Semantics Segmentation for Surgical Scene

Ding Hao, Mingyi Zheng, Zhaoshuo Li

***Abstract*— This report was submitted for CS 682 Deep Learning Final Project. The goal of this project is to successfully segment 12 classes of objects in the surgical scene, using the dataset from EndoVis2018 Challenge. We have experimented with the UNet, AlbuNet and DeepLabV3+ architectures. In addition, we proposed an architecture design named *Super-Label* which can improved the segmentation result, and a data splitting method named *Mini-Data* which encourages the diversity of data. We have demonstrated the efficacy of our approaches with a mean dice score of 0.78 on only 20% of the total data.**

I. INTRODUCTION

Robot-assisted surgeries are currently completely human teleoperated, given a prominent example of the da Vinci surgical robot developed by the Intuitive Surgical Inc. One future direction of surgical assisted surgeries is to increase the level of autonomy, which requires certain level of scene understanding. Therefore, this project will focus on semantic segmentation for 12 categories of objects (Figure 1). Semantics segmentation for surgical scenes have the challenge of change of illumination, small objects such as needles and thread, visual occlusion and similar textures between kidney, kidney parenchyma and background tissues.

0	background-tissue
1	instrument-shaft
2	instrument-clasper
3	instrument-wrist
4	kidney-parenchyma
5	covered-kidney
6	thread
7	clamps
8	suturing-needle
9	suction-instrument
10	small-intestine
11	ultrasound-probe

Fig. 1. Object classes in EndoVis2018.

FCN [1] was the first successful network that combines classification and convolution for segmentation purpose. The idea is that each pixel will be assigned to a classification label. The network upsamples feature map at different resolution to the size of original input to produce from coarse to fine segmentation result.

U-Net [2] is developed using an encoder-decoder design, with the idea being low dimensionality latent representation can be blown up to high-level classification result. Currently, most of the architectures followed similar strategies and have produced great result in various challenges. A simplified U-Net is implemented for our project, which has 3x max-pooling layers and 3x up-conv layers.

AlbuNet34 [3] follows closely to the design of U-Net, except that after down-sampling, the fully-connected layers in the bottleneck part of the network is replaced with a pre-trained ResNet34 before the fully-connected layers. The ResNet34 is used to enhance the capability of dimensionality compression.

DeepLabV3+ [4] used one tunable dilated convolution (named Atrous Spatial Pyramid Pooling, ASPP), in comparison to UNet. On the other hand, it uses skip-connections to concatenate encoding and decoding features. It is an improved version of DeepLabV3 [5]. The network used in this project has ResNet101 as backbone, and applied ASPP on top of it. During the decoding process, two 4x up-sampling were performed instead of directly 8x up-sampling.

In this report, we summarized our comparison of different network performance with the same training parameters trying to solve the challenges present in the surgical scene segmentation. Moreover, we proposed *Super-Label* and *Mini-Data* techniques along with normal image augmentation, which has resulted a dice score 0.78. In Section section II, we will discuss the *Super-Label* and *Mini-Data* approaches. In Section section III, we will discuss the experimental settings and result.

II. METHODOLOGY

A. Super-Label

The idea of using *Super-Label* stems from [6], where the detection and classification problem is decoupled. We adopted this idea for segmentation by adding two different classification layers - super and sub - in the network, with the same encoding and decoding path. The super class is used to encourage the network to learn the difference of appearance of the super categories first, and then learn the intrigue differences within the super class. We think this will mitigate the problem of learning small objects and learn different visual appearances of the objects. The intuition is that classifying general categories will be easier than classifying the sub-categories, and given a general category, sub-classification will become easier. In this case, there are 2 classes in the Super-Label classifier, surgical instrument and background. Background-tissue, kidney-parenchyma, covered-kidney, and small-intestine are defined as the super class of background. The other classes shown in Figure 1 are the sub-classes of surgical instrument. An example of using *Super-Label* for DeepLabV3+ is shown in Figure 2, where two paths of *convolution-upsampling-softmax* classification for super and sub categories growing out from the same low-features.

During the training, the loss function is defined as

$$l_{total} = \lambda \cdot l_{super} + l_{sub} \tag{1}$$

where l indicates loss. The summation is used to facilitate gradient descent for back-propagation.

During inference time, the final score before softmax layer is calculated as

$$s_{total,i,j} = s_{super,i} \cdot s_{sub,j} \tag{2}$$

where s is the score, j is the $j-th$ sub class under $i-th$ super class. The maximum of different scores is used for assigning the final class.

