Improving realism in patient-specific abdominal Ultrasound simulation using CycleGANs

Article *in* International Journal of Computer Assisted Radiology and Surgery · July 2019

DOI: 10.1007/s11548-019-02046-5

CITATIONS
2

READS
536

4 authors:

Santiago Vitale
National University of the Center of the Buenos Aires Province
2 PUBLICATIONS **2** CITATIONS
SEE PROFILE

José Ignacio Orlando
National Scientific and Technical Research Council
41 PUBLICATIONS **547** CITATIONS
SEE PROFILE

Emmanuel Iarussi
National Institute for Research in Computer Science and Control
10 PUBLICATIONS **127** CITATIONS
SEE PROFILE

Ignacio Larrabide
National Scientific and Technical Research Council
131 PUBLICATIONS **1,453** CITATIONS
SEE PROFILE

Noname manuscript No.
(will be inserted by the editor)

Improving realism in patient-specific abdominal Ultrasound simulation using CycleGANs

Santiago Vitale · José Ignacio Orlando · Emmanuel Iarussi · Ignacio Larrabide

Received: date / Accepted: date

Abstract *Purpose:* In this paper we propose to apply generative adversarial neural networks trained with a cycle-consistency loss, or CycleGANs, to improve realism in ultrasound (US) simulation from Computed Tomography (CT) scans.
Methods: A ray-casting US simulation approach is used to generate intermediate synthetic images from abdominal CT scans. Then, an unpaired set of these synthetic and real US images is used to train CycleGANs with two alternative architectures for the generator, a U-Net and a ResNet. These networks are finally used to translate ray-casting based simulations into more realistic synthetic US images.
Results: Our approach was evaluated both qualitatively and quantitatively. A user study performed by 21 experts in US imaging shows that both networks significantly improve realism with respect to the original ray-casting algorithm ($p \ll 0.0001$), with the ResNet model performing better than the U-Net ($p \ll 0.0001$).
Conclusion: Applying CycleGANs allows to obtain better synthetic US images of the abdomen. These results can contribute to reduce the gap between artificially generated and real US scans, which might positively impact in applications such as semi-supervised training of machine learning algorithms and low-cost training of medical doctors and radiologists in US image interpretation.

Keywords Ultrasound · Image simulation · Deep learning

S. Vitale · I. Larrabide
Pladema, UNICEN, Tandil, Argentina
Tel.: +54-249-4385690
E-mail: svitale@conicet.gov.ar

J. I. Orlando
OPTIMA, Department of Ophthalmology,
Medical University of Vienna, Vienna, Austria

E. Iarussi
UTN-FRBA, Buenos Aires, Argentina

S. Vitale · E. Iarussi · I. Larrabide
CONICET, Argentina

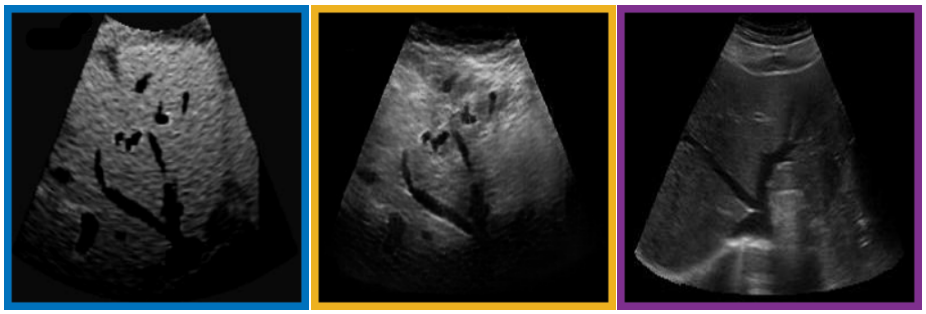
1 Introduction

Ultrasound (US) is a frequently used medical imaging technique that is extensively applied as an examination tool for diagnosis, treatment and in emergentology [17]. Compared to other modalities such as X-ray or Computed Tomography (CT), US imaging is non-invasive, radiation-free and can be acquired in real time using relatively portable and economic devices. In emergency rooms, US is frequently used to obtain a fast scan of the inner organs, allowing a quick assessment of potentially damaged areas [2]. US is also widely applied to diagnose pathological alterations of the abdominal organs, to assess the location of abnormal masses [27], among other clinically relevant tasks. These images require specially trained readers for their interpretation, who must discriminate between the variable echogenicity properties and speckle noise characteristics of the different tissues and imaging artifacts produced e.g. by the frequency of the US wave. As a consequence, a significant effort is being made to develop tools for training radiologists both for image acquisition and interpretation [15].

Medical image simulation is an active field of research that allows to artificially recreate clinical scenarios with abnormal and/or critical events without any risk, in a controlled environment and without any patient's risk. More specifically, US simulation has been vastly explored as an alternative for training radiologists in capturing and interpreting US scans [8,19]. Traditionally, training skills such as recognizing abnormalities require US users to operate a real device and capture images from volunteers with pathologies [8]. This means that the patient has to be present during the whole training exercise, which can be stressful or even not possible in some cases (e.g. under rare conditions or in high risk scenarios). Alternatively, US simulators have proved to allow low-cost training of US operators [19].

