

LC-GAN: Image-to-image Translation Based on Generative Adversarial Network for Endoscopic Images

Shan Lin, Fangbo Qin, Yangming Li, Randall A. Bly, Kris S. Moe, Blake Hannaford*, *Fellow, IEEE*

Abstract—The intelligent perception of endoscopic vision is appealing in many computer-assisted and robotic surgeries. Achieving good vision-based analysis with deep learning techniques requires large labeled datasets, but manual data labeling is expensive and time-consuming in medical problems. When applying a trained model to a different but relevant dataset, a new labeled dataset may be required for training to avoid performance degradation. In this work, we investigate a novel cross-domain strategy to reduce the need for manual data labeling by proposing an image-to-image translation model called live-cadaver GAN (LC-GAN) based on generative adversarial networks (GANs). More specifically, we consider a situation when a labeled cadaveric surgery dataset is available while the task is instrument segmentation on a live surgery dataset. We train LC-GAN to learn the mappings between the cadaveric and live datasets. To achieve instrument segmentation on live images, we can first translate the live images to fake-cadaveric images with LC-GAN, and then perform segmentation on the fake-cadaveric images with models trained on the real cadaveric dataset. With this cross-domain strategy, we fully leverage the labeled cadaveric dataset for segmentation on live images without the need to label the live dataset again. Two generators with different architectures are designed for LC-GAN to make use of the deep feature representation learned from the cadaveric image based instrument segmentation task. Moreover, we propose structural similarity loss and segmentation consistency loss to improve the semantic consistency during translation. The results demonstrate that LC-GAN achieves better image-to-image translation results, and leads to improved segmentation performance in the proposed cross-domain segmentation task.

I. INTRODUCTION

Minimally invasive surgery (MIS) such as endoscopic surgery brings many benefits to patients, but also introduces many challenges to the surgeons. In endoscopy, a tiny camera is inserted into the human body with surgical instruments to provide a real-time view of the surgical site. The endoscopic videos have limited field-of-view (FoV) and reduced depth perception, and this indirect observing method impacts the surgeons' eye-hand coordination ability [1], [2]. Additionally, MIS provides a limited sense of touch so the information a surgeon can get from the surgical site has been further reduced [3], [4].

The challenges of MIS may lead to accidental damage to important structures and suboptimal surgical results [5], [6]. To improve surgical performance, robotic-assisted surgery systems have attracted increasing attention [7]. For example,

Shan Lin, Randall A. Bly, Kris S. Moe and Blake Hannaford are with University of Washington, Seattle, WA 98195, USA. shan13@uw.edu

Fangbo Qin is with Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China qinfangbo2013@ia.ac.cn

Yangming Li is with Rochester Institute of Technology, Rochester, NY 14623, USA Yangming.Li@rit.edu

robots that handle the endoscope or laparoscope have been explored [8], [9], the applications of soft robotic devices in MIS have been studied to reduce patients' pain and damage [5]. Moreover, overlaying pre- and intra-operative imaging with surgical videos could improve surgeons' capabilities [10]. For all these applications, instrument segmentation is a critical component and one of the main challenges [2].

In recent years, deep learning has achieved state-of-the-art performance on instrument segmentation over traditional approaches based on hand-craft features. However, deep learning-based methods rely on large-scale labeled datasets, of which the availability is usually limited for medical applications [11], [2]. When a different but related labeled dataset is available, we explore an alternative and less expensive approach that uses domain adaptation to transfer the knowledge learned by a particular model on a source domain to a related target domain [12]. Recently, domain adaptation is usually implemented by searching for a common feature space of both domains but this strategy suffers from semantic inconsistency, i.e. one type of object may be mapped to another type of object [12].

In this work, we propose an image-to-image translation model LC-GAN that learns the mapping between the cadaveric and live image domains. Our model is developed based on CycleGAN [13] by Zhu *et al.*, which is a model for learning the bijective translations between two image domains using unpaired training data. First, we improve the feature representation capability of one generator by using ResNet-50 [14] pre-trained on the cadaveric domain. Secondly, we propose the structural similarity loss to reduce the image structural changes during translation, and propose the segmentation consistency loss to encourage the segmentations of the real image and the generated fake image to be similar. These two loss functions together improve the semantic consistency of LC-GAN. We then use the deep segmentation models trained on the cadaveric domain to implement segmentation on the fake-cadaveric surgery images translated from the live domain. We demonstrate our method on a sinus surgery dataset with 10 cadaveric surgery videos as the source domain and 3 live surgery videos as the target domain.

To the best of our knowledge, our work is the first to perform image-to-image translation between live and cadaveric surgery images. In addition to the proposed cross-domain strategy, LC-GAN could be used in data augmentation or to generate a larger dataset by combining two smaller ones. Also, our work could bring benefits to the surgery training processes. For example, we could augment the cadaveric

surgery videos in real-time to live surgery-like videos as the surgical trainees gain experience on the cadaver.

II. RELATED WORKS

Generative adversarial networks (GANs) proposed by Goodfellow *et al.* [15] have achieved impressive performance in many vision problems such as image generation, feature learning and image super-resolution [16], [17], [18]. GANs train two networks, a generator and a discriminator, which contest with each other. The discriminator is trained to distinguish the generated data from the true data, while the generator is trained to fool the discriminator.

