Medical Image Analysis 97 (2024) 103263

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Guided image generation for improved surgical image segmentation

Emanuele Colleoni [a,b,*,1], Ricardo Sanchez Matilla [b,1], Imanol Luengo [b], Danail Stoyanov [a,b]

[a] Medtronic Digital Surgery, 230 City Rd, EC1V 2QY, London, United Kingdom
[b] Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London (UCL), 43-45 Foley St, W1W 7TY, London, United Kingdom

ARTICLE INFO

Keywords:
Surgical image synthesis
Surgical image segmentation
Surgical robotics
Surgical vision

ABSTRACT

The lack of large datasets and high-quality annotated data often limits the development of accurate and robust machine-learning models within the medical and surgical domains. In the machine learning community, generative models have recently demonstrated that it is possible to produce novel and diverse synthetic images that closely resemble reality while controlling their content with various types of annotations. However, generative models have not been yet fully explored in the surgical domain, partially due to the lack of large datasets and due to specific challenges present in the surgical domain such as the large anatomical diversity. We propose Surgery-GAN, a novel generative model that produces synthetic images from segmentation maps. Our architecture produces surgical images with improved quality when compared to early generative models thanks to the combination of channel- and pixel-level normalization layers that boost image quality while granting adherence to the input segmentation map. While state-of-the-art generative models often generate overfitted images, lacking diversity, or containing unrealistic artefacts such as cartooning; experiments demonstrate that Surgery-GAN is able to generate novel, realistic, and diverse surgical images in three different surgical datasets: cholecystectomy, partial nephrectomy, and radical prostatectomy. In addition, we investigate whether the use of synthetic images together with real ones can be used to improve the performance of other machine-learning models. Specifically, we use Surgery-GAN to generate large synthetic datasets which we then use to train five different segmentation models. Results demonstrate that using our synthetic images always improves the mean segmentation performance with respect to only using real images. For example, when considering radical prostatectomy, we can boost the mean segmentation performance by up to 5.43%. More interestingly, experimental results indicate that the performance improvement is larger in the set of classes that are under-represented in the training sets, where the performance boost of specific classes reaches up to 61.6%.

1. Introduction

Robotic platforms are becoming a standard paradigm across hospitals, with the number of surgical procedures performed via Robotic Minimally Invasive Surgery (RMIS) increasing every year (Sheetz et al., 2020; Tsui et al., 2013). Despite RMIS benefits, which combine an intuitive and simplified interaction with the surgical site via telemanipulated robotic arms and stereo vision, much can still be done to boost the surgical experience on both the patient and surgeon sides. On this line, correctly and automatically identifying the patient's anatomy within the scene is a priority, when considering the harmful and long-term effects that unwanted damage of anatomical structures such as nerves or arteries can lead to. Following this, computer vision techniques are growing fast within the surgical field and can support surgeons with automatic anatomy detection (Madani et al., 2022). Learning meaningful representations for anatomy localization, however, demands a high number of annotated data that are difficult and expensive to obtain within the medical field (Kumar et al., 2021). Computer vision can answer these needs by providing models that generate synthetic data to be used as a means to support downstream tasks, such as surgical scene and anatomy segmentation. These models, known as generative models, can synthesize realistic images simply starting from a noise source (Karras et al., 2019) or from guide annotations, such as text, frame-level labels, sketches, or segmentation maps (Huang et al., 2022). Recent works on generative models identified normalization layers (Huang and Belongie, 2017) as a key operator for their architectures, carrying improvements in image quality and feature control when used to modulate the model weights (Dhariwal and Nichol, 2021) and giving the possibility to inject information at

* Corresponding author at: Medtronic Digital Surgery, 230 City Rd, EC1V 2QY, London, United Kingdom.
E-mail address: collee3@medtronic.com (E. Colleoni).
[1] These authors equally contributed to the article.

https://doi.org/10.1016/j.media.2024.103263
Received 24 February 2023; Received in revised form 30 May 2024; Accepted 27 June 2024
Available online 3 July 2024

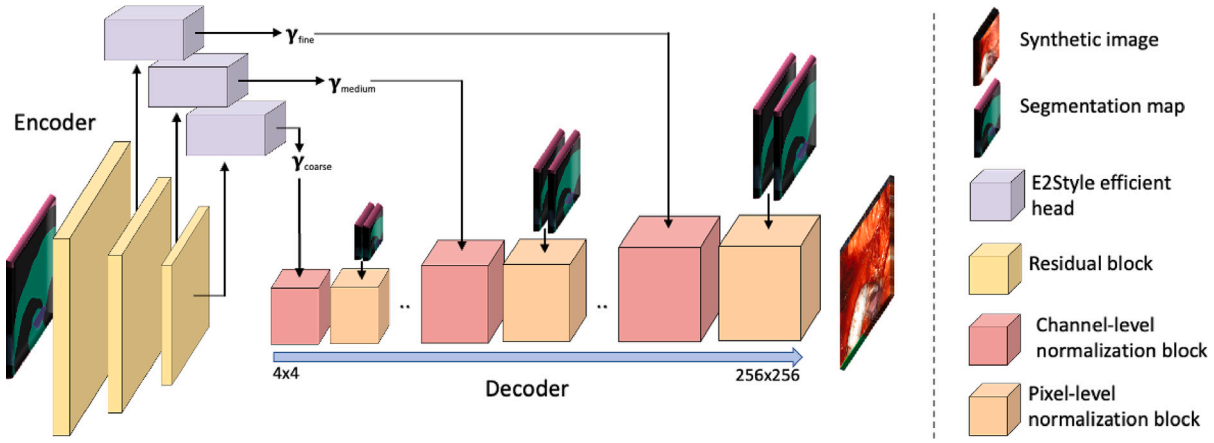

Fig. 1. Overview of the proposed SuGAN generator architecture. Given a *segmentation map* as input, the *encoder* embeds the essential information into a latent representation γ. Then, the *decoder* synthesizes a novel image using both latent representations as well as segmentation maps themselves.

different depths in the model (Park et al., 2019). Normalization layers also support noise injection at different levels in the model, which is shown to be important for synthetic image variability (Sushko et al., 2020; Karras et al., 2020a). This effect has also been confirmed in the medical field (Schonfeld and Veeravagu, 2023; Daroach et al., 2022). However, although research is moving fast in improving the quality of synthetic images, little effort has been put into using this data to improve downstream machine-learning models. While there are some works that investigate the use of simulation and synthetic data to improve image segmentation (Poucin et al., 2021; Fernandez et al., 2022), we believe that more investigation is needed to fully exploit the potential of data generation in the surgical domain. In this paper, we cover the problem of generating surgical images with high quality and adherence to the input map, and how to use them to boost the performance of semantic segmentation models. In summary, the main contributions of our work are:

- Surgery-GAN (SuGAN), a novel image synthesis architecture conditioned to segmentation maps that embraces recent advances in conditional and unconditional pipelines to support the generation of multi-modal (i.e., diverse) surgical images, and to prevent overfitting, lack of diversity, and cartooning, that are often present in synthetic images generated by state-of-the-art models. Particularly, we use channel- and pixel-level normalization blocks, where the former allows for realistic image generation and multi-modality through latent space manipulation, while the latter enforces the adherence of the synthetic images to their input segmentation map. This differs from state-of-the-art methodologies (Wei et al., 2022; Sushko et al., 2020), where these modules have only been used in isolation, often resulting in images of poor quality or not well following the input condition.
- Experiments on image generation, showing that SuGAN is able to synthesize realistic and diverse images in three surgical datasets of different sizes: cholecystectomies, nephrectomies, and prostatectomies. We perform careful comparison and analysis of our synthetic images with the ones from state-of-the-art generative models, both qualitatively and quantitatively.
- Extensive downstream experiments indicating that synthetic images can be used along with real images to boost semantic segmentation performance in the same three types of surgical procedures and using five different segmentation models, CNN- and transformers-based. Unlike other state-of-the-art image generation works (Park et al., 2019; Huang et al., 2022), we believe that it is important to validate the utility of synthetic images on downstream tasks since, as our experiments show, higher image quality may not necessarily translate to greater performance in these tasks.

2. Related work

Image generation pipelines can be generally classified as conditional or unconditional, depending on their training being performed with or without input labels such as segmentation maps, bounding boxes, or text, respectively. Among unconditional methods, that only use random noise as input, Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) have guided the field of image generation up until recent years, being able to produce detailed and realistic synthetic images from a variety of natural subjects, from faces of celebrities (Karras et al., 2019) to urban scenes (Wang et al., 2018). In GANs, a generator and a discriminator compete with each other on antagonistic tasks, one to generate data that cannot be distinguished from real ones, and the other to tell synthetic and real data apart. Among them, StyleGAN (Karras et al., 2019, 2020b,a) represents the first framework with the ability to produce realistic high-resolution images. Likewise, it opened the way to multi-modal image generation, where one can modify the appearance, colours, and luminance of the objects in the image while leaving their semantics unaltered. Other architectures besides GANs have then been designed, from variational autoencoders (VAE) (Higgins et al., 2017), to the recent diffusion models (Dhariwal and Nichol, 2021; Nichol and Dhariwal, 2021). In spite of the achieved outstanding image quality, unconditional approaches' lack of control over the generated images decreases their further utility to support downstream tasks. Conversely, conditional image-to-image translation (*cI2I*) models use a prior condition, such as text, segmentation maps, or sketches, to guide the creation of the synthetic images (Huang et al., 2022). On this track, Product-Of-Experts (POE)-GAN (Huang et al., 2022) combines text, sketches, and segmentation maps as prior conditions for image generation, while conditional-GAN (cGAN) (Isola et al., 2017) proposes a conditional approach where synthetic and real images are used by the discriminator along with their condition to learn a joint annotation-image mapping. Based on cGANs, SPatially Adaptive DEnormalization (SPADE (Park et al., 2019)) and Only Adversarial Semantic Image Synthesis (OASIS (Sushko et al., 2020)) are part of the state-of-the-art for conditional image synthesis, where segmentation maps are encoded as normalization parameters at different levels in the network to serve as a base for feature translation. Compared to unconditional models, cGAN-based models usually produce images of limited quality. A possible reason could be intrinsic to their formulation, where models focus on translating visual features according to the annotation, rather than evaluating the image at a global scale. In a similar direction, CollageGAN (Li et al., 2021) employs pre-trained, class-specific StyleGAN models to improve the generation of finer details on the synthetic images. Other approaches use a learnable encoder jointly with a frozen pre-trained StyleGAN to learn latent space mappings from segmentation

