IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 May 2024, accepted 6 June 2024, date of publication 11 June 2024, date of current version 19 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3412780

RESEARCH ARTICLE

Real-Time Surgical Instrument Segmentation Analysis Using YOLOv8 With ByteTrack for Laparoscopic Surgery

NYI NYI MYO[1], APIWAT BOONKONG[2], KOVIT KHAMPITAK[3], AND DARANEE HORMDEE[1], (Member, IEEE)

[1]Department of Computer Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand
[2]Department of Computer Engineering, Faculty of Engineering, Nakhon Phanom University, Nakhon Phanom 48000, Thailand
[3]Department of Obstetrics and Gynecology, Faculty of Medicine, Khon Kaen University, Khon Kaen 40002, Thailand

Corresponding author: Daranee Hormdee (darhor@kku.ac.th)

This research was supported in part by the Khon Kaen University GMS (Greater Mekong Subregion) Master Scholarship, Thailand.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Center for Ethics in Human Research, Khon Kaen University (KKU), under Application No. #HE641206, and performed in line with the declaration of ethical approval for human research at Srinagarind Hospital, KKU, Thailand.

ABSTRACT As Computer Vision technology has evolved rapidly these days, the implementation of object detection and instance segmentation has been presented in various areas. In computer-aided laparoscopic surgery, the segmentation of surgical instruments is one of the active research areas. This paper presents the implementation and comparative analysis of a real-time surgical instruments segmentation system by incorporating ByteTrack, a powerful object tracking system, within the YOLOv8, a state-of-the-art Deep Learning algorithm for object detection and segmentation, together with an instruments gesture analysis of the practical results. The instrument gestures have been categorized into separating, crossing, and overlapping cases according to the most common instrument gestures during the surgery. The datasets from the ROBUST-MIS 2019 challenge have been applied and annotated for training, validating, and blind testing in this study. Considering trade-offs among model complexity, speed, and accuracy, the medium version (YOLOv8m) has been chosen in this study for its comparative model complexity, design for working in real-time, and relative high accuracy. In order to validate the effectiveness of this research, real-time segmentation of surgical instruments has been performed with the streaming of laparoscopic gynecologic surgery on 5 donated soft-tissue cadaver cases. According to the experimental results, although YOLOv8 can provide very high-accuracy evaluation metrics for both **F1-score** and **mAP** (mean Average Precision), the segmentation accuracy results could have been further improved by incorporating the ByteTrack within the YOLOv8 algorithm. Owing to the 2-association scheme that has been designed for object tracking in ByteTrack, referring to the tracklet from the previous frame could recover missed segmentations that come with too low confidence values. The findings identify that the Modified model of incorporating ByteTrack with YOLOv8 could improve the **F1-score** from 0.89 to 0.92, which outperformed all of the previous studies on the ROBUST-MIS 2019 Challenge, and from 0.82 to 0.88 on the blinded captured dataset from live streaming videos with a real-time segmentation speed of approximately 45 **FPS** (Frames Per Second), which is sufficient for a real-time application as opposed to 60 **FPS** from only the YOLOv8 algorithm. From the instrument gestures result analysis, ByteTrack could improve the segmentation performance in all gesture categories: separating, crossing, and overlapping. However, the remaining segmentation failures mostly lie in crossing and overlapping gestures.

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

INDEX TERMS Real-time instance segmentation, deep learning, surgical instrument, laparoscopic surgery, YOLOv8, ByteTrack.

I. INTRODUCTION

As opposed to conventional open-surgery, laparoscopy has the advantages of being less painful, a use of smaller incisions and more rapid recovery time for the patient. Robotic-Aided Surgery (RAS) has been implemented to mitigate the long-term muscle fatigue due to holding the laparoscope for a long period from the start of the surgery to finish [1], causing unsteady and inaccurate screen monitoring, which could interrupt the surgeon, leading to taking longer time and even putting the patient in danger [2]. Since 1995, the da Vinci System [3], which is the most well-known of RAS systems, has revolutionized Minimally Invasive Surgery (MIS). At the current time, innumerable approaches towards Image-Guided Surgery (IGS) [4] have undergone development within the realm of Computer Vision. As for RAS, organ and surgical instrument detection, segmentation, and tracking have been useful not just during the actual surgery but also for surgical training to improve surgical skills. However, unlike other images containing a wide RGB range, the viewpoint of these images, streaming from the laparoscope, is rather unique, as their color range is quite limited, mostly in a red-brown tone. Furthermore, objects within these images would appear bigger and closer positioned than the actual objects, resulting in what seems to be more rapid motion, which might move across the whole frame in a split second. A large number of studies have been undertaken on surgical instrument detection and tracking [5], along with a prior study [6] that compared a number of Deep Learning methodologies. Recently, various research [7], [8], [9], [10], [11], [12], [13] has progressed towards surgical instrument segmentation as it provides more fine-grained boundaries and regions rather than only identifying specific objects along with their locations [14]. Although the majority (if not all) of previous research offered outstanding performance, the major challenge still remains with close-positioned and overlapping surgical instruments. The purpose of this study is not only to present surgical instrument segmentation performance via the combination of the Deep Learning algorithm, YOLOv8 along with the powerful ByteTrack, but also to perform an analysis on how this model deals with designated objects in this unique viewpoint of laparoscopic surgery. Furthermore, in addition to performing on the dataset, the models have also been experimented on 5 soft-tissue cadaver surgeries to verify their real-time performance.

II. RELATED WORK

Within the realm of Computer Vision combined with Deep Learning, the instance segmentation technique is a significant improvement in image analysis. The segmentation algorithms such as Mask RCNN [15], EfficientNet [16], YOLACT++ [17], and Mask SSD [18] have been developed in recent years. Some prominent areas of instance segmentation models are autonomous driving cars, satellite imaging and agriculture, robotics, and also medical imaging [19]. Although the instance segmentation technique can provide more detailed information about an object's location and even occlusion handling, there are still challenges for it to be real-time segmentation with high speed and reliable accuracy [20].

For surgical instruments segmentation in laparoscopic surgery, Mask RCNN algorithm has been applied with various datasets [7], [8], [9], [10]. Although Mask RCNN can provide reliable accuracy, its inference speed is quite low for real-time applications and it also has failure cases of crossing, overlapping, and occlusion [9], [10].

As for using the EfficientNet network, surgical instruments segmentation has been implemented with different datasets [11], [12], [13]. Although EfficientNet can provide 20 **FPS** speed, its results yield slightly low accuracy, with problems in poor lighting situations and missed segmentations due to overlapping [12], [13].

In 2019, the ROBUST-MIS (Medical Instrument Segmentation) 2019 Challenge [21] was organized as an international benchmarking competition, aiming to find and compare the algorithms with robustness and generalization capabilities. While the top-3 ranked Deep Learning models, DeepLabV3+ [22], OR-UNet [23], and BARNet, or Dense Pyramid Attention Network [24], with the highest accuracy, have been reported [25] for being robust to color changes due to reflections, blur, blood, and smoke, and YOLACT++ [26] has been reported for providing 30 **FPS** speed, there are still critical challenges, including occluded conditions such as close-positioned, overlapping, and crossing of surgical instruments.

Nowadays, YOLO has become the optimal real-time Deep Learning algorithm for object detection, instance segmentation, and identification. Since YOLACT++, later YOLO versions, such as YOLOv5, YOLOv7, and YOLOv8, have been able to provide instance segmentation to train the custom model [27]. This instance segmentation is a modification of the detection architecture with an additional neural network in the head to output the segmentation masks.

In order to perform real-time object segmentation in video sequences, the most current, with high tracking accuracy, object tracking algorithms, StrongSORT [28], OC-SORT [29], and ByteTrack [30], have been considered. With the help of Kalman filter, while StrongSORT and OC-SORT have been designed with respect to SORT (Simple Online and Real-time Tracking), ByteTrack has been implemented, considering speed as a critical factor. Unlike other tracking algorithms, which only associate bounding boxes with higher scores than the confidence threshold and ignore the lower scores completely, ByteTrack will filter both higher and lower scores to work on motion or appearance similarity in order to recover possible occlusions between consecutive frames. This association with both high and low scores not only benefits object tracking but could also be beneficial to

object occlusion. An occluded object might easily be able to get detected; therefore, later, once it is occluded, it could still be detected via the association scheme.

While a number of studies have worked on multi-class surgical instrument detection/segmentation [5], [10], [13], only a few have reported on real-time segmentation [26] and none on detailed analysis, i.e., instrument gestures. To perform real-time surgical instrument segmentation in video sequences, ByteTrack has been considered to be incorporated with the current latest YOLOv8 algorithm, similarly to other previous works [31], [32]. However, some issues need to be addressed (i.e., objects appear to be closer, bigger, and move faster) in order to handle the specific scenarios in the laparoscopic surgery case to form the Modified Y+BT model here.

