

A Nested U-Structure for Instrument Segmentation in Robotic Surgery

Yanjie Xia, Shaochen Wang, and Zhen Kan

Abstract—Robot-assisted surgery has made great progress with the development of medical imaging and robotics technology. Medical scene understanding can greatly improve surgical performance while the semantic segmentation of the robotic instrument is a key enabling technology for robot-assisted surgery. However, how to locate an instrument's position and estimate their pose in complex surgical environments is still a challenging fundamental problem. In this paper, pixel-wise instrument segmentation is investigated. The contributions of the paper are twofold: 1) We proposed a two-level nested U-structure model, which is an encoder-decoder architecture with skip-connections and each layer of the network structure adopts a U-structure instead of a simple superposition of convolutional layers. The model can capture more context information from multiple scales and better fuse the local and global information to achieve high-quality segmentation. 2) Experiments have been conducted to qualitatively and quantitatively show the performance of our approach on three segmentation tasks: the binary segmentation, the parts segmentation, and the type segmentation, respectively. The results show that our method significantly improves the segmentation performance and outperforms state-of-the-art approaches.

I. INTRODUCTION

Robot-assisted systems have revolutionized the minimally invasive surgery to achieve safer, more precise and consistent, and less invasive intervention [1]. For instance, the Da Vinci Xi robot is able to control laparoscopic surgery through remotely operated by surgeons [2]. Since the success of robot-assisted surgery highly relies on the understanding of surgical scene, accurate segmentation of surgical instruments is crucial.

Recent advances of robotics [3] [4] and computer vision technologies promote the intelligent endoscopic vision, which can help surgeons perform precise operation. For instance, the augmented reality (AR) based on endoscopic video can improve surgeon's visual awareness of high-risk targets [5]. The vision-based endoscopic navigation method has been applied in sinus surgery [6]. The work of [7] presents a method for automatically assessing a surgeon's performance by tracking and analyzing tool movements in surgical videos. The 3D dense reconstruction of handheld monocular endoscopic surgery scenes was developed in [8]. However, in the above applications, the endoscope visual perception is inseparable from medical scene understanding,

resulting in the poor performance in extracting necessary visual and regional information for surgical procedures.

Medical scene understanding can significantly improve the surgical performance, since it can expand the surgeon's perception by providing information on internal anatomy and surgical instruments. Such information is usually provided by videos or 2D images consisting of human tissues and surgical instruments that present their position, shape, size and posture intuitively. The location of the instrument's position relative to the patient's anatomy helps surgeons understand surgical scene and operate more accurately. The pose of the instrument can be used to measure the distance to risk structures, evaluate the surgeon's skills, and realize automated surgical operation [9]. Therefore, it is important to extract these valuable information selectively and intelligently, while avoiding unnecessary information that might confuse the surgeons.

To address this challenge, scene segmentation of surgical instruments is a recent research focus, which can separate the instruments from the background tissue and provides important information in surgical procedures for surgeons. Segmentation masks can prevent the covering occlusion apparatus of the rendered tissue and clearly show the position and pose of the surgical instruments in the endoscopic images [10]. Besides, segmentation masks play an important role in instrument tracking systems. Therefore, the semantic segmentation of surgical robotic instruments is highly desired for promoting the cognitive assistance to surgeons. However, due to the complicated medical scene, how to locate an instrument's position and estimate their pose to achieve precise segmentation is an essentially fundamental yet challenging problem [11].

Recently, a variety of vision based methods are developed for the location and tracking of the instruments [12]. Prior methods of instrument-background segmentation utilized color and texture features [13] [14], Haar wavelets [15], and HoG [16]. Later, machine learning algorithms, such as Random Forest [17] and Gaussian Mixture Model [18], were applied to deal with the segmentation problem. However, these models only focus on single binary segmentation problems. More complex segmentation, such as the detection of various parts and types of the instrument, are desired in modern surgery.

To solve this problem, many deep learning based approaches have been developed, showing promising performance in medical areas, especially for tracking, classification

This work was supported in part by the National Natural Science Foundation of China under Grant U2013601 and Grant 62173314 and CAAI-Huawei MindSpore Open Fund.

