Improving Realism in Abdominal Ultrasound Simulation Combining a Segmentation-Guided Loss and Polar Coordinates training

Otras autorías: Vitale, Santiago; Orlando, José Ignacio; Díaz, Alejandro; Larrabide, Ignacio

Autoría ditelliana: Iarussi, Emmanuel

Año: 2025

Nota: Este document es una versión previa. La versión final está publicada en Medical Physics (eISSN: 2473-4209)

¿Cómo citar la versión final de este trabajo?
Vitale S, Orlando JI, Iarussi E, Díaz A, Larrabide I. Improving realism in abdominal ultrasound simulation combining a segmentation-guided loss and polar coordinates training. Med Phys. 2025; 52: 4540–4556. https://doi.org/10.1002/mp.17801

Improving Realism in Abdominal Ultrasound Simulation Combining a Segmentation-Guided Loss and Polar Coordinates training

Santiago Vitale[1,2], José Ignacio Orlando[1,2], Emmanuel Iarussi[1,3], Alejandro Díaz[1,4], Ignacio Larrabide[1,2]

Correspondence: Santiago Vitale, Pladema, UNICEN, Tandil, Argentina. Campus Universitario, Paraje Arroyo Seco. Email: santiago.vitale@pladema.exa.unicen.edu.ar

Abstract

Background: Ultrasound (US) simulation helps train physicians and medical students in image acquisition and interpretation, enabling safe practice of transducer manipulation and organ identification. Current simulators generate realistic images from reference scans. Although physics-based simulators provide real-time images, they lack sufficient realism, while recent deep learning-based models based on unpaired image-to-image translation improve realism but introduce anatomical inconsistencies. **Purpose:** We propose a novel framework to reduce hallucinations from generative adversarial networks (GANs) used on physics-based simulations, enhancing anatomical accuracy and realism in abdominal US simulation. Our method aims to produce anatomically consistent images free from artifacts within and outside the field of view (FoV). **Methods:** We introduce a segmentation-guided loss to enforce anatomical consistency by using a pre-trained Unet model that segments abdominal organs from physics-based simulated scans. Penalizing segmentation discrepancies before and after the translation cycle helps prevent unrealistic artifacts. Additionally, we propose training GANs on images in polar coordinates to limit the field of view to non-blank regions. We evaluated our approach on unpaired datasets comprising 617 real abdominal US images from a SonoSite-M turbo v1.3 scanner and 971 artificial scans from a ray-casting simulator. Data was partitioned at the patient level into training (70%), validation (10%), and testing (20%). Performance was quantitatively assessed with Frechet and Kernel Inception Distances (FID and KID), and organ-specific χ^2 histogram distances, reporting 95% confidence intervals. We compared our model against generative methods such as CUT, UVCGANv2, and UNSB, performing statistical analyses using Wilcoxon tests (FID and KID with Bonferroni-corrected $\alpha = 0.01$, χ^2 with $\alpha = 0.008$). A perceptual realism study involving expert radiologists was also conducted. **Results:** Our method significantly reduced FID and KID by 66% and 89%, respectively, compared to CycleGAN, and by 34% and 59% compared to the leading alternative UVCGANv2 ($p \ll 0.01$). No significant differences ($p > 0.008$) in echogenicity distributions were found between real and simulated images within liver and gallbladder regions. The user study indicated our simulated scans fooled radiologists in 36.2% of cases, outperforming other methods. **Conclusions:** Our segmentation-guided, polar-coordinates-trained CycleGAN framework significantly reduces

[1]National Scientific and Technical Research Council (CONICET), Buenos Aires, Argentina
[2]Pladema Institute, UNICEN, Tandil, Argentina
[3]Laboratory of Artificial Intelligence, University Torcuato Di Tella, Buenos Aires,Argentina
[4]Facultad de Ciencias de la Salud, UNICEN, Olavarría, Argentina

hallucinations, ensuring anatomical consistency and realism in simulated abdominal US images, surpassing existing methods.

I. Introduction

Abdominal ultrasound (US) is an essential non-invasive imaging technique for diagnosing various abdominal conditions[2]. Effective clinical use requires specialists skilled in both image acquisition and interpretation. Typically, this training involves hands-on sessions with patients or volunteers, limiting scalability due to the need for devices and human subjects[3].

US simulation has emerged as a valuable training tool, allowing medical professionals to safely develop technical skills and procedural proficiency without needing real patients or equipment[4,5]. Simulators provide repeatable and controlled scenarios where users practice device manipulation[5], organ localization[6], and complex procedures[7]. Hence, these risk-free platforms contribute to improved clinical outcomes and increased confidence of clinicians to handle the complexities of real-world medical imaging. Additionally, US simulation supports applications like image registration[8] and expands datasets for deep learning[9], highlighting the necessity for realistic simulated images. High-fidelity simulations are crucial for achieving anatomical accuracy in training and clinical applications.

Several methods have been proposed to generate synthetic US images, such as ray-casting algorithms applied to CT volumes[10] or ray-tracing methods on deformable meshes[11,12]. While efficient, these physics-based approaches lack the realism needed for clinical training in image interpretation and diagnosis[13]. Recent generative models using convolutional neural networks have gained considerable attention for their enhanced realism[14]. These models have primarily focused on simulating images from specific areas of interest, such as intravascular[15] or fetal examinations[16], and regions like the brain[8], ovaries[17], kidneys[18], and musculoskeletal structures[19]. More complex regions, such as the abdominal cavity, have been less explored using these techniques. Previously, we applied an unpaired CycleGAN-based translation model[20] to improve ray-casting simulations[1]. While this refinement enhances the overall realism of the generated images, it suffers from hallucinated features typical of distribution matching losses[21]. In particular, the resulting scans include both unexpected organs in anatomically incorrect areas and distorted edges of the observable area captured by the device, typically referred to as the field of view (FoV).

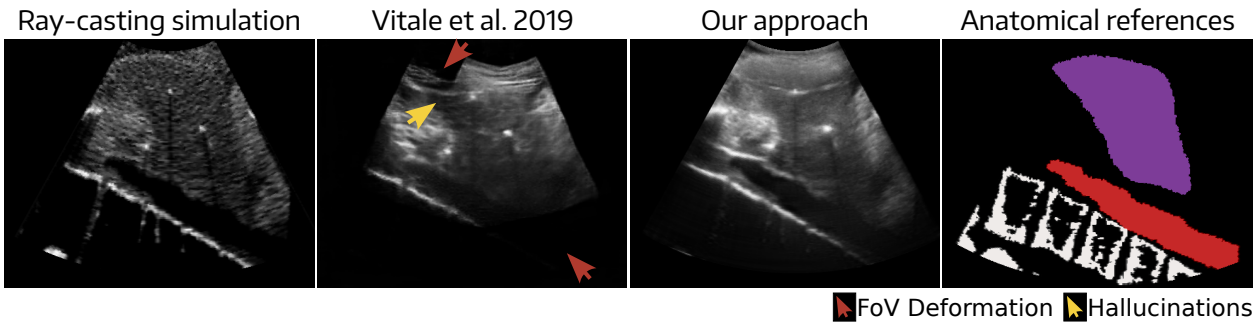


Figure 1: Examples of different artificial US scans obtained with a ray-casting model, our previous approach based on a standard CycleGAN model[1], and our improved method using a segmentation-guided loss and polar coordinates. Anatomical masks are provided as reference.

In this study, we propose some novel changes to our previous approach[1], with the goal of eliminating hallucinations and enabling the generation of anatomically consistent abdominal US scans from ray-casting-based simulations[22]. We achieve this by introducing a novel segmentation-guided loss, which leverages a pretrained Unet[23] segmentation model that penalizes differences between organ segmentations in the input image and its reconstructed versions after completing a full translation cycle. This information propagates through the entire cycle, compelling the fake-to-realistic generator to preserve anatomical consistency in the forward cycle. Otherwise, any hallucinations and unrealistic artifacts introduced will be propagated in the realistic-to-fake generator, and detected by the segmentation network. This aids to eliminate one of the sources of mistake, the hallucinations within organs. Additionally, we propose training our models directly in polar coordinates to remove irrelevant blank areas outside the field of view (FoV) and reduce artifacts in these regions. In summary, our key contributions with respect to our previous CycleGAN approach are threefold:

1) We introduce an objective term that enforces consistency between organ segmentations in the input scan, and its equivalent after the realism improvement transformation. To the best of our knowledge, such an "asymmetrical" approach for backpropagating anatomical knowledge have not been applied before to reduce hallucinations.

2) We adapted the training process to be directly applied to images in polar coordinates, eliminating empty spaces outside the FoV and preventing FoV deformations.

3) We demonstrate the model's generalization capability—unlike our previous patient-specific approach, the new model can be trained on multiple subjects and effectively applied to simulate

new individuals

Experimental results confirm that our approach significantly improves realism and anatomical accuracy over previous CycleGAN-based methods[1] and an improved ray-casting-based simulator.

II. Related work

The proposed model builds on top of our previous approach for improving realism in US simulations[1]. Originally, the method was based on a standard CycleGAN model[20], which allows image-to-image translation with unpaired samples. Technically, this model features two GANs, each defined by its own pair of generators and discriminators. In this context, it formally translates images from the domain $\mathcal{A}$ of **a**rtificially generated US images to another set $\mathcal{R}$ of **r**eal US images (both described in Section IV.A.1.), and viceversa. Formally, let $G_{\mathcal{A}\to\mathcal{R}}$ be the generator that translates an artificial image $a \in \mathcal{A}$ to $\mathcal{R}$, and $D_{\mathcal{R}}$ the discriminator that distinguishes between real images r and the translated ones $G_{\mathcal{A}\to\mathcal{R}}(a)$. On the other hand, let $G_{\mathcal{R}\to\mathcal{A}}$ be the generator that translates an image $r \in \mathcal{R}$ to the domain $\mathcal{A}$ while trying to avoid being detected by a discriminator $D_{\mathcal{A}}$. In the original CycleGAN definition, both pairs of networks are simultaneously trained by optimizing a linear combination of losses, including a standard adversarial penalty $\mathcal{L}_{\text{GAN}}$, a cycle-consistency term $\mathcal{L}_{\text{cyc}}$, and the identity loss $\mathcal{L}_{\text{idt}}$.

