Computer Methods and Programs in Biomedicine (2020)

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.journals.elsevier.com/computer-methods-and-programs-in-biomedicine

Towards realistic laparoscopic image generation using image-domain translation

Aldo Marzullo[a], Sara Moccia[b,c], Michele Catellani[d], Francesco Calimeri[a], Elena De Momi[e]

[a]*Department of Mathematics and Computer Science, University of Calabria, Rende, Italy*
[b]*Department of Information Engineering, Unviersità Politecnica delle Marche, Ancona, Italy*
[c]*Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genoa, Italy*
[d]*Department of urology, European Institute of Oncology (IEO), IRCCS, Milan, Italy*
[e]*Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy*

ARTICLE INFO

Article history:

Keywords: Generative Adversarial Networks, Minimally Invasive Surgery, Image translation, Data Augmentation

ABSTRACT

Background and Objectives Over the last decade, Deep Learning (DL) has revolutionized data analysis in many areas, including medical imaging. However, there is a bottleneck in the advancement of DL in the surgery field, which can be seen in a shortage of large-scale data, which in turn may be attributed to the lack of a structured and standardized methodology for storing and analyzing surgical images in clinical centres. Furthermore, accurate annotations manually added are expensive and time consuming. A great help can come from the synthesis of artificial images; in this context, in the latest years, the use of Generative Adversarial Neural Networks (GANs) achieved promising results in obtaining photo-realistic images. **Methods** In this study, a method for Minimally Invasive Surgery (MIS) image synthesis is proposed. To this aim, the generative adversarial network *pix2pix* is trained to generate paired annotated MIS images by transforming rough segmentation of surgical instruments and tissues into realistic images. An additional regularization term was added to the original optimization problem, in order to enhance realism of surgical tools with respect to the background. **Results** Quantitative and qualitative (i.e., human-based) evaluations of generated images have been carried out in order to assess the effectiveness of the method. **Conclusions** Experimental results show that the proposed method is actually able to translate MIS segmentations to realistic MIS images, which can in turn be used to augment existing data sets and help at overcoming the lack of useful images; this allows physicians and algorithms to take advantage from new annotated instances for their training.

1. Introduction

Minimally invasive surgery (MIS) is currently the elective treatment for many procedures, such as nephrectomy and prostatectomy. By relying on small incisions on patient abdomen, through which surgical tools and endoscope are inserted, MIS attenuates some of the drawbacks of open surgery, such as prolonged patient hospitalization and recovery time [1]. Nevertheless, MIS suffers from some drawbacks in terms of reduced field of view on the surgical site, which may hamper surgeon context awareness, and restricted freedom of movement for surgical action [2]. The surgical data science (SDS) community has been more and more focusing on developing solutions for Computer Assisted Interventions (CAI), with the final goal of increasing context awareness and providing decision support to surgeons [3]. Surgical phase recognition [4], 3-D reconstruction of soft tissues [5], tissue classification [6] and surgical tool segmentation and pose estimation [7] are among the main challenges of SDS. To address these challenges, while tackling

e-mail: marzullo@mat.unical.it (Aldo Marzullo)

the high variability encoded in the laparoscopic images, CAI methods commonly rely on deep learning (DL) [8]. The development of effective DL algorithms requires large annotated datasets, which are currently not available in this field, despite laudable initiatives promoted by some scientific communities (e.g., MICCAI[1] and ISBI[2]). Manual annotation has to be performed, which consists of time-consuming and tedious procedures; indeed, even though it has been shown that untrained anonymous individuals from online communities can generate training data of expert quality, obtaining manual annotation still remains a challenging task [9].

Data augmentation using affine transformations (e.g., rotation and scaling) is often used to try to tackle the small size of annotated datasets in several scenarios. Nonetheless, this approach presents known drawbacks, such as very limited diversities in existing data, especially when a small dataset is processed for augmentation [10]. Researchers in other fields proposed to artificially generate unseen images, while automatically providing semantic label-map annotation, using generative adversarial networks (GANs). GANs have been proven to be able to estimate the underlined distribution of data when dealing with very complicated data structures, and learn to replicate meaningful samples [11, 12, 13].

In this work, we specifically address the problem of generating realistic MIS images with surgical tools in the field of view, using the surgical-tool semantic label maps as constraint for the GAN generation process. This concept, in the literature referred to as image-to-image translation problem, is widely explored for synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images. A way to address image-to-image translation problems is to use conditional GANs (cGANs), which leverage standard GANs in a conditional setting [14]. In the herein addressed scenario, cGANs allow one to generate realistic MIS images, while having the corresponding surgical-tool label masks available. This has great potential, as the number of currently available datasets of surgical instruments in MIS is limited (such as ENDOVIS Grand Challenge[3]), despite a growing number of research works have been published on the topic of instrument-tool analysis during MIS in the last few years [15, 7, 16].

1.1. Contribution

In order to tackle the issues described above, we propose to train a *pix2pix* cGAN [14] to translate semantic segmentation label maps of surgical tools to realistic MIS images with surgical tools in the field of view. To the best of our knowledge, although cGANs have been used for tasks in the MIS field, previous approaches only marginally involved surgical tools in the field of view, being more focused on colonoscopy data (see Sec. 2). As previously reported in the literature [17], cGANs allow to optimize image-translation and RGB-image prediction simultaneously, while remaining independent of camera, lighting and patient characteristics. In order to further condition the image-to-image translation problem, in this work we add a rough tissue-label map to the surgical-tool label map. This is done to tackle the high variability encoded in MIS images, in terms of noise, blur, illumination levels, tissues in the field of view (with different shape and texture), presence of occlusions and surgical tool pose, size and shape (as shown in Fig. 1). It is worth noting that obtaining such rough label map for abdominal tissues is a trivial task, and state- of-the-art algorithms can be used for the purpose (e.g., [6]).

