2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871673

MVD-Net: Semantic Segmentation of Cataract Surgery Using Multi-View Learning

Mingyang Ou[1], Heng Li[*1], Haofeng Liu[1], Xiaoxuan Wang[1],
Chenlang Yi[1], Luoying Hao[1], Yan Hu[*1], Jiang Liu[1,2]

***Abstract*— Semantic segmentation of surgery scenarios is a fundamental task for computer-aided surgery systems. Precise segmentation of surgical instruments and anatomies contributes to capturing accurate spatial information for tracking. However, uneven reflection and class imbalance lead the segmentation in cataract surgery to a challenging task. To desirably conduct segmentation, a network with multi-view decoders (MVD-Net) is proposed to present a generalizable segmentation for cataract surgery. Two discrepant decoders are implemented to achieve multi-view learning with the backbone of U-Net. The experiment is carried out on the Cataract Dataset for Image Segmentation (CaDIS). The ablation study verifies the effectiveness of the proposed modules in MVD-Net, and superior performance is provided by MVD-Net in the comparison with the state-of-the-art methods. The source code will be publicly released.**

I. INTRODUCTION

Cataracts are the leading cause of moderate to severe vision impairment in the world and are mainly responsible for blindness [1]. With the aging of the population, the treatment of cataracts attracts more and more attention from medical institutes and organizations [2], [3]. Considering cataract surgery is the only effective treatment approach for cataracts [4], promoting the development of cataract surgery is of significance. In recent years, Surgery Data Science (SDS) as an emerging scientific field, which combines capturing, analysis, and modeling of surgery data, presents the potential to boost the quality of clinical surgery and disease diagnosis[5]. SDS provides founded technical support for surgery, mainly in decision making, context-aware assistance and surgical training process [6]. In cataract surgery, in-surgery SDS assistance can effectively prevent inexperienced surgeons' mistaken operations, which lead to hazardous situations. However, cataract surgery-aid systems rely on visual demonstration to take effect, where segmentation and object detection play important roles. Particularly, semantic segmentation with pixel-to-pixel annotation is capable to separate and categorize surgical tools and environments, increasing the overal performance of the surgery.

Due to the advantage in image feature representation, Convolutional Neural Networks (CNNs) have attained striking achievement in computer vision [7] and been introduced to segment surgical scenarios. Fully Convolutional Network (FCN) [8] overcome the loss of spatial information by replacing fully-connected layers with convolutional modules, and has been commonly used to address semantic segmentation in surgical scenarios. Nonetheless, FCN has limited capacity on edge detection for subtle objects, which may cause inferior segmentation in surgical scenarios, since some surgical instruments are tiny on the screen. Using skip-connection, lower-level spatial characteristics are forwarded in U-Net [9] to achieve convincing results on medical image segmentation. Based on U-Net, M-Net [10] is developed, which imports multi-scale inputs to refine segmentation. However, as U-Net and M-Net concentrate on medical images, they have not been fully explored to segment surgical instruments.

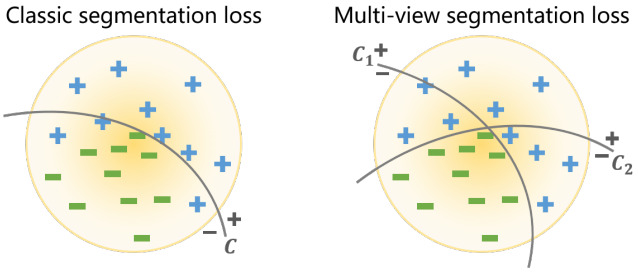

Fig. 1. Illustration of conventional and the proposed segmentation model. Multi-view learning enhances the capacity for feature interpretation.

Furthermore, semantic segmentation in cataract surgery is a troublesome task. According to the public cataract surgery image segmentation dataset, CAIDS [11], the challenges includes: 1) strong lighting in cataract surgery leads to serious uneven reflection, which changes the visual characteristics of surgical instruments; 2) as cataract surgery instruments are designed for micromanipulation, the number of background pixels is far beyond that of foreground pixels, resulting in an image-level class imbalance; 3) various instruments appear at different frequencies in CAIDS, leading to a dataset-level class imbalance. A Dilation Feature-Pyramid (DFP) module was designed in [12] to extract multi-scale and multi-level features in the decoder. However, DFP was only applied to identify the background and foreground of surgical instruments. Repeat Factor Sampling(RF) and Adaptive Sampling(Adapt.) [13] give bias on rarer classes. However, problems like strong lighting in surgery are yet to be settled. And in [14], only two instruments were segmented from a surgical image at a time.