$\mathcal{L}_{\text{GAN}}$ is defined per each pair of generator and discriminator as follows:

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G_{\mathcal{A}\to\mathcal{R}}, D_{\mathcal{R}}, \mathcal{A}, \mathcal{R}) &= \mathbb{E}_{r\sim p_{\text{data}}(r)}[\log(D_{\mathcal{R}}(r)-1)^2] \\ &\quad + \mathbb{E}_{a\sim p_{\text{data}}(a)}[\log(D_{\mathcal{R}}(G_{\mathcal{A}\to\mathcal{R}}(a))^2], \\ \mathcal{L}_{\text{GAN}}(G_{\mathcal{R}\to\mathcal{A}}, D_{\mathcal{A}}, \mathcal{A}, \mathcal{R}) &= \mathbb{E}_{a\sim p_{\text{data}}(a)}[\log(D_{\mathcal{A}}(a)-1)^2] \\ &\quad + \mathbb{E}_{r\sim p_{\text{data}}(r)}[\log(D_{\mathcal{A}}(G_{\mathcal{R}\to\mathcal{A}}(r))^2],\end{aligned} \tag{1}$$

where $\mathbb{E}$ stands for the expected value of each corresponding data distribution, and each term is based on the least-squares GAN loss (LSGAN)[24], which prevents vanishing gradient issues.

To allow unpaired image-to-image translation, the training scheme incorporates an additional cycle-consistency loss $\mathcal{L}_{\text{cyc}}$. This term enforce that translations produced by one generator are reversible and retain the original domain's characteristics (Step 1, Figure 2). Formally, a forward cycle translates an image $a \in \mathcal{A}$ previously translated to domain $\mathcal{R}$ back to $\mathcal{A}$ (that is, $a \to G_{\mathcal{A}\to\mathcal{R}}(a) \to G_{\mathcal{R}\to\mathcal{A}}(G_{\mathcal{A}\to\mathcal{R}}(a)) \approx a$). Similarly, a reverse cycle ensures an image $r \in \mathcal{R}$

translated to domain $\mathcal{A}$ is brought back to $\mathcal{R}$ (by doing $r \rightarrow G_{\mathcal{R}\rightarrow\mathcal{A}}(r) \rightarrow G_{\mathcal{A}\rightarrow\mathcal{R}}(G_{\mathcal{R}\rightarrow\mathcal{A}}(r)) \approx r$). $\mathcal{L}_{\text{cyc}}$ can then be defined as the sum of two losses:

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G_{\mathcal{A}\rightarrow\mathcal{R}}, G_{\mathcal{R}\rightarrow\mathcal{A}}) = \mathbb{E}_{r\sim p_{\text{data}}(r)}[||G_{\mathcal{A}\rightarrow\mathcal{R}}(G_{\mathcal{R}\rightarrow\mathcal{A}}(r)) - r||_1] \\ + \mathbb{E}_{a\sim p_{\text{data}}(a)}[||G_{\mathcal{R}\rightarrow\mathcal{A}}(G_{\mathcal{A}\rightarrow\mathcal{R}}(a)) - a||_1].\end{aligned} \tag{2}$$

The identity loss $\mathcal{L}_{\text{idt}}$ regularizes the generators towards identity mappings, thereby biasing the models towards learning only what is needed to accurately generate realistic images:

$$\begin{aligned}\mathcal{L}_{\text{idt}}(G_{\mathcal{A}\rightarrow\mathcal{R}}, G_{\mathcal{R}\rightarrow\mathcal{A}}) = \mathbb{E}_{a\sim p_{\text{data}}(a)}[||G_{\mathcal{R}\rightarrow\mathcal{A}}(a) - a||_1] \\ + \mathbb{E}_{r\sim p_{\text{data}}(r)}[||G_{\mathcal{A}\rightarrow\mathcal{R}}(r) - r||_1].\end{aligned} \tag{3}$$

III. Methods

Figure 2 depicts a schematic representation of the training and test phases of our abdominal US simulation model. Our approach requires two sets of unpaired images for training, one with intermediate artificial US images ($\mathcal{A}$) and one with real US scans ($\mathcal{R}$). The first one is obtained by applying a ray-casting-based simulation algorithm[22] on cross-sectional 2D slices retrieved from multiple 3D CT scans and their associated 3D segmentation masks, based on the coordinates of an artificial probe. These 2D images are then transformed to polar coordinates to eliminate blank spaces outside the FoV and avoid the generative model hallucinating features outside the area. Images in $\mathcal{A}$, and their associated set of 2D segmentation masks ($\mathcal{M}$), are used offline to train a segmentation model S, which remains fixed later on while training our SG-CycleGAN model. This approach learns to map images from $\mathcal{A}$ to $\mathcal{R}$ and viceversa using two image-to-image translation models $G_{\mathcal{A}\rightarrow\mathcal{R}}$ and $G_{\mathcal{R}\rightarrow\mathcal{A}}$. The optimization minimizes a combined loss: a cycle-consistency term ($\mathcal{L}_{\text{cyc}}$) and a segmentation-guided term ($\mathcal{L}_{\text{sg}}$). The latter penalizes anatomical inconsistencies by comparing the predicted segmentations of the artificial scan and its reconstruction. During the testing phase, we input an intermediate artificial ultrasound image into the $G_{\mathcal{A}\rightarrow\mathcal{R}}$ generator, provided it was generated using the same ray-casting approach utilized during training. Doing so will yield a more realistic version of the original image.

In this study we propose to improve the previous approach by incorporating a novel segmentation-guided term ($\mathcal{L}_{\text{sg}}$) that enforces consistency between segmentation predictions of

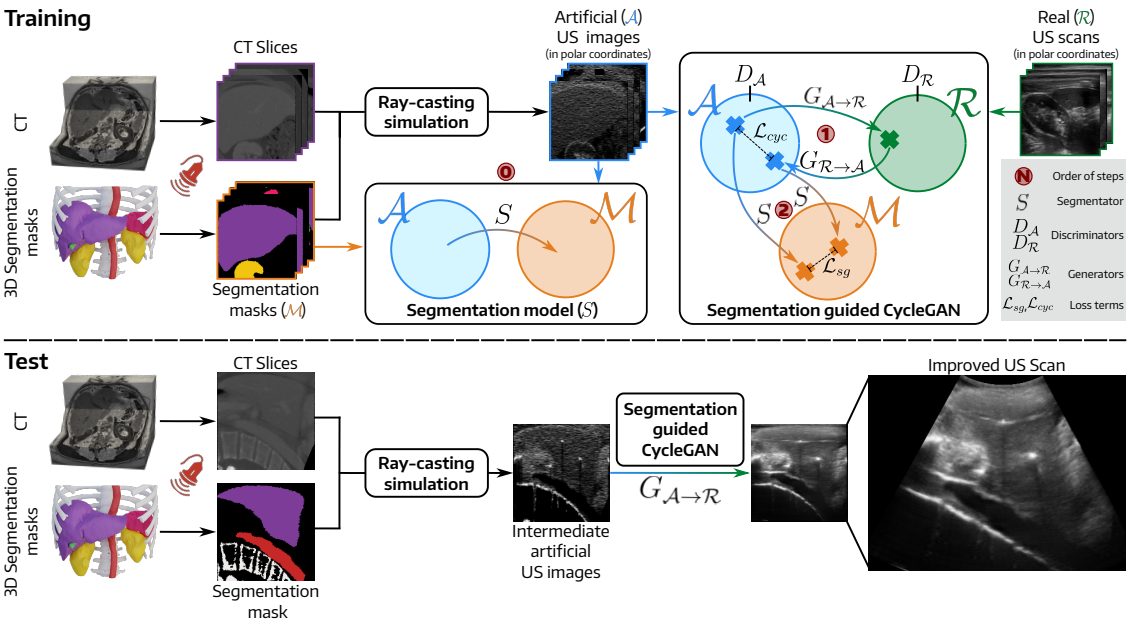


Figure 2: Schematic representation of the training (top) and testing (bottom) phases of our proposed approach for improving abdominal US simulation using a novel anatomically consistent image-to-image translation model.

images from $\mathcal{A}$ and their reconstructed counterparts. By penalizing discrepancies between the segmentation maps of the original and reconstructed fake images, the model is encouraged to maintain realistic anatomical features throughout the cycle during fake-to-real translation process. This consistency reduces the likelihood of introducing unrealistic artifacts and hallucinations, as any deviations from expected anatomical structures are penalized during training.