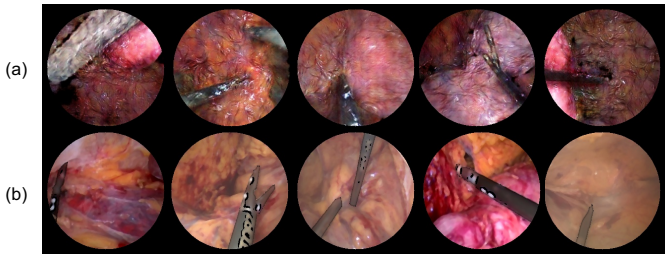

Fig. 1. Samples from the (a) generated and (b) real MIS images.

The original *pix2pix* loss function was here modified to improve the realism of surgical tools by adding an additional regularization parameter, which leverages the *pix2pix* generator to preserve the high-level features of surgical tools (e.g., sharp borders). The generated images were evaluated both analytically and by means of a custom web platform, where surgeons and SDS experts were asked to examine images and try to correctly identify if they were real or "fake".

The remainder of the paper is structured as follows. In Section 2 we survey relevant work in the literature, focusing on cGAN applications for endoscopic images. Section 3 provides a detailed description of the proposed approach; results are presented in Section 4 and discussed in Section 5. Eventually, Section 6 draws our conclusions and summarizes the main contribution of the present work.

2. Related work on cGANs in endoscopy

The GAN framework has gained growing interest in computer vision and medical imaging as a more reliable approach to generate realistic training images rather than using synthetically augmented datasets. Applications include magnetic-resonance images generation [18, 11], liver-lesion classification [19], and pathology-progress simulation [20].

Recently, research in the field of GANs is focusing more and more on image-to-image translation problem, with the goal to generate new, unseen images as a combination of the content of an image and the style of another. Among those methods, the *pix2pix* network [14] received particular attention and has been used a wide range of sectors, including the biomedical domain [17]. The *pix2pix* is a deep learning model that solves the image-translation problem using a cGAN to learn a function to map the input image to the output one. The framework is composed by two main components, the generator and the discriminator. The generator is trained to transform the input image to

[1] http://www.miccai.org/
[2] http://2020.biomedicalimaging.org/
[3] https://endovis.grand-challenge.org/

the target one, while the discriminator measures the output and target similarity to encourage the generator to create a plausible translation. Another approach proposed for image-translation problems is CycleGAN [21], where two generator and two discriminator models are introduced for modeling both the direct and inverse translation mapping. In the specific context of laparoscopic images, both models have achieved remarkable results.

Focusing on the endoscopy domain, Rau *et al.* [17] used the *pix2pix* method to transform endoscopic images into depth maps, to compensate for the lack of labelled training data for deep-learning algorithms for depth estimation. Mathew *et al.* [22] presented a deep-learning framework, called Extended and Directional CycleGAN, for image-to-image translation between optical and virtual colonoscopy, with the final foal of increasing the size of optical colonoscopy data. Similarly, Oda *et al.* [23] used CycleGAN for generating optical colonoscopy data from virtual ones obtained from computerized tomography scans. CycleGAN has also being used by Esteban *et al.* [24], to translate bronchoscopic intra-operative videos to virtual bronchoscopies.

The majority of the approaches in the literature are focused on generating images with only one structure in the field of view (e.g., bronchi or colon). With respect to those endoscopic images, MIS images present higher variability in terms of tissues in the field of view, surgical tools shape, size and pose, and illumination level. Images acquired during MIS often present high noise level, blur, specularities, as well as presence of smoke and blur in the field of view. All this aspects pose challenges to the generation of realistic images. Hence, as the best of our knowledge, this work is the first attempt which investigates the problem of translating semantic label maps into realistic MIS images.

CycleGAN has been observed to work well on tasks that involve color or texture changes, like virtual to optical colonoscopy. However, for tasks that require substantial geometric changes to the image, such as cat-to-dog translations, CycleGAN usually fails [21]. Hence, in this work, the *pix2pix* method is used to produce a mapping between semantic label maps and new, unseen, RGB MIS images.

3. Materials and Methods

In the following, we report a detailed description of background techniques and methods, along with details on the herein proposed approach.

3.1. Dataset Description and Semantic Label Map Preparation

The EndoVis Dataset from the EndoVis Intrument Segmentation and Tracking Challenge 2017[4] was used for this study. Sample images from the dataset are shown in Fig. 1 (b). The dataset is made of 40 2D in-vivo images (image size = 720×576 pixels) from 4 laparoscopic colorectal surgeries, for a total of 160 images with the corresponding surgical-tool label map.

[4]https://endovissub-instrument.grand-challenge.org/Data/

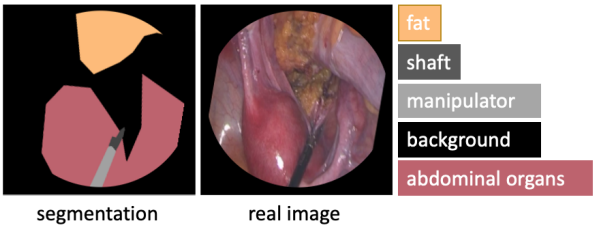

Fig. 2. Schematic representation of the semantic label map used to generate realistic MIS images. The label map includes accurate surgical tool annotation and rough annotation of anatomic structures (i.e., fat and abdominal organs).

