# Cataract surgical videos analysis based on cross domain data

11913001 Mingyang OU 11911633 Xiaoxuan Wang 11910713 Chenlang YI

November 2021

## 1 Introduction

Surgical picture semantic segmentation is attracting increasing attention from the medical image processing community. The goal generally is not to precisely locate tools in images, but rather to indicate which tools are being used by the surgeon at each instant. The main motivation for annotating tool usage is to design efficient solutions for surgical workflow analysis. Analyzing the surgical workflow has potential applications in report generation, surgical training, and even real-time decision support.

We propose an innovative framework on cross-domain data in the clinical application of the cataract surgical tool segmentation. We first go through a frequency space domain randomization methods that transforms cataract surgery images into frequency space and performs domain generalization by identifying and randomizing domain-variant frequency components (DVFs) while keeping domain invariant frequency components (DIFs) unchanged. Based on the DIFs and DVFs, we apply domain Randomization. After that with multi-view methods, shared features of both domains are preserved and therefore improve the performance of semantic segmentation. Then we apply a category-level adversarial network to enforce local semantic consistency during the trend of global alignment.

## 2 Background

Nowadays, cataract surgery is the most common ophthalmic surgery. According to statistics, 19 million cataract surgeries are performed annually[8]. At the same time, there is a need for the optimization of cataract surgery procedures. Hence, cataract surgical tool segmentation has attracted more and more attention in medical image processing for it helps to optimize the surgical workflow. Meanwhile, deep convolutional neural networks trained on large numbers of labeled images provide powerful image representations that can be used for

cataract surgical tool segmentation.

In addition, semantic segmentation is a key problem for many computer vision tasks especially in the field of medical images including surgical tools. While approaches based on convolutional neural networks constantly break new records on different benchmarks, generalizing well to diverse testing environments remains a major challenge. At the same time, deep learning-based methods have achieved great successes in the field of semantic segmentation while large-scale and densely-annotated training data which are required by the deep are deep convolutional networks training are usually extremely expensive and quite time-consuming to collect and annotate. Recent work has shown that generative adversarial networks combined with cycle-consistency constraints are surprisingly effective at mapping images between domains, even without the use of aligned image pairs.

However, the clinical application of semantic segmentation in cataract surgical tool segmentation is insufficient. There is much space for the improvement of the segmentation models. Additionally, recognition models trained along with these representations on one large dataset do not generalize well on other datasets because of the phenomenon of dataset bias and domain shift. Besides, the generalization of the segmentation models can hardly meet the clinical standards for domain shift occurs in the application of the semantic segmentation model on different domains. Therefore, Addressing domain shift by using domain adaptation plays important role in the clinical application of the cataract surgical tool segmentation.

# 3 Related Work

In this section, we briefly review the main ideas from prior work that are relevant to our project.

## 3.1 Semantic Segmentation on cataract surgeries

The continuous evolution of DNN architectures brings great benefits to semantic segmentation. They are usually trained on datasets with dense pixel-level annotations. However, manual annotation is not extensible. Meanwhile, it is quite challenging to capture invariant features in those data. Numerous methods have been developed to improve models by utilizing context information or enlarging receptive fields. CNN(eg. AlexNet, VGG, or ResNet ) are largely used in the field of semantic segmentation. They benefit the semantic Segmentation on cataract surgeries and real-time recognition of surgical tasks[5].

In the presence of surgical scene variation, surgical tool segmentation necessitates precise tool delineation. Occlusion, shadows, reflections, and blurriness are common in images. These aberrations decrease the quality of anticipated

masks and impact the segmentation process. As a result, the process of segmenting surgical tools is deemed difficult.

Tonet et al. suggested one of the first attempts at tracking by altering the visual look of the equipment to make the work easier. This approach, however, had a detrimental influence on tool sterilization. Changing the tools' outside look also necessitates a unique setup. Applicability to pre-existing surgical settings or even recordings is limited by special setups[7]. As a result, semantic segmentation is a reliable and accurate solution to the segmentation problem. It is used to apply a category or class label to each pixel in an image for pixel-level labeling. This can help with proper identification of the surgical tool's various elements as well as delineation of the instrument from the surrounding tissue[9].

However, domain shift problems often occur in semantic segmentation on cataract surgeries when they are applied in different domains. Hence, generalization of the segmentation models can hardly meet the clinical standards for domain shift occurs in the application of the semantic segmentation model on different domains and there isn't a good candidate method to address this problem.