The contributions of this research are as follows:

- The Modified Y+BT model, which has been evaluated on the ROBUST-MIS 2019 Challenge dataset, over the state-of-the-art YOLOv8 algorithm and also the original Y+BT model. The results have also been compared with those from the ROBUST-MIS 2019 Challenge.
- The real-time results from the experiments and evaluation of the Modified Y+BT model on 5 soft-tissue cadaver cases for live-streaming laparoscopic gynecologic surgery over the state-of-the-art YOLOv8 algorithm and also the original Y+BT model.
- The analysis of the practical results according to instrument gesture categories.
- A new dataset, captured images from 5 soft-tissue cadaver cases used in this research.

III. METHODOLOGY

This section commences with the Modified Y+BT model, then how the datasets have been obtained, followed by the details of the experimental process, real-time process, and evaluation metrics. The gesture categories for later analysis are described at the end.

A. THE MODIFIED Y+BT ALGORITHM

The Modified Y+BT algorithm consists of YOLOv8 as the core segmentation and ByteTrack as an adapted module, as shown in Figure 1. Two confidence thresholds (high and low) have to be defined for ByteTrack in order to take advantage of having 2 associations between the current frame and the previous frame, unlike other segmentation algorithms, including YOLOv8, which concern only a specific confidence threshold. Any instance that might get segmented with a low confidence value (as indicated in the red circles) by the YOLOv8 algorithm may be recovered by ByteTrack later. In this study, the optimal instance segmentation model from the training process, together with this optimal confidence threshold, was applied in the ByteTrack algorithm.

Figure 2 illustrates how these 2 confidence thresholds take part in both the first and second associations of ByteTrack. Per the first association, the high confidence threshold serves as the typical confidence threshold of other segmentation algorithms to identify whether an instance is segmented. While in ByteTrack, this segmented instance within high confidence values in the current frame might or might not have a tracked instance (aka tracklet) from the previous frames associated, depending on whether the location of the segmented instance falls down in the predicted location of any tracklet from the previous frame. If so, the tracklet is updated, but if not, a new tracklet is introduced in this frame and held up for a certain time (configured to the capture rate in this study) before faded off. The next step, an exclusive feature in ByteTrack, requires another (low) confidence threshold. Similarly to the first association, this time the association is between a tracklet from the previous frame and an instance in the current frame within low confidence values at the predicted location. Two cases could occur, where the second association only concerns when a tracklet is associated with an instance, promoting what used to be an unsegmented instance due to its low confidence value to become a segmented instance, and the tracklet is updated. Where any other instances with low confidence values do not have any tracklet associated, those instances will then be ignored. Lastly, for those instances with confidence values lower than the low confidence threshold, these instances will be neglected as they might have used to be there but not anymore, or it is a false alarm as they have never been there in the first place.

The main distinction between the Modified Y+BT algorithm used in this study and other YOLOv8 + ByteTrack algorithms is its attempt to make it more reliable for handling the dataset's unique characteristic of laparoscopic surgery. As mentioned in the paper, the dataset from laparoscopic surgery has a uniqueness over other scenarios because objects are in a more closed-up position, hence the objects appear bigger and in more rapid motion. As a result, according to the motion uncertainty of an instance in this scenario, two hyperparameters within the Kalman Filter, position and velocity weights, must be modified to support the size and speed of the designated objects here.

B. DATASETS

As mentioned in Section II, ROBUST-MIS 2019 Challenge datasets [33], which have been considered as the first-occurring large-scale annotated MIS dataset, have been applied, trained, and tested. A total dataset of 10,040 captured images (from 25 **FPS** videos) was used, each with a resolution of 960 × 540 pixels, from 30 minimally invasive surgical procedures from 3 laparoscopic surgery types: Proctocolectomy, Rectal Resection, and Sigmoid Resection (named **C1** to **C10** datasets). Independent testing sets, divided into 3 different stages, have been included in the dataset in order to evaluate the model with different levels of difficulty. All of the captured images were then annotated through the Roboflow [34] online annotation tool by using the polygon annotation type. After annotation of

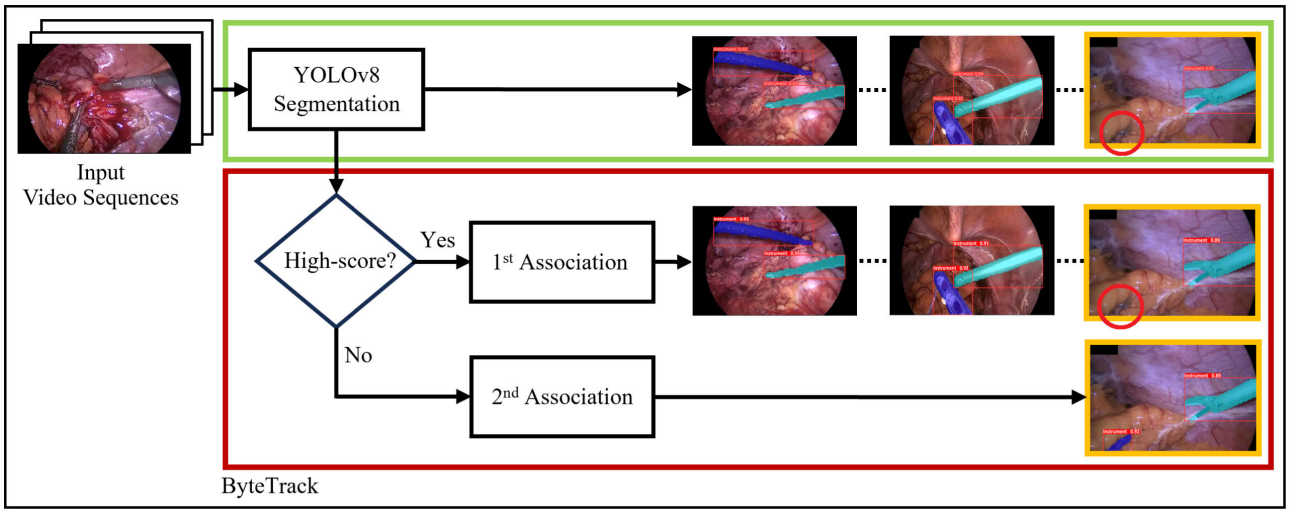


FIGURE 1. Overview diagram of the modified algorithm.

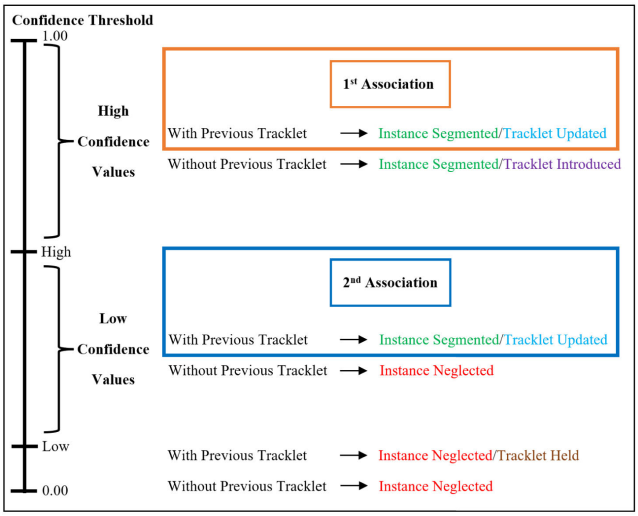


FIGURE 2. Association procedure.

all captured images, augmentation is necessary in order to enlarge the size of training and testing datasets for improving model accuracy and preventing models from overfitting [35]. Additionally, YOLOv8 provides some data augmentations, such as blurring, median blurring, gray scaling, and contrasting, from the Albumentations package, but only for the training dataset [27]; nonetheless, image augmentation techniques in the geometric transformation aspect were not applied. Therefore, further augmentations, including horizontal flipping and ±15° random rotation via Roboflow, have also been applied to all the datasets.

Moreover, per real-time evaluation, the models have been tested in laparoscopic gynecologic surgery undertakings on 5 donated soft-tissue cadaver cases for the purpose of educational research at Srinagarind Hospital, KKU, Thailand, under the Ethical approval #HE641206, waived by the Center for Ethics in Human Research, Khon Kaen University (KKU), on May 13th, 2021. Later, the Streaming dataset comprising 262 annotated images with a resolution of 1,920 × 1,080 pixels acquired from these 5 recorded surgery clips was created and applied as additional testing datasets, named **S1** to **S5**, serving as one of the contributions here and which is available at "https://www.kaggle.com/datasets/nyinyimyo2022/streaming-datasets." Table 1 lists the details of the training-validation (80:20) and testing datasets from the ROBUST-MIS 2019 Challenge and the Streaming dataset with the number of captured images used in this study.