Image-to-Image Translation Image-to-image translation is a vision problem that aims to learn the mapping between two different but relevant image domains [19]. GANs have been successfully applied in various image-to-image translation tasks with paired or unpaired datasets [20], [13], [19].

In medical applications, the availability of certain image modalities may be limited because data collection is usually expensive and time-consuming [11], [2]. Image-to-image translation provides an effective way to estimate the appearance of the desired image modality from a relevant image and has brought impressive results in medical image analysis problems [21], [22], [23], [24]. Wang *et al.* applied conditional generative adversarial network (cGAN) to reduce the metal artifacts from computed tomography (CT) and achieved better segmentation results on the generated CTs than previous segmentation methods [21]. Mahmood *et al.* used GAN to transform true medical images to synthetic-like images and estimated the depth map using a neural network trained on synthetic images [22].

The unreliability of the generated images due to semantic inconsistency in translation is one limitation and concern of the applications of image-to-image translation in medicine. Recent studies worked on introducing constraints to guarantee semantic consistency. In CycleGAN [13], a cycle consistency loss was introduced to ensure the translation is invertible. Liu *et al.* included an assumption that the two image domains share a latent space based on CycleGAN [19]. Cherian *et al.* added two segmentation functions alongside the CycleGAN model to include the semantic consistency constraints in the training paradigm [25].

Surgical Instrument Segmentation Vision-based surgical instrument segmentation has been widely studied. Recently, deep learning-based approaches have achieved state-of-the-art segmentation performance with larger datasets [26], [27], [28], [29], [30], [31], [32], [33], [34]. Shvets *et al.* improved two U-Net [27] family models Ternausnet and LinkNet and achieved top performances in the MICCAI 2017 Endoscopic Vision Sub-Challenge: Robotic Instrument Segmentation [32], [34]. Chen *et al.* proposed an advanced segmentation model DeepLabV3+ by taking advantage of the pyramid pooling module [35] and atrous convolution [31]. Despite these efforts, existing algorithms may not be robust enough under challenging conditions in live surgeries such as strong specular reflection and blood. Additionally, the trained models may not generalize well to different surgical

interventions. To achieve good performance, labeling a new and large dataset is needed, which requires significant expert time and can be very expensive.

III. METHODS

In this work, we explore using image-to-image translation methods to reduce the need for manual data labeling for surgical instrument segmentation. We consider a situation that there exists a labeled dataset and one wants to perform segmentation on another related but different unlabeled dataset. The two datasets used here are cadaveric surgery images and live surgery images. Although we have manually labeled both datasets, we assume the labels of the cadaveric dataset are available while the labels of the live dataset are only used to evaluate the proposed method. Section IV provides more details on the datasets.

The overall framework is shown in Fig. 1. The target task is instrument segmentation on the live surgery images. As shown in Fig. 1(b), the current mainstream strategy is to train and test segmentation models directly on a labeled live dataset, but this strategy requires us to also label the live dataset for good performance. In contrast, we propose a cross-domain strategy that does not use the labels of live dataset by taking advantage of the already existed labeled cadaveric dataset as shown in Fig. 1(b). We propose a model LC-GAN (see Fig. 1(a)) developed based on CycleGAN [13] to learn the mapping between the cadaveric dataset and live dataset. We then perform instrument segmentation on the fake-cadaveric surgery images generated from the real-live surgery images. The segmentation is implemented using deep convolutional neural network (CNN) models trained with the labeled cadaveric dataset. Finally, we evaluate the potential of the cross-domain strategy by comparing its segmentation results with the current mainstream strategy.

A. Live-cadaver GAN (LC-GAN) Architecture

Similar to CycleGAN [13], LC-GAN consists of two generators and two discriminators. Given two image domains X and Y , the two generators G and F are trained to learn the mappings between the image domains, i.e. $G : X \rightarrow Y$ and $F : Y \rightarrow X$. D_X and D_Y are two discriminators trained to distinguish between real and fake images. We define the cadaveric dataset as the X domain and the live dataset as the Y domain.

For images with many challenging conditions as in our dataset, training the generators from scratch is difficult and may require longer training time to converge. Considering that we already have deep segmentation models trained on the cadaveric dataset, we could use the backbone feature extractors in these models to provide better feature maps for the generator. We construct two generators with different architectures as shown in Fig. 2. For the generator G that maps cadaveric surgery images to live surgery images, we use the network backbone Resnet-50 [14] of a DeepLabV3+ segmentation model [31] trained with the labeled cadaveric dataset as the feature extractor. We also adapt part of the DeepLabV3+ model into G . We use the Atrous Spatial