Y. Xia, S. Wang, and Z. Kan (Corresponding Author) are with the Department of Automation at the University of Science and Technology of China, Hefei, Anhui, China, 230026.

and location problems of robotic surgical instrument. Convolutional neural networks (CNN) have been successfully applied, which can realize pixel-level segmentation of images captured by endoscope camera [19]. However, it requires a large size of training data, which limits their success in practice. U-Net [20] with an encoder-decoder network architecture was designed to address this issue and has achieved good performance on different biomedical segmentation applications. In fact, location information is the basis of semantic segmentation. Accurate location information can lead to precise segmentation performance. Deep neural networks (DNN) have been used to combine semantic segmentation with landmark locations [21], which learn better feature representation of the existing input by the training mechanism of layer by layer via data pre-training. In [22], the recurrent neural network was embedded with convolutional neural network to establish dependencies among multiple tags. To further improve segmentation accuracy, [23] fused the information of kinematic pose and convolutional neural networks prediction. Besides, [24] [25] [11] have provided solutions for three sub-problems of instrument segmentation, i.e., binary segmentation, partial segmentation, and type segmentation. While these deep learning-based methods have achieved impressive results, it still leaves room for improvement. Moreover, how to improve the accuracy of surgical instrument segmentation efficiently and make it suitable for multiple segmentation tasks is still a challenge.

In this paper, to facilitate intelligent surgery, we develop a novel nested U-structure framework for surgical instrument semantic segmentation. The goal is to better understand the medical scene and extract the semantic information to promote minimally invasive surgery. The main contributions of this work can be summarized as follows: 1) we take multi-scale feature extraction and multi-level deep feature integration into consideration and proposed a two-level nested U-structure, which fuses the local and global features to realize a more precise segmentation for surgical instrument segmentation tasks. 2) Dilated convolutions were used in our network to maintain high resolution feature maps while increasing the reception field of convolution kernel. 3) Experiments are conducted on the MICCAI EndoVis Challenge 2017 dataset. The results show that our model can greatly improve the performance of segmentation and outperforms other state-of-the-art approaches.

II. METHODS

A. Overview

The understanding of medical scene can improve the surgeon's ability in perception. To have a better scene understanding of the surgical scenario, semantic segmentation is one of the important methods to extract the posture and position information of surgical instruments, which is crucial for the smooth surgical operation. It is especially helpful for medical imaging and robot-assisted surgical system. Given an image captured by the high resolution stereo camera, the goal is to separate the surgical instrument from the background in the image, and segment the parts and types of

surgical instruments semantically. To achieve this objective, we proposed a new deep learning-based solution. The overall architecture of our network is illustrated in Fig. 1, which is a two-level nested U-structure. The images obtained by the laparoscopic system are taken as input to the network, and the output is the semantic segmentation of surgical instruments.

B. Network Architecture

For robot-assisted minimally invasive surgery, when a surgical instrument is moved and operated within the tissue, the robot needs to locate and track the instrument. Due to the complex surgical environment, it is crucial for the model to acquire high-precision segmentation of instruments in the surgical scene. In general, features from multiple deep layers are able to generate better results. Taking into account the memory and the computation budget, we choose a 6-layer deep structure for our network architecture. For semantic segmentation, both local and global contextual information are essential for high precision segmentation. Traditional UNet [20] uses two 3×3 convolutions, rectified linear unit (ReLU), and 2×2 maxpooling operation repeatedly to shrink or expand the feature maps, so as to extract important feature information. However, more detailed feature information is required to achieve accurate segmentation and feature information extraction is limited for simple superposition of convolutional layers. The new modules need to be designed to implement multi-scale feature extraction. Inspired by UNet, the model that we proposed adopts an encoder-decoder architecture which can capture context information by a contraction path and locate accurately by an expansion path. Hence, considering multi-level deep feature integration and multi-scale feature extraction, the network we proposed is a two-level nested U-structure as shown in Fig. 1(a). Each layer of the network structure uses a U-structure instead of a simple superposition of convolutional layers.

