# Surgical Tool Segmentation using Generative Adversarial Networks with Unpaired Training Data

Zhongkai Zhang, Benoît Rosa, Florent Nageotte

# Surgical Tool Segmentation using Generative Adversarial Networks with Unpaired Training Data

Zhongkai Zhang, Benoît Rosa, Florent Nageotte

***Abstract*—Surgical tool segmentation is a challenging and crucial task for computer and robot-assisted surgery. Supervised learning approaches have shown great success for this task. However, they need a large number of paired training data. Based on Generative Adversarial Networks (GAN), unpaired image-to-image translation (I2I) techniques (like CycleGAN and dualGAN) have been proposed to avoid the requirement of paired data and have been employed for surgical tool segmentation. The unpaired I2I methods avoid annotating images for domain changes. Instead of using them directly for the segmentation task, we propose new GAN-based methods for unpaired I2I by embedding a specific constraint for segmentation, namely each pixel of input image belongs to either background or surgical tool. Our methods simplify the architectures of existing unpaired I2I with a reduced number of generators and discriminators. Compared with dualGAN, the proposed strategies have a faster training process without reducing the accuracy of the segmentation. Besides, we show that, using textured tool images as annotated samples to train discriminators, unpaired I2I (including our methods) can achieve simultaneous tool image segmentation and repair (such as reflection removal and tool inpainting). The proposed strategies are validated for image segmentation of a flexible tool and for in vivo images from the EndoVis dataset.**

***Index Terms*—Surgical tool segmentation, deep learning, generative adversarial networks, unpaired data**

## I. Introduction

Surgical tool segmentation aims to separate surgical tools apart from the organs background. It plays a fundamental role in robotic-assisted surgery and has many potential applications. The segmentation can provide surgeons with enhanced context-awareness in the operating room. It can also be employed to identify and locate surgical instruments, which is the prerequisite for visual servoing control of surgical robots. In addition, the segmentation is an indispensable procedure for automatic and quantitative evaluation of surgical skills during robotic surgery [1]. Therefore, there is a compelling requirement to develop efficient algorithms for surgical tool segmentation from endoscopic images. However, accurate surgical tool segmentation from tissue backgrounds is a very challenging task due to the complex environment. The quality of the segmentation may be affected by many factors, such as specular reflections, occlusion by blood or smoke, and blur from tool motion.

Recently, deep learning-based approaches have demonstrated cutting-edge performances on surgical tool segmentation [2]. Image segmentation can also be treated as a problem of image-to-image translation (I2I), which has been achieved using Generative Adversarial Networks (GANs). I2I employs either paired training data (pix2pix[3]) or unpaired training data (CycleGAN [4] and dualGAN [5]). Supervised learning approaches (such as paired I2I) need paired training data, i.e. every image has its corresponding annotated image. When using paired I2I for domain variations, manually annotated images need to be collected again because the deep learning model needs to be re-trained to adapt to the new domain. On the contrary, the employment of unpaired I2I for surgical tool segmentation would allow to re-use the annotated images in one domain to train the model for other domains. We can obtain the labeled data by manual segmentation of images in a random domain or from a CAD model with realistic rendering. For the training on another domain with a different background, manual segmentation will not be necessary. This task can be achieved using existing unpaired I2I frameworks, like CycleGAN and dualGAN.

Image segmentation tasks involve classifying pixels to be either the object or the background. This constraint for each pixel (we term it as 0-1 constraint), however, has not been taken into consideration for surgical tool segmentation using existing I2I methods. Besides, traditional unpaired I2I consists of two generators and two discriminators, which needs a substantial training time. In this paper, we focus on the methodological research of new GAN frameworks which are specifically designed for the image segmentation tasks with unpaired data. The first objective is to explore strategies to embed the 0-1 constraint for unpaired I2I, and show to what extent our methods can simplify the traditional unpaired I2I and improve the segmentation performances.

During the surgical procedure, the recorded tool images usually have low quality because of light reflection and occultation by tissues. Therefore, it could be useful to develop methods to obtain tool images with a higher quality by image repair. GAN has been successfully employed to achieve reflection removal [6] and image inpainting [7]. However, as far as we know, the image repair function has not been investigated for surgical tool segmentation. As our second interest, in this paper, we propose to use unpaired I2I for simultaneous tool image segmentation and repair. This is achieved by using the discriminator trained by tool images with detailed textures.

Two main contributions are introduced in this paper: (1) To the best of our knowledge, this is the first work for image segmentation using unpaired I2I with embedded 0-1 constraint. We propose three methodologies to implement this idea. Compared

This work was supported by the INSERM (French Institute for Health) through the Plan Physics for Cancer (projet ROBOT) and by French state funds managed within the "Plan Investissementsd'Avenir" and by the ANR (IHU reference ANR-10-IAHU-02 and Labex CAMI (ANR-11-LABX-0004)).

Authors are with ICube Laboratory, University of Strasbourg, CNRS UMR 7357, France. Contact: zhongkai.zhangzkz@gmail.com

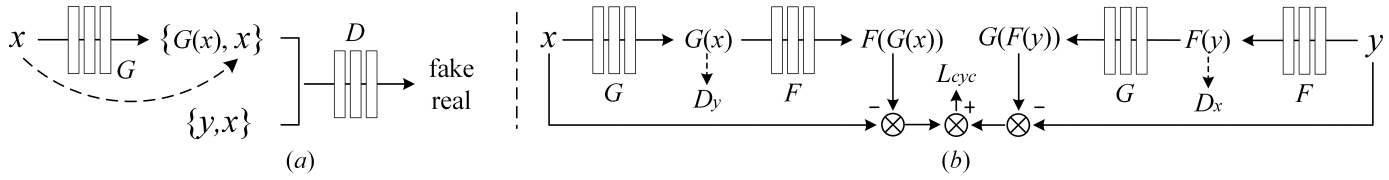


Figure 1. Two training procedures of I2I using GAN. $x$ is input images and $y$ is the training dataset. (a) is the paired I2I framework (pix2pix) [3] which requires paired training data. (b) is the unpaired I2I framework (CycleGAN [4] and dualGAN [5]) using unpaired training data. They consist of two generators $G: x \rightarrow y$ and $F: y \rightarrow x$, and their corresponding two discriminators $D_y$ and $D_x$. $L_{cyc}$ is a cyclic consistency loss which encourages $F(G(x)) \approx x$ and $G(F(y)) \approx y$.

with the general unpaired I2I methods (like dualGAN), our methods have simpler architectures, a faster training process, and a competitive accuracy. (2) We achieve a simultaneous tool image segmentation and repair using unpaired I2I where textured tool images are treated as annotated samples.