Each pixel is labelled as either background, shaft and manipulator.

Rough manual annotation of tissues was added to the original EndoVis label maps in order to explicitly leverage *pix2pix* learning the variability encoded in abdominal tissues. We empirically observed that having such a rough annotation allowed to increase the realism of generated images. Tissues were roughly annotated considering two classes, namely abdominal organs and fat. A sample of the resulting semantic label map, with both surgical tools and abdominal tissues, is shown in Fig. 2.

We also applied geometrical and morphological data augmentation (i.e., rotation, scaling, and shifting) to increase the number of training instances by a factor of 10.

3.2. Conditional Generative Adversarial Networks

As introduced in Sec. 1, GANs are a class of generative models that learn the probability distribution over a dataset [25]. The standard GAN implementation can be described as a zero-sum game, in which two networks, the *Generator* (G) and *Discriminator* (D), compete to one another to maximize its own payoff.

In this adversarial process, the two networks are trained simultaneously: the Generator network produces samples $x = G(z|\theta^{(g)})$, where $\theta^{(g)}$ refers to the parameters of G, and z is a random input noise (samples from a uniform distribution). The Discriminator network estimates $D(x|\theta^{(d)})$, which refers to the probability that x is a real training example, rather than an artificial sample drawn from the model.

Formally, the training objective for a GAN can be described as:

$$G^{\star} = arg \min_{G} \max_{D} \mathcal{L}_{GAN}(G, D) \tag{1}$$

where:

$$L_{GAN}(G, D) = E_x[logD(x)] + E_{x,z}[log(1 - D(x, G(x, z))] \tag{2}$$

The cGANs [21, 26, 14] extend the GANs by conditioning the generator and critic on prior information (c). The training objective for a Conditional GAN then becomes:

$$\begin{aligned} L_{cGAN}(G, D) := E_{x,c}[logD(x, c)] + \\ E_{x,z}[log(1 - D(x, G(x, z))] \end{aligned} \tag{3}$$

Several researchers have pointed out the benefits of combining the cGAN objective with a more traditional loss, such as the L_1 distance [26, 14]. This is formally expressed by:

$$L_1(G) = E_{x,c,z}[\|c - G(x,z)\|_1] \tag{4}$$

This objective can be used to train to produce outputs that look realistic and as close as possible to the real label, that is, it regularizes the generator model to output images that are a plausible translation of the source image. The final objective is then:

$$G^\star = arg \min_G \max_D L_{cGAN}(G,D) + \lambda \mathcal{L}_1(G) \tag{5}$$

where λ is a regularization parameter.

3.3. cGAN for laparoscopic image generation

In this work, we take advantage of a *pix2pix* cGAN [14] to translate a semantic label map relevant to tissue and instruments in realistic, synthetic, laparoscopic images. The *pix2pix* architecture is illustrated in Fig. 3, and described in Table 1 (generator) and Table 2 (discriminator).

In *pix2pix*, the discriminator is a deep CNN that performs conditional-image classification. The discriminator is a CNN that is run at patch level. The difference between the *pix2pix* discriminator and standard discriminators is that, instead of producing output as single scalar vector, the *pix2pix* discriminator generates an $N \times N$ array, when $N \times N$ is the patch size. The average prediction of all patches is used to classify the whole image as real or fake.

The generator follows the U-Net encoder-decoder network architecture, with long skip connections between mirrored layers in the encoder and decoder stacks for recovering the resolution lost in the encoder path [27]. The network incorporates down-scaling on multiple levels to learn different degree of details, from a general representation to more local feature representations. Such architecture has been shown produce excellent results in several relevant scenarios, including medical image segmentation and translation [17, 14, 27]

The generator is trained on pairs (x, c), where x is the semantic label map with tissues and surgical tools and c is the corresponding RGB endoscopic image. The random noise vector z of Eq. 5 is not provided as input, instead, it is simulated through dropout layers with probability 0.5 during both training and inference. Such an approach has been proven to give comparable results, while simplifying the overall implementation [14].

The endoscopic images we aim to simulate in this study can be schematically decomposed in two semantic sets: surgical tools and abdominal tissues. However, because of several differences in light and shapes, learning to simulate such hierarchy could not be a trivial task, resulting in blurred and non-realistic surgical-tool boundaries. In this work, to leverage a more effective image-to-image translation, we explicitly penalise differences between the true and generated surgical tools. Let t be the semantic label map of the surgical tools in an image, the loss function of the cGAN (Eq. 5) was modified by adding a further regularization term:

$$L_t(G) = E_{x,c,t,z}[\|ct - G(x,z)t\|_1] \tag{6}$$

Table 1. Architecture of the generator network

Id	Layer (type)	Output Shape	Connected to
1	Input	$256 \times 256 \times 3$	
2	2DConv	$128 \times 128 \times 64$	1
2a	Leaky ReLU	$128 \times 128 \times 64$	2
3	2DConv	$64 \times 64 \times 128$	2a
3a	Batch Norm	$64 \times 64 \times 128$	3
3b	Leaky ReLU	$64 \times 64 \times 128$	3a
4	2DConv	$32 \times 32 \times 256$	3b
4a	Batch Norm	$32 \times 32 \times 256$	4
4b	Leaky ReLU	$32 \times 32 \times 256$	4a
5	2DConv	$16 \times 16 \times 512$	4b
5a	Batch Norm	$16 \times 16 \times 512$	5
5b	Leaky ReLU	$16 \times 16 \times 512$	5a
6	2DConv	$8 \times 8 \times 512$	5b
6a	Batch Norm	$8 \times 8 \times 512$	6
6b	Leaky ReLU	$8 \times 8 \times 512$	6a
7	2DConv	$4 \times 4 \times 512$	6b
7a	Batch Norm	$4 \times 4 \times 512$	7
7b	Leaky ReLU	$4 \times 4 \times 512$	7a
7	2DConv	$2 \times 2 \times 512$	6b
7a	Batch Norm	$2 \times 2 \times 512$	7
7b	Leaky ReLU	$2 \times 2 \times 512$	7a
8	2DConv	$1 \times 1 \times 512$	7b
8b	Leaky ReLU	$1 \times 1 \times 512$	8
9	2DConvT	$2 \times 2 \times 512$	8b
9a	Batch Norm	$2 \times 2 \times 512$	9
9b	Dropout	$2 \times 2 \times 512$	9a
9c	Concatenate	$2 \times 2 \times 1024$	9b,7b
9d	Leaky ReLU	$2 \times 2 \times 1024$	9c
10	2DConvT	$4 \times 4 \times 512$	9b
10a	Batch Norm	$4 \times 4 \times 512$	10
10b	Dropout	$4 \times 4 \times 512$	10a
10c	Concatenate	$4 \times 4 \times 1024$	10b,6b
10d	Leaky ReLU	$4 \times 4 \times 1024$	10c
11	2DConvT	$8 \times 8 \times 512$	10b
11a	Batch Norm	$8 \times 8 \times 512$	11
11b	Dropout	$8 \times 8 \times 512$	11a
11c	Concatenate	$8 \times 8 \times 1024$	11b,5b
11d	Leaky ReLU	$8 \times 8 \times 1024$	11c
12	2DConvT	$16 \times 16 \times 512$	11b
12a	Batch Norm	$16 \times 16 \times 512$	12
12b	Concatenate	$16 \times 16 \times 1024$	12a,4b
12c	Leaky ReLU	$16 \times 16 \times 1024$	12b
13	2DConvT	$32 \times 32 \times 256$	12b
13a	Batch Norm	$32 \times 32 \times 256$	13
13b	Concatenate	$32 \times 32 \times 256$	13a,3b
13c	Leaky ReLU	$32 \times 32 \times 512$	13b
14	2DConvT	$64 \times 64 \times 128$	13b
14a	Batch Norm	$64 \times 64 \times 128$	14
14b	Concatenate	$64 \times 64 \times 128$	14a,2b
14c	Leaky ReLU	$64 \times 64 \times 256$	14b
15	2DConvT	$128 \times 128 \times 64$	14b
15a	Batch Norm	$128 \times 128 \times 64$	15
15b	Concatenate	$128 \times 128 \times 64$	15a,1b
15c	Leaky ReLU	$128 \times 128 \times 128$	15b
16	2DConvT	$256 \times 256 \times 3$	15c
16	Tanh	$256 \times 256 \times 3$	16