TABLE 1. Datasets details.

Procedure	Train (80%)	Validate (20%)	Test			
			Stage 1	Stage 2	Stage 3	Stream
Proctocolectomy	2,355	588	325	255	-	-
Rectal Resection	2,432	608	338	289	-	-
Sigmoid Resection	-	-	-	-	2,880	-
Gynecology	-	-	-	-	-	262
Total	4,787	1,196	663	514	2,880	262

C. EXPERIMENTAL PROCESS

Two processes contained in this study are the computational process and the real-time process. Figure 3 illustrates the overview diagram for the computational process, comprising model training, validation, and testing for both YOLOv8 and the Modified Y+BT surgical instrument segmentation models. The dataset includes 10,040 captured images from the ROBUST-MIS 2019 Challenge as training-validation-testing datasets and 262 captured images from the streaming gynecologic surgery. With the help of Roboflow, the number of annotated and augmented images has increased to 14,361 images for training, 3,588 for validating, and 12,957 for testing. Furthermore, the number of images within the training dataset has been increased by 57,444 by YOLOv8 albumentation.

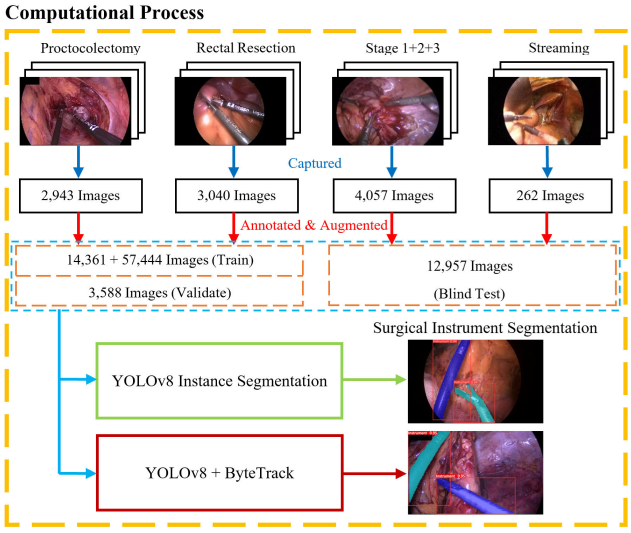


FIGURE 3. Computational process.

Per the real-time streaming process (Figure 4), the trained YOLOv8 along with the Modified Y+BT instance segmentation models have been performed to verify the segmentation speed. The captured images from these live streamings of the laparoscope have been sent back to the computational process once again to evaluate the accuracy, as shown in Figure 3.

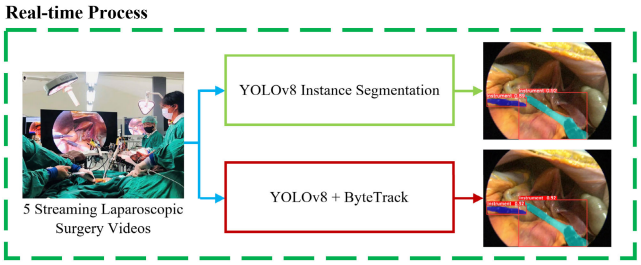


FIGURE 4. Real-time process.

YOLOv8 instance segmentation algorithm has been applied to train the custom model as it is the state-of-the-art Deep Learning algorithm with the optimal speed and accuracy to apply in real-time applications. Five sized versions of YOLOv8, running from lower to higher versions: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x, are available, with the higher versions being capable of providing higher accuracy but lower segmentation speed, and a more powerful machine will be required. In this study, considering the trade-off between speed and accuracy, YOLOv8m was selected in order to apply in real-time surgery as the segmentation model. It has been trained on the Google Colaboratory platform with a pre-trained model that was trained using the COCO val2017 dataset [36]. The following settings parameters were selected for the configuration in the training process: *epochs=300, batch_size=16, image_size=640, workers=8, optimizer=SGD, lr (learning rate)=0.01, momentum=0.937, weight_decay=0.0005 etc.* The size of the input image was set to 640. NVIDIA Tesla A100 GPU has been chosen for the runtime type during the training process on Colab. A total of 300 epochs were trained to generate the last weight, and then the best weight was applied for the testing process.

D. EVALUATION METRICS

The performance evaluation in this study has been measured in 3 schemes: accuracy, speed, and complexity. For the trained model, the validation and testing results have been evaluated according to accuracy evaluation metrics such as **Precision**, **Recall**, **F1-score** (aka **DSC** - Dice Similarity Coefficient) and **mAP** [37]. Intersection Over Union (**IOU**) is responsible for calculating overlap between the predicted bounding box and the ground truth one. The value will be between 0 and 1, according to overlapping regions. For the Confusion Matrix, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are calculated by comparing **IOU** and a predefined threshold value. From these values, **Precision** and **Recall** can be calculated by using Equations (1) and (2).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

From these values, **F1-score**, the harmonic mean of **Precision** and **Recall**, can be calculated by using Equation (3).

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Precision and **Recall** are computed for different Confidence Thresholds for generating Precision-Recall curves. Average Precision (**AP**) is computed as the area under the Precision-Recall curve. While **AP** is computed per class, **mAP** can be calculated as the average over all classes for combining **AP** scores. The optimal Confidence Threshold has also been defined firstly for the model in order to have the highest **F1-score**. Segmentation speed has also been evaluated practically in terms of **Inference Time** and **FPS** in order to be applied in real-time applications. In order to fulfill an optimal segmentation model, a trade-off among accuracy, speed, and model complexity is very important. The model complexity in Deep Learning refers to the number and size of hidden units and layers, activation functions, and learning algorithm parameters. The model complexity comparison among the top models from the ROBUST-MIS 2019 Challenge and the YOLOv8m according to the number of parameters in millions (M) is as listed in Table 2.

E. GESTURE CATEGORIES

Although the existing state-of-the-art instance segmentation algorithms can provide reliable accuracy to deal with and segment blur cases due to reflection or motion and overlay of blood or smoke, there are still some segmentation failures

TABLE 2. Model complexity.

Deep Learning Algorithm	Parameter Size (M)
DeepLabV3+ [22]	59.30
OR-UNet [23]	N/A
BARNet [24]	21.90
YOLOv8m	27.30

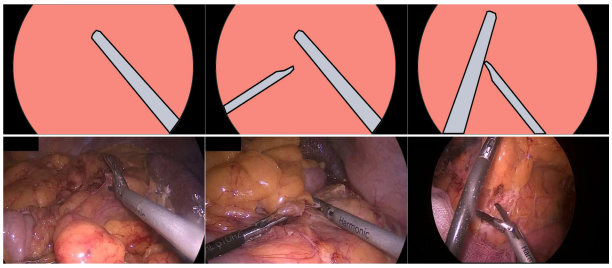


FIGURE 5. Separating cases.

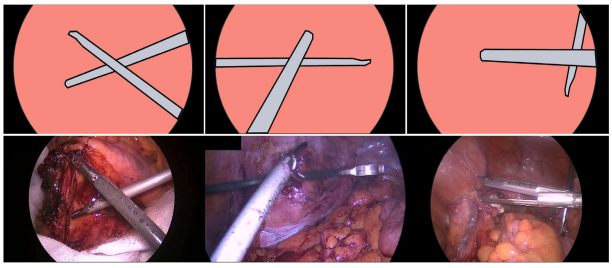


FIGURE 6. Crossing cases.

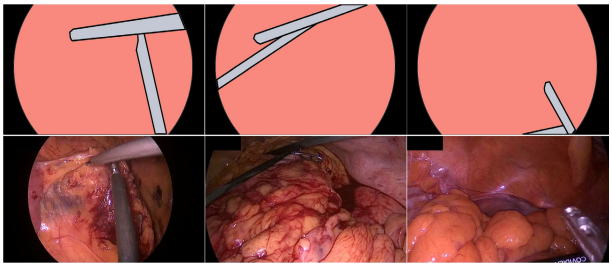


FIGURE 7. Overlapping cases.

due to the position of instruments in the viewpoint [25], [26]. In-depth study and analysis should be performed, therefore, and so three main gesture categories of surgical instruments, those of separating, crossing, and overlapping cases, are classified to be analyzed in this study. The definitions for these three categorized gesture cases, along with a mixed case, are as follows, while Figures 5-8 provide relevant depicted drawings and example images.