This paper is organized as follows. Section II introduces the theoretical background of the loss function and unpaired I2I. Three new strategies are proposed in Section III for the segmentation with unpaired training data. The experiments and results are presented in Section IV. Finally, Section V summarizes the work with discussions and conclusions.

## II. Related Work

### A. Related work of surgical tool segmentation using GAN

Generative Adversarial Networks (GANs) [8] were initially invented for data generation without explicit density models. GAN and its variations have shown great success in the field of medical imaging where it is usually difficult to obtain a large number of labeled samples [9]. For surgical tool segmentation, GAN-based methods have shown their advantages to improve the segmentation quality and reduce the manual annotation efforts. A GAN-based re-colorization method [10] is proposed to retrain segmentation models, which radically reduces the number of labeled images. Adversarial loss is employed to refine the higher-order inconsistency of the feature maps for pixel-wise segmentation of surgical tools [11]. The framework of pix2pix is trained to generate paired annotated images in [12]. Unpaired I2I has been proposed using GAN [4], [5] where a cycle consistency loss ensures an invertible translation. In [13], a CycleGAN network is trained to generate realistic-looking tools from synthetic data. Then, a semantic segmentation neural network is trained using the generated images with their labels. CycleGAN can also be employed for domain adaptation, e.g. learning the mapping from cadaveric to in vivo images [14]. The in vivo images are unlabeled, while the cadaveric images are from a labeled dataset and can be obtained by rendering CAD models of each tool [15]. The implementation of CycleGAN works for surgical tool segmentation with unlabeled data for live images. In a recent work [16], a synthetic dataset is employed to train a deep learning model for surgical tool segmentation. The synthetic dataset is obtained by transforming images from simulation into a real domain using I2I and then blending with a surgical background. The limitations of the above-mentioned approaches are the requirement of either paired training data (paired I2I) or training four networks (unpaired I2I).

### B. An introduction to I2I translation using GAN

The goal of I2I is to learn the mapping between an input image and an output image. GAN has been employed as a general-purpose solution for this translation problem. A comprehensive survey on GAN-based I2I can be found in [17].

The pix2pix framework [3] is a typical solution for I2I using paired training data. The training procedure is shown in Fig. 1 (a) where a conditional GAN, with input image $x$ being the condition, is employed to learn a mapping from $x$ and random noise $z$ to output image $y$. The final objective of pix2pix is [3]

$$\min_D \max_G L_{cGAN}(G,D) + \lambda_{L1} L_{L1}(G) \tag{1}$$

where $L_{cGAN}(G,D)$ is the objective function of a conditional GAN. $L_{L1}(G)$ is a $L1$ distance between the generated output and the ground truth output. $\lambda_{L1}$ is a constant parameter.

CycleGAN [4] and dualGAN [5] can be employed for the tasks where paired datasets are not available. As shown in Fig. 1 (b), unpaired I2I framework has a symmetric structure which consists of two generators $G, F$ and two discriminators $D_y, D_x$. It is trained in an unsupervised manner with the following objective function:

$$\min_{G,F} \max_{D_x,D_y} L_{GAN}(G,D_y,x,y) + L_{GAN}(F,D_x,y,x) + \lambda L_{cyc}(G,F) \tag{2}$$

where $L_{cyc} = L_{L1}(F(G(x)),x) + L_{L1}(G(F(y)),y)$ is the cyclic consistency loss which consists of a forward and a backward cycle-consistency losses.

To avoid the vanishing gradient problem for the Jensen-Shannon divergence, the GAN loss can be formulated based on the Wasserstein distance [18], which requires a Lipschitz condition. The Lipschitz constraint on the discriminator can be enforced by including a gradient penalty directly in the loss function [19]. Then, the minimax objective of Wasserstein GAN with gradient penalty (WGAN-GP) can be written as:

$$\min_D \max_G \mathbb{E}_{z\sim p_z(z)}[D(G(z))] - \mathbb{E}_{x\sim p_x(x)}[D(x)] + \lambda L_{gp}(D) \tag{3}$$

where $z$ is a random input with distribution $p_z(z)$. $\lambda$ is a constant parameter. $L_{gp}$ is employed to penalize $D$ if its gradient norm deviates from 1. We can write it as

$$L_{gp}(D) = \mathbb{E}_{\hat{x}\sim p_{\hat{x}}(\hat{x})}\left[(\| \nabla_{\hat{x}} D(\hat{x}) \|_2 - 1)^2\right] \tag{4}$$

where $\hat{x} \sim p_{\hat{x}}(\hat{x})$ represents random samples with $\hat{x} = \varepsilon x + (1-\varepsilon) G(z)$, and $0 \leq \varepsilon \leq 1$.

## III. Proposed Method

In this section, we first give a brief overview of the three proposed unpaired I2I methods for image segmentation, and then explain them in detail. The details of the network architecture are also provided. Finally, we provide a unified training procedure for our proposed methods.

### A. Overview of proposed methods

Compared with the existing methods for surgical tool segmentation, we make two modifications: (1) The employment of surgical tools with detailed textures as annotation samples; (2) The implementation of 0-1 constraint for the segmentation. We propose three unpaired I2I methods for tool segmentation based on three different ways to implement the 0-1 constraint. I2IS-1D implicitly achieves this constraint, while I2IS-1cD adds a masked image as a condition for an explicit constraint. I2IS-2D allows us to adjust the weight of the 0-1 constraint. All three methods have simpler training procedures with a reduced number of networks, compared with existing unpaired I2I methods. In order to help introducing the proposed methods, Table I summarizes the symbols used in this section.

Table I
Symbols employed in Section III.

| | | |
|---|---|---|
| $x$: input image | $G$: Generator | $M$: image with a mask |
| $y$: annotated sample | $D$: Discriminator | $T$: image with textures |
| $L$: loss function | $L_{gp}$: gradient penalty | |
| $p$: image distribution | $\odot$: Hadamard matrix product | |
| $y_M$: labeled images with masks | | |
| $y_T$: labeled images with textures | | |
| $\lambda$: constant parameter to address gradient penalty | | |
| $\kappa$: constant parameter to address 0-1 constraint | | |
| $G_{M,T}$: map $x$ to its segmented image with masks/textures | | |

*Remark* 1. The 0-1 constraint cannot be achieved directly by adding a specific activation function in $G_M$ (map $x$ to its segmented image with masks), because it will generate a non-differentiable cost function which makes the backpropagation impossible. Instead, we propose to guide the generation using a discriminator which distinguishes images with 0-1 constraint (labeled images with masks $y_M$) and images without 0-1 constraint. The 0-1 constraint is gradually achieved during training. As explained in the next subsection, we use the Hadamard product to ensure that the generated tool image is a label of the input image.