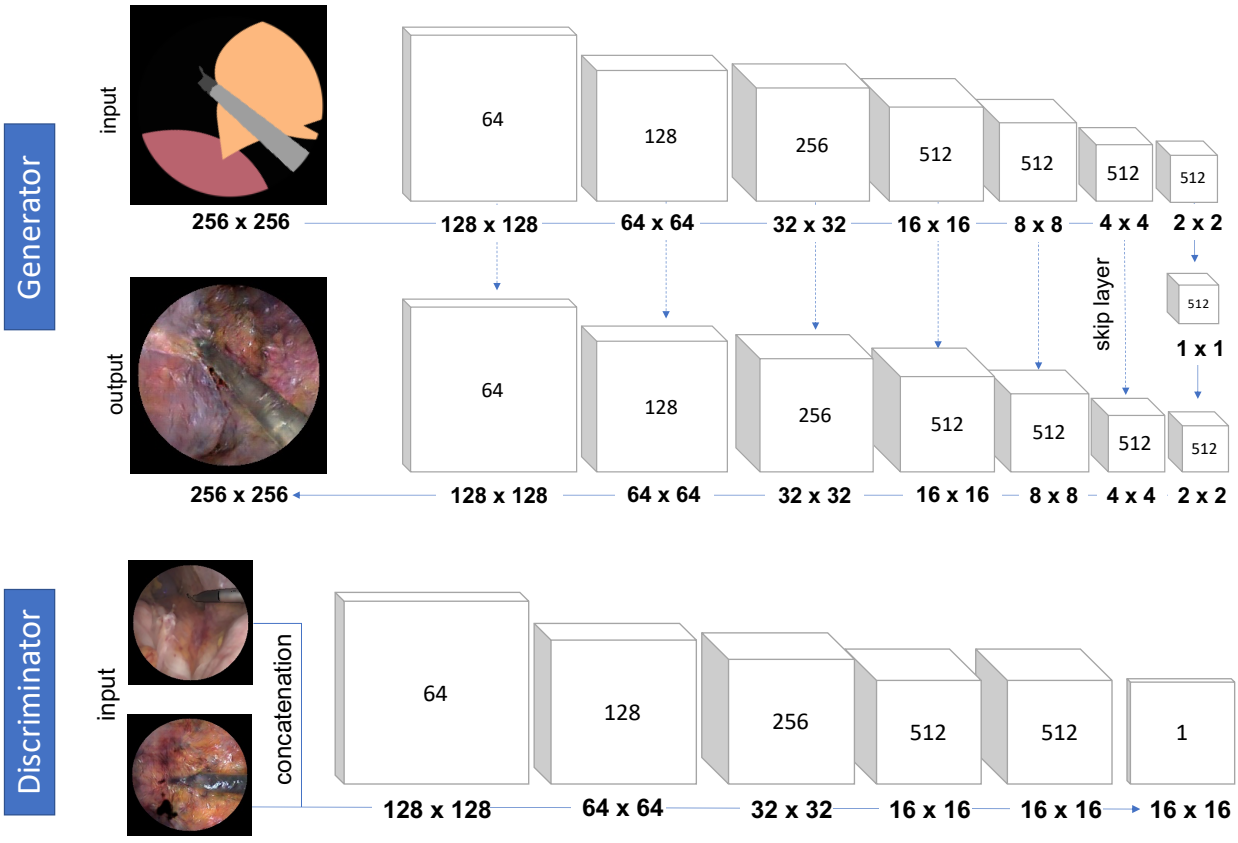

Fig. 3. Proposed *pix2pix* framework for laparoscopic-image generation. The (top) generator and (bottom) discriminator networks are shown. The feature-map size and the number of channels is reported, too.

Table 2. Architecture of the discriminator network

Id	Layer (type)	Output Shape	Connected to
1a	Input	$256 \times 256 \times 3$	
1b	Input	$256 \times 256 \times 3$	
1c	Concatenate	$256 \times 256 \times 6$	1a,1b
2	2DConv	$128 \times 128 \times 64$	1c
2a	Leaky ReLU	$128 \times 128 \times 64$	2
3	2DConv	$64 \times 64 \times 128$	2a
3a	Batch Norm	$64 \times 64 \times 128$	3
3b	Leaky ReLU	$64 \times 64 \times 128$	3a
4	2DConv	$32 \times 32 \times 256$	3b
4a	Batch Norm	$32 \times 32 \times 256$	4
4b	Leaky ReLU	$32 \times 32 \times 256$	4a
5	2DConv	$16 \times 16 \times 512$	4b
5a	Batch Norm	$16 \times 16 \times 512$	5
5b	Leaky ReLU	$16 \times 16 \times 512$	5a
6	2DConv	$16 \times 16 \times 512$	5b
6a	Batch Norm	$16 \times 16 \times 512$	6
6b	Leaky ReLU	$16 \times 16 \times 512$	6a
7	2DConv	$16 \times 16 \times 1$	6b
7b	Leaky ReLU	$16 \times 16 \times 1$	7

Our objective loss was then defined as:

$$G^t = arg \min_G \max_D L_{cGAN}(G, D) + \lambda_1 L_1(G) + \lambda_2 L_t(G) \quad (7)$$

where λ_1 is the overall regularizing weight, while λ_2 is the regularizing weight specific to the surgical tools.

3.4. Experimental settings

The *pix2pix* architecture was implemented and trained using Keras. The Adam optimizer was chosen with a learning rate of 2×10^{-3} for both the generator and discriminator. We empirically set the L_1 ($\lambda_1 = 50$) and L_t ($\lambda_2 = 100$) regularization parameters to reflect the fact that the semantic label map was accurate for the surgical tools, while it was rough for the abdominal tissues. Patch size in the discriminator was set to 70×70 pixels. Images were scaled to 256×256 pixels for memory constraints. The images were normalized in the range [-1,1]. We trained the *pix2pix* for 300 epochs using a batch size of 1. We stopped the training when satisfying visual results were achieved on a small test set of 10, unseen, images (this happened after ~100 epochs). Training took about 6 hours on a NVIDIA Quadro P6000 GPU.

3.5. Experimental Protocol

Currently, there is no established consensus, in the research community, on the best way to evaluate GANs [25, 11]. Indeed,

it is not trivial to obtain the definition of precise mathematical rules allowing one to tell whether an image is eligible for belonging to a given group or not. For instance, considering probability distributions between pixels or regions of the image by using statistical approaches might not be enough, as geometrical relationships between objects in the image should be taken into account. For these reasons, in this study we tested our approach by means of both quantitative measurements and qualitative visual assessments by surgeons and SDS experts over the generated laparoscopic images.

3.5.1. Human evaluation of generated images

A web platform was build, where physicians and experts were called to distinguish between real and synthetic endoscopic images. In order to generate completely unseen synthetic MIS images, a proper Python script was defined to generate semantic-label maps with tissues and surgical tools. The idea was to randomly generate polygons (representing the tissues and tools) over a black background. The trained *pix2pix* was used to translate these semantic-label maps into synthetic MIS images.

Two sets of 160 images (160 synthetic and 160 real) were considered. More precisely, we iteratively showed to each judge 20 pairs consisting of a real image and a generated one, extracted from the two sets with equal probability. The judge has to tell which one was the fake and which one was the real image. During each of the 20 trials, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) and the number of correct response were collected, where positive and negative refer to the generated and original images, respectively.

A total of 54 tests were collected by different users, divided in four categories (17 naive users, 19 non-expert surgeons, 13 expert surgeons, 5 technical developers). In order to guarantee a fair comparison, a postprocessing step (Gaussian blur) was added to the real images, in order to make their visual quality comparable to that of synthetic images. This was done to avoid the evaluators focusing on high-level details (e.g. instrument brand) to spot real images.