Separating Case (S): Each surgical instrument in the image is separating without overlapping each other (Figure 5).

Crossing Case (C): Two surgical instruments in an image are crossing each other, showing both tips (Figure 6).

Overlapping Case (O): One surgical instrument in an image is overlapped by another (Figure 7). Only one tip is pronounced.

Mixed Case (M): Some images might have more than one gesture case (Figure 8). Five mixed cases, found in this study, are: a C+S case; an O+S case; a C+O case

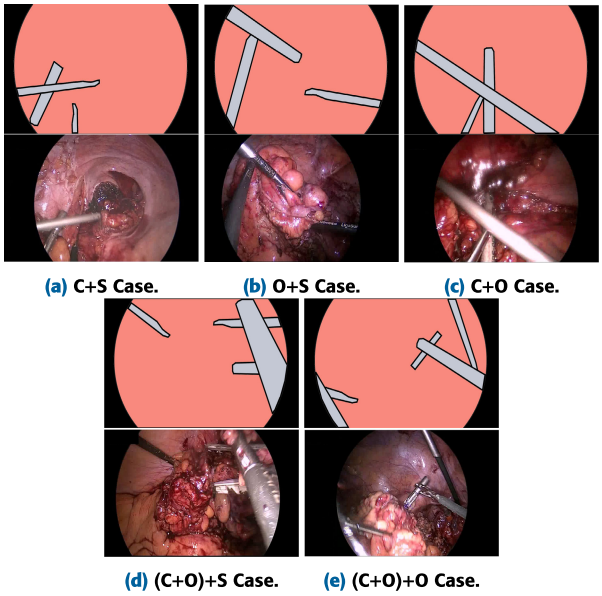


FIGURE 8. Mixed cases.

TABLE 3. Numbers of gesture cases.

Dataset	Separating	Crossing	Overlapping
Stage 3	9,432	126	126
Streaming	1,490	0	33
Total	10,922	126	159

where 3 instruments are involved; a (C+O)+S case where 4 instruments are involved; and a (C+O)+O case where 5 instruments are involved.

An image in each dataset in this study could contain a number of 1-instrument separating cases or a 2-instrument crossing case or a 2-instrument overlapping case or a 3-instrument C+O case or a 3-instrument C+S case or a 3-instrument O+S case or a 4-instrument (C+O)+S case or a 5-instrument (C+O)+O case. The number of each case can be accumulated from its simple case along with its special case within mixed cases as well. It is noted that there is no crossing case in the real-time streaming videos. Table 3 lists the numbers of gesture cases for Stage 3 and Streaming datasets.

IV. RESULTS AND DISCUSSION

This section contains 3 parts, starting with the overall performance evaluation of the YOLOv8, the original Y+BT and the Modified Y+BT models on both the ROBUST-MIS 2019 challenge and the Streaming datasets, then instrument gesture cases analysis, and lastly, the evaluation of the Modified model performance.

A. OVERALL PERFORMANCE EVALUATION

Subsequent to the completion of the training process, a validation of the trained model is important to ensure its performance, providing the opportunity for the model to be improved. For the first association, the high threshold value

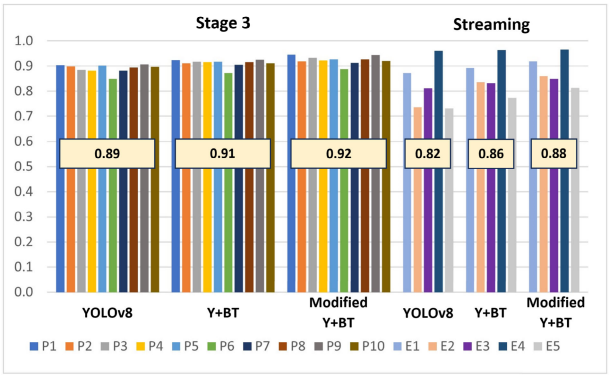


FIGURE 9. F1-score comparison.

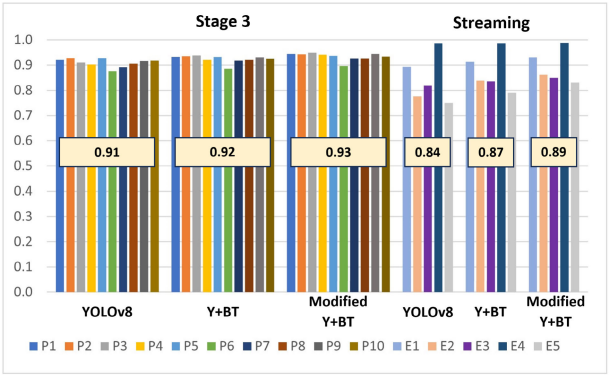


FIGURE 10. mAP comparison.

was set to 0.5 for the original Y+BT model and 0.42 for the Modified model in this study, the same as the optimal confidence threshold, obtained from the validation process, to be the highest accuracy. For the second association, the low threshold value was set to 0.1 for the original model and the minimal value as low as 0.01 for the Modified model to retrieve almost all correct segmentations in the configuration of the ByteTrack algorithm. In order to perform the evaluation for instance segmentation models as described in Table 1 and Figure 3, the YOLOv8m model, the original Y+BT model, and the Modified Y+BT model have been performed on two experiment sets (**P1** to **P10** and **E1** to **E5**), with the blinded ROBUST-MIS 2019 Challenge testing dataset (**C1** to **C10**) and the Streaming dataset (**S1** to **S5**). Figures 9 and 10 present the **F1-score** and **mAP** for these cases. Both evaluation metrics of the Streaming dataset are slightly lower compared to those of the Stage 3 dataset, which is understandable as the captured images from the streaming surgery were originally acquired just for this study from different settings. As can be seen clearly, there is an improvement in the Modified Y+BT model over the YOLOv8, for both metrics and both testing datasets. Details on the benefits of the Modified model are shown and discussed in the following sections.

To benchmark the accuracy for binary segmentation of all models in this study, the comparison with the top 3 ranked models [22], [23], [24] with the highest mean **DSC**s (or **F1-score** in this study) on the ROBUST-MIS 2019 Challenge testing dataset has been declared as listed in Table 4. As can

TABLE 4. Benchmarking with the top 3 previous models.

Deep Learning Algorithm	Mean (DSC)
DeepLabV3+ [22]	0.89
OR-UNet [23]	0.88
BARNet [24]	0.88
YOLOv8 (from the experiments here)	0.89
Y+BT (from the experiments here)	0.91
Modified Y+BT (from this study)	**0.92**

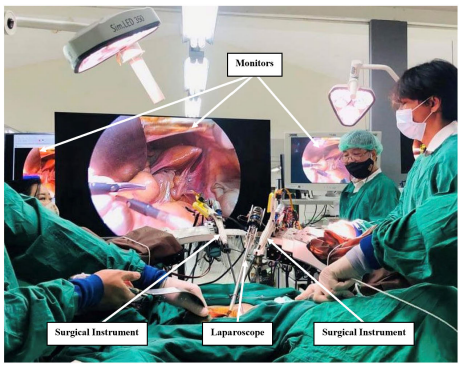


FIGURE 11. Real-timed experiment on a cadaver.

TABLE 5. Computational speed.

Models	Inference Time (ms)	FPS
YOLOv8m Model	16.67	60
Y+BT Models	22.22	45

be seen here, all models in this study, especially the Modified Y+BT model, conquered the results of those top 3. This is because all models benefit from the advantages of the anchor-free approach and augmentation technique during the training process. As for the model complexities of all models experimented and analyzed in this study, they are much the same, while the additional ByteTrack and Modified algorithms are not part of the training process in the Deep Learning model.

Next, as for the real-time process to evaluate the computational speed of the models, 5 practical laparoscopic gynecologic surgeries on donated soft-tissue cadavers have been performed. Figure 11 illustrates a special set-up of the experiment conducted on a cadaver in the operating theater room. With the help of the extra light via a medium-sized incision through the abdomen, the inner view could be seen and observed much more clearly and easily. As can be seen vividly on the main monitor, used by the surgeon to do the operation, two surgical instruments were captured by the laparoscope, while the left monitor was used for showing the real-time segmentation results from the Modified model. All of the results from the streaming cases shown in this paper have come from this set-up. Table 5 reports on the computational speed for the YOLOv8m and Y+BT models performed here, measured in **Inference Time** (ms) and **FPS**. In the use of the PyTorch GPU version with NVIDIA RTX 3050 Ti, with speeds of 60 **FPS** and 45 **FPS** for the YOLOv8m and Y+BT models, respectively, the

results showed that both models could cope well for real-time applications, with 25 **FPS** streaming videos having the same speed as those in the ROBUST-MIS 2019 Challenge dataset.