*Remark* 2. In the following, RGB values for images are normalized[1] in the $[-1,1]$ range. In order to facilitate mask multiplication, the background and the masked tool are respectively coded as a gray color and a white color (in the normalized RGB space, $[0,0,0]$ stands for gray and $[1,1,1]$ for white).

### B. I2IS-1D

I2IS-1D achieves unpaired I2I for surgical tool segmentation by modifying the standard GAN architectures (see Fig. 2). The input of I2IS-1D is the image $x$ which needs to be segmented. $G_M$ learns to map $x$ to $G_M(x)$ which has a gray background and a white tool mask. Then, we translate $G_M(x)$ to its corresponding tool image with detailed textures $G_T(x)$ where the RGB value for each pixel is computed by multiplying the corresponding value on $x$ by that on $G_M(x)$, i.e. $G_T(x) = G_M(x) \odot x$ where $\odot$ is the Hadamard matrix product. Instead of classifying $G_M(x)$ using a discriminator $D$ directly, we classify $G_T(x)$ with the textured annotation sample $y_T$. Normalizing the background RGB value for $y_T$ as $[0,0,0]$ implicitly assures that $G_M(x)$ is a tool image with white mask.

[1] https://pytorch.org/docs/stable/torchvision/transforms.html

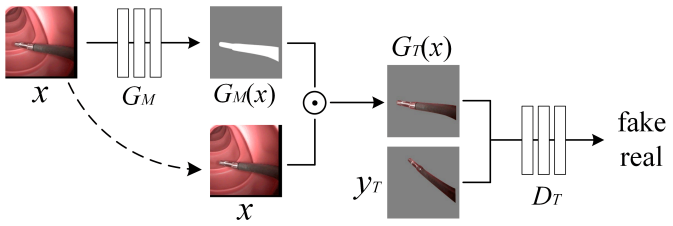


Figure 2. The training procedure of I2IS-1D. I2IS-1D contains one generator $G_M$ and one discriminator $D_T$. The textured tool image $G_T(x)$ is generated by multiplying $x$ by $G_M(x)$ on each pixel. $D_T$ learns to classify $y_T$ with $G_T(x)$.

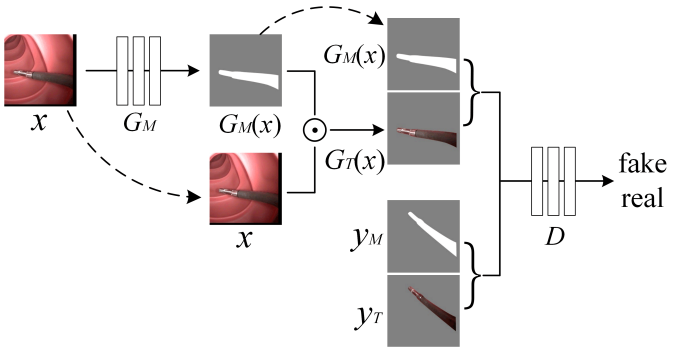


Figure 3. The training procedure of I2IS-1cD. I2IS-1cD consists of one generator $G_M$ and one discriminator $D$. Similarly to I2IS-1D, $G_M$ learns to generate tool images with masks $G_M(x)$. Then, these are converted as textured tool images $G_T(x)$. Instead of classifying $y_T$ with $G_T(x)$ (see Fig.2), $D$ learns to classify the pair of annotation samples $\{y_M, y_T\}$ with that of generated images $\{G_M(x), G_T(x)\}$.

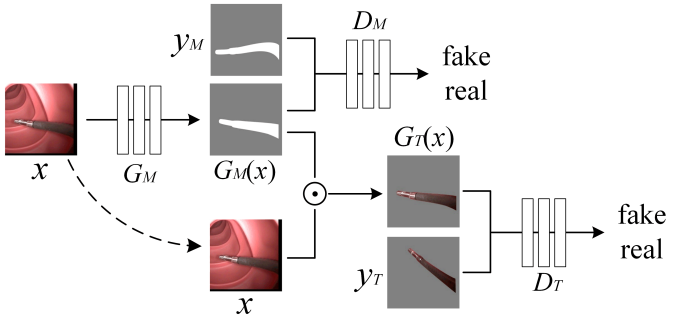


Figure 4. The training procedure of I2IS-2D. I2IS-2D consists of one generator $G_M$, and two discriminators $D_M$ and $D_T$. $G_M$ learns a mapping from $x$ to its corresponding tool image $G_M(x)$ with a mask. $D_M$ classifies $G_M(x)$ and the masked annotation sample $y_M$. $D_T$ classifies the textured annotation image $y_T$ and the textured tool image $G_T(x)$ which is obtained from $G_M(x)$. In this framework, no paired annotation samples are needed.

We employ the strategy of WGAN-GP[19] to train both $G_M$ and $D_T$ at the same time. WGAN-GP reduces the mode dropping phenomenon and relieves the requirements of a careful training balance of $G_M$ and $D_T$. The distributions of $x$ and $y_T$ are denoted respectively as $x \sim p(x)$ and $y_T \sim p(y_T)$. Then, the loss function of I2IS-1D can be written as:

$$\begin{aligned} L(G_M, D_T) =& \mathbb{E}_{x \sim p_x(x)} [D_T (G_M (x) \odot x)] - \\ & \mathbb{E}_{y_T \sim p_{y_T}(y_T)} [D_T (y_T)] + \lambda L_{gp} (D_T) \end{aligned} \tag{5}$$

where $L_{gp}$ is the gradient penalty explained in (4). The final objective is

$$G^* = \arg \min_{D_T} \max_{G_M} L(G_M, D_T) \tag{6}$$

For image segmentation tasks, each pixel on the input image belongs to either the background or the object. This poses a constraint, which we term as a 0-1 constraint, for each pixel. The translation from $G_M(x)$ to $G_T(x)$ implicitly imposes this 0-1 constraint during image generation, and does not increase the architecture's complexity. However, without using this translation, the network generates random and textured tool images which do not come from $x$. Although I2IS-1D has a similar architecture with pix2pix shown in Fig. 1 (a), it does not need paired training data, thanks to this implementation of the 0-1 constraint.