maps (Richardson et al., 2021; Wei et al., 2022). However, the quality of these models is bound to StyleGAN's capability to fully cover the training distribution. Moving away from GAN-based methods, the most recent advances in conditional image synthesis are establishing diffusion models as the new state-of-the-art in terms of image quality (Wang et al., 2022b,a), with stable and latent diffusion models dominating the field of text-to-image generation, but also achieving impressive results when conditioning images to other inputs, such as segmentation maps or bounding boxes (Rombach et al., 2022). In this work, however, we focused our experiments on GAN-based approaches as inference with diffusion models is currently too slow for generating large datasets. Moreover, training diffusion models require a large amount of computational resources compared to GANs. Although stable diffusion proposes to use an autoencoder to generate lower-dimensional encoded features, thus speeding up training and inference, we believe the improvement is still insufficient. Also, latent diffusion models have been tested mostly on large datasets such as Imagenet and COCO, thus putting a further constraint on the use of these methods with reduced data, as in the surgical context. Recently, studies have investigated the optimal use of simulation data for supporting segmentation (Poucin et al., 2021); however, among the described cI2I approaches, none use GAN-based synthetic images to improve downstream tasks. Considering applications in the medical field, a few attempts have been made to translate visual media generation in this environment. Research mostly focused on unpaired I2I (Colleoni et al., 2022; Pfeiffer et al., 2019; Rivoir et al., 2021), where models are trained to learn feature translation at a dataset level, while the previously discussed cI2I focus on image-level translation. In particular, most of these models require rich input conditions, such as images generated by high-quality virtual simulators, in order to correctly translate features. Conditional approaches have also been used to support depth estimation and anatomy segmentation in colonoscopy (Rau et al., 2019; Thambawita et al., 2022) as well as to increase surgical training realism (Engelhardt et al., 2019). In a similar fashion, cGANs have been employed as means to replace data for medical image classification (Kovalev and Kazlouski, 2019) and for MRI image segmentation (Fernandez et al., 2022). To the best of our knowledge, the work from Marzullo et al. (2021) represents the only attempt that investigates the use of cGANs for surgical image generation. Here, synthetic images have no further scope, such as to improve downstream-task models.

3. Proposed method

Let $\mathbf{x} \in \{0, 255\}^{W,H,3}$ be an RGB image with width W, height H, and 3 colour channels, and let $\mathbf{y} \in \{0, C-1\}^{W,H}$ be a pixel-wise segmentation map with C different semantic classes. Let $G(\cdot) : \mathbf{y} \rightarrow \mathbf{x}_s$ be a generator that, given as input a segmentation map $\mathbf{y}$, generates a synthetic image, $\mathbf{x}_s$. Our goal is to design $G(\cdot)$ so that it can generate realistic multi-modal images conditioned to an input annotation $\mathbf{y}$. This directly translates to the problem of preserving the image *content* while varying the *style*, where content refers to semantic features in the image, namely objects' shape, and location, while style relates to the appearance of the object such as colours, texture, and luminance. Next, we describe the proposed model, named SuGAN, as well as presenting an overview of how we leverage SuGAN to generate realistic and multi-modal synthetic images to boost the performance of segmentation models on real test data.

3.1. Surgery-GAN

We propose SuGAN, an adversarially trained model that can produce novel, realistic, and diverse surgical images conditioned to semantic segmentation maps. We use an encoder–decoder-discriminator structure, where the task of the encoder is to extract the essential information from the input segmentation map while the decoder generates synthetic images from the input segmentation maps and the output of the encoder. Finally, a discriminator determines whether generated synthetic images are real or synthetic. This is required to enable adversarial training. An overview of the proposed architecture is illustrated in Fig. 1. Next, we describe the specifics of each module.

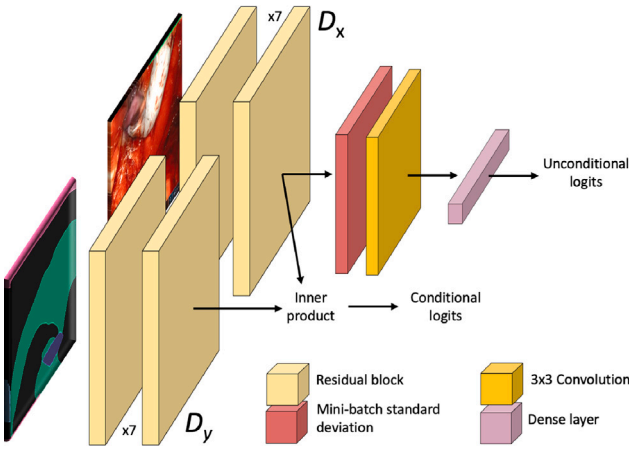

Fig. 2. Overview of the proposed discriminator architecture. First, two feature extractors D_x and D_y encode images and segmentation maps into a feature space, respectively. Then, conditional and unconditional logits are calculated via inner product and mini-batch standard deviation followed by convolutional layers, respectively. Unconditional logits are used to enforce the model to generate realistic synthetic images. Conditional logits are used to enforce that the synthetic image satisfies the segmentation condition.

Encoder. Following conditional image generation pipelines (Isola et al., 2017), we use an encoder $E(\cdot)$,

$$E(\cdot) : \mathbf{y} \rightarrow \gamma, \tag{1}$$

to project segmentation maps $\mathbf{y}$ into a latent feature $\gamma \in \mathbb{R}^{M \times 512}$ where M is the number of feature vectors. Our encoder follows E2Style architecture (Wei et al., 2022). Specifically, a map $\mathbf{y}$ is first processed into three consecutive residual blocks and the output of each block is further refined by E2Style efficient heads, which consists of an average pooling layer followed by a dense layer, to obtain γ. As such, γ is composed of three sets of latent vectors, each one produced by a different head and controlling a different set of features in the final synthetic image, namely coarse (γ_{coarse}), medium (γ_{medium}), and fine features (γ_{fine}). Although not novel "per se", the E2Style encoder was developed in the context of GAN inversion, where it was trained jointly with a frozen StyleGAN to find a bijective relationship between latent vectors and synthetic images. To the best of our knowledge, this represents a first attempt to use this module for conditional image synthesis with a trainable decoder.

Decoder. We propose a novel decoder architecture that supports image generation conditioned to segmentation maps. The decoder architecture $D(\cdot)$ is defined as,

$$D(\cdot) : (\gamma, \mathbf{y}) \rightarrow \mathbf{x}_s, \tag{2}$$

and takes γ and $\mathbf{y}$ as inputs to generate the synthetic image $\mathbf{x}_s$. Our decoder is built by sequentially nesting two blocks of normalization layers: the first is a pixel-level normalization block (Park et al., 2019), where a normalization parameter (mean and standard deviation) for each feature pixel is computed by processing the input segmentation map with 3x3 convolutional layers and then used to modulate sets of features at different depths in the model. The second is a channel-level normalization block (Karras et al., 2020a), which is composed of two sets of 3×3 convolutional layers whose weights are modulated using two 512 latent vectors from the encoder as input. In this setting, each filter in the 3×3 convolution is first modulated to have variance equal to 1 and then demodulated by multiplying it

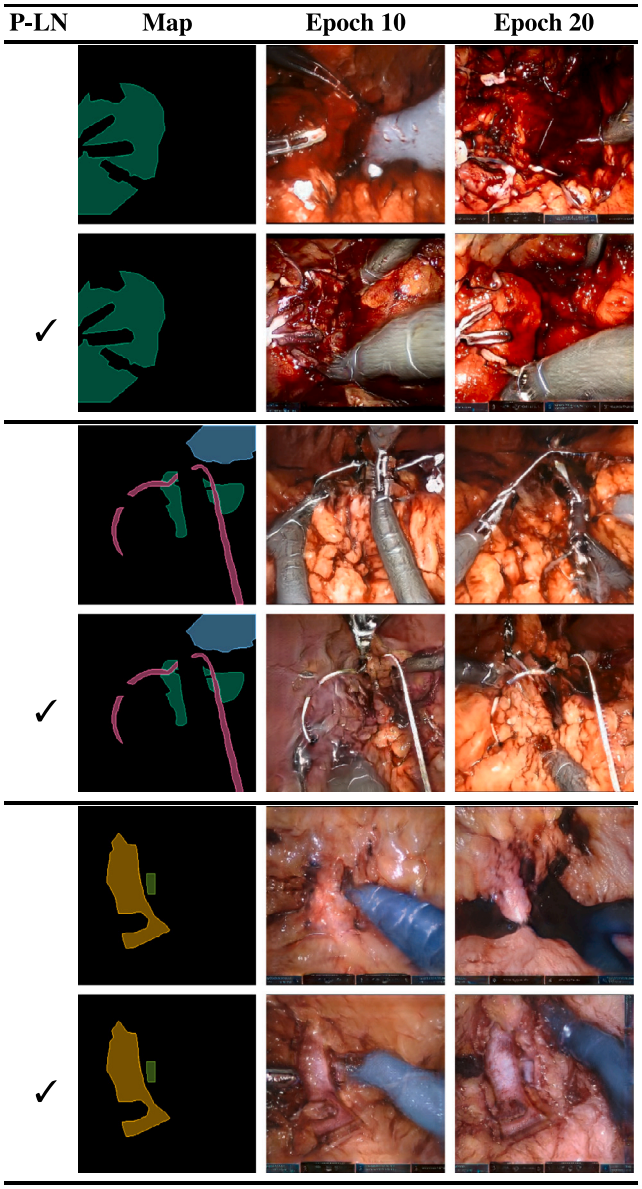

Fig. 3. Sample synthetic images generated by the proposed SuGAN with and without pixel-level normalization (P-LN) blocks at training epochs 10 and 20 in Partial Nephrectomy dataset (■ kidney, ■ liver, ■ renal vein, ■ renal artery, ■ ignore, and ■ background). Input maps are *from-train* set. Without P-LN, the model struggles to generate images that adhere to the input map, particularly when following the map boundaries (e.g. kidney in 1st and 2nd rows or renal artery in 5th and 6th rows) or when dealing with thin or small class instances (e.g. tools/background in 1st and 2nd rows, ignore in 3rd and 4th rows and renal vein in 5th and 6th rows).

by one element in the latent vector. For this reason, all layers inside a channel-level normalization block have 512 channels. To the best of our knowledge, this represents the first attempt to merge these two typologies of normalization layers into a single architecture, thus taking advantage of all their capabilities. These include the precise matching of the generated image with the input segmentation map via pixel-level normalization layers while allowing for the separability and manipulability of content and style features in the synthetic image thanks to channel-level normalization.

