Cross-Scene Semantic Segmentation for Medical Surgical Instruments Using Structural Similarity-Based Partial Activation Networks

Zhengyu Wang, Ziqian Li, Xiang Yu, Zirui Jia, Xinzhou Xu, *Member, IEEE*,
and Björn W. Schuller, *Fellow, IEEE*

***Abstract*—Robot-assisted minimally invasive surgery requires accurate segmentation for surgical instruments in order to guide surgical robots on tracking the target instruments. Nevertheless, it is difficult to perform surgical-instrument semantic segmentation in unknown scenes with extremely insufficient intra-scene surgical data, despite of the attempts for general semantic segmentation tasks. To address this issue, we propose a cross-scene semantic segmentation approach for medical surgical instruments using structural similarity based partial activation networks in this paper. The proposed approach includes a main branch for multi-level feature extraction, a segmentation head global consistency, and a structural similarity based loss function to provide high-level information acquisition, which improves the generalisation performance for the cross-scene segmentation task. Then, the experimental results in cross-scene surgical-instrument semantic segmentation cases show the effectiveness of the proposed approach compared with state-of-the-art semantic segmentation ones, using the newly established endoscopic simulation dataset.**

***Index Terms*—Cross-scene semantic segmentation, surgical instruments, structural similarity, partial activation networks.**

I. Introduction

RECENT progress has witnessed the effectiveness of semantic segmentation for surgical instruments in robot-assisted minimally invasive surgery, through mining visual information in surgical scenes [1], [2]. As a critical procedure in surgical-robot control, these instruments' segmentation aims to separate the instruments' foreground from the organ background, which can be applied to instrument tracking and pose estimation tasks [3], [4]. In addition to these tasks, the masks generated by the segmentation process also contribute to workflow analysis and process optimisation for surgical applications, for the purpose of reducing the workload of doctors and improving surgical safety [5], [6].

Within the solutions to surgical-instrument semantic segmentation, deep learning based approaches make the segmentation robust and adaptive, through refining target information on data resources [7], [8]. Most of the approaches employ U-Net frameworks containing *Convolutional Neural Networks* (CNNs) [9], focusing on improving the segmentation's accuracy and response speed in complex surgical environments (e. g., occlusion, motion artefacts, blood stains, and smoke). To this end, the networks consider improvements on lightweight operations [1], [10], [11], novel attention modules [12], [13], [14], multi-scale or multi-level fusion [2], [5], [15], and prior knowledge [16], [17]. In addition, the inclusion of a Swin Transformer [18] also improves the performance of semantic segmentation for surgical instruments, benefiting from its strong generalisation ability [19].

Further, on the aspect of data preparation, the past organised *Medical Image Computing and Computer Assisted Intervention* (MICCAI) sub-challenges on surgical-instrument segmentation provide high-quality labelled images of endoscopic procedures, in order to meet the requirements for large-scale fine-grained labelled surgical instrument data [20], [21], [22]. Considering the challenge on the lack of large-scale finely-annotated datasets, computer-generated data [23], [24] and cross-domain segmentation networks [25], [26] have been employed to alleviate training issues, arising from the scarcity of the finely-annotated data.

This work was supported in part by the Natural Science Foundation of China under Grant 52175221; in part by the China Postdoctoral Science Foundation under Grant 2022M711693; in part by the Fundamental Research Funds for the Central Universities under Grant 020414380195; and in part by the German Research Foundation (DFG) Reinhart Koselleck-Project AUDI0NOMOUS under Agreement 442218748. *(Corresponding authors: Zhengyu Wang; Xinzhou Xu.)*

Zhengyu Wang, Ziqian Li, Xiang Yu, and Zirui Jia are with the School of Mechanical Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: wangzhengyu_hfut@hfut.edu.cn; 2021110095@mail.hfut.edu.cn; 2021170200@mail.hfut.edu.cn; 2021110151@mail.hfut.edu.cn).

Xinzhou Xu is with the School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China, also with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and also with the Key Laboratory of Modern Acoustics, MOE, Nanjing University, Nanjing 210093, China (e-mail: xinzhou.xu@njupt.edu.cn).

Björn W. Schuller is with the Chair of Health Informatics, Technische Universität München, 80333 Munich, Germany, also with the Group on Language, Audio, & Music, Imperial College London, SW7 2BX London, U.K., and also with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: schuller@tum.de).

Digital Object Identifier 10.1109/TMRB.2024.3359303

In spite of these achievements, two deficiencies still exist in current research on semantic segmentation for surgical instruments. First, conventional approaches on surgical-instrument segmentation fail to sufficiently model cross-scene surgical-instruments segmentation on generalisation and adaptation, especially when confronting unfamiliar or unknown surgical scenarios divergent from their training environments. Second, conventional surgical-instrument segmentation models fail to jointly consider local nuances, global context, and structure information, which may lead to incomplete representation of the intricate details of surgical instruments especially in the cross-scene cases. In this regard, we propose a *Structural Similarity based Partial Activation Network* (SSPAN)[1] in cross-scene semantic segmentation for medical surgical instruments, intended to make up the deficiencies on exploring the connection between simulated and real-world surgical scenes, and designing partial activation networks with *Structural SIMilarity* (SSIM) [27] for the cross-scene cases, in order to jointly fuse local and global features, while focusing on structure information.

The proposed approach consists of three modules: First, the main branch contains a multi-sub-branch architecture for extracting pixel-level and region-level image features, in order to jointly construct local and global representations. Then, the segmentation head improves global consistency through introducing a partial attention mechanism, based on both of these representations, while retaining local specificity. This can help to alleviate variability (e. g., texture, size, and shape) between target instruments from the same class in different scenes. Finally, to further improve the generalisation ability, a loss function based on structural similarity is designed to consider both low-level and high-level information for the cross-scene segmentation. In comparison with related surgical-instrument segmentation works, the proposed approach focuses on cross-scene cases, through considering structural similarity based partial activation networks.

The contributions of this work are shown as follows:

- We propose a novel approach of SSPAN in cross-scene semantic segmentation for medical surgical instruments.
- Within the proposed SSPAN approach, we design a multi-branch structure and partial attention mechanism to jointly extract and fuse local and global feature information.
- Within the proposed SSPAN approach, we also design an SSIM-based loss function for cross-scene learning.

The remainder of this paper is organised as follows. Section II presents related work. The specific methodology of the proposed approach is then presented in Section III. Finally, Sections IV and V present the experiments, analysis, and conclusion.

II. Related Work

A. Learning-Based Surgical Instrument Segmentation

Learning-based surgical instrument segmentation aims at constructing segmentation models with the knowledge from collected data in surgical cases, due to the excellent adaptive capability [5], [6], [9].

Recent efforts have concentrated on fusing features across varying scales and hierarchical levels, as well as applying diverse attention mechanisms [28], [29], to enhance the models' representational capacity. Cerón et al. [30] proposed a method that aggregates multi-scale semantic information by fusing feature maps from the last four blocks of ResNet-101, augmented with multiple attention modules to boost representational strength and segmentation accuracy. Yang et al. [31] employed a residual path for contextual feature fusion and integrated a non-local attention block in the bottleneck layer, introducing dual attention modules to suppress irrelevant features and improve local feature representation, while Shen et al. [32] utilised both spatial and channel attention mechanisms to merge semantic insights from various levels, capturing extensive contextual details and thus improving segmentation precision. Furthermore, Qin et al. [33] employed a multi-angle feature aggregation approach, enhancing the model's robustness to directional variations in instruments by aligning features across different rotational angles.

Further research on cross-scene medical surgical instrument segmentation primarily addresses the performance degradation of models caused by domain shift between different surgical scenarios, for the purpose of improving models' generalisation for binary segmentation tasks in similar scenes. This can be tackled through various techniques such as style transformation [25], prototype learning [26], and incorporating prior knowledge [34]. Note that these works require sufficient real-world surgical data in training the models.

