

國立臺灣大學電機資訊學院資訊網路與多媒體研究所

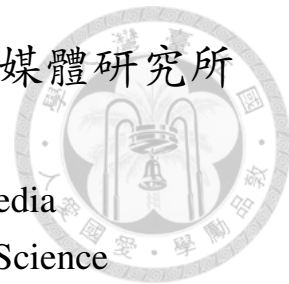
碩士論文

Graduate Institute of Networking and Multimedia

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis



內視鏡手術的語意分割:利用資料增強與中間層監督達到以少量資料訓練深層網路

Semantic Segmentation in Endoscopy Surgery: Using Data Augmentation and Intermediate Layer Supervision to Train Deep Neural Net with Few Data

江洵安

Hsun-An Chiang

指導教授：施吉昇 博士

Advisor: Chi-Sheng Shih, Ph.D.

中華民國 107 年 7 月

July, 2018

國立臺灣大學碩士學位論文
口試委員會審定書

內視鏡手術的語意分割:利用資料增強與中間層監督達到以少量資料訓練深度網路

Semantic Segmentation in Endoscopy Surgery: Using
Data Augmentation and Intermediate Layer Supervision to
Train Deep Neural Network with Few Data

本論文係江洵安君（學號 R04944017）在國立臺灣大學資訊網路與多媒體研究所完成之碩士學位論文，於民國一百零七年七月卅一日承下列考試委員審查通過及口試及格，特此證明

口試委員：

施吉昇 (簽名)

(指導教授)

徐慧中

楊佳玲

楊佳玲

所長：



Acknowledgments

I would like to express my gratitude to professor Chi-Sheng Shih for his counsellings. This thesis would not be finished without him. Also, I would like to express my gratitude to professor Chia-Lin Yang and professor Wei-Chung Hsu for their kind advices on my thesis.

Taipei, July, 2018

Hsun-An Chiang



摘要

中文摘要

隨著以內視鏡微創手術的興起，許多研究試著透過分析處理內視鏡影像來提供手術人員即時的協助。其中，我們試著處理內視鏡影像的語意分割問題，因為語意分割可以為許多應用提供重要的資訊，像是虛擬實境(VR)、擴增實境(AR)或是內視鏡的同時定位與建地圖(SLAM)。語意分割在都市場景或自然場景上已經累積了大量的研究，但是因為缺乏充足的學習資料，鮮少有研究觸及內視鏡手術。在這篇研究中，我們提出了一個資料增強的方法，和一種作用在網路中間層的監督模式，用來解決使用少量資料訓練深度網路的問題。實驗結果證明提出的方法可以比常用的資料增強方法更有效的提升網路的準確度。

關鍵字 - 內視鏡手術;微創手術;語意分割;資料增強;深度學習;類神經網路;中間層監督



Abstract

English Abstract

With the increasing popularity of endoscope-based minimal-invasive surgery, many have tried to provide surgeons real-time assistance by processing video frames from endoscope. We aim at a particular problem, endoscopy semantic segmentation, which can provide important information for other applications like VR or endoscopy-SLAM. While semantic segmentation in other scenarios, e.g. urban scene or natural scene, has been intensively studied, seldom has reach the area of endoscopy surgery due to lack of large-scale, finely annotated dataset. In this work, we tried to solve the problem of training deep neural network with few training data in the case of endoscopy surgery, by introducing an aggressive data augmentation technique, and additional loss term which is applied on intermediate layers of network. Experiment results show that our proposed methods can improve network performance more effectively than commonly used data augmentation on endoscopy surgery dataset, and improve performance of state-of-the-art network by 4.67 in terms of mIoU(%)

Keywords - Endoscopy Surgery; Robot-assisted Surgery; Semantic Segmentation; Deep Learning; Neural Network; Data Augmentation; Auxiliary Loss; Intermediate Layer Supervision.



Contents

口試委員會審定書	i
Acknowledgments	ii
摘要	iii
Abstract	iv
1 Introduction	1
1.1 Motivation	1
1.2 Contribution	3
1.3 Thesis Organization	4
2 Background and Related Work	5
2.1 Semantic Segmentation	5
2.2 Data Augmentation	6
2.3 Computer-assisted Endoscopy Surgery	7
3 Problem Definition and System Architecture	9
3.1 Problem Definition	9
3.1.1 Semantic Segmentation in Endoscopy Environment	9
3.1.2 Issues in Endoscopy Environment	10
3.2 System Architecture	11
3.2.1 Network Architecture	12
3.2.2 ResNet	12
3.2.3 Pyramid Pooling Module	14

4	Design and Implementation	15
4.1	Collage	15
4.2	Auxiliary Loss With Meta-Class	17
5	Performance Evaluation	20
5.1	Experiment Environment	20
5.2	Evaluation Results	21
5.2.1	Ablation Study for Commonly Used Data Augmentation	21
5.2.2	Ablation Study for Collage	21
5.2.3	Ablation Study for Auxiliary Loss with Meta-Class	23
5.2.4	Combined Study	24
6	Conclusion	26
	Bibliography	27





List of Figures

3.1	System Architecture	11
3.2	Network Architecture Used in This Work.	12
3.3	Building Block of ResNet	12
4.1	Example of Collage	16
4.2	Example of Meta-Class	18
4.3	Training Network with Auxiliary losses and Meta-Classes	19
5.1	Qualitative Results.	25



List of Tables

3.1	Architecture of ResNet101 Used in This work	13
4.1	Class and Meta-Class	17
5.1	Evaluation of commonly used image processing for data augmentation.	22
5.2	Experiment of collage with different setting.	22
5.3	Experiment of class specific augmented collages.	23
5.4	Empirically tuning parameter λ_1 and λ_2	24
5.5	Ablation experiment for meta-class.	24
5.6	Experiment on proposed method alone and combined.	24



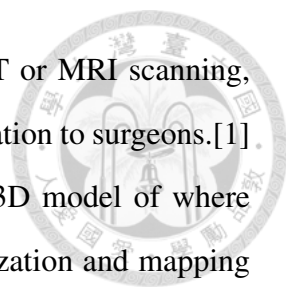
Chapter 1

Introduction

1.1 Motivation

Minimally invasive surgeries (MIS) are surgical procedures that only require minimum size of incision, and thus have advantages like fewer pain, less recovery time, and less infection risk compared to traditional open surgeries, which require considerable size of incision. Most MIS involves putting an endoscope along with other surgical tools inside patient's body through small holes, and surgeon will perform the surgery with video captured by endoscope as guidance. This kind of endoscope-based surgery have very different surgery environment from open surgery and thus extra training is demanded. For example, in endoscopy surgery surgeon usually needs to control mechanism at one end of a long shaft to perform sophisticated operation at the other end of the shaft with only video from endoscope as guidance, which is very counter-intuitive for most people.