### 3.2 Domain Adaptation

Domain adaptation focuses on transferring knowledge from one or multiple source domains to a target domain with a different data distribution, which possesses a long history. Recently, deep neural networks are largely used in many methods to solve the problem of domain adaptation and can be classified into three categories. The first one includes those that try to find a mapping between source and target dis-tributions[6]. The second one seeks to find a shared latent space for source and target distribu-tions[8]. The third one regularizes a classifier trained on a source distribution to work well on a target distribution[1].

The CoGAN technique, which was recently published, used GANs to solve the domain transfer problem by training two GANs to create the source and target pictures, respectively[2]. By linking the high-level layer parameters of the two GANs, the technique produces a domain invariant feature space and proves that the identical noise input may generate a matching pair of pictures from the two distributions. Domain adaptation was accomplished by using the discriminator output to build a classifier that was then applied to shifts between the MNIST and USPS digit datasets. However, the generators must identify a mapping from the common high-level layer feature space to complete pictures in both domains for this strategy to work. This is particularly useful for digits, which may be challenging in more different contexts.

Since domain adaptation is closely related to finding connection between the source domain and the target domain, feature-level alignment and pixel-level alignment are two common ways. There are also techniques that encourages

alignment on both domains concurrently but only is only limited to small images size and restricted domain shifts. How to utilize both levels at the same time effectively remains to be explored further. Meanwhile, semantic consistency problem is also critical in domain adaptation. J, Hoffman, et al. have proposed a CyCADA methods to align pixel and feature level while preserving semantic consistency at the same time. It applies CycleGan method on semantic segmentation and achieves convincing performance.

The applications of domain adaptation in many fields such as urban streetscape segmentation develop fast and have gained numerous achievements. However, the application of domain adaptation in surgical cataract surgical videos analysis remains a huge space for improvement. We propose an innovative framework on cross- domain data in the clinical application of the cataract surgical tool segmentation.

# 4 Datasets and Preprocessing

There are three datasets used in our experiments. They are all in coco format and are created manually or automatically. Two datasets are based on cataract surgery videos and the other dataset which will be created and annotated by our team members is sampled from a cataract surgery video game.

## 4.1 Datasets of cataract surgery videos

### 4.1.1 Datasets Introduction

Dataset 1 of cataract surgery videos( Iris_pupils and Insegact ) includes annotated 600 different images. Iris_pupils and Insegact are two dataset from the same source. Hence, we combine the annotations of the two data set into our whole data set.

Dataset 2 of cataract surgery videos are CaDIS-dataset for Image Segmentation. It involves 4738 annotated images which are sampled from 50 videos of cataract surgeries performed in Brest University Hospital between January 22, 2015, and September 10, 2015.

Datasets URL: $http://ftp.itec.aau.at/datasets/ovid/InSegCat/$

Different cataract surgical tasks are shown in Figure1: (a) incision; (b) rhexis; (c) hydrodissection; (d) phacoemulsification; (e) epinucleus removal; (f) viscous agent injection; (g) implant setting-up; (h) viscous agent removal; (i) stitching up; (j) miscellaneous.

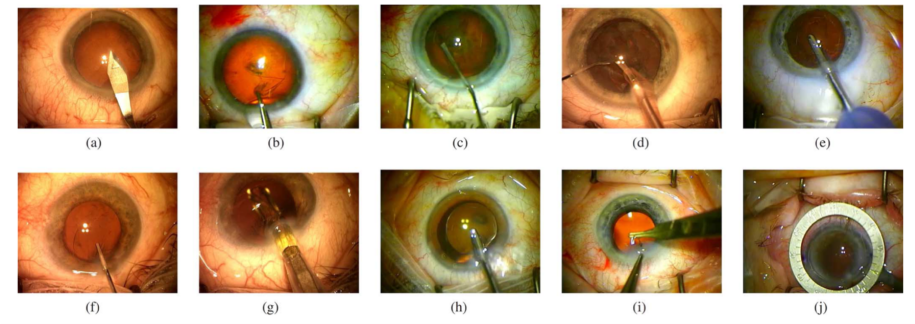


Figure 1: Different cataract surgical tools

| Data sets | Picture numbers | Numbers of categories |
|---|---|---|
| CaDIS training set | 3600 | 6 |
| CaDIS validation set | 400 | 6 |
| CaDIS test set | 582 | 6 |
| Iris_pupils and Insegact training set | 314 | 6 |
| Iris_pupils and Insegact validation set | 120 | 6 |
| Iris_pupils and Insegact test set | 88 | 6 |

Table 1: Cataract surgery videos dataset after data cleaning

#### 4.1.2 Data cleaning

Since the two data sets need to be cleaned in order to apply them in the program, we do the following steps to make it right.