B. Surgical-Instrument Datasets

Conventional surgical-instrument datasets for semantic segmentation tasks contain fine-grained labelled data collected in real-world surgical scenes. For example, for the endoscopic surgical-instrument datasets, a series of MICCAI sub-challenges (EndoVis 2015 [20], EndoVis 2017 [21], and EndoVis 2019 [35]) present available visual data derived from endoscopic surgical videos. These datasets are labelled with rigid surgical or robotic-surgical instruments, and contain complex environmental factors (e. g., smoke, blood stains, motion artifacts, and occlusions).

Considering the limitations on diversity and adaptation for real-world data [36], simulation-supervised image synthesis has been investigated in surgical-instrument segmentation, primarily using computer simulators of dV-Trainer [37], RobotiX mentor [38], and *Asynchronous Multi-Body Framework* (AMBF) [39] to generate images of simulated surgical instruments. Due to the style discrepancies between the synthetic and real-world data, the style-transfer technique of CycleGAN is employed to convert synthesized images into more lifelike data [23], while Colleoni et al. [40] introduced attention mechanisms into the CycleGAN to better focus on domain-specific features.

III. Methodology

We propose the SSPAN approach in cross-scene semantic segmentation for surgical instruments, as shown in Fig. 1, with

[1] https://github.com/Li71226006/SSPAN

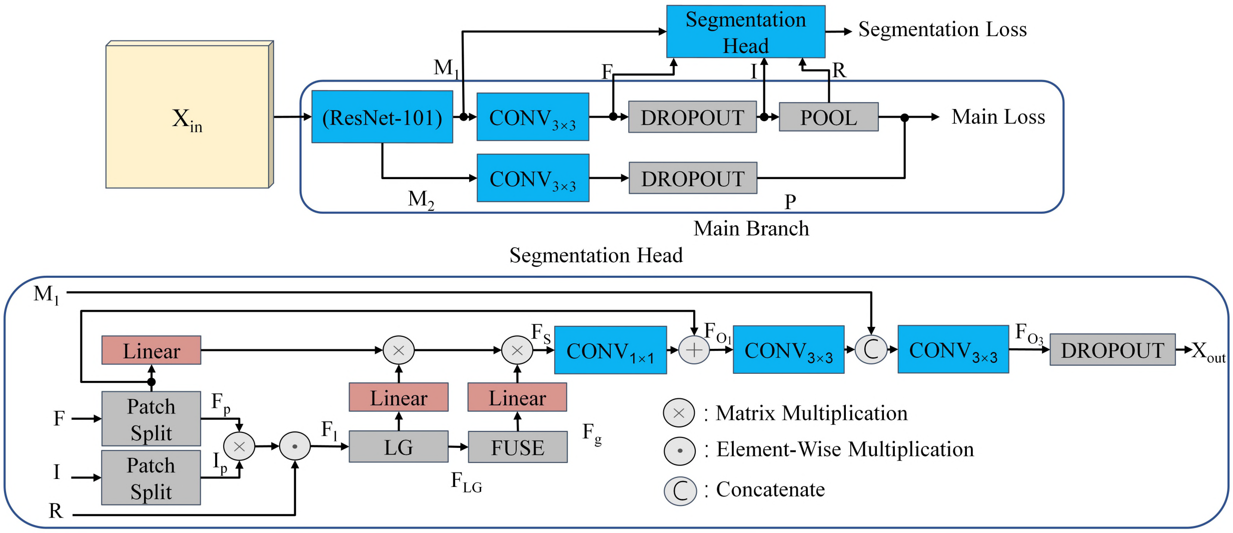


Fig. 1. Diagram of the network for the proposed approach, including a main branch performing processing based on the ResNet-101 backbone, and a segmentation head using the input feature maps from the main branch.

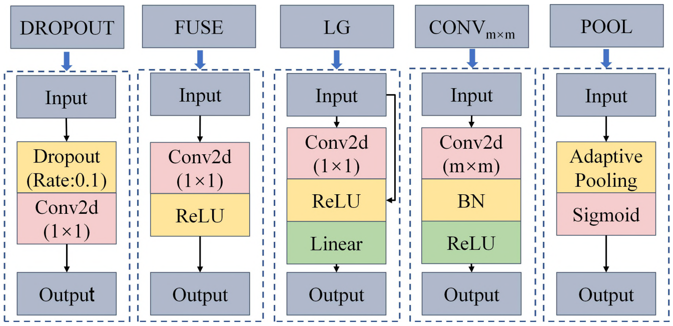


Fig. 2. Description for the five modules of DROPOUT, FUSE, LG, CONV$_{m\times m}$, and POOL in the network diagram, where m indicates the convolutional kernel size.

Fig. 2 introducing the six modules appearing in Fig. 1. As a two-branch structure, the proposed SSPAN contains a main branch designed for integrating pixel-level and region-level features, in order to jointly extract global and local semantic information. In addition to the main branch, a segmentation head is developed based on partial attention mechanisms, including region partitioning and self-attention mechanisms applied to the feature map.

A. Main Branch

The main branch serves to enhance the network's representational and generalisation capabilities by enabling the model to learn feature representations at different granularities through pixel-level and region-level tasks. It also provides the segmentation head with image features from multiple hierarchical levels.

Note that an arbitrary input image $\mathbf{X}_{in}(H_{in} \times W_{in} \times C_{in})$ contains C_{in} channels, with the height H_{in} and the width W_{in}. Then, the image is processed by the ResNet-101 backbone to extract its feature maps $\mathbf{M}_1(H \times W \times C_1)$ and $\mathbf{M}_2(H \times W \times C_2)$ from the last two blocks of ResNet-101 containing C_1 and C_2 convolutional channels, respectively, with the height H and the width W [41], [42]. The two blocks' outputs can help to prevent supervision vanishing and reduce internal redundancy within the model, which ultimately enhances the model's generalisation ability [43], [44].

To jointly extract local and global information, the main branch contains two sub-branches to extract pixel-level and region-level information. The pixel sub-branch utilises the CONV$_{m\times m}$ and DROPOUT modules (with C_3 and K channels, respectively) to obtain the pixel-level feature map $\mathbf{P}(H \times W \times K)$ written as

$$\mathbf{P} = \text{Softmax}(\text{DROPOUT}(\text{CONV}_{3\times3}(\mathbf{M}_2))), \qquad (1)$$

performing a Softmax operation on the K segmentation classes, where CONV$_{m\times m}$ and DROPOUT represent the different operational modules as in Fig. 2 with a *Rectified Linear Unit* (ReLU) activation and *Batch Normalisation* (BN), where m is the kernel size.

In addition to CONV$_{m\times m}$ and DROPOUT modules, the region sub-branch additionally employs the POOL module to divide the feature map into non-overlapping regions, obtaining the region-level feature map $\mathbf{R}(K \times S \times S)$ with the division size S in adaptive pooling. As a crucial component of the POOL module, adaptive average pooling divides the input feature map into predefined regions and performs averaging within each to yield regional representations. These representations are subsequently normalised to a range between 0 and 1 through a sigmoid function.

In this sub-branch, semantic features $\mathbf{F}(H \times W \times C_4)$ and semantic information $\mathbf{I}(H \times W \times K)$ (with C_4 and K channels, respectively) are obtained through

$$\mathbf{R} = \text{POOL}(\underbrace{\text{DROPOUT}(\underbrace{\text{CONV}_{3\times3}(\mathbf{M}_1)}_{\mathbf{F}})}_{\mathbf{I}}), \qquad (2)$$

where the POOL module in Fig. 2 employs a sigmoidal function with an adaptive average-pooling operator.

B. Segmentation Head

The segmentation head first divides the feature maps generated by the main branch into non-overlapping regions in a fixed quantity. Then, it implements the partial attention mechanism through facilitating feature interaction, fusion, and attention calculation within these regions. This process allows for joint consideration of local specificity and global consistency, so as to enhance the model's generalisation performance.

