Machine Learning Task Report

Chao Wang

Project Code Link: https://github.com/WANGCHAO0116/SAR-RARP50

Introduction

In this task, we need to complete a work related to semantic segmentation. The dataset SAR-RARP50 [1] is provided, which contain tens of recording videos of surgery. Our goal is to segment surgical tool parts from the background, such as surgical needles, surgical clips, surgical threads, etc. It involves processing RGB video frames and assigning semantic labels at the pixel level. A convolutional neural network is built to implement this task.

Data Preprocess

First of all, the dataset is pre-processed. The original dataset consists of two parts, the train set (10128 images) and the test set (3252 images). In SAR-RARP50 dataset, there are totally ten segmentation classes in this dataset. They are “background”, “tool clasper”, “tool wrist”, “tool shaft”, “suturing needle”, “thread”, “suction tool”, “needle holder”, “clamps”, “catheter”. The provided videos of surgery are encoded at a constant FPS of 60 with the width of 1920 and the height of 1080. Each video has a corresponding directory containing instrumentation semantic segmentation masks for video frames samples at 1 Hz, which will be served as labels for the subsequent training. Therefore, we should extract frames from the provided videos at 1 Hz as well. The Python script that extracts frames is provided in the GitHub repository.

Method

Our model is a convolutional neural network which has an “encoder-decoder” structure. The encoder part of the model is to compress the size of latent embeddings and extract the semantic information from the input images, and it is also called “backbone”. We select Resnet-50 [2] as the backbone of our model. The decoder part of the model is to upsample latent embeddings to output masks, and it is also called “segmentation head”. The decoder of UperNet [3] is used as the segmentation head of our model. Therefore, our model is a combination between ResNet-50 and UperNet.

In this project, we utilized an open-sourced semantic segmentation framework called “MMSegmentation” to construct and train our model. SGD is used as the optimizer. Momentum, dropout and regularization techniques are also applied to improve the performance of the model. In total, our model is trained 100,000 iterations (with batch

size = 8) for 12 hours using the GPU Nvidia A100.

Result

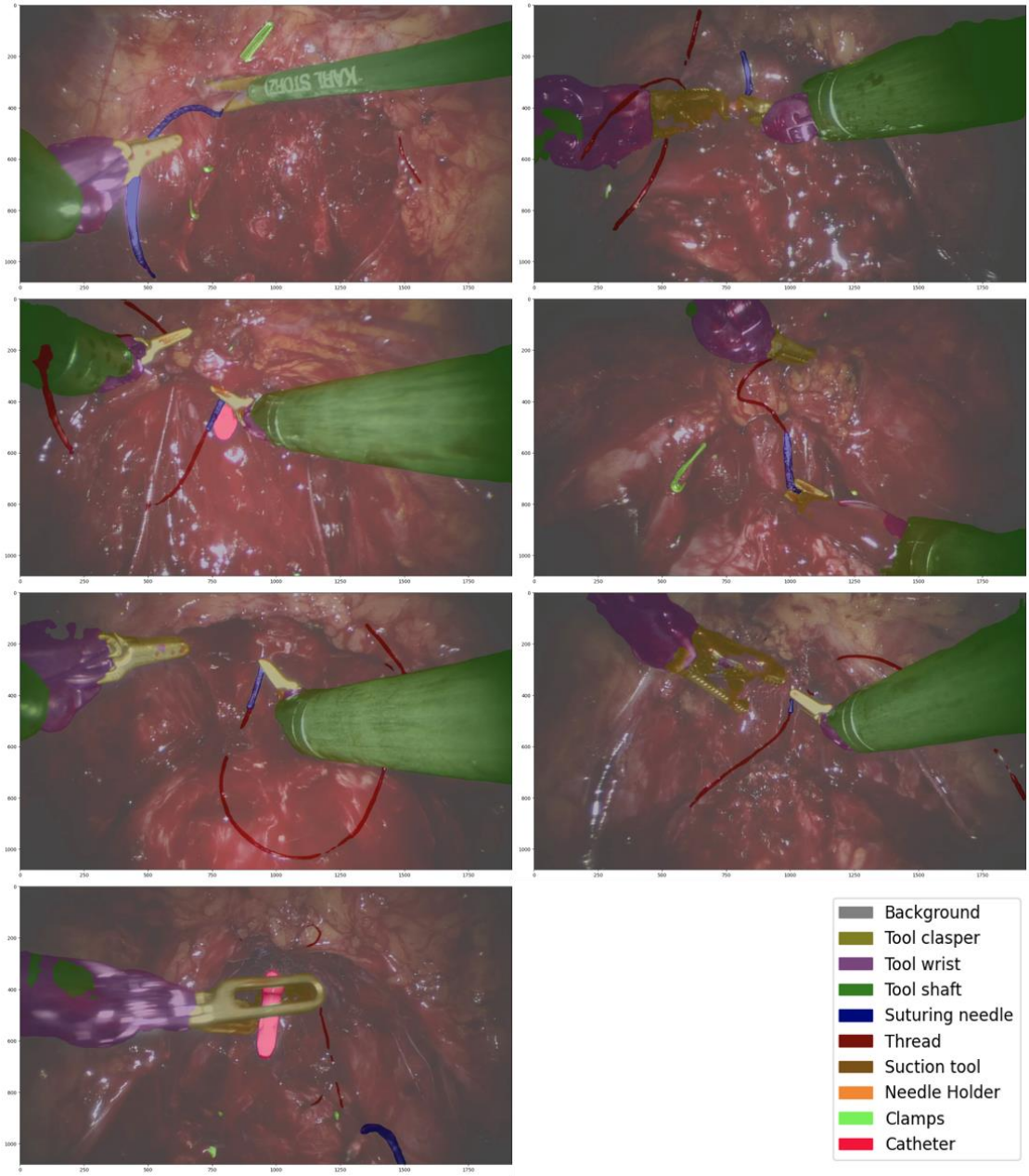

Figure 1. The visualization of predicted results of the trained model. (Masks are merged with their corresponding original images)

Figure 1 visualized the evaluation results. The pixels of surgical tools are marked by a corresponding colour. Intersection of union (IOU) and accuracy (ACC) are utilized as the metrics to evaluate the performance of the trained model. As table 1 shows, it performs well in predicted the pixels of classes “background”, “tool shaft” and “catheter”. As for classes “thread”, “suction tool” and “clamps”, its values of IOU and ACC are relatively lower than other classes.

CLASS	IOU	ACC
BACKGROUND	95.65	99.29
TOOL CLASPER	64.51	72.78
TOOL WRIST	67.49	74.62
TOOL SHAFT	88.24	91.64
SUTURING NEEDLE	63.03	72.11
THREAD	49.89	56.16
SUCTION TOOL	45.51	50.17
NEEDLE HOLDER	72.43	76.89
CLAMPS	37.46	40.19
CATHETER	91.33	95.09
MEAN VALUE	67.55	72.89

Table 1. The evaluation results of the trained model.

Reference

[1] Dimitrios Psychogyios. (2024). SAR-RARP50: Segmentation of surgical instrumentation and Action Recognition on Robot-Assisted Radical Prostatectomy Challenge.

[2] Koonce, B. (2021). ResNet 50. In: Convolutional Neural Networks with Swift for Tensorflow.

[3] Tete Xiao. (2018). Unified Perceptual Parsing for Scene Understanding.