B. Mini-Data

The idea of the *Mini-Data* was due to the limited training data available (2235 image in total) and the distinct distribution of object visual appearances of different video sequences. Therefore, it is very challenging for the network to predict an object with something it never sees. This is illustrated in Figure 3. Therefore, we thought about this problem in a different way. Given the same man-power, how can we improve the network's experience in a meaningful way? Thus, we proposed to split data and validation by sampling only 20% of a given video sequence and generalizing the network to the rest of 80% in order to include more diversity with the same amount of annotated data. Considering the relatively difficulty of recording many videos and labeling partial of the video frames versus recording a few videos and labeling all the frames, it is sensible that the former one will be more cost-effective. If the idea of *Mini-Data* works well, we think it will be more beneficial to label a subset of different video sequences for learning, and thus learning a better representation of surgical instrument and scene segmentation.

One critic on this approach might be the overfitting on one video sequence given the similarity between the sequences in the video. However, this is still much variation within a sequence since the video is sampled at 2 Hz. As shown in Figure 4, the frames in the same video sequence shows a greater variation compared to a video frame from another video sequence. Therefore, being able to generalize on the rest of 80% of the data is still a good indication of the generality of a network.

III. EXPERIMENT AND RESULT

A. Hyper Parameters and Data Pre-processing

All training are seeded such that the experiment is repeatable. Images are downsampled from 1280×1024 to 320×256 for memory saving. The training batch size is set to be 10. The learning rate is set to be 0.001. The number of epochs is set to be 30 epochs without augmentation and 60 epochs with augmentation. Adam optimizer is implemented with learning rate scheduler. The weight decay is set to be 5×10^{-4}, the learning rate decays by 0.95epoch. The dice loss is used. There are two super classes for this experiment, tissue and surgical tools, with a super class weight of $\lambda = 0.5$. Data augmentations of affine transformation, horizontal flip, dropout, spectral highlight, brightness, contrast, HSV and motion blur are used when necessary. The code is available at here.

B. Super-Label

To test if *Super-Label* helps with the segmentation learning, we compared the performance on two networks, AlbuNet and DeepLabV3+, and the UNet is used to serve as baseline. The validation results by leaving 3 video sequences out is summarized in Table II. The dice score of AlbuNet and DeepLabV3+ can be increased by 0.09 and 0.03 with *Super-Label*. Similar trends can be seen in other experiments such as **??**.

In the extreme case, the *Super-Label* approach can increase the dice score by 0.24 for an unseen validation image during our experiment. It can be seen in Figure 5 that the *Super-Label* can discourage random sparse prediction and enhance uniformity since attention was given to both global and local context. AlbuNet also outperforms the DeepLabV3+ due to a shallower network, therefore less prone to overfitting. Due to page limit, the high training mean dice scores are not included, which are signs for overfitting.

C. Mini-Data

To find a baseline for *Mini-Data*, we first performed two experiments. Firstly, we split the data by leaving 20% video sequences out. Therefore, the distribution of the training data and validation data are very distinct. This serves as a lower limit for our comparison. Then, we extracted 80% of the frames from each video sequence, serving as the gold standard since it overfits well on the training data. This serves as an upper limit for our comparison.

Then, for *Mini-Data*, we uniformly sampled 20% of the frames from each video sequence, resulting the same number of training images as before. The performance of the trained network is only reported for DeepLabV3+ since it outperformed other architectures.

It has shown that with only limited data, the network is able to segment the rest of the video sequence. However, for small objects with only a few pixels, such as the needle class (only 9 frames containing this classes, with a total of 4754 pixels), we only successfully segmented it with the "upper limit" data splitting technique. This indicates that more data may be beneficial for learning this class, for which augmentation and increased dice loss weight could not help.

D. Other

Interestingly, we have also found some mislabelling from the dataset, which the network outputs a more correct result. For example, for video sequence 1 frame 94 (Figure 6), the label does not include any suturing threads, while the Upper Limit DeepLabV3+SuperLabel network predicts the suturing threads successfully. May there be any other errors, the network may not have learned correctly.