To remit the segmentation challenges in cataract surgery, the multi-view algorithm is introduced to develop a MVD-

*Corresponding authors.
[1]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China.
[2]Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen 518055, China.

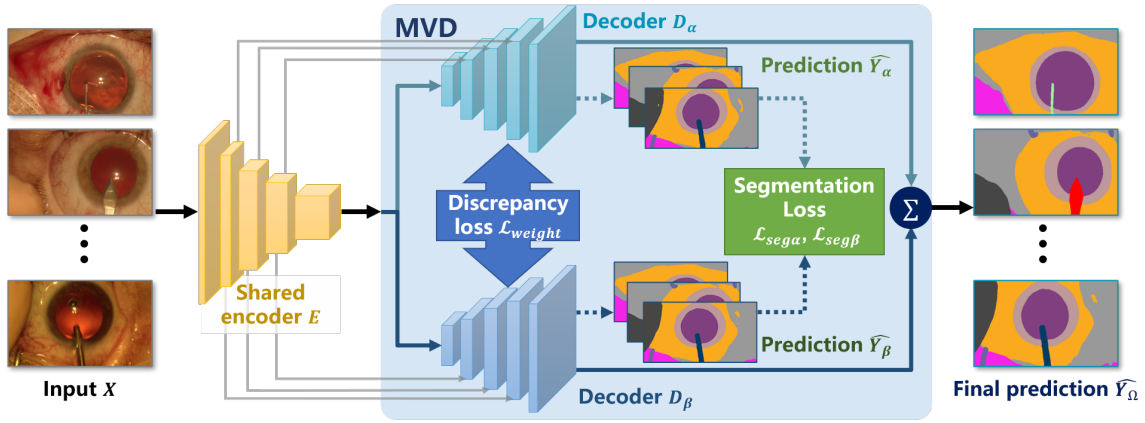

Fig. 2. Overview of the proposed MVD-Net. The encoder E is shared by the two decoders in MVD. To conduct multi-view segmentation, L_{weight} enforces the discrepancy of the decoders, and segmentation loss is respectively calculated in the decoders.

Net to segment instruments in cataract surgery. The multi-view algorithm aims to enforce learners to learn different attribute sets and cooperates with each other to improve prediction quality. For properly segment street scenes, a category-level adversarial network (CLAN) [15] was established by implementing multi-view training on FCN. Inspired by CLAN, the decoder in U-Net is replaced by multi-view decoders (MVD) to boost the instrument segmentation in cataract surgery as shown in Fig. 1. The main contributions of this study are as follows:

- A segmentation network, termed MVD-Net, is designed using multi-view learning to segment the entire semantic categories in cataract surgery.
- By implementing MVD in U-Net, the advantages of multi-view and skip-connection are joined to perform superior segmentation performance.
- In the experiment, MVD-Net not only reasonably segments every semantic category in CAIDS, but also present outperforms the state-of-the-art methods.

II. METHODOLOGY

A segmentation network, named MVD-Net is proposed in this study to remit the challenges from uneven reflection and class imbalance. As demonstrated in Fig. I, the encoder in MVD-Net is shared by two decoders to conduct multi-view learning in the segmentation. To enforce the decoders of MVD to be diverse, a discrepancy loss is calculated. And the segmentation loss is respectively quantified by the outputs of the two decoders.

A. Multi-view learning

In real-world applications, a particular single view is sometimes insufficient to comprehensively understand an item, such that descriptions from different observation views are often collected to interpret items [16]. For instance, color information and texture information are two different kinds of features in an image, which can be regarded as two-view descriptions of the image.

Multi-view learning attempts to integrate descriptions from different views and jointly optimizes the model to improve the generalization performance. Inspired by implementing multi-view learning with different classifiers in [15], the decoder in U-Net is replaced by two discrepant decoders to boost the generalizability of the segmentation model in this study.

B. Network Architecture

As exhibited in Fig. I, the proposed MVD-Net consists of two major components i.e. a shared encoder E and an MVD module. The encoder E embeds input images and forwards the embeddings to the decoders in MVD. Following the architecture of U-Net, skip-connections are bridged between the layers in the shared encoder and both decoders of MVD.