Discriminator. We use a projection discriminator (Miyato and Koyama, 2018) to support adversarial training. The full discriminator is composed of two feature extractors D_x and D_y defined as

$$D_x(\cdot) : \mathbf{x} \rightarrow \mathbf{d}_x, \tag{3}$$

and

$$D_y(\cdot) : \mathbf{y} \rightarrow \mathbf{d}_y, \tag{4}$$

where $\mathbf{d}_x$, $\mathbf{d}_y \in \mathbb{R}^{512\times4\times4}$ are features respectively extracted from (synthetic or real) images $\mathbf{x}$, and segmentation maps $\mathbf{y}$. Differently from its original formulation (Miyato and Koyama, 2018), the architecture of each discriminator module follows the one proposed in StyleGAN (Karras et al., 2019), consisting of several residual blocks in series. Then, we use the extracted features $\mathbf{d}_x$ and $\mathbf{d}_y$ to determine whether $\mathbf{x}$ is real or synthetic and whether the image satisfies the segmentation condition or not. To determine whether the image is real or synthetic, we compute the unconditional logits $\mathbf{p}_x$ as

$$\mathbf{p}_x = \alpha(\mathbf{d}_\mathbf{x}), \tag{5}$$

where $\alpha(\cdot)$ is a mini-batch standard deviation layer followed by a convolutional and a dense layer. To determine whether the image follows the segmentation map, we compute the conditional logits $\mathbf{p}_{x,y}$ as

$$\mathbf{p}_{x,y} = \langle \mathbf{d}_x , \mathbf{d}_y \rangle, \tag{6}$$

where $\langle \cdot, \cdot \rangle$ is the inner product operator. Please refer to Fig. 2 for an overview of our model's discriminator.

Training procedure. We use a conditional adversarial loss as the main objective function for training our model. In this setting, our generator and discriminator are trained on antagonistic losses defined as:

$$L(G, D, \mathbf{x}_s, \mathbf{y})_{Gen} = S(-\mathbf{p}_{x_s}) + S(-\mathbf{p}_{x_s,y}) \tag{7}$$

$$L(G, D, \mathbf{x}_r, \mathbf{y})_{Disc} = S(\mathbf{p}_{x_s}) + S(\mathbf{p}_{x_s,y}) + S(-\mathbf{p}_{x_r}) + S(-\mathbf{p}_{x_r,y}) \tag{8}$$

where $\mathbf{x}_s$ and $\mathbf{x}_r$ refer to synthetic and real images respectively and S is the softplus function defined as

$$S(a) = log(1 + e^a), \tag{9}$$

where a is the input logits to the function. Along the adversarial loss, we also used a lazy R1 regularization objective to train the discriminator, as defined in Karras et al. (2019):

$$L_{R_1} = \|\nabla W_D\| \tag{10}$$

where ∇W_D is the gradient of the discriminator weights. L_{Gen} and L_{Disc} are computed and applied in an alternative fashion for the generator and the discriminator.

Design process. Next, we discuss the design process that helped us reach the previously presented SuGAN implementation. Initially, we designed our model's architecture with an encoder–decoder structure. Inspired by recent StyleGAN's achievements in terms of image quality, and different from the final formulation described in , we tried a pre-trained, frozen StyleGAN model, where channel-level normalization blocks represent the standard building block. This model was trained using *LPIPS* (Zhang et al., 2018) and L_2 losses between target and synthetic images with no adversarial loss, thus without discriminator. However, results showed that this setting was sub-optimal, with *LPIPS* and L_2 losses imposing an excessive constraint over the training procedure and leading to overfitting and poor visual quality. Next, we translated our framework towards a conditional adversarial approach. As such, we made the StyleGAN decoder learnable during training, thus not requiring a pre-training step, and we introduced the discriminator as described in to support adversarial training. For further details on how the adversarial loss was defined, please refer to Miyato and Koyama (2018). After these modifications, visual quality and realism improved, but the model showed to struggle in generating

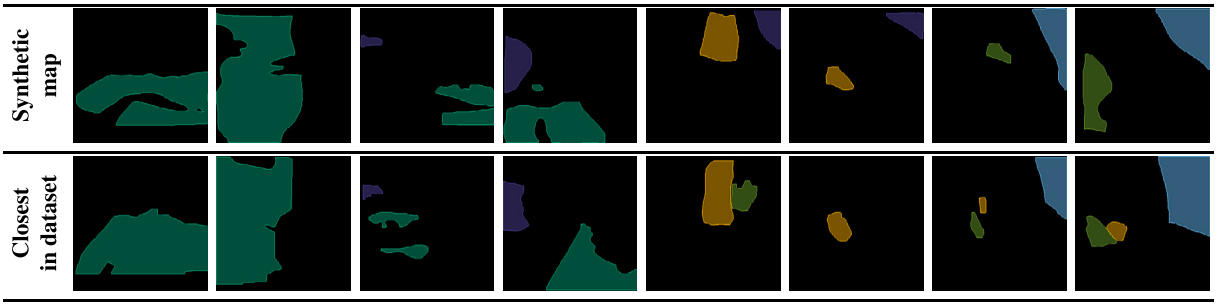

Fig. 4. Sample synthetic maps from a trained StyleGAN2 model. The produced maps are varied, both in relation to each other (1st row) and to the training ones, for which the closest map in terms of Intersection over Union is displayed (2nd row). Maps from **PN** dataset (■ kidney, ■ liver, ■ renal vein, ■ renal artery, ■ spleen, and ■ background).

Table 1
Summary statistics of the datasets used in our experiments: CholecSeg8k (LC) (Hong et al., 2020), Robotic Partial Nephrectomy (PN), and Robotic Prostatectomy (RP). # indicates the number of items.

Dataset	# videos				# frames				# classes
	Train	Val	Test	Total	Train	Val	Test	Total	
LC	14	–	3	17	7,360	–	720	8,080	13
PN	115	9	13	137	43,795	2,954	5,456	52,205	6
RP	1549	85	167	1801	166,152	10,194	18,047	194,393	11

images properly conditioned to the input segmentation map. Consequently, we embedded pixel-level normalization (Park et al., 2019) blocks between channel-level normalization ones to inject semantic maps directly into the model at different depths, leading to SuGAN, our final configuration. A visual example of the benefits introduced by using pixel-level normalization within channel-level ones is provided in Fig. 3. As we will show in the next sections, the proposed model is able to control content features via pixel-level normalization layers while supporting multi-modal image generation by manipulating latent vectors at inference stage. Note that removing channel-level normalization blocks would recover SPADE implementation, for which comparison experiments will be described in the next sections.

3.2. Generating new data with SuGAN

Channel-level normalization blocks can support multi-modal image generation via latent space manipulation, giving our model the capability of generating images with different styles while maintaining unaltered the content defined by the input segmentation map. To produce multi-modal images we use a style randomization procedure. Once the model is trained, we first encode all segmentation maps from the training dataset into their latent space representation and we fit a multivariate Gaussian distribution over each feature vector γ_i, $i \in [0, M]$, where $M = 14$ for our standard SuGAN configuration. At inference time, given an input segmentation map $\mathbf{y}$, we encode it into the latent space $\gamma = E(\mathbf{y})$ and we substitute the last $m \in [0, 13]$ feature vectors by sampling from the multivariate distributions: $\gamma \rightarrow \dot{\gamma}$. Finally, we produce a synthetic image $\mathbf{x}_s = D(\dot{\gamma}, \mathbf{y})$. Note that sampling multiple times from the multivariate distribution will generate different $\dot{\gamma}$, and thus multiple modalities while leaving the content unchanged.

We propose a second approach to generate original images using SuGAN based on semantic map generation. A class-conditional StyleGAN2 is first trained on training segmentation maps, allowing the generation of novel semantic views of the surgical scene. We choose this model as one of the most reliable in terms of image quality and variability, although any generative model could be used for this purpose. The hyperparameters for training our StyleGAN2 model match the default ones from Karras et al. (2020a). Once trained, an infinite number of segmentation inputs can be produced and fed into SuGAN to generate new labelled images. Sample-generated maps are presented in Fig. 4, showing that diverse and novel semantic views can be produced. The semantic map generation is performed by simply feeding class labels and noise into the trained StyleGAN2. An example of generated input maps along with synthesized images is presented in Fig. 5.

4. Validation

4.1. Experimental setup

Datasets. We validate the performance of the proposed model in three datasets: CholecSeg8k (LC), Robotic Partial Nephrectomy (PN), and Robotic Prostatectomy (RP). LC (Hong et al., 2020) is a public dataset of laparoscopic cholecystectomy surgeries focusing on the resection of the gallbladder. The dataset is composed of 8080 frames (17 videos) from the widely used Cholec80 (Twinanda et al., 2016) dataset. LC provides pixel-wise segmentation annotations of 8 classes,[2] including background, five anatomical classes, and two surgical instruments. PN is a private dataset of robotic partial nephrectomy surgeries focusing on the resection of the kidney. The dataset is composed of 52,205 frames from 137 videos. PN provides pixel-wise segmentation annotations of five anatomical classes and a background class which also includes surgical instruments.[3] RP is a private dataset of robotic prostatectomy surgeries for the resection of the prostate. The dataset is composed of 194,393 frames from 1801 videos. RP provides pixel-wise segmentation annotations of 10 anatomical classes and a background class which does not include any surgical instrument.[4] Table 1 shows further dataset statistics. In PN and RP datasets we target the generation of surgical images of anatomical structures and disregard the presence of surgical instruments, as instrument annotations are not provided. For LC, where instrument annotations are provided, we focus on the generation of anatomical structures as well as surgical instruments. We do not use validation images to train our model, thus relying only on training images.

Performance measures. We assess the performance of the proposed algorithm on two tasks. First, we evaluate image quality and diversity of synthetic images using a variation of Fréchet Inception Distance (FID) (Heusel et al., 2017). This metric is widely used for

[2] LC dataset classes: *background*; five anatomical structures: *abdominal wall*, *liver*, *gastrointestinal tract*, *fat*, *gallbladder*; and two surgical instruments: *grasper*, and *hook*

[3] PN dataset classes: *background*, *kidney*, *spleen*, *liver*, *renal vein*, and *renal artery*.