Generating realistic scans is essential to ensure a smooth transition of human trainees from US simulations to real US acquisitions, specially for image interpretation. However, current approaches still struggle to produce truthful artificial scans. Early US simulators partially overcome this difficulty by navigating through pre-recorded 2D US images [26,12]. These scans were joined into individual volumes and traversed using slice resampling based on the (simulated) transducer location. However, although images are not synthetic, realism is lost as soon as the transducer moves away from the original position in the acquired volume. Moreover, these simulators do not explicitly model other US parameters such as frequency and gain, so a new volume has to be acquired for each specific configuration. Other recent approaches produce synthetic scans by exploiting geometries extracted from alternative imaging modalities such as CT [24,15]. These models are more sophisticated as they intend to recreate the complex interaction of sound waves coming out from the transducer probe as they traverse the body. However, as previously described in [23], they struggle to realistically model imaging artifacts (e.g. shadows, reverberations, comet tail artifacts, etc.), anatomical features (e.g. muscle fibers, fat streaks, microcalcifications, etc.) or tissue interfaces. This is partially because of their hand crafted nature, which force them to explicitly model every artifact.

Deep learning is currently the state-of-the-art tool for automated medical image analysis in numerous applications, including image segmentation, classification and synthesis [16]. Generative Adversarial Neural Networks (GANs) [10] are a specific type of deep neural networks that allows synthesizing new images either



(a) (b) (c)

Fig. 1: US image simulations of the liver. (a) Ray-casting based simulation. (b) CycleGAN result. (c) Real US image from approximately the same area.

.

by sampling from a noise distribution [10] or based on an input image [30]. In this work we focus on this second type of networks, where the output sample is conditioned by the characteristics of the input. This task is usually referred to in computer vision as "image-to-image" translation. The first approach introducing GANs for this task is the so-called *pix2pix* [14], which learns to transfer the style of an image from one domain to another based on paired training samples. Alternatively, Zhu *et al.* [30] introduced a cycle-consistency loss that allows learning similar transformations but from unpaired sets of images. This setting reduces the burden of collecting paired samples for translation, which is extremely expensive or even unfeasible in medical imaging applications. As a result, this approach has been successfully applied to translate magnetic resonance images (MRIs) to CT scans [28, 13] and to improve surgical phantoms [9]. To the best of our knowledge, CycleGANs have not been applied yet in the context of US image simulation.

In this paper we introduce the first approach for realistic patient-specific abdominal US simulation based on a combination of standard ray-casting based simulation and deep convolutional neural networks (Fig. 1). In particular, we assess the viability of using GANs trained with a cycle-consistency loss, or CycleGANs, to improve realism on the outputs of a ray-casting approach [15, 23]. We also analyze the architecture influence of the generator in the final outcomes by training two alternative models based on a U-Net [22] and a ResNet [11], respectively. We validated our approach through a user study performed over a cohort of experts in US image analysis. We observed that both architectures allow to improve realism on the output images, with the ResNet reporting the best results. We also qualitatively evaluate the limitations of our approach and propose further lines of research to improve them.

2 Methods

Fig. 2 depicts our simulation pipeline. Our inputs are an abdominal CT volume and a voxel-wise segmentation of the organs. A ray-casting approach (Section 2.1) is applied on these inputs to retrieve synthetic US images based on the position of

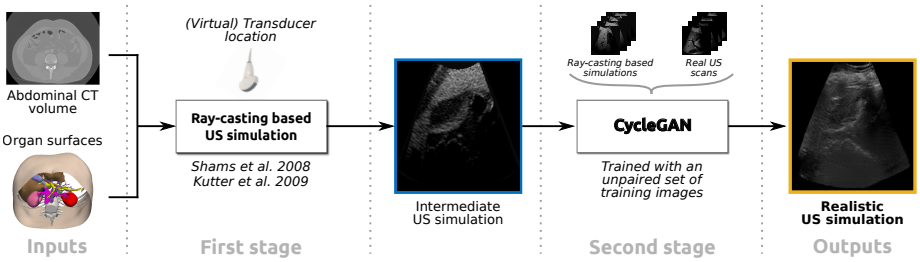


Fig. 2: Schematic representation of our US simulation pipeline. A patient-specific abdominal CT volume and its associated organ segmentations are used as inputs for a ray-casting based simulator. The resulting scan is used as input for a CycleGAN, trained with an unpaired set of real US images and other intermediate ray-casting based simulation, to retrieve a more realistic synthetic US image.

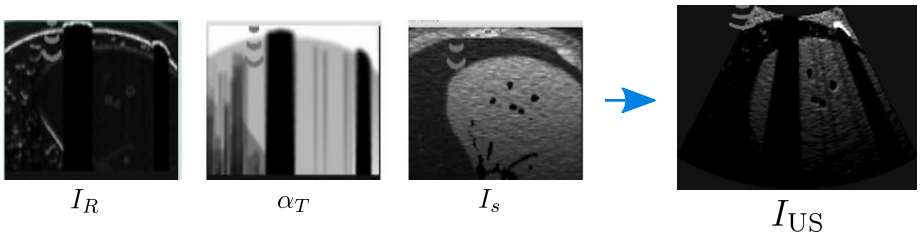


Fig. 3: Ray-casting based simulation. From left to right: (in polar coordinates) buffers for reflection I_R and transmision α_T, associated slice from the scattering volume I_s and (in cartesian coordinates) resulting I_{US} image.

a virtual transducer. A CycleGAN model (Section 2.2) trained on an unpaired set of these intermediate images and real US scans, is then used to obtain the more realistic output.