The one-way ANOVA test was performed to assess if any statistical difference between the groups of evaluators existed. Group normality and homoscedasticity assumptions were evaluated by means of the Shapiro-Wilk and Bartlett's test, respectively (significance level = 5%).

3.5.2. Quantitative image-quality assessment

In order to evaluate the similarity between the two groups (i.e., the real and generated images), the distributions of real and synthetic datasets were estimated by means of the Kernel Density Estimation (KDE) function, and their likelihoods were compared. Similar approaches were already applied in the context of GANs in [11] and [25]. In our approach, we computed the similarity between the two datasets estimating their distribution by means of the Kernel Density function, so that similar datasets should be represented by similar distributions. More in detail, the probability of the synthetic data is estimated by fitting a Gaussian Parzen window to the generated samples and reports the likelihood under this distribution. The bandwidth of the Gaussians is obtained by cross-validating the training data. In addition, for a visual comparison of the two groups, Principal Component Analysis (PCA) embedding representation of real and synthetic images was computed [28]. Several embedding dimension (d, from 10 to 100 with a 10-unit increments) were analysed without noticing any significant visual difference. Thus, $d = 100$ was used for the experiments.

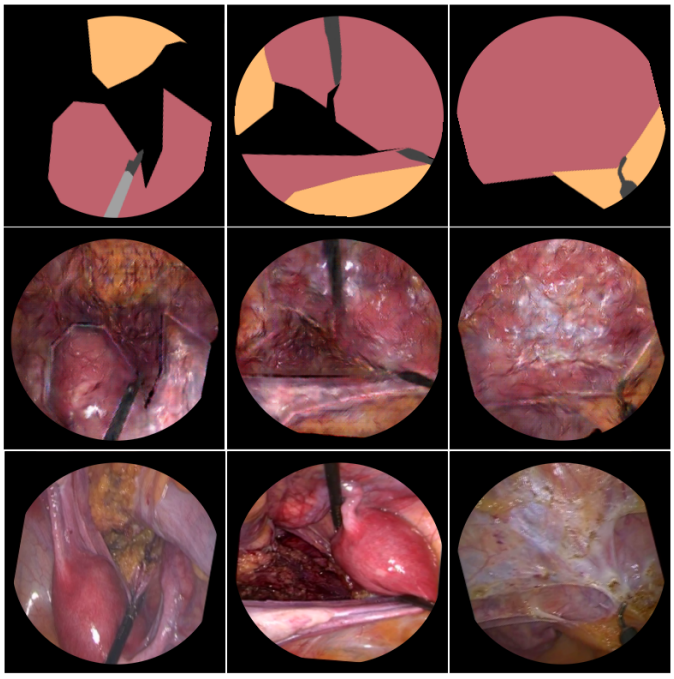

Fig. 4. Sample from the image translation results. Starting from the top row, the input semantic label map, the corresponding translation and the ground truth are shown.

3.5.3. Surgical tool segmentation

In order to quantitatively assess the informative content of the generated images, we evaluated whether the proposed method can serve as a data augmentation method for surgical-tool segmentation tasks. A total of 1600 images were generated by means of the proposed approach and used to train a standard U-Net architecture [27]. The network was trained for 50 epochs using binary cross-entropy as loss function and *Adam* optimizer (learning rate = 0.0001, batch size = 16). The model was implemented in Keras and trained using a NVIDIA Quadro P6000 GPU. After training, the best model was chosen as the one that minimized the loss on the validation set (30% of the whole dataset).

The model was finally tested on 40 images from the original MIS dataset. We calculate five evaluation indexes respectively: Søresen Dice Coefficient (*Dice*), Jaccard Similarity (*Jaccard*), Precision (*Precision*), Recall (*Recall*) and F1-score ($F1$):

$$Dice = \frac{2TP}{2TP + FN + FP} \tag{8}$$

$$Jaccard = \frac{TP}{TP + FN + FP} \tag{9}$$

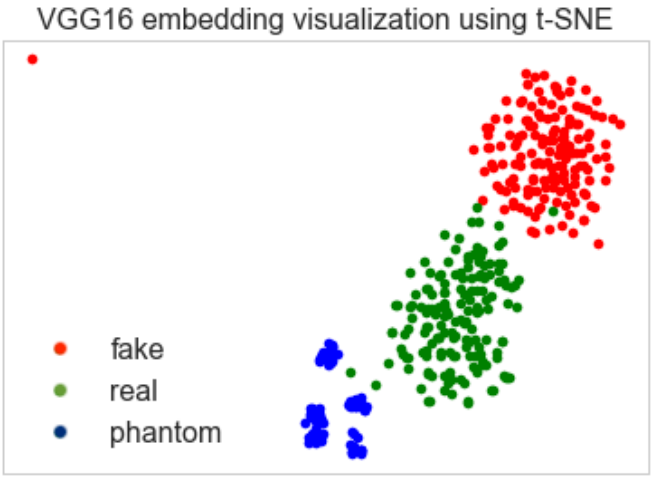

Fig. 5. Bi-dimensional Principal Component Analysis (PCA) embedding representation of real and synthetic images.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{12}$$

It is important to point out that the main goal of the study was not to achieve high segmentation performance; rather, we wanted to evaluate the informative content of the artificially generated images. For this reason, no further parameter tuning was performed to improve segmentation results.