[4] RP dataset classes: *background*, *bladder*, *dorsal venous complex*, *vas deferens and seminal vesicles*, *prostate*, *foley catheter*, *endopelvic fascia*, *pubic bones*, *urethra*, *rectum*, and *neurovascular bundle*.

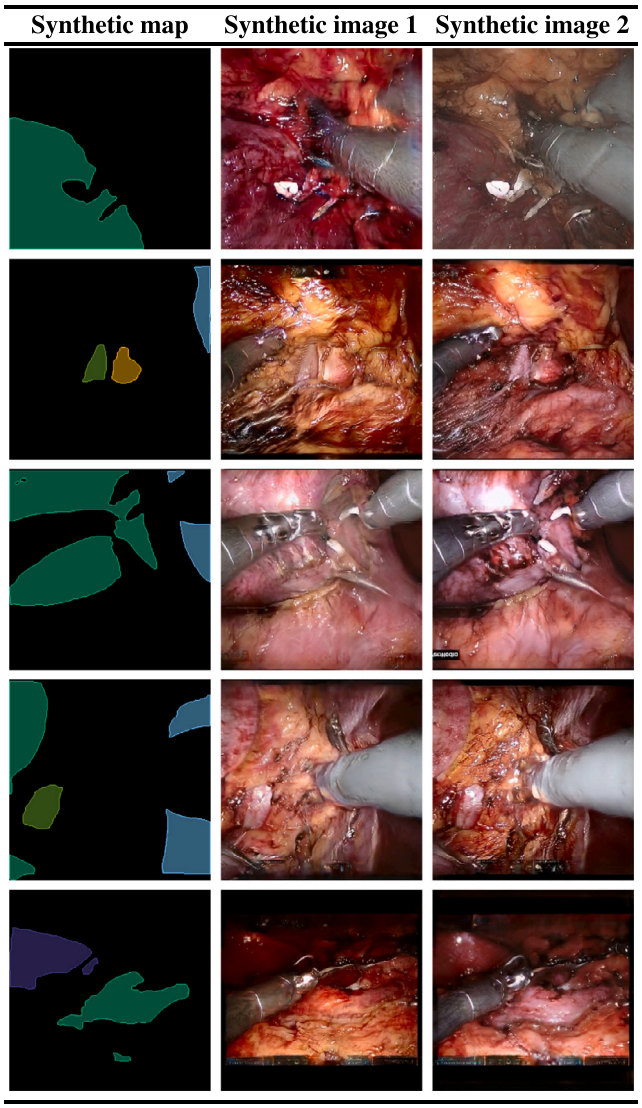

Fig. 5. Sample synthetic images generated by the proposed SuGAN using synthesized input maps from a trained StyleGAN2 (Karras et al., 2020a) model. StyleGAN2 trained with Partial Nephrectomy training maps (■ kidney, ■ liver, ■ renal vein, ■ renal artery, ■ spleen, and ■ background).

assessing image quality and diversity of synthetic images as it captures the similarity between synthetic and real data at a dataset level. Notably, we use FID infinite (Chong and Forsyth, 2020) in view of recent studies that show that FID is an intrinsically biased metric. In the following, we refer to FID infinite simply as FID. Second, we evaluate the performance of segmentation models per each class and image using Intersection over Union (IoU) $IoU_c^i = \frac{\bar{\mathbf{y}}_c^i \cap \mathbf{y}_c^i}{\bar{\mathbf{y}}_c^i \cup \mathbf{y}_c^i}$, where $\mathbf{y}_c^i$ is the annotated segmentation map and $\bar{\mathbf{y}}_c^i$ is the segmentation map predicted by the model for class c on image i. Per-class IoU is computed by averaging across images as $IoU_c = \frac{1}{N}\sum_{i=1}^{N} IoU_c^i$, where N is the total number of images. In addition, mean Intersection over Union (mIoU) aggregates scores across classes and reports a single performance value, calculated as $mIoU = \frac{1}{C}\sum_{c=1}^{C} IoU_c$ where C is the total number of classes in the dataset.

Algorithms under comparison. For the task of image generation, we compare SuGAN against two state-of-the-art image generation models, namely SPADE (Park et al., 2019) and OASIS (Sushko et al., 2020). We use these models as they represent the gold standard in generating high-quality conditional synthetic images from semantic segmentation annotations, and because their code is publicly available. We could not compare against other state-of-the-art models such as POE-GAN and Collage-GAN as their code is not publicly available at the time of submission. For the task of semantic segmentation, we compare five different segmentation models, three convolutional-based models, namely DeepLab (Chen et al., 2017), HRNet32 and HRNet48 (Wang et al., 2020), and two transformer-based models, namely Swin Transformer Small and Swin Transformer Base (Liu et al., 2021).

Implementation details. Regarding SuGAN, we follow StyleGAN2 training settings (Karras et al., 2020a), using Adam as an optimizer with a learning rate of 0.0025 and training for up to 50 epochs. The batch size is 16. Most experiments are performed with images at 256 × 256 resolution, but we also show that our model can successfully produce images of higher resolution (512 × 512). As augmentation, we use only vertical and horizontal flips as well as 90° rotations in order to not introduce black edges. Additionally, we use all non-leaking augmentations from Karras et al. (2020a) to stabilize the training. We use a class-balanced sampler. In the definition of our latent space (Section 3.1), we set $\gamma_{coarse} = \gamma_{0-6}$, $\gamma_{medium} = \gamma_{7-10}$ and $\gamma_{fine} = \gamma_{11-14}$. We choose these indexes grouping for our latent space experimentally. We train SPADE and OASIS using their default settings. We train SPADE along with a VAE to support multi-modal generation, following a variant of its original formulation (Park et al., 2019). Once SuGAN is trained, we generate synthetic images as described in Section 3.2 with m=5 for PN and RP and m=10 for LC. The number of feature vectors was chosen after experimental validation as the best trade-off between style randomization and reduction of artefacts arising from content modification. Regarding the implementation details for segmentation tasks, we train all models for 25 epochs using AdamW optimiser, with a OneCycle scheduler that reaches the maximum learning rate of 0.0005 after 1.25 epochs. We run inference using models at the last epoch for all experiments. To enable a fair comparison between models trained with real data only and real and synthetic data, all models are trained using the same batch size, number of steps per epoch, and number of epochs. The batch size is set to 32. When training only using real images all 32 samples are real. However, when training using real and synthetic, 16 real images are randomly sampled from the full training dataset and 16 synthetic images are randomly sampled from the generated synthetic dataset and concatenated within the same batch. Note that, in both cases, we use all the available training data. We believe these training settings can effectively expose the role of using synthetic data while discarding the possibility of different performances deriving from a different number of training steps or of available real images. We use a carefully crafted augmentation for all segmentation experiments, with images being resized to 256 × 256 resolution, maintaining the aspect ratio, performing random rotations between −20 and 20 degrees, applying 25% of the times motion blur, and modifying the hue and saturation by a random value between −10 and 10. Real images are scaled randomly between 0.7 and 2.0 times, while synthetic images are not scaled. This set of augmentations is the result of several prior experiments and it represents an empirical gold standard for training our baseline segmentation models. All training procedures are repeated five times with different seeds to reduce stochasticity in our results. For LC, we sample a third of the training dataset in each epoch using a random sampler, while for PN and RP we sample 12,000 images in each epoch with a class-balanced sampler. An empirical study showed that using a balanced sampler, when possible, has a significant impact on the performance boost carried by the use of synthetic images in the training pipeline. We believe this is expected: SuGAN effectively generates diverse images for unrepresented classes and using a balanced sampler ensures fair representation of new data, while a shuffling sampler may undermine SuGAN's contributions by maintaining minority class representation, given a fixed number of training steps.