- Remove all the wrong annotations according to the cataract surgical books
- Use Coco annotator to annotate parts that were not annotated in the data sets
- Combine some categories according to the cataract surgical books
- Remove all the pictures that are without any annotations.
- Re-annotate the cataract surgical parts which were annotated wrongly
- Re-arrange the data sets include train, test and valid
- Resize the data in order to fit the networks

The detailed information of the our cataract surgical videos data sets after and before the data cleaning are given in Table 1.

### 4.1.3 Label Alignment

We align the two dataset annotations. The classes in these two datasets are annotated differently on the same cataract surgical tool which may possess different proper nouns in the field of cataract surgery.

Meanwhile, the number of annotated classes of the two datasets is inequal. Since we lack medical knowledge, we preprocessed the dataset to align the annotations according to the book *Color map of cataract microsurgery*. The annotations after aligning are shown in Figure 2

| CaDIS | Insegact/Iris pupils | Cataract Game | Label | Color |
|---|---|---|---|---|
| Skin | Skin | Eyes skin | L1 | 9B9B9B 155 155 155 |
| Iris | iris | Iris part | L2 | DDA0DD 221 160 221 |
| Conera | - | Conera part | L3 | 141414 20 20 20 |
| Pupils | pupil | Pupil | L4 | 9437FF 148 55 255 |
| Eye retractor | i20:eye retractor | Eye retractor | L5 | 4F86F7 79 134 247 |
| Instrument (All sugrical tools) | Instruments (i 1 - i 12) | Instruments | L6 | FFA500 255 165 0 |

Figure 2: Annotations after label aligning

## 4.2 Dataset of the cataract surgery video game

We are working on creating a cataract surgeries dataset based on a cataract surgery video game. It is a flash game and there are two main challenges for the work :

- There is no API available for the game;
- Extracting the required surgical flow images from the game is difficult.

We decompile the cataract surgery game to extract the data we need. Then we divide surgical scenes and align the game scenes to real-time cataract surgery scenes. By modifying the image location of surgical instruments and some methods like Gaussian blur, salt noise and brightness adjustment we generate more data to mimic a real surgical scenario. After that, we do the annotations and align all the categories to the two datasets of cataract surgery videos. The final dataset information is given in Table 2.

| Data sets | Picture numbers | Numbers of categories |
|---|---|---|
| Cataract surgical game training set | 1350 | 6 |
| Cataract surgical game validation set | 200 | 6 |
| Cataract surgical game test set | 150 | 6 |

Table 2: Cataract surgery game dataset

# 5 Methodology

Our model contains three parts : Spectrum analysis, Randomization and multi-view adversaries model. The workflow of the method is shown in Figure 3 [3]

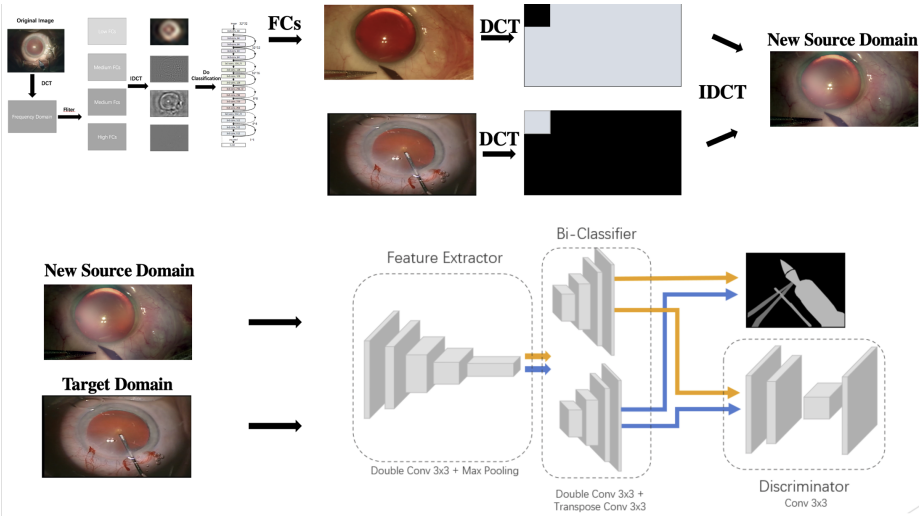


Figure 3: Our proposal methods

## 5.1 Spectrum Analysis

Inspired by the idea of *JPEG* that converts spatial images into multiple frequency components (FCs), Discrete Cosine Transform (DCT) is used to convert source domain images to the frequency space. In this part, spectrum analysis is used to identify domain invariant frequency components (DIFs) and domain variant frequency components (DVFs).