In order to represent local specificity, the segmentation head employs a patch-split strategy to divide $\mathbf{F}$ and $\mathbf{I}$ into non-overlapping regions $\mathbf{F}_p(h \times w \times C_4 \times N_p)$ and $\mathbf{I}_p(h \times w \times K \times N_p)$, where $N_p = S^2$ represents the total number of the regions, with the size of $h = \frac{H}{S}$ and $w = \frac{W}{S}$.

Afterwards, akin to OCRNet in class-feature correlation calculation [45], the weighted class-feature affinity matrix $\mathbf{F}_l^{(i)}$ of the ith region ($i = 0, 1, \ldots, N_p - 1$) can be obtained through

$$\mathbf{F}_l^{(i)} = (\hat{\mathbf{R}}^{(i)} \mathbf{e}_{C_4}^\top) \odot (\mathrm{Softmax}(\hat{\mathbf{I}}_p^{(i)})^\top \hat{\mathbf{F}}_p^{(i)}), \tag{3}$$

where the ith region's reshaped feature maps of $\mathbf{I}^{(i)}(h \times w \times K)$ and $\mathbf{F}^{(i)}(h \times w \times C_4)$ can be represented as $\hat{\mathbf{I}}_p^{(i)}(N \times K)$ and $\hat{\mathbf{F}}_p^{(i)}(N \times C_4)$, where $N = hw$. Then, a Softmax operator is performed across the K segmentation classes for each row of $\hat{\mathbf{I}}_p^{(i)}$. Within Eq. (3), we also utilise the region-level feature map $\mathbf{R}$ as the weights to enhance the class-relevant features, while weakening the irrelevant features for each region, with '$\odot$' indicating element-wise multiplication. Hence, $\mathbf{R}$ is reshaped as $\hat{\mathbf{R}}(K \times N_p)$, and $\hat{\mathbf{R}}^{(i)}(K \times 1)$ contains the weights of the i-th region. By duplicating $\hat{\mathbf{R}}^{(i)}$ along the column direction, we transform it into a shape of $K \times C_4$ using the column vector $\mathbf{e}_{C_4} = [1, 1, \ldots, 1]^\top$.

Afterwards, the weighted class-feature affinity matrix $\mathbf{F}_l(K \times C_4 \times N_p)$ can be obtained using $\mathbf{F}_l^{(i)}$s. Further, in order to improve the generalisation of local regions for each class, we employ the *Local Gather* (LG) module (see Fig. 2) to enable information interchange across different regions [42], [46]. This is achieved by performing 1×1 convolution on the N_p regions. Then, the convolution results are passed through a ReLU activation combined with $\mathbf{F}_l$. Then, the feature map is fed to a linear operation with C_4 nodes on the C_4 channels, resulting in outputting the feature map $\mathbf{F}_{\mathrm{LG}}(K \times C_4 \times N_p)$. The residual structure and linear transformation are incorporated, aiming at facilitating gradient propagation and strengthening the association between classes and features. Using the LG module's output $\mathbf{F}_{\mathrm{LG}}$, we fuse different local regions through the FUSE module (see Fig. 2) and obtain global features $\mathbf{F}_g(K \times C_4)$. The FUSE module also includes 1×1 convolution on the N_p regions with 1 channel and a ReLU activation, as employed in [47], [48]. This design merges multiple channels into one, using an activation function for nonlinear integration of weighted features, for the purpose of enhancing representation and complexity management in the fused feature map.

For the ith patch, we feed the obtained $\hat{\mathbf{F}}_p^{(i)}(N \times C_4)$, $\mathbf{F}_{\mathrm{LG}}^{(i)}(K \times C_4)$ (from the ith patch in $\mathbf{F}_{\mathrm{LG}}(K \times C_4 \times N_p)$), and $\mathbf{F}_g(K \times C_4)$ to three linear mappings $W_p(\cdot)$, $W_{\mathrm{LG}}(\cdot)$, and $W_g(\cdot)$ (each with C_5 nodes) on the C_4 channels, respectively. This leads to calculating the ith patch's enhanced feature maps represented as

$$\mathbf{F}_s^{(i)} = \mathrm{Softmax}\Big(W_p\big(\hat{\mathbf{F}}_p^{(i)}\big) W_{\mathrm{LG}}\big(\mathbf{F}_{\mathrm{LG}}^{(i)}\big)^\top \Big) W_g\big(\mathbf{F}_g\big), \tag{4}$$

considering the non-local structure in [49], with the Softmax performing across the K classes. Hence, we obtain the reshaped enhanced features $\hat{\mathbf{F}}_s(N \times C_5 \times N_p)$ and its non-reshaped form $\mathbf{F}_s(h \times w \times C_5 \times N_p)$.

The aforementioned process constitutes the partial attention mechanism, through segmenting the feature map into fixed-size non-overlapping regions. Then, the target is to improve intra-region and inter-region feature interactions (via the LG module), fusion (via the FUSE module), and attention mechanisms (employing non-local structures), in order to achieve global consistency and local specificity.

Then, we perform $\mathrm{CONV}_{1\times1}$ (with C_4 channels) for $\mathbf{F}_s$ across its C_5 channels, and obtain

$$\mathbf{F}_{o_1}(h \times w \times C_4 \times N_p) = \mathrm{CONV}_{1\times1}(\mathbf{F}_s) + \mathbf{F}_p, \tag{5}$$

and reshape it into $\mathbf{F}_{o_2}(H \times W \times C_4)$.

When further performing $\mathrm{CONV}_{3\times3}$ (with C_4 channels) across the channels, we form $\mathbf{F}_{o_3}(H \times W \times C_4)$ as

$$\mathbf{F}_{o_3} = \mathrm{CONV}_{3\times3}\big(\mathrm{Concat}(\mathrm{CONV}_{3\times3}(\mathbf{F}_{o_2}), \mathbf{M}_1)\big), \tag{6}$$

where the $\mathrm{Concat}(\cdot)$ indicates concatenation on the channels.

Finally, we employ the DROPOUT module with K channels to acquire the output of the segmentation head as

$$\mathbf{X}_{out}(H \times W \times K) = \mathrm{Softmax}(\mathrm{DROPOUT}(\mathbf{F}_{o_3})). \tag{7}$$

Note that Equations (5) and (6) pertain to feature fusion operations, where features are either added or concatenated to enhance feature representation. Equation (7) introduces a DROPOUT module, designed to improve models' robustness against overfitting.

C. Loss Function

The loss function L of the network can be divided into main loss L_{main} and segmentation loss L_{seg}, aiming at guiding the main branch to extract and fuse local and global information, and assisting the segmentation head to better segment the target, respectively, forming the loss function as

$$L = L_{main} + L_{seg}. \tag{8}$$

The **main loss** L_{main} can be further divided into two terms: The pixel-level loss function utilises *Cross Entropy* (CE) loss, while the region-level loss function considers *Focal Loss* (FL) shown as

$$L_{main} = l_{pixel} + l_{region}. \tag{9}$$

Note that the one-hot ground-truth segmentation annotation of an image is represented as $\mathbf{Y}(H_{in} \times W_{in} \times K)$. Meanwhile, we perform up-sampling for the feature map $\mathbf{P}(H \times W \times K)$ using linear interpolation, leading to $\mathbf{Q}(H_{in} \times W_{in} \times K)$. We thereby calculate

$$l_{pixel} = -\frac{\lambda_1}{H_{in} W_{in}} \sum_{j_1=1}^{H_{in}} \sum_{j_2=1}^{W_{in}} \sum_{k=1}^{K} Y_{j_1 j_2 k} \log Q_{j_1 j_2 k}, \tag{10}$$

where the weight $\lambda_1 = 0.4$, while $Y_{j_1j_2k}$ and $Q_{j_1j_2k}$ correspond to the (j_1, j_2, k)th elements of $\mathbf{Y}$ and $\mathbf{Q}$, respectively.

