IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 May 2023, accepted 28 June 2023, date of publication 7 July 2023, date of current version 19 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3293006

APPLIED RESEARCH

Performing and Evaluation of Deep Learning Models for Uterus Detection on Soft-Tissue Cadavers in Laparoscopic Gynecology

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

[1]Department of Computer Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand

[2]Department of Obstetrics and Gynecology, Faculty of Medicine, Khon Kaen University, Khon Kaen 40002, Thailand

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

This work was supported in part by the Nakhon Phanom University Ph.D. Scholarship, Thailand.

ABSTRACT Nowadays, with the current technological forces that have been shaping our bright future, one of these is Computer Vision. This statement is true across various matters, including laparoscopic gynecology, where computer-aided procedures for object recognition could offer surgeons the opportunity to ease up on on-going surgeries and/or to practice their surgical skills with offline surgeries. However, most of the previous work has been retrospective and focused on methodology from a computational viewpoint with minimal datasets showing how Computer Vision can be utilized for laparoscopic surgery. The main purpose of this paper is not just to evaluate state-of-the-art object detection models for uterus detection, but also to emphasize clinical application via the collaboration between surgeons and peopleware which is important in the further development and adoption of this technology, leading to improved clinical outcomes in Laparoscopic Gynecology. Two experiment phases have been conducted. Phase#1 applied 8 different Deep Learning models for uterus detection and were tested on the dataset, obtained from 42 public YouTube videos in Laparoscopic Gynecologic Surgery. In order to prove this new technology before performing on patients, and also due to the ethics of human experimentation, extensive testing on soft-tissue cadavers has been used, because theoretically, a soft-tissue cadaver is considered the closest to human in terms of shape and structure. Therefore Phase#2 has been performed on the best models from the first experiment phase serving a real-time streaming feed during 4 soft-tissue cadaver laparoscopic surgeries. Four models, pre-trained on the COCO 2017 Dataset on TensorFlow Model Zoo: CenterNet; EfficientDet; SSD; and Faster R-CNN; plus YOLOv4 on Darknet Framework, along with YOLOv4, YOLOv5 and YOLOv7 on Pytorch have been scrutinized here. The inference time (in FPS: Frame Per Second), F1-score and AP (Average Precision) have been used as evaluation metrics. The results exhibited that all 3 YOLOs on PyTorch outperformed all effectiveness metrics, including with great inference speed which is suitable for real-time surgeries. Lastly, a by-product but also useful contribution of this work, is the annotated dataset on uterus detection from both public videos and live feed on cadaver surgeries.

INDEX TERMS Object detection, uterus detection, laparoscopic surgery, deep learning, soft-tissue cadaveric surgery.

I. INTRODUCTION

In laparoscopic surgery, a type of Minimally Invasive Surgery (MIS), the surgeon uses a digital camera, known as a laparoscope, which sends images of the inside of the abdomen or pelvis to a monitor [1]. In doing so, such a camera affords surgeons the opportunity to gain assistance from image guided surgery systems. Image interpretation is essential for these systems, namely a computer is required to be capable of understanding what is being seen by the laparoscope.

The performance of RAS (Robot Assisted Surgery) has been greatly improved by the use of Computer Vision and

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

Artificial Intelligence (AI) tools, mainly in the efficiency of medical instruments and healthcare safety. Recently, consideration was given to a Deep Learning (DL) approach for the detecting of organs required for a gynecologic laparoscopic surgery, however most substantial studies for surgical instruments and anatomy detection in computational publication have concentrated on algorithms with little clinically relevant information. The review [2] reported current technology has shown mostly above 85% for the accuracy in comprehensive classification of surgical instruments, organs, and surgical procedures. Despite live surgery, performing surgical procedures on cadavers is essential for safe and effective medical practice. Thanks to all human cadaveric donors, the benefits of the soft-tissue cadaver are obvious, including soft and pliable visceral organs, bright and realistic color tone and a 'lifelike' feel.

This paper firstly focuses on applying and comparing various DL models for uterus detection from captured laparoscopic images from public gynecologic surgery videos.

The major contribution of this paper is on extensively conducting the selected object detection models from the first experiment phase on soft-tissue cadavers, to verify whether those models work properly and efficiently for both accuracy and speed aspects on the live-streaming feed in the laparoscopic surgery.

The remainder of this paper has been organized as follows. Section II reviews related work. Details of research methodology and its implementation along with a series of comparative simulations are described in Section III. Then Section IV presents experimental results and outcome discussion. Finally, the conclusions are in Section V.

II. LITERATURE REVIEW

This section gives a brief information on various related work, followed by state-of-the-art of object detection models. Then evaluation metrics are presented.

A. PREVIOUS WORK

In 2017, Convolutional Neural Network (CNN) was used as a method for differentiating uterine arteries from ureters [3] and segmenting the liver from other anatomy [4]. Later, in 2018, CNN was also applied for recognizing the most frequent surgical actions in laparoscopic gynecology including dissection, coagulation, cutting, injection, suction and irrigation, and suturing [5]. In the same year, a dataset, LapGyn4 [6], was introduced for 4 use cases: surgical actions, anatomical structures, actions on anatomy, and instrument count. This paper also presented a quantitative base line on evaluations for image classification using each dataset in LapGyn4 with GoogLeNet architecture. Since then, this dataset has been used widely, including in this work. Recently, anatomical landmarks have been detected using YOLOv3 as the 'critical view of safety' in order to avoid Bile Duct Injury (BDI) during laparoscopic cholecystectomy. In the same year, 2020, object segmentation on uterus, ovaries and surgical instruments were undertaken via Mask R-CNN with the accuracy of 97%, 24% and 86%, respectively. This work was evaluated on their SurgAI dataset [7], which has been intended to be made public upon the paper's acceptance; however, as of this year (2023), it has not been publicized. Last but by no means least in this section, a list of proofs [2] showing how DL has been used in laparoscopic surgery for a variety of purposes between the years 2012-2020 has been reviewed. Around half of the work has been reported on surgical instrument detection with 15% on anatomy detection. The common tested procedures were cholecystectomy (51%) with 26% on gynecology — mainly hysterectomy and myomectomy.