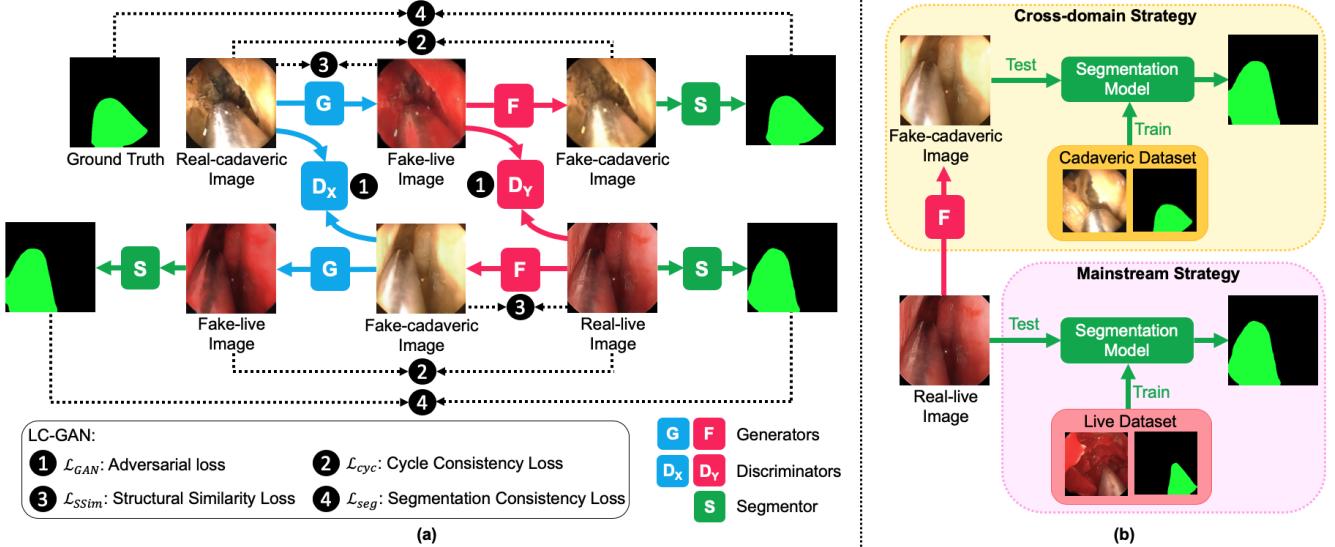


Fig. 1. Overall framework. (a) Schematic of LC-GAN. The generator G performs cadaver-to-live translation, while generator F performs live-to-cadaver translation. The discriminators D_X and D_Y are used to distinguish the fake images from the real images in cadaveric and live domains, respectively. The segmentor S is a deep segmentation model trained on the real cadaveric dataset. (b) Schematic of the proposed cross-domain strategy and the mainstream strategy for instrument segmentation on the live dataset. The predicted or ground truth instrument regions in the segmentation masks are shown in green.

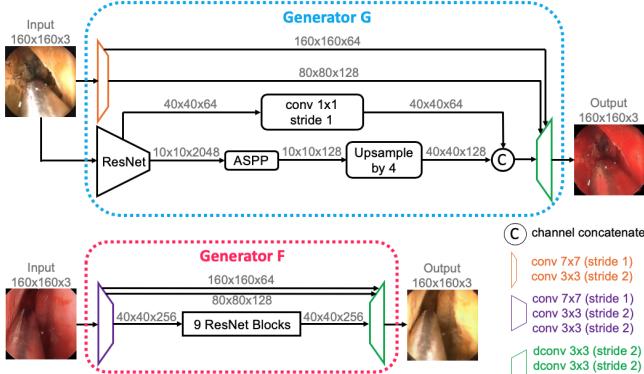


Fig. 2. Generator architectures of LC-GAN (example for 160x160 input image). Each trapezoid represents a series of convolution or deconvolution operations. The sizes (width-height-channel) of the feature maps are shown on top of the corresponding arrows.

Pyramid Pooling (ASPP) module [31] that extracts multi-scale information from the output of Resnet-50, and concatenate this information with low-level features from Resnet-50. The parameters of Resnet-50 are fixed during training. For another generator F , no trained feature extractor is available so we train it from scratch. We choose a ResNet with two stride-2 convolutions, nine residual blocks and two fractionally-strided convolutions with stride $\frac{1}{2}$ as the generator F [36], [13]. For the discriminator, we use the 70×70 PatchGAN [20], [13] to determine the real 70×70 image patches from fake patches.

B. Loss Functions for LC-GAN

CycleGAN was proposed with the adversarial loss \mathcal{L}_{GAN} and the cycle consistency loss \mathcal{L}_{cyc} . The details of \mathcal{L}_{GAN} and \mathcal{L}_{cyc} are provided in [13]. Although CycleGAN shows compelling results in many datasets, the semantic consistency might not be guaranteed for some complex scenes. Fig. 3(a) shows an example of semantic inconsistency between

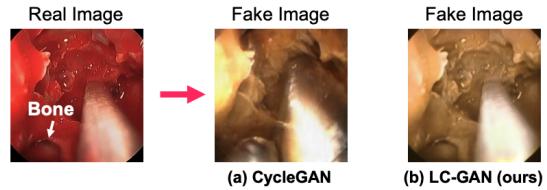


Fig. 3. Translation from a real-live surgery image to a fake-cadaveric surgery image. (a) The result of CycleGAN is an example of semantic inconsistency. The bone region pointed by the white arrow is translated into an instrument and the instrument becomes much larger in the fake image than its true size. (b) The proposed LC-GAN generates a fake image with better semantic consistency.

the real-live surgery image and the fake-cadaveric surgery image generated using CycleGAN. In Fig. 3(a), the bone is translated to an instrument and the true instrument becomes much larger in the fake image. To ensure the semantic consistency, we propose structural similarity loss \mathcal{L}_{SSim} and segmentation consistency loss \mathcal{L}_{seg} alongside with the cycle consistency loss as shown in Fig. 1(a).