Many efforts have been made to aid surgeons when performing MIS with the help of technology. For example, the da Vinci Surgical System provides surgeon a more comfortable surgery environment. It has two cameras on it's endoscope and provides stereo videos to surgeons, which helps them to have better sense of hand-eye coordination with surgical tools. It also has a better interface to control the surgical tools, which enables surgeons to perform surgeries more precisely and stably. Many other works involve processing images from the endoscope and provide additional information to surgeons. Some try to recognise and track target organ during surgery, and fit interior information of the target organ

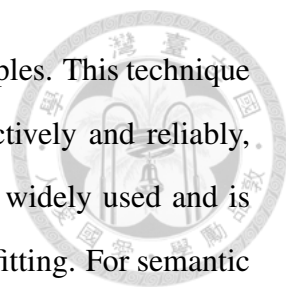


like location of blood vessels or lesion, acquired by pre-surgical CT or MRI scanning, with endoscope image using tracking results, to provide more information to surgeons.[1] [2] try to track the location of endoscopy camera and reconstruct 3D model of where the surgery took place at the same time, namely simultaneous localization and mapping (SLAM), which is commonly used in robotic applications. Results of endoscopy-SLAM include 3d reconstructed model of surgical site, or 'map' of surgical site, and frame-by-frame position of endoscope in the 3d model. The reconstructed 3d model can then be used to match with pre-surgical CT or MRI scan reconstruction to fuse information from both side together, so that surgeons can have more thorough understanding during surgery. Among all these applications, one of the most important preliminary requirement is to recognize objects within view of endoscope.

Object recognition is one of the most intensively-studied research topics in the domain of computer vision, and has achieved great breakthrough thanks to the development of Deep Neural Network (DNN). Networks like YOLO [3] or Faster R-CNN [4] can accurately locate objects of interest in a given image in bounding-box fashion. On the other hand, networks like FCN [5], DeepLab [6], PSPNet [7], etc, which can do prediction on pixel level, provide finer details about the scene in a given image, for example which part of the scene is road and which part is cars. With these information, machine or robot can gain better understanding about the world around it, which is vital in some applications that require accurate and detailed understanding about the environment, e.g. autonomous driving and robot-assisted surgery.

Although above-mentioned networks can provide scene information on pixel level, they require a large amount of finely annotated data for training. In the domain of autonomous driving, there are many publicly available datasets, for example KITTI [8], Cityscapes [9] and CamVid [10], which contain a large volume of data about urban street scene taken from cameras mounted on cars, and pixel-level annotations which can be used to train deep neural network. However, data for robot-assisted surgery is still scarce.

There are many issues for training deep network with only few data, e.g. overfitting, resulting degraded network performance. One effective technique to prevent these issue is

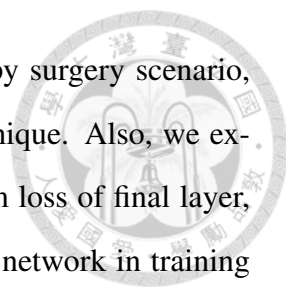


data augmentation, which tries to increase the number of training samples. This technique can help classifier learns invariance of the given dataset more effectively and reliably, which is important for generalization. Data augmentation has been widely used and is proved to be effective for network generalization and preventing overfitting. For semantic segmentation, random scaling, horizontal flipping and random cropping are commonly-used technique for data augmentation. [11] proposed to apply random elastic transformation on microscopical images because non-rigid movement is frequently seen on cells or tissues.

Moreover, scenario of robot-assisted endoscopy surgery exhibits some characteristics that might be beneficial for applying data augmentation on small dataset. First, objects needed to recognize usually have high appearance invariance in endoscopy surgery. During a surgery, surgical instruments, target organ of the surgery, and lesion or anomaly are three major categories of objects that might be interesting to surgeons. Here we will focus on the first two categories, because recognition of lesion or anomaly usually requires diagnosis from professionals, and thus is beyond the scope of our research interest at this moment. Instruments used in surgeries of the same type seldom vary, and thus a relative small number of data should be enough to contain most variance. Similarly, organs or tissues that might be observed in surgeries of the same type are mostly the same. Although appearance of these organs or tissues might change from patient to patient, it's difference won't be as much as vehicles or buildings from street scene scenario. Thus a relative small number of data should be enough to demonstrate to the network how these organs or tissues might vary. Second, scenes in endoscopy surgery are relatively simple because the number of objects that might appear at the same time is small. Also, distance between objects and camera in endoscopy surgery is limited, thus the scale of objects only change marginally.

1.2 Contribution

In this thesis, we propose a new data augmentation technique called collage, which aggressively generates new training samples by collecting different parts of original train-



ing data. We evaluated the effectiveness of collage in the endoscopy surgery scenario, and compared it with other commonly used data augmentation technique. Also, we extend the concept of auxiliary loss proposed by [7], which, apart from loss of final layer, adds loss computed with results from intermediate layer to help the network in training phase. We design the auxiliary loss to guide our network to predict meta-classes of the input data in intermediate layer, which will help successive layers predict the actual class better. A meta-class is a group of classes that share similar characteristics, for example, all the organs are in the same meta-class, so are the surgical instruments. Experiment results show that proposed data augmentation can improve network performance better than other commonly used data augmentation methods. Further, collage and auxiliary loss combined can improve performance of state-of-the-art network by 4.67 in terms of mIoU(%)

1.3 Thesis Organization

The rest of this thesis is organized as follow: background and many related research will be described in chapter 2, main architecture of our endoscopy semantic segmentation system will be described in chapter 3, design details of our algorithm will be elaborated in chapter 4, and experiment setting and results will be presented in chapter 5. Finally, conclusion is made in chapter 6.



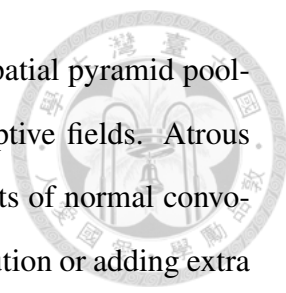
Chapter 2

Background and Related Work

2.1 Semantic Segmentation

Semantic segmentation is a problem that tries to recognize and segment objects of interest simultaneously in an image. Unlike bounding-box object detection which describes object location by a rectangle that bounds the target precisely, semantic segmentation will do pixel-wise classification on image, so each and every pixel in an image will have its own label after classification. With this dense pixel-wise classification, semantic segmentation can provide information with finer detail, e.g. shape of the objects or the exact boundaries between each object or between object and background.

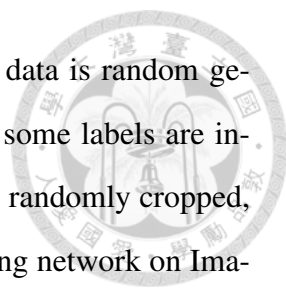
Long et. al [5] proposed a novel architecture called fully convolutional network (FCN) that removes fully-connected layer commonly seen in the end of a network which is served as classification layer in a image classification network, and replace it with convolutional layer with kernel size 1×1 , which is equivalent to do per-pixel classification on the output of previous layer, and thus enables the network to output pixel-wise prediction. Since then, many works based on fully convolutional architecture have been proposed for semantic segmentation. Some adopted encoder-decoder architecture, in which encoder gradually down-samples the size of feature map and extracts features of higher-level with larger receptive field, while decoder gradually recovers extracted feature map or prediction to original scale of input image [11], [12]. Skip connections between encoder and decoder of the same scale are often used in such architecture to forward fine details which



may get lost during down-sample to decoders. [6] proposed atrous spatial pyramid pooling (ASPP) which can capture features of different scales and receptive fields. Atrous convolution, or dilated convolution, adds holes between sample points of normal convolution, and thus increase field of view without sacrificing image resolution or adding extra computation. [7] proposed pyramid pooling module, which do global pooling of different scales on the feature map outputted by a deep network, and then rescale and concatenates results of global pooling with original feature map, which is then used for final prediction. It also proposed to use auxiliary loss in intermediate layer to help deep networks learn better. Based on [7], [13] proposed a multi-scale cascade network architecture, which has image of different scales as input, and only the image with lowest resolution will go through the main branch of the network, while images of higher resolution will only go through a sub-branch of the network to provide finer information to the main branch. This cascade network can achieve real-time performance in machine with single modern GPU, with only minor accuracy drop. Our segmentation architecture is based on [7].