2.1 Ray-casting based simulation

Our first stage consists of a ray-casting algorithm based on the approach presented in [24, 15]. We use our own group implementation [23] that features changes, due to implementation decisions, with respect to the original implementation. Generally speaking, ray-casting based simulation models the sound wave movement through the body to obtain an initial synthetic US image scan with occlusions and large scale reflections effects (Fig. 3). As a pre-processing, offline step, this method requires to compute a scattering volume, a 3D representation of the body in which a different scattering coefficient is assigned to each organ, according to their (known) tissue properties [24, 15]. The interested reader could refer to [5] for specific details regarding of our implementation of the scattering pre-computation. A virtual transducer is positioned within this 3D space of both the CT and the scattering volume. Subsequently, the algorithm computes values for reflection, attenuation and speckle noise of the sound waves casted from the US probe. When a wave travels in homogeneous tissue and reaches an interface between two media

with different acoustic impedance, energy is reflected back to the transducer. The amount of reflected energy is determined using a reflection coefficient, α_R, given by:

$$\alpha_R = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1} \right)^2 \tag{1}$$

where Z_1 and Z_2 are the CT Hounsfield unit approximated acoustic impedance of the different medias, which assumes that there is an approximately linear relationship between them. This relationship can be derived from different mapping functions [29,21]. Our implementation uses a direct conversion from Hounsfield units to acoustic impedance as in [23]. The remaining energy passing through the interface to the second medium is referred as transmission, α_T, obtained as $\alpha_T = 1 - \alpha_R$.

The reflection at tissue interfaces is diffuse and subject to scattering. To simulate this effect, we used a Lambertian model as in [24,15]. As pointed out in [24], a Rayleight model would be a better choice since the interface dimensions are much smaller than a wavelength. Nevertheless, we used a Lambertian model due to its simpler computational implementation and efficiency. This type of models assumes the brightness of a surface to be independent from the viewing angle but dependent on the angle of incidence between the ray and the surface of the organ to traverse. Therefore, the intensity of the reflected signal can be modelled as:

$$I_R(x) \propto \alpha_R(x) \frac{I_i^2}{I_0} \left| r(x) \cdot n(x) \right| \tag{2}$$

where $I_R(x)$ is the reflected intensity, r is a unit vector indicating the direction of the wave, n is the normal to the organ surface and $\frac{I_i^2}{I_0}$ is the cumulative attenuation at the interface, being I_0 the initial ray intensity and $I_i(x)$ the intensity at a location x [24,15]. The speckle pattern and view-dependent effects are modelled using the scattering image by sampling along the same reflection wave path.

Finally, the synthetic US image I_{US} at a location x is calculated based on the reflection, transmission and scattering buffers (Fig. 3) by doing:

$$I_{\text{US}}(x) = \left(w_1 G_{\sigma_1}(x) \cdot I_R(x) + w_2 G_{\sigma_2}(x) \cdot \alpha_T(x) \right) I_s(x) \tag{3}$$

where $I_R(x)$ and $I_s(x)$ are the reflection and scattering images, respectively, w_1 and w_2 are blending coefficients that are hyperparameters of the method and G are Gaussian filters with 0 mean and adjustable deviations σ_1 and σ_2 [24,15]. A log-compression method is applied at the end to reduce the dynamic range of the output [23].

2.2 Improving realism with CycleGANs

The aim of our CycleGAN [30] is to translate images from the domain X of all possible outputs of the ray-casting model to the domain Y of all possible real US images.

A standard GAN normally consists of two convolutional neural networks, a generator $G : X \rightarrow Y$ and a discriminator D_Y. G is used to translate the input x to the domain Y–in our case, I_{US} to its (more) realistic analogue, I^y_{US}–. D_Y,

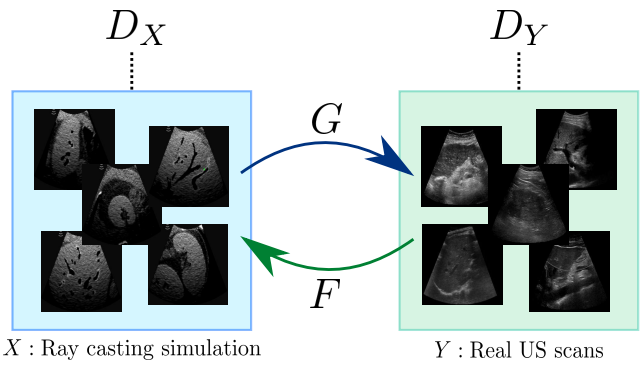


Fig. 4: Our CycleGAN model for improving US simulation realism.

on the other hand, is used during training to recognize fake images $\{G(x)\}$ from real ones $\{y\}$. Both networks are trained in an end-to-end fashion by optimizing an adversarial loss:

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \backsim p_{\text{data}}(y)}[\log D_Y(y)] \\ + \mathbb{E}_{x \backsim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))],\end{aligned} \tag{4}$$

where $\mathbb{E}$ stands for the expected value of each corresponding data distribution [10]. This loss models a competition in which G tries to deceive D_Y by producing realistic outputs and D_Y tries to overpass the scam.

Under this setting, a GAN needs to be trained using paired samples (x, y). In our context, this means that we need real US images ($y = I^Y_{\text{US}}$) perfectly registered with the outputs of the first simulation stage ($x = I_{\text{US}}$), which is unfeasible. A CycleGAN [30] gets rid of this limitation by including a second pair of generator/discriminator networks and a cycle consistency loss (Fig. 4). The new generator $F : Y \to X$ learns how to translate realistic US images to the outputs of the ray-casting based simulation by fooling a discriminator D_X using an adversarial loss complementary to (4). Additionally, the cycle consistency loss ensures that any generated images can bring back the original input of its associated generator, e.g. by following the forward cycle $x \to G(x) \to F(G(x)) \approx x$ and the backward cycle $y \to F(y) \to G(F(y)) \approx y$. This loss is given by the sum of two losses, one per each cycle [30]:

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \backsim p_{\text{data}}(x)}[||F(G(x)) - x||_1] \\ + \mathbb{E}_{y \backsim p_{\text{data}}(y)}[||G(F(y)) - y||_1].\end{aligned} \tag{5}$$

Hence, $\mathcal{L}_{\text{cyc}}(G, F)$ acts as a regularizer for G and F, aiding them to approximately match the data distributions $p_{\text{data}}(y)$ and $p_{\text{data}}(x)$ without relying on a training set of paired samples.

