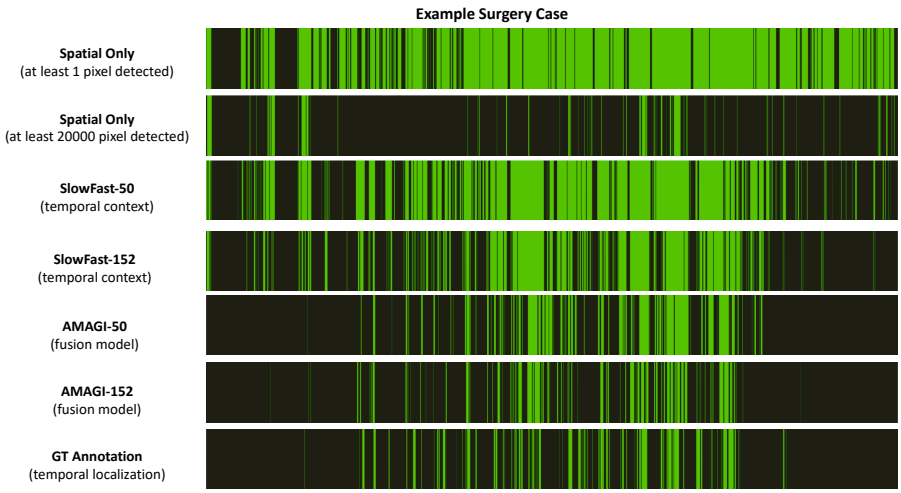


Figure 6. Model prediction comparisons of active bleeding framework which we defined as the temporal localization task. Example surgery case is from the validation set.

Thus, we reflected on the trade-offs using two diagnostic tools, ROC and Precision-Recall curves, and provided a rich analysis of our active bleeding framework. Briefly saying, ROC (receiver operating characteristic) curve has two parameters to plot, TPR (true positive rate) and FPR (false positive rate) in different thresholds. Similarly, the precision-recall curve needs precision and recall to plot the trade-off of model robustness. We control the robustness of the bleeding prediction model by changing the decision threshold of temporal localization results. Both two curves are important to interpret the trade-off of the model prediction.

In the first row of Figure 5, we draw two types of curves for AMAGI and SlowFast. We averaged the same threshold results of the test cases in 3-fold cross-validation. We use a threshold step of 0.01 from zero to one for the plot. Since precision tends to have a low value in the graphs, it seems that the active bleeding task has difficulty inherently. Nevertheless, our model has superior performance to the existing model, SlowFast. Likewise, the second row of Figure 5 indicates the graphs of the segmentation model OCR. Previously, we judged if the frame has any active bleeding classified pixels in the default setting, but we verified the various threshold steps in the range of 0 to 70000. We use threshold steps of 2000 for plotting. The ROC and PR curves of the segmentation model are located lower than the curves of the models in the first row. To summarize, our fusion model AMAGI has superior performance compared to SlowFast and OCR.

Table 4 supports our argument as a basis. We calculate the AUCs (area under the curve) of ROC and PR curves to compare which model has better performance in reflecting the trade-offs. Our performance excellence is more pronounced when we give more finely-detailed steps on thresholds.

4.3. Visualizations

We have two visualizations to demonstrate the significance of our fusion model. First, visualization of temporal localization in Figure 6 indicates active bleeding prediction results of an example surgery case. Our fusion-based model AMAGI best matches the annotation of temporal localization discussed in the surgical analysis index of Table 1.