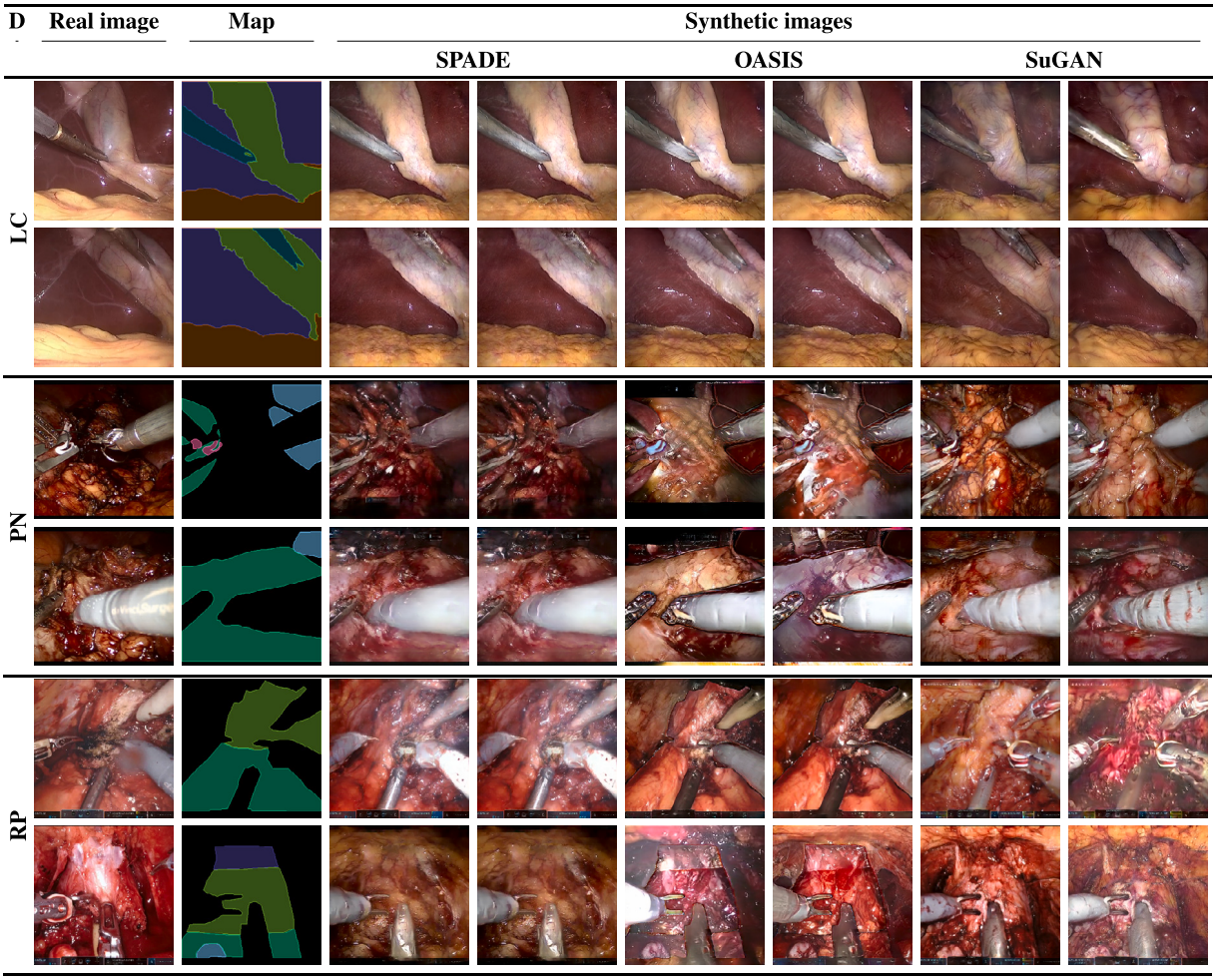

Fig. 6. Sample synthetic images generated *from-test* maps using SPADE, OASIS, and the proposed SuGAN in Laparoscopic Cholecystectomy (LC: liver, gallbladder, fat, and grasper), Partial Nephrectomy (PN: kidney, liver, and ignore), and Radical prostatectomy (RP: bladder, dorsal venous complex, vas deference and seminal vesicles, and prostate) datasets (D). background in all datasets.

4.2. Synthetic image generation

We evaluate and compare the quality and diversity of the images generated by the proposed SuGAN with the ones generated by SPADE and OASIS. Following previous works (Park et al., 2019; Huang et al., 2022; Sushko et al., 2020), we use only real segmentation maps as inputs for this experiment. Once the generative models are trained, we generate two sets of synthetic images for each model and dataset, namely *from-test* and *from-train*. *From-test* set is generated using segmentation maps from the test set without style randomization. *From-train* set is produced using segmentation maps from the training set as input and randomizing the style 10 times, thus creating 10 times larger training sets. We choose to shape the datasets in such a way as to test the models' capabilities in handling maps that were not used during training (*from-test*) as well as being able to use synthetic images for segmentation models training without leaking information from test-set ground-truth (*from-train*). Also, *from-train* dataset could be useful to highlight overfitting, where the model falls into generating mono-modal images that resemble original training ones.

In Table 2 we report the FID obtained comparing *from-test* images from SPADE, OASIS and SuGAN to real ones. The lower the FID, the most similar the synthetic data distribution is to the real one, thus indicating an overall good performance and generalization ability of the generative model. Also, note that the absolute value of FID is dataset-dependent, thus it should be compared within the same dataset, while inter-dataset FID comparison is not meaningful. In general and as can be appreciated in Fig. 6, both SuGAN and OASIS models manage to produce original multi-modal images, while SPADE seems constrained to mono-modal generation. Regarding LC dataset, SPADE and OASIS obtain the worst FID, 137.99 and 115.08, respectively, while our model achieved an FID of 107.70. We believe this result reflects a superior performance of SuGAN in terms of image quality and diversity, as it can be appreciated in Fig. 6, 1st and 2nd rows, where SuGAN is the only model that can generate diverse images on LC when using *from-test* input maps. Regarding PN dataset, OASIS, closely followed by SuGAN, obtains the best FID, indicating a better generalization capability compared to SPADE. OASIS, however, consistently generates cartoon-like instance blending within images (Fig. 6, 3rd and 4th rows). This is not the case with SuGAN, whose discriminator seems to discard such images from the synthetic image distribution. SPADE obtains the worst FID in the same dataset and consistently shows only mono-modal images, thus showing inferior generation capabilities compared

Table 2
Image generation results. The input segmentation maps used to synthesize the images are *from-test* set. We compare SPADE, OASIS, and the proposed SuGAN in three different datasets LC, PN, and RP in terms of FID (↓). Bold indicates the best result.

Dataset	SPADE	OASIS	SuGAN
LC	137.99	115.08	**107.70**
PN	34.33	**27.10**	28.38
RP	22.35	11.87	**8.03**

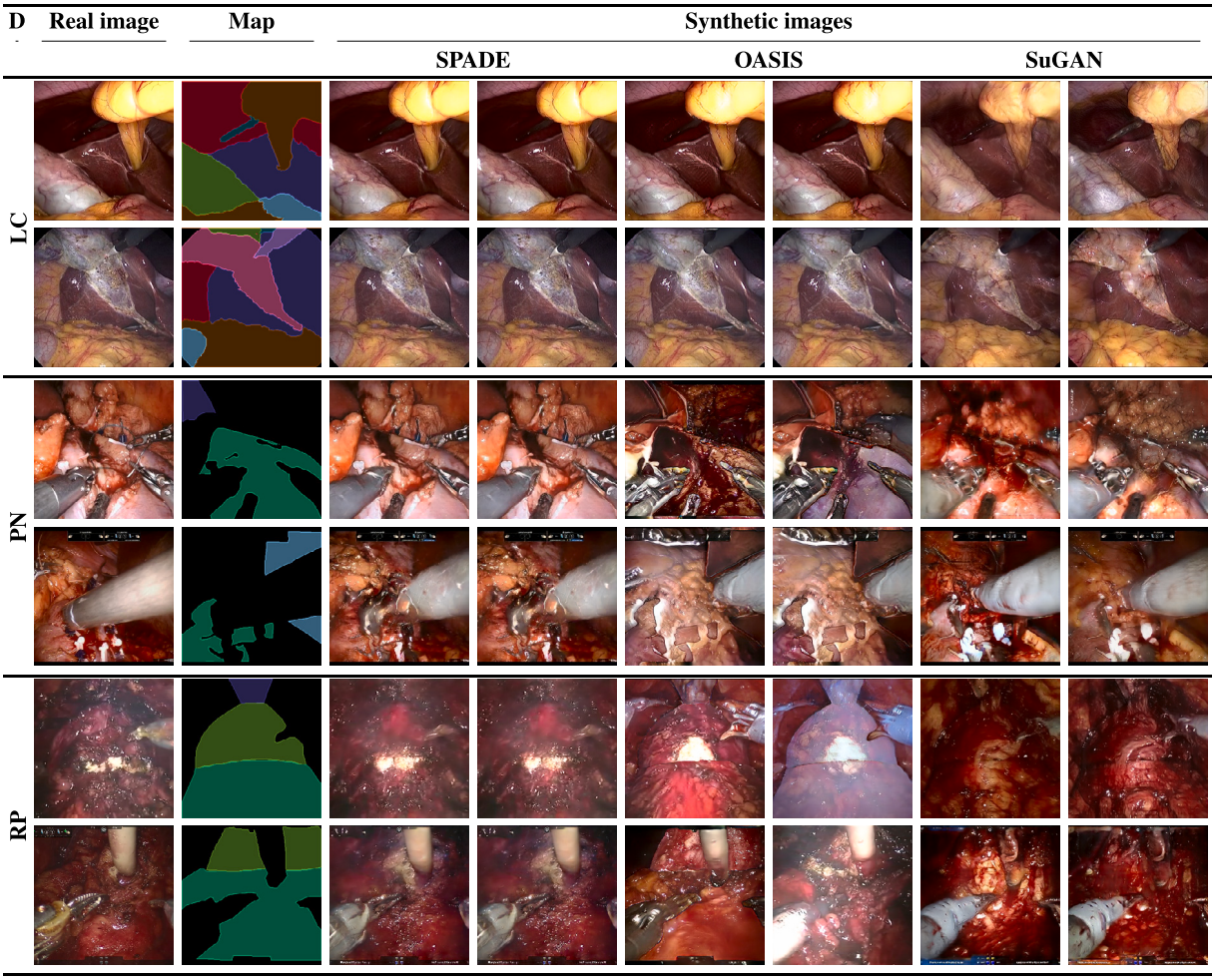

Fig. 7. Sample synthetic images generated from maps *from-train* set using SPADE, OASIS, and the proposed SuGAN in Laparoscopic Cholecystectomy (LC: liver, gallbladder, abdominal wall, fat, grasper, connective tissue, and L-hook electrocautery), Partial Nephrectomy (PN: kidney, spleen, liver) and Radical prostatectomy (RP: bladder, dorsal venous complex, and prostate) datasets (D). background in all datasets. Note that images generated with SPADE are often overfitted to the real images; OASIS might generate images with artefacts; while SuGAN generates realistic, non-overfitted, and diverse images.

to SuGAN and OASIS. Finally, on RP dataset, SuGAN leads with a score of 8.03, followed by OASIS (11.87) and SPADE (22.35). Here, probably due to the large size of the dataset, all models show the ability to produce images dissimilar to the training ones, although with different levels of diversity, and still reflecting the issues described above for OASIS and SPADE. When generating images from *from-train* maps, which we also use to train segmentation models in our next experiments, SPADE shows to be unable to generate images with different styles in all datasets while usually producing images similar to the real ones. OASIS can generate diverse images in both PN and RP datasets; however, the images usually contain cartoon-like artefacts, as can be seen in Fig. 7. We believe that a possible cause for these artefacts can be related to the local nature of OASIS discriminator, whose design does not allow for an evaluation of the image at a global scale, thus accepting as real images those that come as an unnatural mosaic of different image parts. Also, OASIS shows no ability to generate diverse samples on LC, probably due to the small size of the dataset. Notably, SuGAN is the only model under comparison that produces diverse, realistic, and artefact-free images in all datasets, including the small LC (Fig. 7, 7th and 8th columns).