Figure 2 illustrates the proposed additional asymmetric objective, which integrates information about tissue locations within $a \in \mathcal{A}$ and enforces anatomical consistency between the original input and its reconstructed counterpart. Let $S(x)$ represent a deep neural network that produces a pixel-wise multiclass segmentation of a given input image x. The model S is trained offline using images $a \in \mathcal{A}$ and the corresponding segmentation masks, remaining fixed during the CycleGAN training phase (Step 0, Figure 2). During CycleGAN training, each image $a \in \mathcal{A}$ is translated into the $\mathcal{R}$ domain by the generator $G_{\mathcal{A}\to\mathcal{R}}$. The resulting image is subsequently translated back into the original domain by the generator $G_{\mathcal{R}\to\mathcal{A}}$ to obtain the reconstructed image (Step 1, Figure 2). Both the original image a and its reconstruction are segmented by S, yielding anatomical masks which are subsequently compared for consistency (Step 2, Figure 2). Formally, our proposed loss function, $\mathcal{L}_{\text{sg}}$, penalizes differences between $S(a)$ (the segmentation

map of an image $a \in \mathcal{A}$) and $S(G_{\mathcal{R}\to\mathcal{A}}(G_{\mathcal{A}\to\mathcal{R}}(a)))$ (the segmentation map of the reconstructed image after completing the full cycle):

$$\mathcal{L}_{\text{sg}}(G_{\mathcal{A}\to\mathcal{R}}, G_{\mathcal{R}\to\mathcal{A}}, S) = -\sum S(a)\log\left(S(G_{\mathcal{R}\to\mathcal{A}}(G_{\mathcal{A}\to\mathcal{R}}(a)))\right), \tag{4}$$

By means of this term, anatomical knowledge is transferred between generators, forcing $G_{\mathcal{A}\to\mathcal{R}}$ to preserve organs shape so that the reverse cycle through $G_{\mathcal{R}\to\mathcal{A}}$ does not produce an inconsistent sample.

In summary, the proposed training scheme is defined as a linear combination of the CycleGAN loss terms and the novel objective introduced above, namely:

$$\begin{aligned}\mathcal{L}(G_{\mathcal{A}\to\mathcal{R}}, G_{\mathcal{R}\to\mathcal{A}}, D_{\mathcal{A}}, D_{\mathcal{R}}, S) = {} & \mathcal{L}_{\text{GAN}}(G_{\mathcal{A}\to\mathcal{R}}, D_{\mathcal{R}}, \mathcal{A}, \mathcal{R}) \\ & + \mathcal{L}_{\text{GAN}}(G_{\mathcal{R}\to\mathcal{A}}, D_{\mathcal{A}}, \mathcal{R}, \mathcal{A}) \\ & + \lambda_{\text{cyc}} \cdot \mathcal{L}_{\text{cyc}}(G_{\mathcal{A}\to\mathcal{R}}, G_{\mathcal{R}\to\mathcal{A}}) \\ & + \lambda_{\text{idt}} \cdot \mathcal{L}_{\text{idt}}(G_{\mathcal{A}\to\mathcal{R}}, G_{\mathcal{R}\to\mathcal{A}}) \\ & + \lambda_{\text{sg}} \cdot \mathcal{L}_{\text{sg}}(G_{\mathcal{A}\to\mathcal{R}}, G_{\mathcal{R}\to\mathcal{A}}, S),\end{aligned} \tag{5}$$

where λ_{cyc}, λ_{idt} and λ_{sg} are hyperparameters that control the relative importance of each term in the final loss. Supplementary materials provide a flow chart with a visual representation of the calculation of the global loss throughout the training process.

Notice that the identity loss and the segmentation-guided loss serve different purposes in the model. The identity term enforces that each generator maintains features from the target domain that are already present in the source domain. On the other hand, our segmentation-guided loss focuses on preserving anatomical structure when transitioning from one domain to another.

IV. Experimental setup

IV.A. Materials

IV.A.1. Artificial US dataset

We generated a set of simulated images using 13 contrast-enhanced CT volumes (60% male) from the VISCERAL's Anatomy3 Challenge dataset[25]. To standardize the images, we manually cropped them to retain only the abdominal cavity, from the thoracic diaphragm to the pelvic inlet. Hounsfield Units (HUs) were then normalized to $[0, 1]$ using histogram equalization. A 2D Gaussian smoothing kernel of size 50×50 pixels (ranging from 34×34 mm to 37.5×37.5 mm, depending on voxel size) with a standard deviation of 2.5 pixels (approximately 1.7–1.875 mm) was applied to reduce high-frequency noise and improve uniformity.

For intermediate simulation, an artificial probe was placed at various abdominal locations to extract clinically relevant cross-sectional slices from both the CT scans and their segmentation masks (see Segmentation masks dataset). These slices served as inputs for a modified version of the ray-casting simulation algorithm by Rubí *et al.*[22] (see supplementary materials for further details). This process generated 926 artificial scans.

IV.A.2. Segmentation masks dataset

The anatomical masks used correspond to the cross-sectional slices extracted from the silver corpus segmentations of the 13 CT volumes in Artificial US dataset. The original dataset included segmentations of the spleen, liver, gallbladder, aorta, and kidneys. To provide additional anatomical references, we manually segmented the rib cage and spine.

IV.A.3. Real US scan dataset

Our real US dataset comprised 617 prospectively collected images from 11 volunteers (60% male, age $= 27 \pm 3$ years) with no known abdominal conditions. A specialist acquired these scans during routine abdominal exams using a SonoSite-M turbo v1.3 US Scanner (FUJIFILM, Bothell, USA). The scanning parameters differed from those used in the ray-casting model, as there is no direct correspondence between the device and the algorithm. All images were exported in JPEG format at 640×480 pixels.

IV.A.4. Dataset preprocessing and partition

To standardize spatial dimensions and align with the transducer's curvature, we applied a Cartesian-to-Polar transformation to both artificial and real ultrasound scans. This process involved calculating the center, inner and outer radii, and angular range (θ) for each image. For simulated images, these parameters were derived from the ray-casting algorithm, while for real scans, they were manually extracted using FoV masks. This transformation corrected the transducer's curvature and removed non-informative areas (see supplementary materials for a graphical explanation). The final images were resized to 256×256 pixels and randomly partitioned at the patient level into training (70%, 8 patients), validation (10%, 2 patients), and test (20%, 3 patients) subsets.

IV.B. Architectures

IV.B.1. Generator architecture

We evaluated three generator architectures, all based on a Unet encoder-decoder network. The first was a standard Unet[23] (Unet in our experiments), where the decoder was replaced with nearest-neighbor upsampling followed by a convolutional layer to prevent checkerboard artifacts[1]. The second was a modified Unet with bottleneck layers and residual connections[26] (ResUnet in our experiments), implemented in two width variations. Lastly, we included the densely connected image-to-image translation generator by Dangi *et al.*[27] (DenseUnet in our experiments). All generators used a tanh activation function. Further architectural details are provided in the supplementary materials.

IV.B.2. Discriminator architecture

Following previous studies[16,17,28], we employed a 70×70 patchGAN[29] as the discriminator. The network consists of four convolutional blocks, each with a 4×4 kernel and a stride of 2. Instance normalization was used instead of batch normalization, as it has been shown to enhance diversity and prevent mode collapse[30,31]. Each block applies Leaky-ReLU activation, as commonly done in patchGANs[29], progressively reducing spatial dimensions while increasing feature maps to 64, 128, 256, and 512, respectively. A final 1-filter convolution, followed by a sigmoid activation function, produces the output probability for each patch.

IV.B.3. Segmentation model

The segmentation network S is based on a Unet architecture. The encoder consists of four convolutional blocks with 64, 128, 256, and 512 filters, each followed by 2×2 max-pooling for downsampling. Each block comprised a sequence of a 3×3 convolutional layer, a batch normalization operation, and a ReLU activation, repeated twice. A bottleneck layer with 1024 filters precedes the decoder, which uses nearest-neighbor upsampling followed by convolutional layers with progressively fewer filters, from 512 down to 64. A final 1×1 convolutional layer produces class logits, converted into probabilities using softmax activation. The network was trained to segment the liver, spleen, gallbladder, aorta, and kidneys. Since the kidney consists of two ultrasound-differentiable structures—the hyperechoic renal pelvis and the hypoechoic renal cortex—we treated them as separate classes, using weak annotations for each (see supplementary for further details).

IV.C. Model configuration

Hyperparameters were empirically selected based on validation set performance using Fréchet Inception Distance (FID). In tied cases, we visually inspected the results and chose parameters that produced more realistic and anatomically consistent images. Coefficients λ_{cyc}, λ_{idt}, and λ_{sg} were experimentally fixed to 10, 0.5, and 0.5, respectively. We found that a higher λ_{cyc} improved cycle consistency in image translations. We trained the model for 200 epochs using Adam[32] optimization with an initial learning rate of 2×10^{-4} and a batch size of 4. After 100 epochs, the learning rate was reduced linearly by $\frac{1}{101}$. The segmentation network S was trained offline using a multiclass cross-entropy objective, Adam optimization with an initial learning rate of 1×10^{-4}, and a batch size of 16 for 150 epochs. The learning rate was decreased by a factor of 0.5 every time that the performance plateaued for 20 epochs, measured by the average Dice coefficient. Hyperparameters were selected to maximize the average Dice score for all organs in the validation set.

All CNNs, including the segmentation network, were implemented in Pytorch 1.10.0 and trained on a desktop workstation with an AMD Ryzen 9 5900X CPU and an NVIDIA GeForce RTX 3060 GPU (12GB RAM).

IV.D. Baselines for comparison

We compared SG-CycleGAN with the ray-casting-based method[10] used to generate the input images and four state-of-the-art image-to-image translation models. Given the limited number of models available for unpaired datasets in this task, we focused on CycleGAN-based approaches, which have shown promise in US simulation. First, we compared SG-CycleGAN to our previously published CycleGAN[1], trained with images in Cartesian coordinates. Second, we included the Contrastive Unpaired Translation (CUT)[33] model, which has been used as a baseline for obstetric US simulation[34]. To incorporate recent advances, we tested the UNet Vision Transformer cycle-consistent GAN (UVCGANv2)[35], which integrates a U-Net with a Vision Transformer encoder. Finally, we included the Unpaired Neural Schrödinger Bridge (UNSB)[36], a diffusion-based model that provides an alternative to GANs and has been applied to US simulation[37]. This selection covers both standard approaches and recent innovations in generative modeling for US simulation.