It is well known that the encode process is essentially a process of feature extraction while reducing the spatial size of feature maps. Since modern CNN models have been successfully used and UNet [20] has greatly promoted the development of deep learning in the field of medical imaging, we introduce Resnet into UNet and call it ResUNet. We use ResUNet as the main encoder backbone of the network architecture. The starting unit of encoder is a combination of Resnet18 and UNetplusplus [26] which we call it ResUNetpp (Fig. 1(b)). It can be interpreted as UNetplusplus using Resnet18 as the encoder. Subsequently, there are 4 stages of encoder which consists of ResUNet34. Similarly, ResUNet34 can be interpreted as UNet using Resnet34 as the encoder. Each step in the contraction path contains alternating convolution and pooling operations, and the feature maps are gradually downsampled with stride 2 while increasing the number of feature maps at each layer. After 4 times of downsampling, the resolution of the feature map has been greatly reduced. The feature information will be lost if we continue on downsampling. Therefore, we adopt the RSU-4F [27] module at the bottom of the structure to keep the

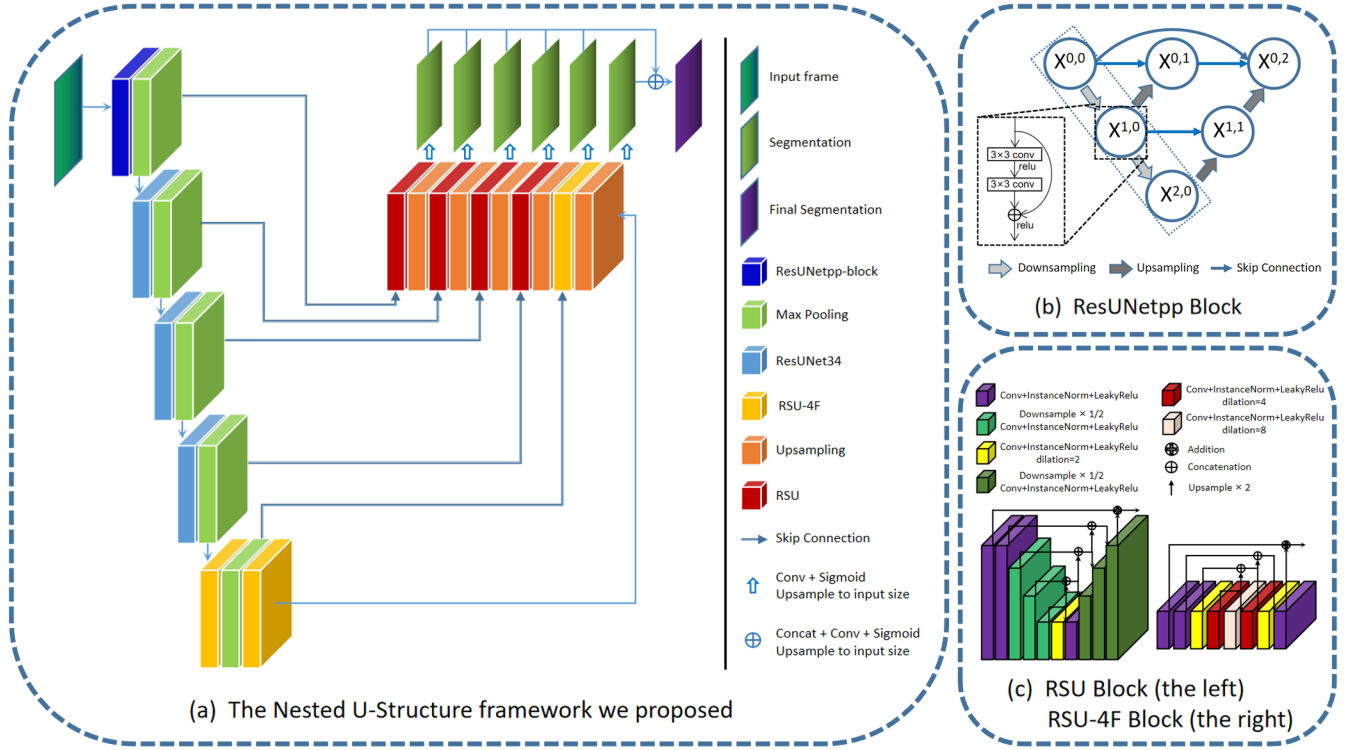


Fig. 1. The illustration of the proposed network for the semantic segmentation of surgical instruments.