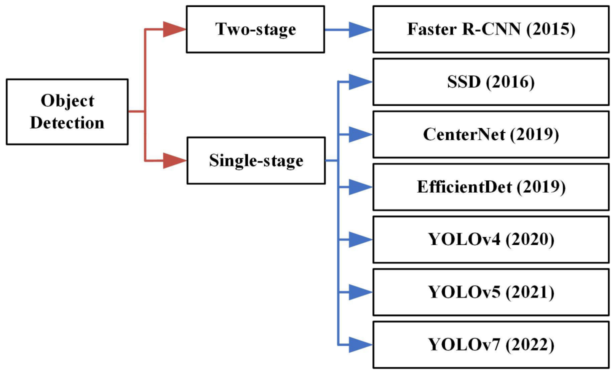


FIGURE 1. Two main approaches in object detection.

As for object detection, CNN has always been the fundamental DL model, which later has been improved within many new architectures. The following section explains the chosen models that have been explored in this paper.

B. STATE OF THE ART

The current state-of-the-art (SoA) on Computer Vision proposes and explores different strategies for object classification and detection based on 2D images. Generally, there are two main approaches in object detection based on DL, namely two-stage and single-stage detections [8]. Figure 1 lists all 7 models that have been explored in this work, where 5 models have been chosen from the survey of modern DL based object detection models [9], with Faster R-CNN [10] as a representative of two-stage detector while Single-Shot MultiBox Detector (SSD) [11], CenterNet [12], EfficientDet [13], and 'You Only Look Once' (YOLOv4) [14] are one-stage detectors. The other 2 models are YOLOv5 [15] and YOLOv7 [16] which have recently been released.

C. EVALUATION METRICS

Firstly, two primary terminologies in accuracy evaluation are **Ground Truth** (green box) and **Bounding Box** [17] (red box), as illustrated in Figure 2, which are used in object detection, with the outcomes of the detection process being examined for their accuracy against what is in the present.

The next term is **Confusion Matrix** [18] which interprets each Bounding Box against Ground Truth. Figure 3 depicts 4 cases of Confusion Matrix. Predicted by a detection model,

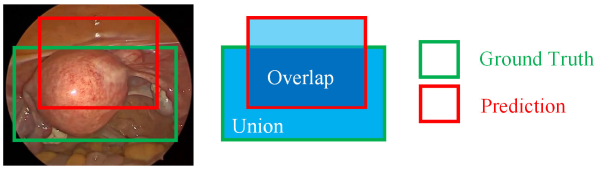
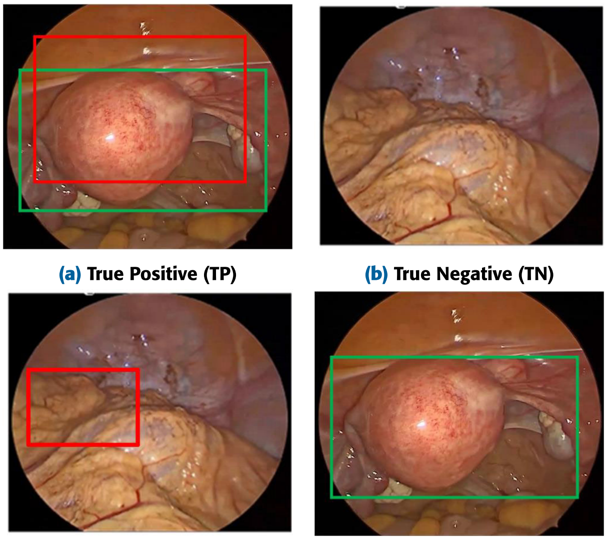


FIGURE 2. Ground truth vs prediction bounding boxes.



(a) True Positive (TP) (b) True Negative (TN)

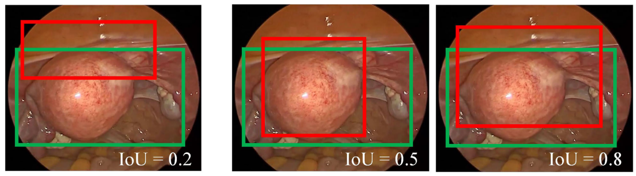
(c) False Positive (FP) (d) False Negative (FN)

FIGURE 3. Confusion matrix.

a positive or negative group can be categorized within these prediction results as being true or false, respectively. With a **TP** - True Positive (Figure 3(a)), a uterus is correctly predicted; while with a **TN** - True Negative (Figure 3(b)), there is a correct prediction that there is not a uterus; on the other hand with a **FP** - False Positive (Figure 3(c)), the model predicts there is a uterus, but this is actually not correct; and lastly, with a **FN** - False Negative (Figure 3(d)), there is an incorrect prediction of there being no uterus.

Another terminology, the **IoU** - Intersection over Union ratio [17] illustrates how much there is an overlap between the Bounding Box surrounding a predicted uterus and the Bounding Box surrounding the Ground Truth. Figure 4 provides some examples of the results of the same image with **IoU** of 0.2, 0.5 and 0.8, respectively. This **IoU** indicates how much the prediction overlaps with the **Ground Truth**, to determine if a predicted result is either a **TP** or a **FP**. Considering Figure 4(b), if the **IoU** threshold is predefined as 0.5 and below, this would be a **TP**. While if the threshold is above 0.5, this would yield to an **FP**.

Two more performance indicators of object detection, **Precision** and **Recall**, can be explained as follows. **Precision** [17] (as can be calculated as in Equation (1)) gives the ratio of the number of **TPs** in respect to the total number of positive predictions. **Recall** (as in Equation (2)) gives the ratio of the number of **TPs** in respect of the total number of actual (and relevant) objects.



(a) IoU = 0.2 (b) IoU = 0.5 (c) IoU = 0.8

FIGURE 4. IoU.

Two types of metrics used to evaluate object detection models in this work are in relation to accuracy and speed. While **F1-score** and **AP** (Average Precision) are used to evaluate the accuracy-wise performance, **Inference Time** is used to evaluate the speed-wise performance.