IV.E. Evaluation metrics & statistical analysis

Assessing the quality and realism of simulated US scans is challenging, as in any image generation task[38,39]. The most widely used metrics are Fréchet Inception Distance (FID)[40] and Kernel Inception Distance (KID)[41], which have been applied in various US studies[16,17,34,42]. Both metrics quantify the statistical distance between feature distributions of real and artificial images, extracted from an Inception v3[43] network pretrained on ImageNet. This comparison captures macro-level differences in speckle noise texture. A lower FID score indicates that the generated images better resemble real US scans in terms of noise and echogenicity. We used the intermediate 768-feature layer to avoid highly specialized low-level descriptors[34]. For evaluation, we used the validated TorchFidelity implementation[44]. Statistical significance was assessed using one-tailed Wilcoxon signed-rank tests, with Bonferroni correction[45] adjusting the significance level from 0.05% to 0.01% (5 comparisons). Effect sizes were evaluated using Cohen's d[46], which measures differences relative to the pooled standard deviation. According to Cohen's criteria, 0.2 represents a small effect size and indicates that the difference between groups is noticeable but not substantial; 0.5 represents a medium effect size, suggesting a moderate difference that is likely to be meaningful in most contexts; and 0.8 represents a large effect size, indicating a substantial difference between groups, which is often considered to be practically significant. Very small effects (below 0.2) indicate negligible differences that may not have practical relevance.

Values greater than 1, on the other hand, are considered very large, and highlight a difference that is both statistically and practically significant.

The χ^2 distance[47], commonly used in US simulation[16], quantifies dissimilarities between image histograms:

$$\chi^2(h_A \| h_B) = \frac{1}{2} \sum_{l=1..d} \frac{(h_A[l] - h_B[l])^2}{h_A[l] + h_B[l]}, \tag{6}$$

where d is the number of histogram bins (50 in our case). While alternatives like Jensen-Shannon (JS) divergence[48] compare entire histograms, we opted for χ^2 as it is more sensitive to relative differences in individual bins.

Histogram-based methods are affected by intensity shifts and contrast variations[49]. To evaluate potential mismatches in echogenicity, we compared intensities locally within segmented gallbladder, liver, and kidney regions. Segmentation masks were slightly eroded using a 5×5 structuring element to reduce edge irregularities. Pairwise χ^2 distances from real US images were used as reference values. To ensure fair comparisons, scans with minimal tissue representation were excluded, and histograms were normalized by the number of pixels within each mask. Statistical significance was tested using a one-tailed Wilcoxon rank-sum test with a Bonferroni-corrected threshold of 0.0083 (6 comparisons), alongside effect size analysis via Cohen's d. Notice that if simulations are realistic, the χ^2 distance distribution for each organ should closely match that of real scans, showing no significant differences. For all metrics, 95% confidence intervals (95% CI) were computed using bootstrap resampling ($N = 1000$).

Finally, we assessed anatomical accuracy by comparing segmentation masks from our method and standard CycleGAN using mean Intersection over Union (mIoU). These masks were created by manually segmenting a set of 16 simulated images and comparing the resulting organ masks with those used as input to the physical model.

IV.F. User study-based evaluation

We further evaluated our approach through a custom-made online user study, implemented using the jsPsych JavaScript library[50] (see supplementary materials for further details). The study comprised two experiments. The first experiment assessed experts' ability to distinguish real

from simulated US images. Participants were shown a US scan and asked to classify it as real or simulated. The dataset included 45 images: 15 real US scans, 15 generated by our approach, and 15 by the original CycleGAN model. Classification accuracy was measured as the fraction of correctly identified real and fake images. The second experiment evaluated anatomical preservation in the generated images. Experts were presented with two simulated scans—one generated with and one without the segmentation-guided term—and asked to select the scan with better anatomical preservation. The original segmentation mask was provided as a reference. This test included 10 scan pairs covering typical abdominal capture windows such as intercostal, subcostal margin, longitudinal, oblique, and transverse views. A total of 16 clinicians, all experts in US imaging, participated in the study, most of whom were affiliated with Sociedad Argentina de Ultrasonido en Medicina y Biología (SAUMB).

V. Results

We conducted a comprehensive evaluation of the proposed approach, using both qualitative and quantitative approaches. Our method was compared to state-of-the-art techniques outlined in Subsection IV.D., elaborated upon in Subsection V.A.2.. Additionally, an ablation study was carried out to evaluate the impact of each design choice on the final results, detailed in Subsection V.B..

V.A. Simulation performance

V.A.1. Qualitative evaluation

Figure 3 presents example simulations generated using the original CycleGAN in Cartesian coordinates[1], the same model in polar coordinates, and our SG-CycleGAN. The samples correspond to different abdominal windows commonly used in clinical analyses.

The Cartesian CycleGAN results exhibit FoV deformations in all scans, mainly as irregular edges (Figures 3 (a) and (d)). In some cases, these distortions remove anatomical structures, such as part of the liver (Figures 3 (a) and (c)), the aorta (Figures 3 (c) and (d)), or the kidneys (Figures 3 (a) and (e)). Alternatively, using polar coordinates ensures images that are consistent with the input FoV, with both the standard CycleGAN and our proposed SG-CycleGAN, preserving

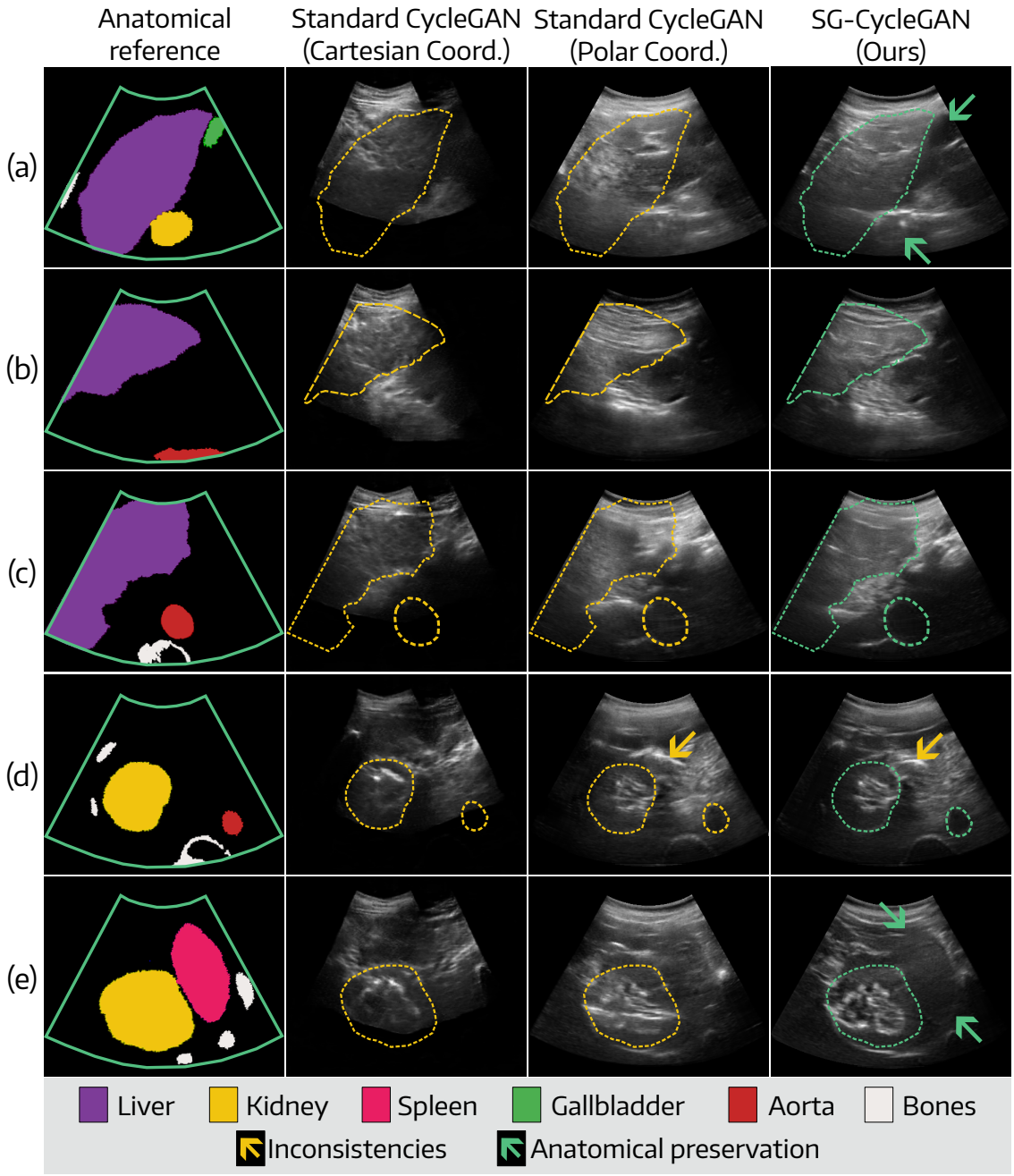


Figure 3: Qualitative results for abdominal US simulation obtained using a standard CycleGAN trained in Cartesian and polar coordinates and our proposed SG-CycleGAN approach. Dotted lines indicate inconsistent organs (yellow) and their improved counterparts (green). From top to bottom: (a) right subcostal margin, (b) longitudinal, (c) oblique, and (d,e) right and left intercostal acquisition windows.

all the organs that are present in the images.

Figures 3 (a)–(c) show that standard CycleGANs introduce inhomogeneities in the liver, appearing as hallucinated shadows (Figures 3 (a) and (c)) or anatomically inconsistent hyperechoic structures (Figures 3 (b) and (c)). Our segmentation-guided approach preserves liver structure, maintaining homogeneous echogenicity (green contours).

