applied sciences

MDPI

Article

Surgical Tool Detection in Open Surgery Videos

Ryo Fujii [1,*], Ryo Hachiuma [1], Hiroki Kajita [2] and Hideo Saito [1]

[1] Faculty of Science and Technology, Keio University, Yokohama 223-8852, Japan
[2] Department of Plastic and Reconstructive Surgery, Keio University School of Medicine, Tokyo 160-8582, Japan
* Correspondence: ryo.fujii0112@keio.jp

Abstract: Detecting surgical tools is an essential task for analyzing and evaluating surgical videos. However, most studies focus on minimally invasive surgery (MIS) and cataract surgery. Mainly because of a lack of a large, diverse, and well-annotated dataset, research in the area of open surgery has been limited so far. Open surgery video analysis is challenging because of its properties: varied number and roles of people (e.g., main surgeon, assistant surgeons, and nurses), a complex interaction of tools and hands, various operative environments, and lighting conditions. In this paper, to handle these limitations and difficulties, we introduce an egocentric open surgery dataset that includes 15 open surgeries recorded with a head-mounted camera. More than 67k bounding boxes are labeled to 19k images with 31 surgical tool categories. Finally, we present a surgical tool detection baseline model based on recent advances in object detection. The results of our new dataset show that our presented dataset provides enough interesting challenges for future methods and that it can serve as a strong benchmark to address the study of tool detection in open surgery.

Keywords: surgical tool detection; open surgery; egocentric camera; surgical video analysis; deep neural network

Citation: Fujii, R.; Hachiuma, R.; Kajita, H.; Saito, H. Surgical Tool Detection in Open Surgery Videos. *Appl. Sci.* **2022**, *12*, 10473. https://doi.org/10.3390/app122010473

Academic Editor: Valeriu Surlin

Received: 11 September 2022
Accepted: 7 October 2022
Published: 17 October 2022

1. Introduction

Automated analysis of surgery videos has been indispensable for collecting information about surgery. Combined with computer vision and machine learning, it can improve performance and contribute to enhancing surgical safety and efficiency. A great deal of attention to surgical video analysis has sparked active research in the medical computer vision community.

Most research on surgery video analysis has been conducted in the area of minimally invasive surgery (MIS) and cataract surgery. MIS has especially gained great attention, and many tasks have been proposed to analyze surgery, such as workflow analysis [1], phase recognition [2], type recognition [3], video segmentation [4], and video summarization [5].

Here, we focus on surgical tool detection, which can provide an estimation of the identification and position of each surgical tool that appears in a frame (Figure 1). This task is fundamental for recognizing the surgical scene. It can be used for a lot of downstream applications, such as tool tracking [6–8], tool pose estimation [9–11], prediction of the remaining surgery duration [12], and surgical technical skill assessment [13]. Surgical tool detection has also been well investigated in MIS [13–16] and cataract surgery [17] as there are various datasets available in public.

However, tool detection in the open surgery field has been limited because of the dataset collection and privacy-preserving costs. In MIS and cataract surgery, the surgeon sees the surgery through an endoscope camera, and videos can be easily recorded. The recordings only include surgical fields and tools and do not contain information that may identify individuals. This enables large and well-annotated datasets in MIS [18–20] and cataract surgery [17]. On the other hand, in open surgery, the surgeon sees the surgery with their own eyes. Thus, head-mounted cameras or cameras in the operating room are needed for recording videos. The recordings may include surgeon and patient faces and

information that may identify individuals. Thus, the datasets annotated for open surgery are small and often collected in simulations [21–23], which is difficult to be transferred for use in real-world surgery. Moreover, the computer vision task in open surgery has challenging inherent properties: varied number and roles of people (e.g., surgeons, anesthesiologists, perfusionists, nurses), the complex interaction of tools and hands, various operative environments, and lighting conditions. Hence, a diverse and large-scale dataset is indispensable.

In this paper, we introduce a large-scale egocentric open surgery dataset that contains densely annotated bounding boxes and their categories of tools. Compared with the conventional open surgery dataset for tool detection [22–25], our dataset contains a tremendous number of annotated frames and is captured at an actual surgery scene. The exampled images in the dataset are shown in Figure 1. Our data were collected by six surgeons in real-world open surgery at our university's school of medicine. We recorded 15 videos of seven different types of surgery. Different from MIS and cataract surgery, there are several choices of the recording viewpoint (e.g., view from the head-mounted camera, the camera attached to the light [26], the surveillance camera). We choose a first-person view as the viewpoint of recording because the data from an egocentric viewpoint are suitable for capturing the details during the surgery. The dataset has 15 h of recording from a camera attached to the surgeon's head, densely annotated bounding boxes, and the corresponding category of the surgical tool. In total, there are 19,560 frames with the bounding boxes of the 31 categories, which is relatively large-scale compared to conventional open surgery datasets.

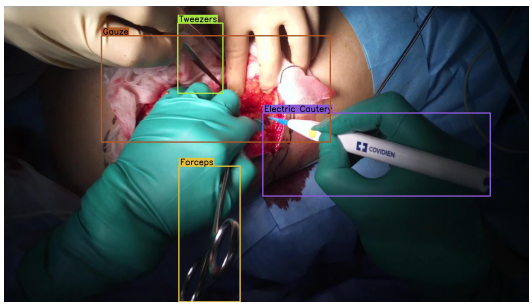


Figure 1. An example of surgical tool detection in an open surgery video.

We provide an extensive evaluation of the conventional object detection methods that have been trained with our proposed dataset. Because only a few prior works have tackled object detection with large-scale open surgery datasets, we employ two well-known object detection methods—Faster R-CNN [27] and RetinaNet [28] —with different backbones.

2. Related Work

2.1. Surgical Tool Detection in Mis and Cataract Surgery

Surgical tool detection has been worked on using various approaches. These include sensor-based methods such as radiofrequency identification (RFID) [29] or endoscopic image-based methods. The most conventional image-based methods rely on the defined handcrafted features, such as color [30,31], gradient [32], and texture [33].

With the rapid development of convolutional neural networks (CNN), most surgical tool detection algorithms started to utilize deep learning techniques. Twinanda et al. [18] introduced a baseline model called EndoNet that performs both tool presence detection and phase recognition tasks in a multi-task manner. Choi et al. [14] employed the YOLO architecture [34], which can directly predict the location of tools as a representation of the boundary box in real-time but has no outstanding accuracy. Sarikaya et al. [15] utilized a region proposal network and a multi-modal two-stream convolutional network for the surgical tool detector. Kurmann et al. [9] proposed a U-Net architecture-based model performing tool detection and 2D pose estimation jointly. A. Jin et al. [13] achieved high detection accuracy using a region-based CNNs (R-CNNs) [27].

Moreover, many approaches that consider the long-term temporal information and model the temporal dependency between the frames have been proposed. Particularly, a method combining CNN with a recurrent neural network (RNN) has been a trend [6,35,36]. Al Hajj et al. [36] applied a CNN-RNN model to detect a tool used in the surgical videos; they employed a boosting mechanism instead of end-to-end training. Nwoye et al. [6] developed an end-to-end approach composed of CNN + convolutional LSTM (ConvLSTM) neural networks that can perform tool presence detection and tool tracking using tool binary labels. Wang et al. [37] proposed taking advantage of both 3D CNNs and graph convolutional networks (GCNs) for tool presence detection, thereby considering the relationship between tools. Y. Jin et al. [38] presented a multi-task recurrent convolutional network with correlation loss (MTRCNet-CL) for tool presence detection and surgical phase recognition.

2.2. Computer Vision Research in the Open Surgery Domain

A small body of research exists in the open surgery domain. Most works have been related to recording techniques. Shimizu et al. [26] proposed a novel surgical recording system using multiple cameras mounted on a surgical lump where computer vision-based region segmentation and recognition techniques are applied to automatically select the camera with the best view, hence producing a single video without occlusion. Hachiuma et al. [39] improved this camera selection algorithm using CNNs. Saito et al. [40] presented self-supervised learning for camera selection by taking advantage of a first-person view. Yoshida et al. [41] estimated the incision scenes in long-duration open surgery videos using learning gaze speed, hand movements, number of hands, and background movements in egocentric surgical videos.