F1-score [19] (as in Equation (3)) constitutes a weighted average of both the **Precision** and **Recall** values. By measuring the balance between **Precision** and **Recall**, ranging from 0 to 1, a resulting value of 1 represents the highest degree of accuracy. When the value of **F1-score** is high, this means both the **Precision** and **Recall** are high. A lower **F1-score** score means a greater imbalance between **Precision** and **Recall**. The mathematical definitions of these are as follow:

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

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

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

AP [18] is another well-known metric in measuring the accuracy of object detection, which can be computed from the Area Under the Curve (**AUC**) [19] of the **Precision** - **Recall** relationship (as in Equation (4)), providing the **AP** per class for a set of predictions. The average of this value, taken over all classes, is termed as mean Average Precision (**mAP**). The model with the highest **AUC** is the best performing model. While **F1-score** is usually used for a single-class object detection, **mAP** is more popular when it comes to evaluating multi-class detection models.

$$AP = \sum_{k=0}^{k=n-1} [Recalls(k) - Recalls(k+1)] \times Precisions(k) \tag{4}$$

Lastly, **Inference Time** can be measured in **FPS** - Frame Per Second, to define how long each image can be processed to generate the desired output during the testing process by a detection model.

III. METHODOLOGY

Figure 5 depicts an overview of the uterus detection pipeline used in this work. It comprises four processes; a training process to reach the DL model, a validation process to optimize the test model, a testing process to obtain the raw results, and an evaluating process to translate the experimental result

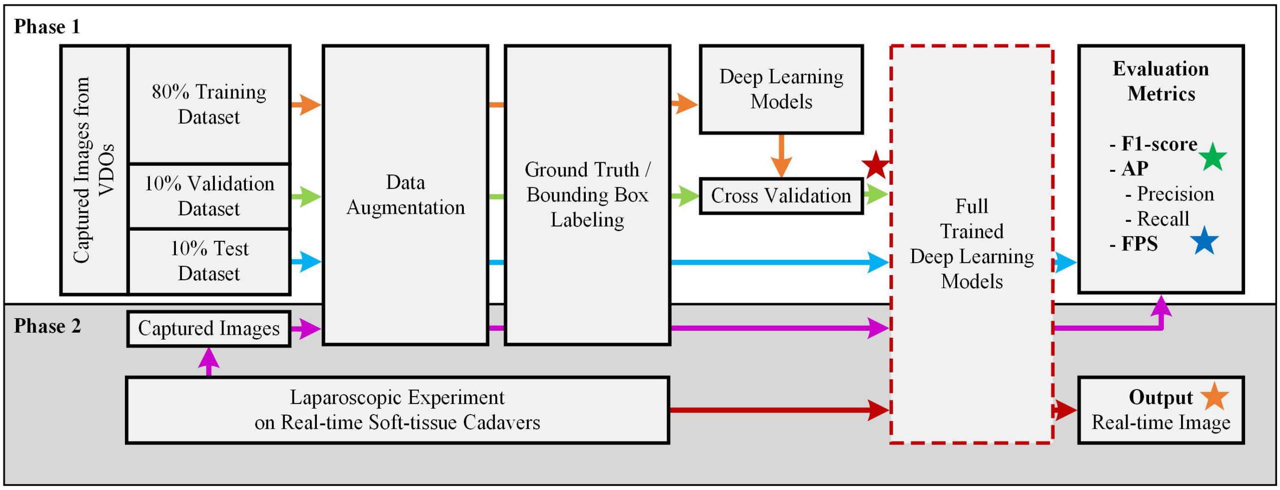


FIGURE 5. Overview of the uterus detection process.

TABLE 1. Dataset used in this work.

Dataset	Training	Validation	Test	Total
42 YouTube VDOs	1,640	205	205	2,050
4 Live-streaming Performed on Cadavers	-	-	100	100

TABLE 2. Dataset augmentation.

Transformation	Characteristics
Rotation	-30°, -15°, 0°, +15°, +30°
Flip	Image Mirroring
Blackbody Temperature	3,400 K
Blur (Gaussian Filter)	Radius = 5

into meaningful information. This architecture is explained in detail in the following subsections.

A. DATA ACQUISITION

The dataset used in the experiment obtained from 2 data sources is as listed in Table 1.

1) 42 PUBLIC VIDEOS

Forty two public videos in Laparoscopic Gynecologic Surgery from Mario Nutis: Public YouTube Channel [20], varying from hysterectomy, myomectomy, removal of ovary and ovarian cyst, etc. with mixed resolutions of 1080p (2 clips), 720p (34 clips) and 540p (6 clips), were used. Retrieved images were selected only when the uterus was not attacked and remained in one piece. Stratified Sampling for splitting a dataset was used to alleviate the problem of Random Sampling in datasets with an imbalanced distribution in each of the training, validation, and test datasets. In total, 2,050 images were used in this work with the ratio of 80:10:10 as training, validating and test data.

2) 4 LIVE-STREAMING VIDEOS

Four live-streaming videos, which later were recorded to obtain 100 captured images as further test datasets, with the resolution of 720p, performing on soft-tissue cadavers at Srinagarind Hospital, Khon Kaen University, Thailand, were used for ultimate blind testing only.

B. DATASET GENERATION

In accordance to the conclusion of a number of studies [21], to increase the size and variety of the dataset, data transformations have been applied on the original images for data augmentation. In total, 2,050 images from public YouTube videos have been augmented and manually annotated. Table 2 lists the characteristics of different types of transformations used in this work.

For the training purpose, we divided the dataset into two sets: a training set of 1,640 images (80%) and a validation set of 205 images (10%). Regardless of **TN**s, the training set contained 64,158 augmented images while the validation set contained 8,076 augmented images. For the testing purpose of the trained models, on top of 8,100 augmented images (205 original test images which are 10%) from the same dataset public YouTube source, an extra set of 4,000 annotated images (100 captured images) were acquired in the same condition from 4 recorded live-streaming videos performing on soft-tissue cadavers. This dataset might have been a by-product from the main research objectives, however sometimes even collecting training images could be difficult. Moreover, there are many legal restrictions for working with healthcare data, and obtaining it requires a lot of effort. Hence this research with its annotated dataset of uterus images could be a contribution for others who would like to exploit it further. The generated dataset can be found at "https://www.kaggle.com/datasets/apiwatboon/laparoscopic-uterus-detection".

C. DESIGN EXPERIMENTS

Two phases of the experiments have been conducted in this work.