Compared with existing unpaired I2I methods, like CycleGAN or dualGAN shown in Fig. 1 (b), I2IS-1D consists of only one generator and one discriminator, which dramatically reduces the complexity of the training procedure. The employment of annotated tools with detailed textures for $D_T$ training not only assures that the segmented tool $G_T(x)$ has a correspondence with the input image $x$, but also provides a possibility to repair the segmented images, such as to get rid of reflections and predict hidden parts of the tool.

### C. I2IS-1cD

Similarly to I2IS-1D, only one generator $G_M$ and one discriminator $D$ are needed for the implementation of I2IS-1cD. However, I2IS-1cD explicitly imposes the 0-1 constraint to generate tool images (see Fig. 3). The generated tool image $G_M(x)$ with mask is employed as a condition for the textured tool image $G_T(x)$. $G_M(x)$ and $G_T(x)$ are by definition paired between each other, which is denoted as $\{G_M(x), G_T(x)\}$. As for the annotation samples $\{y_M, y_T\}$, the masked tool images $y_M$ and their textured images $y_T$ need also to be paired[2]. Then, $D$ is trained to discriminate the generated image pair $\{G_M(x), G_T(x)\}$ from the annotated image pair $\{y_M, y_T\}$.

We still use WGAN-GP[19] to train both $G_M$ and $D$. The loss function for I2IS-1cD can be written as:

$$\begin{aligned} L(G_M, D) =& \mathbb{E}_{x \sim p_x(x)} [D(\{G_M(x), G_T(x)\})] \\ & - \mathbb{E}_{(y_T, y_M) \sim p_{(y_T, y_M)}(y_T, y_M)} [D(\{y_M, y_T\})] \\ & + \lambda L_{gp} (D) \end{aligned} \tag{7}$$

[2]It does not make the method as a paired I2I (i.e. $y_M$ and $y_T$ are paired together, but they are not paired with the input image $x$).

where $G_T(x) = G_M(x) * x$ and $L_{gp}$ is the gradient penalty shown in (4). The final objective is

$$G^* = \arg \min_{D} \max_{G_M} L(G_M, D) \tag{8}$$

The sample $y_M$ can be easily converted from $y_T$ without involving human annotation. During the training, we need the annotation pair image $\{y_M, y_T\}$ instead of the paired images between $x$ and $y$. As an extension of I2IS-1D, I2IS-1cD is proposed to increase the segmentation accuracy by explicitly imposing the 0-1 constraint.

### D. I2IS-2D

Similarly to I2IS-1cD, I2IS-2D not only explicitly addresses the 0-1 constraint but also learns roughly the object-background pixel distribution. I2IS-2D adds one more discriminator $D_M$ which allows to balance the weight between masked images and textured images. $y_M$ (masked tool images) and $y_T$ (textured tool images) are respectively the annotated samples for $D_M$ and $D_T$. The generated tool image $G_T(x)$ with textures is computed by $G_T(x) = G_M(x) \odot x$ (see Section III-B for details). $D_M$ is trained to distinguish $y_M$ and the generated image $G_M(x)$, while $D_T$ learns to classify $y_T$ and $G_T(x)$. By training $G_M$ and $D_M$ competitively, $G_M$ finally generates realistic masked tool images. The employment of $D_T$ ensures the correspondence between the generated images $G_T(x)$ and the input images $x$.

We still employ the idea of WGAN-GP [19] for the training. However, the original loss function is modified as:

$$\begin{aligned} L(G_M, D_M, D_T) =& \kappa \mathbb{E}_{x \sim p_x(x)} [D_M (G_M (x))] \\ & - \kappa \mathbb{E}_{y_M \sim p_{y_M}(y_M)} [D_M (y_M)] \\ & + \mathbb{E}_{x \sim p_x(x)} [D_T (G_T (x))] \\ & - \mathbb{E}_{y_T \sim p_{y_T}(y_T)} [D_T (y_T)] \\ & + \lambda L_{gp} (D_M, D_T) \end{aligned} \tag{9}$$

where $\kappa \geq 0$ is a constant parameter to balance the weight to generate binary image and textured tool image. I2IS-1D is a special case of I2IS-2D at $\kappa = 0$. $L_{gp}(D_M, D_T)$ is the gradient penalty for both $D_M$ and $D_T$, and the same as (4).

The final objective can be written as:

$$G^* = \arg \min_{D_M, D_T} \max_{G_M} L(G_M, D_M, D_T) \tag{10}$$

As both I2IS-1D and I2IS-1cD, I2IS-2D enables unpaired I2I and generates masked tool images directly, which can be converted to textured tool images using the Hadamard matrix product. Although requiring two discriminators, I2IS-2D allows balancing the loss of the 0-1 constraint and of the textured images by setting the value of $\kappa$. A smaller $\kappa$ enables image repair, while a larger one allows for explicitly imposing the 0-1 constraint. However, this adjustment is impossible by using I2IS-1cD, which means that I2IS-1cD has no ability to repair images during segmentation.

### E. Network Architectures

We employ the "U-Net" architecture [20] for generators in both I2IS-1D and I2IS-2D. "U-Net" is an encoder-decoder

network with skip connections between mirrored downsampling and upsampling layers. It has been shown to have great advantages for image segmentation tasks with a higher accuracy and a lower computation cost [20]. Similarly to pix2pix framework [3], the random noise $z$ is not provided explicitly as $G$'s input. Instead, it is generated by randomly zeroing some of the elements (dropout) on several layers at both training and test time. For discriminators in both frameworks, we use PatchGANs [3] but with a larger reception field $382 \times 382$.

We employ the architecture from [3] with minor changes to adapt to the WGAN-GP framework. The input images have a resolution of $256 \times 256$ with 3 channels. For $G$ in both I2IS-1D and I2IS-2D, the encoder consists of

$$CR64 - CBR128 - CBR256 - CBR512 - CBRD512 \\ - CBRD512 - CBRD512 - CRD512$$

and the U-Net decoder has the form of

$$CBRD512 - CBRD1024 - CBRD1024 - CBRD1024 \\ - CBR1024 - CBR512 - CBR256 - CT128$$

where $C$, $R$, $B$, $D$, and $T$ denote respectively Convolution, ReLU, Batch Normalization, Dropout, and Tanh layers. The number $k$ is the size of feature maps at the input (for the encoder) and output (for the decoder) of each layer. The kernel size for all convolutions is $4 \times 4$ with stride 2 and padding 1. In the encoder, ReLU are leaky with a slope of 0.2, while ReLUs are not leaky in the decoder. A convolution is applied to the last layer in the decoder to map to the number of output channels (3 for RGB images).