Few studies have focused on research related to surgical tool detection in open surgery videos [22–25,42]. Ref. [42] performs operating hands detection and tracking in open surgery videos. Ref. [24] proposed detecting two similar surgical tools using hand information in an open surgery video recorded by an egocentric camera. Basiev et al. [22] performed surgical tool detection using a multi-camera setting that can help prevent the surgical tool from being totally invisible in open surgery videos, where occlusion often occurs. In the same approach as MIS, the multi-task framework has been often utilized in open surgery video analysis. Goodman et al. [25] proposed a multi-task network for action segmentation, tool detection, hands detection, and hand pose keypoints estimation. Goldbraikh et al. [23] proposed a multi-task network that solves both tool localization and tool hand interaction detection task.

2.3. Surgical Dataset

Because a large-scale dataset is essential for the improvement of the data-driven deep learning algorithms, a dataset for surgical tool detection in the open surgery domain is not available; however, several datasets in other surgical domains for various tasks have been released in recent years.

Most datasets released so far are related to MIS or cataract surgery. The Cholec80 dataset [18] consists of endoscopic videos of 80 cholecystectomy procedures labeled with phase and tool presence annotations. There are seven categories of surgical tools, including grasper, hook, clipper, bipolar, irrigator, scissors, and specimen bag. Some of the videos in Cholec 80 are included in the M2CAI16-tool dataset [19], which consists of 15 cholecystectomy videos with ground truth binary annotations of the tools. A. Jin et al. [13] introduced a new dataset, M2CAI16-tool-locations, which extends the M2CAI16- tool dataset with the coordinates of spatial bounding boxes around surgical tools. In the same way, Shi et al. [43] extended the Cholec80 dataset [18] with spatial annotations of surgical tool. ATLAS Dione [15] consists of 99 action video clips recorded in a simulated surgical scene of 10 surgeons performing six surgical tasks on the da Vinci Surgical System (dVSS) with surgical tool annotations. The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [44,45] was released as the first public dataset containing video and kinematic

recordings of robotic surgical demonstrations, as well as gesture and skill annotations. Bawa et al. [20] introduced the ESAD dataset for surgeon action detection in real-world endoscopic videos. This dataset is annotated with a surgical tool bounding box and its action label. The CATARACTS dataset [17] contains 50 videos of cataract surgeries, which involves removing a clouded natural lens and replacing it with an artificial lens. Here, the presence of 21 surgical tools was manually annotated by two experts.

Recently, some datasets for open surgery video analysis have been proposed to perform object detection. Surgery hands [42] was developed for hand detection in open surgery videos composed of publicly available videos of open surgery collected from YouTube and the annotation of spatial bounding boxes of the surgeon's hands. Shimizu et al. [24] recorded seven different types of actual open surgery with an egocentric camera for detecting two similar surgical tools: scissors and needle holders. Basiev et al. [22] collected the videos taken from two different angles simulating injured open bowel repairing surgery for a tool and hand detection task. These videos contain four surgical tools: needle holder, forceps, scissors, and mosquito forceps. Another open surgery dataset is Annotated Videos of Open Surgery (AVOS), which contains 1997 videos scraped from YouTube and annotated with bounding boxes for surgical tools, including electrocautery, needle drivers, forceps, and hands, and 21 joint key points for hands and action. Goldbraikh et al. [23] introduced a dataset recorded by the open surgery suturing simulation system and provided tool bounding box and tool and hand bounding box annotations including three surgical tools: needle drivers, forceps, and scissors. In these existing open surgery datasets, unlike ours, the videos are often collected from a simulator [22,23], and if the videos are collected from actual open surgery, they only provide, at most, four types of surgical tool class annotations [24,25].

Behavioral analysis during surgery also has many potential applications for research, quality improvement, and education. The Multi View Operating Room (MVOR) dataset [46] is the first multi-view pose estimation dataset generated from real surgery recordings obtained in an operating room (OR) using three different views. The dataset has been manually annotated to provide both 2D and 3D upper-body poses.

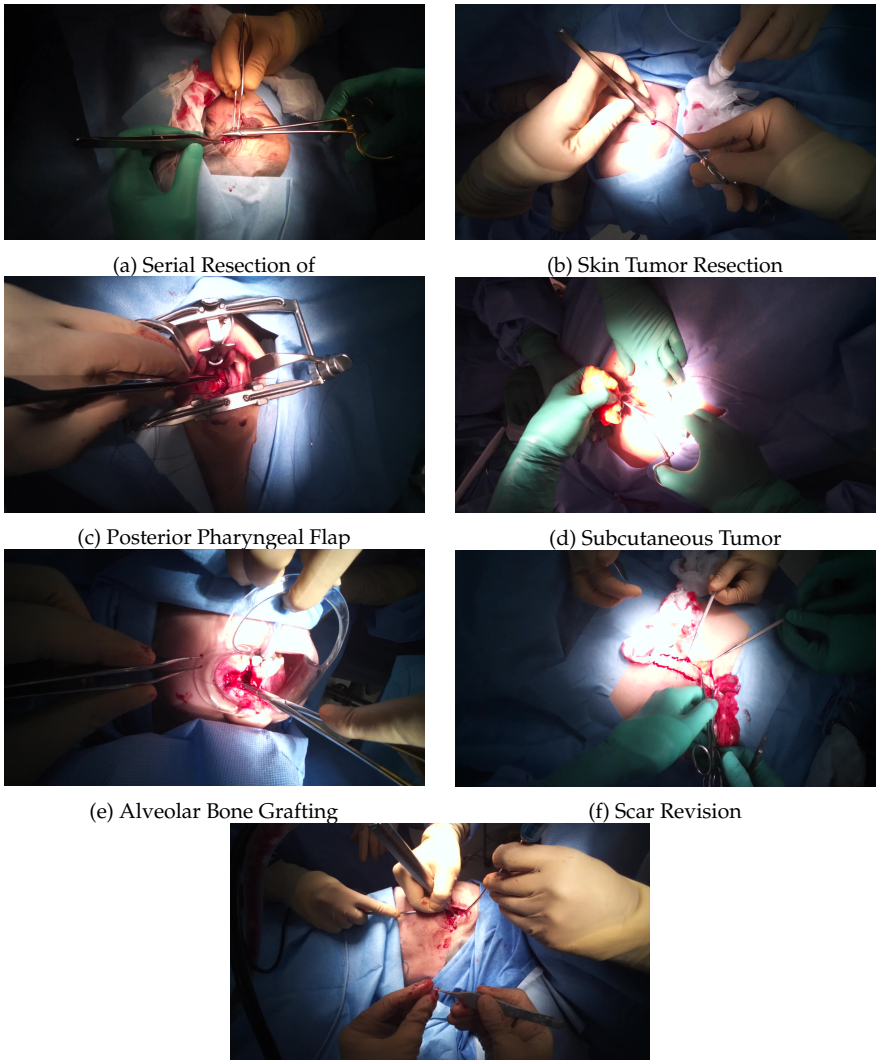
3. Dataset

In this section, we present the details of the collection, annotation, statistics, and quality of the dataset, respectively.

3.1. Dataset Collection and Annotation

Because there is no dataset available that contains open surgery recordings via an egocentric camera, we record our own dataset. The actual plastic surgeries were recorded using Tobii cameras attached to the surgeon's head. The surgeries were recorded at Keio University Hospital. Video recording of the patients was approved by the Keio University School of Medicine Ethics Committee, and written informed consent was obtained from all patients or their parents.

Our dataset contains 15 videos of seven different types of surgery, including serial resection of skin lesions, skin tumor resection, posterior pharyngeal flap, subcutaneous tumor resection, alveolar bone grafting, scar revision, and open reduction and internal fixation, which were performed by six surgeons. Images from each type of open surgery video are shown in Figure 2. Each surgery video is about 30 to 90 min long and was recorded at 25 frames per second (FPS). The videos are downsampled to 1 fps for processing to avoid frames where the scene changes very little. Each instance is annotated using a bounding box and its surgical tool category label. The frame size of each video is 1920 $\times$ 1080 (pixels).