1) WITH PUBLIC SURGERY CLIPS

In order to investigate applying DL algorithms on uterus detection, 8 modern object detection algorithms listed in the survey [9] which were also listed in Model Zoo [22], a collection of detection models pre-trained on the COCO 2017 Dataset [23] - Faster R-CNN, SSD, CenterNet, EfficientDet, YOLOv4 (both on DarkNet and PyTorch), YOLOv5 and YOLOv7 algorithms - were chosen in this study with ResNet50 V1 as a common backbone, and similar input size of 512×512 and 640×640. As for the frameworks used for each of the models, the first 4 architectures, namely Faster R-CNN, SSD, CenterNet and EfficientDet were pre-trained on Tensorflow Framework [24], while YOLOv5s and YOLOv7 were pre-trained on Pytorch framework, where there are 2 versions of YOLOv4, pre-trained on DarkNet [25] and Pytorch [26] frameworks.

These 3 frameworks (Tensorflow, Darknet and Pytorch) are very powerful and mature DL libraries with strong visualization capabilities and several options to use for high-level model development. The whole of the training and testing procedures were experimented on Google Colab(oratory) [27] with 16 GB memory. For fine-tuning the pre-trained models, the default values of the pre-training pipeline, adjusting the batch size for the capacity of the available GPU have been considered. The evaluation metrics used are the inference time (in FPS: Frame Per Second), F1-score and AP (Average Precision). All training sessions ran for 3,000 epochs as the experiments with all of the 8 models proved that they did not need more than 3,000 epochs to converge to the best solution in the solution space. Table 3 lists all object detection models with their training information used in this study.

TABLE 3. Training information for each model.

Model	Backbone	Image Size	Batch Size
CenterNet (Tensorflow)	ResNet50 V1	512x512	32
EfficientDet (Tensorflow)	EfficientNet B0	512x512	16
Faster R-CNN (Tensorflow)	ResNet50 V1	640x640	1
SSD (Tensorflow)	ResNet50 V1	640x640	32
YOLOv4 (Darknet)	DarkNet53	416x416	64
YOLOv4 (Pytorch)	DarkNet53	512x512	16
YOLOv5 (Pytorch)	DarkNet53	512x512	16
YOLOv7 (Pytorch)	ELAN	512x512	16

2) WITH SOFT-TISSUE CADAVERS

In order to test how effectively DL models can perform in a real-time surgery, the best detection models, considering both accuracy and speed, from Phase#1 were chosen to run on live-streaming laparoscopic experiments on 4 soft-tissue cadavers.

IV. RESULTS AND DISCUSSION

This section presents the evaluation result of 8 object detection models to detect the uterus in laparoscopic gynecology. As mentioned in Section I-III, the trained models were evaluated using the following evaluation metrics:- Confusion Matrix (**TP**, **TN**, **FP**, and **FN**), **Precision**, **Recall**, **F1-score**, **Precision**×**Recall** curve, **AUC**, **AP** and **FPS**. From the Process in Figure 5, the experimental results have been listed in four sets (labeled with colored stars) as follow:

A. OBTAINING CALIBRATED CONFIDENCE THRESHOLD

Figure 6 shows the evolution of **TP**(6(a)), **FP**(6(b)), and **FN**(6(c)) with the variation of the Confidence Threshold. Prior to proceeding on the evaluation of the performance of object detection models, at the cross validate stage labeled with the red star in Figure 5, defining for the most optimal Confidence Threshold for each model is required, in order to maximize the **F1-score** for the best balance between the **Precision** and **Recall**, optimizing the number of **TP**s while avoiding the **FP**s and **FN**s. The first fact that can be observed here is the number of **TP**s and **FN**s are added up to the number of Ground Truths for each Confidence. As for the results on **TP**s (Figure 6(a) with the higher value, the better) and **FN**s (Figure 6(c)) with the lower value, the better), the most obvious ones to be the winners for the most accurate indications of the presence of the object correctly (**TP**s) and incorrectly (**FN**s), are YOLOv4-P, YOLOv5 and YOLOv7, with the remained models being in the order of; Faster R-CNN, SSD, YOLOv4-D, EfficientDet, and with CenterNet coming last, for both **TP**s and **FN**s.

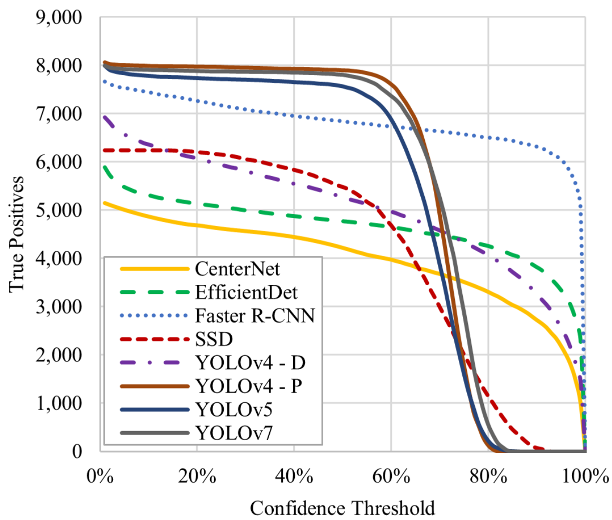
Furthermore, special attention should be given to SSD and Faster R-CNN, for the highest (worst) **FP**s especially during Confidence rates <20%. This particularity leads to misleading fault detecting, indicating the absence of the object incorrectly, and this could incur more serious consequences than completely miss, certainly affecting various evaluation metrics, yielding to low **Precision**, low **F1-score** and perhaps low **AUC**, hence **AP**.

The evolution of the **F1-score** across the increasing Confidence Threshold for cross-validation is illustrated in Figure 7.

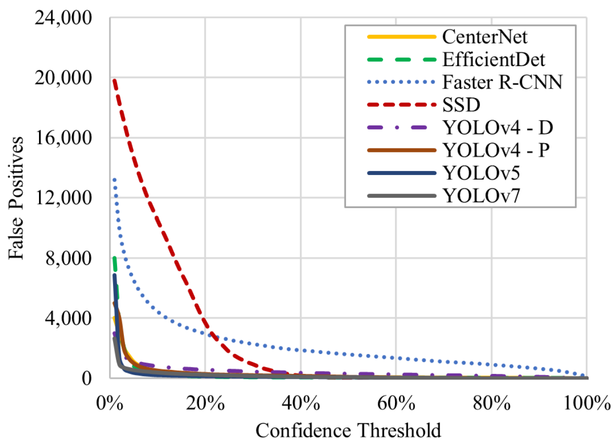
Table 4 listed the highest **F1-score** of each DL object detection model on validation datasets with its corresponding Confidence Threshold.