For the region-level term, we reshape $\mathbf{Y}$ and perform adaptive average-pooling on each region, resulting in $\mathbf{Z}(K \times S \times S)$. Then, using the region-level feature map $\mathbf{R}$ with Softmax processing, we further obtain its annotation-related representation $\mathbf{R}'(K\times S\times S)$ with the corresponding (k', s_1, s_2)-element $R'_{k's_1s_2} = R_{k's_1s_2}$ when the corresponding element $Z_{k's_1s_2} = 1$ (from $\mathbf{Z}$), otherwise equal to $1 - R_{k's_1s_2}$. Therefore, the region-level loss is shown as

$$l_{region} = -\frac{\lambda_2}{S^2}\sum_{k'=1}^{K}\sum_{s_1=1}^{S}\sum_{s_2=1}^{S}(1 - R'_{k's_1s_2})^{\gamma}\log R'_{k's_1s_2}, \quad (11)$$

where the weight $\lambda_2 = 1$, and $\gamma = 2$ is used to adjust the balance between positive and negative elements.

For the **segmentation loss** L_{seg}, we set a joint form of

$$L_{seg} = l_{SSIM} + l_{CE}, \quad (12)$$

using an SSIM-loss term l_{SSIM} and a CE-loss term l_{CE}.

The SSIM is commonly used as a loss function and evaluation metric in image reconstruction tasks [50], [51]. It measures the structural similarity between images, taking three aspects of luminance, contrast, and structure into account on image quality, resulting in general representation of structural features [52], [53].

In contrast to these works, this paper extends SSIM for use in the loss function of multi-class semantic segmentation tasks, considering to capture structure information through fitting statistics [54], shown as

$$l_{SSIM} = 1 - \frac{1}{K}\sum_{j=1}^{K} D^{(j)}_{MSSIM}, \quad (13)$$

where $D^{(j)}_{MSSIM}$ refers to the mean structural similarity between the prediction $\mathbf{X}_{out}(H \times W \times K)$ and the ground truth $\mathbf{Y}(H_{in} \times W_{in} \times K)$ on the jth segmentation class.

In this way, we first perform up-sampling (using linear interpolation) and then normalisation (using Softmax on the K segmentation classes) on $\mathbf{X}_{out}$, leading to the transformed prediction $\tilde{\mathbf{X}}_{out}(H_{in} \times W_{in} \times K)$. We go on splitting $\tilde{\mathbf{X}}_{out}$ and $\mathbf{Y}$ into N_r regions using a 11×11 sliding window with the stride of 1. Then, we obtain the mean structural similarity

$$D^{(j)}_{MSSIM} = \frac{1}{N_r}\sum_{j=1}^{N_r}\frac{\left(2\mu_x^j\mu_y^j + \tau_1\right)\left(2\sigma_{xy}^j + \tau_2\right)}{\left((\mu_x^j)^2 + (\mu_y^j)^2 + \tau_1\right)\left((\sigma_x^j)^2 + (\sigma_y^j)^2 + \tau_2\right)}, \quad (14)$$

where μ_x^j and $(\sigma_x^j)^2$ correspond to the mean and standard deviation of $\tilde{\mathbf{X}}_{out}$, respectively, while μ_y^j and $(\sigma_y^j)^2$ are the mean and standard deviation of $\mathbf{Y}$, respectively. σ_{xy}^j indicates the covariance between $\tilde{\mathbf{X}}_{out}$ and $\mathbf{Y}$. We also set the biases $\tau_1 = 10^{-4}$ and $\tau_2 = 0.03$.

In addition to l_{SSIM}, the CE loss l_{CE} is included in L_{seg}, using $\tilde{\mathbf{X}}_{out}$ and $\mathbf{Y}$. For the inference step, we use $\tilde{\mathbf{X}}_{out}$ as the segmentation output for an arbitrary test sample, to achieve its final semantic-segmentation prediction.

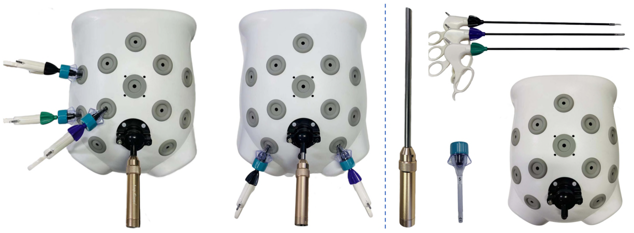
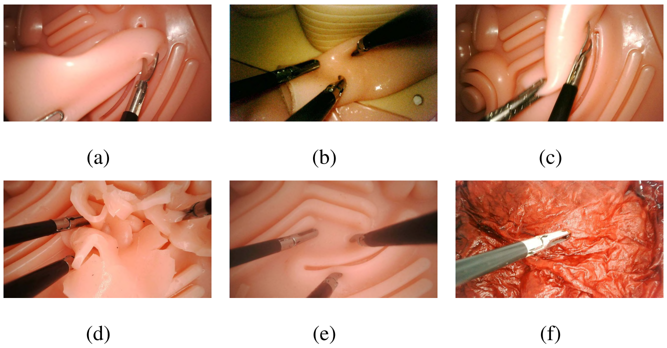


Fig. 3. The laparoscopic surgical-simulation platform for the simulation dataset, with the right part showing the components, while the left part is showing the ways of standing.



(a) (b) (c)

(d) (e) (f)

Fig. 4. Six basic challenging scenarios considered in the endoscopic simulation dataset, including (a) category imbalance, (b) illumination imbalance, (c) motion artifacts, (d) occlusion, (e) smoke, and (f) blood stains.

IV. Experiments

A. Data Preparation

The Simulation Dataset: The *Artificial Laparoscopic Instrument Dataset* (ALID)[2] is a fine-grained labelled dataset built by us, using a simulated operating table at different surgical stations. More specifically, we have used the BellySim laparoscopic simulation training system [60], for the purpose of simulating multi-station clinical training in thoracoscopy and laparoscopy in an artificial-pneumoperitoneum form.

As shown in Fig. 3, the laparoscopic simulation training system consists of trocars, a 30-degree focusable endoscope, and a pneumoperitoneum morphology experimental platform. In addition to these, the system also contains three types of rigid surgical instruments (scissors, grasping forceps, and detachment forceps), consistent with clinical specifications. Due to the fact that endoscopic surgical instruments usually have different insertion points, two stations have been designed in order to simulate different orientations of the surgical instruments in real-world cases.

Then, considering the influence from complex surgical environment for instruments' segmentation, we take into account multiple challenging scenarios frequently arising in laparoscopic surgery when acquiring the data. As can be seen from Fig. 4 and Table I, the ALID dataset contains 1 250 images with the size of 1080 × 1920 pixels, involving six scenarios of occlusion, illumination imbalance, category imbalance, smoke, motion artefacts, and blood stains, also including their pair-wise combinations. In addition, we define the segmentation target as both of the shaft and manipulator for

[2] https://github.com/Li71226006/ALID

TABLE I
THE PAIR-WISE COMBINATIONS OF THE SCENARIO CATEGORIES FOR THE SIMULATION DATASET, WITH CORRESPONDING NUMBERS OF IMAGE SAMPLES FOR EACH PAIR

Category	Normal	Occlusion	Smoke	Blood Stains	Category Imbalance	Illumination Imbalance	Motion Artifacts
Normal	150	120	60	60	130	120	60
Occlusion	-	-	20	20	60	60	60
Smoke	-	-	-	30	20	20	20
Blood Stains	-	-	-	-	20	20	20
Category Imbalance	-	-	-	-	-	60	60
Illumination Imbalance	-	-	-	-	-	-	60

a surgical instrument used in building the dataset. As for the annotation, we employ the *Efficient Interactive Segmentation* (EISeg) toolbox to obtain the surgical-instrument masks by labellers [61], considering the inclusion of review and revision by the annotators and experts, in order to further ensure correct annotations.

The Real-World Datasets: We set the EndoVis 2015 and EndoVis 2017 datasets (from MICCAI EndoVis Challenge 2015 and 2017, respectively) [20], [21] as the real-world datasets in the experiments. The EndoVis 2015 dataset contains 300 surgical pictures of rigid surgical instruments from in-vivo environments, with the image size of 480×640 pixels, 160 of which are used for training and 140 for testing. The segmentation target is the shafts and manipulators of the instruments. The EndoVis 2017 dataset is built on endoscopic surgery, acquired by a DA VINCI XI surgical system. It contains 3 000 images with a resolution of 1080×1920 pixels, 1 800 of the images are used for training and 1 200 for testing. The segmentation target is the shafts, articulated wrists, and claspers of surgical instruments.

