Submission report should be between 1 and 5 pages and a word count not to exceed 1000. Use the following as a rough guide. You do not need to use this exact outline or its headings, but if you do, please flesh them out.

A Practical Two-Stage Method for Surgical Tool Classification and Localization

Team name: Medibot

Authors: Xiaoliu Ding

Affiliations: Shanghai Microport MedBot(Group) Co., Ltd R&D

Link to code repo: https://github.com/dxlcnm/surgvu25-cat1-submission.git

(Please provide sufficient description to re-train your model from scratch either in this report or your repo)

Private dataset used (e.g. private videos, additional annotations of challenge training set, etc): No

If answered yes above, link to private dataset used (this can also be in the code repo):

Introduction

Robotic surgery is becoming increasingly popular as technology develops. Accurate localization and classification of surgical tools is critical for computer-assisted interventions, surgical workflow analysis, and real-time decision support in the robotic surgery [1]. Our approach is motivated by the need for a solution that is accurate, real-time, and generalizable across diverse surgical environments, such as in vitro, in vivo and in animals.

Motivated by the above, we propose the following simple but effective solutions

(1) A two-stage surgical tool localization and classification method.

(2) A efficient and accurate semi-automatic data annotation method.

Methodology & Results

Method

According to the experience, we proposed a two-stage method to do the tool location and classification respectively. We use the yolov8l model [2] for the tool location because of its real-time and maturity. The location model only distinguishes whether it is a tool, but does not determine the category. This can improve the recall of the model. Then, the ResNet50 model is used to classify the location boxes into 12 categories.

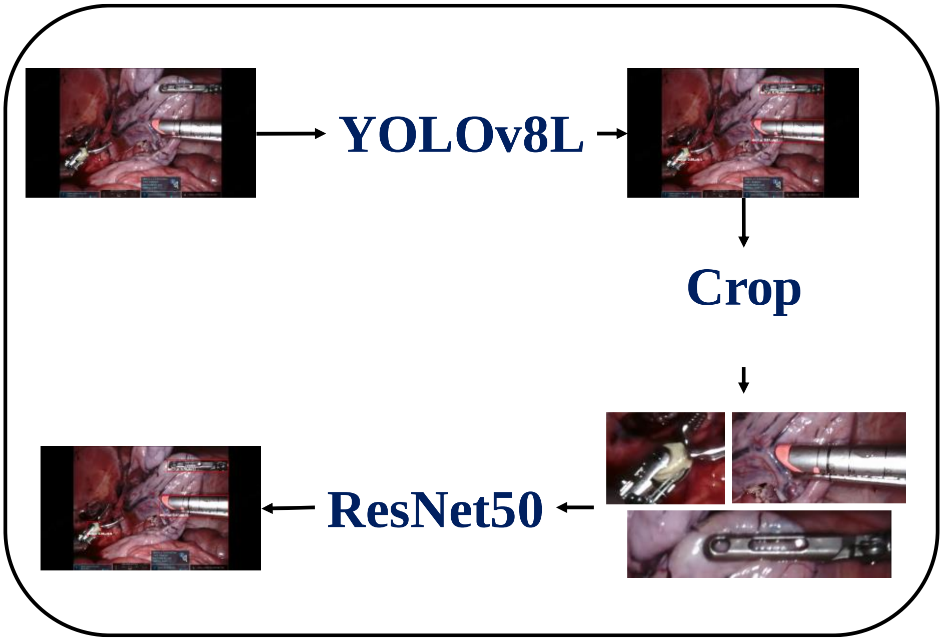

Fig1: a two-stage surgical tool localization and classification method.

Data

For the surgical tools location and classification task, the accurate labeling of surgical tools in every frame is very time-consuming but import for the model training. To generate the high quality labels, we trained a location and classification model with yolov8l using the Cat 1 Validation set. Then we clip out the segment from the videos according to the stop and end time in the task csv file. We got the first frame of these clips only to create a dataset. We do the prediction on the dataset to obtain the pseudo labels. We manually modify the incorrect categories of these pseudo labels to the correct ones. We try not to modify the box unless there is an obvious error. After counting the number of boxes in each category of all data , we found that some categories had too few boxes. We supplemented some data on tools with fewer categories in the tool csv file and different backgrounds with the same way.

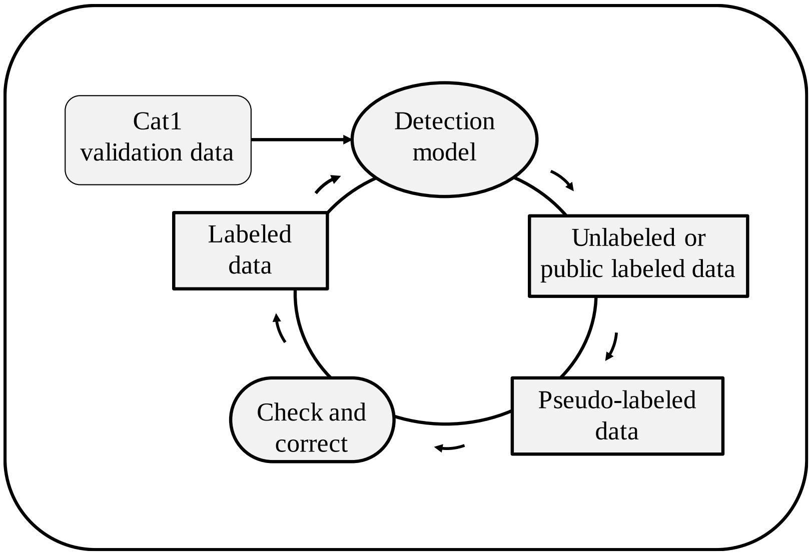

Fig2: an efficient and accurate semi-automatic data annotation method

Then, the training and validation dataset is formed by splitting the data by 8:2 portion. In order to increase the amount of data in the training set, the open source dataset EndoVis17 and the Cat 1 validation set are all added in the training set. Finally, we got the 10997 images in training set and 979 images in validation set. The boxes count of every category are list below.