Source-domain image after converted into frequency space are decomposed into 128 FCs by a band-pass filter. Then,we identify DIFs and DVFs in FCs by a set of control experiments. We group the 128FCs into 4 parts and for each source image, we first filter out FCs with indexes between certain lower or

upper thresholds with a band- reject filter and then train models with remaining FCS. Then apply the trained model to the target images to examine the domain invariance and generalization of the filtered source domain FCs.

### 5.2 Randomization

After we get the DIFS and DVFS in source-domain images, we then use DCT to convert source-domain images and target-domain images to the frequency space. After that, we keep the DIFS in source-domain images while replacing the DVFS in source-domian images with the corresponding frequency band of the target-domain images. Finally, we use Inverse DCT to convert the source-domain images back to the picture space. So now we get some source-domain images with target style.

### 5.3 Apply multi-view adversaries methods

The decisive step would be proper alignment between the source domain and target domain, where a multi- view method tends to be applied to high-quality segmentation results. As the ultimate objective of this research is to enhance surgery performance via deploying a model pre-trained with synthesis data, training with artificial labeled gaming data and unlabeled surgery data would be one of the significant challenges. Under these circumstances, multi-view methods are one of the prevalent ways to utilize unlabeled data.

The crucial point of such methods is disagreements between different classifiers. In application cases, attributes are not isolated. Instead, attributes with similar characteristics or represent corresponding aspects of an instance are themselves a set(attribute set). Each attribute set can be treated as one view on the dataset and different attribute sets focus on distinguished parts. A multi-view method forces classifiers to focus on dissimilar parts, which leads to a more comprehensive understanding of domain invariants.

Inspired by paper[4], it is noticed that multi-view methods would be beneficial to adapting game synthesis data to cataract surgery data. With this approach, shared features of both domains are preserved and therefore improve the performance of semantic segmentation.

## 6 Experiments

### 6.1 Platform

The platform information that we used to do all the experiments is shown in Table 3 and Table 4.

| | |
|---|---|
| CPU | Intel Xeon Gold 5222 @ 3.80GHz, 4 cores, x2 |
| GPU | NVIDIA TITAN RTX 24GB, x2 |
| Memory | 8*32 GB, DDR4 2133Mhz |
| Hard Disk | 240GB SSD + 1.92TB NVMe |

Table 3: Hardware Information

| | |
|---|---|
| OS | Ubuntu 18.04.5 LTS |
| Compiler | GNU Compiler v7.5.0; CUDA Compiler v10.0 |
| MPI | Open MPI 4.0.4 |
| Python version | Python 3.6.9 |

Table 4: software Information

## 6.2 Spectrum Analysis

### 6.2.1 Frequency space decomposing

We decompose frequency space of the source-domain image into 270 FCs using different band-pass filters. Then we group the 270 FCs into 6 parts for each source image shown in Figure 4

- $[0, 1)$ FCs : low frequency components
- $[1, 3)$ FCs : low frequency components
- $[3, 8)$ FCs: low frequency components
- $[8, 16)$ FCs: medium frequency components
- $[16, 32)$ FCs: medium frequency components
- $[32, 64)$ FCs: high frequency components
- $[64, 128)$ FCs: high frequency components
- $[128, 270)$ FCs: high frequency components

### 6.2.2 Classification task for identifying DIFs

We analyze and identify domain variant and invariant frequency components (FCs) by training models with certain FCs on source domain (synthetic), and testing with both source domain and target domain images in the classification task.

Because of the unbalanced distribution of the annotations, we chose Secondary Knife to do classification and the results are given in Table **??** . From the table we can tentatively figure out that $[8, 64)$ are DIFs to do the following experiments.