Figures 3 (d) and (e) present results for windows that include part of the kidney. Training with Cartesian coordinates produces unrealistic kidneys, with artifacts such as hyperechoic reflections that are inconsistent with this anatomical area (Figure 3 (d)), or intensities of the renal pelvis below the usual echogenicities (Figure 3 (e)). Similarly, Figures 3 (c) and (d) show poor aorta representations, which disappear into larger anechoic areas. While polar coordinates mitigate this issue, they still generate anatomical inconsistencies (e.g., hyperechoic streaks in the kidney or diffuse spleen edges in Figure 3 (e)). Our approach better preserves organs, yielding anatomically accurate results for the gallbladder (Figure 3 (a)), aorta (Figures 3 (b) and (c)), bones (Figures 3 (c)–(e)), kidneys (Figures 3 (a), (d), and (e)), and spleen (Figure 3 (e)). On this last area, a better scattering effect can be observed on top of the artifact generated by the skin (top green arrow), as well as more defined interfaces at the bottom (bottom green arrow).

Figure 4 visually compares our method to other baselines. Further qualitative results are provided in the supplementary materials. The previous CycleGAN model reduces the FoV, removing image regions (e.g., the missing backbone in Figure 4 (c) or the truncated kidney in Figure 4 (e)). CUT better preserves anatomical structures but still producing hallucinations such as a hyperechoic artifact in the liver (Figure 4 (a)) and an anechoic tubular formation in the kidney (Figure 4 (c)). It also fails to maintain spleen integrity (Figure 4 (e)). UVCGANv2 struggles to maintain structures, reducing gallbladder size (Figure 4 (b)) and distorting kidneys (Figures 4 (c) and (e)). The UNSB model preserves structures like the liver, gallbladder, and vessels (see Figure 4 (a), (b) and (d), respectively), but struggles with kidney structures, where it hallucinates anechoic formations (Figure 4 (c) and (e)). Additionally, it fails to simulate the skin layer artifacts, which are captured in the other models. Finally, our model corrects the FoV limitations observed in our previous version, while also preserving all the anatomical structures provided in the ray-casting based input.

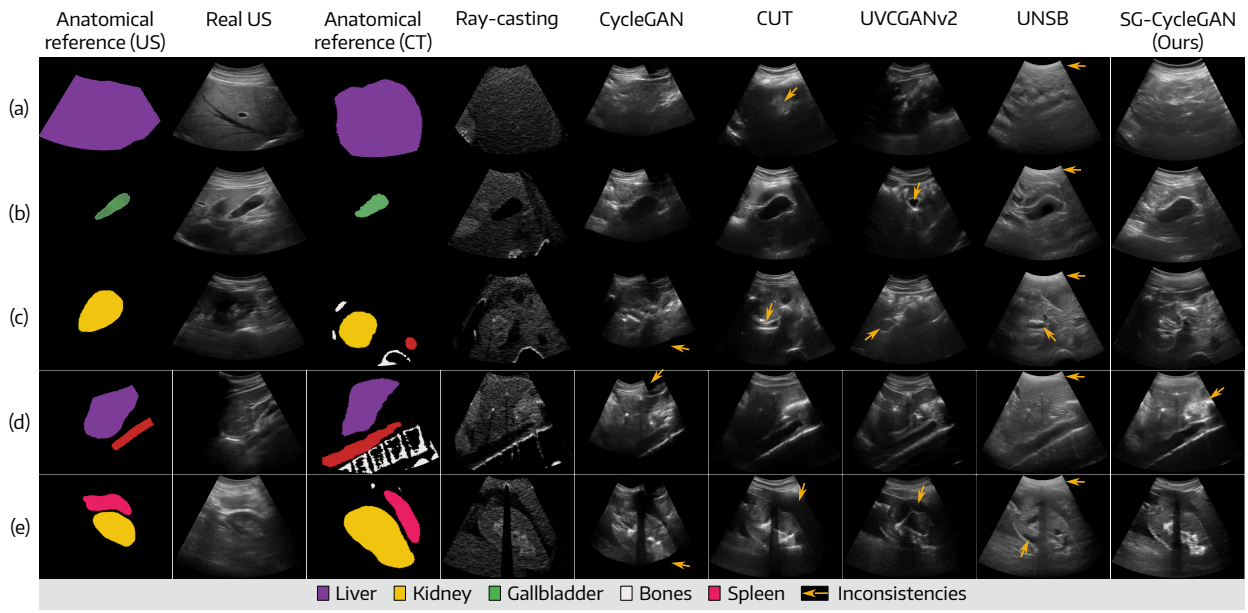


Figure 4: Qualitative examples for each model and their associated segmentations as reference.Yellow arrows indicate inconsistencies.

V.A.2. Quantitative evaluation

Table 1 compares our approach to all baselines detailed in Section IV.D.. While all generative models outperform the physics-based simulator, our SG-CycleGAN achieves statistically significant reductions in FID (80%) and KID (97%) ($p < 0.01$). Very large effect sizes (Cohen $d = 84.24$ and 79.97) further support these findings. Among baselines, UVCGAN performed best, but still lags behind our method, with substantial Cohen effect sizes ($d = 10.43$ for FID and $d = 9.20$ for KID).

Our SG-CycleGAN also exhibits χ^2 distances within the liver and gallbladder that closely resemble those observed between real images (Table 1). Figure 5 (A) provides a detailed analysis of this metric for each tissue, with colored boxplots representing the distribution of pairwise χ^2 distances between simulated and real US images, and gray boxplots representing the reference distribution between real scans. Although these cannot be compared directly one other for being calculated using different samples, it can be observed that methods incorporating generative approaches achieve χ^2 distances that distribute approximately similar as in real images, for all organs. All generative models produce echogenicities in the gallbladder that are statistically indistinguishable from those in real US images, with p values greater than 0.021. However, it should be noted that our model, like CUT and UVCGANv2, presents closer mean values and a very low Cohen's d value (< 0.09), indicating a very small effect size compared to the rest

models, which have values close to 0.2. Within the liver, our SG-CycleGAN and UNSB model achieved distances comparable to the distances observed between real images. In this case, the statistical tests performed between these models and real US images showed no statistically significant differences, with $p > 0.127$ for all comparisons. On the contrary, performing the same comparison between CUT, CycleGAN and UVCGANv2 exhibited statistically significant differences ($p < 0.008$). Nonetheless, all models exhibit a small effect size (Cohen's $d < 0.16$), with the UNSB model standing out with a Cohen's d of 0.01. In the the kidney, the CycleGAN and the UNSB did not exhibit statistically significant differences when compared to real US images, with $p > 0.183$, showing a very small effect size (Cohen $d < 0,09$).

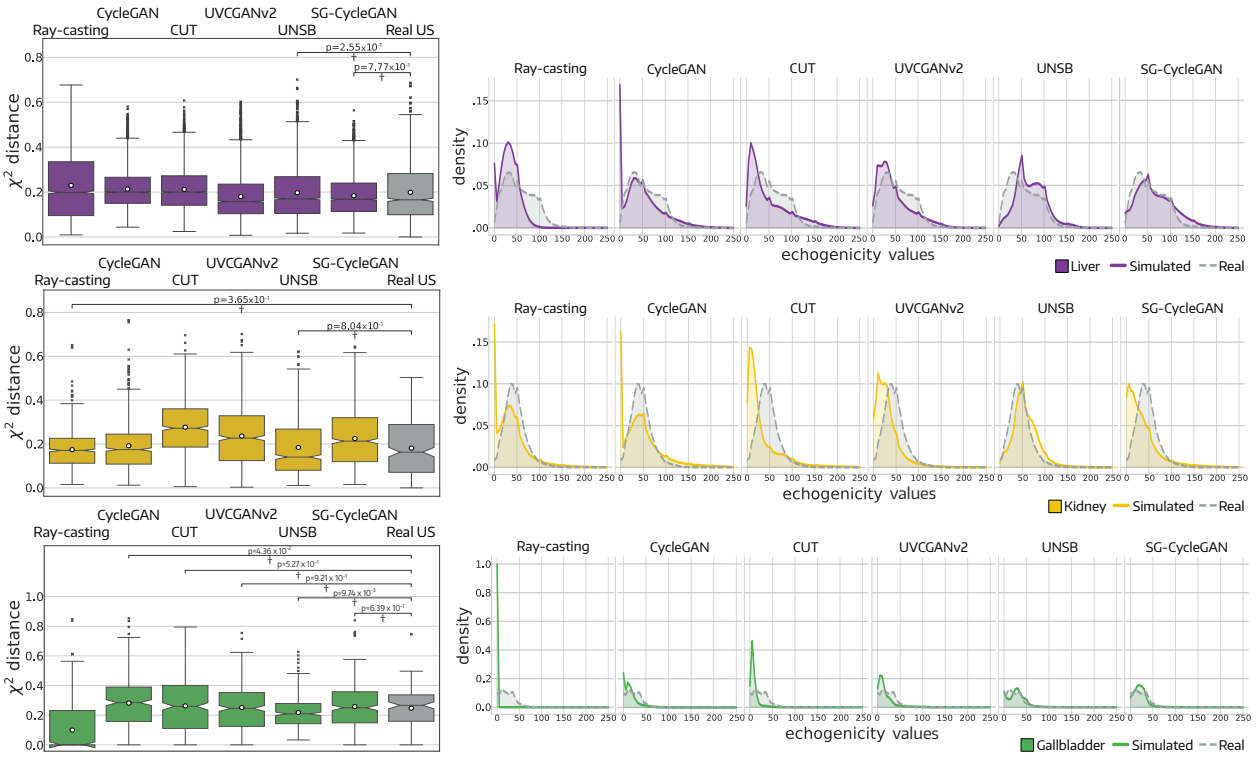
To further illustrate echogenicity similarities, Figure 5 (b) presents histograms of cumulative intensity distributions for each organ. These histograms differ from those used for organ-specific χ^2 comparisons in Table 1 and Figure 5 (A). Consistent with previous observations, our model produces intensities that closely resemble real images, particularly in the liver and gallbladder. For the kidney, UNSB outputs are more similar to real images.