The three discriminators (one in I2IS-1D and two in I2IS-2D) have the same architecture:

$$CR64 - CIR128 - CIR256 - CIR521 - CIR1024 - CIR2048 - C1$$

where $I$ denote InstanceNorm. $R$ is LeakyReLU with a slope of 0.2. The kernel size is the same as the one in $G$.

### F. Training Procedure

We introduce a unified training procedure for our three methods: I2IS-1D, I2IS-1cD, and I2IS-2D (see Algorithm 1). It has been noted that I2IS-1D is a special case of I2IS-2D when $\kappa = 0$, and I2IS-1cD is an extension of I2IS-1D adding an image condition. We denote the parameters of $G$, $D_T$ ($D$ for I2IS-1cD) and $D_M$ as $\theta$, $\omega_T$ and $\omega_M$, respectively. These parameters are optimized by modifying the training procedure of WGAN-GP [19]. $D_T$ and $D_M$ are trained $n_{critic}$ steps for one step on $G$. In our work, we set $n_{critic} = 1$. We perform the optimization for all the networks with the Adam solver [21], with a learning rate $\alpha = 0.0002$. The gradient penalty coefficient is set as $\lambda = 10$. The batch size for training is $m = 16$.

**Algorithm 1** A unified training procedure for I2IS-1D, I2IS-1cD, and I2IS-2D.

**Requirement**1: Dataset of input images $X : x \in X$, annotated samples: $Y_M : y_M \in Y_M$ and $Y_T : y_T \in Y_T$.
**Requirement**2: Parameters $\alpha, \beta_1, \beta_2, \lambda, m, n_{critic}, \varepsilon$ and $\kappa$
**Requirement**3: Randomly initialize $\theta$, $\omega_T$ and $\omega_M$
1. **While** $\theta$ has not converged **do**
2. **If** I2IS-1D or I2IS-2D **do**
3. $g \leftarrow G_T(x)$, $g_m \leftarrow G_M(x)$, $y \leftarrow y_T$
4. **Else** I2IS-1cD **do**
5. $g \leftarrow \{G_M(x), G_T(x)\}$, $y \leftarrow \{y_M, y_T\}$
6. **End if**
7. $L_{gp} \leftarrow D_T, g, y$
8. $L^i \leftarrow D_T(g) - D_T(y) + \lambda L_{gp}$
9. $\omega_T \leftarrow \text{Adam}(\frac{1}{m}\sum_{i=1}^{m} L^i, \omega_T, \alpha, \beta_1, \beta_2)$
10. **If** I2IS-1D or I2IS-1cD **do**
11. $\kappa = 0$
12. **Else** I2IS-2D **do**
13. $L_{gp,M} \leftarrow D_M, g_m, y_M$
14. $L_M^i \leftarrow D_M(g_m) - D_M(y_M) + \lambda L_{gp,M}$
15. $\omega_M \leftarrow \text{Adam}(\frac{1}{m}\sum_{i=1}^{m} \kappa L_M^i, \omega_M, \alpha, \beta_1, \beta_2)$
16. **End if**
17. $L_G^i \leftarrow \kappa D_M(g_m) + D_T(g)$
18. $\theta \leftarrow \text{Adam}(\frac{1}{m}\sum_{i=1}^{m} L_G^i, \theta, \alpha, \beta_1, \beta_2)$
19. **End while**

## IV. Experiments and Results

In this section, we first introduce the datasets we used for testing the methods, and then three metrics to evaluate the segmentation performances. Finally, we present the experimental results qualitatively and quantitatively.

### A. Dataset

We test our methods on four datasets (see Fig. 5) where all images have been segmented manually and reshaped to be $256 \times 256$. Dataset 1 and 2 are recorded using the STRAS robot [22], which is a flexible robotic system for endoscopic surgery. A soft tool is telemanipulated to bend, translate and rotate inside two phantoms [23]. Dataset 4 is built by adding disturbance pixels on both the background and the tool, while the segmented tool images are kept the same as in dataset 1. It is employed to test the performance of image repair. There are 220 images in each dataset 1, 2, and 4, which consist of 176 for training and 44 for test. In order to test our methods on real in vivo images, we employ the EndoVis dataset [24] which has eight surgical scenes with different rigid surgical tools. We separate this dataset into 1616 training images and 185 test images.

For the validation of I2IS-1cD and I2IS-2D, we need two annotated data for the training of discriminators: one with masks (see M in Fig. 5), the other with textures (see T in Fig. 5). For I2IS-1D, the labels with masks are not necessary for the training. The tool images in test datasets are segmented manually and used as ground truth for the evaluation.

### B. Evaluation Metrics

To measure the accuracy of mask generation for surgical tool segmentation, we use two commonly used metrics [25], [26]: (1) Intersection over Union (IoU), (2) Dice index. The two indices measure the overlap between the predicted ($A$) and true ($B$) masks by $\text{IoU}(A,B) = \| A \cap B \| / \| A \cup B \|$ and $\text{Dice}(A,B) = 2 \| A \cap B \| / (\| A \| + \| B \|)$. Both IoU and the Dice indices are bounded between 0 and 1. The lower the

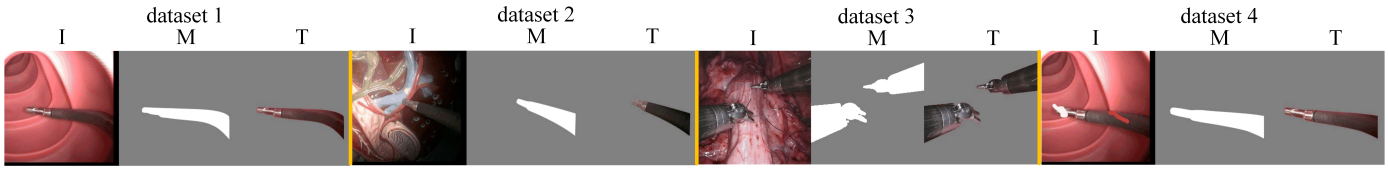


Figure 5. Samples from four datasets. Dataset 1 and dataset 4 have the same annotated samples to train discriminators.