(a) Serial Resection of (b) Skin Tumor Resection

(c) Posterior Pharyngeal Flap (d) Subcutaneous Tumor

(e) Alveolar Bone Grafting (f) Scar Revision

(g) Open Reduction and Internal Fixation

Figure 2. Examples of each type of open surgery in seven different types of surgery.

We conducted the annotation using the coordinates of spatial bounding boxes around the tools using eight graduate students in the Department of Information and Computer Science; all annotations have been examined by an expert surgeon. We used the Virtual object Tagging Tool (VoTT) for annotation. VoTT is a Microsoft open-source tool for the annotation and labeling of image and video assets. A screenshot of its graphical user interface is shown in Figure 3.

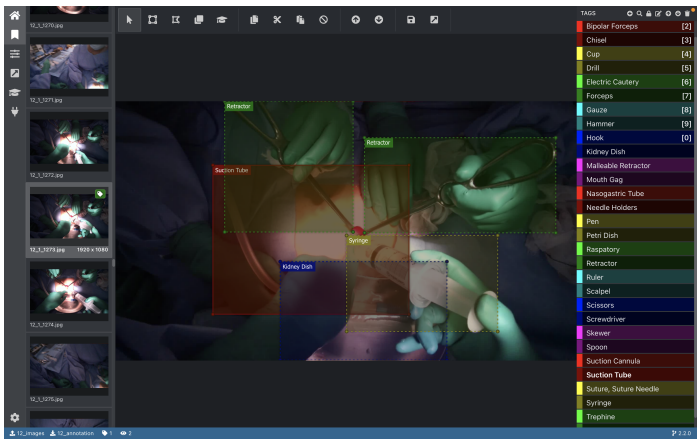


Figure 3. Screenshot captured while using VoTT for annotation.

All surgical tool categories that appeared in the recording video are annotated. Then, we eliminated the surgical tool categories that appear less than 10 times. As a result, our dataset contains 31 surgical tool categories for initial annotation. The list of categories is shown in Table 1, and examples of each surgical tool category are shown in Figure 4. We have focused on the annotation of the surgical tool in use and excluded the annotation of the surgical tool not in use (e.g., surgical tools placed on the operating table).

Table 1. List of surgical tool categories in our dataset, with the number of instances in each of the training, validation, and test sets.

Category	Training Set	Validation Set	Test Set	Total
BiClamp	297	0	0	297
Bipolar Forceps	453	55	205	713
Chisel	42	0	11	53
Cup	65	18	11	94
Drill	75	0	0	75
Electric Cautery	1460	101	162	1723
Forceps	2885	155	4038	7078
Gauze	5480	469	1921	7870
Hammer	26	0	2	28
Hook	1114	191	167	1472
Kidney Dish	33	0	34	67
Malleable Retractor	121	0	0	121
Mouth Gag	4596	1208	1325	7129
Nasogastric Tube	1672	0	0	1672
Needle Holders	3449	531	1327	5307
Pen	21	0	0	21
Petri Dish	106	0	44	150
Raspatory	750	86	105	941
Retractor	3423	45	364	3832
Ruler	14	0	0	14
Scalpel	784	168	173	1125
Scissors	1869	422	620	2911
Screwdriver	79	0	0	79
Skewer	212	103	29	344
Spoon	27	0	0	27
Suction Cannula	3911	622	827	5360
Suction Tube	24	0	46	70
Suture, Suture Needle	3979	419	2023	6421
Syringe	347	96	144	587
Trephine	6	0	34	40
Tweezers	8114	1119	2833	12,066
Total Object Instances	45,434	5808	16,445	67,687

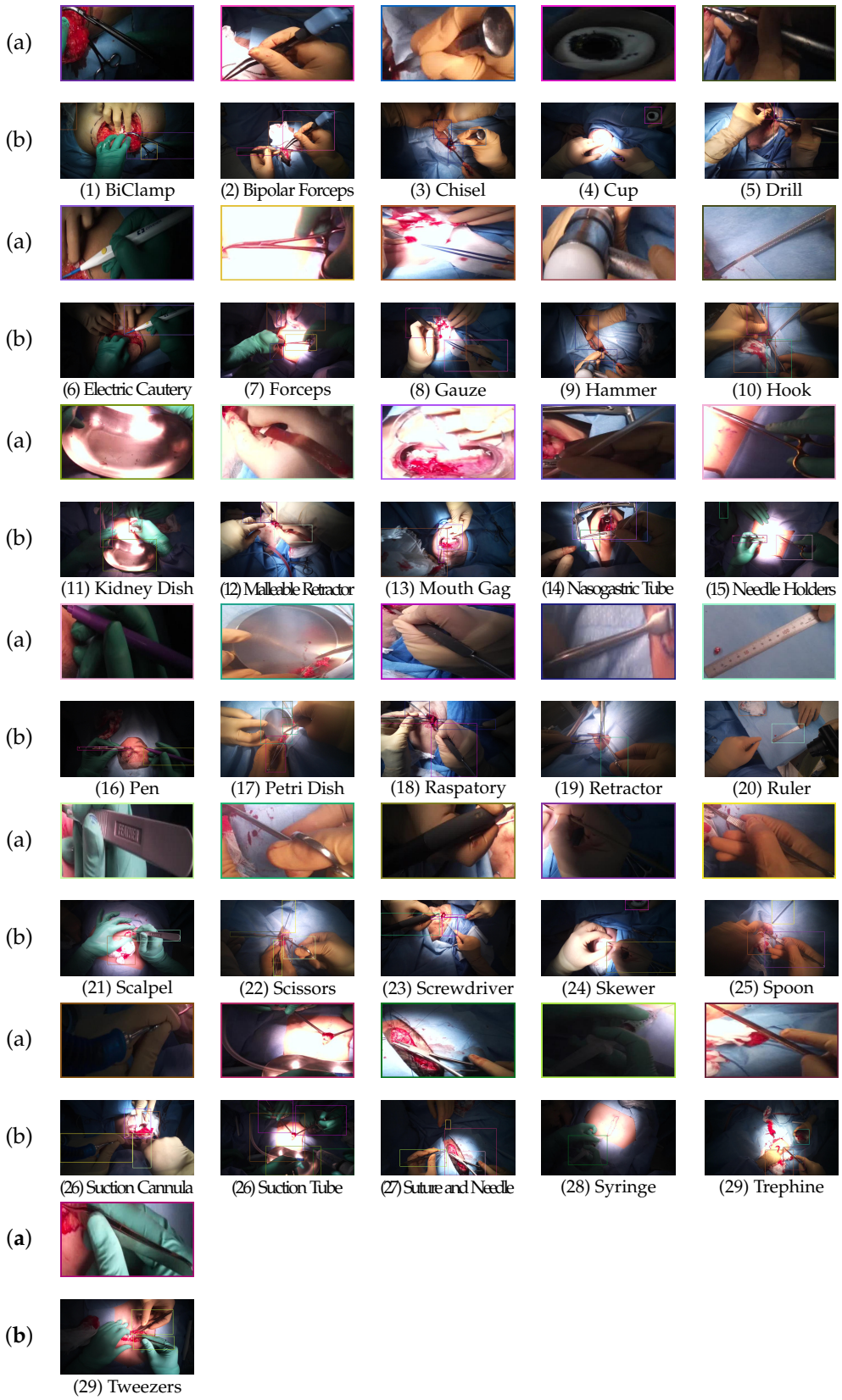


Figure 4. (**a**) Examples of the 31 surgical tools in our open surgery dataset; (**b**) Example frames with their spatial tool annotations. Color of the bounding box corresponds to tool identity.

Surgical tool detection in open surgery exhibits several specific features. First, in open surgery, the surgical tools are severely occluded by the surgeon's hand or other surgical

tools (Figure 5). Second, even though there is no occlusion of the tools, it is difficult to classify the tools because some have similar shapes and textures (Figure 6). These features make surgical tool detection in open surgery challenging.