We also report training and inference time comparisons in the supplementary material. SG-CycleGAN increased training time from 95 seconds (standard CycleGAN) to 127 seconds per epoch, similar to CUT and notably faster than UVCGANv2 and UNSB. For inference, SG-CycleGAN and CycleGAN were the fastest at 0.0813 seconds per scan, while other models required 2–3 times longer.

V.B. Ablation analysis

V.B.1. Quantitative evaluation

Table 2 presents results from CycleGAN models trained with different strategies. Models using polar coordinates (rows 2 to 4) achieved better FID and KID scores than the Cartesian-based model (row 1). However, improvements in χ^2 distances appeared only in the gallbladder and liver when combined with the segmentation-guided loss and LSGAN objective. Regarding adversarial loss, LSGAN outperformed the vanilla loss (Table 2 rows 2 and 3). The best results were achieved by incorporating the segmentation-guided loss (row 4), which further improved FID and KID scores. In terms of anatomical preservation relative to ground truth label maps, our model achieved a higher overall mIoU (0.68) than the standard CycleGAN (0.59). For individual organs



(a) (b)

Figure 5: Organ-wise quantitative evaluation. (A) Box plots illustrating the distribution of pairwise χ^2 distances between pairs of simulated and real US images for each organ of interest (colored), and between pairs of real US images (gray).p-values are included for all comparison where no statistical differences observed. (B) Histograms representing the distribution of echogenicity values for each organ, for simulated (colored) and real (gray) images.

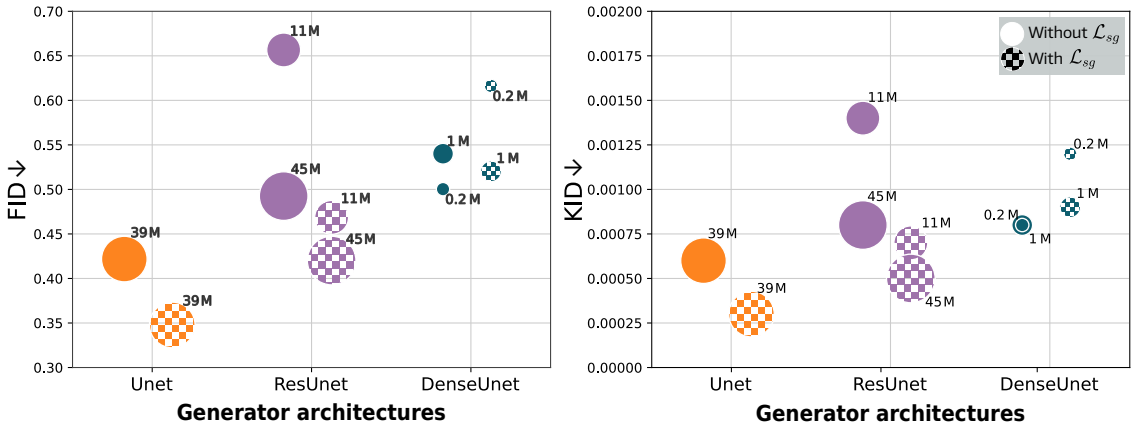


Figure 6: FID and KID results for different architectures of generator models. Each network was trained with (right) and without (left) our proposed loss term. The bubble size is proportional to the number of parameters of each model, indicated in millions (M) on top of each one.

Table 1: Quantitative comparison of the proposed model with respect to other alternatives in terms of FID and KID distances (lower value, marked as ↓, is better), and mean χ^2 distances for different organs of interest. Asterisks (*) next to FID and KID values indicate statistically significant differences, when compared to our approach ($p < 0.01$). χ^2 distances between pairs of real scans are included as a reference (closer to this reference is better). Daggers (†) in χ^2 distances indicate no statistical differences with the real scans ($p > 0.008$). Sub-indices indicate Cohen's d values. Best values are indicated in bold. Last row corresponds to the number of real and simulated US used to calculate each metric.

Model	FID ↓ [95% CI]	KID ↓($\times 10^{-3}$) [95% CI]	χ^2 [95% CI] Liver	Kidney	Gallbladder
Ray-casting[22]	1.73 [1.69 - 1.77]*$_{84.24}$	5.02 [4.85 - 5.19]*$_{79.97}$	0.23 [0.02 - 0.54]$_{0.21}$	0.17 [0.03 - 0.34]†$_{0.06}$	0.09 [0.0 - 0.50]$_{0.90}$
CycleGAN[1]	0.99 [0.96 - 1.03]*$_{48.63}$	2.61 [2.48 - 2.74]*$_{46.83}$	0.21 [0.07 - 0.41]$_{0.13}$	0.19 [0.03 - 0.47]$_{0.09}$	0.28 [0.02 - 0.65]†$_{0.22}$
CUT[33]	0.80 [0.76 - 0.84]*$_{27.25}$	1.90 [1.74 - 2.06]*$_{26.79}$	0.21 [0.06 - 0.42]$_{0.12}$	0.28 [0.05 - 0.55]$_{0.74}$	0.26 [0.0 - 0.66]†$_{0.09}$
UVCGANv2[35]	0.48 [0.45 - 0.51]*$_{10.43}$	0.69 [0.59 - 0.79]*$_{9.20}$	0.17 [0.03 - 0.43]$_{0.16}$	0.23 [0.03 - 0.52]$_{0.41}$	0.25 [0.00 - 0.54]†$_{0.03}$
UNSB[36]	0.95 [0.90 - 0.99]*$_{36.23}$	2.42 [2.26 - 2.58]*$_{37.31}$	**0.19 [0.04 - 0.46]†$_{0.01}$**	0.18 [0.02 - 0.50]†$_{0.03}$	0.22 [0.05 - 0.45]†$_{0.23}$
SG-CycleGAN (ours)	**0.33 [0.32 - 0.35]**	**0.28 [0.25 - 0.31]**	0.18 [0.05 - 0.40]†$_{0.13}$	0.22 [0.03 - 0.48]$_{0.33}$	**0.25 [0.00 - 0.53]†$_{0.07}$**
Real US	-	-	0.19 [0.00 - 0.51]	0.18 [0.00 - 0.45]	0.24 [0.00 - 0.48]
Number of scans $\mathcal{R}\|\mathcal{A}$	213|213	213|213	40|90	16|48	12|28

(liver, kidney, gallbladder), our model outperformed CycleGAN with IoU values of 0.84, 0.93, and 0.86, respectively, compared to 0.75, 0.89, and 0.81, demonstrating superior anatomical fidelity. We also analyzed the impact of different generator architectures by comparing FID and KID metrics across network types and backbone sizes (Figure 6). The standard Unet consistently outperformed ResUnets and DenseUnets in FID and KID scores. Additionally, adding $\mathcal{L}_{sg}$ improved performance across all networks, except for the DenseUnet with the smallest capacity (0.2 million parameters).

V.B.2. Qualitative effect of using polar coordinates

To assess the impact of using polar instead of Cartesian coordinates for training, Figure 7 compares input simulations from the ray-casting algorithm with their improved versions using both alternatives. All scans share the same FoV, outlined in green. With Cartesian coordinates, the model either restricts the original FoV (left edge of image (a)) or introduces organs outside of it (bottom of both scans). In Figure 7 (b), the network hallucinates large shadowed areas near the contours while partially preserving original image details (yellow arrow, left side), creating false tissue reflections beyond the incorrect FoV. In contrast, images generated in polar coordinates remain confined to the pre-defined FoV, free of deformations or hallucinated artifacts. Figure 7

Table 2: Evaluation of the ablation test in terms of FID, KID (lower value, marked as ↓, is better) and mean χ^2 distances for different organs. Asterisks (*) next to FID and KID values indicate statistically significant differences ($p < 0.016$), when compared to our approach. χ^2 distances between pairs of real scans are included as a reference (closer to this reference is better).Daggers (†) in χ^2 distances indicate no statistical differences with the real scans ($p > 0.012$). Sub-indices indicate Cohen's d values. The best values are indicated in bolds. The last row corresponds to the number of real and simulated US images used to calculate each metric respectively.

Model	Adversarial loss	Coordinate space	FID ↓ [95% CI]	KID ↓ ($\times 10^{-3}$) [95% CI]	χ^2 Liver	χ^2 Kidney	χ^2 Gallbladder
CG	Vanilla	C	0.99 [0.96-1.03]*$_{48.63}$	2.61 [2.48 - 2.74]*$_{46.83}$	0.21 [0.07 - 0.41]$_{0.13}$	**0.19 [0.03 - 0.46]**$_{0.09}$	0.28 [0.02 - 0.65]$_{0.22}$
CG	Vanilla	P	0.73 [0.71 - 0.76]*$_{36.12}$	1.82 [1.72 - 1.92]*$_{38.93}$	0.26 [0.09 - 0.52]$_{0.49}$	0.29 [0.03 - 0.56]$_{0.85}$	0.23 [0.00 - 0.53]$_{0.08}$
CG	LSGAN	P	0.42 [0.40 - 0.44]*$_{7.48}$	0.38 [0.33 - 0.43]*$_{9.64}$	0.21 [0.05 - 0.44]†$_{0.05}$	0.22 [0.04 - 0.47]$_{0.25}$	0.27 [0.00 - 0.54]$_{0.04}$
SG	LSGAN	P	**0.33 [0.32 - 0.35]**	**0.28 [0.25 - 0,31]**	**0.18 [0.05 - 0.40]**†$_{0.13}$	0.22 [0.03 - 0.48]$_{0.34}$	**0.25 [0.00 - 0.53]**†$_{0.07}$
Real US			-	-	0.19 [0.00 - 0.51]	0.18 [0.00 - 0.45]	0.24 [0.00 - 0.48]
Number of scans ($\mathcal{R}$|$\mathcal{A}$)			213|293	213|293	40|90	16|48	6|27

Abbreviations: CG, Standard CycleGAN; SG, SG-CycleGAN; Vanilla, Jensen-Shannon divergence loss; LSGAN, least squares GAN loss; C, cartesian; P, Polar

also includes patches illustrating speckle noise patterns. Unlike input simulated scans, Cartesian-based outputs exhibit randomly oriented patterns, misaligned with the US transducer. Polar coordinates mitigate this issue, producing more realistic lateral speckle orientations consistent with the convex transducer's azimuthal angle.