Table 1. The box count of every category in training set and validation set

Category	Trainset Boxes count	Valset Boxes count
1	6696	1110
2	4074	258
3	949	25
4	674	94
5	3085	258
6	4486	266
7	563	36
8	502	15
9	1024	68
10	699	30
11	568	19
12	294	1

Results

Yolov8l-1cls detection

We found that compared with the yolov8l 12-class detection model, the yolov8l single-class detection model performed better. So we selected the yolov8l single-class model as the detection model. We evaluated the detection performance on validation set. We trained the model initialized with pretrained yolov8l weights for 1000 epochs with early stopping strategy. The image size is 640. The batch size is 300 with 6 A100 Gpus.

Table 2. The one-class vs twelve-class yolov8l result vs on validation set

	One-class	twelve-class
AP@[IoU=0.50:0.95]	0.490	0.435
AP@[IoU=0.50]	0.776	0.708
AR@[IoU=0.50:0.95]	0.558	0.497

ResNet50 classification

For all categories, we cropped regions using bounding boxes scaled by a factor of two relative to the original size, following the approach in [1]. We initialized the model with pretrained ResNet-50 weights and then trained it on our dataset. We trained the model with an input size of 224 and a batch size of 600 for 400 epochs. In contrast to standard classification model training, we resized each cropped region directly to 224 × 224, without applying random scale cropping. We enabled model exponential moving average (EMA) ,Mixed-precision training (AMP) and evaluation with EMA weights.. In addition, the training process adopted ThreeAug data augmentation [4].The classification model result is list below

Table 3. The 12-class resnet50 result on validation set

Max accuracy 1	EMA accuracy 1
95.88%	93.27%

Two-stage location and classification

Finally, we combined the results of the single-class detection model and the 12-class classification model for evaluation, and the final evaluation results are listed below.

Table 4. The 12-class location and classification mean mAP on validation set

Category	AP@[IoU=0.50:0.95]	AP@[IoU=0.50]
1: Needle driver	0.491	0.796
2: Monopolar curved scissors	0.665	0.905
3: Force bipolar	0.315	0.697
4: Clip applier	0.297	0.644
5: Cadiere forceps	0.248	0.489
6: Bipolar forceps	0.680	0.893
7: Vessel sealer	0.220	0.546
8: Permanent cautery hook/spatula	0.432	0.765
9: Prograsp forceps	0.167	0.460
10: Stapler	0.646	0.855
11: Grasping retractor	0.324	0.528
12: Tip-up fenestrated grasper	0.000	0.000
Average	0.374	0.632

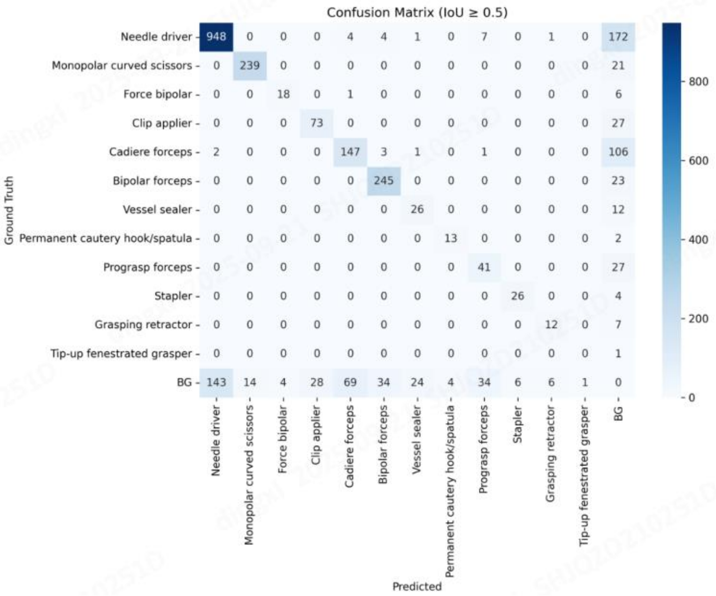

Fig 1: The confusion matrix of the classification result

Preliminary phase and Final phase

As the number of training data increased from 3543 to 10016, the indicators of the preliminary stage also continued to rise. The result of the submissions are shown in the following table

Table 5: submission result of different training data in different phases

Preliminary phase				Final phase
Image Number	2628	4845	10016	10976
Mean mAP	0.4587	0.6073	0.7124	0.5018

Conclusion & Discussion

The two-stage framework demonstrates strong performance in both detection and classification. Compared with the multi-class detector, the single-class detection model achieves superior accuracy. Moreover, a limited set of high-quality annotations proves to be more valuable for model training than a large quantity of noisy pseudo-labeled data. Within a certain range, increasing the amount of accurately labeled data consistently leads to improved model performance.

References

[1]. Intuitive Surgical SurgToolLoc Challenge Results：2022-2023

[2]. Dillon Reis, etc. Real-Time Flying Object Detection with YOLOv8. 2023.

[3]. Kaiming He. Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. 2015.

[4]. Hugo Touvron, Matthieu Cord, Hervé Jégou. DeiT III: Revenge of the ViT. 2022.