2.2 Data Augmentation

Data augmentation is a practical technique that is widely used in most of the machine learning applications, here we will focus on methods used in image classification applications, especially semantic segmentation. The motivation behind data augmentation is simple, an image of dog, for example, will still be an image of dog even if it undergoes some simple image processing like scaling, rotating or gray scale conversion. Thus we can safely add those processed images labelled dog into the dataset that will be used to train a dog classifier, expanding the number of training samples. Benefit of data augmentation is twofold. First, by doing data augmentation, we are effectively adding variance into training set, which in turn can help the classifier learns the invariance of the data robustly, and thus make the classifier generalize better. Second, if we have prior knowledge about how testing data will distribute, we can synthesize some training data that reflect such distribution and add them into training set, so that the distribution of training set will be more similar to that of testing set.



One of commonly used data augmentation techniques on image data is random geometrical transformation or color conversion. As mentioned above, some labels are invariant under certain image transformations or color conversion. [14] randomly cropped, horizontally flipped and altered color intensity of images when training network on ImageNet dataset, which is crucial for their network to fight against overfitting. [15] proposed to do elastic transformation on training data when training network on MNIST dataset. Elastic transformation will randomly compute a displacement field, which indicates how each pixel will transform to a new location. [11] also used elastic transformation on microscopical image dataset, and it described elastic transformation as "the key concept to train a segmentation network with very few annotated images".

Another commonly used technique is data synthesis, which usually involves in generating training data through rendering 3d models [16] [17] [18] [19] [20]. One of the main advantages of data synthesis is that ground truth of synthesized training data can be acquired automatically, which relieves human labor needed for annotation, and makes collecting large quantity of training data possible. Network pre-trained on synthesized data can achieve improved performance in real data [16].

Data augmentation is usually data-dependent, and requires expert knowledge to choose the best data augmentation technique for a given dataset. However, recently [21] proposed a new data augmentation technique called mixup that is experimentally proven to be effective for many different scenario, from image classification to speech recognition. Mixup linearly combines two training samples and their labels which can be seem as generating new training sample by linear interpolation on training distribution.[21] argues that using mixup encouraging classifier, or any learning model, to behave linearly in-between training examples, which helps model generalize better since linear model is the most simple model possible.

2.3 Computer-assisted Endoscopy Surgery

In recent years, endoscopy surgery has grown in popularity thanks to its advantages over traditional open surgery, i.e. minimal incision requirement and thus reduced pain and

recovery time. As a result, many researches focusing on videos from endoscopy surgery emerged. [22] provided a thorough survey on researches focusing on endoscope video or image.

Robot-assisted surgery is one of the active topic in endoscopy research, which aims to provide surgeons more capability, dexterity and stability during surgery with the help of computer and microprocessor. Currently, passively-assisting robotic surgery system has been piratically used for years [23] [24], where the robotic system passively enhances surgeon's ability and has no autonomy. Many have tried to extend autonomy of robotic system. For example, [25] [26] [27] [28] studied the ability of autonomous endoscope holding, and [29] [30] use virtual fixture to define forbidden area and guide robotic instruments not to perform erroneous action. More high-leveled behaviors are also studied [31] [32] [33]. All these applications require detailed understanding of surgical scene around robotic instruments, which can be provided by surgical scene semantic segmentation.



Chapter 3

Problem Definition and System

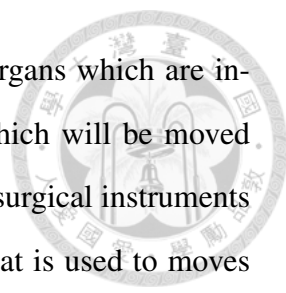
Architecture

3.1 Problem Definition

In this work we try to solve the problem of semantic segmentation on endoscopy environment. While extensive researches have been conducted in the domain of semantic segmentation on urban street environment or generic natural environment, but related works on endoscopy environment remains scarce. In below we are going to define the problem in detail, including definition of semantic segmentation in endoscopy environment, and specific issues one will encounter in endoscopy environment which our work tries to solve.

3.1.1 Semantic Segmentation in Endoscopy Environment

The problem of semantic segmentation is to, given an image $I \in R^{h \times w \times c}$, predict the class for each pixel $L = C(I)$, $L \in R^{h \times w}$, where L is the per-pixel prediction, C is the classifier, and m, n, c are height, width and number of channels of the image respectively. In the case of endoscopy surgery, the classes we are trying to predict can be divided into four categories. The first category is organs which are the primary targets in the surgery, e.g. gall bladder for cholecystectomy surgery (removal of gall bladder), and kidney for

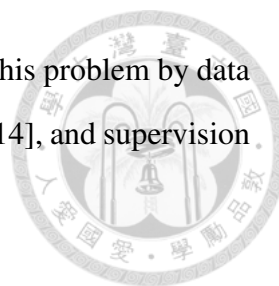


nephrectomy surgery (removal of kidney). The second category is organs which are involved in the surgery although not the primary target, e.g. liver which will be moved around during cholecystectomy surgery. The third category is major surgical instruments which conduct most of the surgical procedures, including grasper that is used to move and hold still organs, cauterizer or scissors that is used to cut tissue apart, etc. The final category is supplementary or supportive instruments, for example thread or clip which is used to block ducts. In this work, we will focus on nephrectomy procedures performed on porcine using da Vinci Xi surgical system, and the classes we are interested in include kidney, small intestine, da Vinci surgical instruments, Suturing thread, and clamps.

3.1.2 Issues in Endoscopy Environment

Most state-of-the-art methods use deep neural network (DNN) as classifier in the problem of semantic segmentation, which requires a large quantity of finely annotated, pixel-wise annotations as training labels for supervised training. However, collecting training data for endoscopy surgery is very difficult due to two reasons. First, using videos from endoscopy surgery is under regulation. One will need approval from authority to use these data for research purpose, which requires drafting proposal to justify the usage, and go through examinations from bureaucracy. These procedures will halt the pace of research for an indefinite time. Further, such approval usually need to apply for one type of surgery at a time, which make collecting data from a variety of surgeries very inefficient, forcing researches to focus on one type of surgeries at a time, making learning-based classifiers prone to overfitting to one specific type of surgery, and lacking in the ability to generalize to other surgeries with similar settings. Second, annotating data from endoscopy surgery is difficult. It might require expert knowledge to annotate anatomical objects correctly because many of them have similar appearance and thus it is hard to separate one from the other.

As a result, data available for semantic segmentation in endoscopy environment is few, which makes it difficult to train deep neural network, and thus prevents us from utilizing deep neural net architectures like [6], [7], which have be proved to be very effective in



the domain of semantic segmentation. In this work, we try to handle this problem by data augmentation, which is key to train deep network with few data [11] [14], and supervision on intermediate layers.