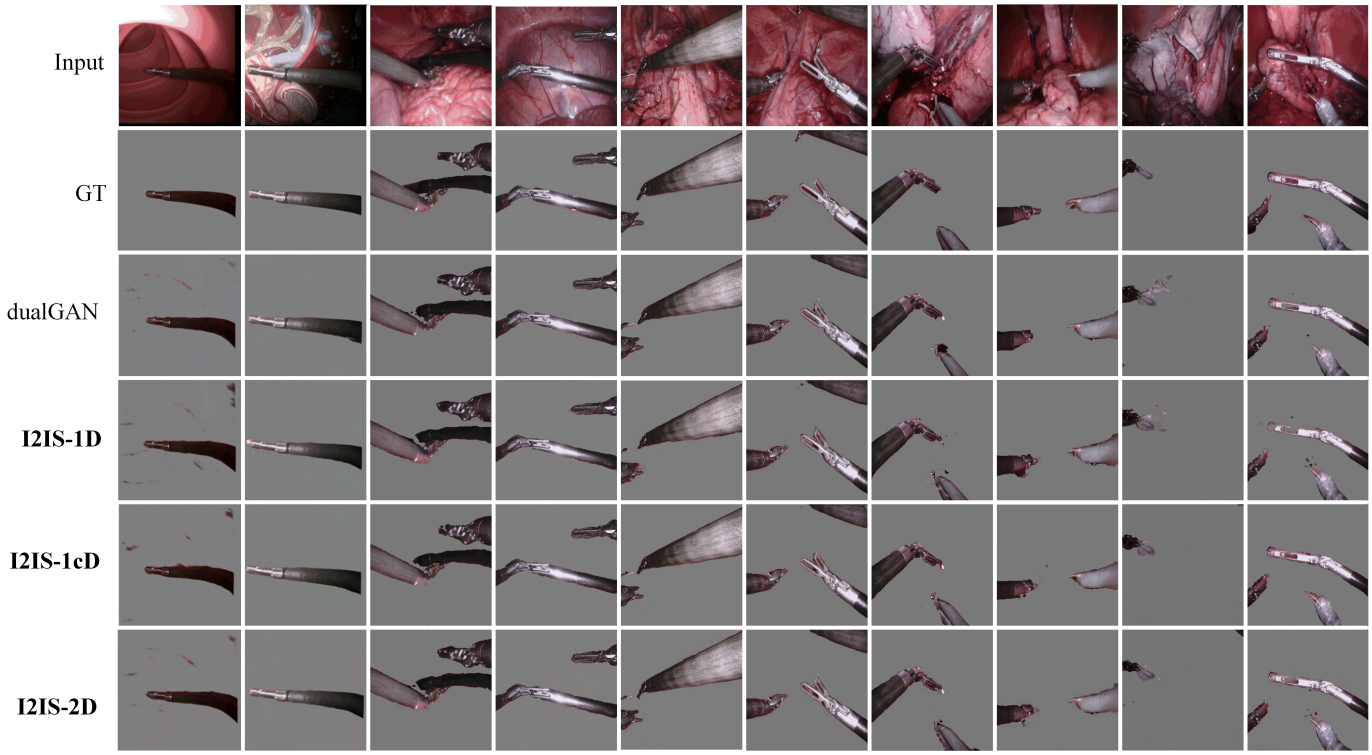


Figure 6. Qualitative results for surgical tool segmentation. From left to right: input image, ground truth (GT), dualGAN [5], I2IS-1D, I2IS-1cD, and I2IS-2D. Generated tool images with textures are visualized for three different datasets (see Fig. 5 where dataset 3 consists of eight surgical videos).

Table II
QUANTITATIVE COMPARISON BETWEEN OUR PROPOSED METHODS, AND THREE OTHER METHODS: SUPERVISED LEARNING(SL), PIX2PIX, AND DUALGAN. THE RED AND BLUE NUMBERS CORRESPOND TO RESPECTIVELY THE BEST AND THE WORST INDEX AMONG THE UNPAIRED METHODS.

| | Dataset 1 | | | | | | Dataset 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SL | pix2pix | dualGAN | **I2IS-1D** | **I2IS-1cD** | **I2IS-2D** | SL | pix2pix | dualGAN | **I2IS-1D** | **I2IS-1cD** | **I2IS-2D** |
| time(h) | $\approx 0.5$ | $\approx 1.9$ | $\approx$ **4.2** | $\approx$ **1.9** | $\approx$ **1.9** | $\approx 3.1$ | $\approx 0.5$ | $\approx 1.9$ | $\approx$ **4.2** | $\approx$ **1.9** | $\approx$ **1.9** | $\approx 3.1$ |
| IoU | 0.936 | 0.932 | **0.920** | 0.928 | 0.925 | **0.932** | 0.943 | 0.937 | **0.884** | **0.884** | **0.918** | 0.897 |
| Dice | 0.966 | 0.964 | **0.957** | 0.962 | 0.960 | **0.964** | 0.971 | 0.967 | 0.936 | **0.934** | **0.955** | 0.942 |
| $10 \times L_1$ | 0.054 | 0.056 | **0.064** | 0.074 | **0.105** | 0.072 | 0.050 | 0.056 | 0.087 | 0.101 | **0.102** | **0.082** |

indices, the worse the prediction result. If the prediction is completely correct, IoU = Dice = 1. We first convert the RGB images of true and predicted masks into gray images, which are further binarized to images with pixels value either 0 (background) or 1 (tool). Then, the average IoU and Dice indices are computed for the whole test dataset.

We also employ L1 loss to measure the accuracy of the image translation where the goal is to generate tool images with textures. We compute $L_1$ loss by $L_1 = \frac{1}{N}\sum_{i=1}^{N} \mid y_{pred} - y_{true} \mid$ where $y_{pred}$ and $y_{true}$ are respectively the generated tool image with textures, and its ground truth. $N$ is the number of elements in the image array, including $256 \times 256$ pixels with 3 channels.