B. PERFORMANCE EVALUATION

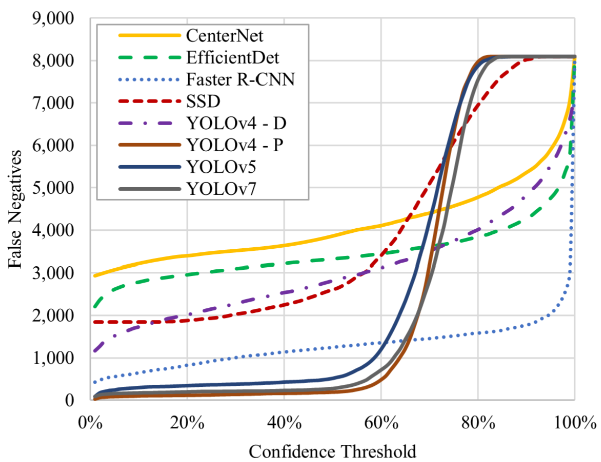
Considering the previously performed process as a pre-filtering process on the validation dataset, later, these



(a) True Positives.



(b) False Positives.



(c) False Negatives.

FIGURE 6. Evolution of the number of TPs, FPs, and FNs with the increase of the confidence threshold.

computed Confidence values were used for fully characterizing the models for object detection purposes (labeled with the green star in Figure 5).

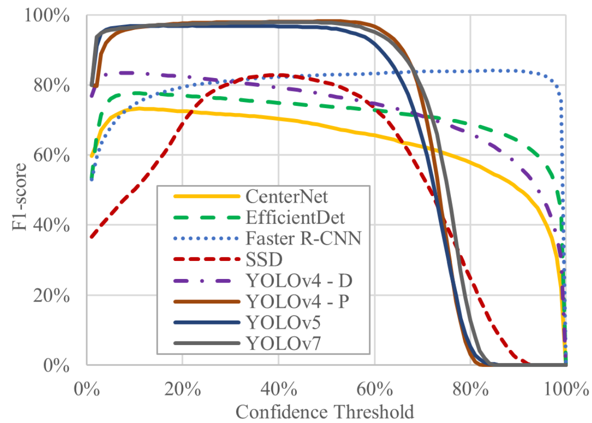


FIGURE 7. Evolution of the F1-score with the variation of the confidence threshold for all DL models on the validation dataset.

TABLE 4. Confidence threshold for each DL model that optimizes the F1-score metric.

Model	F1-score	Confidence
CenterNet (Tensorflow)	73.20%	12%
EfficientDet (Tensorflow)	77.64%	11%
Faster R-CNN (Tensorflow)	84.15%	88%
SSD (Tensorflow)	82.88%	39%
YOLOv4 (Darknet)	83.50%	7%
YOLOv4 (Pytorch)	98.23%	51%
YOLOv5 (Pytorch)	96.92%	20%
YOLOv7 (Pytorch)	97.93%	43%

Two **Precision** × **Recall** curves were built on the test dataset for uterus detection in images of laparoscopic gynecology. These curves established the compromise between the **Precision** and **Recall** rates, considering all the predictions (Confidence Threshold at 0%) (shown in Figure 8) and with calibrated Confidence Thresholds from Table 4 (shown in Figure 9). The difference between these two graphs was the fact that the results in Figure 9 had masked out those with a Confidence rate lower than the chosen calibrated threshold, therefore the final performance evaluation results might not be completely the same with models providing better results continuing to perform better. Whereas YOLO series on Pytorch performed well throughout the wide spectrum, other models' results had dropped dramatically towards the end, resulting in much smaller **AUC**s. As for **F1-score**s, it was obvious and sensible that those of the best Confidence rates provided better results than those of 0% for all cases.

Next, Figures 10-11 depict the results on **Precision**s and **Recall**s, respectively, across all experimented detection models. The test results, corresponding to labeled models, have been organized into sets of 2 bars: on validation datasets vs on blind test datasets. To thoroughly evaluate the performance of a detection model, both **Precision** and **Recall** should be examined. Unfortunately, **Precision** and **Recall**

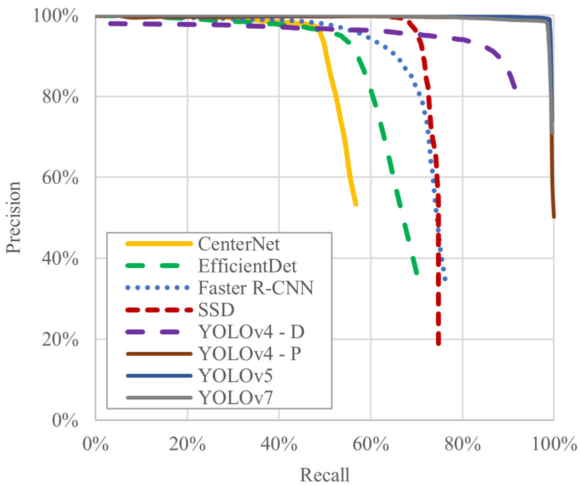


FIGURE 8. Precision × recall curve in the test dataset considering all the predictions.

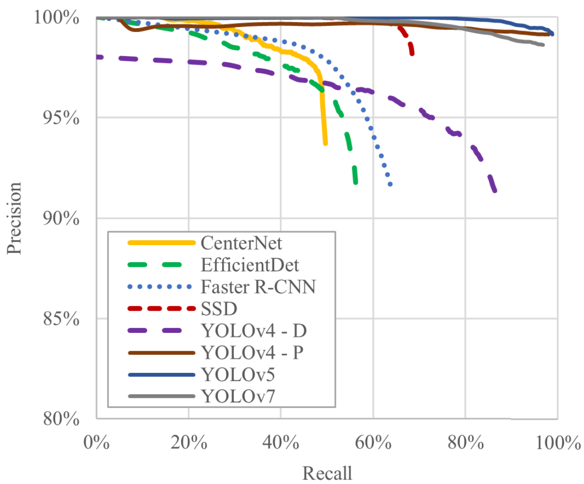


FIGURE 9. Precision × recall curve in the test dataset using the calibrated confidence threshold.

are often in tension. That is, improving **Precision** typically brings **Recall** down and vice versa. Also the case with low **Recall** but high **Precision** implies that all predicted boxes are correct, but most Ground Truths have been missed (high **FN**s), hence a low **F1-score**. Furthermore, with the evolution of the Confidence Threshold, increasing Confidence rates is likely to increase the **Precision** but decrease the **Recall** in their predictions.