\mathcal{L}_{SSim} estimates the structural similarity between the input and output of a generator. Because the live and cadaveric surgery images have very different colors, the color information should be excluded when calculating \mathcal{L}_{SSim} . Therefore, we convert the image from RGB to YUV color space and use the Y channel that consists of luminance information [37] to estimate the structural similarity. Besides color features, the live surgery images have different lighting conditions and the presence of fluids including blood compared to the cadaveric surgery images. Therefore, \mathcal{L}_{SSim} should focus on overall structural similarity while allowing differences in image brightness, contrast and details. Inspired by [38], we propose a multi-scale method that compares the image structural information at different resolutions. We iteratively downscale the image by a factor of 2 and get a total of n_s scaled images which are $1, \frac{1}{2}, \dots, \frac{1}{2^{n_s}}$ of the original size,

respectively. The expression of \mathcal{L}_{SSim} is

$$\begin{aligned}\mathcal{L}_{SSim}(G, F) = & [1 - \sum_{i=0}^{n_s} \gamma_i C(x_{Y,i}, G(x)_{Y,i})] \\ & + [1 - \sum_{i=0}^{n_s} \gamma_i C(y_{Y,i}, F(y)_{Y,i})]\end{aligned}\quad (1)$$

where $x \in X$ is an image from the X domain and $y \in Y$ is an image from the Y domain, the subscripts Y, i of an image denote that we extract the Y channel of the image and scale it to $\frac{1}{2^i}$ of its original size, γ_i are the multi-scale weights and are normalized to $\sum_{i=0}^N \gamma_i = 1$. $C(a, b)$ is the zero-normalized cross-correlation (ZNCC) [39], [40] between images a and b

$$C(a, b) = \frac{\sigma_{ab} + \epsilon}{\sigma_a \sigma_b + \epsilon} \quad (2)$$

where σ_{ab} is the covariance between a and b , and is defined as [40]

$$\sigma_{ab} = \frac{1}{m-1} \sum_{i=1}^m (a_i - \mu_a)(b_i - \mu_b) \quad (3)$$

m is the number of pixels in a and b , a_i and b_i are the i th pixel of a and b . μ_a , μ_b and σ_a , σ_b correspond to the mean intensity and standard deviations of a and b . ϵ is a constant to stabilize the division when $\sigma_a \sigma_b$ is close to zero. ZNCC is less sensitive to the differences of illumination conditions and contrast of the two compared images.

To further improve semantic consistency, we introduce a segmentation consistency loss

$$\begin{aligned}\mathcal{L}_{seg}(G, F) = & \mathcal{L}_{CE}(x_m, S(F(G(x)))) \\ & + \mathcal{L}_{CE}(S(y), S(G(F(y))))\end{aligned}\quad (4)$$

where x_m is the ground truth segmentation mask of image x , S is a segmentation model trained on the labeled cadaveric dataset, $\mathcal{L}_{CE}(a, b)$ is the naive cross-entropy loss. For the segmentation consistency loss between y and $G(F(y))$, we use the segmentation on the image y as the target mask because the labels of the live dataset are not available according to our assumption.

Finally, the overall objective function is given by

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = & \mathcal{L}_{GAN}(G, D_Y) + \mathcal{L}_{GAN}(F, D_X) \\ & + \lambda_1 \mathcal{L}_{cyc}(G, F) + \lambda_2 \mathcal{L}_{SSim}(G, F) \\ & + \lambda_3 \mathcal{L}_{seg}(G, F)\end{aligned}\quad (5)$$

where λ_i are hyper-parameters that balance the impact of the losses. The generators are trained to minimize the overall objective function and the discriminators are trained to maximize it.

C. Instrument Segmentation Methods

Instrument segmentation is implemented using four deep CNN models include DeepLabV3+ [31], TernausNet [30], MFF+SPP [41] and LWANet [42], and a Multi-angle Feature Aggregation (MAFA) strategy [43]. MAFA was proposed to provide richer visual cues for segmentation and can be

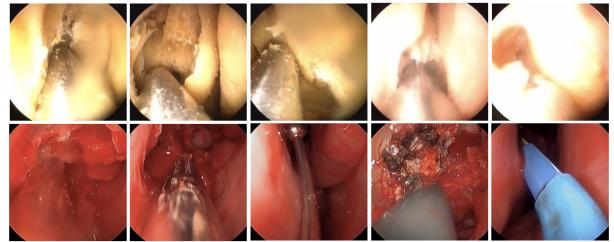


Fig. 4. Examples of center-cropped video frames in our dataset. The top row is from the cadaveric surgery videos and the bottom row is from the live surgery videos.

flexibly combined with these four deep CNN models. In MAFA, each input image is first rotated by several different angles and the rotated images are sent to a shared feature extraction module. The feature maps extracted from the rotated images are then rotated back by the same angle and aggregated together.