MVD contains two decoders, D_α and D_β. To provide different descriptions, a discrepancy loss, $\mathcal{L}_{weight}$ is captured between the weights of D_α and D_β to enforce the multi-view segmentation. Subsequently, individual segmentation losses, $\mathcal{L}_{seg\alpha}$ and $\mathcal{L}_{seg\beta}$ are respectively computed from the predictions by the outputs of the last layers in D_α and D_β to optimize the network. The final prediction is acquired by combining the outputs of D_α and D_β.

C. Training Objective

The objective of MVD-Net is composed of three losses, including $\mathcal{L}_{weight}$, the discrepancy loss between the two decoders and $\mathcal{L}_{seg\alpha}$, $\mathcal{L}_{seg\beta}$, the individual segmentation losses of the decoders.

Multi-view learning is conducted by MVD-Net to learn a generalizable segmentation model. To implement multi-view learning, the decoders in MVD-Net are driven to discriminately comprehend the latent features from the encoder. Specifically, during the training stage, the discrepancy loss $\mathcal{L}_{weight}$ is employed to enforce the decoders converge diversely. The cosine similarity metric is leveraged to measure the structural distance between the two decoders. Given the weights in the first three layers of D_α and D_β as $\overrightarrow{w_\alpha}$ and $\overrightarrow{w_\beta}$, the discrepancy loss is defined as:

$$\mathcal{L}_{weight}(E, D_\alpha, D_\beta) = \frac{\overrightarrow{w_\alpha} \cdot \overrightarrow{w_\beta}}{||\overrightarrow{w_\alpha}||||\overrightarrow{w_\beta}|| + 1} \cdot \gamma_{dacay}, \quad (1)$$

where $||\cdot||$ is the modulo of a weight vector, and γ_{decay} denotes a decaying coefficient.

For an input image $x \in X$ and ground truth $y \in Y$, the output of D_α are denoted as $D_\alpha(E(x))$. Thus the predictions by them are given by $\widehat{y_\alpha} = \mathrm{softmax}(D_\alpha(E(x)))$, and the segmentation loss of D_α is defined as:

$$\mathcal{L}_{seg\alpha}(E, D_\alpha) = \sum_{i=0}^{W}\sum_{j=0}^{H} -\log\left(\frac{\exp(\widehat{y_\alpha}(i,j)[y(i,j)])}{\sum_{c=0}^{C}\exp(\widehat{y_\alpha}(i,j)[c])}\right), \tag{2}$$

where $\widehat{y_\alpha}(i,j)$, $y(i,j)$ represents the prediction and ground truth of class probability on the pixel at (i,j). c denotes a class in the total C classes, and $[\cdot]$ refers to the probability of a specific class. W and H are the width and height of the input images. The other segmentation loss $\mathcal{L}_{seg\beta}$ is defined in the identical pipeline.

All the loss functions together comprise the total loss function of our network:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{weight} + \lambda_2(\mathcal{L}_{seg\alpha} + \mathcal{L}_{seg\beta}), \tag{3}$$

where λ_1 and λ_2 refer to the coefficients of the losses. The goal of training procedure is to find the weights of model such that the total loss $\mathcal{L}_{total}$ is minimized. And the final prediction by MVD-Net is given by:

$$\widehat{y_\Omega} = \mathrm{softmax}(D_\alpha(E(x)) + D_\beta(E(x))). \tag{4}$$

III. EXPERIMENTS

A. Dataset And Metric

The public cataract surgery image segmentation dataset, CAIDS [11] is used for the experiment. It contains a total of 4,670 cataract surgery images, and in this study 3,550, 543, and 577 images are respectively used as the training, validation, and test data. Thirty-six item classes are annotated in CAIDS, including 29 surgical instrument classes, 4 anatomy classes, and 3 classes of other objects appearing in the scene. Due to the class imbalance in CADIS, frequency weighted intersection over union (FWIoU) is adopted to properly assess the performance, which is given by:

$$FWIoU = \frac{1}{p_N}\cdot\sum_{m=0}^{C}\frac{p_{mm}}{\sum_{n=0}^{C}p_{mn} + \sum_{n=0}^{C}p_{nm} - p_{mm}} \tag{5}$$

where p_N represents total number of pixels, C represents the number of categories and p_{mn} is the number of pixels of class m that are predicted as n by the model. Accordingly, the evaluation metric assigning frequency weights to classes is able to comprehensively assess the segmentation performance on CADIS.