We also incorporated an identity loss as in [25,30] to further regularize our generators:

$$\begin{aligned}\mathcal{L}_{\text{identity}}(G, F) = \mathbb{E}_{x \backsim p_{\text{data}}(x)}[||F(x) - x||_1] \\ + \mathbb{E}_{y \backsim p_{\text{data}}(y)}[||G(y) - y||_1].\end{aligned} \tag{6}$$

Notice that $\mathcal{L}_{\text{identity}}$ penalizes changes in images from the target domain, that should not be altered. This enforces each generator to avoid hallucinating unrealistic features.

The full training objective is equivalent to the one in [30], given by:

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = &\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) \\ &+ \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) \\ &+ \lambda_{\text{cyc}} \cdot \mathcal{L}_{\text{cyc}}(G, F) \\ &+ \lambda_{\text{idt}} \cdot \mathcal{L}_{\text{identity}}(G, F)\end{aligned} \tag{7}$$

where λ_{cyc} and λ_{idt} are hyperparameters that control the relative importance of the loss terms.

The final models are obtained by solving the optimization problem:

$$G^*, F^* = \arg\min_{G,F} \max_{D_X, D_Y} \mathcal{L}(G, F, D_X, D_Y). \tag{8}$$

Notice that only the generator G is used in test time, as our intention is refining the simulated outputs I_{US} and not altering a real US scan to retrieve its simulated equivalent.

3 Experimental setup

3.1 Materials

Our simulations were performed as in [15], using an abdominal CT volume from patient 11 in the 3D-IRCADb-01 data set [1]. It comprises a scan from asubject without any liver pathologies, with a resolution of $512 \times 512 \times 132$ voxels with a size of $0.72 \times 0.72 \times 1.6$ mm. This data includes the segmentations of all abdominal organs, except the stomach and the pancreas.

The set X was built using 817 simulated US images, pre-computed using the ray-casting algorithm described in Section 2.1. The images were manually collected from the volume described above, simulating a typical abdomen scanning procedure with different transducer position, rotation and inclination angle. A separate set of 33 simulated images from the main views in a typical abdomen ultrasound exploration, (none of them used in the CycleGAN training set) was used for the user study (Section 3.4).

The set Y of real US scans was built using a total number of 992 real images from the abdomen. 719 were manually collected from the ETHZ data set [20, 7].[1] To increase the size of the original set, an additional set of 273 images was collected from the Deep Ultrasound website. [2] The final set was manually curated by resizing and cropping the images to approximately the same pixel resolution (256 x 256 pixels), and to remove any text and device information outside the field of view.

[1] `http://www.vision.ee.ethz.ch/datasets_extra/usliverseq.zip`

[2] `http://deepultrasound.ai/`

3.2 Simulation parameters

We used the C++ CPU implementation of the Kutter *et al.* approach [15] presented in [23]. Such a model uses fixed reflection coefficients α_R for bones (0.5), air (0.6), arteries (0.0001) and the portal vein (0.1). The coefficients w_1 and w_2 were experimentally fixed to 0.9 and 0.1, respectively. The deviations σ_1 and σ_2 were set to 1.5 and 3. The logarithmic compression at the end was applied using a coefficient of 12 [23].

3.3 CycleGAN configuration

We used the Pytorch 0.4.0 implementation of the original CycleGAN paper [30] as the baseline of our neural network implementation. [3] The two original generators, a U-Netand a ResNetwere modified to avoid the classical checkerboard artifacts of the transposed convolutions in the decoder branches [18]. Instead, we replaced them with a combination of a bilinear upsampling followed by a convolution with a kernel size of 3 and a stride of 1. Two different CycleGAN models, $\text{CycleGAN}_{\text{U-Net}}$ and $\text{CycleGAN}_{\text{ResNet}}$, were trained for comparison purposes, each of them using one of the two architectures for the generators. The coefficients λ_{cyc} and λ_{idt} were experimentally fixed to 10 and 0.5. The models were trained using Adam optimization for 400 epochs. An initial learning rate of 2×10^{-5} was used. After the first 200 epochs, the learning rate was iteratively reduced by a fixed value of $\frac{1}{201}$. A NVIDIA Titan X GPU was used, and the mini-batch size was set to the maximum possible value to fit both the models and the data on the GPU memory (4 images for the ResNet architecture and 8 for the U-Net).

3.4 User evaluation study setup

A blind user study was performed to analyze the improvement in the perceptual realism of the simulated US images. A custom online tool was developed to this end, using the jsPsych JavaScript library [6]. The study was iteratively designed based on the comments from independent volunteers, whose responses were not included in the final test. After finishing each design trial, they were asked about aspects such as usability of the user interface of the tool, tediousness of the test and missing features. To account for criticisms regarding the time length of the test and the presence of too similar images, we limited the amount of images to a series of clinically representative sequences, with sufficient variability one another. All these images corresponded to the main views in a typical abdominal US exploration, namely intercostals, subcostals margin, longitudinal, oblique and transverse scans.