4.3. Semantic segmentation with synthetic images

We investigate whether training segmentation models with synthetic images, together with real images, can boost segmentation performance when evaluated on real test images. Firstly, we explore two different approaches to use synthetic images along with real ones, namely *SR* and *S>R*. Both approaches use synthetic images from real training maps. SR (i.e., synthetic and real) simultaneously uses synthetic (from real train input maps) and real images to train the segmentation model, as described in Section 4.1. Instead, S>R (i.e., synthetic then real) uses synthetic images to pre-train the segmentation model and then fine-tune it with real images. This latter experiment is only performed in the PN dataset due to computational and time constraints.PN results in Table 3 indicate that synthetic data can successfully be used to improve segmentation performance with either approach. Unexpectedly, S>R seems to favour transformer-based architectures, while SR rewards more convolutional-based models. Overall, the use of synthetic data shows improved average performances with both approaches. Secondly, we compare SR models trained using synthetic images from real training maps (*RM*), synthetic maps from StyleGAN2 (*SM*) or a combination of the two (*RM+SM*), as described in Section 3.2, on PN dataset. Results are presented in the PN section of Table 3 for *RM* and Table 4 for *SM* and *RM+SM*. The use of real maps as input for the generative model leads to the best results for all CNN-based models, while the use of generated maps shows to benefit Swin-base architecture. Using a combination of both real and generated maps does not show evidence of consistent improvement. The experiments carried over LC and RP datasets (Table 3) use SR approach with real input maps for the generative model as these settings provide, on average, the most consistent and higher improvements and require only

Table 3
Segmentation results using different training approaches: *R* trains with real images only; *S>R* first pre-trains with synthetic images only and then fine-tunes with real images only; and *SR* trains with real and synthetic images simultaneously. *RM* refers to the use of real training maps as input for the synthetic dataset generation. The experiment is performed in LC, PN, and RP datasets. Results are reported on the test set as mean and standard deviation (std) based on 5 runs with different seeds. Bold indicates the highest mIoU. Green indicates an improvement with respect to only using real images. *S>R* is performed only on PN due to computational constraints.

Dataset	Model	Source	mIoU		
			Mean (↑)	Std (↓)	Diff (↑) %
LC	DeepLab	R	0.7887	0.0150	–
		SR RM	**0.8016**	0.0226	1.64
	HRNet32	R	0.7928	0.0205	–
		SR RM	**0.8358**	0.0080	5.42
	HRNet48	R	0.7800	0.0370	–
		SR RM	**0.8263**	0.0139	5.94
	Swin-Small	R	0.8321	0.0094	–
		SR RM	**0.8537**	0.0050	2.60
	Swin-Base	R	0.8283	0.0096	–
		SR RM	**0.8450**	0.0198	2.02
PN	DeepLab	R	0.6130	0.0052	–
		S>R RM	0.6190	0.0039	0.98
		SR RM	**0.6270**	0.0074	2.28
	HRNet32	R	0.6238	0.0072	–
		S>R RM	0.6284	0.0105	0.74
		SR RM	**0.6447**	0.0114	3.35
	HRNet48	R	0.6257	0.0057	–
		S>R RM	0.6297	0.0054	0.64
		SR RM	**0.6546**	0.0074	4.61
	Swin-Small	R	0.6474	0.0063	–
		S>R RM	**0.6635**	0.0042	2.49
		SR RM	0.6589	0.0052	1.77
	Swin-Base	R	0.6430	0.0080	–
		S>R RM	**0.6561**	0.0060	2.02
		SR RM	0.6546	0.0074	1.79
RP	DeepLab	R	0.4484	0.0117	–
		SR RM	**0.4665**	0.0032	4.03
	HRNet32	R	0.4606	0.0063	–
		SR RM	**0.4741**	0.0065	2.93
	HRNet48	R	0.4542	0.0040	–
		SR RM	**0.4788**	0.0040	5.43
	Swin-Small	R	0.4495	0.0122	–
		SR RM	**0.4648**	0.0047	3.42
	Swin-Base	R	0.4425	0.0087	–
		SR RM	**0.4608**	0.0063	4.14

one training stage instead of two. Results show that using synthetic images generated with SuGAN along with real ones improves the segmentation performance of all architectures for all datasets. Regarding LC, the smallest segmentation performance increase is achieved by DeepLab (1.64%) and the largest one by HRNet48 (5.94%). On PN, the smallest performance increase is achieved by Swin-Small (1.77%) and the largest one by HRNet48 (4.61%). Regarding RP, finally, the smallest performance increase is achieved by HRNet32 (2.93%) and the largest one by HRNet48 (5.43%). The highest overall mean mIoU in each dataset is achieved using HRNet48 in all the datasets. We also report the class-specific performance improvements in Fig. 8. For all models and datasets, the large majority of classes improve with a small number of exceptions, where performance remains unaltered or slightly decreased. Next, we summarize the largest decrease and increase in per-class performance. In LC, *L-hook electrocautery* with DeepLab decreased from 80.73 to 78.54 (2.7%), and *Gastrointestinal tract* with HRNet48 increased from 38.31 to 61.91 (61.6%). In PN, *Liver* with Swin-Small decreased from 69.06 to 68.64 (0.6%), and *Renal Artery* with HRNet48 increased from 31.79 to 40.64 (27.8%). In RP, *Foley catheter* with Swin-Small decreased from 40.92 to 38.25 (6.5%), and *Public Bones* with

Table 4
Segmentation results using different training approaches: *R* trains with real images only, and *SR* trains with real and synthetic images simultaneously. *SM* refers to the use of synthetic maps from StyleGAN2, as described in Section 3.2, as input for the synthetic dataset generation. The experiment is performed in PN dataset. Results are reported on the test set as mean and standard deviation (std) based on 5 runs with different seeds. Bold indicates the highest mIoU. Green indicates an improvement with respect to only using real images.

Model	Source	mIoU		
		Mean (↑)	Std (↓)	Diff (↑) %
DeepLab	R	0.6130	0.0052	–
	SR SM	0.6214	0.0040	1.36
	SR RM+SM	**0.6221**	0.0095	1.48
HRNet32	R	0.6238	0.0072	–
	SR SM	**0.6349**	0.0073	1.78
	SR RM+SM	0.6346	0.0054	1.73
HRNet48	R	0.6257	0.0057	–
	SR SM	**0.6433**	0.0087	2.81
	SR RM+SM	0.6339	0.0057	1.30
Swin-Small	R	0.6474	0.0063	–
	SR SM	0.6532	0.0058	0.89
	SR RM+SM	**0.6577**	0.0045	1.58
Swin-Base	R	0.6430	0.0080	–
	SR SM	**0.6570**	0.0088	2.17
	SR RM+SM	0.6567	0.0044	2.12

Swin-Base increased from 10.46 to 16.91 (61.6%). Interestingly, results seem to indicate an inverse correlation between the amount of per-class real data (indicated in the x-axis of Fig. 8) and the improvement when using synthetic images, particularly in RP. This suggests that synthetic data can effectively be used to boost performance over under-represented classes or, in general, class-unbalanced datasets. However, the performance boost affects only a subset of these classes, thus suggesting that class representation may be not the only parameter that drives higher or lower contributions from the use of synthetic data. A possible key factor could also lie in the variability and complexity of the class itself, where classes that have features that are easy to detect, e.g. liver (PN) or L-hook (LC), may benefit less than classes that have a more complex structure. In order to highlight the contribution carried by the use of synthetic images against simply using more common approaches such as image augmentation techniques, we compare the segmentation performance when training segmentation models using only real images (R), using synthetic images (SR), and using real data with different image augmentation techniques. First, we define a batch of five different base colour-based augmentations,[5] namely random gamma contrast, random colour temperature, random multiply and add to brightness, random multiply hue and saturation, and random add hue and saturation. Then, we compare R and SR with segmentation models trained using three colour-based augmentations that we refer to as A_I, $I \in [1, 2, 3]$, where I is the number of base augmentations applied consecutively at each training step. We randomly sample and apply I augmentations from the pool of five and the intensity of each base augmentation is uniformly drawn. This leads to stronger augmentations with increasing I. We train all five models over all three datasets and always replicate each training five times with different seeds. Overall, results are presented in Fig. 9. The best results using real data only are achieved by R. The rest of all colour-based augmentations show reduced performance compared with R over LC and PN test sets, independently from I. On RP, only A_1 shows little benefits for large models, i.e. HRNet48 and Swin-base, while degrading results on the remaining models and for all experiments with A_2 and A_3. These results suggest that colour-based augmentations do not generally advantage the performance for anatomy segmentation, as well as highlighting the gap between the use of common augmentations and the proposed

[5] https://imgaug.readthedocs.io/en/latest/source/overview/color.html

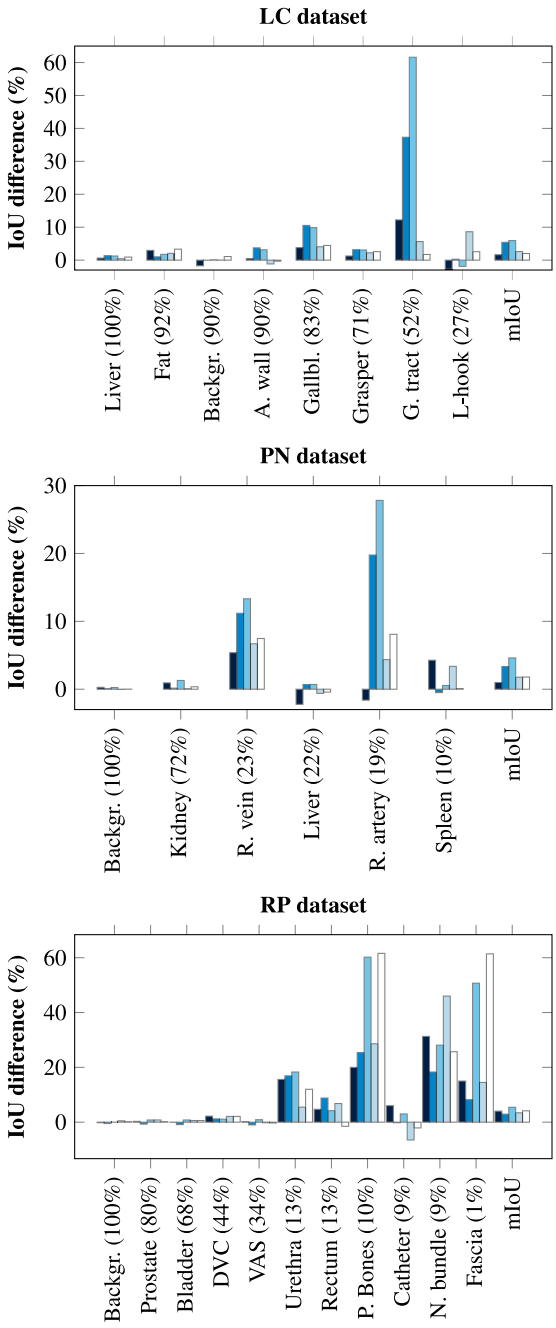

Fig. 8. Per-class perceptual intersection over union (IoU) difference when segmentation models are trained with real data only (R) or real and synthetic data simultaneously (SR). Results reported on the test set with three datasets LC, PN, and RP; and five segmentation models: ■ DeepLab, ■ HRNet32, ■ HRNet48, ■ Swin Small, and □ Swin-Base. A positive difference indicates that using synthetic data, along with real data, improves the segmentation results at testing time. The (%) in the *x*-axis refers to the presence of each class in the training set. Results are averaged across 5 runs with different seeds.