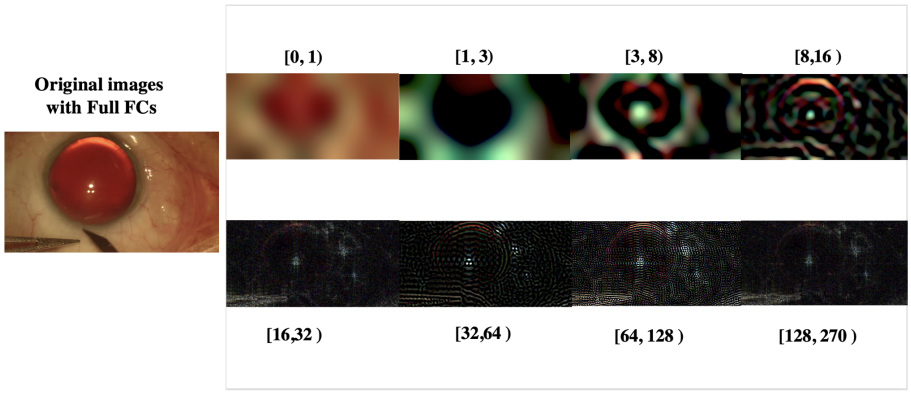


Figure 4: Different Frequency Components

| Rejected band | Source accuracy(%) | Target accuracy(%) |
|---|---|---|
| Null | 96.2 | 61.7 |
| $[0, 1)$ | 96.4 | 62.9 |
| $[1, 3)$ | 96.6 | 62.1 |
| $[3, 8)$ | 94.1 | 61.2 |
| $[8, 16)$ | 94.1 | 58.7 |
| $[16, 32)$ | 94.5 | 59.2 |
| $[32, 64)$ | 95.8 | 59.3 |
| $[64, 128)$ | 97.3 | 62.7 |
| $[128, 270)$ | 97.4 | 62.4 |

Table 5: Band-reject Spectrum analysis

#### 6.2.3 Frequency Composed

We combine the DIFs from Source domain and DVFs from the target domain. A demo of the randomisation results is shown in Figure 5. The test results based on the classification is given in Table 6.We chose $[3, 128)$ from the source domain based on the final results.

### 6.3 Randomization

#### 6.3.1 Preliminary results

We then input the image(source domain) which are processed by frequency replacement into the segmentation network based on ResNet101. We set batch

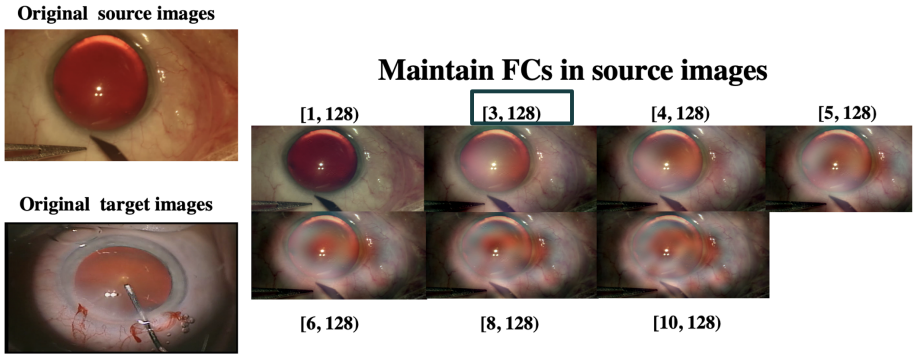


Figure 5: Different randomization results

| Source FCs | Target FCs | Accuracy % |
|---|---|---|
| $[1, 128)$ | $[0, 1) and [128, 270)$ | 72.3 |
| $[3, 128)$ | $[0, 3) and [128, 270)$ | 73.6 |
| $[4, 128)$ | $[0, 4) and [128, 270)$ | 71.1 |
| $[5, 128)$ | $[0, 5) and [128, 270)$ | 67.8 |
| $[6, 128)$ | $[0, 6) and [128, 270)$ | 60.3 |
| $[8, 128)$ | $[0, 8) and [128, 270)$ | 58.4 |
| $[10, 128)$ | $[0, 10) and [128, 270)$ | 57.2 |

Table 6: Band-reject Spectrum analysis based GTA5 as source domain and Cityscapes as target domain

size = 1 and trained 20000 times. We have selected the pictures with relatively good segmentation effect for display(Figure 6).

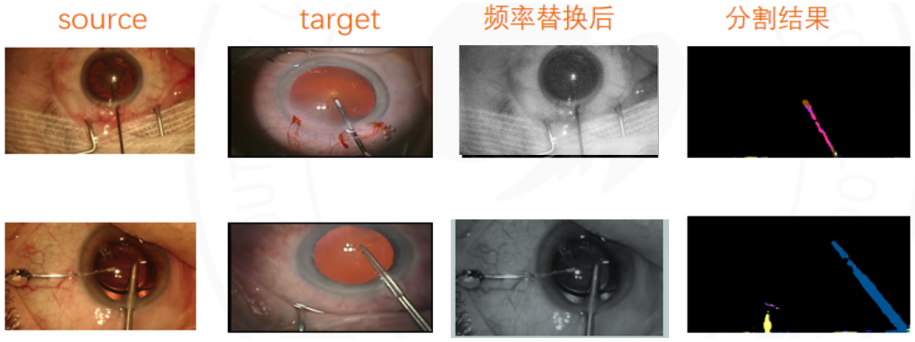


Figure 6: Randomization and segmentation result

### 6.3.2 Preliminary Analysis

Since the proportion of surgical instruments in the whole image is too small and there are many bright spots and shadows, the segmentation of the surgical instruments is difficult. Later we will increase the proportion of surgical instruments in the image by changing the size of the image. Also, we may use image normalization to Remove shadows and bright spots.