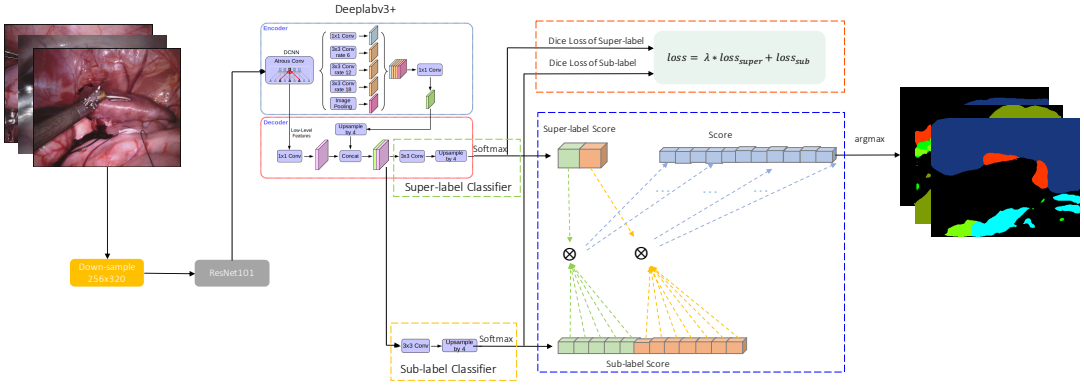


Fig. 2. Super-label Network

TABLE I

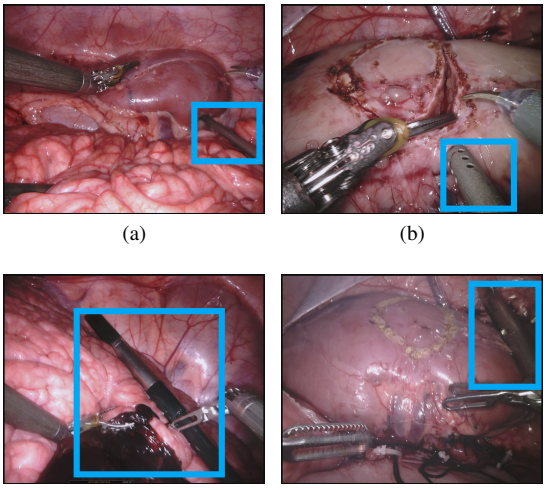
COMPARISON OF VALIDATION DICE SCORE WITH/WITHOUT SUPER LABEL

	0	1	2	3	4	5	6	7	8	9	10	11	Mean
UNet	0.66	0.87	0.76	0.77	0.41	0.22	0.35	0.22	0.00	0.09	0.53	0.00	0.41
AlbuNet	0.69	0.90	0.76	0.78	0.51	0.29	0.38	0.15	0.00	0.23	0.59	0.01	0.44
AlbuNet+SuperLabel	0.75	0.94	0.79	0.84	0.60	0.43	0.43	0.45	0.00	0.45	0.62	0.00	**0.53**
DeepLabV3+	0.74	0.89	0.76	0.80	0.65	0.29	0.30	0.40	0.00	0.06	0.56	0.00	0.45
DeepLabV3+SuperLabel	0.74	0.92	0.78	0.83	0.64	0.33	0.33	0.39	0.00	0.20	0.59	0.00	**0.48**

TABLE II

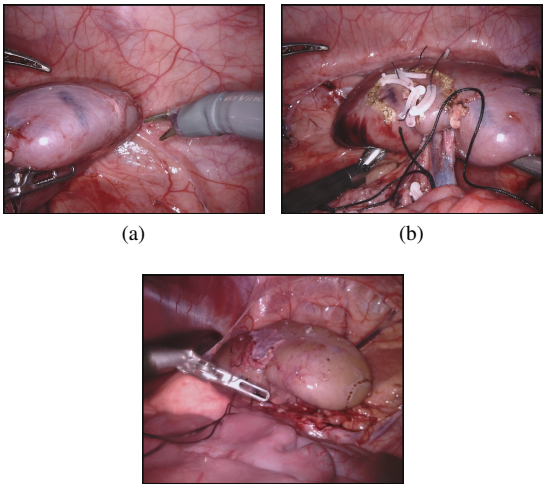
COMPARISON OF VALIDATION DICE SCORE FOR DEEPLABV3+ WITH DIFFERENT DATA SPLITTING