Next, the focus was on the results tested with the best Confidence rates on the blind test dataset, which were very close to those on the validation dataset for all conducted experiments here.

Figures 12 and 13 illustrate the results on **F1-scores** and **AP**s, respectively, again in the same format across all experimented detection models. It can be seen that those from blind test datasets are quite similar to those from validation datasets.

Table 5 is the experimental results comparison of accuracy (calculated in **F1-score** and **AP**) vs speed tradeoff (indicating via **FPS**) for each model and its corresponding Confidence

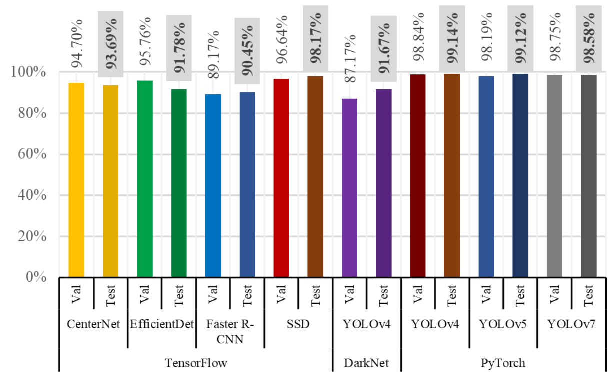


FIGURE 10. Precision.

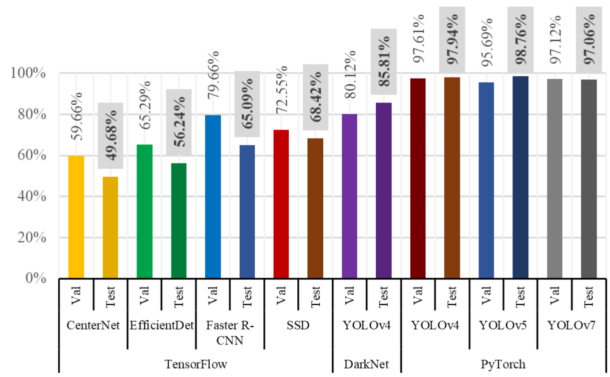


FIGURE 11. Recall.

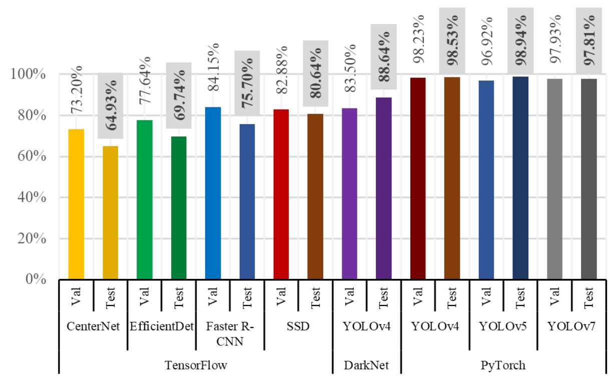


FIGURE 12. F1-score.

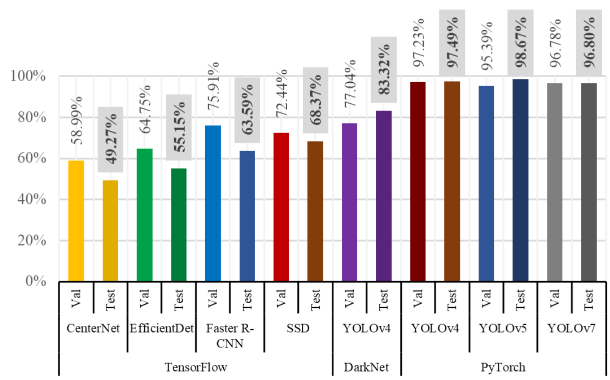


FIGURE 13. AP.

Threshold. Three groups have been categorized, according to each model's best **F1-score** and **AP**, where the performance of CenterNet and EfficientDet were in the poor group

TABLE 5. The F1-score, AP and FPS for each DL model with calibrated confidence threshold.

Model	F1-score	AP	FPS
CenterNet (Tensorflow)	64.93%	49.27%	31
EfficientDet (Tensorflow)	69.74%	55.15%	18
Faster R-CNN (Tensorflow)	75.70%	63.59%	19
SSD (Tensorflow)	80.64%	68.37%	17
YOLOv4 (Darknet)	88.64%	83.32%	31
YOLOv4 (Pytorch)	98.53%	97.49%	66
YOLOv5 (Pytorch)	98.94%	98.67%	142
YOLOv7 (Pytorch)	97.81%	96.80%	111

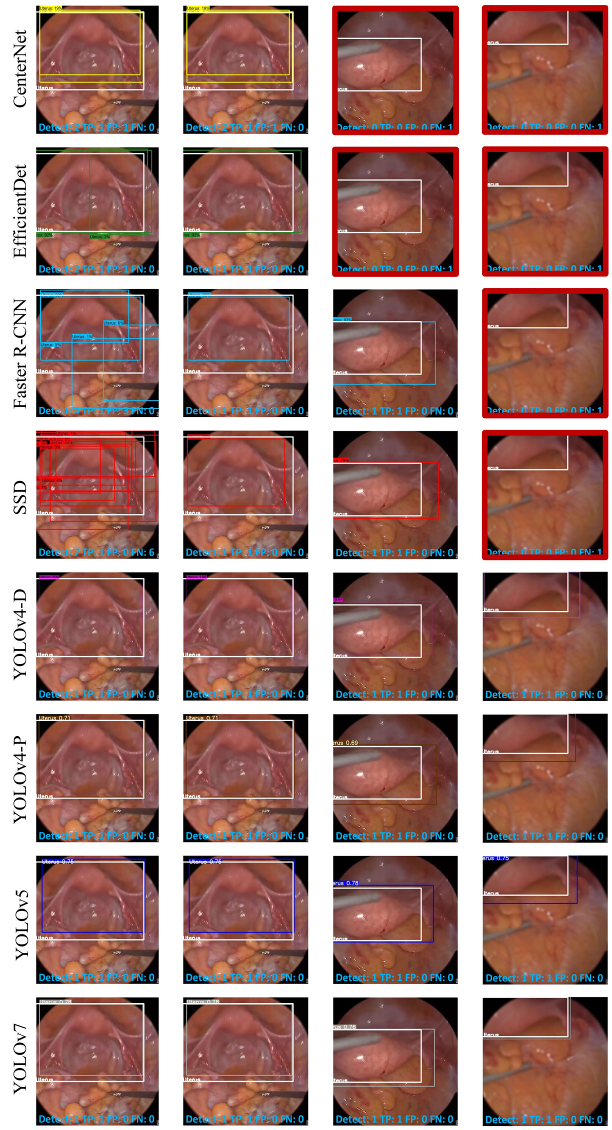
(with F1-score <70% and AP <60%) due to very high **FN**s (Figure 6(c)), which means these two models had so little confidence and failed to agree on detecting objects when there should have been one. While, the performance of Faster R-CNN, SSD and YOLOv4 (DarkNet) fell down in the average group (with **F1-score** <90% and **AP** <85%), the excellent group (with **F1-score** >90% and **AP** >85%) consists of the YOLO series on PyTorch with a great contribution on Albumentation [28], a fast and flexible image augmentation library.