B. Experiment Details

The network construction and experiments in this study are all conducted on the Pytorch platform. Considering the computation resource, the input images are resized to 480 $\times$ 270 from 960 $\times$ 540. The models are all trained for 80 epochs with a batch size of 7. And in order to prevent models from over-fitting, Adam is chosen as the optimization strategy with decay learning. The initial learning rate is $2 \times e^{-4}$ and starts to decay at the $60th$ epoch. Time cost of the training process is approximately 7.3 hours.

C. Ablation Study

To demonstrate the effectiveness of the proposed modules in MVD-Net, an ablation study is presented. The multi-view decoders, discrepancy loss, and individual segmentation loss are respectively installed on the backbone of U-Net. The segmentation performance is summarized in Table I, where pixel accuracy (PA) and FWIoU are calculated to quantitively verify the effectiveness.

TABLE I
ABLATION STUDY OF THE PROPOSED METHOD.

D_β	$\mathcal{L}_{weight}$	$\mathcal{L}_{seg\beta}$	PA (%)	FWIoU (%)
			93.82	88.22
√			94.57	89.86
√	√		94.62	89.93
√	√	√	**94.69**	**90.06**

The segmentation performance is improving steadily with assistance from the proposed modules. The additional decoder D_β imports extra interpretation of the features from the encoder. Subsequently, the discrepancy loss $\mathcal{L}_{weight}$ enforces the training of a generalizable model by multi-view learning. Finally, the individual segmentation loss is used to integrally optimize the encoder and separately optimize the decoders.

D. Comparison with state-of-the-art methods

To demonstrate the performance of MVD-Net, a comparison with state-of-the-art methods is conducted. The segmentation neural networks, i.e. FCN-8, -16, -32 [8], U-Net [9], M-Net [10], and DFP [12], are selected as the benchmark to compare with MVD-Net. The segmentation results are exhibited in Fig. 3, and a quantitative comparison is summarized in Table II.

As shown in Fig. 3, compared to the anatomy classes, cataract surgery instruments are tiny on the screen. Moreover, the uneven reflection leads to variant brightness of the identical items. As a result, fragmented and misclassified segmentation predictions are common in Fig. 3. The state-of-the-art methods confuse the surgical instruments, such as the hydrodissection cannula and phacoemulsifier handpiece. The superior performance of MVD-Net premise outstanding segmentation results.

The quantitative comparison is provided in Table II with independent intersection over union (IoU) of each category and the FWIoU of the entire classes. According to Table II, MVD-Net performs the best in all of the 10 most frequent classes, and outperforms the state-of-the-art methods in the rest of the classes and the overall dataset.

FCN networks present superior results than U-Net, and M-Net enhances the performance of the U-Net backbone such that it outperforms FCN networks. Although DFP improved the binary segmentation of cataract surgery images, it shows mediocrity for segmenting the entire classes. The multi-view

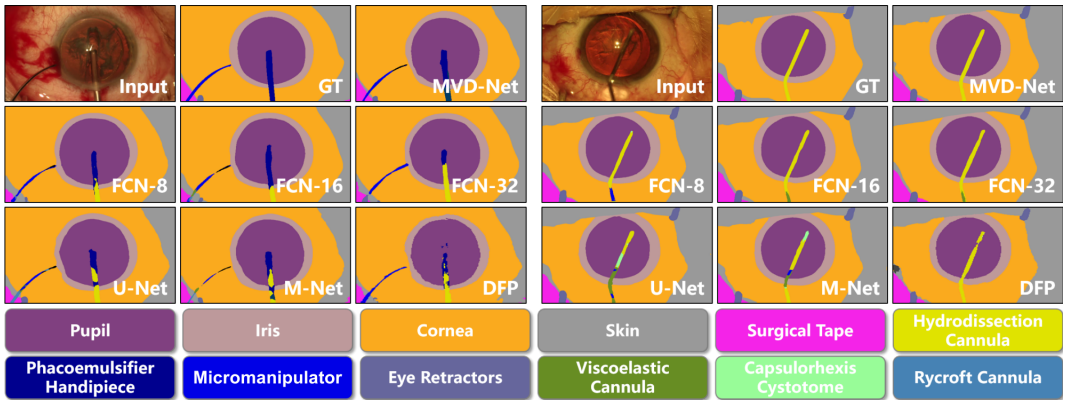

Fig. 3. Visualized results of the segmentation of cataract surgery. The classes presented the colors in the masks are enumerated.

TABLE II

QUANTITIVE COMPARISON WITH STATE-OF-THE-ART METHODS USING THE METRIC OF IOU (%).