4. Results

Samples from the generated images starting from the semantic label map with accurate surgical-tool segmentation and rough tissue segmentation can be seen in Figure 4. The average results from human evaluation are reported in Table 3. No differences were found between the assessor groups, except for the technical developers. The difficulty to differentiate between the real and generated images is well underlined by the low values of correct instances guessed by humans in the tests. Expert surgeons achieved the highest results. However, a similar performance was achieved by both naive users and non-expert surgeons. It is interesting to observe the high score achieved by the technical developer, which were able to spot the typical artefacts of synthetic images produced by GANs.

Values of 33.5765 and 33.5791 were obtained respectively by comparing the log-likelihoods for the 160 true and 160 synthetic data. This result suggests that the two groups are similar but not identical. Such similarity can be observed by the visual representation of the first two dimensions of the embedding (d = 100) illustrated in Figure 5. The two distributions do not overlap, suggesting that the *pix-to-pix* did not replicate images from the original set.

Average U-Net segmentation results are reported in Table 4: the synthetic images provided high enough informative content to let the model generalize on real instances, achieving encouraging results.

5. Discussion

In this work, we presented a deep-learning approach based on cGANs for generating realistic laparoscopic images with surgical instruments. Because annotated MIS datasets are difficult to obtain, due to high costs in terms of money and time required to perform the annotation, the proposed approach provides an effective method to augment MIS endoscopic datasets, with a view to generate SDS methods that rely on instrument-tool analysis. We took advantage of the *pix2pix* architecture for image-domain translation, incorporating instrument and (rough) tissue segmentation as priors. For *pix2pix* training, we exploited the publicly-available images of the EndoVis Intrument Segmentation and Tracking Challenge 2017.

Results were evaluated by means of quantitative and qualitative approaches. Fake images distribution was observed to be close to the real dataset, suggesting that our method was able to generate realistic contents without replicating the original images. In addition, we addressed the problem of image segmentation by training a well-known architecture used in the medical field, i.e. U-Net, using only synthetic images. Such simple training allowed to achieve promising segmentation results on real images, showing that the proposed method can be effectively applied for augmenting surgical datasets for segmentation tasks.

Naive users, surgeons and experts in GAN development were called to discriminate between real and newly generated images. In order to have a more detailed feedback, we also asked them to write down comments describing how did they tell the difference between real and synthetic images. Both naive users and surgeons were mostly unable to discriminate the images. It is worth to note that some of them raised concerns about the lower image quality with respect to that commonly available during their actual surgical practice, which, in fact, constitutes a limitation of our work. For a more fair comparison, indeed, we slightly modified real-images to avoid the evaluators focusing on high-frequency details (e.g. instrument brand) to discriminate between fake and real images. However, we found such processing not representing a limitation for the test, according to the high results and comments collected by technical developers and some of the surgeons. Technical developers (non-clinical), indeed, were significantly better at discriminating real and generated images. This may be probably attributed to their background, which made them able to to spot features relevant to GAN generation.

A second limitation of this study may be seen in the relatively small size of the training dataset. However, the lack of large and annotated publicly available datasets is a well known-problem in the SDS community [3]. This problem was addressed by performing linear transformations to the original available images. It is worth to note that methods for training image-to-image translation system that do not require paired examples exist. In this context, several experiments using CycleGAN [21] were performed but not reported in this paper. In particular, images generated by CycleGAN were observed to be only small

Table 3. Results of the human evaluation in terms of True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), number of correct response (Correct). Results are averaged across all users (all) and for each level of expertise (± standard deviation).

	count	No. Fake	No. Real	Correct	TN	TP	FN	FP
overall	54	10.19 ± 2.47	9.81 ± 2.47	7.94 ± 5.37	4.74 ± 3.8	3.2 ± 3.1	6.61 ± 3.21	5.44 ± 3.14
naive users	17	10.12 ± 2.6	9.88 ± 2.6	7.29 ± 5.51	3.94 ± 3.36	3.35 ± 3.18	6.53 ± 3.06	6.18 ± 3.21
non-expert surgeon	19	9.68 ± 1.6	10.32 ± 1.6	6.63± 4.42	3.63 ± 2.5	3.0 ± 2.96	7.32 ± 3.0	6.05 ± 2.68
expert surgeon	13	10.0 ± 2.55	10.0 ± 2.55	7.85 ± 5.49	5.15 ± 4.18	2.69 ± 3.45	7.31 ± 2.9	4.85 ± 3.51
technical developer*	5	12.8 ± 3.63	7.2 ± 3.63	**15.4 ± 1.67**	10.6 ± 3.71	4.8 ± 2.68	2.4 ± 2.61	2.2 ± 1.48

*number of correct responses significantly different from other groups

Table 4. Segmentation results in terms of Dice Coefficient (*Dice*), F1-score ($F1$), Jaccard Similarity (*Jaccard*), Precision (*Precision*) and Recall (*Recall*).

	Dice	$F1$	*Jaccard*	*Precision*	*Recall*
mean	0.71	0.97	0.59	0.81	0.68
std	0.22	0.02	0.22	0.18	0.26

variations of the same image (mode collapse problem) and we were not able to overcome it with the current implementation. Such results were probably related to a common issue of CycleGAN, which usually fails with tasks that require substantial geometric changes [21]. CycleGAN was successfully exploited in the majority of the approaches in the literature [22, 23, 24], which, however, focused on generating images with only one structure in the field of view (e.g., bronchi or colon). In this study, instead, we focused on images presenting higher variability, in terms of tissues in the field of view, surgical tools shape, size and pose, and illumination level, for which CycleGAN was challenging to train.