IV. DATASETS

Our sinus surgery dataset consists of 10 cadaveric surgery videos ranging from 5 minutes to 23 minutes with a resolution of 320×240 and 3 live surgery videos ranged from 12 minutes to 66 minutes with a resolution of 1920×1080 . More details of data collection are provided in the author's previous publication [44]. Fig. 4 shows some examples of video frames. The challenging conditions of this dataset include specular reflections, blur from motion, blood, smoke, instruments in shadow and occlusions by tissues.

The video frames were sampled from the cadaveric surgery videos at 0.5 Hz resulting in a total of 4413 frames. For the live surgery videos, due to blood, blur from motion and specular reflection, the instrument regions in some video frames cannot be identified without temporal information from neighboring frames. The temporal information has not been considered in both the image-to-image translation and segmentation methods in this paper, so such images were excluded. We captured frames from the live surgery videos at 1 Hz and removed the aforementioned challenging images resulting in a total of 4660 frames.

Endoscopic images have a large black border region without useful information. To speed up the training process for LC-GAN, we cropped image patches within the endoscopic area and downsampled these patches to 160×160 . For training, we selected frames with less specular reflections and blurriness from both cadaveric and live datasets. 901 cadaveric surgery frames were selected for the X domain and 3174 live surgery frames were selected for the Y domain.

For surgical instrument segmentation, we center-cropped the frames and downsampled them to 240×240 . The cadaveric surgery images were separated into a training set of 2917 frames from 7 videos and a test set of 1496 frames from the remaining 3 videos. The live surgery images were separated into a training set of 3506 frames from 2 videos and a test set of 1154 frames from the remaining 1 video. The instrument contours in all frames were manually labeled.

The dataset will be available at <https://github.com/SURA23/Sinus-Surgery-Endoscopic-Image-Datasets>.

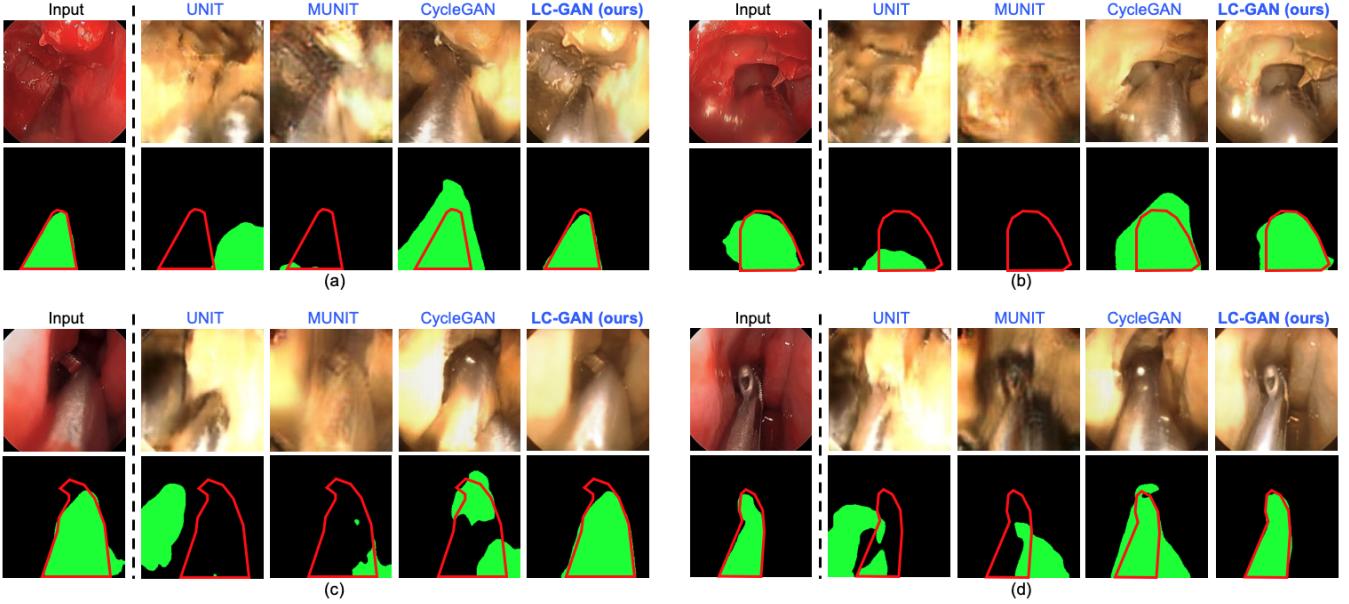


Fig. 5. Examples of results from the cross-domain strategy and mainstream strategy. In each subfigure, the last four columns of the top row show the fake-cadaveric images translated from the input real-live image using UNIT [19], MUNIT [45], CycleGAN [13] and LC-GAN (ours). The second row shows the corresponding instrument segmentation results obtained using DeepLabV3+ [31] with MAFA [43]. In the bottom row, the first segmentation is the result of the mainstream strategy and the last four segmentations are from the cross-domain strategy. The predicted instrument regions are shown in green and the ground truth of the instrument contours are shown as red lines.