3.2 System Architecture

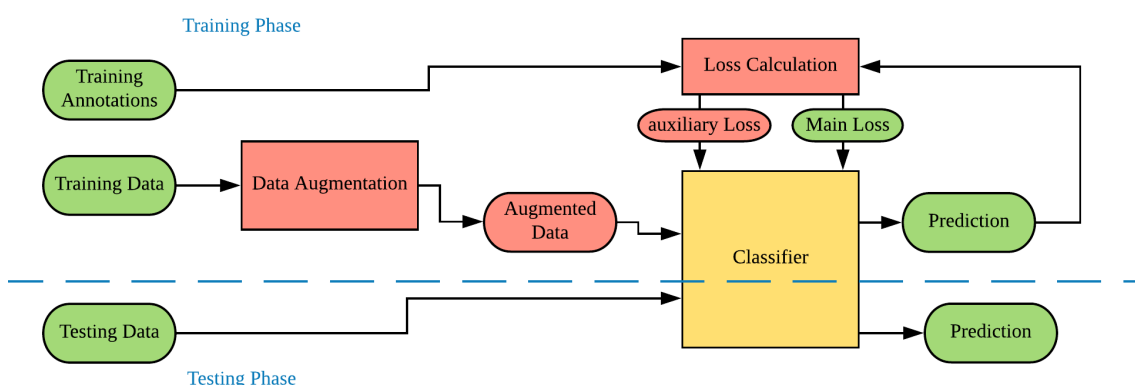


Figure 3.1: System Architecture

Figure 3.1 shows our system architecture. The core of our system is a classifier for semantic segmentation on images from endoscopy surgery, which is expected to be a deep neural net. Before the classifier is a data augmentation model which will drastically expand the original training set by aggressive data augmentation which will be elaborated in chapter 4, so that our deep neural net classifier can be trained effectively on the original low-volume training set. During training, we additionally use two auxiliary losses [7] on two different intermediate layers to guide our network. We use category or 'meta class' of the original label instead to compute auxiliary losses. We argue that this is effectively guiding the network to learn a hierarchical class predicting procedure, that shallow layer will first try to predict the broadest class, i.e. artificial or anatomical objects in our case, and the prediction will become narrower in deeper layers, and the final layer will have the narrowest prediction, i.e. the original class. With this hierarchical prediction and auxiliary losses, the network can learn representation more efficiently and thus increase the accuracy.



3.2.1 Network Architecture

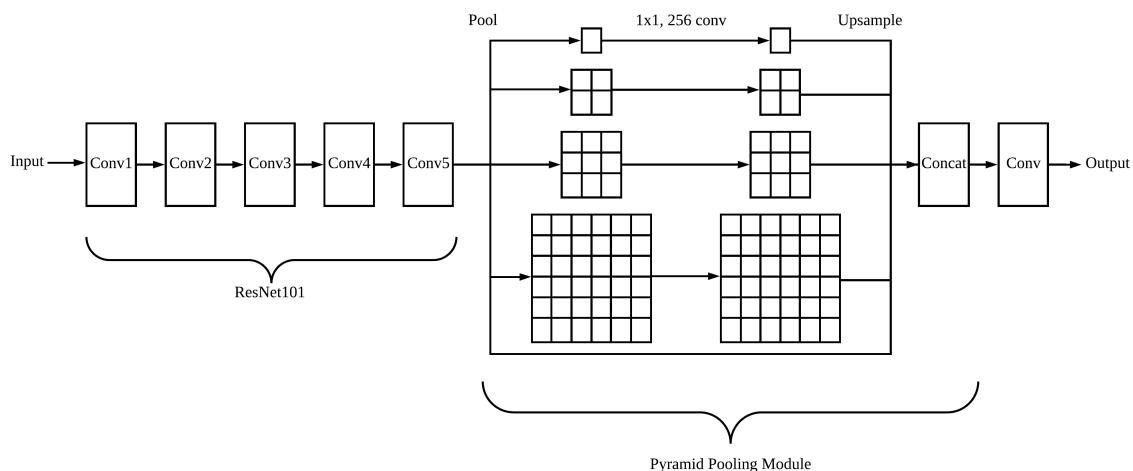


Figure 3.2: Network Architecture Used in This Work.

The classifier used in our work is PSPNet101 [7], which is built upon ResNet [34] and adds a global pyramid pooling module at the end of the network. Network architecture is shown in figure 3.2.

3.2.2 ResNet

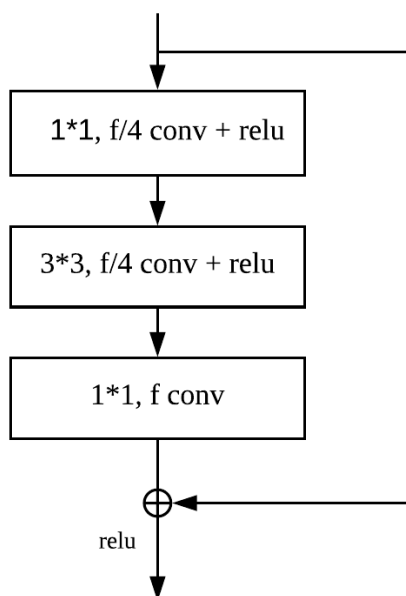
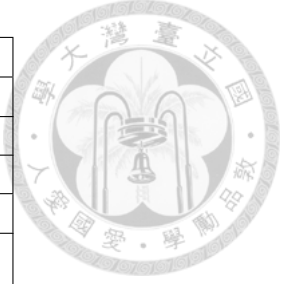


Figure 3.3: Building Block of ResNet

ResNet [34] introduced shortcut connection between layers and residual learning and



Layer Name	Constitution
Conv1	$3 \times 3, 64, \text{stride } 2$
	$3 \times 3, 64, \text{stride } 1$
	$3 \times 3, 128, \text{stride } 1$
Conv2	$3 \times 3 \text{ max pool, stride } 2$
	$1 \times 1, 64$
	$3 \times 3, 64 \times 3$ $1 \times 1, 64$
Conv3	$1 \times 1, 128$
	$3 \times 3, 128 \times 4$
	$1 \times 1, 512$
Conv4	$1 \times 1, 256$
	$3 \times 3, 256, \text{dilaterate} = 2 \times 23$
	$1 \times 1, 1024$
Conv5	$1 \times 1, 512$
	$3 \times 3, 512, \text{dilaterate} = 4 \times 3$
	$1 \times 1, 2048$

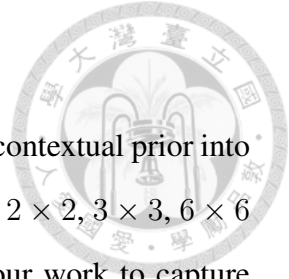
Table 3.1: Architecture of ResNet101 Used in This work

thus made training a very deep network possible. Basic building block for ResNet with bottleneck and identity shortcut connection is shown in figure 3.3, where f is the size of feature, and the whole ResNet is constructed by stacking many blocks together. The network first reduces the size of feature by a convolution with kernel sized 1×1 , then applies actual feature extraction on the narrowed feature map, and finally increase the size of feature back to original size, leaving a wide-narrow-wide feature map transformation behavior, thus called bottleneck. This bottleneck design is useful for reducing number of parameters and thus reducing the difficulty of network training. After, output of the bottleneck is added by its input, so the network is not directly learning the representation, but its residual compared to previous result.

Table 3.1 lists detailed architecture of ResNet101 used in our work. Inspired by [35], the first convolutional layer with kernel sized 7×7 in original ResNet is replaced by three conv layers with 3×3 kernel. In Conv4 and Conv5, dilated convolution [6] is used to increase receptive field without reducing feature map resolution.