Since both models were equipped with YOLOv8, which is a single-stage algorithm, processing the prediction and classification in a single pass with the assistance of a high-performance PyTorch GPU version, this results in a fast speed capable of real-time applications. As a result, the Y+BT model could continue associating without the segmentor delay, yielding 45 **FPS**, which still serves well along with 25 **FPS** streaming video from the laparoscope, despite the fact that its overall segmentation speed decreased due to the additional processing time in the ByteTrack.

B. GESTURE CASES ANALYSIS

Per gesture cases analysis, all the results came from the YOLOv8 vs. Modified models. Figure 12 shows example images of successfully segmented results from the ROBUST-MIS 2019 Challenge and Streaming datasets for the different gesture categories of separating, crossing, overlapping, and mixed cases. As noted in Table 3, only separating and overlapping cases existed in Streaming datasets. The left column illustrates ground truth from annotation, and the right column illustrates practical results from YOLOv8 surgical instruments segmentation. These results are also the same as those from the Y+BT and Modified models.

Although YOLOv8 can segment surgical instruments accurately with high evaluation metrics, a number of segmentation failures can still be found in all gesture cases. Figure 13 illustrates example images of unsuccessfully segmented results by YOLOv8, which can be recovered by ByteTrack for different gesture categories, together with corresponding ground truth images from the ROBUST-MIS 2019 Challenge and Streaming datasets. From the results, the most straightforward case to gain the least missed segmentations was the separating case; however, there are a few factors that could cause segmentation failures, including:

- Instance Tininess: a problem which occurs when the instrument is shown in a very small size that could be caused by occlusion.
- Occlusion (by tissues and/or instruments): a problem which occurs when the designated instance is covered by some tissues and/or other instances. Per the instance tininess and occlusions, the failures can be seen in Figures 13(a-d) and (f).
- Underexposed Instance Color: a problem which occurs when the color of the surgical instrument is similar to or not obviously different from the background color, as can be seen in Figures 13(d) and 13(e).

The next section shows how ByteTrack could recover these failures.

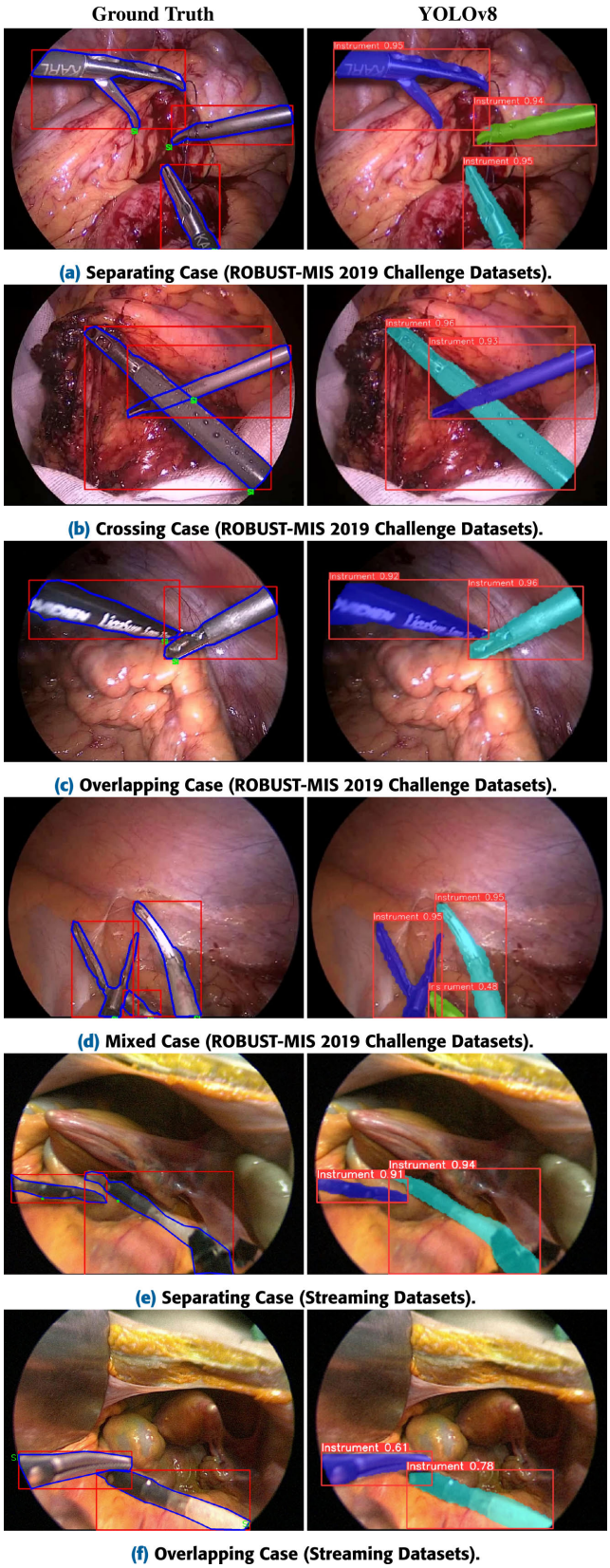


FIGURE 12. Successful segmentations.

C. EVALUATION ON THE MODIFIED MODEL PERFORMANCE

In order to exhibit the benefit of ByteTrack, sequences of frames in Figures 14-16, show results from the YOLOv8 model (left column) and results from the Modified Y+BT model (right column) for each gesture case. It is visible that

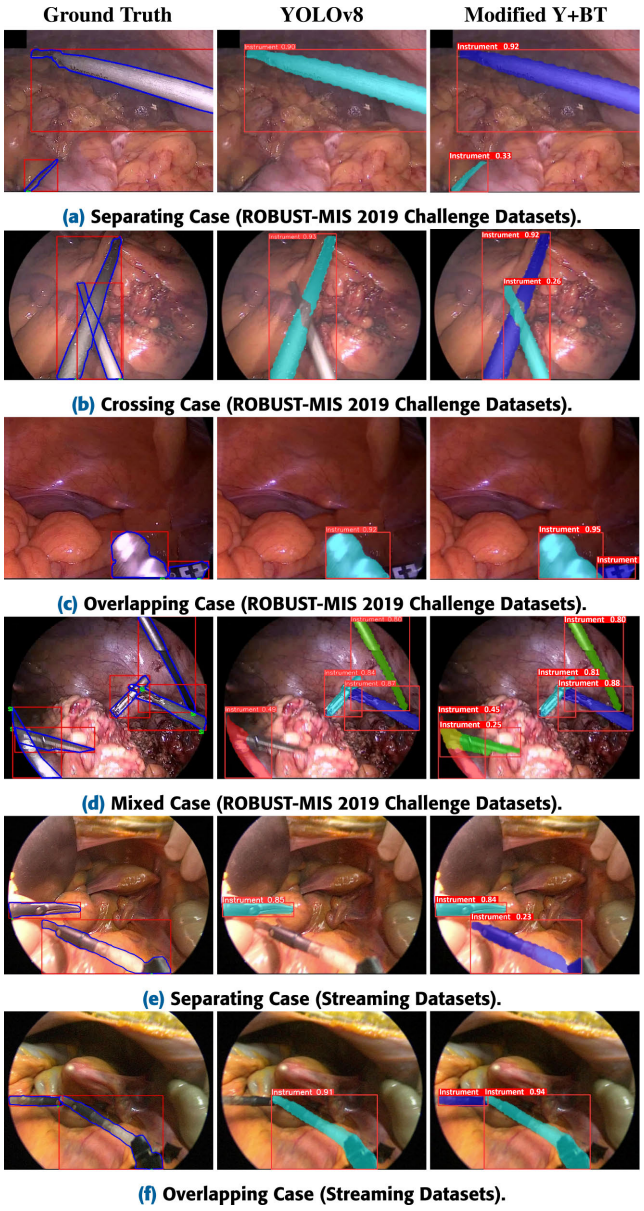


FIGURE 13. Segmentation improvements.