In order to facilitate the process, the size of EndoVis 2017 and ALID has been scaled down to 512 × 768 pixels. The articulated parts in EndoVis 2017 have been used as the manipulators of surgical instruments in order to harmonise the classes. Afterwards, for the validation procedures, we perform training on ALID and test on the test sets of the EndoVis 2015 and EndoVis 2017 datasets, in accordance with the setups in the EndoVis challenges. In addition, for the approaches with strong generalisation capabilities for the cross-scene cases, we also perform intra-scene training, and perform testing on the same test sets of the two real-world datasets as in the cross-scene cases, which is designed to provide segmentation's upper bounds for these approach.

B. Experimental Setups

Training Details: Within the training procedures, we employ a 'poly' learning rate policy [62], with the initial learning rates set as $\{0.0001, 0.001, 0.002, 0.005, 0.01, 0.02, 0.1\}$, for the total loss function L. The learning rates are also multiplied by $(1 - \frac{n_{iter}}{n_{max}})^{0.9}$, with n_{iter} and n_{iter} indicating the current and maximum numbers of iteration. The optimiser is chosen as *Stochastic Gradient Descent* (SGD) [63], with a momentum 0.9 and a weight decay of 10^{-4}. The batch size is set to 4. The numbers of channels C_1 to C_5 are set to 2 048, 1 024, 256, 512, and 256, respectively. The region-number parameter $S = 4$, resulting in $N_p = 16$, and the feature maps' size is set as $H = \frac{H_{in}}{8}$ and $W = \frac{W_{in}}{8}$. All the training samples are enhanced by random horizontal and vertical flipping for each epoch.

Evaluation Indicators: In order to measure the similarity between predicted and ground-truth segmentation, the Jaccard index (noted as 'J'; or equivalently *Intersection over Union* (IoU)) and the Dice score (noted as 'Dice') are used in performance evaluation for the semantic segmentation approaches. Higher Jaccard-index results and Dice scores indicate better performance, represented as

$$\mathrm{J} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{15}$$

$$\mathrm{Dice} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{16}$$

where TP, FP, and FN represent true-positive, false-positive, and false-negative subsets of pixels, respectively. Note that we aim to separately segment both of the shaft and manipulator for a surgical instrument, which leads to multi-class segmentation tasks in the experiments [64]. For each class, we calculate the J and Dice coefficients. To obtain a comprehensive metric evaluation, we use a macro-average, which means averaging the metrics across all class.

C. Experimental Results

Comparison with Existing Approaches: First, in order to show the performance of the proposed approach for the cross-scene cases, we present the experimental results on the EndoVis 2015 and EndoVis 2017 datasets in Table II, in terms of the J and Dice indicators for the multi-class segmentation. Note that for the cross-scene case, the models are trained and tested on the ALID dataset and the test set of each real-world dataset, respectively. The compared semantic segmentation approaches include U-Net [9], *Attention-guided LightWeight Network* (LWANet) [11], *Residual Attention U-Net* (RAUNet) [56], *Refined Attention Segmentation Network* (RASNet) [57], Swin Transformer [19], ISNet [58], DeeplabV3+ [59] and *Partial Class Activation Attention* (PCAA) [42], considering ResNet [65], MobileNetV2 [55], and UperNet [66] as the backbone networks. For the experimental results on the EndoVis 2015 and EndoVis 2017 datasets, the proposed SSPAN achieves the best results when using both of the indicators (J = 56.0% & Dice = 67.9% for EndoVis 2015, and J = 65.5% & Dice =

TABLE II
CROSS-SCENE SEMANTIC-SEGMENTATION RESULTS IN TERMS OF JACCARD INDEX (J; %) AND DICE SCORE (Dice; %) AS INDICATORS ON THE ENDOVIS 2015 AND ENDOVIS 2017 DATASETS, USING THE PROPOSED AND STATE-OF-THE-ART APPROACHES. NOTE THAT THE INDICATORS IN THE BRACKETS REFER TO THE INTRA-SCENE CASES AS THE UPPER BOUNDS FOR THE TEST SETS OF THE TWO REAL-WORLD DATASETS

Approach	**Backbone**	**EndoVis 2015**		**EndoVis 2017**	
		J	Dice	J	Dice
U-Net [9]	-	40.5 (62.3)	51.7 (73.9)	35.6 (89.5)	48.7 (94.3)
U-Net [9], [55]	MobileNetV2	46.2 (73.3)	57.3 (83.4)	49.8 (89.4)	61.9 (94.2)
LWANet [11]	MobileNetV2	45.6 (73.3)	56.3 (83.3)	50.8 (89.5)	63.1 (94.3)
RAUNet [56]	ResNet-34	45.4 (75.2)	55.7 (84.7)	44.2 (90.8)	52.4 (95.0)
RASNet [57]	ResNet-50	43.8 (75.9)	55.1 (85.3)	44.3 (91.2)	54.7 (95.3)
Swin Transformer [19]	UperNet	47.1 (75.3)	58.4 (84.8)	53.0 (91.3)	64.7 (95.3)
ISNet [58]	ResNet-50	47.9 (74.9)	59.0 (84.5)	55.5 (90.4)	67.8 (94.8)
ISNet [58]	ResNet-101	50.6 (77.2)	61.8 (86.2)	62.5 (90.3)	74.4 (94.7)
DeepLabV3+ [59]	ResNet-101	51.8 (79.2)	63.3 (87.6)	62.0 (90.1)	74.0 (94.6)
PCAA [42]	ResNet-50	48.2 (76.7)	60.1 (85.8)	51.1 (89.3)	63.3 (94.1)
PCAA [42]	ResNet101	52.2 (78.1)	63.9 (86.8)	64.0 (90.0)	75.6 (94.6)
SSPAN (Proposed)	-	**56.0** (77.2)	**67.9** (86.2)	**65.5** (90.9)	**77.1** (95.1)

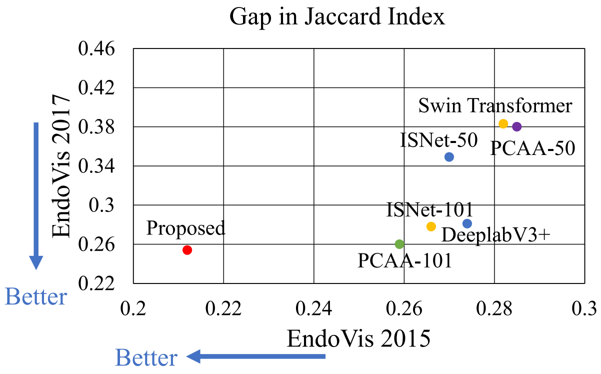


Fig. 5. Performance demonstration for the proposed and well-performed compared approaches, in terms of the gap of the Jaccard index (J) indicator between cross-scene and intra-dataset cases, with the horizontal and vertical axis representing the gaps when testing on the EndoVis 2015 and EndoVis 2017 datasets, respectively.

77.1% for EndoVis 2017). It is also observed from the upper-bound results that the proposed SSPAN obtains similar results for the intra-scene cases (J = 77.2% & Dice = 86.2% for EndoVis 2015, and J = 90.9% & Dice = 95.1% for EndoVis 2017), compared with the well-performed PCAA approaches. This strongly shows the cross-scene segmentation performance for the proposed SSPAN, in view of the gaps between the SSPAN and compared approaches in the cross-scene cases.