3.2.3 Pyramid Pooling Module

Pyramid pooling module is proposed by [7] to incorporate global contextual prior into the network. Following [7], 4 pooling operations with bin size 1×1 , 2×2 , 3×3 , 6×6 respectively are applied to the output feature map of resnet101 in our work to capture global representation in different scales. Averaging pooling is used because it performs better than max pooling. After pooling, 1×1 convolution is applied to each pooled representations to reduce their feature dimension to $1/4$ of original size to maintain weight of global context. Finally, all pooled representations are upsampled to original size and are fused together with input feature map by concatenating them altogether.





Chapter 4

Design and Implementation

4.1 Collage

Here we describe proposed data augmentation technique named "collage", which generate new training data and label by gluing together pieces that are cut from other training samples. The name is originated from a technique of art production, where the artwork is made from an assemblage of different forms, thus creating a new whole.

Algorithm 1 is the algorithm of collage. It will iterate through all possible label l , randomly select a training image I_i and annotation L_i pair, and copy all pixels with class l in I_i and L_i to collaged image and annotation I_c, L_c . Figure 4.1 are some example images and annotations generated by collaging. We don't explicitly try to fill each and every pixel during collaging, because it would be computation-demanding, and even impossible in some cases. So there might be blank areas in the final collaged image, which will be ignored during training. During collaging, we will take contextual relationship into consideration, i.e. background in the original images will still be background in collaged images, and other objects will be added upon it, and artificial objects will always be on top of anatomical objects.

One of the major advantage of collage is that it increase inter-class contextual variance of training data. In endoscopy surgery scenario, training data is usually collected by first choosing many sequences, each with one minute or so duration, from video of one or many surgeries. Then annotations are made on down-sampled frames, e.g. annotation

might be made on every thirty frames, reducing annotation rate to 2Hz from original 60Hz video. Because each sequence has very short duration compared to the whole surgery, frames in each sequence might have highly biased inter-object correlations. Such correlations might not be desirable, for example preliminary experiment shows that pixels near surgical instruments are more likely to be predicted as organ of primary surgery target, because in training set these classes have high co-occurrence, even though the pixels are actually from other anatomical objects. By randomly collaging objects from different frames of different sequence, we can reduce undesirable inter-class correlations, and encourage network to focus more on intra-class invariance.

Algorithm 1 Collage

Input: The training image set, I_t ; The training annotation set, L_t ;

Output: The collaged image, I_c and corresponding annotation L_c ;

- 1: **for** l in all Label **do**
 - 2: randomly choose an Image I_i from I_t , with it's corresponding annotation L_i .
 - 3: $I_c(p) = I_i(p)$ and $L_c(p) = L_i(p) \forall p \in P$, where P is the set of all pixels in I_i that are in class l .
 - 4: **end for**
 - 5: **return** I_c, L_c ;
-

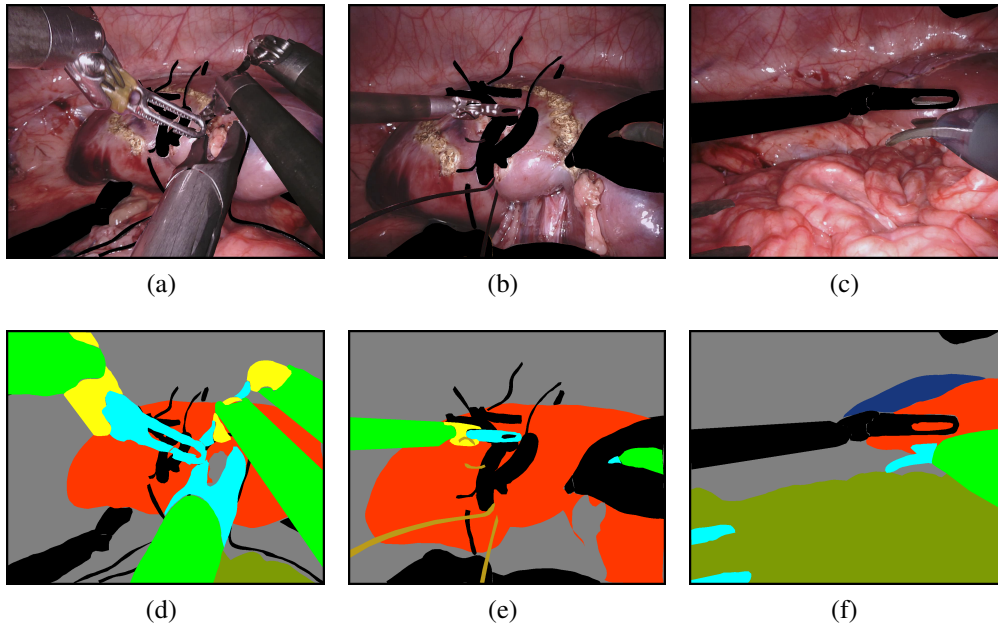



Figure 4.1: Example of Collage

4.2 Auxiliary Loss With Meta-Class



Meta-Class 1	Meta-Class 2	Class
Artificial Objects	Surgical Instruments	Instrument Shaft
		Instrument Clasper
		Instrument Wrist
	Other Artificial Objects	Thread
		Clamp
		Suction Instrument
Anatomical Objects	Kidney	Kidney
	Small Intestine	Small Intestine
	Others	Others

Table 4.1: Class and Meta-Class

Zhao, et. al [7] proposed to use auxiliary loss in intermediate layer to help deep network optimize better. Here we further extend this idea by substituting class label involved in auxiliary loss computation with meta-class label. Meta-classes are collections of classes that share similar properties. Table 4.1 lists all class and meta-class that we are interested in, and their relation. We use two layers of meta-class, and hence two auxiliary losses during training. The first meta-classes categorize if the object is artificial or anatomical, and are the broadest classifications. The second meta-classes narrows down the classification according the property of each class. Figure 4.2 shows an example of a frame labelled in different level of meta-class.

Figure 4.3 shows the details of how auxiliary losses are applied to the network in training step. Auxiliary loss for the first meta-classes is applied on res4b11 residual block, and the second one is applied on res4b22 residual block. The auxiliary losses are defined as cross entropy between respective meta-classes and predicted meta-classes. With these auxiliary losses, we are effectively guiding the network to learn representation hierarchically, so the shallow layers will learn features that can best predict the first meta-class of each pixel which is the easiest among all three, while the middle layers will learn features upon those computed by the shallow layers that can best predict the second meta-class of each pixel, and finally the deep layers will learn features upon those computed by the middle layers that can best predict the original class of each pixel which is the most difficult among all three. The network can optimize more efficiently with this hierarchical

representation learning scheme, which is especially crucial because we are training the network with only few data. Weights are added to auxiliary losses to balance their importance. The network will predict meta-class for auxiliary losses only in training phase. During testing, the network only outputs predictions from the last layer.