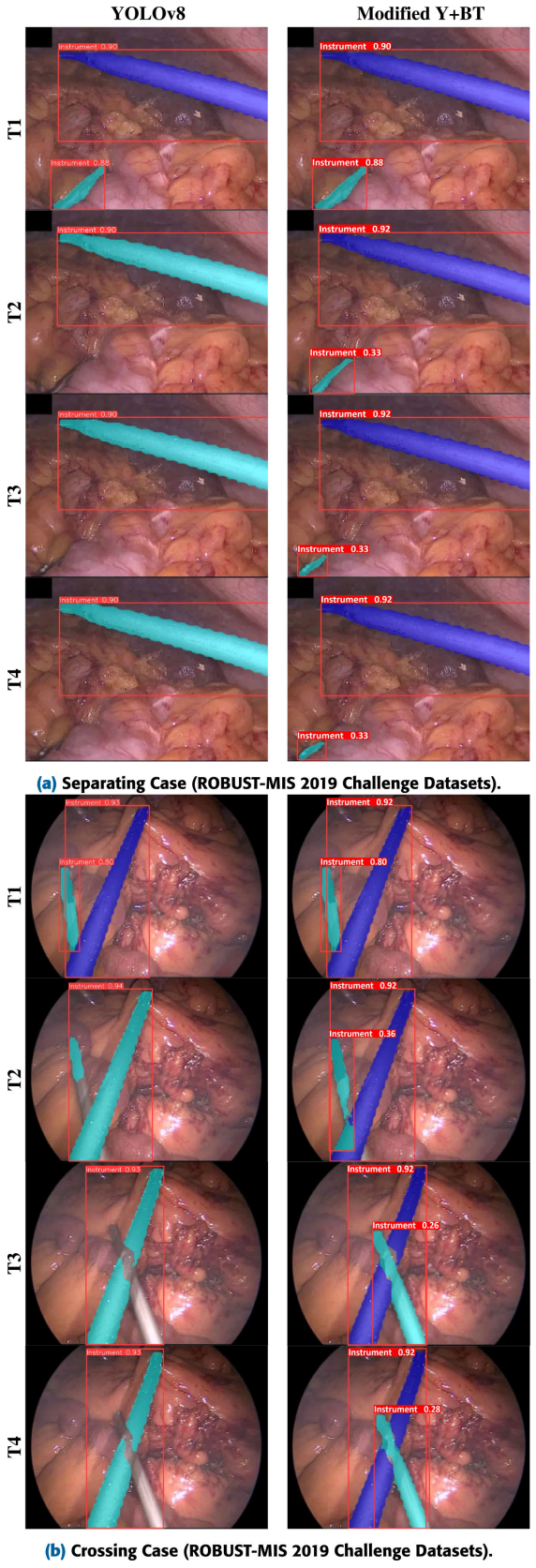
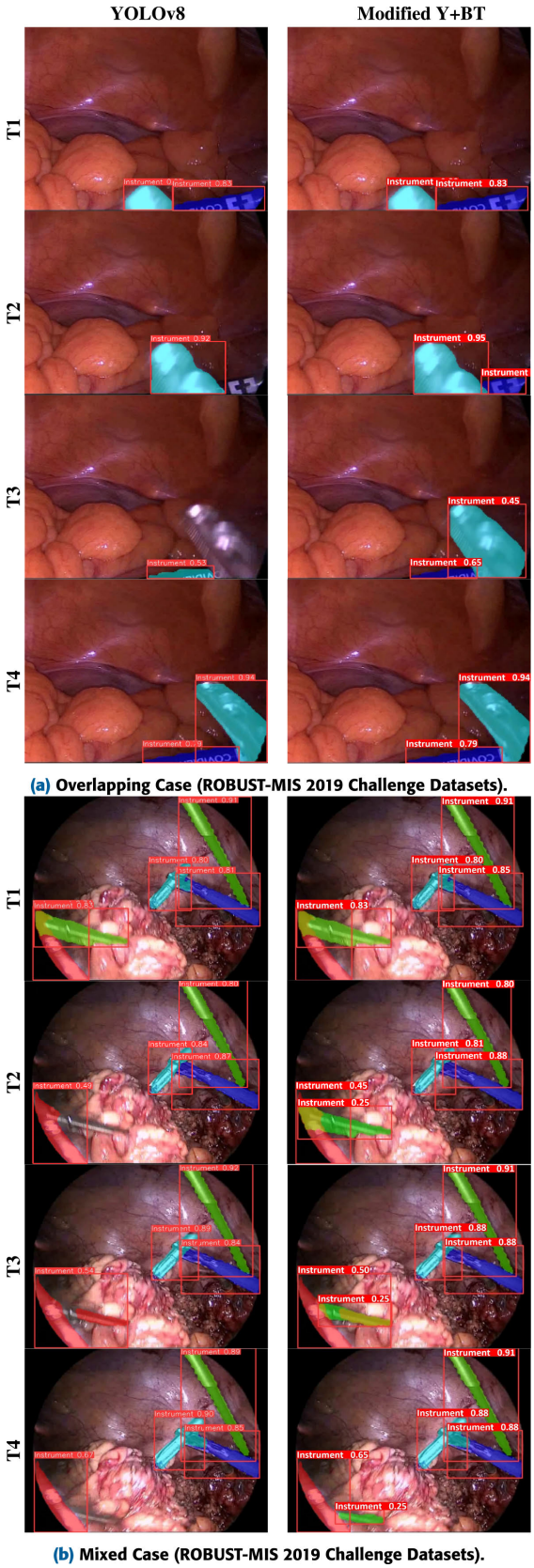


FIGURE 14. Frame sequences showing ByteTrack performance.

only T1 (for all cases) and T4 (for Figures 15(a) and 16(b)) frames were segmented accurately for both columns.

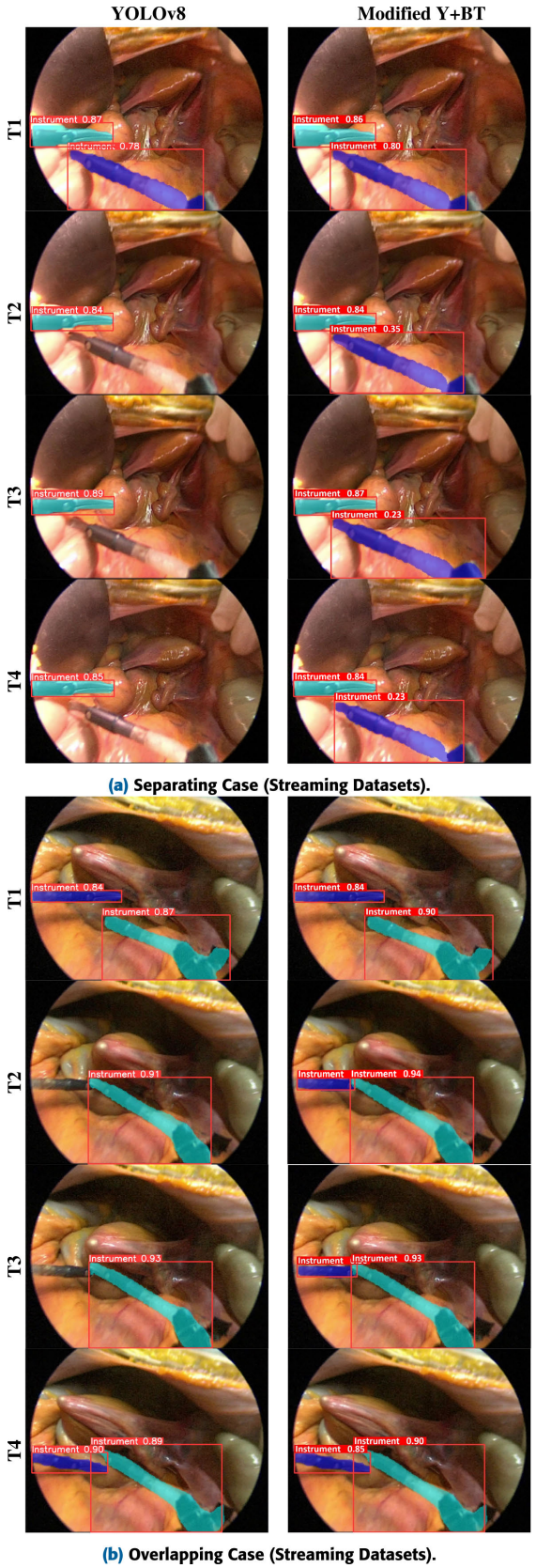
However, during T2-T4 frames (in Figures 14, 15(b) and 16(a)) and T2-T3 frames (in Figures 15(a) and 16(b)), any features causing segmentation failures have taken place, resulting in low confidence values on one of those instruments, hence the segmentation failure for the YOLOv8 model. While instances have been segmented correctly in the previous frame (T1), in ByteTrack, tracklets have been introduced or updated as important information for the following frames (T2, and so on). For instance, in the T2 frame, with a low confidence value for an instrument in the first association, causing a segmentation miss, ByteTrack could recover that particular instrument segmentation by using the held tracklet in the second association if the low score bounding box of the missing instrument is matched correctly to the held tracklet.



(a) Overlapping Case (ROBUST-MIS 2019 Challenge Datasets).

(b) Mixed Case (ROBUST-MIS 2019 Challenge Datasets).