### 6.3.3 Further results

We stylized each source domain image with five target domain images, resulting in a new dataset that is five times the size of the original dataset.Then We compared the segmentation results of Unet network and Resnet network, and it can be seen that Unet's segmentation results on medical images are better than Unet's. Therefore, we finally chose Unet as the basic framework of our neural network.

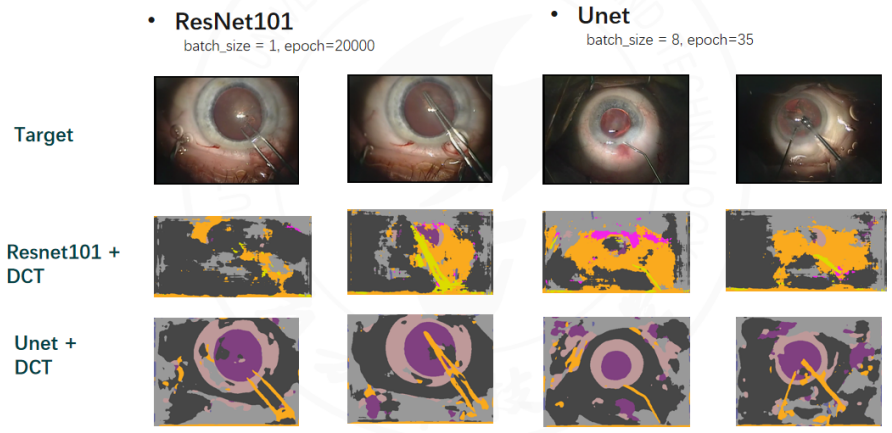


Figure 7: segmentation results of Unet and Resnet

Resnet101 +DCT

| Category | Pupil | Iris | Surgical Tools |
|---|---|---|---|
| Iou | 0.119 | 0.112 | 0.096 |

Unet +DCT

| Category | Pupil | Iris | Surgical Tools |
|---|---|---|---|
| Iou | 0.490 | 0.411 | 0.210 |

Figure 8: MIOU of Unet and Resnet

### 6.3.4 Further results

We obtain the segmentation results after domain randomization based on Unet network. We find that the segmentation results on the target domain by using domain randomization method are better than those by using Unet network alone, which proves that our domain randomization method is effective.

## 6.4 Multi-View Domain Adaptation Model: U2D

### 6.4.1 Model of Multi-View Method

In previous report, we have suggested applying multi-view method on cataract surgery semantic segmentation by demanding two distinctive classifiers to have different views on the same dataset but have similar segmentation prediction. We now propose our, though not mature enough, model, which utilizes U-Net as backbone and adopts GAN methods for domain adaptation.

U-Net has been widely used in medical semantic segmentation due to its outstanding performance. A U-Net model, which has similar to FCN model, consists of an encoder and a decoder. The encoder is responsible for extracting high-level domain features from lower-level, projecting a raw input image into high dimensional feature space. Following is the decoder which is in charge of making prediction based on unearthed features. While both in encoder and decoder, an elementary unit is a double-convolution layer. The encoder is the cross combination of max-pooling layers and double-convolution layers, while decoder layer is the combination of transpose-convolution layers and double-convolution layers. Another prominent enhancement of U-Net is the skip-connection method. Since spatial information attends to be lost in the process of max-pooling, skip-connection attaches outputs of double-convolution layers of decoder on the same level, providing more precise feature information in low-level.

Due to the challenges mentioned, U2D domain adaptation segmentation model is proposed. The model is consist of two major part 1) U2D segmentation based on the structure and scheme of traditional Unet 2) A classic pix2pix discriminator for adversarial training. In U2D segmentation model, decoder is the same as Unet model while in decoding part, two decoders are with identical architecture but distinguish parameters. On the other hand, discriminator is designed under the base frame of pix2pix discriminator. What should be paid attention to is that the kernel size is dedicated selected to be 1x1 in order to give evaluation on segmentation quality.