C. SPEED EVALUATION

From the process, labeled with the blue star in Figure 5, speed is essential for real-time operation. Considering most human eyes can perceive between 30 to 60 FPS [29], this means the human brain would process the video streaming as one steady stream, rather than a series of constant flickering lights. As a consequence, three models (in magenta in Table 5), namely EfficientDet, Faster R-CNN and SSD, would not serve the objective of real-time application here, regardless of quite high **F1-score**s for Faster R-CNN and SSD. Whereas the remaining models would be fine as for their adequate **FPS**s. As for those two models resulting in low **F1-score**, CenterNet and EfficientDet, with low **Recall** but high **Precision** it implies that all predicted boxes are correct, but most Ground Truths have been missed (high **FN**s), hence low **F1-score**.

Figure 14 presents examples of the results from four cases performed on all 8 tested models, where the column 14(a) is with Confidence Threshold >0%, column 14(b) is of the same image but with the calibrated Confidence Thresholds providing the best **F1-score** (as listed in Table 5) for each model, while columns 14(c) and 14(d) are also with the same calibrated Confidence rates but of different images. Generally speaking, with the best Confidence rate for each model, the detection became more accurate, yielding better **FP**s and as a result better **F1-score**.

It can be seen in column 14(a) that a number of models seemed to be over detected on the same target, hence after getting one correct (=1 **TP**) the rest resulted in high **FP**s.



(a) #1 Before (b) #2 After (c) #3 (d) #4

FIGURE 14. Examples of test results using filtered images (calibrated confidence threshold).

Not only in the pictures shown here, but for a larger number of cases, showing that SSD and Faster R-CNN (also, but not so obvious, CenterNet and EfficientDet) seemed to jump to the conclusion of the detection too easily, therefore very high **FP**s (Figure 6(b)) especially at the beginning of the growth of Confidence Thresholds. Lastly, columns 14(c) and 14(d) show the detection failure (framed in red) with each model.

D. PERFORMING ON SOFT-TISSUE CADAVERS

After performing and analyzing the experimental results on the test dataset as presented earlier, these detection models were conducted with these four live feeds from the laparoscope on soft-tissue cadavers (labeled with the orange star in Figure 5). Figure 15 shows the set up and the environment in the operation room. Ethical approval #HE641206 for this study was waived by the Center for Ethics in Human

FIGURE 15. Laparoscopic surgery experiments on soft-tissue cadavers.

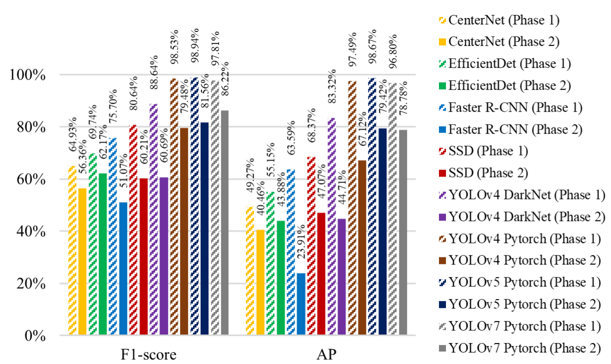


FIGURE 16. F1-score and AP.

Research, Khon Kaen University (KKU), on May 13th, 2021, because this work has been conducted on cadavers donated for educational research to Srinagarind Hospital, KKU, Thailand.

As for the speed, it can be clearly seen that EfficientDet, Faster R-CNN and SSD struggled with this 25 **FPS** live-streaming, while the rest could keep up rather well. This observation was consistent with the previous result in Table 5.

Later, images from these laparoscopic experiments were captured then augmented, annotated, and put into the test with those trained models to evaluate further.

The comparisons between Confidence of 0% (Phase#1) vs the best Confidence Threshold (Phase#2) from the previous test for both **F1-score**s and **AP**s with these blind captured images are illustrated in Figure 16. Though all of the results from this experiment phase might have been quite low compared to those of the first phase, that was because the used models in this phase were the trained models with Confidence Threshold from the training with public images which was in a completely different environment to the experiments in the phase.

Thus, it can be concluded that YOLO series running on Pytorch have performed excellently for all evaluation metrics: **F1-score**, **AP** and also the speed (in **FPS**), meeting the requirement of real-time application. The rest failed completely for the speed and/or offered rather low performance.

V. CONCLUSION

This paper has presented the evaluation results on various Deep Learning models covering state-of-the-art object detection models, from two-stage detectors to one-stage detectors. Those models were run on 3 different frameworks: TensorFlow, pre-trained on the COCO 2017 Dataset (Faster R-CNN, CenterNet, EfficientDet, SSD); DarkNet (YOLOv4-D) and Pytorch (YOLOv4-P, YOLOv5 and YOLOv7). The target object in this study was the uterus in laparoscopic gynecology. It is injudicious to compare results shoulder-to-shoulder from different papers, as those experiments are undertaken in different settings or have different targets which are not purposed for direct comparisons.

To begin with this study, an annotated dataset of the uterus must be generated. The first contribution of this paper on the uterus dataset is not just in respect of the captured images from 42 existing public YouTube videos with the ratio of 80:10:10 for training: validating and testing, but also from 4 live-streamings, operated on soft-tissue cadavers, which is considered the closest environment to the real human body.