Using the final version of the test, new volunteers were invited to rank the realism of the images using a Likert scale from 1-5: *Fake* (1), *Rather fake* (2), *Cannot decide* (3), *Rather real* (4) and *Real* (5). 21 experts in US participated from the user study: 7 medical doctors (MD), 2 US technicians (Tec) and 12 bio-engineers/computer scientists (BEng) with experience in US imaging (Fig. 7). A

[3] `https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix`

total of 44 images were used during the test, with 11 being real US scans, 11 produced with the ray-casting based algorithm, and 22 processed with the CycleGANs. From this set, 11 images were obtained with the U-Net architecture, and the remaining 11 were computed using the ResNet. The first 4 images corresponded to each of the categories and were presented as training examples, so the responses to them were discarded. Therefore, the responses to the remaining 40 images were taken for subsequent analysis. These scans were randomly presented to the experts without indicating their origin or any ratio between simulated and reals. The time spent in analyzing each image was also registered on a per-volunteer setting to assess the participants' reliability.

4 Results and Discussion

4.1 Qualitative analysis

Fig. 5 depicts qualitative examples of the outputs of the stages of our US simulation approach, including the results obtained with the $\text{CycleGAN}_{\text{U-Net}}$ and the $\text{CycleGAN}_{\text{ResNet}}$. All the images correspond to regions that are typically analyzed during abdominal US examinations.

The ray-casting simulator images exhibits the organs in a clearer way than the neural network outputs, as it uses the segmentations to generate the synthetic scans. The reflection coefficients used for bones, air and blood vessels help to better represent these structures, as seen in Fig. 5(b), although the resulting artifacts are exaggerated. When applying the $\text{CycleGAN}_{\text{U-Net}}$, the overall structure of the images is maintained but the intensity distribution is changed to a new appearance that better resembles a real US scan. Organs remain visible as in the ray-casting based input, and no significant artifacts are introduced by the network. The $\text{CycleGAN}_{\text{ResNet}}$, on the contrary, produces more aggressive changes to the synthetic input. Although organs are still recognizable, bright areas resembling echos and noise, characteristic of US images, are spread over the images. We hypothesize that the multiscale skip connections in the U-Net grant a better reconstruction of the inputs than the ResNet, which only has two of these connections in the first two resolution blocks. This has effects in the realism outcomes (see Section 3.4), as it produces images that are hard to interpret. The improvement in the tissue appearance near the probe becomes evident when introducing any CycleGAN module, as the ray-casting algorithm do not explicitly incorporate this artifact in the model. Furthermore, it is worth mentioning that the $\text{CycleGAN}_{\text{ResNet}}$ does not introduce any features in black areas without echogenic response (Fig. 5(d), first and second columns) but an artificial acoustic shadowing, which is in line with the real response of the tissue. The $\text{CycleGAN}_{\text{U-Net}}$, on the contrary, introduces bright but low contrast features in the area, which are inconsistent with these attenuation artifacts.

The effect of the λ_{idt} parameter can be observed in Fig. 6. When the identity loss is ignored ($\lambda_{\text{idt}} = 0$), the generator introduces bright artifacts that remain constant in the same image location, regardless the position of the transducer and without anatomical coherence (see arrows in Fig. 6). When λ_{idt} is increased, the artifacts become more realistic and appear only at organ interfaces (e.g., in the

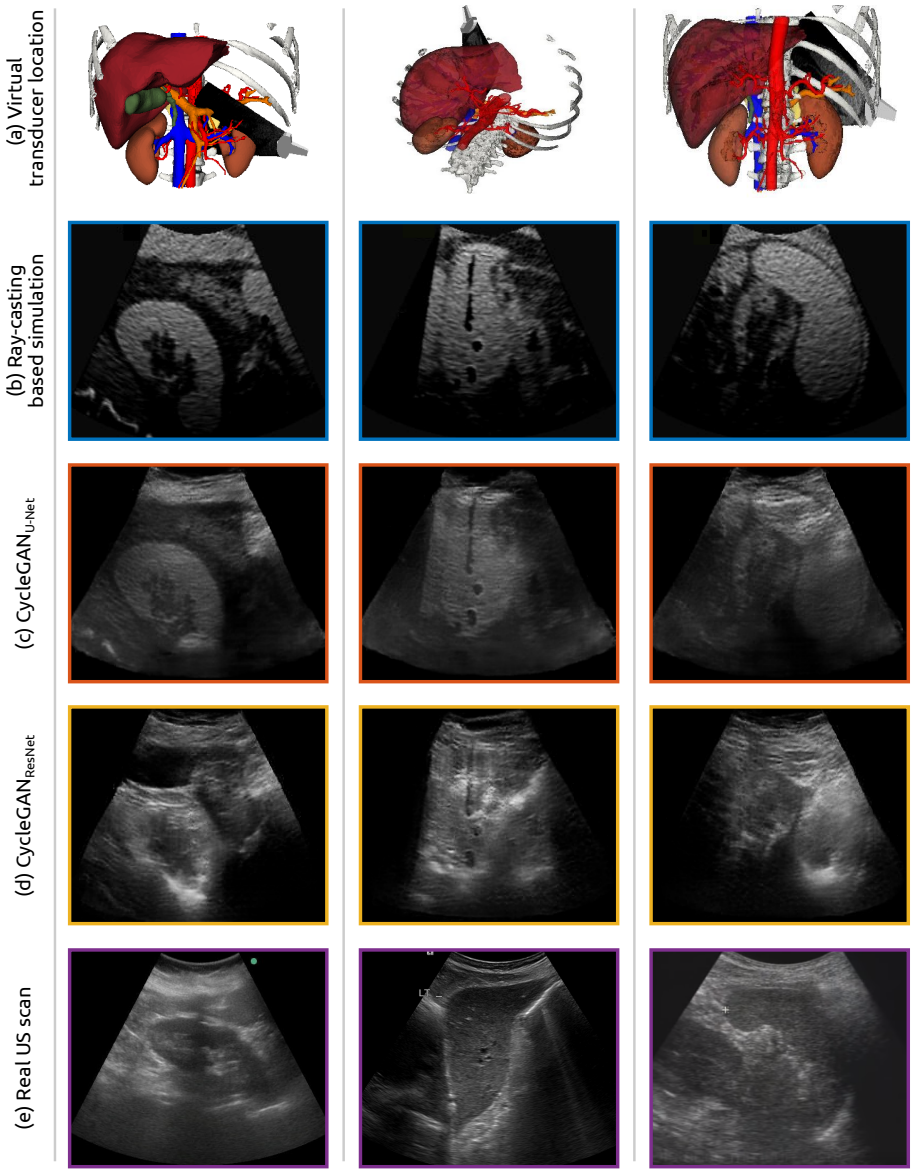
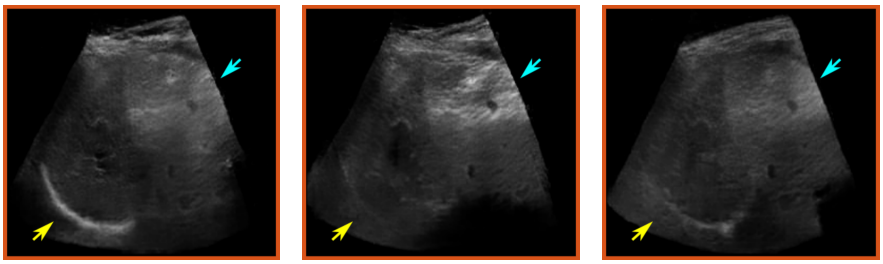