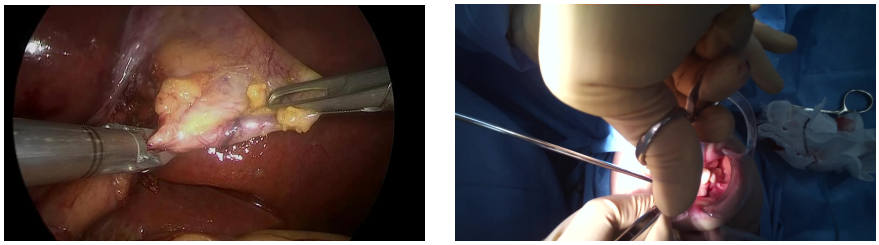


Figure 5. Comparison of typical occlusion between the images of endoscopic surgery [13] (**left**) and open surgery (**right**). In endoscopic surgery, the tools are often occluded by tissues. On the other hand, in open surgery, the tools are often occluded by the surgeon's hands and tools.

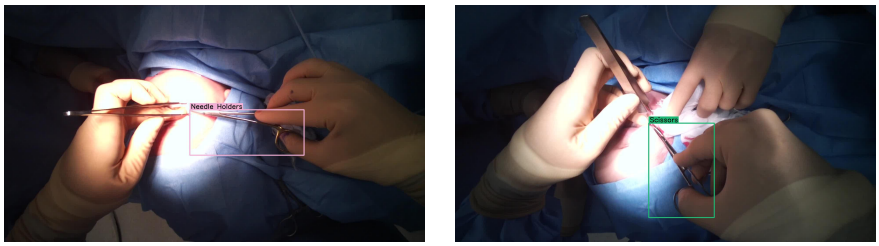


Figure 6. An example of two different surgical tools that have a similar shape and texture.

3.2. Statistics

Based on the setting of Section 3.1, we annotated 19,560 images among the 27,208 total frames. As a result, 13,537 instances are annotated in our dataset. The size of the dataset is comparable, and the number of annotated instances is the largest, even in the existing MIS surgical tool detection dataset, as shown in Table 2.

Table 2. A survey table of the surgical tool detection dataset and a comparison with our proposed dataset.

Dataset	Real Env.?	Bbox Annotated?	Number of Annotated Frames	Number of Annotated Instances	Avg. Instances per Frame	Number of Surgical Tool Categories
Minimally invasive surgery (MIS):			-	-	-	-
m2cai16-tool [19]	✓		23,000	-	-	1
Cholec80 [18]	✓		86,000	-	-	7
ATLAS Dione dataset [15]		✓	22,467	43,227	1.9	1
m2cai16-tool-locations [13]	✓	✓	2532	3141	1.2	7
Cholec80-locations [43]	✓	✓	4011	6471	1.6	7
Cataract surgery:						
CATARACTS dataset. [17]	✓		957,884	-	-	21
Open surgery:						
Shimuzu et al. [24]	✓	✓	2300	-	-	2
Basiev et al. [22]		✓	11,500	-	-	4
Goldbraikh et al. [23]		✓	1124	-	-	3
AVOS dataset [25]	✓	✓	3348	2843	0.85	3
Ours	✓	✓	19,560	67,687	3.5	31

3.3. Data Splits

The dataset was divided into a training set (10 videos), a validation set (two videos), and a test set (three videos). There are 12,715 images for a training set (65%), 1807 images for a validation set (9%), and 5038 images for a test set (26%). The distribution of the number of instances per surgical tool category in the three splits is shown in Figure 7. Because not all categories appear in all videos, we ensured all categories were present in the training set. We split the rest of the videos between the validation and test sets. The distribution of the category in each set is different because data are split by the *video-based split* , as shown in Figure 7. Another way to perform a split, the *frame-based split*, can generate a more uniform distribution among the three sets. However, the *frame-based split* would have the

model be trained and evaluated on almost the same data. In real-world surgery, the data distribution drastically change every surgery, or video due to the various surgery type, operative environments, and lighting conditions. To tackle the tool detection in these challenging settings, we chose *video-based split*.

Figure 7 indicates that the dataset is highly imbalanced; consequently, accurate instrument classification is more challenging. *Tweezers* contains the largest number of instances because surgeons use them commonly for all procedures. On the other hand, *Ruler* contains the smallest number of instances, as surgeons use it only for specific purposes (e.g., measuring tumors). Furthermore, there are other visual challenges because of the high inter-category similarity among tools. For instance, *Forceps*, *Needle Holders*, and *Scissors* categories have similar appearances, as shown in Figure 4.

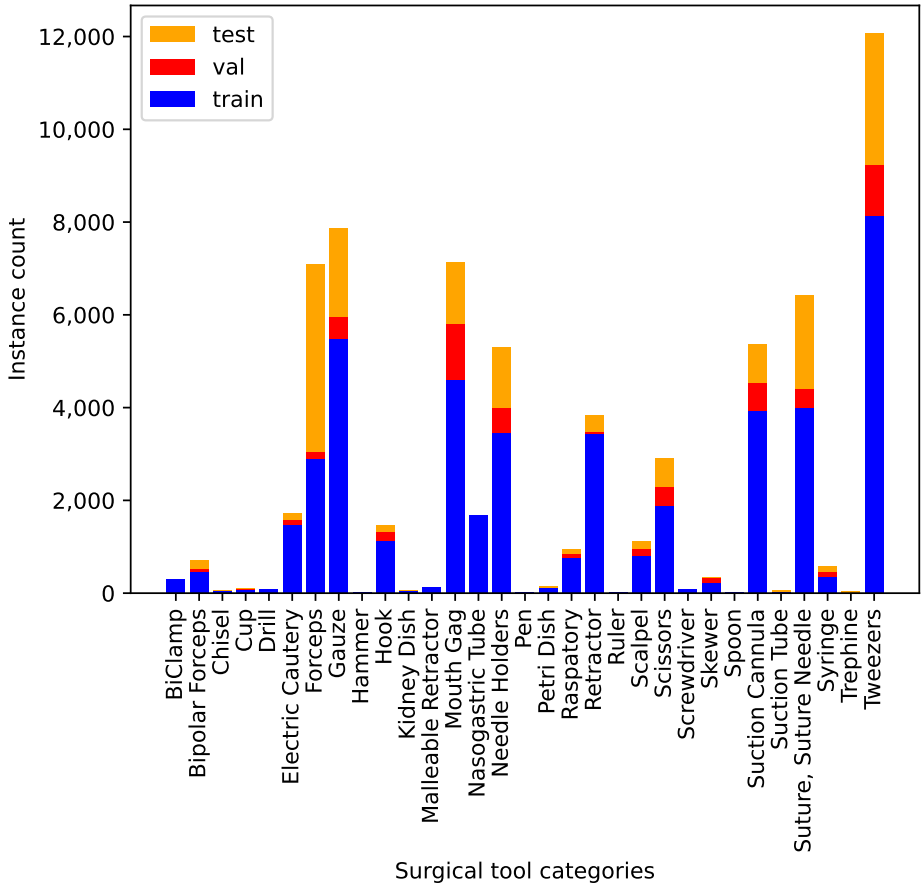


Figure 7. Distribution of the number of samples per surgical tool category in our training, validation, and test sets, represented by blue, red, and orange bars in the diagram, respectively. Our dataset has a high class imbalance.

3.4. Number of Instances per Frame

We provide a comparison between our dataset and M2CAI16-tool-locations [13], which is a common dataset for surgical tool detection for MIS in terms of the number of instances per image. According to Figure 8, our dataset is more dense than m2cai16-tool-locations [13]. In general, MIS contains three surgeons performing operations. Two surgeons operate the laparoscopic tools, and one surgeon controls the laparoscopic camera. Thus, the number of surgical tools that appear in a frame is at most four. On the other hand, in open surgery, in addition to the surgeons, other people (e.g., anesthesiologists, perfusionists, and nurses) controlling surgical tools can appear in a frame at the same time. The maximum number of surgical tools that appear in our dataset is 15. Moreover, the number of surgical tools appear in a frame drastically varies based on the surgery type. This makes the variance of

the number of instances per image from our open surgery dataset significantly larger than the existing MIS dataset.