resolution consistent between the input and output feature maps. As shown in Fig. 1(c), RSU-4F is a four layer structure similar to UNet which consists of a 3×3 convolutions, a Batch Normalization(BN) and a rectified linear unit (ReLU). To maintain high resolution feature maps while increasing the reception field of convolution kernel, we use dilated convolutions to upsample and downsample. The expanding path increases the resolution of the feature maps by upsampling and the backbone of decoder also follows a similar architecture of UNet. The decoder consists of 4 ResSdual Ublocks(RSU) [27] and the architecture of RSU is shown in Fig. 1(c). It is a variant of a 5-layer UNet which enables the network architecture to extract the features of multiple scales from each residual block. Skip-connections have been applied to combine the feature maps from contracting path and expanding path. It is worth mentioning that we have replaced the classical activation function ReLU with LeakyReLU and substitute InstanceNorm2d for BatchNorm2d in our model. The network takes RGB images as input and generate pixel-level segmentation prediction pictures. We perform three segmentation tasks by setting the number of output channels of the network structure.

C. Lossfunction

Since image segmentation tasks can be regarded as a classification problem of pixels, the overlap rate between the predicted masks and the corresponding ground truth represents the probability that the pixel belongs to each category. The segmentation loss function is based on Jaccard index, which indicates the similarity between two sets. In

this work, we introduce a common loss function, denoted as H . In different segmentation tasks, H represents different loss functions. For binary segmentation task, H adopts the BCEWithLogitsLoss. For multi-class segmentation problem, H represents the Cross-Entropy Loss. We define the generalized segmentation loss function as follows:

$$Loss = H - \log\left(\frac{1}{n} \sum_{i=1}^n \left(\frac{m_i n_i}{m_i + n_i - m_i n_i}\right)\right), \quad (1)$$

where m_i and n_i represent the ground truth and the predicted output for the pixel i , respectively. In order to perform the segmentation task better, it has to minimize the generalized segmentation loss function via maximizing the probability of correctly predicting pixels.

III. EXPERIMENTS

In this section, experiments are conducted to show the performance of the proposed network architecture in three types segmentation tasks. To evaluate the accuracy of this nested U-structure for segmentation qualitatively and quantitatively, we calculate the Intersection Over Union (IOU), also referred to Jaccard Index, as the evaluation criteria. Meanwhile, we use Dice coefficient (Dice) as another evaluation metrics. The segmentation accuracy is proportional to the numerical value of IOU and Dice. We compared our approach with the state-of-the-art methods on EndoVis 2017 dataset, and analyzed the experimental results.