Fig. 5: Qualitative examples of our results. (a) Virtual transducer location in the CT volume, (b) Ray-casting based simulation, (c) CycleGAN$_{\text{U-Net}}$, (d) CycleGAN$_{\text{ResNet}}$ and (e) Real US image.



(a) $\lambda_{\text{idt}} = 0.0$ (b) $\lambda_{\text{idt}} = 0.5$ (c) $\lambda_{\text{idt}} = 0.8$

Fig. 6: Effect of the λ_{idt} value in the CycleGAN$_{\text{U-Net}}$ results. Yellow arrow: artifact in the liver interface. Light blue arrow: echogenic artifact.

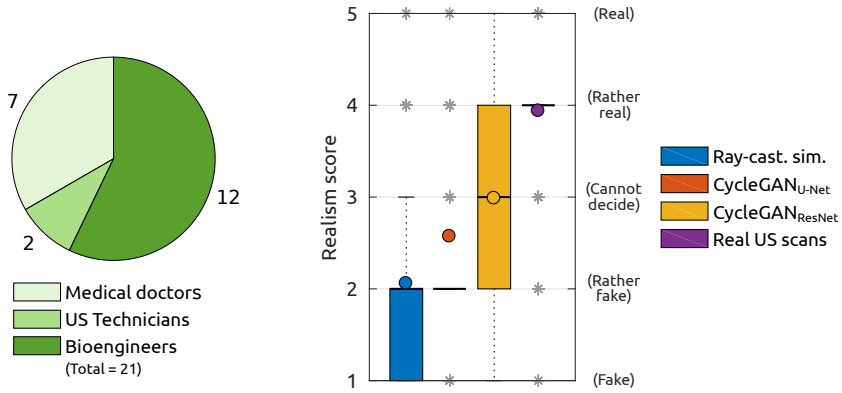


Fig. 7: Left: Distribution of expert volunteers participating in the user study. Right: Box plots with the distribution of realism scores assigned by the group of 21 volunteers to each set of images in the user study. Mean responses and outliers are indicated with circles and stars, respectively.

edges of the liver–yellow arrow–) or they move through the image following the transducer position (e.g. the bright artifact indicated by the light blue arrow).

4.2 User study

To assess the degree of matching in the distribution of answers between the three group of experts (MD, Tec and BEng), a series of two-tailed Wilcoxon rank sum tests was performed between groups, per each image category. Bonferroni correction was applied to account for multiple comparisons, adjusting the significance level 5% to 1.66% (3 comparisons). No statistical significant differences were observed in the responses between groups of experts for all types of images, except between the BEng and Tec groups for the ray-casting generated images ($p = 0.0104$).

Fig. 7 presents the results of the user study. Responses from the different groups of experts were merged for statistical analysis due to the high agreement in their distribution. One-tail Wilcoxon sign rank tests were performed to assess the