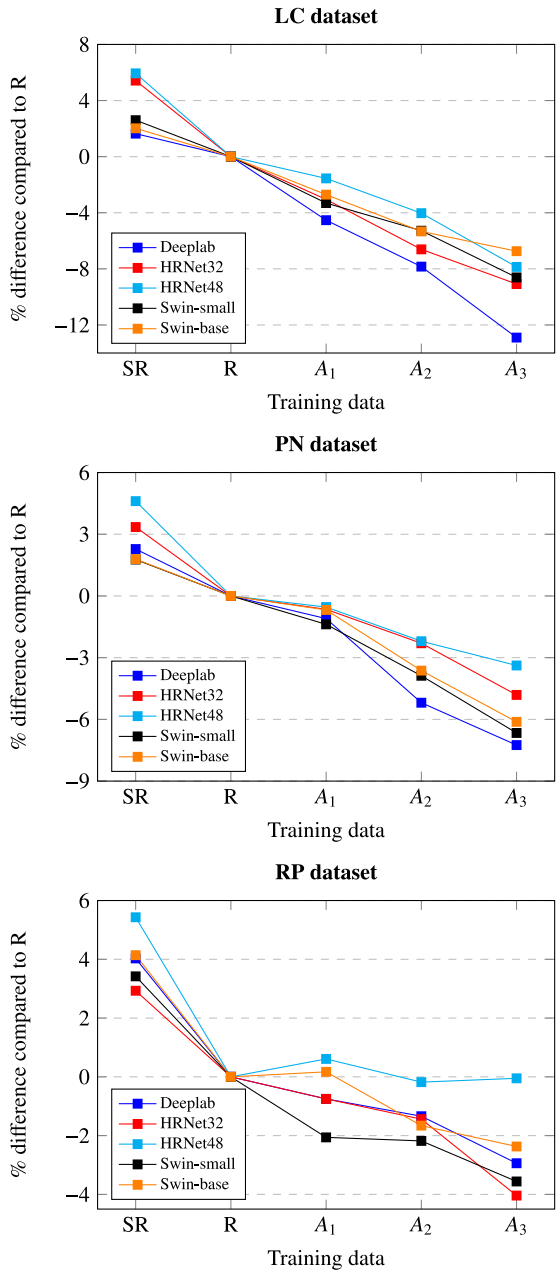

Fig. 9. Percentual intersection over union (IoU) difference over segmentation models trained with different data, synthetic and real (SR), real only (R), and real only with different augmentations (A_1, A_2 and A_3), as described in Section 4.3. Results are reported on the test set with three datasets: LC, PN, and RP; and five segmentation models: DeepLab, HRNet32, HRNet48, Swin Small, and Swin-Base. A positive difference indicates that there is an improvement over the baseline R. Results are averaged across 5 runs with different seeds.

method for improving segmentation performance. Next, we perform SR experiments using synthetic images generated by OASIS or SPADE. This experiment is performed by training DeepLab and Swin-Small in LC, PN, and RP datasets. Note that we use these two architectures only due to computational constraints, while still covering both CNN- and transformer-based models. Results in Table 5 indicate that while using synthetic images generated by SuGAN always improves mIoU, images from SPADE or OASIS do not always lead to such a positive result. Focusing on Deeplab, both SPADE and OASIS decrease the performance in LC by up to 2.44%, while they increase by 0.63% (SPADE) and decrease by 0.48% (OASIS) in PN. In RP dataset, with both Deeplab and Swin-Small, OASIS shows to generate the most suitable images to boost segmentation (Deeplab: +5.66%, Swin-Small: +6.51%, compared to R), followed by SPADE (Deeplab: +4.86%, Swin-Small: +5.92%), and SuGAN (Deeplab: +4.03%, Swin-Small: +3.42%). This result is unexpected as it goes against the assumption that synthetic images with higher quality would always lead to a better segmentation performance, especially when considering the absence of diversity in the images produced by SPADE and the artefacts generated by OASIS. We briefly discuss insights to further investigate this behaviour as well as other promising lines of investigation in Section 5. On final

Table 5
Segmentation results with two different training approaches: *R* trains with real images only, and *SR* trains with real and synthetic images simultaneously. Synthetic images generated by SPADE, OASIS, and the proposed SuGAN. Results are reported on the test set of LC, PN, and RP datasets. Models are Deeplab (CNN-based) and Swin-Small (Transformer-based). Results are reported as mean intersection over union (IoU) based on 5 runs with different seeds. Bold indicates the best mIoU.

Dataset	Source	mIoU (↑)	
		Deeplab	Swin-Small
LC	R	0.7887	0.8321
	SR SPADE	0.7694	0.8463
	SR OASIS	0.7799	0.8441
	SR SuGAN	**0.8016**	**0.8537**
PN	R	0.6130	0.6474
	SR SPADE	0.6169	0.6515
	SR OASIS	0.6101	0.6428
	SR SuGAN	**0.6270**	**0.6589**
RP	R	0.4484	0.4495
	SR SPADE	0.4702	0.4761
	SR OASIS	**0.4738**	**0.4787**
	SR SuGAN	0.4665	0.4648

analysis, it appears that the images produced by SuGAN have the most impact when training segmentation models on datasets of small (LC) and medium (PN) sizes, while frames from SPADE and OASIS have the most significant effect on larger datasets (RP).

4.4. Synthetic dataset size variability study

We analyse the contribution of using SuGAN synthetic images for training segmentation models while varying the size of the synthetic dataset, and thus the number of available synthetic images. Using as reference the size of each of the real training datasets, and following the same training settings described in Section 4.1, we train segmentation models using all the available real training data along with synthetic datasets as big as x0.25, x0.5, x1, x5 and x10 times the size of the real training set. The synthetic images composing each dataset are produced from segmentation maps sampled from the training set. Again, all trainings are repeated five times with different seeds and we set image resolution at 256 × 256. Results, Fig. 10, indicate that the use of synthetic data, in any amount, and for any dataset and model, contributes to an increase in performance with respect to only using real images (i.e. x0.0 case). Overall, the use of synthetic data shows a consistent performance increase, with a general trend of achieving the best scores when more synthetic data is used, i.e. x5 and x10. Particularly for small and medium datasets such as LC and PN, the best performance is achieved on all models but Swin-base and HRNet32, respectively, when using bigger synthetic datasets. On RP, on the other side, only Swin-small and HRNet48 followed the general trend, while resulting in a very similar performance for all synthetic dataset sizes in Swin-base and privileging a minor use of synthetic images when training Deeplab (best x1) and HRNet32 (best x0.25).

4.5. Sugan at double resolution

We investigate whether SuGAN can produce images at double resolution (512 × 512). To achieve this, we add one more set of pixel-level and channel-level normalization blocks on top of our standard architecture and we increase the number of latent vectors to 16 to support one more up-sampling step. This model is trained on PN and, similarly to previous experiments, we compare segmentation results (Deeplab) when training in R and SR configurations. Results indicate that also in this scenario, using synthetic images generated from SuGAN, we can enhance segmentation results (Table 6). Notably, using images with higher resolution leads to better results when training in both R and SR settings, while the improvement introduced by using synthetic images slightly increases, moving from 2.28% at 256 × 256 resolution to 2.42% at 512 × 512 resolution.

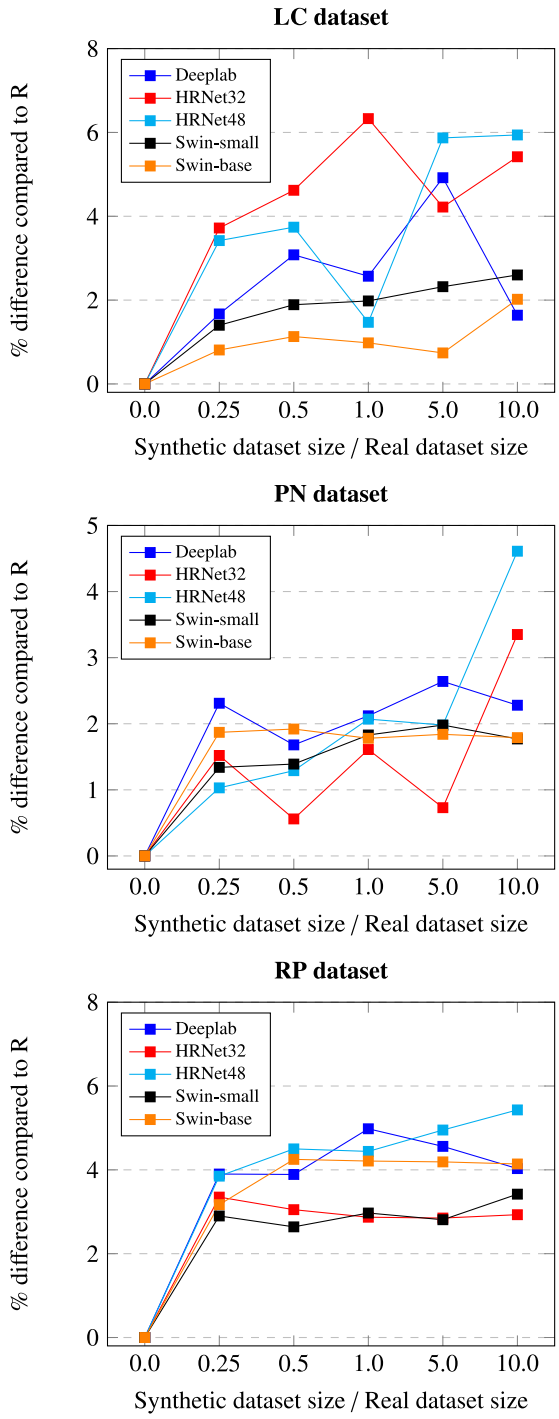

Fig. 10. Percentual intersection over union (IoU) difference over segmentation models trained with real and synthetic datasets of different sizes (e.g. 0.25 refers to a model trained with all real data and a synthetic dataset with size 25% of the real dataset size). Results reported on the test set with three datasets LC, PN, and RP; and five segmentation models: DeepLab, HRNet32, HRNet48, Swin Small, and Swin-Base. A positive difference indicates that there is an improvement over the baseline R. Results are averaged across 5 runs with different seeds.