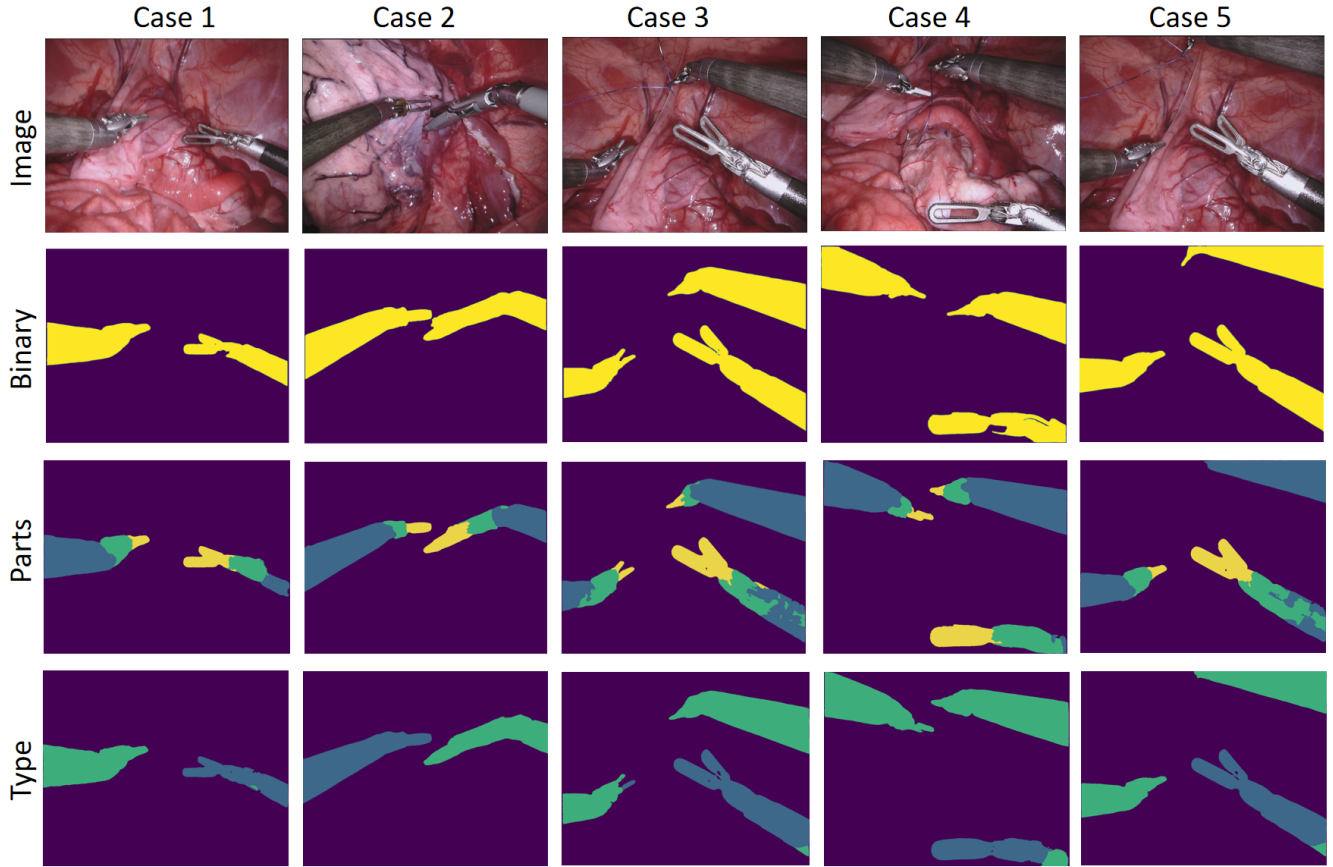


Fig. 2. Examples of visual segmentation results of the proposed model.

TABLE I
COMPARISON OF INSTRUMENT SEGMENTATION RESULTS ON THE THREE TASKS(MEAN \pm STD).

Methods	Binary segmentation		Parts segmentation		Type segmentation	
	IOU(%)	Dice(%)	IOU(%)	Dice(%)	IOU(%)	Dice(%)
U-Net	75.44 \pm 18.18	84.37 \pm 14.58	48.41 \pm 17.59	60.75 \pm 18.21	15.80 \pm 15.06	23.59 \pm 19.87
TernausNet	81.14 \pm 19.11	88.07 \pm 14.63	62.23 \pm 16.48	74.25 \pm 15.55	34.61 \pm 20.53	45.86 \pm 23.20
LinkNet-34	82.36 \pm 18.77	88.87 \pm 14.35	34.55 \pm 20.96	41.26 \pm 23.44	22.47 \pm 35.73	24.71 \pm 37.54
PlainNet	81.86 \pm 15.85	88.96 \pm 12.98	64.73 \pm 17.39	73.53 \pm 16.98	34.57 \pm 21.93	44.64 \pm 25.16
Ours	82.94 \pm 16.82	89.42 \pm 14.01	58.38 \pm 19.06	69.59 \pm 18.66	41.72 \pm 33.44	48.22 \pm 34.46

TABLE II
THE NUMERICAL RESULTS OF OUR METHOD AND COMPARISON WITH OTHER METHODS IN BINARY SEGMENTATION OF ROBOTIC TOOLS.