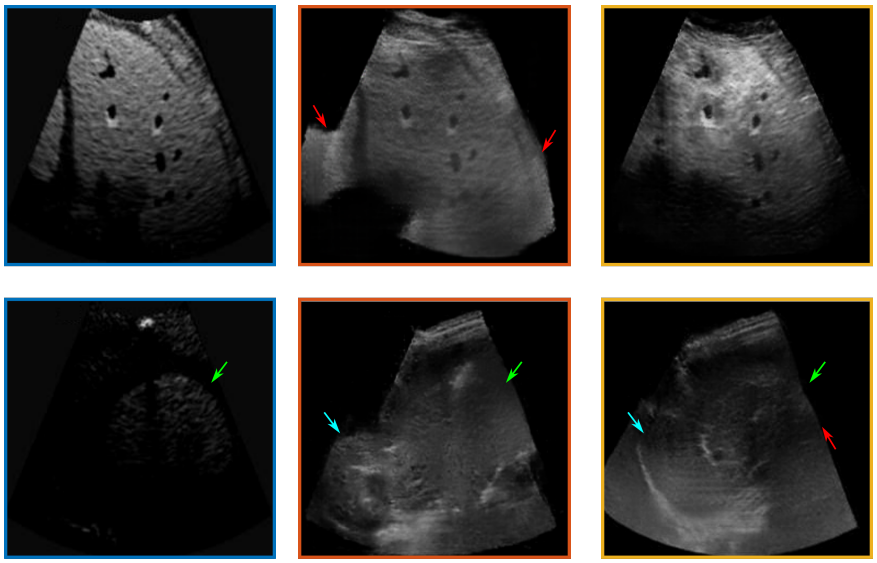
statistical significance of the differences in the responses for each group of images. Bonferroni correction was also applied in this case, adjusting the significance level of 5% to 1.25% (4 comparisons).

The outputs of the ray-casting algorithm were ranked with statistical significantly lower scores than the images generated with the CycleGAN$_{\text{U-Net}}$ ($p = 2.96 \times 10^{-7}$), the CycleGAN$_{\text{ResNet}}$ ($p = 3.04 \times 10^{-14}$) and the real US scans ($p = 6.14 \times 10^{-29}$). Introducing the CycleGAN$_{\text{U-Net}}$ reduced the variability of the responses, with most of them labeled as "rather fake" and with an increment in the mean Likert scale. This can be associated to the evident changes in the image intensity distribution that were qualitatively observed in Fig. 5 (c). The CycleGAN$_{\text{ResNet}}$, on the other hand, achieved higher realism values than the CycleGAN$_{\text{U-Net}}$, with a much wider range of responses and higher mean and median values. These differences were also statistically significant according to the hypothesis test ($p = 3.51 \times 10^{-5}$). The improvement in realism can be explained by the ability of the ResNet based model to reproduce bright artifacts such as those illustrated in Fig. 5 (d). On the other hand, these artifacts turn more difficult to interpret the content of the images, which could be associated to the median response ("cannot decide") and to the extra time taken by the volunteers to provide the answers. Finally, it is worth mentioning that the real US scans still reported statistically significant higher scores than our best simulation model ($p = 8.68 \times 10^{-15}$).

4.3 Model limitations

Fig. 8 presents some limitations of the current approach. We observed that both models eventually introduce slight deformations in the edges of the field of view (red arrows). This could be improved by training the models using images in polar coordinates. Since this transformation results in images without empty spaces at the boundaries, the generators will not produce fake responses there. On the other hand, in challenging poor quality scenarios such as the one illustrated in the second row of Fig. 8, both networks produce outputs with realistic artifacts (light blue arrows) but uncorrelated with the anatomical location. Also the organs with low contrast (green arrow) do not appear in the CycleGANs outputs (green arrows), and the networks are both introducing fake organs (light blue arrows) such as a kidney (Fig. 8b) or part of the liver (Fig. 8c). Both settings are in line with the observations in [4] regarding fake features produced by GANs trained with distribution matches losses. This might be prevented using a higher λ_{idt} coefficient for the identity loss or by incorporating further regularization based on organs locations. Improving the diversity of the real image set by adding US scans from alternative regions could also helped to overcome this limitation.

It is also worth mentioning that we trained our method using the fixed configuration of the ray-casting based simulation used in [23]. These parameters were previously chosen based on a qualitative analysis of the generated scans. Nevertheless, improving this configuration with more realistic acoustic parameters and scattering distributions might aid the CycleGAN model to further optimize the cycle consistency losses, ensuring even more realistic outcomes. However, this would demand an intensive hyperparameter search.



(a) Ray casting (b) U-Net (c) ResNet

Fig. 8: Failed cases of the simulation method. Red arrow: deformations outside the field of view. Green: missing organs. Light blue: hallucinated features.

5 Conclusions

In this work we introduced CycleGANs to improve realism in US simulation. We observed that incorporating this additional stage allows generating more realistic representations of the abdominal cavity than using only a ray-casting based algorithm. This conclusion is also supported by a preliminary user study performed by 21 experts in US imaging, which indicates that a ResNet architecture performs better than a U-Net when applied as model's generator. Further research in network architectures might significantly improve the quality of the results. To the best of our knowledge, this is the first study applying CycleGANs in this context. This method paves the way towards efficient, realistic and patient-specific simulation, which might be applied to improve simulators for training radiologists in US image interpretation and to produce synthetic data sets for training deep neural networks e.g. for US image segmentation[3].

Acknowledgements This work was funded by ANPCyT PICT 2016-0116 and PID-UTN SIUTNBA0005139. A NVIDIA GPU hardware grant supported this researchwith the donation of a Quadro P6000 graphic card. JIO is now a Postdoctoral Fellow at MedUniWien, funded by WWTF AugUniWien/FA7464A0249 (Medical University of Vienna); VRG12-009 (University of Vienna). We thank Lucia Llan de Rosos, Constantine Butakoff, Lidia Quinteros, Sergio Sánchez Martínez, Diego Pegoraro, Debbie Zhao, Matthieu De Craene, Gaurav Phadke and all the anonymous volunteers who participated in the user study. We also thank Claudia Marinelli and Rosana Cepeda from Instituto Multidisciplinario sobre Ecosistemas y Desarrollo Sustentable (UNICEN, Tandil, Argentina) for their assistance with the statistical analysis.