5. Future directions

Although showing promising results, we believe that there is still further room to be explored from a data handling point of view, to fully exploit the potential of image synthesis to support segmentation models training. Firstly, this covers the investigation of how image quality and

Table 6
Per-class segmentation results with two different training approaches: *R* trains with real images only, and *SR* trains with real and synthetic images simultaneously. Experiment at 512 × 512 resolution using DeepLab in PN dataset. Results are reported as mean and standard deviation (std) based on 5 runs with different seeds. Bold indicates the best mIoU. Green/red indicates an increase/decrease in performance with respect to only using real images.

Class	Source	mIoU		
		Mean (↑)	Std (↓)	Diff (↑) %
Background	R	0.9111	0.0018	–
	SR	**0.9132**	0.0019	0.24
Kidney	R	0.7460	0.0061	–
	SR	**0.7559**	0.0075	1.33
Spleen	R	**0.5838**	0.0144	–
	SR	0.5760	0.0236	−1.34
Liver	R	0.6946	0.0110	–
	SR	**0.7027**	0.0095	1.16
Renal vein	R	0.4391	0.0193	–
	SR	**0.5155**	0.0231	17.38
Renal artery	R	0.3537	0.0162	–
	SR	**0.3553**	0.0072	0.45
mIoU	R	0.6214	0.0061	–
	SR	**0.6364**	0.0084	2.42

variability can affect the training procedure of downstream tasks since our experiments showed that these two factors can have a negative correlation under certain conditions. Secondly, we believe that other ways to generate synthetic data could be explored other than style manipulation, e.g. synthesizing images from automatically generated custom segmentation maps. This may include compositing new segmentation maps by randomly removing class instances or mixing class-specific labels from other existing maps, which has been explored in other works outside the surgical domain (Zhu et al., 2021; Su et al., 2022). On the same line, following our experiments that use GAN-generated maps as input for image generation, we believe it is important to explore this setting with more advanced generative models. Then, further inspection on how to leverage synthetic data to train downstream tasks and obtain higher performance with real test data should be carried out. This can also include changing the training settings to explore the contribution of synthetic images when training until convergence. Further on this, a deeper investigation could be carried out to explore how synthetic images can be used to further tackle class-unbalanced datasets. Finally, we believe investigating conditional diffusion models applications for downstream tasks in the surgical field is of great interest for future research.

6. Conclusion

In this paper, we presented SuGAN, a novel architecture for multimodal surgical data generation that competes with state-of-the-art models in terms of visual quality while avoiding common drawbacks present in existing models such as generation of overfitted images to the training ones, images with a lack of diversity, and images with artefacts. We explored two approaches for synthetic data handling when training segmentation models, with the final goal of improving the performance of semantic segmentation models of anatomical structures. Experimental results on three surgical datasets indicated that, unlike other state-of-the-art generative models, the proposed SuGAN is less prone to overfitting in both small and large datasets the proposed. In addition, experiments on surgical image segmentation show that using synthetic images, in addition to real ones, improves the segmentation performance in all considered three datasets and with any of the five considered segmentation models. We believe this work can positively impact research in the field and hopefully support future works in this direction.

CRediT authorship contribution statement

Emanuele Colleoni: Conceptualization, Methodology, Investigation, review & editing. **Ricardo Sanchez Matilla:** Conceptualization, Methodology, Investigation, review & editing. **Imanol Luengo:** Conceptualization, Methodology, Investigation, review & editing, Supervision. **Danail Stoyanov:** Conceptualization, Methodology, Investigation, review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Financial support was provided by Medtronic.

Data availability

The data that has been used is confidential.

Acknowledgements

The work was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145Z/16/Z]; Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080 /1, EP/P012841/1]; The Royal Academy of Engineering Chair in Emerging Technologies Scheme; and Horizon 2020 FET (GA 863146). For the purpose of open access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

References

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. PAMI 40 (4), 834–848.

Chong, M.J., Forsyth, D., 2020. Effectively unbiased fid and inception score and where to find them. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6070–6079.

Colleoni, E., Psychogyios, D., Van Amsterdam, B., Vasconcelos, F., Stoyanov, D., 2022. SSIS-Seg: Simulation-supervised image synthesis for surgical instrument segmentation. IEEE Trans. Med. Imaging.

Daroach, G.B., Duenweg, S.R., Brehler, M., Lowman, A.K., Iczkowski, K.A., Jacobsohn, K.M., Yoder, J.A., LaViolette, P.S., 2022. Prostate cancer histology synthesis using stylegan latent space annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 398–408.

Dhariwal, P., Nichol, A., 2021. Diffusion models beat gans on image synthesis. Adv. Neural Inf. Process. Syst. 34, 8780–8794.

Engelhardt, S., Sharan, L., Karck, M., Simone, R.D., Wolf, I., 2019. Cross-domain conditional generative adversarial networks for stereoscopic hyperrealism in surgical training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 155–163.

Fernandez, V., Pinaya, W.H.L., Borges, P., Tudosiu, P.D., Graham, M.S., Vercauteren, T., Cardoso, M.J., 2022. Can segmentation models be trained with fully synthetically generated data? In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer, pp. 79–90.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63 (11), 139–144.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inf. Process. Syst. 30.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A., 2017. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=Sy2fzU9gl.

Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S., 2020. CholecSeg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv:2012.12453.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510.

Huang, X., Mallya, A., Wang, T.C., Liu, M.Y., 2022. Multimodal conditional image synthesis with product-of-experts gans. In: European Conference on Computer Vision. Springer, pp. 91–109.

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

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T., 2020a. Training generative adversarial networks with limited data. Adv. Neural Inf. Process. Syst. 33, 12104–12114.

Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T., 2020b. Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119.

Kovalev, V., Kazlouski, S., 2019. Examining the capability of GANs to replace real biomedical images in classification models training. In: International Conference on Pattern Recognition and Information Processing. Springer, pp. 98–107.

Kumar, R., Wang, W., Kumar, J., Yang, T., Khan, A., Ali, W., Ali, I., 2021. An integration of blockchain and AI for secure data sharing and detection of CT images for the hospitals. Comput. Med. Imaging Graph. 87, 101812.

Li, Y., Li, Y., Lu, J., Shechtman, E., Lee, Y.J., Singh, K.K., 2021. Collaging class-specific gans for semantic image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14418–14427.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: CVPR. pp. 10012–10022.

Madani, A., Namazi, B., Altieri, M.S., Hashimoto, D.A., Rivera, A.M., Pucher, P.H., Navarrete-Welton, A., Sankaranarayanan, G., Brunt, L.M., Okrainec, A., et al., 2022. Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. Ann. Surg..

Marzullo, A., Moccia, S., Catellani, M., Calimeri, F., De Momi, E., 2021. Towards realistic laparoscopic image generation using image-domain translation. Comput. Methods Programs Biomed. 200, 105834.

Miyato, T., Koyama, M., 2018. Cgans with projection discriminator. arXiv preprint arXiv:1802.05637.

Nichol, A.Q., Dhariwal, P., 2021. Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. PMLR, pp. 8162–8171.

Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2337–2346.

Pfeiffer, M., Funke, I., Robu, M.R., Bodenstedt, S., Strenger, L., Engelhardt, S., Roß, T., Clarkson, M.J., Gurusamy, K., Davidson, B.R., et al., 2019. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 119–127.

Poucin, F., Kraus, A., Simon, M., 2021. Boosting instance segmentation with synthetic data: A study to overcome the limits of real world data sets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 945–953.

Rau, A., Edwards, P., Ahmad, O.F., Riordan, P., Janatka, M., Lovat, L.B., Stoyanov, D., 2019. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int. J. Comput. Assist. Radiol. Surg. 14 (7), 1167–1176.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D., 2021. Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2287–2296.

Rivoir, D., Pfeiffer, M., Docea, R., Kolbinger, F., Riediger, C., Weitz, J., Speidel, S., 2021. Long-term temporally consistent unpaired video translation from simulated surgical 3d data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3343–3353.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695.

Schonfeld, E., Veeravagu, A., 2023. Demonstrating the successful application of synthetic learning in spine surgery for training multi–center models with increased patient privacy. Sci. Rep. 13 (1), 12481.

Sheetz, K.H., Claflin, J., Dimick, J.B., 2020. Trends in the adoption of robotic surgery for common surgical procedures. JAMA Netw. Open 3 (1), e1918911.

Su, W., Ye, H., Chen, S.Y., Gao, L., Fu, H., 2022. Drawinginstyles: Portrait image generation and editing with spatially conditioned stylegan. IEEE Trans. Vis. Comput. Graphics.

Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B., Khoreva, A., 2020. You only need adversarial supervision for semantic image synthesis. arXiv preprint arXiv:2012.04781.

Thambawita, V., Salehi, P., Sheshkal, S.A., Hicks, S.A., Hammer, H.L., Parasa, S., Lange, T.d., Halvorsen, P., Riegler, M.A., 2022. SinGAN-Seg: Synthetic training data generation for medical image segmentation. PLoS One 17 (5), e0267976.

Tsui, C., Klein, R., Garabrant, M., 2013. Minimally invasive surgery: national trends in adoption and future directions for hospital strategy. Surg. Endosc. 27 (7), 2253–2257.

Twinanda, A.P., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N., 2016. Single- and multi-task architectures for surgical workflow challenge at M2CAI 2016. arXiv:1610.08844.

Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., Li, H., 2022a. Semantic image synthesis via diffusion models. arXiv preprint arXiv:2207.00050.

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

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al., 2020. Deep high-resolution representation learning for visual recognition. TPAMI 43 (10), 3349–3364.

Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F., 2022b. Pretraining is all you need for image-to-image translation. arXiv preprint arXiv:2205.12952.

Wei, T., Chen, D., Zhou, W., Liao, J., Zhang, W., Yuan, L., Hua, G., Yu, N., 2022. E2Style: Improve the efficiency and effectiveness of StyleGAN inversion. IEEE Trans. Image Process. 31, 3267–3280.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595.

Zhu, P., Abdal, R., Femiani, J., Wonka, P., 2021. Barbershop: GAN-based image compositing using segmentation masks. ACM Trans. Graph. 40 (6), 1–13.