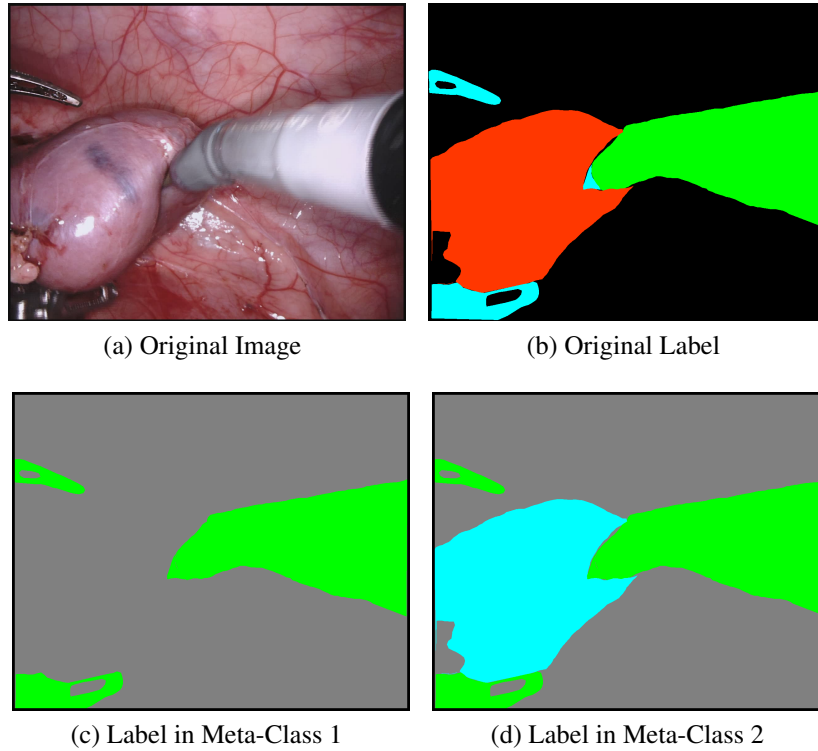
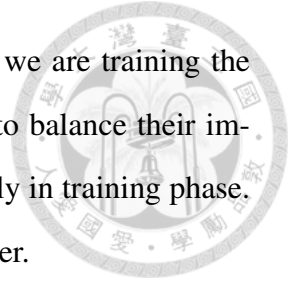


Figure 4.2: Example of Meta-Class

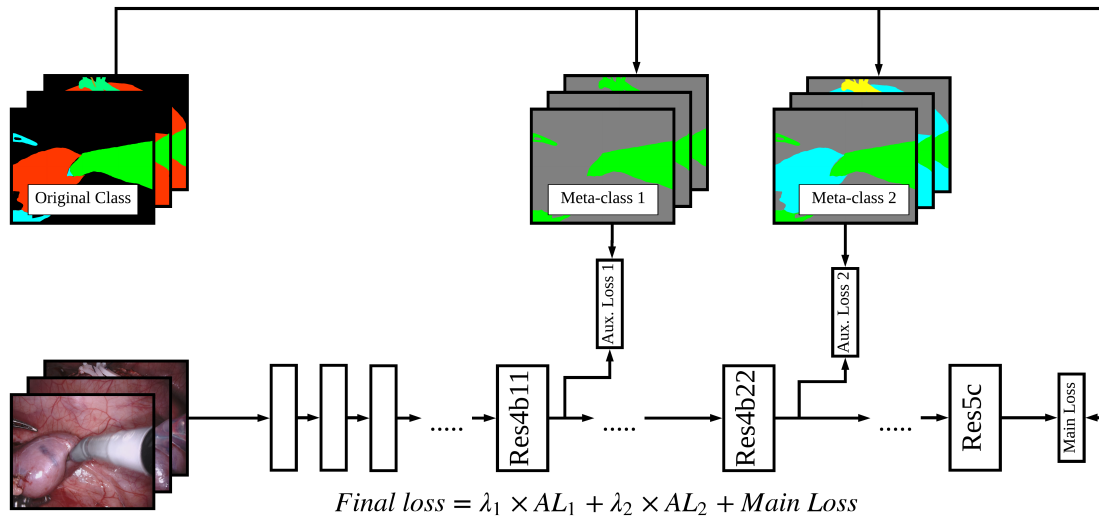


Figure 4.3: Training Network with Auxiliary losses and Meta-Classes



Chapter 5

Performance Evaluation

5.1 Experiment Environment

We adopted the Tensorflow [36] implementation of PSPNet [37]. The batch size is 1 because of GPU memory limitation. Learning rate, momentum and weight decay are 0.00005, 0.9 and 0.0001 respectively. Poly learning rate policy [6] is used in training phase, where the learning rate will be multiplied by $(1 - \frac{step}{max_step})^{power}$, and power is set to 0.9. Networks are pre-trained on ImageNet dataset, and are trained for 35K iterations in all experiments.

Dataset used in the experiments are from Miccai 2018 Robotic Scene Segmentation Sub-Challenge [38], which contains 16 images sequences recorded in robotic porcine nephrectomy procedures using da Vinci Xi systems. Annotation frame rate is 2Hz out of originally 60Hz videos, and each sequence has 149 annotated frames. We used 4 of the 16 sequences for experiments, 3 of which are used as training set and 1 of which is used as validation set. Mean-intersection-over-union(mIoU) is used as evaluation metrics.



5.2 Evaluation Results

5.2.1 Ablation Study for Commonly Used Data Augmentation

Data augmentation is beneficial for network to generalize better, but effectiveness of each data augmentation is data dependent. Here we evaluate many image processing methods that are commonly used in semantic segmentation as data augmentation, including random horizontal mirroring, random scaling, and random distortion [15], and the per-class IoU result is shown in table 5.1.

As shown in table 5.1, while all three methods are beneficial to the network in terms of mIoU, their improvement is insignificant because of the properties of endoscopy dataset. First, in the dataset, endoscopes always approach to surgical cite from the same direction, so that random mirroring will not provide much useful variance like it does to other scenario, e.g. urban street. Also, there is strong invariance of from which side of frame will a certain type of surgical instrument appears, and this invariance will be break by random mirroring, which could explain the minor drop of instrument clasper's mIoU after applied random mirroring. Second, because all objects will be very close to endoscope during surgery, their scale can only change little, so that random scaling also won't be as effective as in other dataset. Finally, while random distortion benefits a little for classes exhibit non-rigid transformation, e.g. clamps or small intestine, it adds undesirable variance to rigid objects like most of surgical instruments. Concluded from above, these commonly used data augmentation methods are not effective in endoscope surgery scenario.

5.2.2 Ablation Study for Collage

The proposed data augmentation technique, collage, can effectively increase the volume of training data and help the network to fight against overfitting. To evaluate collage, we generate training data by collage with different factor f , which controls the number of collaged data such that additional $f \times n$ collaged images will be added to training set, where n is the number of original training set. A baseline network is trained with only original training data while the other is trained with collaged data and original training



Model	No Data Augmentation	Random Mirroring	Random Scaling	Random Distortion
Background Organ IoU(%)	72.65	71.76	73.64	72.39
Instrument Shaft IoU(%)	68.73	69.23	70.58	68.02
Instrument Clasper IoU(%)	48.14	47.29	53.17	50.67
Instrument Wrist IoU(%)	36.65	40.91	42.17	30.77
Kidney IoU(%)	43.36	47.43	47.29	46.88
Thread IoU(%)	19.30	18.22	16.41	18.28
Clamp IoU(%)	5.28	9.16	9.98	11.68
Suction Instrument IoU(%)	5.02	1.49	2.63	2.23
Small Intestine IoU(%)	79.38	77.16	66.89	80.43
mean IoU(%)	42.06	42.52	42.47	42.31

Table 5.1: Evaluation of commonly used image processing for data augmentation.

Collage Factor f	mean IoU(%)
Baseline(0x)	42.06
1x	42.16
2x	42.37
4x	42.85
8x	43.77
16x	43.96
32x	42.62

Table 5.2: Experiment of collage with different setting.

Model	mIoU(%)
Not Aug	45.42
Aug	35.07



Table 5.3: Experiment of class specific augmented collages.

data combined.