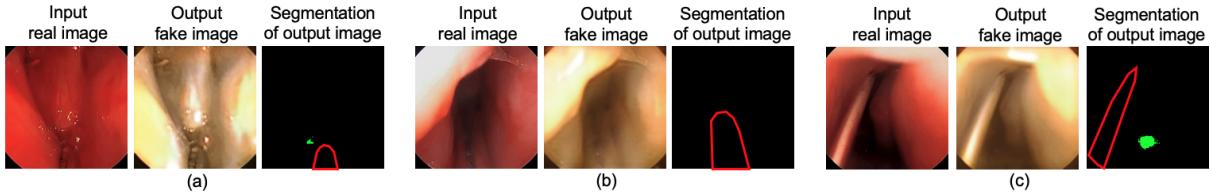


Fig. 6. Failed live-to-cadaver translation examples given by LC-GAN. The corresponding instrument segmentations are obtained using DeepLabV3+ [31] with MAFA [43]. The predicted instrument regions are shown in green and the ground truth of the instrument contours are shown as red lines.

V. IMPLEMENTATION DETAILS

We implemented image-to-image translation models using Tensorflow on a single Nvidia Tesla T4 GPU. The segmentation models were implemented on a 3.70GHz Intel i7-8700K CPU and two Nvidia GTX2080ti GPUs.

LC-GAN We trained the network with 6 epochs through Adam optimizer using an initial learning rate of 0.0001 for the first half of training, and then linearly decaying to zero over the remaining epochs. Each batch consisted of one image from domain X and one image from domain Y. The hyper-parameters λ_1 , λ_2 and λ_3 were set to 10, 0.5 and 1, respectively. For the structural similarity loss \mathcal{L}_{SSim} , we chose $n_s = 4$ and empirically set $\gamma_1 = 0.05$, $\gamma_2 = 0.33$, $\gamma_3 = 0.35$ and $\gamma_4 = 0.27$.

Compared image-to-image translation models We compared LC-GAN with three image-to-image translation models CycleGAN [13], UNIT [19] and MUNIT [45]. We first trained all three models until 250 epochs and chose the minimum epochs that made the models converge. We found CycleGAN stabilized at 200 epochs. UNIT and MUNIT failed to converge within 250 epochs. We then chose 9 epochs for these two models because after 9 epochs they became more unstable.

VI. EXPERIMENTS AND RESULTS

We compared our proposed method LC-GAN against the baseline and state-of-the-art models include CycleGAN [13], UNIT [19] and MUNIT [45].

Qualitative evaluation Fig. 5 shows examples of the image-to-image translation results. We found that LC-GAN generally retained the shape and location of the instruments and the structures of sinus tissue in the resultant fake images. A majority of images in our live dataset have the instruments merged into the background due to strong specular reflection and blood. LC-GAN could successfully handle such cases most of the time as shown in Fig. 5(a,b), while CycleGAN tended to increase the size of instruments. Also, CycleGAN was less robust and might translate part of the background to an instrument as shown in Fig. 5(c). The other two comparison methods, UNIT and MUNIT, did not converge to the correct correspondence between the two image domains.

Although our image-to-image translation results are promising, the performance is still not satisfactory under challenging conditions. Fig. 6 shows some typical fail cases. Fig. 6(a) shows an example of an instrument merged in blood. Fig. 6(b) shows an example of an instrument in shadow with a red appearance due to specular reflection.

TABLE I
SEGMENTATION PERFORMANCES ON LIVE SURGERY DATASET

Image-to-image Translation Model	Segmentation Train Set	Segmentation Test Set	Segmentation Performance (mDice(%) / mIoU(%))			
			DeepLabV3+ [31]	TernausNet-16 [30]	MFF+SPP [41]	LWANet [42]
UNIT [19]	Cadaver	Live	32.9/24.8	34.2/25.4	33.9/25.0	35.6/26.2
MUNIT [45]			25.2/20.1	19.5/14.1	19.5/14.2	28.2/20.1
CycleGAN [13]			63.3/51.9	60.5/49.1	60.9/49.3	60.0/47.9
LC-GAN (ours)			78.7/72.3	75.4/68.0	72.8/64.4	71.4/62.0
N/A	Live	Live	82.7/75.5	82.4/75.7	81.2/73.0	82.1/74.6

* All segmentation methods were implemented with MAFA [43]. The first four rows show the results of the cross-domain strategy, and the last row shows the results of the mainstream strategy. The bold font indicates the best performance of the cross-domain strategy in each column.

The instruments in Fig. 6(a,b) are translated to regions similar to the surrounding background and lead to incorrect segmentations. Fig. 6(c) represents another type of failure that happens when the instrument appears red and does not exist in the cadaveric dataset. In such a case, the color information is not reliable and the segmentation models should use shape information for successful segmentation. The segmentation models trained with the cadaveric dataset have not seen this or similar instrument before, resulting in segmentation failure.