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6	Dataset 7	Dataset 8	Dataset 9	Dataset 10	mIOU
NCT	0.784	0.788	0.926	0.934	0.701	0.876	0.846	0.881	0.789	0.899	0.843
UB	0.807 [*]	0.806	0.914	0.925	0.740	0.890	0.930	0.904	0.855	0.917	0.875
BIT	0.275	0.282	0.455	0.310	0.220	0.338	0.404	0.366	0.236	0.403	0.326
MIT	0.854	0.794	0.949	0.949	0.862	0.922	0.856	0.937	0.865	0.905	0.888
SIAT	0.625	0.669	0.897	0.907	0.604	0.843	0.832	0.513	0.839	0.899	0.803
UCL	0.631	0.645	0.895	0.883	0.719	0.852	0.710	0.517	0.808	0.869	0.785
TUM	0.760	0.799	0.916	0.915	0.810	0.873	0.844	0.895	0.877	0.909	0.873
Delhi	0.408	0.524	0.743	0.782	0.528	0.292	0.593	0.562	0.626	0.715	0.612
UA	0.413	0.463	0.703	0.751	0.375	0.667	0.362	0.797	0.539	0.689	0.591
UW	0.337	0.289	0.483	0.678	0.219	0.619	0.325	0.506	0.377	0.603	0.461
Ours	0.877	0.814	0.962	0.959	0.849	0.892	0.812	0.956	0.855	0.917	0.889

A. Dataset

The dataset we used in this paper is provided by the Endoscopic Vision Challenge 2017 [28]. The training dataset

consists of 8 robotic surgical videos acquired from da Vinci Xi surgical system in different procedures, and each video is divided into a sequence of 225 images. To avoid data

redundancy, video sampling rate of 2 Hz in the training sequences was provided. The RGB stereo channels from the left and right cameras together form these video sequences. The images taken by the camera on the left provide the hand-labeled ground truth for every robotic instrument, but the right frames was not provided. Therefore, the training images are from the left channel. Different surgical instruments such as rigid shafts, articulated wrists, claspers, drop-in ultrasound probe, and a laparoscopic instrument, have all been labelled by hand. A surgical instrument can be roughly divided into three parts: shaft, wrist, and clasper which are also labelled in the frames.

The testing dataset consists of 8×75 frame sequences sampled immediately after each training sequence and two full 300-frame sequences. These sequences were sampled at the same rate as the training set, resulting in ten test datasets with a total of 1200 images.

B. Training

Before starting the training, the input images are pre-processed. Every RGB images generated from surgical video sequences have a high resolution of 1920×1080 pixels. In order to crop out the black canvas in the frames, images should be reduced to 1280×1024 which are necessary to be cropped at the position of (320, 28). Besides, several simple augmentations (e.g., PadIfNeeded, RandomCrop, Flip Horizontal and Flip vertical) are used for dataset's pre-processing in order to improve the performance of semantic segmentation. The dataset within each image channel is normalized and the mean value of each channel is subtracted to get a zero-average image.

To compare directly and fairly, IOU is a standard performance measure for segmentation problems. The IoU value is calculated as

$$IOU = \frac{1}{n} \sum_{i=1}^n \left(\frac{m_i n_i}{m_i + n - m n_i} \right), \quad (2)$$

where m_i and n_i represent the ground truth value and the predicted output for the pixel i , respectively.

To measure the similarity of the sets, Dice is defined as :

$$Dice = \sum_{i=1}^n \frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i}, \quad (3)$$

where the n represent the number of images in the dataset for evaluation, and TP_i , FP_i , and FN_i denote the numbers of true positives, the false positives and the false negatives for every image, respectively. The value of Dice is between 0 and 1, which indicates the similarity of the semantic segmentation.

In experiments, we set the learning rate as $1e-4$. During the training, the optimizer we used in our proposed model is AdamW. In order to evaluate the generality of the model to the data, the K-fold cross-validation is employed in the course of training, which divides the dataset into 4 folds. Our framework is implemented in MindSpore with 2 NVIDIA GTX3090 GPUs for training. We trained all the models for

100 epochs with the batch size set to 2. Multiple GPUs are allowed to accelerate the training.