Then, for the sake of investigating generalisation ability of these approaches, we also present intra-scene results when training and testing on two real-world datasets in Table II without involving the ALID data (noted as 'cross-scene J or Dice (intra-scene J or Dice)'), showing the gaps between cross-scene and intra-scene cases in Fig. 5. The gaps in the indicator J are presented in the figure for EndoVis 2015 (horizontal) and EndoVis 2017 (vertical) datasets, with the points corresponding to Swin Transformer, ISNet with ResNet-50 (noted as 'ISNet-50'), ISNet with ResNet-101 (noted as 'ISNet-101'), DeeplabV3+, PCAA with ResNet-50 (noted as 'PCAA-50'), PCAA with ResNet-101 (noted as 'PCAA-101'), and the proposed SSPAN (noted as 'Proposed'). The results indicate that our proposed SSPAN achieves strong generalisation performance compared with the other approaches.

We further demonstrate the visulisation of the segmentation examples from the test sets of the EndoVis 2015 and EndoVis 2017 datasets in Fig. 6, when using Swin Transformer, ISNet-50, ISNet-101, DeeplabV3+, PCAA-50, PCAA-101, and our proposed SSPAN. The examples show that although the proposed approach yields better segmentation performance in separating foreground surgical instruments from the backgrounds, although not well-performing on distinguishing shafts and manipulators in the instruments.

To analyse class-wise performance, we present the confusion matrix for SSPAN alongside the foreground J and Dice coefficients, compared with the PCAA-101 model due to its performance from Table II, as shown in Fig. 7. The results indicate that it is not dominant for the proposed SSPAN to classify shafts and manipulators in the foreground areas. Then, we showcase samples from three datasets along with their annotations. Fig. 8 illustrates that, in the ALID dataset, there is a pronounced color distinction between the shaft and the manipulator, marked by an evident visual boundary. However, in real surgical datasets, the color differentiation between the shaft and effector can be subtle or even absent. Fig. 6 reveals that under such conditions, the model finds it challenging to distinguish between the shaft and the manipulator. To investigate binary segmentation, we consolidate the three-class predictive confusion matrices into a binary format by considering the shaft and manipulator as a single foreground class, resulting in calculating the foreground J and Dice coefficients. As depicted in Table III, SSPAN outperforms PCAA-101 on these foreground evaluation metrics, indicating SSPAN's

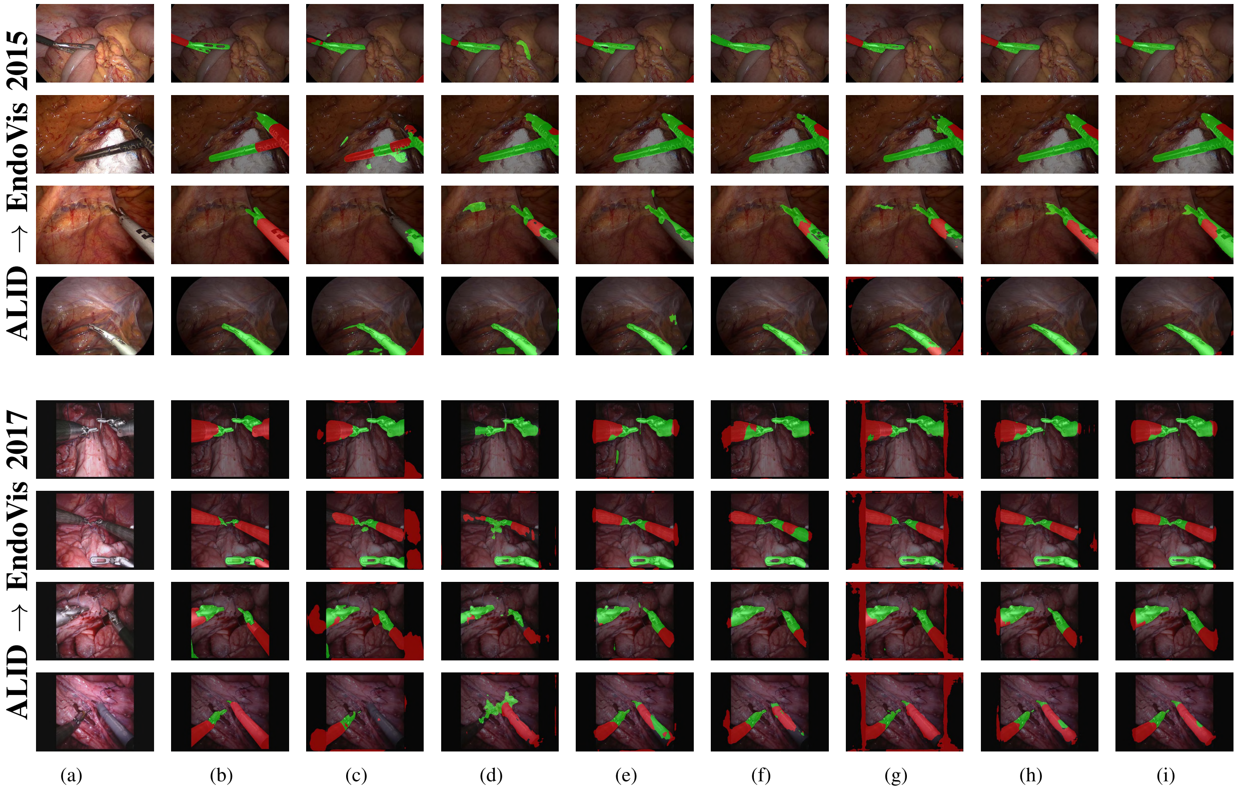


Fig. 6. Visualisation of cross-scene semantic segmentation results on Endovis 2015 (ALID → EndoVis 2015) and Endovis 2017 (ALID → EndoVis 2017) datasets. (a) Input images without segmentation, (b) corresponding ground truth, and (c) to (i) segmentation results obtained using Swin Transformer, ISNet-50, ISNet-101, DeeplabV3+, PCAA-50, PCAA-101, and the proposed SSPAN method, where the masks in red and green correspond to the segmentation for the shafts and manipulators, respectively.

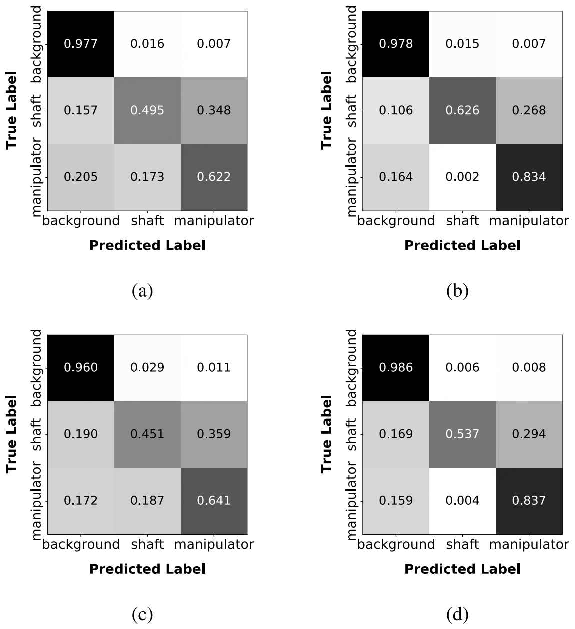


Fig. 7. The cross-scene three-class confusion matrices for SSPAN and PCAA-101 on the EndoVis 2015 (ALID → EndoVis 2015) and EndoVis 2017 (ALID → EndoVis 2017), where (a) and (b) depict the cross-scene segmentation performance of SSPAN on EndoVis 2015 and EndoVis 2017, respectively, while (c) and (d) represent the cross-scene segmentation performance of PCAA-101 on EndoVis 2015 and EndoVis 2017, respectively.