Experiment result in table 5.2 shows that performance improved as the number of collaged data increased while saturated at factor 16x, and the best performance is 1.90 (in terms of mIoU(%)) better than baseline.

Inspired by observation of above data augmentation experiments that each class has different responses to different data augmentation techniques, i.e. some augmentations have positive effect on given class while others are not, we tried to evaluate if it is helpful to use different data augmentation on different class. Thus we proposed a naive combination of class specific data augmentation technique and collage, which will apply different augmentation on different classes during collage, attempting to optimize effectiveness of each data augmentation. For example, when collage an instrument part, random scaling will be applied but not distortion, and when collage small intestine, only random distortion is applied. Result of this class specific augmented collages is shown in table 5.3, which suggests that this naive combination is not suited for boosting network performance.

5.2.3 Ablation Study for Auxiliary Loss with Meta-Class

Training network with proposed auxiliary loss with meta-class helps the network to learn representation hierarchically, which is especially effective when training deep network with few data. Here we evaluate the effectiveness of both auxiliary loss and meta-class, by comparing network trained with both auxiliary loss, and meta-class with baseline network, and network trained with auxiliary loss but without meta-class.

Table 5.4 shows that auxiliary loss training with $\lambda_1 = 0.5, \lambda_2 = 1.0$ yields the best performance, with improvement of 3.99 in terms of mIoU(%) over baseline. Table 5.5 shows that supervision with meta-class is crucial for auxiliary loss training, because naive auxiliary loss training without meta-class results in degraded performance, which is even

worse than baseline.



Model	mean IoU(%)
No AL	42.06
$\lambda_1 = 0.2, \lambda_2 = 0.6$	40.00
$\lambda_1 = 0.4, \lambda_2 = 0.2$	43.37
$\lambda_1 = 0.4, \lambda_2 = 0.4$	43.12
$\lambda_1 = 0.4, \lambda_2 = 0.6$	42.63
$\lambda_1 = 0.4, \lambda_2 = 0.8$	43.49
$\lambda_1 = 0.5, \lambda_2 = 0.8$	43.98
$\lambda_1 = 0.5, \lambda_2 = 1.0$	46.05
$\lambda_1 = 0.6, \lambda_2 = 0.4$	42.98
$\lambda_1 = 0.6, \lambda_2 = 0.6$	41.44
$\lambda_1 = 0.6, \lambda_2 = 0.8$	43.84
$\lambda_1 = 0.8, \lambda_2 = 0.6$	40.05
$\lambda_1 = 1.0, \lambda_2 = 1.0$	41.53
$\lambda_1 = 1.0, \lambda_2 = 2.0$	42.78
$\lambda_1 = 2.0, \lambda_2 = 1.0$	43.16

Table 5.4: Empirically tuning parameter λ_1 and λ_2 .

Model	mean IoU(%)
AL with meta-classes($\lambda_1 = 0.4, \lambda_2 = 0.6$)	42.63
AL without meta-classes($\lambda_1 = 0.4, \lambda_2 = 0.6$)	35.64

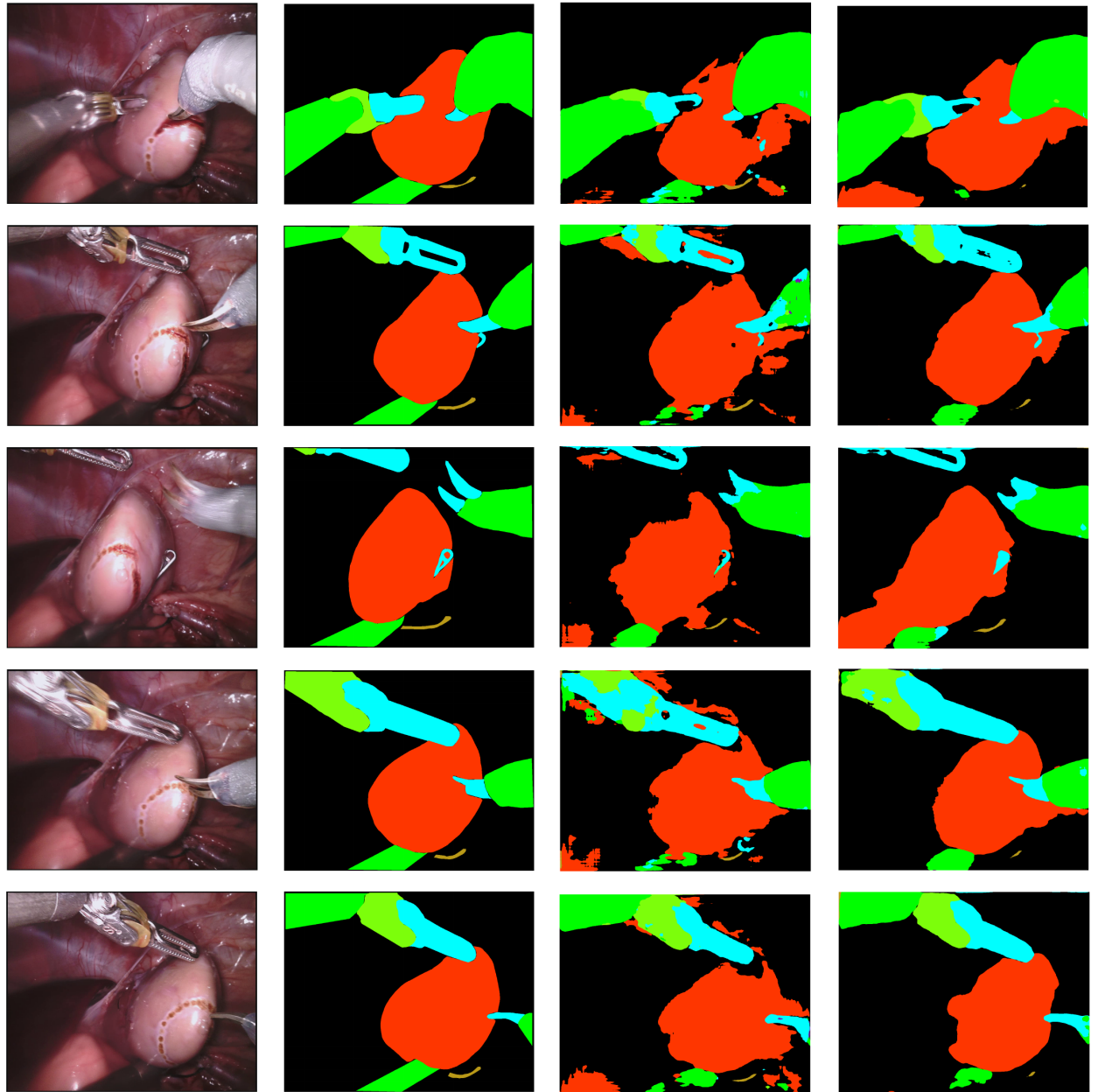
Table 5.5: Ablation experiment for meta-class.

5.2.4 Combined Study

Table 5.6 shows performance comparison of proposed methods alone and combined. Experiment results show that optimal λ_1 and λ_2 are different for networks trained with collaged dataset and without collaged dataset. And with proper parameters, combined method can achieve better performance, which outperform baseline by 4.67 in terms of mIoU(%).

Model	mIoU(%)
PSPNet101	42.06
PSPNet101+AL($\lambda_1 = 0.5, \lambda_2 = 1.0$)	46.05
PSPNet101+Collages(8x)	43.77
PSPNet+AL($\lambda_1 = 0.5, \lambda_2 = 1.0$)+Collages(8x)	44.53
PSPNet+AL($\lambda_1 = 1.0, \lambda_2 = 2.0$)+Collages(8x)	46.73

Table 5.6: Experiment on proposed method alone and combined.