6. Conclusions

A laparoscopic image generation method was proposed in this study. The conditional generative adversarial network system *pix2pix* was exploited for generating synthetic images starting from an (accurate) surgical-tool and a (rough) abodminal-tissue label map. Achieved results suggest that the method can be effectively used for augmenting surgical datasets for segmentation tasks. As future work is concerned, we aim at improving the quality of our results by exploiting advanced generative models such as *pix2pixHD* [29], also dealing with unpaired datasets. In addition, despite our analysis being focused on images with surgical tools, the proposed framework could be easily extended to other object classes. Hence, we plan to investigate the performance of the proposed approach also in other anatomical fields, e.g., for generating images with placenta membrane in fetoscopic images to support surgeons with context awareness [30].

Acknowledgments

This study did not need any ethical approval. The authors declare no competing interests. The study was partially supported by the European Commission and the Regione Calabria within the program POR Calabria FESR-FSE 2014/2020, topic "ICT e Terziario Innovativo". We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P6000 GPU used for this research.

References

[1] J. H. Palep, Robotic assisted minimally invasive surgery, Journal of Minimal Access Surgery 5 (1) (2009) 1–7.

[2] M. Cianchetti, A. Menciassi, Soft robots in surgery, in: Soft Robotics: Trends, Applications and Challenges, Springer, 2017, pp. 75–85.

[3] L. Maier-Hein, S. S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, et al., Surgical data science for next-generation interventions, Nature Biomedical Engineering 1 (9) (2017) 691–696.

[4] D. Katić, C. Julliard, A.-L. Wekerle, H. Kenngott, B. P. Müller-Stich, R. Dillmann, S. Speidel, P. Jannin, B. Gibaud, LapOntoSPM: an ontology for laparoscopic surgeries and its application to surgical phase recognition, International Journal of Computer Assisted Radiology and Surgery 10 (9) (2015) 1427–1434.

[5] V. Penza, J. Ortiz, L. S. Mattos, A. Forgione, E. De Momi, Dense soft tissue 3D reconstruction refined with super-pixel segmentation for robotic abdominal surgery, International Journal of Computer Assisted Radiology and Surgery 11 (2) (2016) 197–206.

[6] S. Moccia, S. J. Wirkert, H. Kenngott, A. S. Vemuri, M. Apitz, B. Mayer, E. De Momi, L. S. Mattos, L. Maier-Hein, Uncertainty-aware organ classification for surgical data science applications in laparoscopy, IEEE Transactions on Biomedical Engineering 65 (11) (2018) 2649–2659.

[7] 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).

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88.

[9] L. Maier-Hein, T. Ross, J. Gröhl, B. Glocker, S. Bodenstedt, C. Stock, E. Heim, M. Götz, S. Wirkert, H. Kenngott, et al., Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 616–623.

[10] M. B. Lee, Y. H. Kim, K. R. Park, Conditional generative adversarial network-based data augmentation for enhancement of iris recognition accuracy, IEEE Access 7 (2019) 122134–122152.

[11] F. Calimeri, A. Marzullo, C. Stamile, G. Terracina, Biomedical data augmentation using generative adversarial neural networks, in: International Conference on Artificial Neural Networks, Springer, 2017, pp. 626–634.

[12] P. Andreini, S. Bonechi, M. Bianchini, A. Mecocci, F. Scarselli, Image generation by GAN and style transfer for agar plate image segmentation, Computer Methods and Programs in Biomedicine 184 (2020) 105268.

[13] T. Shen, C. Gou, F.-Y. Wang, Z. He, W. Chen, Learning from adversarial medical images for x-ray breast mass segmentation, Computer Methods and Programs in Biomedicine 180 (2019) 105012.

[14] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[15] E. Colleoni, S. Moccia, X. Du, E. De Momi, D. Stoyanov, Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers, IEEE Robotics and Automation Letters 4 (3) (2019) 2714–2721.

[16] M. Allan, S. Ourselin, D. J. Hawkes, J. D. Kelly, D. Stoyanov, 3-D pose estimation of articulated instruments in robotic minimally invasive surgery, IEEE Transactions on Medical Imaging 37 (5) (2018) 1204–1213.

[17] A. Rau, P. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, D. Stoyanov, Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy, International Journal of Computer Assisted Radiology and Surgery 14 (7) (2019) 1167–1176.

[18] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434 (2015).

[19] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, GAN-based synthetic medical image augmentation for increased cnn performance in liver lesion classification, Neurocomputing 321 (2018) 321–331.

[20] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, J. Barfett, Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2018, pp. 990–994.

[21] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[22] S. Mathew, S. Nadeem, S. Kumari, A. Kaufman, Augmenting colonoscopy using extended and directional CycleGAN for lossy image translation, arXiv preprint arXiv:2003.12473 (2020).

[23] M. Oda, K. Tanaka, H. Takabatake, M. Mori, H. Natori, K. Mori, Realistic endoscopic image generation method using virtual-to-real image-domain translation, Healthcare Technology Letters 6 (6) (2019) 214–219.

[24] A. Esteban-Lansaque, C. Sanchez, A. Borras, D. Gil, Augmentation of virtual endoscopic images with intra-operative data using content-nets, BioRxiv (2019) 681825.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[26] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).

[27] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[28] H. Abdi, L. J. Williams, Principal component analysis, Wiley interdisciplinary Reviews: Computational Statistics 2 (4) (2010) 433–459.

[29] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8798–8807.

[30] A. Casella, S. Moccia, E. Frontoni, D. Paladini, E. De Momi, L. S. Mattos, Inter-foetus membrane segmentation for TTTS using adversarial networks, Annals of Biomedical Engineering 48 (2) (2020) 848–859.