V.B.3. Qualitative effect of the generator architecture

Figure 8 compares results from SG-CycleGAN using different generator architectures. All generative models enhance overall brightness, but the ResUnet introduces bright artifacts that are anatomically inconsistent, such as in the renal pelvis (Figure 8 (a), yellow arrow) and an unsegmented region (Figure 8 (b), green arrow). Additionally, ResUnet produces an overly blurred and poorly defined speckle pattern. In contrast, the Unet and DenseUnet backbones yield better intensity distributions while preserving organ shapes and boundaries (e.g., the aorta in Figure 8 (a), red arrow). The kidney (Figure 8 (a), yellow arrows) also shows well-defined interfaces both externally and within the renal pelvis. These networks generate more realistic speckle noise patterns (e.g., in the liver, Figure 8 (b)), though DenseUnet hallucinates interfaces in unsegmented areas compared to Unet (green arrow).

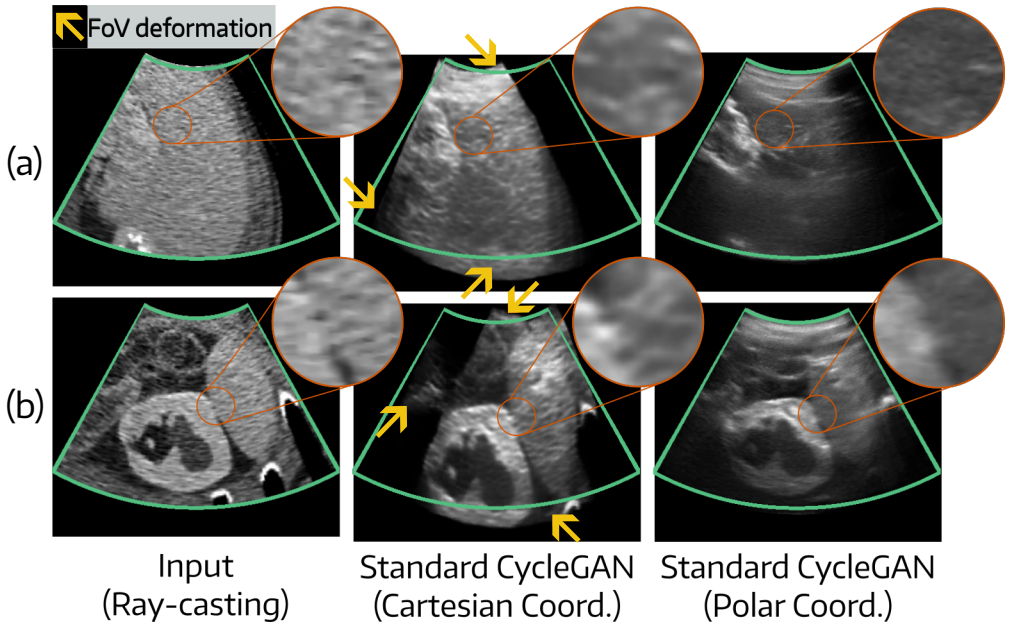


Figure 7: Comparison of simulated images with CycleGANs trained on different coordinate systems. Green boundaries indicate the original FoV.

V.B.4. User study

Figure 9 presents the user survey results. Figure 9.A) shows bar charts of user accuracy in classifying images—generated by CycleGAN, SG-CycleGAN, or real US, as fake or real. The average and standard deviation for each type are also included. Lower accuracy indicates more frequent misclassification of fake images as real and vice versa. Most participants correctly identified CycleGAN-generated images as fake with high accuracy (98%), reflecting their lower realism. However, for SG-CycleGAN images, accuracy averaged 63.75%, meaning 36.25% were mistaken for real. This trend is also evident in real US scan classification, where expert accuracy averaged below 80%. Figure 9.B) presents a pie chart summarizing radiologists' responses on anatomical preservation. When asked about the preservation of the anatomy in fake images generated with both synthetic methods, 81.6% of cases favored SG-CycleGAN to be more anatomically consistent over CycleGAN.

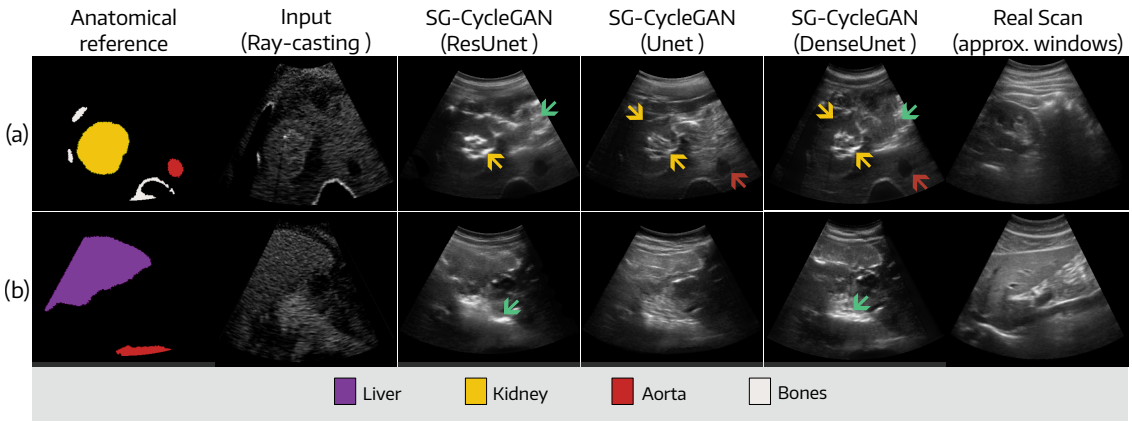


Figure 8: Comparison of simulation results obtained using an SG-CycleGAN with Unet, ResUnet, or DenseUnet based generator.

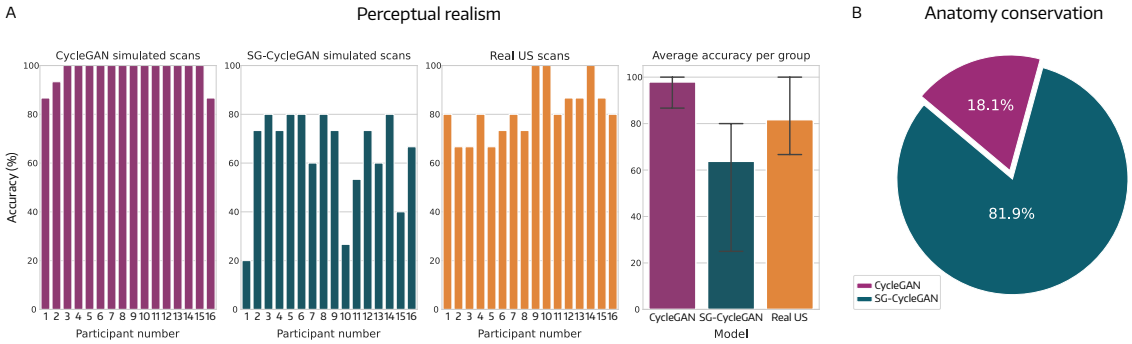


Figure 9: User study results. A) Classification accuracy for each simulation model and real scans as a bar per participant. Additionally, a bar plot with average accuracy per method. B) Pie chart comparing responses about which generative model performs better in terms of anatomy conservation.

VI. Discussion

VI.A. Effect of our segmentation-guided loss

Simulating abdominal US images is challenging. While physics-based approaches generate anatomically plausible images, their echogenicities remain unrealistic. In contrast, CycleGANs enhance visual quality but introduce hallucinated artifacts that distort the underlying anatomy[1]. These inconsistencies appear as non-uniform echogenicity patterns within organs (yellow dotted lines in Figure 3), a common issue in unpaired models relying on distribution-matching losses[21].

To alleviate this issue, we proposed a segmentation-guided loss, penalizing segmentation mismatches before and after completing the cycle. This term prevents the generator $\mathcal{G}_{\mathcal{A}\to\mathcal{R}}$ to introduce artifacts that cannot be removed through the reversed cycle $\mathcal{G}_{\mathcal{R}\to\mathcal{A}}$, without any extra

annotation. The anatomical labels from ray-casting simulations suffice for training. As seen in Figure 3 (green lines), our approach produces well-defined organ interfaces and homogeneous speckle noise patterns. Compared to existing methods (Figure 4), our loss function preserves anatomical structures while preventing hallucinated patterns within them.

Quantitatively, our model significantly reduces FID and KID scores by 66% and 89%, respectively ($p \ll 0.01$), as shown in Table 2. Our model not only presents the lowest FID and KID values, but when comparing with the others, we obtain high Cohen's d values (> 9.20), which imply a very large effect size between the simulations of our model and the others. Lower FID scores suggest improved statistical similarity to real images, resulting from the reduction in hallucinations and unusual artifacts in the constrained areas. This ensures that simulated images closely resemble real ones, making them more valuable for medical training. Furthermore, our segmentation-guided loss enhances anatomical accuracy, improving mIoU by up to 15.3% over standard CycleGAN. This advantage is reinforced by our user study, where SG-CycleGAN was rated as more anatomically consistent in 81.9% of cases compared to the standard CycleGAN.