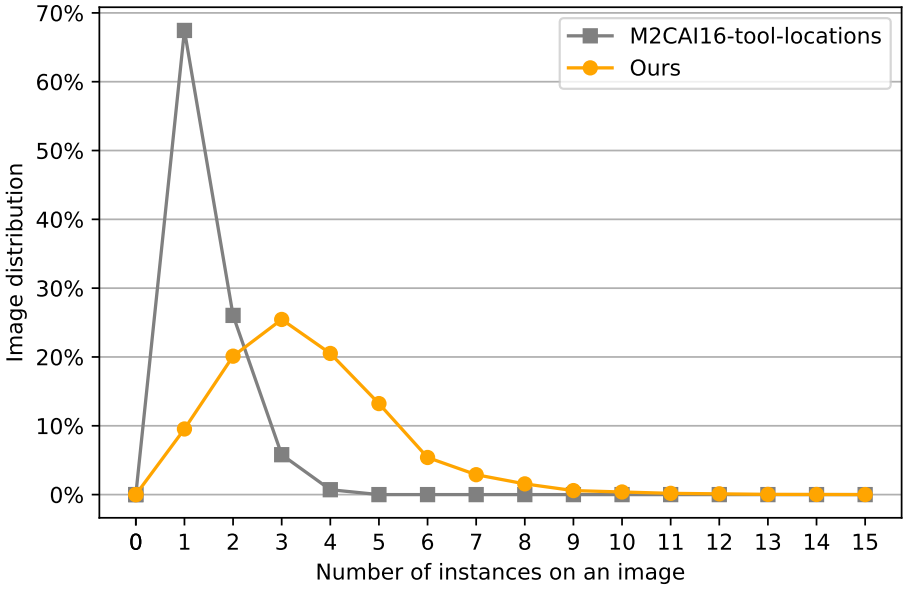
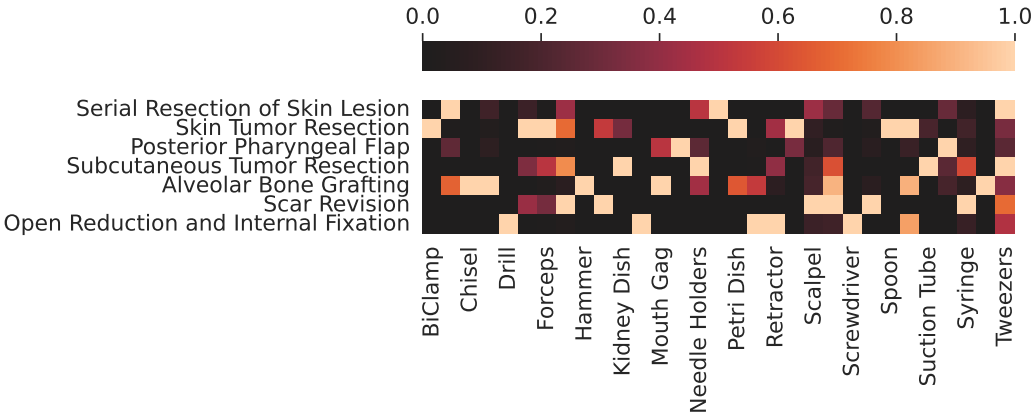


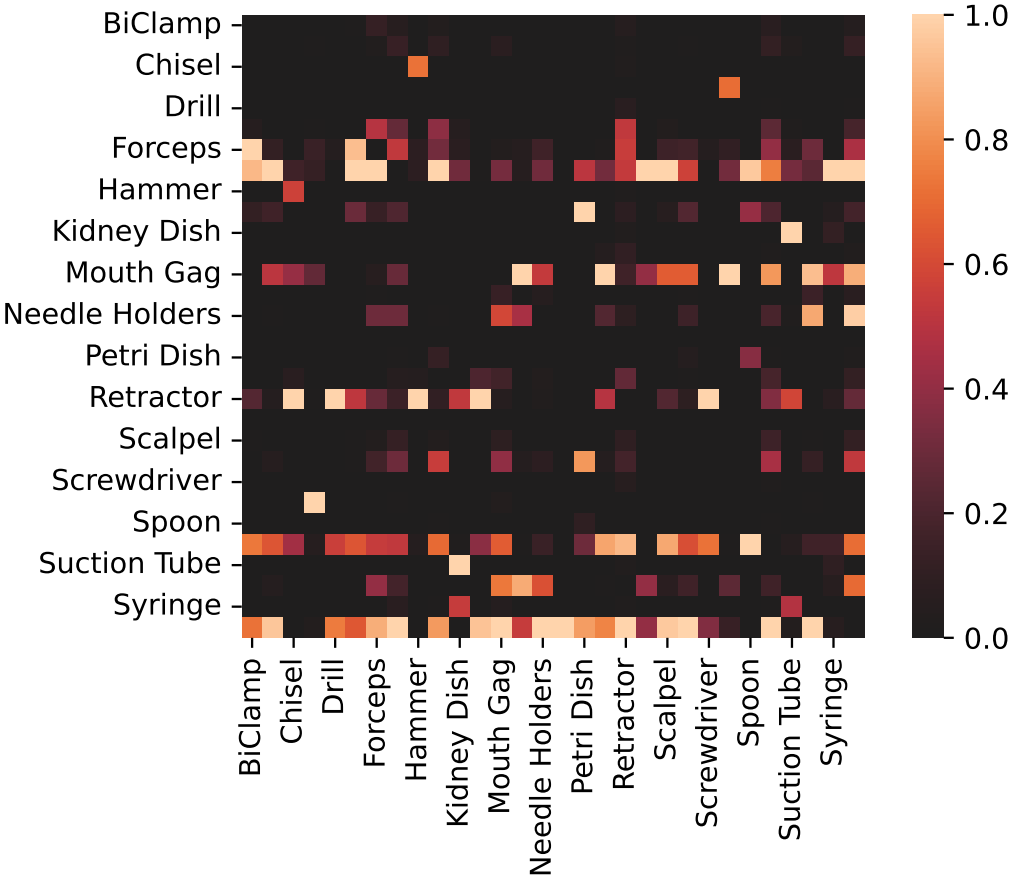
Figure 8. A comparison between M2CAI16-tool-locations, the endoscopic surgical spatial tool detection dataset [13] , and our open surgery dataset based on the statistics of the number of instances per image. Our dataset is denser with a modal of five categories per image with a flatter curve. Although the maximum number of instances per image of M2CAI16-tool-locations is four, that of our dataset is fifteen.

3.5. Co-Occurrences of the Surgical Tools

We also study the co-occurrences in our dataset. In Figure 9a, the co-occurrence matrix of surgical tools and surgical types, we can see some tools are only used in particular surgery types. As is obvious, we can see *Mouth Gag* is only used in the types of surgery performed around the mouth (*Posterior Pharyngeal Flap* and *Alveolar Bone Grafting*), and *Skewer* is only used for *Scar Revision*. On the other hand, we can confirm *Tweezers* are used in all surgeries. Therefore, since the surgical tool type can be the main indicator of the type of surgery performed, it is important to distinguish between the types of surgical tools. In Figure 9b, the c0-occurrence matrix of the surgical tools and surgical tools, we can see some sets of tools appear at the same time. For example, *Cup* and *Skewer*, *Chisel* and *Hammer* and *Needle Holders* and *Suture*, *Suture Needle* are often used together. On the other hand, *Gauze*, *Suction Cannula* and *Tweezers* appear with any type of tool. Therefore, it indicates that the information of one tool can help the detection of other tools.



(**a**) Co-occurrence matrix of surgical tools and surgical types



(**b**) Co-occurrence matrix of surgical tools and surgical tools

Figure 9. We showthe co-occurrence matrix of surgical tools and surgical types and surgical tools and surgical tools.