When input images are from source domain, such input are first passed through the feature extractor and the two differentiate classifiers, following by discriminator giving judgement on it. When input images are derived from target domain, analogous process is operated except that there is no segmentation quality evaluation since target input has no ground truth label. Furthermore,

outputs of discriminator would be multiplied by a weighed map for the purpose of encouraging model to give more consideration on small categories.

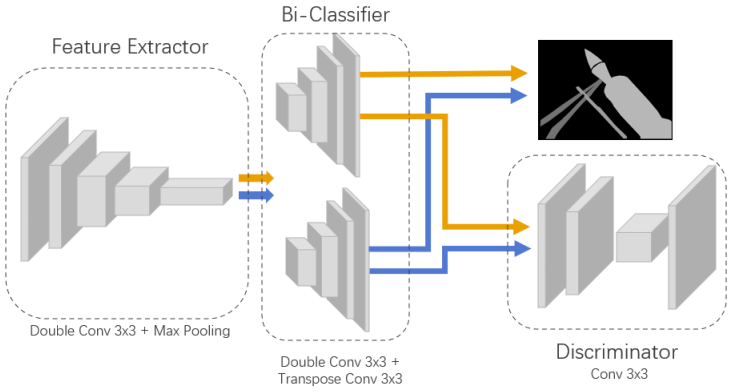


Figure 9: U2D Model

### 6.4.2 Metric Measurements

Two major metric methods are adopted for model evaluation. In semantic segmentation, there are four categories of pixel-level prediction, respectfully true positive($TP$), true negative($TN$), false positive($FP$) and false negative($FN$):

| | True | False |
|---|---|---|
| Positive | True Positive(TP) | False Positive(FP) |
| Negative | True Negative(TN) | False Negative(FN) |

Table 7: Four Categories of Prediction Evaluation

Metric $mIoU$ is broadly used for evaluation of semantic segmentation model selection, as it calculates intersection of the ground truth and prediction over union of them for each class. Finally, a mean value is token from the sum of IoU of all classes. The formula is presented as follows($p_{ij}$ indicates number of pixels of class i being predicted as class j):

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}}$$

### 6.4.3 Experiment Schedule and Results

U2D segmentation model is first trained on the source domain and its performance is compared with traditional Unet model. Experiments prove the

feasibility of U2D model. Besides, due to the multi-view property, U2D tends to grasp more latent features than Unet, as IoU results of U2D model shows enhancement on majority of categories. After validation experiment on source

IoU: Unet

| Pupils | Iris | Sugical Tape | Hand | Eye Retractors |
|---|---|---|---|---|
| 0.941 | 0.879 | 0.907 | 0.748 | 0.848 |

| Skin | Cornea | Other Surgical Tools |
|---|---|---|
| 0.855 | 0.956 | 0.801 |

Table 8: Intersection Over Union: Unet

IoU: U2D

| Pupils | Iris | Sugical Tape | Hand | Eye Retractors |
|---|---|---|---|---|
| 0.943 | 0.880 | 0.907 | 0.803 | 0.854 |

| Skin | Cornea | Other Surgical Tools |
|---|---|---|
| 0.860 | 0.956 | 0.803 |

Table 9: Intersection Over Union: U2D

domain, adversarial training is blended into the model training process. The experiment demonstrate convincing adaptation results. Advantage of U2D can be proved in visual results. From bottom to top, each row represent Unet results on target domain, U2D results on target domain, U2D with domain adaptation results on target domain and the input images correspondingly. The results indicates that U2D model retain more domain invariant information. As a result, its perform better segmentation than Unet. Adding adversarial training, segmentation achieves excellent outcomes. Enhancement on metrics IoU matches with the visual results, where U2D with adversarial training could increase by 106.30% on pupils, 167.84% on Iris and 46.74% on other surgical tools respectively.

| IoU | | | | Enhancement | | | |
|---|---|---|---|---|---|---|---|
| Model | Pupil | Iris | Surgical Tools | Model | Unet | U2D | U2D+DA |
| Unet | 0.349 | 0.227 | 0.184 | Pupil | 0 | 24.93% | 106.30% |
| U2D | 0.436 | 0.237 | 0.208 | Iris | 0 | 4.41% | 167.84% |
| U2D(DA) | 0.720 | 0.608 | 0.270 | Surgical Tools | 0 | 13.04% | 46.74% |