(a) Image

(b) Ground Truth

(c) PSPNet

(c) PSPNet+AL+Collage

Figure 5.1: Qualitative Results.



Chapter 6

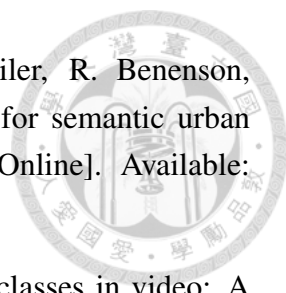
Conclusion

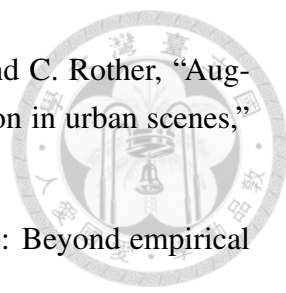
In this thesis, a new data augmentation technique and a training scheme are proposed. The proposed data augmentation randomly samples objects or background from different images and then collage them together. Experiment results showed that collage is more effective than many commonly used data augmentation methods in endoscopy surgery scenario. The proposed training scheme uses additional supervision in intermediate layers with meta-class is proposed, which encourages network to train representation hierarchically. The two methods combined help us achieve improved performance on state-of-the-art network by a large margin.

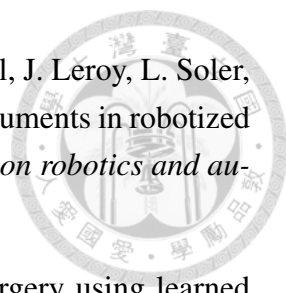


Bibliography

- [1] J. Song, J. Wang, L. Zhao, S. Huang, and G. Dissanayake, "MIS-SLAM: real-time large scale dense deformable SLAM system in minimal invasive surgery based on heterogeneous computing," *CoRR*, vol. abs/1803.02009, 2018. [Online]. Available: <http://arxiv.org/abs/1803.02009>
- [2] N. Mahmoud, I. Cirauqui, A. Hostettler, C. Doignon, L. Soler, J. Marescaux, and J. M. M. Montiel, "Orb slam-based endoscope tracking and 3d reconstruction," *CoRR*, vol. abs/1608.08149, 2016. [Online]. Available: <http://arxiv.org/abs/1608.08149>
- [3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [5] 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.
- [6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
- [7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.
- [8] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

- 
- [9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” *CoRR*, vol. abs/1604.01685, 2016. [Online]. Available: <http://arxiv.org/abs/1604.01685>
- [10] G. J. Brostow, J. Fauqueur, and R. Cipolla, “Semantic object classes in video: A high-definition ground truth database,” *Pattern Recognition Letters*, vol. xx, no. x, pp. xx–xx, 2008.
- [11] 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.
- [12] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *CoRR*, vol. abs/1511.00561, 2015. [Online]. Available: <http://arxiv.org/abs/1511.00561>
- [13] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, “Icnet for real-time semantic segmentation on high-resolution images,” *CoRR*, vol. abs/1704.08545, 2017. [Online]. Available: <http://arxiv.org/abs/1704.08545>
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [15] P. Y. Simard, D. Steinkraus, and J. C. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” in *null*. IEEE, 2003, p. 958.
- [16] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, “Virtual worlds as proxy for multi-object tracking analysis,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4340–4349.
- [17] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla, “Understanding real world indoor scenes with synthetic data,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4077–4085.
- [18] Y. Movshovitz-Attias, T. Kanade, and Y. Sheikh, “How useful is photo-realistic rendering for visual learning?” in *European Conference on Computer Vision*. Springer, 2016, pp. 202–217.
- [19] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.

- 
- [20] H. A. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, “Augmented reality meets deep learning for car instance segmentation in urban scenes,” in *British Machine Vision Conference*, vol. 1, 2017, p. 2.
- [21] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.
- [22] B. Münzer, K. Schoeffmann, and L. Böszörményi, “Content-based processing and analysis of endoscopic images and videos: A survey,” *Multimedia Tools and Applications*, vol. 77, no. 1, pp. 1323–1362, 2018.
- [23] V. Vitiello, S.-L. Lee, T. P. Cundy, and G.-Z. Yang, “Emerging robotic platforms for minimally invasive surgery,” *IEEE reviews in biomedical engineering*, vol. 6, pp. 111–126, 2013.
- [24] G. S. Guthart and J. K. Salisbury, “The intuitive/sup tm/telesurgery system: overview and application,” in *Robotics and Automation, 2000. Proceedings. ICRA’00. IEEE International Conference on*, vol. 1. IEEE, 2000, pp. 618–621.
- [25] A. Casals, J. Amat, and E. Laporte, “Automatic guidance of an assistant robot in laparoscopic surgery,” in *Robotics and Automation, 1996. Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 895–900.
- [26] S. Voros, G.-P. Haber, J.-F. Menudet, J.-A. Long, and P. Cinquin, “Viky robotic scope holder: Initial clinical experience and preliminary results using instrument tracking,” *IEEE/ASME transactions on mechatronics*, vol. 15, no. 6, pp. 879–886, 2010.
- [27] S. Voros, J.-A. Long, and P. Cinquin, “Automatic detection of instruments in laparoscopic images: A first step towards high-level command of robotic endoscopic holders,” *The International Journal of Robotics Research*, vol. 26, no. 11-12, pp. 1173–1190, 2007.
- [28] X. Zhang and S. Payandeh, “Application of visual tracking for robot-assisted laparoscopic surgery,” *Journal of Robotic systems*, vol. 19, no. 7, pp. 315–328, 2002.
- [29] P. Marayong, M. Li, A. M. Okamura, and G. D. Hager, “Spatial motion constraints: Theory and demonstrations for robot guidance using virtual fixtures,” in *Robotics and Automation, 2003. Proceedings. ICRA’03. IEEE International Conference on*, vol. 2. IEEE, 2003, pp. 1954–1959.
- [30] S. Park, R. D. Howe, and D. F. Torchiana, “Virtual fixtures for robotic cardiac surgery,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2001, pp. 1419–1420.

- 
- [31] A. Krupa, J. Gangloff, C. Doignon, M. F. De Mathelin, G. Morel, J. Leroy, L. Soler, and J. Marescaux, “Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing,” *IEEE transactions on robotics and automation*, vol. 19, no. 5, pp. 842–853, 2003.
- [32] N. Padoy and G. D. Hager, “Human-machine collaborative surgery using learned models,” in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5285–5292.
- [33] C. Staub, T. Osa, A. Knoll, and R. Bauernschmitt, “Automation of tissue piercing using circular needles and vision guidance for computer aided laparoscopic surgery,” in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4585–4590.
- [34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [35] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [36] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “Tensorflow: a system for large-scale machine learning.” in *OSDI*, vol. 16, 2016, pp. 265–283.
- [37] hellochick, “hellochick/pspnet-tensorflow,” [Online; accessed 20-July-2018]. [Online]. Available: <https://github.com/hellochick/PSPNet-tensorflow>
- [38] C. for Open Medical Image Computing, “Endovissub2018-roboticscenese segmentation,” [Online; accessed 20-July-2018]. [Online]. Available: <https://endovissub2018-roboticscenese segmentation.grand-challenge.org/home/>