### C. Segmentation Accuracy on Datasets 1 and 2

The three proposed methods (I2IS-1D, I2IS-1cD, and I2IS-2D ($\kappa = 0.1$)) are compared with three traditional strategies: supervised learning (SL), pix2pix [3] and dualGAN [5]. It must be noted that our methods and dualGAN belong to unpaired I2I, while pix2pix is an I2I method using paired training data. We provide a quantitative comparison (see Tab. IV-A) of segmentation accuracy using all these methods. The indexes IoU, Dice are employed to evaluate the generation of tool images with masks, while $L_1$ is a measurement of the textured image generation including the background. All methods are trained to generate masked tool images, which can then be

converted to textured images. We assume that the annotated samples are rich enough to capture the tool's configurations. To simplify data gathering, the samples are segmented manually from the training dataset as ground truths. However, for training discriminators, the segmented tool image and the input image are not paired. In Fig. 6, we show the segmented surgical tool with textures on randomly selected test images from datasets 1, 2, and 3 using unpaired methods (dualGAN, I2IS-1D, I2IS-1cD, and I2IS-2D).

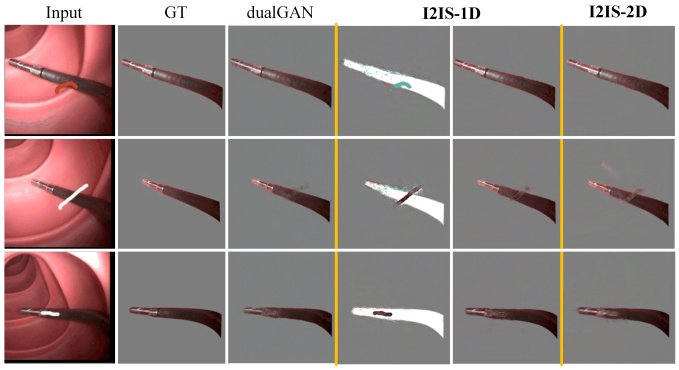


Figure 7. The performances of image repair using different methods. From left to right: input image, ground truth (GT), dualGAN, I2IS-1D (including the generated masked tool $G_M(x)$ and the converted textured tool $G_T(x)$), and I2IS-2D ($\kappa = 0.1$). Random disturbances are added to the test images to simulate the reflections and the hidden section on the surgical tool. All the methods use textured tool images as annotated samples to train the discriminators.

Table III
QUANTITATIVE COMPARISON OF SEGMENTATION ACCURACY USING ENDOVIS DATASET BETWEEN OUR PROPOSED METHODS, AND DUALGAN. THE RED AND BLUE NUMBERS CORRESPOND TO RESPECTIVELY THE BEST AND THE WORST INDEX. THE INDEXES ARE THE AVERAGE INDEXES FOR EIGHT SCENES IN ENDOVIS DATASET.

| Dataset 3 | dualGAN | **I2IS-1D** | **I2IS-1cD** | **I2IS-2D** |
|---|---|---|---|---|
| time(h) | $\approx$ **39** | $\approx$ **17** | $\approx$ **17** | $\approx$ 27 |
| Average IoU | 0.857 | **0.841** | **0.867** | 0.855 |
| Average Dice | 0.917 | **0.908** | **0.924** | 0.916 |
| Average $10 \times L_1$ | 0.118 | **0.129** | **0.114** | 0.128 |

Table IV
SUMMARY OF METHODS MENTIONED IN THIS PAPER. IR: IMAGE REPAIR; UT: UNPAIRED TRAINING; MA: MASK ANNOTATION; CS: RANK OF CONVERGENCE SPEED; AS: RANK OF ARCHITECTURE SIMPLICITY; ACC.(M): RANK OF ACCURACY OF MASKED IMAGE GENERATION; ACC.(T): RANK OF ACCURACY OF TEXTURED IMAGE GENERATION; Y: YES; N: NO. THE RED AND BLUE NUMBERS CORRESPOND TO RESPECTIVELY THE BEST AND THE WORST INDEX.

| method | IR | UT | MA | CS | AS | Acc.(M) | Acc.(T) |
|---|---|---|---|---|---|---|---|
| SL | **N** | **N** | **N** | **1** | **1** | **1** | **1** |
| pix2pix | **N** | **N** | **N** | 2 | 2 | 2 | 2 |
| dualGAN | **Y** | **Y** | **N** | **6** | **6** | 4 | 3 |
| **I2IS-1D** | **Y** | **Y** | **N** | 2 | 2 | **6** | 4 |
| **I2IS-1cD** | **N** | **Y** | **Y** | 2 | 2 | 3 | 4 |
| **I2IS-2D** | **Y** | **Y** | **Y** | 5 | 5 | 4 | 4 |

We use the same generator's architecture for the deep network in SL, and generators in the other five methods. All the GAN-based methods have the same network architecture for discriminators. We also set the same training parameters (see Section III-F) for all mentioned methods. Momentum parameters are set to be $\beta_1, \beta_2 = (0, 0.9)$ for the Adam solver. The number of epochs is set to 400, which allows all these methods to converge. SL and pix2pix, which are both fully supervised, have a faster convergence speed and a higher accuracy than the other methods. For the unpaired I2I, our methods have a competitive accuracy with dualGAN. The accuracy of masked image generation for the three proposed methods is ranked as: I2IS-1cD > I2IS-2D($\kappa = 0.1$) > I2IS-1D. A reasonable explanation is the explicit consideration of 0-1 constraint during training. We also noted that our methods have a much faster training speed than dualGAN (see Tab. IV-A, line 2) because of the reduced number of generators and discriminators.

### *D. Segmentation Accuracy on Dataset 3*

The same network structures and training parameters (except momentum parameters $\beta_1, \beta_2 = (0.9, 0.99)$) are employed to test the performance of our three methods using the EndoVis dataset. Results on eight randomly selected images from each surgical video are shown in Fig. 6. We also provide a quantitative comparison (see Tab. III) of the segmentation accuracy using unpaired I2I. Overall, our methods have a similar accuracy for tool segmentation as the traditional dualGAN. However, the training time using our methods is much lower (30-55% faster depending on the method).

### *E. Image Repair Performances*

A simultaneous tool image segmentation and repair can be achieved using unpaired I2I if textured tool images are employed to train the discriminator. In this experiment, dualGAN, I2IS-1D and I2IS-2D ($\kappa = 0.1$) are trained using dataset 4 (see Fig. 5). We randomly select three input images for the test and a qualitative visualization is shown in Fig. 7.