4. Experiments

4.1. Object Detection Models

We provide the baseline results on two widely used detectors, Faster R-CNN [27] and RetinaNet [28], as our benchmark testing algorithms.Faster R-CNN [27] is a two-stage-based object detection method. A feature extraction network takes an RGB image as input and extracts features. Then, a region proposal network (RPN) is used to generate regions of interest (ROIs) and features are pooled over these ROIs before being passed to a final classification and bounding box refinement network. We employ the FPN feature extraction backbone to extract the features from multiple-resolution feature maps. RetinaNet [28] is a

single-stage-based object detection method. A RetinaNet detector is made up of a backbone network and two subnets, one for object classification and the other for object localization. FPN is used at the end of a RetinaNet backbone network. RetinanNet adopts Focal Loss which is designed to handle the problem of class imbalance. For the experiment, we exploit three kinds of backbones, i.e., ResNet-50, ResNet-101 [47], and ResNe-X101 [48].

4.2. Experiment Setup

We used the latest PyTorch implementations of Faster R-CNN and RetinaNet, released by Facebook research [49]. For training, we fine-tune models pretrained on MS-COCO [50] with our training data. We trained our detector on an NVIDIA RTX A5000 GPU, with a batch size of 8. The input image size is 1920×1080. For training, we use the same setting defined in detectron2 [49] for COCO except for the learning rate schedule. We set the learning rate to 0.02. The networks were trained for 50K iterations with a learning rate drop of a factor of 10 after 33K and 44K iterations.

5. Results

The performance of the baseline models for the validation and test sets is represented in Table 3. Overall, Faster-RCNN, which is a two-stage detector, clearly outperforms RetinaNet, which is a single-stage detector. In general, two-stage models are considered to be able to provide better detection accuracy. We can find this tendency in our results. The average precision (AP) of the test set was found to be much lower than that of the validation set for all models. This is caused by the way of performing a split. Because we split training, validation, and test sets with *video-based split* considering the practical scenario, the distribution of categories in the three-set is different. Therefore, a gap in AP between validation and test sets occurs.

We conducted experiments with different backbones to explore the dependency of performance from the backbone. Faster-RCNN with a ResNet-X101 backbone presents the highest AP, and RetinaNet with a ResNet-50 backbone presents the lowest for both validation and test sets. The increase in depth and parameters slightly improves the AP in Faster-RCNN. On the other hand, for RetinaNet, the increase in depth and parameters have little effect on the AP.

Table 3. AP, AP_{50} and AP_{75} per model for validation and test sets. The best results are highlighted in bold.

Model	Backbone	Iters	Validation Set			Test Set		
			AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Faster-RCNN	ResNet-50	40K	48.9	63.6	55.0	27.2	41.7	29.7
	ResNet-101	37K	**51.3**	**66.0**	56.6	28.6	42.9	32.0
	ResNet-X101	40K	**51.3**	64.5	**57.0**	**29.7**	**44.2**	**32.8**
RetinaNet	ResNet-50	41K	**48.1**	**62.4**	**52.2**	26.7	41.0	29.1
	ResNet-101	42K	48.0	61.8	52.1	**27.8**	**41.5**	**29.9**

The AP per category given in Tables 4 and 5 shows that each surgical tool was detected with varying degrees of accuracy. This is mainly because of heavily imbalanced data, which naturally incur bias in the model. As a whole, the greater the number of annotated instances we can get, the better AP. For example, the baseline models successfully detected *Mouth Gag*, *Needle Holders*, *Suction Cannula*, and *Tweezers* categories, which have a relatively large number of instances. On the other hand, they fail to detect *Chisel*, *Hammer*, *Pen*, and *Suction Tube* categories, which have a small number of instances. The class imbalance is the main reason why there is a large gap between the AP of validation data and test data. As shown in Table 1, compared to the validation data, the test data includes the category in which the number of instances is relatively small. This may result in better AP of validation sets rather than that of the test set.

There are cases where the models fail to detect the categories that have profound instances because of their challenging characteristics. First, the deformable surgical tool category is difficult to be detected. *Gauze* has the second largest number of instances in the dataset, but the AP of the best baseline model, Faster-RCNN ResNet-X101, is 15.7. This is much worse than the *Mouth Gag* category, which has almost the same number of instances in the dataset. The reason for this may be that the *Gauze* category is deformable and its appearance drastically changes every scene. *Suture, Suture Needle* also has this characteristic. Second, the surgical tool category that has a variety of appearances through instances is difficult to detect. *Retractor*, which is used to separate the edges of a surgical incision, has a different shape and size depending on the situation. The *Retractor* category also has a relatively large number of instances, but the AP of the best baseline model is only 11.3. Finally, the surgical tool category, which has different usages is difficult to detect. For instance, the *Forceps* category, which is used to grasp and hold objects, is mainly utilized to grasp the tissues or organs but sometimes surgical drape. In the former usage, *Forceps* is handled by the surgeon's hand, and in the latter usage, *Forceps* is laid down around the surgical fields. Thus, depending on the usage, the visual feature drastically changes. This affects the surgical tool detection results.

Table 4. AP of baseline models per category for validation set.

Category	Faster-RCNN			RetinaNet	
	ResNet-50	ResNet-101	ResNet-X101	ResNet-50	ResNet-101
BiClamp	-	-	-	-	-
Bipolar Forceps	62.9	59.5	59.4	55.4	57.8
Chisel	-	-	-	-	-
Cup	0.1	0.2	0.0	0.6	0.6
Drill	-	-	-	-	-
Electric Cautery	86.3	86.2	89.2	88.1	88.6
Forceps	14.7	20.0	22.8	17.2	14.1
Gauze	19.4	18.3	18.9	23.8	19.1
Hammer	-	-	-	-	-
Hook	32.8	34.9	42.2	39.6	37.9
Kidney Dish	-	-	-	-	-
Malleable Retractor	-	-	-	-	-
Mouth Gag	71.4	72.8	71.5	73.2	73.3
Nasogastric Tube	-	-	-	-	-
Needle Holders	74.4	77.6	78.6	76.3	78.9
Pen	-	-	-	-	-
Petri Dish	-	-	-	-	-
Raspatory	61.5	66.9	61.7	57.2	56.7
Retractor	1.7	14.3	3.9	1.0	5.4
Ruler	-	-	-	-	-
Scalpel	78.8	82.0	87.7	75.9	72.8
Scissors	56.33	59.5	58.7	57.4	57.6
Screwdriver	-	-	-	-	-
Skewer	83.8	83.6	83.5	84.9	61.5
Spoon	-	-	-	-	-
Suction Cannula	67.8	70.8	67.3	63.3	65.3
Suction Tube	-	-	-	-	-
Suture, Suture Needle	4.5	7.1	7.0	4.5	6.5
Syringe	48.3	49.2	51.6	39.1	52.4
Trephine	-	-	-	-	-
Tweezers	66.6	68.3	63.9	65.8	65.0
Average	48.9	51.3	51.3	48.1	48.0

Table 5. AP of baseline models per category for test set.

Category	Faster-RCNN			RetinaNet	
	ResNet-50	ResNet-101	ResNet-X101	ResNet-50	ResNet-101
BiClamp	-	-	-	-	-
Bipolar Forceps	46.4	48.7	50.2	39.8	47.6
Chisel	0.19	0.0	0.5	0.3	0.1
Cup	18.5	14.7	11.0	22.0	19.1
Drill	-	-	-	-	-
Electric Cautery	49.7	46.7	58.4	49.4	45.4
Forceps	6.8	8.9	8.9	7.1	7.6
Gauze	17.3	18.5	15.7	19.1	16.7
Hammer	0.0	0.0	0.0	0.0	0.0
Hook	28.5	44.9	46.2	29.4	29.5
Kidney Dish	10.4	12.3	14.2	17.3	3.7
Malleable Retractor	-	-	-	-	-
Mouth Gag	73.3	73.1	73.7	73.9	74.0
Nasogastric Tube	-	-	-	-	-
Needle Holders	30.6	29.1	30.3	29.2	31.1
Pen	-	-	-	-	-
Petri Dish	0.5	0.3	0.0	3.7	2.8
Raspatory	56.1	56.0	60.9	54.5	56.4
Retractor	12.1	12.6	11.3	8.7	10.2
Ruler	-	-	-	-	-
Scalpel	61.0	59.4	64.0	55.2	59.9
Scissors	19.4	21.4	24.6	19.5	21.6
Screwdriver	-	-	-	-	-
Skewer	63.5	71.6	77.3	60.6	66.2
Spoon	-	-	-	-	-
Suction Cannula	46.1	48.9	46.9	42.3	42.8
Suction Tube	0.6	1.1	0.9	1.2	0.5
Suture, Suture Needle	0.9	0.5	0.3	0.7	0.7
Syringe	25.0	21.9	27.9	14.5	18.0
Trephine	0.0	0.0	0.0	1.1	0.5
Tweezers	57.9	58.3	58.1	55.9	57.2
Average	27.2	28.6	29.7	26.7	27.8