Compliance with ethical standards. This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

1. Ircad data set. `https://www.ircad.fr/research/3dircadb/`. Accessed: 2018-12-26
2. American College of Emergency Physicians: Use of ultrasound imaging by emergency physicians. Annals of emergency medicine **38**(4), 469 (2001)
3. Behboodi, B., Rivaz, H.: Ultrasound segmentation using u-net: learning from simulated data and testing on real data. arXiv preprint arXiv:1904.11031 (2019)
4. Cohen, J.P., Luck, M., Honari, S.: Distribution matching losses can hallucinate features in medical image translation. arXiv preprint arXiv:1805.08841 (2018)
5. D'Amato, J.P., Lo Vercio, L., Rubí, P., Fernández Vera, E., Barbuzza, R., del Fresno, M., Larrabide, I.: Efficient scatter model for simulation of ultrasound images from computed tomography data. In: 11th International Symposium on Medical Information Processing and Analysis, vol. 9681, p. 968105. International Society for Optics and Photonics (2015)
6. De Leeuw, J.R.: jspsych: A JavaScript library for creating behavioral experiments in a web browser. Behavior research methods **47**(1), 1–12 (2015)
7. De Luca, V., Tschannen, M., Székely, G., Tanner, C.: A learning-based approach for fast and robust vessel tracking in long ultrasound sequences. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 518–525. Springer (2013)
8. Dinh, V.A., Fu, J.Y., Lu, S., Chiem, A., Fox, J.C., Blaivas, M.: Integration of ultrasound in medical education at United States medical schools: a national survey of directors' experiences. Journal of Ultrasound in Medicine **35**(2), 413–419 (2016)
9. Engelhardt, S., De Simone, R., Full, P.M., Karck, M., Wolf, I.: Improving surgical training phantoms by hyperrealism: Deep unpaired image-to-image translation from real surgeries. arXiv preprint arXiv:1806.03627 (2018)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680 (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
12. Heer, I., Middendorf, K., Müller-Egloff, S., Dugas, M., Strauss, A.: Ultrasound training: the virtual patient. Ultrasound in Obstetrics and Gynecology **24**(4), 440–444 (2004)
13. Hiasa, Y., Otake, Y., Takao, M., Matsuoka, T., Takashima, K., Carass, A., Prince, J.L., Sugano, N., Sato, Y.: Cross-modality image synthesis from unpaired data using CycleGAN. In: International Workshop on Simulation and Synthesis in Medical Imaging, pp. 31–41. Springer (2018)
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
15. Kutter, O., Shams, R., Navab, N.: Visualization and GPU-accelerated simulation of medical ultrasound from CT images. Computer methods and programs in biomedicine **94**(3), 250–266 (2009)
16. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
17. Noble, J.A.: Reflections on ultrasound image analysis (2016)
18. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill **1**(10), e3 (2016)
19. Østergaard, M.L., Ewertsen, C., Konge, L., Albrecht-Beste, E., Nielsen, M.B.: Simulation-based abdominal ultrasound training–a systematic review. Ultraschall in der Medizin-European Journal of Ultrasound **37**(03), 253–261 (2016)
20. Petrusca, L., Cattin, P., De Luca, V., Preiswerk, F., Celicanin, Z., Auboiroux, V., Viallon, M., Arnold, P., Santini, F., Terraz, S., Scheffler, K., Becker, C.D., Salomir, R.: Hybrid ultrasound/magnetic resonance simultaneous acquisition and image fusion for motion monitoring in the upper abdomen. Investigative radiology **48**(5), 333–340 (2013)

21. Pham, A.H., Stage, B., Hemmsen, M.C., Lundgren, B., Pedersen, M.M., Jensen, J.A.: Simulation of shadowing effects in ultrasound imaging from computed tomography images. In: 2011 IEEE International Ultrasonics Symposium, pp. 1411–1414 (2011). DOI 10.1109/ULTSYM.2011.0349
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer (2015)
23. Rubi, P., Vera, E.F., Larrabide, I., Calvo, M., D'Amato, J., Larrabide, I.: Comparison of real-time ultrasound simulation models using abdominal CT images. In: 12th International Symposium on Medical Information Processing and Analysis, vol. 10160, p. 1016009. International Society for Optics and Photonics (2017)
24. Shams, R., Hartley, R., Navab, N.: Real-time simulation of medical ultrasound from ct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 734–741. Springer (2008)
25. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. International Conference on Learning Representations, ICLR 2017 (2017)
26. Terkamp, C., Kirchner, G., Wedemeyer, J., Dettmer, A., Kielstein, J., Reindell, H., Bleck, J., Manns, M., Gebel, M.: Simulation of abdomen sonography. evaluation of a new ultrasound simulator. Ultraschall in der Medizin **24**(04), 239–244 (2003)
27. Walcher, F., Weinlich, M., Conrad, G., Schweigkofler, U., Breitkreutz, R., Kirschning, T., Marzi, I.: Prehospital ultrasound imaging improves management of abdominal trauma. British Journal of Surgery: Incorporating European Journal of Surgery and Swiss Surgery **93**(2), 238–242 (2006)
28. Wang, C., Macnaught, G., Papanastasiou, G., MacGillivray, T., Newby, D.: Unsupervised learning for cross-domain medical image synthesis using deformation invariant cycle consistency networks. In: International Workshop on Simulation and Synthesis in Medical Imaging, pp. 52–60. Springer (2018)
29. Wein, W., Kamen, A., Clevert, D.A., Kutter, O., Navab, N.: Simulation and fully automatic multimodal registration of medical ultrasound. pp. 136–43 (2007). DOI 10.1007/978-3-540-75757-3_17
30. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)