FIGURE 15. Frame sequences showing ByteTrack performance.



(a) Separating Case (Streaming Datasets).

(b) Overlapping Case (Streaming Datasets).

FIGURE 16. Frame sequences showing ByteTrack performance.

Furthermore, in addition to the overall accuracy evaluation metric, **F1-score**, in this study, the accuracy percentage has been used to report on the performance of each gesture category in order to determine the strengths and weaknesses

TABLE 6. Accuracy Percentage on Gesture Categories.

Dataset	Separating		Crossing		Overlapping	
	YOLOv8	Y+BT	YOLOv8	Y+BT	YOLOv8	Y+BT
Stage 3	86%	93%	24%	38%	73%	78%
Streaming	81%	89%	N/A	N/A	12%	52%
Total	86%	92%	24%	38%	60%	72%

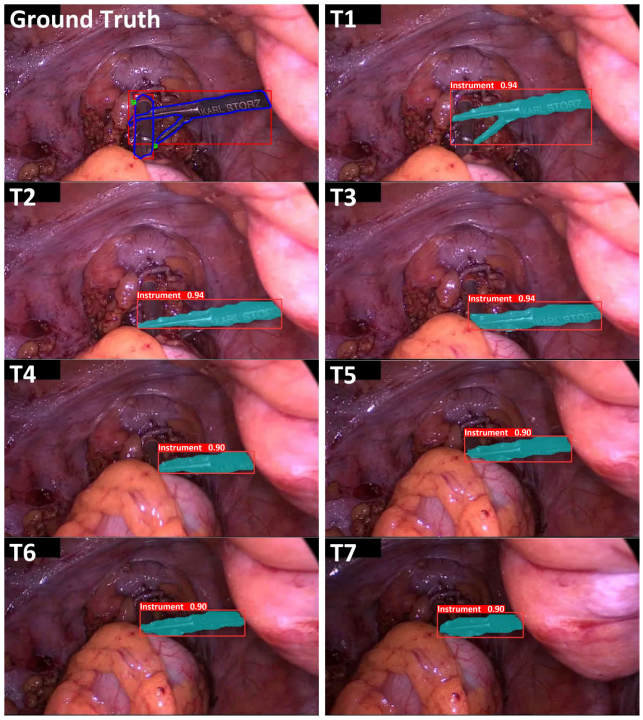


FIGURE 17. Limitation of the modified model.

of the models. Table 6 lists the results of segmentation accuracy percentages for three main gesture categories: separating, crossing, and overlapping. Notice that only 3 different gesture categories are shown, whereas the final mixed case did not specifically exist, as each sub-case in the mixed cases has been presented separately along with its main case. The Modified algorithm can improve accuracy percentages for all gesture categories. The highest accuracy percentage (92%) went to the separating case, which means 92% of this case has been correctly segmented. Furthermore, the accuracy percentage for the overlapping case is 72%. However, the crossing case still remains challenging, with only 38%.

As can be seen clearly, with the help of the two-association scheme of ByteTrack, the Modified Y+BT algorithm could improve segmentation accuracy by using the information from the previous tracklet. However, there are still some existing limitations.

Figures 17 and 18 present two examples of the limitations of ByteTrack, with two frame sequences (T1-T7), commencing from relevant ground truth, showing segmentation failures as limitations of ByteTrack. In Figure 17, only one instance can be correctly segmented as its confidence value was higher

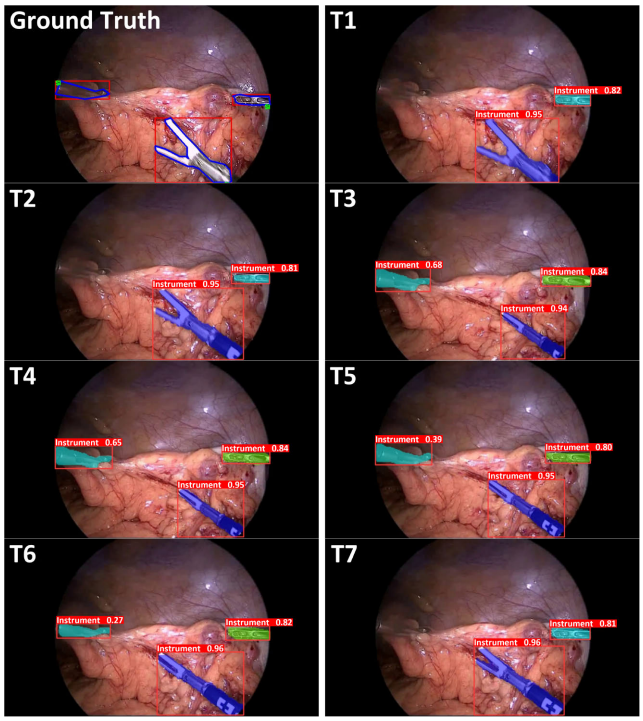


FIGURE 18. Limitation of the modified model that can be recovered.

than the confidence threshold at all times, whereas the other existing instance failed throughout T1-T7 frames due to its too low confidence value. On the other hand, Figure 18 shows how any instances could come back as segmented instances. This sequence commenced from an undetected instance due to low confidence values in T1 and T2, where the correct segmentations occurred in T3 and T4 as the confidence value reached the high confidence threshold; hence, the tracklet was introduced and can be held to support instances with low confidence values in the following frames. Later, though with low confidence values, the instance could still be segmented, as shown in the T5 and T6 frames. However, with a too low confidence value, that instance was undetected completely again, as in the T7 frame.

V. CONCLUSION

Up until now, a large number of object segmentation algorithms have been developed, and high accuracy could be reached, but only a few could deal with it in real-time. Furthermore, object occlusion has still been a critical challenge for all, including surgical instruments segmentation. This would not be easily solved unless the root cause had been analyzed. The main purpose of this research was to investigate real-time surgical instruments segmentation in laparoscopic surgeries, which was commenced by applying YOLOv8, the state-of-the-art instance segmentation algorithm, and testing the model on ROBUST-MIS 2019 Challenge datasets; hence, the results from this could be compared with those of other previous studies. However, unlike other images, the viewpoint of images from a laparoscope is somehow unique due to their red-brown tone range limitation and

more rapid motion from a close-up position. Although YOLOv8 could have offered an excellent result for all of the evaluation metrics—the overall accuracy (**F1-score** and **mAP**), the model complexity, and also segmentation speed, which is very important for real-time operations—from the results analysis, three features (instance tininess, occlusion, and underexposed instance color) remain critical issues for segmentation failures. With the modification added on to the combination of the state-of-the-art YOLOv8 and ByteTrack, the experimental results show that the Modified model could easily serve in real-time and expressed the best results in both **F1-score** and **mAP** with the comparative model complexity. All results were then further inspected according to three different instrument gesture categories: separating, crossing, and overlapping, to find that the segmentation difficulties lie mostly in crossing and overlapping. From the close-up observation of frame sequences, when an occlusion occurs, usually instances have transitions before and after the occlusion. Therefore, instead of dealing with segmentation from still images, analyzing the segmentation from a sequence of frames was considered. ByteTrack, the high-performance object tracking algorithm, was opted for incorporating with YOLOv8 along with some modifications to deal with the unique character of the dataset for having close-up and fast-moving objects to accomplish better segmentation. Owing to the 2-association scheme of ByteTrack, segmentation failures of a current image frame could be recovered by the tracklet from the previous frame, resulting in higher accuracy with a small price to pay for computational time, which is still efficient for the real-time application. Real-time experiments conducted on 5 soft-tissue cadaver cases to validate the effectiveness of YOLOv8, the original Y+BT, and the Modified Y+BT models have also guaranteed the measured results. Captured images from these streaming surgeries have also been utilized as blinded test data and have been publicized. As for future research, further effort to investigate how to gain better instance segmentation and tracking in laparoscopy should be considered.

ACKNOWLEDGMENT

The authors express their gratitude to the individuals who gave their bodies for anatomical research and study, which could contribute to an improvement of overall knowledge and effective patient care. Therefore, these donors and their families deserve their greatest gratitude.

REFERENCES

[1] A. Moglia, K. Georgiou, E. Georgiou, R. M. Satava, and A. Cuschieri, "A systematic review on artificial intelligence in robot-assisted surgery," *Int. J. Surg.*, vol. 95, Nov. 2021, Art. no. 106151, doi: 10.1016/j.ijsu.2021.106151.

[2] A. Agustinos, R. Wolf, J. A. Long, P. Cinquin, and S. Voros, "Visual servoing of a robotic endoscope holder based on surgical instrument tracking," in *Proc. 5th IEEE RAS/EMBS Int. Conf. Biomed. Robot. Biomechatronics*, Aug. 2014, pp. 13–18, doi: 10.1109/BIOROB.2014.6913744.