Category	FCN-8 [8]	FCN-16	FCN-32	U-Net [9]	M-Net [10]	DFP [12]	MVD-Net (ours)
Pupil	92.90	93.40	93.16	93.69	94.02	93.50	**94.09**
Surgical Tape	82.42	82.50	82.02	84.22	83.05	75.05	**84.32**
Hand	57.19	81.48	83.86	68.69	91.76	46.00	**89.70**
Eye Retractors	79.25	75.77	73.87	83.31	82.41	78.69	**85.08**
Iris	80.76	82.01	81.01	79.29	83.23	79.42	**82.70**
Skin	88.31	87.47	87.29	87.18	88.21	85.03	**88.64**
Cornea	92.63	92.86	92.47	91.63	93.27	91.61	**93.37**
Viscoelastic Cannula	54.21	53.70	53.25	51.08	48.19	52.50	**61.09**
I/A Handpiece	61.30	69.71	66.34	68.40	55.94	56.40	**70.94**
Micromanipulator	48.00	49.01	45.35	47.12	47.75	44.40	**57.93**
Other items	66.72	68.56	65.84	68.05	64.49	59.01	**75.52**
Overall (FWIoU)	88.53	88.93	88.46	88.22	89.59	86.64	**90.06**

learning by MVD-Net allows the performance to improve from 88.42% of U-Net to 90.06%.

IV. CONCLUSION

In surgery scenarios, semantic segmentation is a significant task for computer-aided systems. To precisely segment surgical instruments and anatomies in cataract surgery, a network termed MVD-Net is developed in this study to present a generalizable segmentation. Two decoders are implemented in MVD-Net, and the diversity of decoders is enforced by a discrepancy loss. An ablation study is conducted in the experiment to demonstrate the effectiveness of MVD-Net, and compared with the state-of-the-art methods, MVD-Net achieves superior performance.

ACKNOWLEDGMENT

This work was supported in part by Guangdong Basic and Applied Fundamental Research Fund Committee (2020A1515110286), Guangdong Provincial Department of Education (2020ZDZX3043), Guangdong Provincial Key Laboratory (2020B121201001), and Shenzhen Natural Science Fund (JCYJ20200109140820699 and 20200925174052004).

REFERENCES

[1] Jaimie D Steinmetz et al., "Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to vision 2020: the right to sight: an analysis for the global burden of disease study," *Lancet Glob. Health*, vol. 9, no. 2, pp. e144–e160, 2021.

[2] Tao Li et al., "Applications of deep learning in fundus images: A review," *Medical Image Analysis*, p. 101971, 2021.

[3] Heng Li et al., "An annotation-free restoration network for cataractous fundus images," *TMI*, pp. 1–1, 2022.

[4] Shalu Jain et al., "Effects of cataract surgery and intra-ocular lens implantation on visual function and quality of life in age-related cataract patients: a systematic review protocol," *Systematic reviews*, vol. 8, no. 1, pp. 1–6, 2019.

[5] Heng Li et al., "Ct scan synthesis for promoting computer-aided diagnosis capacity of covid-19," in *ICIC*, Berlin, Heidelberg, 2020, p. 413–422, Springer-Verlag.

[6] Lena Maier-Hein et al., "Surgical data science: enabling next-generation surgery," *arXiv preprint arXiv:1701.06482*, 2017.

[7] Heng Li et al., "Restoration of cataract fundus images via unsupervised domain adaptation," in *ISBI*. IEEE, 2021, pp. 516–520.

[8] Jonathan Long et al., "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[9] Olaf Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[10] Raghav Mehta et al., "M-net: A convolutional neural network for deep brain structure segmentation," in *ISBI*. IEEE, 2017, pp. 437–440.

[11] Maria Grammatikopoulou, Evangello Flouty, Abdolrahim Kadkhodamohammadi, Gwenol'e Quellec, Andre Chow, Jean Nehme, Imanol Luengo, and Danail Stoyanov, "Cadis: Cataract dataset for image segmentation," *arXiv preprint arXiv:1906.11586*, 2019.

[12] Yiming Wang et al., "Surgical instrument segmentation based on multi-scale and multi-level feature network," in *EMBC*. IEEE, 2021, pp. 2672–2675.

[13] Theodoros Pissas et al., "Effective semantic segmentation in cataract surgery: What matters most?," in *MICCAI*, 2021.

[14] Zhen-Liang Ni et al., "Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments," in *ICONIP*. Springer, 2019, pp. 139–149.

[15] Yawei Luo et al., "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *CVPR*, 2019, pp. 2507–2516.

[16] Jing Zhao et al., "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.