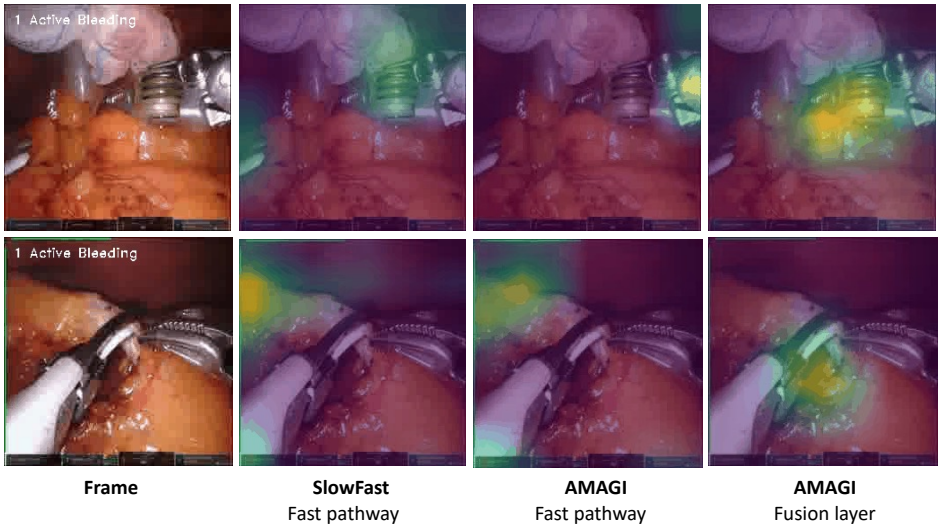


Figure 7. GradCAM Visualization.

Additionally, we have a visualization for understanding how the deep learning model makes a decision. In Figure 7, we compare the gradient-weighted class activation mapping (Grad-CAM) visualization of AMAGI and SlowFast. Grad-CAM is a visual explanation algorithm to interpret the trained deep learning model. This algorithm is applied to an already-trained model and produces a heatmap of the important regions for decisions in a given clip input. We visualize two types of deep neural network layers: the fast pathway of SlowFast and the fusion layer of AMAGI. The fast pathway is a part of the video backbone, and both our fusion and temporal context-only models (SlowFast only setting) have this structure identically. Figure 5 has two example cases in the validation set. The second and third columns display the GradCAM heatmap of the last layer of the fast pathway. The fourth column is the visualization of the fusion layer that our model AMAGI contains only. Overall, heatmaps of the fusion layer concentrate on possibly-bleeding areas of the organ. We can interpret that the fusion layer is much more aware of the bleeding area than the SlowFast layer because the segmentation branch highlights important regions as we intended.

5. Discussion and Conclusion

Detecting active bleeding among iAEs is essential in establishing a CAS environment related to image-guided surgery, postoperative surgical analysis, and patient outcome prediction. More focusing on postoperative surgical analysis, we proposed an AMAGI model that appropriately fused spatiotemporal features in the training and inference stages and evaluated it from the entire surgical video analysis perspective. The AMAGI model was evaluated in terms of TAL and achieved improved performance in terms of recognition accuracy and error of surgical index. We expect our model to help analyze the surgical videos in the automated system. Nevertheless, the feature fusion for the AMAGI model still has potential for improvement, and further analysis of the correlation between training data quantity and recognition performance is needed. We expect that from the training methodology of the proposed AMAGI model, other approaches that consider effective spatiotemporal supervision for iAB detection at the

same time will be extended.

Acknowledgements.

This research was funded by the Ministry of Health Welfare, Republic of Korea (grant number : 1465035498 / HI21C1753000022).

References

Aoyama T, KANO K, Numata M, ATSUMI Y, HARA K, Kazama K, KOUMORI K, Murakawa M, HASHIMOTO I, Maezawa Y, et al. 2020. The impact of intraoperative blood loss on the long-term prognosis after curative resection for borrmann type iv gastric cancer: A retrospective multicenter study. Anticancer Research. 40:405–412.

Bohnen J, Mavros M, Ramly E, Chang Y, Yeh D, Lee J, Moya M, King D, Fagenholz P, Butler K, et al. 2016. Intraoperative adverse events in abdominal surgery: What happens in the operating room does not stay in the operating room. Annals of surgery. 265.

Bonrath E, Dedy N, Zevin B, Grantcharov T. 2013. Defining technical errors in laparoscopic surgery: A systematic review. Surgical endoscopy. 27.

Carreira J, Zisserman A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 4724–4733.

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence. 40(4):834–848.

Contributors M. 2020. Openmmlab's next generation video understanding toolbox and benchmark; [https://github.com/open-mmlab/mmaction2].

Feichtenhofer C. 2020. X3d: Expanding architectures for efficient video recognition. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR):200–210.

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

Fu Y, Zhang W, Mandal M, Meng MQH. 2014. Computer-aided bleeding detection in wce video. IEEE Journal of Biomedical and Health Informatics. 18(2):636–642.