The 0-1 constraint helps to achieve an unpaired I2I for image segmentation. However, it results in a direct copy of the tool pixels from the input image, which does not help to repair tool images. To solve this problem, we set a smaller weight (like $\kappa = 0.1$ for I2IS-2D) on the 0-1 constraint. In this case, the 0-1 constraint feature of the generated image ($G_M(x)$ in Fig. 4) will be sacrificed partially in order to make sure that the output image $G_T(x)$ has a correct texture (see the difference between the two columns for I2IS-1D in Fig. 7).

It is noted that I2IS-1cD set the same weight to 0-1 constraint and texture loss, which results in the failure of image repair. Similarly, I2IS-2D fails in this task if a larger $\kappa$ (such as $\kappa = 1$) is chosen. However, I2IS-1D works for this task because it addresses the constraint implicitly. From the comparison in Fig. 7, we found that dualGAN outperforms our methods in terms of image repair. A possible reason is that, in this case, dualGAN was trained to generate textured tool images directly without being affected by the 0-1 constraint.

### *F. Summary of the Mentioned Methods*

In this subsection, we present a summary (see Table IV) to explicit the advantages and disadvantages of using the proposed methods for surgical tool segmentation. Our methods have a

simple training procedure which results in faster training. The embedded 0-1 constraint increases the accuracy of generated masked tool images. However, they reduce the accuracy of generated textured tool images, compared with other unpaired I2I methods (like dualGAN). Among our proposed three methods, we found that the accuracy to generate masked tool images can be increased by increasing the weight of 0-1 constraint. For example, I2IS-1cD has the highest accuracy to generate masked tool images. The selection of I2IS-1D, I2IS-1cD, or I2IS-2D for applications could be based on the availability of masked labels, the necessity of image repair, and the type of the target segmented tool images (with masks or with textures).

## V. Discussion and Conclusion

This work contributes to developing new unpaired I2I methods for surgical tool segmentation by embedding 0-1 constraint for each pixel. We propose three methods (I2IS-1D, I2IS-1cD, and I2IS-2D) to generate segmented surgical tool images with masks, as well as providing tool images with textures. I2IS-1D implicitly embeds the 0-1 constraint, while I2IS-1cD and I2IS-2D explicitly achieve this constraint guided by the discriminator. Compared with traditional unpaired I2I methods which consist of two generators and two discriminators, our proposed methods reduce the number of networks (I2IS-1D and I2IS-1cD consist of only one generator and one discriminator, and I2IS-2D with one generator and two discriminators). The simpler architecture decreases the training time. Besides, increasing the weight of embedded 0-1 constraint slightly increases the accuracy of masked tool image generation. Instead of using masked tool images as annotated samples, we employ textured tool images to train the discriminators. The benefit of this strategy, together with unpaired I2I methods, allows to repair tool images, such as reflection removal and hidden parts inpainting. We perform some experiments to validate our methods for the segmentation of a flexible surgical tool inside two phantoms, and rigid tools in various surgical settings (EndoVis dataset).

As our future work, we plan to obtain annotated samples from simulation and extend the methods to segment different parts of the surgical tool [27].

## References


[1] D. Lee, H. W. Yu, H. Kwon, H.-J. Kong, K. E. Lee, and H. C. Kim, "Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations," *Journal of clinical medicine*, vol. 9, no. 6, p. 1964, 2020.

[2] S. Bodenstedt, M. Allan, A. Agustinos, X. Du, L. Garcia-Peraza-Herrera, H. Kenngott, T. Kurmann, B. Müller-Stich, S. Ourselin, D. Pakhomov, *et al.*, "Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery," *arXiv preprint arXiv:1805.02475*, 2018.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[5] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, pp. 2849–2857, 2017.

[6] R. Abiko and M. Ikehara, "Single image reflection removal based on gan with gradient constraint," in *Asian Conference on Pattern Recognition*, pp. 609–624, Springer, 2019.

[7] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," *arXiv preprint arXiv:1803.07422*, 2018.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[9] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019.

[10] T. Ross, D. Zimmerer, A. Vemuri, F. Isensee, M. Wiesenfarth, S. Bodenstedt, F. Both, P. Kessler, M. Wagner, B. Müller, *et al.*, "Exploiting the potential of unlabeled endoscopic video data with self-supervised learning," *International journal of computer assisted radiology and surgery*, vol. 13, no. 6, pp. 925–933, 2018.

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

[12] A. Marzullo, S. Moccia, M. Catellani, F. Calimeri, and E. De Momi, "Towards realistic laparoscopic image generation using image-domain translation," *Computer Methods and Programs in Biomedicine*, vol. 200, p. 105834, 2021.

[13] I. Azqueta-Gavaldon, F. Fröhlich, K. Strobl, and R. Triebel, "Segmentation of surgical instruments for minimally-invasive robot-assisted procedures using generative deep neural networks," *arXiv preprint arXiv:2006.03486*, 2020.

[14] S. Lin, F. Qin, Y. Li, R. A. Bly, K. S. Moe, and B. Hannaford, "Lc-gan: Image-to-image translation based on generative adversarial network for endoscopic images," *arXiv preprint arXiv:2003.04949*, 2020.

[15] D. Pakhomov, W. Shen, and N. Navab, "Towards unsupervised learning for instrument segmentation in robotic surgery with cycle-consistent adversarial networks," *arXiv preprint arXiv:2007.04505*, 2020.

[16] E. Colleoni and D. Stoyanov, "Robotic instrument segmentation with image-to-image translation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 935–942, 2021.

[17] A. Alotaibi, "Deep generative adversarial networks for image-to-image translation: A review," *Symmetry*, vol. 12, no. 10, p. 1705, 2020.

[18] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 214–223, 2017.

[19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, pp. 5767–5777, 2017.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2014.

[22] L. Zorn, F. Nageotte, P. Zanne, A. Legner, B. Dallemagne, J. Marescaux, and M. de Mathelin, "A novel telemanipulated robotic assistant for surgical endoscopy: preclinical application to esd," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 4, pp. 797–808, 2017.

[23] C. da Costa Rocha, N. Padoy, and B. Rosa, "Self-supervised surgical tool segmentation using kinematic information," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8720–8726, IEEE, 2019.

[24] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt, *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

[25] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 624–628, IEEE, 2018.

[26] T. Eelbode, J. Bertels, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3679–3690, 2020.

[27] D. Pakhomov, V. Premachandran, M. Allan, M. Azizian, and N. Navab, "Deep residual learning for instrument segmentation in robotic surgery," in *International Workshop on Machine Learning in Medical Imaging*, pp. 566–573, Springer, 2019.