Table 10: Domain Adaptation IoU & Enhancement

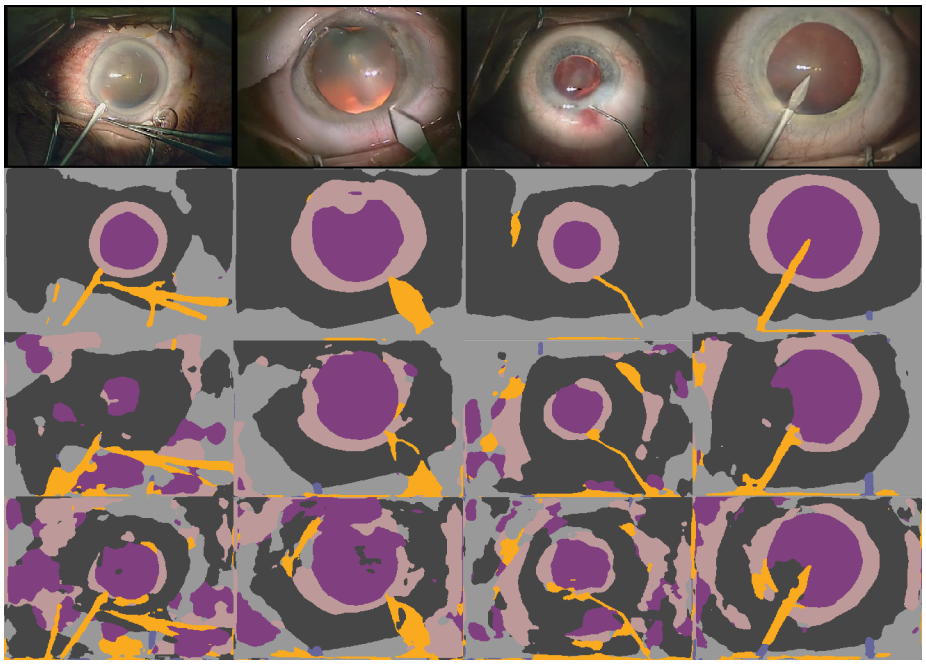

Figure 10: DA Visual Results

# 7 Arrangements

Our arrangements according to the program are given in Figure 11.

| Date | 区明阳 | 王晓轩 | 易辰朗 | Goals |
|---|---|---|---|---|
| 1st Sep. to 25th Nov. | • De-compilation of surgery game and synthesis dataset generation<br>• Cataract surgery game label generation<br>• Real cataract dataset label generation and transformation<br>• Domain adaptation model construction<br>• Baseline model verification | • Data cleaning and label alignment of cataract surgery videos datasets<br>• Decompile the cataract surgery video game and annotate the data.<br>• Research method of Spectrum Analysis<br>• Test the method of spectrum analysis and do pre-experiments.<br>• Get the accuracy results | • Doing research on Fourier Domain Adaptation(FDA) for Semantic Segmentation and other related literatures<br>• De-compiation of cataract surgery game and process the pictures by generating labels and data augmented<br>• A preliminary study on the code structure of FDA method | • Background knowledge study<br>• Background research<br>• Data collection<br>• Methodology discussion<br>• Data alignment<br>• Model Construction<br>• Baseline experiment and verification |
| 25thNov to 30th Dec | • Model integration with other two parts<br>• Get experimental test on proposed model<br>• Model tuning<br>• Result analysis | • Optimize the spectrum analysis<br>• Combine the spectrum analysis with the other two parts<br>• Do the experiments on our innovative framework<br>• Collect and rejudge the results | • Combine method with the other two methods<br>• Make experiments using our proposed model<br>• Revise our model<br>• Analyze the results | • Model experiment and improvement<br>• Experiment result analysis<br>• Model fine tuning |

Figure 11: Arrangements

# References


[1] Alessandro Bergamo and Lorenzo Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. *Advances in neural information processing systems*, 23:181–189, 2010.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[3] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE international conference on computer vision*, pages 3362–3371, 2017.

[4] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[5] Gwénolé Quellec, Mathieu Lamard, Béatrice Cochener, and Guy Cazuguel. Real-time segmentation and recognition of surgical tasks in cataract surgery videos. *IEEE transactions on medical imaging*, 33(12):2352–2360, 2014.

[6] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[7] Oliver Tonet, TU Ramesh, Giuseppe Megali, and Paolo Dario. Tracking endoscopic instruments without localizer: image analysis-based approach. *Studies in health technology and informatics*, 119:544–549, 2006.

[8] Sameer Trikha, AMJ Turnbull, RJ Morris, DF Anderson, and Parwez Hossain. The journey to femtosecond laser-assisted cataract surgery: new beginnings or a false dawn? *Eye*, 27(4):461–473, 2013.

[9] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2):119–135, 2017.