		0	1	2	3	4	5	6	7	8	9	10	11	Mean
Lower Limit	**DeepLabv3+**	0.74	0.89	0.76	0.80	0.65	0.29	0.30	0.40	0.00	0.06	0.56	0.00	0.45
	DeepLabv3+SuperLabel	0.74	0.92	0.78	0.83	0.64	0.33	0.33	0.39	0.00	0.20	0.59	0.00	**0.48**
Upper Limit	**DeepLabV3+**	0.96	0.95	0.89	0.87	0.96	0.96	0.69	0.9	0.37	0.84	0.97	0.86	0.85
	DeepLabv3+SuperLabel	0.97	0.96	0.89	0.87	0.96	0.96	0.7	0.9	0.38	0.82	0.96	0.89	**0.86**
Mini Data	**DeepLabv3+**	0.91	0.93	0.81	0.82	0.94	0.87	0.51	0.60	0.00	0.76	0.92	0.73	0.73
	DeepLabv3+SuperLabel	0.93	0.93	0.83	0.79	0.91	0.90	0.64	0.85	0.00	0.79	0.92	0.82	0.77
	DeepLabv3+Aug	0.90	0.94	0.80	0.84	0.94	0.84	0.53	0.68	0.00	0.59	0.81	0.81	0.72
	DeepLabV3+SuperLabel_Aug	0.94	0.93	0.83	0.81	0.92	0.92	0.64	0.84	0.00	0.81	0.94	0.83	**0.78**



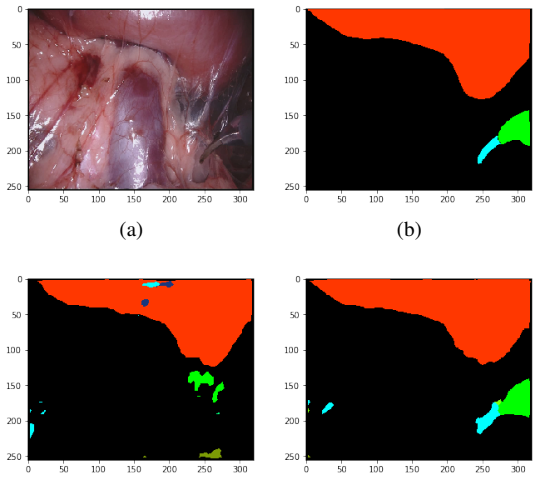
(a) (b) (c) (d)

Fig. 3. Example images from training and validation set with objects of interest enclosed in blue rectangles. (a)(b) Training and validation images of the suction tool, (c)(d) Training and validation image of the ultrasound probe.



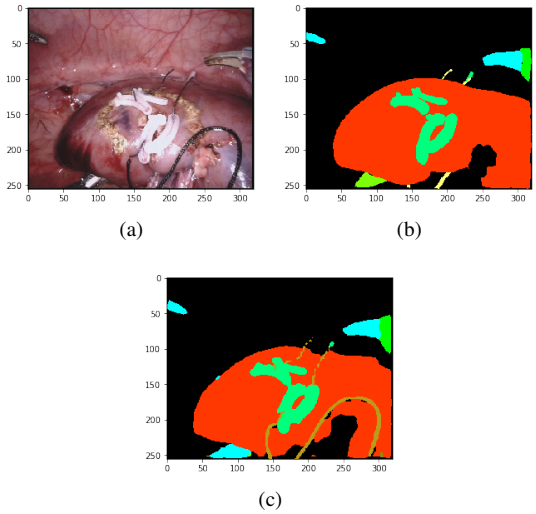
(a) (b) (c)

Fig. 4. Examples from two video sequence, (a) and (b) are from the same video sequence, while (c) is from another video sequence. Visually, (a) and (c) look similar compared to (b), which is against the intuition that frames from the same video share more similarity.



(a) (b) (c) (d)

Fig. 5. Example outputs with and without *Super-Label*. (a)(b) Input and label images, (c) output from models without *Super-Label*(d) output from models with *Super-Label*.



(a) (b) (c)

Fig. 6. A wrongly labeled data, (a) Input, (b) Label, (c) Network prediction

IV. CONCLUSIONS AND FUTURE WORK

In conclusion, we have proposed and evaluated two approaches for the challenges in surgical scene segmentation. The *Super-Label* can help the segmentation by learning the details of the small objects, while *Mini-Data* can encourage the learning process with fewer data.

In the future, it will be great to compare the performance of proposed approach with the 3D U-Net, by incorporating temporal information. Also, the network's performance against the full resolution images has not yet tested. If the test data used for the challenge can be made available, further evaluation can be done. Moreover, a few ideas that we came up with but didn't get them implemented due to time limit such as using puzzle solving for pre-training and random cropping the original image.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[3] A. A. Shvets, V. I. Iglovikov, A. Rakhlin, and A. A. Kalinin, "Angiodysplasia detection and localization using deep convolutional neural networks," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 612–617.

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[5] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[6] B. Singh, H. Li, A. Sharma, and L. S. Davis, "R-fcn-3000 at 30fps: Decoupling detection and classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1081–1090.