C. Results

In order to verify the performance of the network qualitatively and quantitatively, we conducted a series of segmentation experiments based on EndoVis 2017 dataset. We performed a qualitative comparison with state-of-the-art models, such as U-Net [20], TerausNet [25], LinkNet-34 [11] and PlainNet [10], and listed the results of three segmentation tasks in the TABLE I. As we shown in the TABLE I, for binary segmentation, our method achieves the best result, i.e., the IOU score of 82.94% and the Dice score of 89.42%. In particular, 1) compared to U-Net, our model achieves an improvement of 7.5 points for IOU and 5.05 points for Dice. 2) compared to prior advanced methods, we still obtain the best segmentation performance. For parts segmentation, our method does not show the best performance compared to TerausNet and LinkNet-34, but it still greatly improves its segmentation performance over the UNet for 9.97 points. For the task of multi-class class instrument segmentation, as we list in the table, our network achieves the best performance with the IOU score of 41.72% and the Dice score of 48.22% and the result is far superior to others methods. Since several of the seven categories of instruments only appear a few times in the training dataset, the performance of this task is overall lower. The result suggests that increasing the size of the dataset for the corresponding problem can effectively improve the performance.

We have also demonstrated a more intuitive result by visualizing the result of the segmentation tasks of our model on the dataset in Fig. 2. There are three different sub-tasks, i.e., binary segmentation (2 classes), part of instrument segmentation (4 classes), and instrument type segmentation (8 classes). For binary and parts segmentation, we encode the ground truth labels with values (10, 20, 30, 40, 0) to distinguish the background and every part of an instrument. Besides, the instrument type labels are used to classify different surgical instruments, and they are encoded with an incremental numerical value starting from 1 to 7. In order to display the segmentation effect more clearly, we convert the image of the segmentation result into color. As shown in the figure, the instruments and backgrounds are distinguished by purple and yellow, respectively. Three parts of each instrument can be identified individually by different colors while yellow represents the clasper, green represents the wrist and blue represents the shaft, respectively. For the type segmentation, the seven classes of instruments are also distinguished by different colors. Fig. 2 shows that our model can basically complete the detection and segmentation of instrument edges and types well. For case 3 and case 5, for parts segmentation, the surgical instrument in the lower right corner of the picture is not segmented very well. The possible explanation is that it was caused by the reflection of the light from the instrument.

In addition, we evaluate our trained network for instrumentation segmentation on ten different test video sequences,

which consists of 8×75 frame sequences and two full 300-frame sequences. TABLE II lists the performances of our method and that of other ten teams. We show the test result of our proposed model in ten datasets and compare it to the results of ten teams. As shown in the table, it is noticeable that the method we proposed achieves the highest IOU score in 6 datasets. This result demonstrates the high efficiency and accuracy of our model for the semantic segmentation task.

IV. CONCLUSION

In this work, we present a novel model for robotic surgical instrument segmentation, which can address three kinds of surgical instrument segmentation tasks in surgical scenes. The model we proposed is a nested U-structure which is based on the network architecture of UNet. Our method is compared with the existing state-of-the-art models in terms of IOU and Dice, and can achieve efficient and accurate segmentation. We also present comparative analysis of multiple deep network models through experimental data. The experimental results suggest that our model can highly optimize the surgical instrument segmentation and has achieved highly competitive performance for three sub-tasks, especially for the binary instrumentation segmentation and the type instrument segmentation.