The second contribution must be the comparison results among 8 different cutting-edge object detection models for real-time live feeds. The most important question is not which detector is the best in accuracy performance, and which may also not be possible to answer. The real question is which detector and what configurations provide the best balance of speed and accuracy that designated real-time application needs. All of the results pointed out that the YOLO models running on Pytorch performed excellently for both accuracy and speed aspects, whereas EfficientDet, Faster R-CNN and SSD offered rather too low inference time, which cannot meet the needs of real-time detection.

Last by no means least, as for the third contribution, this paper not only applied Deep Learning for uterus detection and tested on the dataset captured from public videos, but also, in order to prove how effective these detection models can perform in real surgery, all 8 models have been conducted with real-time streaming feeds during laparoscopic surgeries on 4 different soft-tissue cadavers. With a complete difference in the setup and experimental environment, it is quite understandable that the results from the completely blind experiments on soft-tissue cadavers have deteriorated. Although the results from this experiment Phase#2 demonstrated lower performance for all models, the ultimate results remained the same for YOLO model series on the Pytorch framework for being the most efficient models for both speed and accuracy performances, with the obvious flickering streaming for EfficientDet, Faster R-CNN and SSD, which suggests that these three models might not be implemented in applications that require real time. The observation from real-time operations has also been confirmed by the final experiment on captured images from those 4 live feeds as another test dataset.

The limitation of this entire research was in detecting the regular uterus and even irregular uterus with different colors/textures and sizes from both public surgeries clips and cadaver surgeries but as a whole piece. The evaluation results showed a number of DL models worked excellently and effectively for real-time operations. However, DL not only is capable of performing object detection, but also could be exploited to detect features of target objects as well, i.e. detecting separate parts or poses of the object even in occluded condition. Therefore, this issue still remains a notable challenge to investigate.

ACKNOWLEDGMENT

The authors would like to thank those who donated their bodies so that anatomical study and research could be performed to potentially increase overall knowledge that can then improve patient care. Therefore, these donors and their families deserve their highest gratitude.

REFERENCES

[1] K. Mishra, R. Sathish, and D. Sheet, "Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2233–2240.
[2] R. Anteby, N. Horesh, S. Soffer, Y. Zager, Y. Barash, I. Amiel, D. Rosin, M. Gutman, and E. Klang, "Deep learning visual analysis in laparoscopic surgery: A systematic review and diagnostic test accuracy meta-analysis," *Surgical Endoscopy*, vol. 35, no. 4, pp. 1521–1533, Apr. 2021.
[3] B. Harangi, A. Hajdu, R. Lampe, and P. Torok, "Recognizing ureter and uterine artery in endoscopic images using a convolutional neural network," in *Proc. IEEE 30th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2017, pp. 726–727.
[4] E. Gibson, M. R. Robu, S. Thompson, P. E. Edwards, C. Schneider, K. Gurusamy, B. Davidson, D. J. Hawkes, D. C. Barratt, and M. J. Clarkson, "Deep residual networks for automatic segmentation of laparoscopic videos of the liver," *Proc. SPIE*, vol. 10135, pp. 423–428, Mar. 2017.
[5] S. Petscharnig, K. Schöffmann, J. Benois-Pineau, S. Chaabouni, and J. Keckstein, "Early and late fusion of temporal information for classification of surgical actions in laparoscopic gynecology," in *Proc. IEEE 31st Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2018, pp. 369–374.
[6] A. Leibetseder, S. Petscharnig, M. J. Primus, S. Kletz, B. Münzer, K. Schoeffmann, and J. Keckstein, "LapGyn4: A dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology," in *Proc. 9th ACM Multimedia Syst. Conf.*, Jun. 2018, pp. 357–362.
[7] S. M. Zadeh, T. Francois, L. Calvet, P. Chauvet, M. Canis, A. Bartoli, and N. Bourdel, "SurgAI: Deep learning for computerized laparoscopic image understanding in gynaecology," *Surgical Endoscopy*, vol. 34, no. 12, pp. 5377–5383, Dec. 2020.
[8] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019.
[9] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digit. Signal Process.*, vol. 126, Jun. 2022, Art. no. 103514.
[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[11] W. Liu, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, vol. 9905, Dec. 2015, pp. 21–37.
[12] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6568–6577.
[13] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.
[14] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," Apr. 2020, *arXiv:2004.10934*.
[15] *YOLOv5*. Accessed: Feb. 24, 2023. [Online]. Available: https://github.com/ultralytics/yolov5
[16] C. Wang, A. Bochkovskiy, and H. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," Jul. 2022, *arXiv:2207.02696*.
[17] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Jul. 2020, pp. 237–242.
[18] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating trust prediction and confusion matrix measures for web services ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020.
[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[20] M. Nutis. *Mario Nutis YouTube*. Accessed: Jun. 9, 2022. [Online Video]. Available: https://www.youtube.com/c/mnutis/videos
[21] S. A. Magalhães, L. Castro, G. Moreira, F. N. dos Santos, M. Cunha, J. Dias, and A. P. Moreira, "Evaluating the single-shot MultiBox detector and YOLO deep learning models for the detection of tomatoes in a greenhouse," *Sensors*, vol. 21, no. 10, p. 3569, May 2021.
[22] *TensorFlow 2 Detection Model Zoo*. Accessed: Jun. 22, 2022. https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf2_detection_zoo.md
[23] T.-Y. Lin, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 8693, Sep. 2014, pp. 740–755.
[24] *TensorFlow*. Accessed: Jun. 22, 2022. [Online]. Available: https://www.tensorflow.org
[25] *Darknet Framework*. Accessed: Jun. 22, 2022. [Online]. Available: https://pjreddie.com/darknet
[26] *PyTorch*. Accessed: Feb. 24, 2023. [Online]. Available: https://pytorch.org
[27] *Google Colab*. Accessed: Jun. 22, 2022. [Online]. Available: https://colab.research.google.com
[28] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020.
[29] F. Pazhoohi and A. Kingstone, "The effect of movie frame rate on viewer preference: An eye tracking study," *Augmented Hum. Res.*, vol. 6, no. 1, pp. 1–5, Dec. 2021.

APIWAT BOONKONG received the B.Eng. and M.Eng. degrees in computer engineering from Khon Kaen University, Thailand, in 2012 and 2017, respectively, where he is currently pursuing the Ph.D. degree in machine learning with an application in laparoscopic gynecologic surgery.

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.