Garcia-Martinez A, Vicente Samper J, Sabater-Navarro J. 2017. Automatic detection of surgical haemorrhage using computer vision. Artificial Intelligence in Medicine. 78.

Hong SG, Choi MK. 2021. Blockwise temporal-spatial pathway network. In: 2021 IEEE International Conference on Image Processing (ICIP). p. 3677–3681.

Hong WY, Kao CL, Kuo YH, Wang JR, Chang WL, Shih CS. 2020. Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. ArXiv. abs/2012.12453.

Hua S, Gao J, Wang Z, Yeerkenbieke P, Li J, Wang J, He G, Jiang J, Lu Y, Yu Q, et al. 2022. Automatic bleeding detection in laparoscopic surgery based on a faster region-based convolutional neural network. Annals of Translational Medicine. 10(10). Available from: https://atm.amegroups.com/article/view/95011.

Jo K, Choi B, Choi S, Moon Y, Choi J. 2016. 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). p. 1260–1263.

Li S, Zhang J, Ruan C, Zhang Y. 2019. Multi-stage attention-unet for wireless capsule endoscopy image bleeding area segmentation. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). p. 818–825.

Li SJ, AbuFarha Y, Liu Y, Cheng MM, Gall J. 2020. Ms-tcn++: qu-stage temporal con-

volutional network for action segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence:1–1.

Lin J, Gan C, Han S. 2019. Tsm: Temporal shift module for efficient video understanding. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). p. 7082–7092.

Lin T, Goyal P, Girshick R, He K, Dollar P. 2020. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 42(02):318–327.

Numata M, Aoyama T, Kazama K, ATSUMI Y, IGUCHI K, SAWAZAKI S, SATO S, KANO K, Oshima T, YAMADA T, et al. 2021. Impact of intraoperative blood loss on the survival of patients with stage ii/iii colorectal cancer: A multicenter retrospective study. In Vivo. 35:3483–3488.

Oh SW, Lee JY, Xu N, Kim SJ. 2019. Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. p. 9226–9235.

Okamoto T, Ohnishi T, Kawahira H, Dergachyava O, Jannin P, Haneishi H. 2019. Real-time identification of blood regions for hemostasis support in laparoscopic surgery. Signal, Image and Video Processing. 13.

Rabbani N. 2022. Video-based computer-aided laparoscopic bleeding management: a space-time memory neural network with positional encoding and adversarial domain adaptation. In: MIDL.

Sekachev B, Manovich N, Zhiltsov M, Zhavoronkov A, Kalinin D, Hoff B, TOsmanov, Kruchinin D, Zankevich A, DmitriySidnev, et al. 2020. opencv/cvat: v1.1.0; Aug. Available from: https://doi.org/10.5281/zenodo.4009388.

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

Usman MA, Satrya G, Usman MR, Shin S. 2016. Detection of small colon bleeding in wireless capsule endoscopy videos. Computerized Medical Imaging and Graphics. 54:16–26.

Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X, et al. 2019. Deep high-resolution representation learning for visual recognition. TPAMI.

Wei H, Rudzicz F, Fleet D, Grantcharov T, Taati B. 2021. Intraoperative adverse event detection in laparoscopic surgery: Stabilized multi-stage temporal convolutional network with focal-uncertainty loss. In: Jung K, Yeung S, Sendak M, Sjoding M, Ranganath R, editors. Proceedings of the 6th Machine Learning for Healthcare Conference; (Proceedings of Machine Learning Research; vol. 149); 06–07 Aug. PMLR. p. 283–307.

Xiao T, Liu Y, Zhou B, Jiang Y, Sun J. 2018. Unified perceptual parsing for scene understanding. In: ECCV. p. 432–448.

Yuan Y, Chen X, Wang J. 2020. Object-contextual representations for semantic segmentation. In: ECCV. p. 173–190.

Zegers M, de Bruijne M, Keizer B, Merten H, Groenewegen P, van der Wal G, Wagner C. 2011. The incidence, root-causes, and outcomes of adverse events in surgical units: Implication for potential prevention strategies. Patient safety in surgery. 5:13.