6. Conclusions

Surgical tool detection is an indispensable task to be tackled for the surgical scene. However, mainly because of the lack of a large-scale and diverse dataset, there is only a small body of research work on open surgery videos. We presented our large-scale egocentric dataset for open surgery and the extensive evaluation of conventional object detection methods. We collected 15 open surgery videos of 7 different types of surgery using a camera attached to the surgeon's head. The dataset consists of 67,687 annotated images with a bounding box around the surgical tool and its category label. Our dataset is superior in size and number of tool types compared with the existing open surgery dataset, which is often collected using a simulator. The statistics presented for the dataset illustrate that the dataset is imbalanced because many surgical tools are used for various purposes.

Despite the baseline experiments showing promising results for the detection of surgery tools in open surgery, the AP is not still acceptable for critical medical applications. One promising direction in future work is finding strategies to overcome the difficulty in the detection of heavily imbalanced data. Planned future work will collect surgery videos of which surgery type is less frequent in our dataset considering the revealed fact that the occurrence of some surgery tools depends on the surgery type.

Author Contributions: Conceptualization, R.F.; data curation, R.F. and H.K.; formal analysis, R.F.; funding acquisition, H.K. and H.S.; investigation, R.F.; methodology, R.F.; project administration, H.K. and H.S.; resources, H.S.; supervision, H.S.; validation, R.F.; visualization, R.F.; writing—original draft, R.F.; writing—review and editing, R.H. and H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by MHLW Health, Labour, and Welfare Sciences Research Grants Research on Medical ICT and Artificial Intelligence Program Grant Number 20AC1004, the MIC/SCOPE #201603003, and JSPS KAKENHI Grant Number 22H03617.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Keio University School of Medicine.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Primus, M.J.; Putzgruber-Adamitsch, D.; Taschwer, M.; Münzer, B.; El-Shabrawi, Y.; Böszörmenyi, L.; Schoeffmann, K. *Frame-Based Classification of Operation Phases in Cataract Surgery Videos*; MultiMedia Modeling; Schoeffmann, K., Chalidabhongse, T.H., Ngo, C.W., Aramvith, S., O'Connor, N.E., Ho, Y.S., Gabbouj, M., Elgammal, A., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 241–253.
2. Zisimopoulos, O.; Flouty, E.; Luengo, I.; Giataganas, P.; Nehme, J.; Chow, A.; Stoyanov, D. *DeepPhase: Surgical Phase Recognition in CATARACTS Videos*; Medical Image Computing and Computer Assisted Intervention; Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 265–272.
3. Kannan, S.; Yengera, G.; Mutter, D.; Marescaux, J.; Padoy, N. Future-State Predicting LSTM for Early Surgery Type Recognition. *IEEE Trans. Med. Imaging* **2020**, *39*, 556–566. https://doi.org/10.1109/TMI.2019.2931158.
4. Volkov, M.; Hashimoto, D.A.; Rosman, G.; Meireles, O.R.; Rus, D. Machine learning and coresets for automated real-time video segmentation of laparoscopic and robot-assisted surgery. In Proceedings of the International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 754–759. https://doi.org/10.1109/ICRA.2017.7989093.
5. Liu, T.; Meng, Q.; Vlontzos, A.; Tan, J.; Rueckert, D.; Kainz, B. Ultrasound Video Summarization Using Deep Reinforcement Learning. In *Medical Image Computing and Computer Assisted Intervention*; Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 483–492.
6. Nwoye, C.I.; Mutter, D.; Marescaux, J.; Padoy, N. Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. *Int. J. Comput. Assist. Radiol. Surg.* **2019**, *14*, 1059–1067.
7. Du, X.; Allan, M.; Dore, A.; Ourselin, S.; Hawkes, D.; Kelly, J.D.; Stoyanov, D. Combined 2D and 3D tracking of surgical instruments for minimally invasive and robotic-assisted surgery. *Int. J. Comput. Assist. Radiol. Surg. (IJCARS)* **2016**, *11*. https://doi.org/10.1007/s11548-016-1393-4.
8. Chen, Z.; Zhao, Z.; Cheng, X. Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context. In Proceedings of the Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017; pp. 2711–2714. https://doi.org/10.1109/CAC.2017.8243236.
9. Kurmann, T.; Marquez Neila, P.; Du, X.; Fua, P.; Stoyanov, D.; Wolf, S.; Sznitman, R. Simultaneous Recognition and Pose Estimation of Instruments in Minimally Invasive Surgery. In *Medical Image Computing and Computer-Assisted Intervention*; Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 505–513.
10. Colleoni, E.; Moccia, S.; Du, X.; De Momi, E.; Stoyanov, D. Deep Learning Based Robotic Tool Detection and Articulation Estimation With Spatio-Temporal Layers. *IEEE Robot. Autom. Lett.* **2019**, *4*, 2714–2721. https://doi.org/10.1109/LRA.2019.2917163.
11. Du, X.; Kurmann, T.; Chang, P.L.; Allan, M.; Ourselin, S.; Sznitman, R.; Kelly, J.D.; Stoyanov, D. Articulated Multi-Instrument 2-D Pose Estimation Using Fully Convolutional Networks. *IEEE Trans. Med. Imaging* **2018**, *37*, 1276–1287. https://doi.org/10.1109/TMI.2017.2787672.
12. Rivoir, D.; Bodenstedt, S.; von Bechtolsheim, F.; Distler, M.; Weitz, J.; Speidel, S. Unsupervised Temporal Video Segmentation as an Auxiliary Task for Predicting the Remaining Surgery Duration. In *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*; Zhou, L., Sarikaya, D., Kia, S.M., Speidel, S., Malpani, A., Hashimoto, D., Habes, M., Löfstedt, T., Ritter, K., Wang, H., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 29–37.
13. Jin, A.; Yeung, S.; Jopling, J.; Krause, J.; Azagury, D.; Milstein, A.; Fei-Fei, L. Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2018; pp. 691–699. https://doi.org/10.1109/WACV.2018.00081.
14. Choi, B.; Jo, K.; Choi, S.; Choi, J. Surgical-tools detection based on Convolutional Neural Network in laparoscopic robot-assisted surgery. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea, 11–15 July 2017; pp. 1756–1759. https://doi.org/10.1109/EMBC.2017.8037183.
15. Sarikaya, D.; Corso, J.J.; Guru, K.A. Detection and Localization of Robotic Tools in Robot-Assisted Surgery Videos Using Deep Neural Networks for Region Proposal and Detection. *IEEE Trans. Med. Imaging* **2017**, *36*, 1542–1549. https://doi.org/10.1109/TMI.2017.2665671.
16. Vardazaryan, A.; Mutter, D.; Marescaux, J.; Padoy, N. Weakly-supervised learning for tool localization in laparoscopic videos. In *MICCAI LABELS*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 169–179.
17. Al Hajj, H.; Lamard, M.; Conze, P.H.; Roychowdhury, S.; Hu, X.; Maršalkaitė, G.; Zisimopoulos, O.; Dedmari, M.A.; Zhao, F.; Prellberg, J.; et al. CATARACTS: Challenge on automatic tool annotation for cataRACT surgery. *Med. Image Anal.* **2019**, *52*, 24–41. https://doi.org/https://doi.org/10.1016/j.media.2018.11.008.
18. Twinanda, A.P.; Shehata, S.; Mutter, D.; Marescaux, J.; de Mathelin, M.; Padoy, N. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Trans. Med. Imaging* **2017**, *36*, 86–97. https://doi.org/10.1109/TMI.2016.2593957.