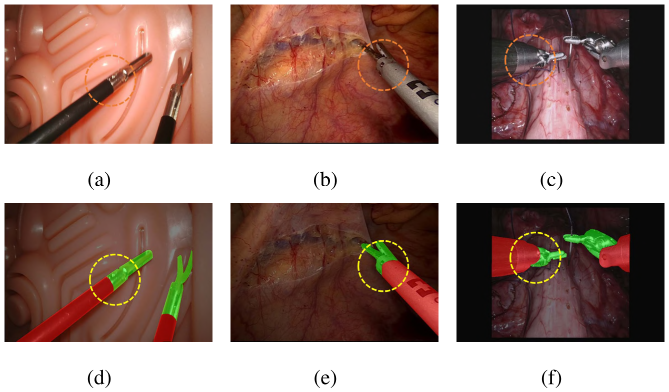


Fig. 8. Samples and annotated images from three datasets of (a) ALID without annotation, (b) EndoVis 2015 without annotation, (c) EndoVis 2017 without annotation, (d) ALID with annotation, (e) EndoVis 2015 with annotation, and (f) EndoVis 2017 with annotation, respectively.

superior ability to distinguish foreground from background. In addition, when making comparison on time complexity, the proposed SSPAN leads to similar training time and inference speed compared with PCAA-101 in the same condition of the experiments.

Quantitative Analysis: Following the performance comparison, we first aim to investigate the effectiveness of the proposed loss function. Within the segmentation loss L_{seg}, we consider the inclusion of CE loss (noted as 'CE'), SSIM loss (noted as 'SSIM'), and *Dice Loss* (DL) [56]. In this regard, we show the loss-term combinations of CE, CE & Dice (CE+Dice), CE & Dice & SSIM (CE+Dice+SSIM), and the proposed CE & SSIM (CE+SSIM) in Table III, using the J and Dice indicators. It is seen from the table that the

TABLE III
CROSS-SCENE SEMANTIC SEGMENTATION RESULTS FOR THE FOREGROUND OF SSPAN AND PCAA-101 IN TERMS OF JACCARD INDEX (J; %) AND DICE SCORE (Dice; %), ON THE ENDOVIS 2015 AND ENDOVIS 2017 DATASETS

Approach	**EndoVis 2015**		**EndoVis 2017**	
	J	Dice	J	Dice
PCAA-101	78.6	88.0	82.3	90.3
SSPAN	**80.7**	**89.3**	**85.4**	**92.1**

TABLE IV
SEGMENTATION-PERFORMANCE COMPARISON WHEN CONSIDERING DIFFERENT LOSS-FUNCTION COMBINATIONS FOR THE PROPOSED APPROACH ON THE ENDOVIS 2015 AND ENDOVIS 2017 DATASETS, IN TERMS OF JACCARD INDEX (J; %) AND DICE SCORE (DICE; %) AS INDICATORS

Segmentation Loss	**EndoVis 2015**		**EndoVis 2017**	
	J	Dice	J	Dice
CE	52.2	63.9	64.0	75.6
CE+Dice	53.4	65.3	64.1	75.8
CE+Dice+SSIM	53.7	65.5	**65.8**	**77.4**
CE+SSIM (Proposed)	**56.0**	**67.9**	65.5	77.1

TABLE V
SEGMENTATION PERFORMANCE WHEN CONSIDERING DIFFERENT SELECTIONS OF THE REGION-NUMBER PARAMETER S AND THE CONVOLUTIONAL KERNEL SIZE m IN THE CONV$_{m\times m}$, FOR THE PROPOSED APPROACH ON THE ENDOVIS 2015 AND ENDOVIS 2017 DATASETS, IN TERMS OF JACCARD INDEX (J; %) AND DICE SCORE (DICE; %) AS INDICATORS

Parameters (S, m)	**EndoVis 2015**		**EndoVis 2017**	
	J	Dice	J	Dice
$(2, 3)$	51.3	63.0	57.7	70.0
$(8, 3)$	52.6	64.2	58.8	70.9
$(4, 5)$	52.8	64.7	64.3	76.1
$(4, 7)$	52.0	63.7	60.5	72.6
$(4, 3)$ (Proposed)	**56.0**	**67.9**	**65.5**	**77.1**

proposed CE & SSIM loss-term setup outperforms the other combinations on the EndoVis 2015 dataset, and the results on the EndoVis 2017 dataset also demonstrate the effectiveness for the inclusion of the SSIM loss. The results indicate that the SSIM loss in the proposed SSPAN contributes to better cross-scene segmentation performance for surgical instruments.

Finally, we also investigate the influence of the region-number parameter S and different convolutional kernel size m (for all the CONV$_{m\times m}$ modules with $m > 1$) on the segmentation performance. These experiments aim to show the detailed design for the partial attention mechanism and the network's architecture. Table V shows the J and Dice results when considering different parametric combinations of (S, m) for the proposed SSPAN, using the ranges of $\{2, 4, 8\}$ and $\{3, 5, 7\}$ for the two parameters, respectively. It can be drawn from the table that the combination of $(S, m) = (4, 3)$ in the proposed approach corresponds to the best performance. This implies that the proposed SSPAN has to choose a slightly small kernel size, while considering to set a suitable number of the regions.

V. CONCLUSION

In this paper, we proposed a cross-scene semantic segmentation approach for medical surgical instruments using *Structural Similarity based Partial Activation Networks* (SSPAN). The proposed approach contained the modules of a main branch, a segmentation head, and a loss function, with regard to learning multi-level information and improving global consistency for the cross-scene segmentation task. Experimental results for cross-scene surgical cases indicated that the proposed approach outperformed state-of-the-art ones when using different segmentation indicators, in view of the better generalisation performance.

In relation to the proposed approach for the cross-scene semantic segmentation task, our further works may focus on two aspects as follows. First, domain adaptive segmentation for surgical instruments can be investigated when providing sufficient unlabelled intra-scene real-world data [67], on the basis of unsupervised domain adaptation. Furthermore, the works in this paper also show the possibility to perform surgical-instrument segmentation in few-shot cases with synthesised scenes for training sets.

REFERENCES

[1] Y. Sun, B. Pan, and Y. Fu, "Lightweight deep neural network for real-time instrument semantic segmentation in robot assisted minimally invasive surgery," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 3870–3877, Apr. 2021.

[2] T. Mahmood, S. W. Cho, and K. R. Park, "DSRD-Net: Dual-stream residual dense network for semantic segmentation of instruments in robot-assisted surgery," *Expert Syst. Appl.*, vol. 202, Sep. 2022, Art. no. 117420.

[3] Y. Kassahun et al., "Surgical robotics beyond enhanced dexterity instrumentation: A survey of machine learning techniques and their role in intelligent and autonomous surgical actions," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, pp. 553–568, Apr. 2016.

[4] M. K. Hasan, L. Calvet, N. Rabbani, and A. Bartoli, "Detection, segmentation, and 3-D pose estimation of surgical tools using convolutional neural networks and algebraic geometry," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101994.

[5] L. Yu, P. Wang, X. Yu, Y. Yan, and Y. Xia, "A holistically-nested U-Net: Surgical instrument segmentation based on convolutional neural network," *J. Dig. Imag.*, vol. 33, pp. 341–347, 2020.

[6] 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., (SMC)*, 2017, pp. 3373–3378.

[7] L. Yang, Y. Gu, G. Bian, and Y. Liu, "DRR-Net: A dense-connected residual recurrent convolutional network for surgical instrument segmentation from endoscopic images," *IEEE Trans. Med. Robot. Bion.*, vol. 4, no. 3, pp. 696–707, Aug. 2022.

[8] L. Yang, H. Wang, Y. Gu, G. Bian, Y. Liu, and H. Yu, "TMA-Net: A transformer-based multi-scale attention network for surgical instrument segmentation," *IEEE Trans. Med. Robot. Bion.*, vol. 5, no. 2, pp. 323–334, May 2023.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.

[10] M. Islam, D. A. Atputharuban, R. Ramesh, and H. Ren, "Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 2188–2195, Apr. 2019.

[11] Z. Ni, G. Bian, Z. Hou, X. Zhou, X. Xie, and Z. Li, "Attention-guided lightweight network for real-time segmentation of robotic surgical instruments," in *Proc. IEEE Int. Conf. Rob. Autom.*, 2020, pp. 9939–9945.

[12] X. Wang et al., "PaI-Net: A modified U-net of reducing semantic gap for surgical instrument segmentation," *IET Image Process.*, vol. 15, no. 12, pp. 2959–2969, 2021.