Quantitative evaluation We used the instrument segmentation performances on the fake-cadaveric surgery images translated from the real-live surgery images for quantitative evaluations. The segmentation models were trained on the cadaveric dataset. We applied four segmentation models (with MAFA [43]) include DeepLabV3+ [31], TernausNet [30], MFF+SPP [41] and LWANet [42]. As a comparison, we also performed the mainstream strategy that trains and tests these segmentation models directly with the labeled live surgery dataset. The implementation details of the segmentation models are provided in [43]. We used the Dice similarity coefficient (DSC) and Intersection over Union (IoU) to evaluate the instrument segmentation performance [46]

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}, IoU = \frac{|X \cap Y|}{|X \cup Y|}$$

where X and Y are the predicted and ground truth instrument segmentations, respectively.

Table I shows the segmentation results. We found that segmentation on fake-cadaveric surgery images translated by LC-GAN were better than the results obtained using other compared image-to-image translation models. Our method achieved 11%~15% better mDice and 14%~20% better mIoU than using CycleGAN. Compared with the mainstream strategy that uses the labeled live dataset, our method achieved 4%~11% lower mDice and 3%~13% lower mIoU.

VII. DISCUSSION

We propose an image-to-image translation model LC-GAN that achieves better semantic consistency using constraints that encourage structural similarity. Our method reduces the need to manually label new large datasets for a cross-domain segmentation task by translating a labeled dataset to a relevant but different domain. The training of the proposed image-to-image translation model only requires

an unpaired dataset, which can be easily extracted from the surgery videos.

Fig. 5 shows that LC-GAN surpasses other comparison methods to provide the best image-to-image translation results. In contrast, CycleGAN tends to increase instrument sizes in the fake-cadaveric surgery images. This can be explained by the fact that in our dataset the instruments in the cadaveric domain are generally larger than the instruments in the live domain. UNIT and MUNIT fail to capture the correct mapping between the cadaveric and live surgery images. UNIT and MUNIT are built based on a shared-latent space assumption, i.e. each pair of corresponding images from the two domains can be mapped to a shared-latent space. However, this assumption may be too strict for our dataset because the scenes in the two domains have many differences in both the instrument types and backgrounds.

Compared with the mainstream strategy, the cross-domain strategy achieved the lower performance. This result is as expected and the current gaps are acceptable especially when DeepLabV3+ [31] is used for segmentation. The live and cadaveric datasets have different types of instruments and lightning conditions, so the distribution of the fake-cadaveric images is still different from the real-cadaveric images. Also, although we propose two loss functions to improve semantic consistency between the real images and their corresponding fake images, the semantic consistency has not been fully guaranteed. To mitigate the gap, we plan to use a small set of fake-cadaveric images with labels to fine-tune the deep segmentation models. Moreover, the proposed live surgery dataset introduces multiple challenges. When we labeled the dataset, we found that it is much easier to decide the instrument locations by referring to neighboring frames. This points out a future work direction, by estimating the motion flow between the neighboring frames in the video, we can potentially obtain clues of semantic consistency for image-to-image translation.

VIII. CONCLUSIONS

In this work, we propose an image-to-image translation model LC-GAN to learn the mapping between two different but relevant image domains. We introduce structural similarity loss and segmentation consistency loss for LC-GAN to improve the semantic consistency during translation. We demonstrate the proposed model in a sinus surgery dataset of cadaveric and live surgery images. Our results show that the proposed method can potentially reduce the need to label

more data for surgical instrument segmentation. These results have major implications on the ability to automatically segment and track surgical instruments, leading to improved analysis of surgery as well as enhanced in surgical training.