19. Twinanda, A.P.; Mutter, D.; Marescaux, J.; de Mathelin, M.; Padoy, N. Single-and multi-task architectures for tool presence detection challenge at M2CAI 2016. *arXiv* **2016**, arXiv:1610.08851.
20. Bawa, V.S.; Singh, G.; KapingA, F.; Skarga-Bandurova, I.; Oleari, E.; Leporini, A.; Landolfo, C.; Zhao, P.; Xiang, X.; Luo, G.; et al. The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods. *arXiv* **2021**, arXiv:2104.03178.
21. Zia, A.; Sharma, Y.; Bettadapura, V.; Sarin, E.L.; Clements, M.A.; Essa, I. Automated Assessment of Surgical Skills Using Frequency Analysis. In *Medical Image Computing and Computer-Assisted Intervention*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 430–438.
22. Basiev, K.; Goldbraikh, A.; Pugh, C.M.; Laufer, S. Open surgery tool classification and hand utilization using a multi-camera system. *arXiv* **2021**, arXiv:2111.06098.
23. Goldbraikh, A.; D'Angelo, A.L.; Pugh, C.M.; Laufer, S. Video-based fully automatic assessment of open surgery suturing skills. *Int. J. Comput. Assist. Radiol. Surg. (IJCARS)* **2022**, *17*, 437–448.
24. Shimizu, T.; Hachiuma, R.; Kajita, H.; Takatsume, Y.; Saito, H. Hand Motion-Aware Surgical Tool Localization and Classification from an Egocentric Camera. *J. Imaging* **2021**, *7*, 15. https://doi.org/10.3390/jimaging7020015.
25. Goodman, E.D.; Patel, K.K.; Zhang, Y.; Locke, W.; Kennedy, C.J.; Mehrotra, R.; Ren, S.; Guan, M.; Downing, M.; Chen, H.W.; et al. A real-time spatiotemporal AI model analyzes skill in open surgical videos. *arXiv* **2021**, arXiv:2112.07219.
26. Shimizu, T.; Oishi, K.; Hachiuma, R.; Kajita, H.; Takatsume, Y.; Saito, H. Surgery recording without occlusions by multi-view surgical videos. In *VISAPP, Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), Valetta, Malta, 27–29 February 2020*; Farinella, G., Radeva, P., Braz, J., Eds.; SciTePress: Setubal, Portugal, 2020; pp. 837–844.
27. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In A*dvances in Neural Information Processing Systems (NIPS)*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2015; Volume 28.
28. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. https://doi.org/10.1109/TPAMI.2018.2858826.
29. Kranzfelder, M.; Schneider, A.; Fiolka, A.; Schwan, E.; Gillen, S.; Wilhelm, D.; Schirren, R.; Reiser, S.; Jensen, B.; Feussner, H. Real-time instrument detection in minimally invasive surgery using radiofrequency identification technology. *J. Surg. Res.* **2013**, *185*. https://doi.org/10.1016/j.jss.2013.06.022.
30. Haase, S.; Wasza, J.; Kilgus, T.; Hornegger, J. Laparoscopic instrument localization using a 3-D Time-of-Flight/RGB endoscope. In Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV), Clearwater Beach, FL, USA, 15–17 January 2013; pp. 449–454. https://doi.org/10.1109/WACV.2013.6475053.
31. Reiter, A.; Allen, P.K. An online learning approach to in-vivo tracking using synergistic features. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; pp. 3441–3446. https://doi.org/10.1109/IROS.2010.5650852.
32. Bouget, D.; Benenson, R.; Omran, M.; Riffaud, L.; Schiele, B.; Jannin, P. Detecting Surgical Tools by Modelling Local Appearance and Global Shape. *IEEE Trans. Med. Imaging* **2015**, *34*, 2603–2617. https://doi.org/10.1109/TMI.2015.2450831.
33. Reiter, A.; Allen, P.K.; Zhao, T. Feature Classification for Tracking Articulated Surgical Tools. In *Medical Image Computing and Computer-Assisted Intervention*; Ayache, N., Delingette, H., Golland, P., Mori, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 592–600.
34. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.
35. Mishra, K.; Sathish, R.; Sheet, D. Learning Latent Temporal Connectionism of Deep Residual Visual Abstractions for Identifying Surgical Tools in Laparoscopy Procedures. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 2233–2240. https://doi.org/10.1109/CVPRW.2017.277.
36. Al Hajj, H.; Lamard, M.; Conze, P.H.; Cochener, B.; Quellec, G. Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. *Med. Image Anal.* **2018**, *47*, 203–218. https://doi.org/10.1016/j.media.2018.05.001.
37. Wang, S.; Xu, Z.; Yan, C.; Huang, J. Graph Convolutional Nets for Tool Presence Detection in Surgical Videos. In *Information Processing in Medical Imaging*; Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 467–478.
38. Jin, Y.; Li, H.; Dou, Q.; Chen, H.; Qin, J.; Fu, C.W.; Heng, P.A. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Med. Image Anal.* **2020**, *59*, 101572. https://doi.org/10.1016/j.media.2019.101572.
39. Hachiuma, R.; Shimizu, T.; Saito, H.; Kajita, H.; Takatsume, Y. Deep Selection: A Fully Supervised Camera Selection Network for Surgery Recordings. In *Medical Image Computing and Computer Assisted Intervention*; Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 419–428.
40. Saito, Y.; Hachiuma, R.; Saito, H.; Kajita, H.; Takatsume, Y.; Hayashida, T. Camera Selection for Occlusion-Less Surgery Recording via Training With an Egocentric Camera. *IEEE Access* **2021**, *9*, 138307–138322. https://doi.org/10.1109/ACCESS.2021.3118426.

41. Yoshida, K.; Hachiuma, R.; Tomita, H.; Pan, J.; Kitani, K.; Kajita, H.; Hayashida, T.; Sugimoto, M. Spatiotemporal Video Highlight by Neural Network Considering Gaze and Hands of Surgeon in Egocentric Surgical Videos. *J. Med Robot. Res.* **2021**, *7*, 2141001, https://doi.org/10.1142/S2424905X21410014.
42. Zhang, M.; Cheng, X.; Copeland, D.; Desai, A.; Guan, M.; Brat, G.; Yeung, S. Using Computer Vision to Automate Hand Detection and Tracking of Surgeon Movements in Videos of Open Surgery. *AMIA Annu. Symp. Proc.* **2021**, *2020*, 1373–1382.
43. Shi, P.; Zhao, Z.; Hu, S.; Chang, F. Real-Time Surgical Tool Detection in Minimally Invasive Surgery Based on Attention-Guided Convolutional Neural Network. *IEEE Access* **2020**, *8*, 228853–228862. https://doi.org/10.1109/ACCESS.2020.3046258.
44. Ahmidi, N.; Tao, L.; Sefati, S.; Gao, Y.; Lea, C.; Haro, B.B.; Zappella, L.; Khudanpur, S.; Vidal, R.; Hager, G.D. A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 2025–2041. https://doi.org/10.1109/TBME.2016.2647680.
45. Gao, Y.; Vedula, S.S.; Reiley, C.E.; Ahmidi, N.; Varadarajan, B.; Lin, H.C.; Tao, L.; Zappella, L.; Béjar, B.; Yuh, D.D.; et al. Jhu-Isi Gesture and Skill Assessment Working Set (Jigsaws): A Surgical Activity Dataset for Human Motion Modeling. *MICCAI Workshop: M2cai*. 2014; Volume 3, p. 3. Available online: https://cirl.lcsr.jhu.edu/research/hmm/datasets/jigsaws_release/ (accessed on 31 August 2022)
46. Kadkhodamohammadi, A.; Gangi, A.; de Mathelin, M.; Padoy, N. A Multi-view RGB-D Approach for Human Pose Estimation in Operating Rooms. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 363–372. https://doi.org/10.1109/WACV.2017.47.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
48. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2017; pp. 5987–5995. https://doi.org/10.1109/CVPR.2017.634.
49. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. Detectron2. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on accessed on 31 August 2022).
50. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755.