[13] M. Xue and L. Gu, "Surgical instrument segmentation method based on improved MobileNetV2 network," in *Proc. Int. Symp. Comput. Inf. Process. Technol., (ISCIPT)*, 2021, pp. 744–747.

[14] Z. Ni et al., "Pyramid attention aggregation network for semantic segmentation of surgical instruments," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11782–11790.

[15] Y. Dong, H. Wang, J. Luo, Z. Lai, F. Wang, and J. Wang, "Semantic segmentation of surgical instruments based on enhanced multi-scale receptive field," in *Proc. J. Phys. Conf. Ser.*, 2021, Art. no. 012006s.

[16] 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. IEEE Int. Conf. Robot. Autom.*, 2019, pp. 9821–9827.

[17] 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. Image Comput. Comput. Assist. Interv.*, 2019, pp. 440–448.

[18] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10012–10022.

[19] X. Sun, Y. Zou, S. Wang, H. Su, and B. Guan, "A parallel network utilizing local features and global representations for segmentation of surgical instruments," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 17, no. 10, pp. 1903–1913, 2022.

[20] S. Bodenstedt et al., "Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery," 2018, *arXiv:1805.02475*.

[21] M. Allan et al., "2017 robotic instrument segmentation challenge," 2019, *arXiv:1902.06426*.

[22] M. Allan et al., "2018 robotic scene segmentation challenge," 2020, *arXiv:2001.11190*.

[23] M. Pfeiffer et al., "Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2019, pp. 119–127.

[24] J. Cartucho, S. Tukra, Y. Li, D. S. Elson, and S. Giannarou, "VisionBlender: A tool to efficiently generate computer vision datasets for robotic surgery," *Comput. Methods Biomech. Biomed. Eng. Imag. Vis.*, vol. 9, no. 4, pp. 331–338, 2021.

[25] M. Kalia, T. A. Aleef, N. Navab, P. Black, and S. E. Salcudean, "Co-generation and segmentation for generalized surgical instrument segmentation on unlabelled data," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2021, pp. 403–412.

[26] J. Liu, X. Guo, and Y. Yuan, "Graph-based surgical instrument adaptive segmentation via domain-common knowledge," *IEEE Trans. Med. Imag.*, vol. 41, no. 3, pp. 715–726, Mar. 2022.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.

[28] Y. Wang, Z. Qiu, Y. Hu, H. Chen, F. Ye, and J. Liu, "Surgical instrument segmentation based on multi-scale and multi-level feature network," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, 2021, pp. 2672–2675.

[29] Z. Ni et al., "BARNet: Bilinear attention network with adaptive receptive fields for surgical instrument segmentation," 2020, *arXiv:2001.07093*.

[30] J. C. Á. Cerón, G. O. Ruiz, L. Chang, and S. Ali, "Real-time instance segmentation of surgical instruments using attention and multi-scale feature fusion," *Med. Image Anal.*, vol. 81, Oct. 2022, Art. no. 102569.

[31] L. Yang, Y. Gu, G. Bian, and Y. Liu, "An attention-guided network for surgical instrument segmentation from endoscopic images," *Comput. Biol. Med.*, vol. 151, Dec. 2022, Art. no. 106216.

[32] W. Shen et al., "Branch aggregation attention network for robotic surgical instrument segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 11, pp. 3408–3419, Nov. 2023.

[33] F. Qin, S. Lin, Y. Li, R. A. Bly, K. S. Moe, and B. Hannaford, "Towards better surgical instrument segmentation in endoscopic vision: Multi-angle feature aggregation and contour supervision," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 6639–6646, Oct. 2020.

[34] H. Ding, J. Zhang, P. Kazanzides, J. Y. Wu, and M. Unberath, "CaRTS: Causality-driven robot tool segmentation from vision and kinematics data," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2022, pp. 387–398.

[35] T. Ross et al., "Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101920.

[36] B. van Amsterdam, M. J. Clarkson, and D. Stoyanov, "Gesture recognition in robotic surgery: A review," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 6, pp. 2021–2035, Jun. 2021.

[37] C. Perrenot et al., "The virtual reality simulator DV-trainer® is a valid assessment tool for robotic surgical skills," *Surg. Endoscopy*, vol. 26, pp. 2587–2593, Apr. 2012.

[38] G. Whittaker et al., "Validation of the RobotiX mentor robotic surgery simulator," *J. Endourol.*, vol. 30, pp. 338–346, Mar. 2016.

[39] A. Munawar, N. Srishankar, and G. S. Fischer, "An open-source framework for rapid development of interactive soft-body simulations for real-time training," in *Proc. IEEE Int. Conf. Rob. Autom.*, 2020, pp. 6544–6550.

[40] E. Colleoni, D. Psychogyios, B. Van Amsterdam, F. Vasconcelos, and D. Stoyanov, "SSIS-Seg: Simulation-supervised image synthesis for surgical instrument segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3074–3086, Nov. 2022.

[41] J. Li, S. Zha, C. Chen, M. Ding, T. Zhang, and H. Yu, "Attention guided global enhancement and local refinement network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 3211–3223, 2022.

[42] S.-A. Liu, H. Xie, H. Xu, Y. Zhang, and Q. Tian, "Partial class activation attention for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16836–16845.

[43] Y. Guo, J. Chen, Q. Du, A. V. D. Hengel, Q. Shi, and M. Tan, "The shallow end: Empowering shallower deep-convolutional networks through auxiliary outputs," 2016, *arXiv:1611.01773*.

[44] R. Feng et al., "SSN: A stair-shape network for real-time polyp segmentation in colonoscopy images," in *Proc. IEEE Conf. Symp. Biomed. Imag.*, 2020, pp. 225–229.

[45] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 173–190.

[46] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 433–442.

[47] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2018, pp. 7132–7141.

[48] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[49] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[50] B. Huang et al., "Simultaneous depth estimation and surgical tool segmentation in laparoscopic images," *IEEE Trans. Med. Robot. Bion.*, vol. 4, no. 2, pp. 335–338, May 2022.

[51] E. Colleoni and D. Stoyanov, "Robotic instrument segmentation with image-to-image translation," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 935–942, Apr. 2021.

[52] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BasNet: Boundary-aware salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7479–7489.

[53] S. Yu, B. Zhang, J. Xiao, and E. G. Lim, "Structure-consistent weakly supervised salient object detection with local saliency coherence," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3234–3242.

[54] H. Huang et al., "MTL-ABS3Net: Atlas-based semi-supervised organ segmentation network with multi-task learning for medical images," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 3988–3998, Aug. 2022.

[55] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[56] Z. Ni et al., "RAUNet: Residual attention U-net for semantic segmentation of cataract surgical instruments," in *Proc. Int. Conf. Neural Inform.*, 2019, pp. 139–149.

[57] Z. Ni, G.-B. Bian, X. Xie, Z. Hou, X. Zhou, and Y. Zhou, "RASNet: Segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, 2019, pp. 5735–5738.

[58] Z. Jin, B. Liu, Q. Chu, and N. Yu, "ISNet: Integrate image-level and semantic-level context for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7189–7198.

[59] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[60] C. Du, J. Li, B. Zhang, W. Feng, T. Zhang, and D. Li, "Intraoperative navigation system with a multi-modality fusion of 3-D virtual model and laparoscopic real-time images in laparoscopic pancreatic surgery: A preclinical study," *BMC Surg.*, vol. 22, no. 1, p. 139, 2022.

[61] Y. Hao et al., "EISeg: An efficient interactive segmentation annotation tool based on PaddlePaddle," 2022, *arXiv:2210.08788*.

[62] L. 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[63] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, no. 3, pp. 400–407, 1951.

[64] S. Nema and L. Vachhani, "Unpaired deep adversarial learning for multi-class segmentation of instruments in robot-assisted surgical videos," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 19, no. 4, p. e2514, 2023.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[66] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.

[67] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3764–3773.