REFERENCES

- [1] B. Singh, N. Sellappan, and P. Kumaradhas, "Evolution of industrial robots and their applications," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 5, pp. 763–768, 2013.
- [2] J. Burgner-Kahrs, D. C. Rucker, and H. Choset, "Continuum robots for medical applications: A survey," *IEEE Trans. Robot.*, vol. 31, no. 6, pp. 1261–1280, 2015.
- [3] S. Wang, Z. Zhou, and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE Robotics Autom. Lett.*, vol. 7, no. 3, pp. 8170–8177, 2022.
- [4] Z. Li, G. Li, X. Wu, Z. Kan, H. Su, and Y. Liu, "Asymmetric cooperation control of dual-arm exoskeletons using human collaborative manipulation models," *IEEE Transactions on Cybernetics*, vol. 52, no. 11, pp. 12 126–12 139, 2021.
- [5] G. A. Puerto-Souza, J. A. Cadeddu, and G.-L. Mariottini, "Toward long-term and accurate augmented-reality for monocular endoscopic videos," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2609–2620, 2014.
- [6] S. Leonard, A. Sinha, A. Reiter, M. Ishii, G. L. Gallia, R. H. Taylor, and G. D. Hager, "Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data," *IEEE Trans. Med. Imag.*, vol. 37, no. 10, pp. 2185–2195, 2018.
- [7] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, and L. Fei-Fei, "Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*. IEEE, 2018, pp. 691–699.
- [8] N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, "Live tracking and dense reconstruction for hand-held monocular endoscopy," *IEEE Trans Med. Imag.*, vol. 38, no. 1, pp. 79–89, 2018.
- [9] S. Bodenstedt, M. Allan, A. Agustinos, X. Du, L. Garcia-Peraza-Herrera, H. Kenngott, T. Kurmann, B. Müller-Stich, S. Ourselin, D. Pakhomov *et al.*, "Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery," *arXiv:1805.02475*. [Online]. Available: <https://arxiv.org/abs/1805.02475>, 2018.
- [10] Y. Jin, K. Cheng, Q. Dou, and P.-A. Heng, "Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video," in *Proc. Int. Conf. Med. Imag. Comput. Assist. Interv.* Springer, 2019, pp. 440–448.
- [11] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.* IEEE, 2018, pp. 624–628.
- [12] S. Wang, Z. Zhou, H. Wang, Z. Li, and Z. Kan, "Unsupervised representation learning for visual robotics grasping," in *IEEE Int. Conf. Adv. Robot. Mechatronics*. IEEE, 2022, pp. 57–62.
- [13] S. Speidel, M. Delles, C. Gutt, and R. Dillmann, "Tracking of instruments in minimally invasive surgery for surgical skill analysis," in *Medical Imaging and Augmented Reality: Third International Workshop, Shanghai, China, August 17-18, 2006 Proceedings 3*. Springer, 2006, pp. 148–155.
- [14] C. Doignon, F. Nageotte, and M. De Mathelin, "Segmentation and guidance of multiple rigid objects for intra-operative endoscopic vision," in *Dynamical Vision: ICCV 2005 and ECCV 2006 Workshops, WDV 2005 and WDV 2006, Beijing, China, October 21, 2005, Graz, Austria, May 13, 2006. Revised Papers*. Springer, 2007, pp. 314–327.
- [15] R. Sznitman, R. Richa, R. H. Taylor, B. Jedynek, and G. D. Hager, "Unified detection and tracking of instruments during retinal microsurgery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1263–1273, 2012.
- [16] N. Rieke, D. J. Tan, C. A. di San Filippo, F. Tombari, M. Alsheikhali, V. Belagiannis, A. Eslami, and N. Navab, "Real-time localization of articulated surgical instruments in retinal microsurgery," *Med. Image Anal.*, vol. 34, pp. 82–100, 2016.
- [17] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, and P. Jannin, "Detecting surgical tools by modelling local appearance and global shape," *IEEE Trans. Med. Imag.*, vol. 34, no. 12, pp. 2603–2617, 2015.
- [18] Z. Pezzementi, S. Voros, and G. D. Hager, "Articulated object tracking by rendering consistent appearance parts," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2009, pp. 3940–3947.
- [19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3431–3440.
- [20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Imag. Comput. Assist. Interv.* Springer, 2015, pp. 234–241.
- [21] I. Laina, N. Rieke, C. Rupprecht, J. P. Vizcaino, A. Eslami, F. Tombari, and N. Navab, "Concurrent segmentation and localization for tracking of surgical instruments," in *Proc. Int. Conf. Med. Imag. Comput. Assist. Interv.* Springer, 2017, pp. 664–672.
- [22] M. Attia, M. Hossny, S. Nahavandi, and H. Asadi, "Surgical tool segmentation using a hybrid deep cnn-rnn auto encoder-decoder," in *Proc. IEEE Int. Conf. Syst. Man Cybern.* IEEE, 2017, pp. 3373–3378.
- [23] F. Qin, Y. Li, Y.-H. Su, D. Xu, and B. Hannaford, "Surgical instrument segmentation for endoscopic vision with data fusion of cnn prediction and kinematic pose," in *Proc. Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2019, pp. 9821–9827.
- [24] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.
- [25] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv preprint arXiv:1801.05746*, 2018.
- [26] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, 2019.
- [27] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognit.*, vol. 106, p. 107404, 2020.
- [28] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.