REFERENCES

- [1] R. Bogdanova, P. Boulanger, and B. Zheng, "Depth perception of surgeons in minimally invasive surgery," *Surgical Innovation*, vol. 23, no. 5, pp. 515–524, 2016.
- [2] D. Bouget, M. Allan, D. Stoyanov, *et al.*, "Vision-based and markerless surgical tool detection and tracking: a review of the literature," *Medical Image Analysis*, vol. 35, pp. 633–654, 2017.
- [3] Y.-H. Su, K. Huang, and B. Hannaford, "Multicamera 3d reconstruction of dynamic surgical cavities: Camera grouping and pair sequencing," in *Proc. Int. Symp. Med. Robot.*, pp. 1–7, 2019.
- [4] J. Rosen, B. Hannaford, M. P. MacFarlane, *et al.*, "Force controlled and teleoperated endoscopic grasper for minimally invasive surgery-experimental performance evaluation," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 10, pp. 1212–1221, 1999.
- [5] M. Runciman, A. Darzi, and G. P. Mylonas, "Soft robotics in minimally invasive surgery," *Soft Robotics*, vol. 6, no. 4, pp. 423–443, 2019.
- [6] A. Khanna and A. Sama, "Managing complications and revisions in sinus surgery," *Current Otorhinolaryngology Reports*, vol. 7, no. 1, pp. 79–86, 2019.
- [7] D. T. Friedrich, M. O. Scheithauer, J. Greve, *et al.*, "Recent advances in robot-assisted head and neck surgery," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 13, no. 2, p. e1744, 2017.
- [8] M. Sahu, D. Moerman, P. Mewes, *et al.*, "Instrument state recognition and tracking for effective control of robotized laparoscopic systems," *Int. J. Mech. Eng. Robot. Res.*, vol. 5, no. 1, p. 33, 2016.
- [9] M. C. Capolei, H. Wu, N. A. Andersen, *et al.*, "Positioning the laparoscopic camera with industrial robot arm," in *Proc. Int. Conf. Control Automat. Robot.*, pp. 138–143, 2017.
- [10] M. Allan, S. Kondo, S. Bodenstedt, *et al.*, "2018 robotic scene segmentation challenge," *arXiv preprint arXiv:2001.11190*, 2020.
- [11] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.
- [12] J. Hoffman, E. Tzeng, T. Park, *et al.*, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017.
- [13] J.-Y. Zhu, T. Park, P. Isola, *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2223–2232, 2017.
- [14] K. He, X. Zhang, S. Ren, *et al.*, "Identity mappings in deep residual networks," in *Proc. Europ. Conf. Comput. Vis.*, pp. 630–645, 2016.
- [15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.
- [16] T. Xu, P. Zhang, Q. Huang, *et al.*, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1316–1324, 2018.
- [17] F. Qin, Y. Li, Y.-H. Su, *et al.*, "Surgical instrument segmentation for endoscopic vision with data fusion of CNN prediction and kinematic pose," in *Proc. Int. Conf. Robot. Autom.*, pp. 9821–9827, 2019.
- [18] C. Ledig, L. Theis, F. Huszár, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4681–4690, 2017.
- [19] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, pp. 700–708, 2017.
- [20] P. Isola, J.-Y. Zhu, T. Zhou, *et al.*, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.
- [21] J. Wang, Y. Zhao, J. H. Noble, *et al.*, "Conditional generative adversarial networks for metal artifact reduction in CT images of the ear," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pp. 3–11, 2018.
- [22] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2572–2581, 2018.
- [23] Y. Li and L. Shen, "cC-GAN: A robust transfer-learning framework for HEp-2 specimen image segmentation," *IEEE Access*, vol. 6, pp. 14 048–14 058, 2018.
- [24] I. Funke, S. Bodenstedt, C. Riediger, *et al.*, "Generative adversarial networks for specular highlight removal in endoscopic images," in *Medical Imaging: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 10576, p. 1057604, 2018.
- [25] A. Cherian and A. Sullivan, "Sem-GAN: Semantically-consistent image-to-image translation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, pp. 1797–1806, 2019.
- [26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3431–3440, 2015.
- [27] 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.*, pp. 234–241, 2015.
- [28] L. C. García-Peraza-Herrera, W. Li, L. Fidon, *et al.*, "Toolnet: holistically-nested real-time segmentation of robotic surgical tools," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, pp. 5717–5722, 2017.
- [29] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Visual Commun. Image Process.*, pp. 1–4, 2017.
- [30] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv preprint arXiv:1801.05746*, 2018.
- [31] L.-C. Chen, Y. Zhu, G. Papandreou, *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Europ. Conf. Comput. Vis.*, pp. 801–818, 2018.
- [32] A. A. Shvets, A. Rakhlis, A. A. Kalinin, *et al.*, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, pp. 624–628, 2018.
- [33] M. Islam, Y. Li, and H. Ren, "Learning where to look while tracking instruments in robot-assisted surgery," *arXiv preprint arXiv:1907.00214*, 2019.
- [34] M. Allan, A. Shvets, T. Kurmann, *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.
- [35] H. Zhao, J. Shi, X. Qi, *et al.*, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2881–2890, 2017.
- [36] K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.
- [37] A. Ford and A. Roberts, "Colour space conversions," *Westminster University, London*, vol. 1998, pp. 1–31, 1998.
- [38] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conf. Signals Syst. Comput.*, vol. 2, pp. 1398–1402, 2003.
- [39] A. Giachetti, "Matching techniques to compute image motion," *Image and Vision Computing*, vol. 18, no. 3, pp. 247–260, 2000.
- [40] Z. Wang, A. C. Bovik, H. R. Sheikh, *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
- [41] M. Islam, D. A. Atputharuban, R. Ramesh, *et al.*, "Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 2188–2195, 2019.
- [42] Z.-L. Ni, G.-B. Bian, Z.-G. Hou, *et al.*, "Attention-guided lightweight network for real-time segmentation of robotic surgical instruments," *arXiv preprint arXiv:1910.11109*, 2019.
- [43] F. Qin, S. Lin, Y. Li, *et al.*, "Towards better surgical instrument segmentation in endoscopic vision: multi-angle feature aggregation and contour supervision," *arXiv preprint arXiv:2002.10675*, 2020.
- [44] S. Lin, F. Qin, R. A. Bly, *et al.*, "Automatic sinus surgery skill assessment based on instrument segmentation and tracking in endoscopic video," in *Proc. Int. Workshop Multiscale Multimodal Med. Imag.*, pp. 93–100, 2019.
- [45] X. Huang, M.-Y. Liu, S. Belongie, *et al.*, "Multimodal unsupervised image-to-image translation," in *Proc. Europ. Conf. Comput. Vis.*, pp. 172–189, 2018.
- [46] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, 2015.