[3] *Robotic-Assisted Surgery for Patients | Intuitive*. Accessed: Mar. 21, 2023. [Online]. Available: https://www.intuitive.com/en-us/patients/da-vinci-robotic-surgery

[4] T. Langø, G. A. Tangen, R. Mårvik, B. Ystgaard, Y. Yavuz, J. H. Kaspersen, O. V. Solberg, and T. A. N. Hernes, "Navigation in laparoscopy—Prototype research platform for improved image-guided surgery," *Minimally Invasive Therapy Allied Technol.*, vol. 17, no. 1, pp. 17–33, Jan. 2008, doi: 10.1080/13645700701797879.

[5] Y. Wang, Q. Sun, Z. Liu, and L. Gu, "Visual detection and tracking algorithms for minimally invasive surgical instruments: A comprehensive review of the state-of-the-art," *Robot. Auto. Syst.*, vol. 149, Mar. 2022, Art. no. 103945, doi: 10.1016/j.robot.2021.103945.

[6] N. N. Myo, A. Boonkong, D. Hormdee, S. Sonsilphong, A. Sonsilphong, and K. Khampitak, "Laparoscope manipulating robot (LMR) navigation using deep learning-based surgical instruments detection," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2022, pp. 1142–1147, doi: 10.23919/APSIPAASC55919.2022.9980059.

[7] S. Kletz, K. Schoeffmann, J. Benois-Pineau, and H. Husslein, "Identifying surgical instruments in laparoscopy using deep learning instance segmentation," in *Proc. Int. Conf. Content-Based Multimedia Indexing (CBMI)*, Sep. 2019, pp. 1–6, doi: 10.1109/CBMI.2019.8877379.

[8] M. Fox, M. Taschwer, and K. Schoeffmann, "Pixel-based tool segmentation in cataract surgery videos with mask R-CNN," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 565–568, doi: 10.1109/CBMS49503.2020.00112.

[9] G. Ciaparrone, F. Bardozzo, M. D. Priscoli, J. L. Kallewaard, M. R. Zuluaga, and R. Tagliaferri, "A comparative analysis of multi-backbone mask R-CNN for surgical tools detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8, doi: 10.1109/IJCNN48605.2020.9206854.

[10] K. Lam, F. P.-W. Lo, Y. An, A. Darzi, J. M. Kinross, S. Purkayastha, and B. Lo, "Deep learning for instrument detection and assessment of operative skill in surgical videos," *IEEE Trans. Med. Robot. Bionics*, vol. 4, no. 4, pp. 1068–1071, Nov. 2022, doi: 10.1109/TMRB.2022.3214377.

[11] R. Sanchez-Matilla, M. Robu, I. Luengo, and D. Stoyanov, "Scalable joint detection and segmentation of surgical instruments with weak supervision," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 501–511, doi: 10.1007/978-3-030-87196-3_47.

[12] D. Banik, K. Roy, and D. Bhattacharjee, "EM-Net: An efficient M-net for segmentation of surgical instruments in colonoscopy frames," *Nordic Mach. Intell.*, vol. 1, no. 1, pp. 14–16, Nov. 2021, doi: 10.5617/nmi.9122.

[13] T. Kurmann, P. Márquez-Neila, M. Allan, S. Wolf, and R. Sznitman, "Mask then classify: Multi-instance segmentation for surgical instruments," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 7, pp. 1227–1236, Jun. 2021, doi: 10.1007/s11548-021-02404-2.

[14] *Object Recognition Vs Object Detection Vs Image Segmentation*. Accessed: Mar. 21, 2023. [Online]. Available: https://www.kaggle.com/getting-started/169984

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[16] M. Tan, R. Pang, and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, May 2019, pp. 6105–6114.

[17] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT++ better real-time instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1108–1121, Feb. 2022, doi: 10.1109/TPAMI.2020.3014297.

[18] H. Zhang, Y. Tian, K. Wang, W. Zhang, and F.-Y. Wang, "Mask SSD: An effective single-stage approach to object instance segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 2078–2093, 2020, doi: 10.1109/TIP.2019.2947806.

[19] *What Is Instance Segmentation*. Accessed: Mar. 25, 2023. [Online]. Available: https://www.v7labs.com/blog/instance-segmentation-guide

[20] D. M. Chilukuri, S. Yi, and Y. Seong, "A robust object detection system with occlusion handling for mobile devices," *Comput. Intell.*, vol. 38, no. 4, pp. 1338–1364, Feb. 2022, doi: 10.1111/coin.12511.

[21] *Robust Medical Instrument Segmentation (ROBUST-MIS) Challenge (Synapse.org)*. Accessed: Jun. 15, 2023. [Online]. Available: https://www.synapse.org/#!Synapse: syn18779624/wiki

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with Atrous separable convolution for semantic image segmentation," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
[24] Z. L. Ni et al., "Barnet: Bilinear attention network with adaptive receptive fields for surgical instrument segmentation," in *Proc. Twenty-Ninth Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 832–838, doi: 10.24963/ijcai.2020/116.
[25] T. Roß, "Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101920, doi: 10.1016/j.media.2020.101920.
[26] J. Angeles-Ceron, G. Ochoa-Ruiz, L. Chang, and S. Ali, "Attention YOLACT++ for real-time instance segmentation of medical instruments in endoscopic procedures," in *Proc. LatinX AI at Comput. Vis. Pattern Recognit. Conf.*, Jun. 2021, doi: 10.52591/lxai2021062511.
[27] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extraction*, vol. 5, no. 4, pp. 1680–1716, Nov. 2023, doi: 10.3390/make5040083.
[28] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng, "StrongSORT: Make DeepSORT great again," *IEEE Trans. Multimedia*, vol. 25, pp. 8725–8737, 2023, doi: 10.1109/TMM.2023.3240881.
[29] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric sort: Rethinking sort for robust multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9686–9696, doi: 10.1109/cvpr52729.2023.00934.
[30] Y. Zhang, "ByteTrack: Multi-object tracking by associating every detection box," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–21, doi: 10.1007/978-3-031-20047-2_1.
[31] *Real Time Object Tracking and Segmentation Using YOLOv8 With Strongsort, Ocsort and ByteTrack*. Accessed: May 10, 2023. [Online]. Available: https://siddharthksah.medium.com/real-time-object-tracking-and-segmentation-using-yolov8-with-strongsort-ocsort-and-bytetrack-180eef43354a
[32] *Football YOLOv8 Segmentation ByteTrack*. Accessed: May 10, 2023. [Online]. Available: https://www.kaggle.com/code/stpeteishii/football-yolov8-segmentation-bytetrack
[33] L. Maier-Hein et al., "Heidelberg colorectal data set for surgical data science in the sensor operating room," *Sci. Data*, vol. 8, no. 1, p. 101, Apr. 2021, doi: 10.1038/s41597-021-00882-2.
[34] *Roboflow (Version 1.0) [Software]*. Accessed: Jan. 18, 2023. [Online]. Available: https://roboflow.com
[35] *A Complete Guide To Data Augmentation*. Accessed: Mar. 28, 2023. [Online]. Available: https://www.datacamp.com/tutorial/complete-guide-data-augmentation
[36] *Ultralytics YOLO (Version 8.0.0) [Computer Software]*. Accessed: May 3, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[37] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electronics*, vol. 10, no. 3, p. 279, Jan. 2021, doi: 10.3390/electronics10030279.

NYI NYI MYO received the B.Eng. degree in electronic engineering from the University of Portsmouth, U.K., in 2018. He is currently pursuing the M.Eng. degree in computer engineering with Khon Kaen University, Thailand. His research interests include machine learning, medical imaging, and robotics.

APIWAT BOONKONG received the B.Eng. and M.Eng. degrees in computer engineering from Khon Kaen University, Thailand, in 2012 and 2017, respectively. Since then, he has started working at the Department of Computer Engineering, Faculty of Engineering, Nakhon Phanom University, Thailand.

KOVIT KHAMPITAK is currently a Professor with the Department of Obstetrics and Gynecology, Khon Kaen University, Thailand. His research interests include gynecologic surgery, biomedical equipment, and laparoscope manipulating robots.

DARANEE HORMDEE (Member, IEEE) received the B.Eng. degree in computer engineering from Khon Kaen University, Thailand, in 1996, and the M.Sc. and Ph.D. degrees from The University of Manchester, U.K., in 1998 and 2002, respectively. She is currently an Assistant Professor with the Department of Computer Engineering, Khon Kaen University. Her research interests include embedded system design and mechatronics.