VI.B. Impact of training in polar coordinates

Another key contribution of our work is migrating CycleGAN training from Euclidean to polar coordinates. As illustrated in Figure 3 and highlighted by the yellow arrows in the intermediate column of Figure 7, CycleGANs trained in Euclidean coordinates produce jagged edges, distort the FoV, or introduce warped regions. This occurs because the network lacks prior knowledge of the region of interest, making it difficult to distinguish between acoustic shadows and empty areas outside the FoV.

Training in polar coordinates addresses this issue by constraining the network's focus to the relevant area while excluding blank spaces. This prevents the model from having to learn the FoV shape itself, allowing for better utilization of its capacity. As a result, the model more accurately mimics speckle noise patterns (see zoomed patches in Figure 7) and better leverages the segmentation-guided loss, as evidenced by improvements in FID and KID values (Table 2). Additionally, since areas outside the FoV are absent in the input, the network naturally avoids generating artifacts in those regions. This is evident in Figure 7, where all images exhibit consistent FoVs without irregularities or hallucinations beyond the designated area.

VI.C. Influence of the generator architecture

Our approach proves effective across different generator architectures and network sizes, consistently improving FID and KID values when using the segmentation-guided loss (Figure 6). Among the tested architectures, the standard Unet outperformed ResUnet and DenseUnet, aligning with previous findings[1]. As illustrated in Figure 8, Unet generates anatomically more coherent outputs than ResUnet. This discrepancy is likely due to the absence of skip connections in ResUnet's bottleneck layers. Without these connections, the decoder must reconstruct anatomical structures using only low-level features from earlier layers, leading to information loss. The bottleneck acts as a lossy compression of the input, making it difficult for the decoder to reconstruct organs without introducing unrealistic artifacts.

VI.D. Advantages of SG-CycleGAN

Integrating all our proposed modifications into the standard CycleGAN framework resulted in a robust generative model that outperforms several state-of-the-art approaches in realism. We compared SG-CycleGAN against recent deep learning models, including Vision Transformers (UVCGANv2) and conditional diffusion models (UNSB). As shown in Table 1, these methods reduced FID and KID scores relative to the ray-casting model, with Vision Transformers achieving the largest improvement. However, SG-CycleGAN achieved the lowest FID and KID values ($p = 0.33 \times 10^{-3}$ and $p = 0.25 \times 10^{-3}$, respectively), with a very large effect size (Cohen's $d > 9.20$). Our model also closely matches real ultrasound (US) echogenicity distributions. As shown in Table 2, χ^2 tests indicate no statistically significant differences in liver and gallbladder echogenicities between SG-CycleGAN-generated images and real scans ($p > 0.008$). The effect size is minimal (Cohen's $d = 0.07$ for the gallbladder and $d = 0.13$ for the liver), suggesting that our model generates tissue echogenicities within the natural variability of real US images. While UNSB achieves a slightly better match for the liver ($d = 0.01$), our approach still performs competitively, as showed in Figure 5 (B). From a qualitative perspective, SG-CycleGAN produces more realistic scans. If the generated images were easily distinguishable from real ones, expert classification accuracy would approach 100%. While this was true for standard CycleGAN, experts misclassified 36% of SG-CycleGAN images as real (Figure 9). This suggests that our model generates anatomically consistent and realistic US scans, making it a promising tool for improving ultrasound training applications.

VI.E. Limitations

The primary limitation of this study is its focus on healthy subjects, as all experiments were conducted on individuals without pathologies or lesions. While we have demonstrated that our approach reduces hallucinations in simulated scans, we cannot guarantee the same for pathological cases or lesions. Future work should extend the evaluation to pathological cases to assess the method's robustness in simulating complex anatomical variations. Nevertheless, preventing hallucinations in healthy cases is already a promising step forwards, as it avoids introducing unrealistic artifacts that could be interpreted as pathologies.

It should be pointed out also that, despite the model exhibiting a substantial reduction in hallucinations compared to its original counterpart, we still observed unrealistic features occurring outside the segmented areas (e.g., around organ interfaces in Figure 3 (d)). In our current setup, we utilized masks for six different tissues available in our set of volumetric segmentations, so anatomical inconsistencies outside these regions are to be expected. In particular, we observed this phenomenon to occur in areas such as the stomach or the pancreas, which are not segmented in our training set. Clinically, these inaccuracies could affect the usefulness of the simulations in training scenarios where detailed anatomy of these regions is critical, such as as in surgical planning or procedural training, where a precise understanding of the anatomical structures is crucial.

Nevertheless, notice that the proposed approach is general enough to include any other organ without considerable modifications, should they are already available for the ray-casting based simulator (e.g. by segmenting the organs from the input CT scans). While these masks are essential for training the segmentation-guided CycleGAN, notice they do not increase the annotation costs beyond that already incurred in the first stage of the pipeline. Furthermore, these input segmentations are obtained from CT scans and not from US images, as it is needed for other US simulation approaches[17,18]. Therefore, accurate CT segmentation models such as TotalSegmentator[51] and Auto3DSeg[52] might be a promising alternative to automate this step and ease the incorporation of new simulation cases.

Notice that our image translation approach was trained and evaluated using images simulated with a single ray-casting approach with a fixed configuration, and with real scans obtained from a single US device. Consequently, it does not generalize to produce images from other probes or

devices. However, notice also that our proposed model is general enough to be retrained with images from other sources. Hence, by changing $\mathcal{A}$ and/or $\mathcal{R}$ with sets of artificial and/or real scans generated with a different simulator or US device, respectively, or under different imaging setups, the model would adapt to produce new artificial images for other practical applications.

As with all generative models, another limitation of this study is the lack of a trustworthy automated evaluation metric. The best approach for assessing the performance of US simulation algorithms is to run user tests with US experts, where individual images are analyzed and ranked based on their realism, without knowing their source. However, this becomes impractical for ablation studies, which require a substantial number of comparisons across multiple models and images. Furthermore, it is affected by subjective factors such as the level of experience of the human graders and their fatigue while performing the assessment. Although we conducted a user study with participants who are professionals specializing in abdominal ultrasound to add reliability to our findings, we acknowledge that a larger sample size could provide additional insights into the generalizability of the results. While the sample size is small, it enabled us to obtain meaningful insights that allowed to complement the validation of our approach. Furthermore, it is important to notice that most user studies in US simulation research use even smaller sample sizes (between 4 and 6[17,18,42,53]) than the one presented in this work (16). To the best of our knowledge, only one study used more experts for the validation than ours[34].

Measuring the quality of results obtained using unpaired generative models is inherently complex since it cannot be done using standardized metrics, such as SSIM and SNR, which require ground truth matching between real and artificial scans[16]. In an effort to provide a quantitative evaluation, we employed several metrics commonly used in the context of US simulation. These metrics enable the assessment of different aspects of the generated images from multiple complementary perspectives[16,17,34]. FID and KID allow to evaluate scans at a macro level, characterizing their texture patterns using filters from a pre-trained convolutional neural network. The χ^2 distance in particular is commonly employed for tissue characterization in paired image patches[16]. Alternatively, we used it to characterize intensities using segmentation masks to extract organ histograms (Section IV.E.). To complement this analysis, we also compared the cumulative distribution of echogenicities of each organ of interest (Figure 5.B). For homogeneous structures, such as the liver and the gallbladder, the histograms from SG-CycleGAN outputs were more alike to the ones computed from real scans. However, some notorious differences persisted in the kidney. The kidney has a complex internal anatomical structure (renal pelvis, renal cortex,

etc.) which might be the cause of these differences. Considering the presence or absence of these structures separately, might be a way to account for these differences.

The fact that US images obtained in DICOM format are, by default, JPEG compressed, is a drawback. JPEG is a lossy compression format that introduces artifacts in the images. As our models were trained to produce artificial scans that match the target distribution, it is expected for them to also feature these artifacts. This does not compromise our proposed model nor its evaluation, since they are compared to images presenting the same artifacts. In a more general context, image data used in the training of the proposed model should be consistent in the characteristics of the data where it will be applied. Failing to do so, might notoriously affect the results.

VII. Conclusions

In this paper we introduce a series of contributions to improve anatomical consistency and reduce artifacts in hybrid abdominal US simulators than combine ray-casting-based methods and CycleGANs. Our approach preserves anatomical structures and reduces hallucinations both inside organs and outside the FoV. We demonstrated that the weakly supervised segmentation-guided loss prevents significant alterations in anatomical areas, by penalizing differences in predicted masks obtained from a pre-trained Unet before and after the cycle consistency term. Additionally, training with images in polar coordinates constrains the FoV, enabling the model to focus on relevant content within non-blank areas. Our model demonstrated to be able to generate synthetic US images with fewer unrealistic artifacts, scattering patterns that are compatible with the acquisition probe's azimuthal angle, and a consistent FoV, closely resembling real scans. This approach enhances the realism of simulators, aiding in training and localization of abdominal organs. We believe future research can further improve these results by incorporating more organs and simulating abnormalities such as liver tumors or cysts, benefiting training for clinicians. Additionally, eliminating the ray-casting stage by training paired models directly from segmentation masks could lead to end-to-end trainable simulators. We encourage researchers to explore these promising directions to advance this field.

Acknowledgments

This work is funded by ANPCyT PICTs 2020-0045 and PIP GI 2021-2023-11220200102472CO. A Kaggle Open Data Research Grant also supported us with a financial grant to purchase the GPU used for this research. We thank all the expert radiologists who participated in the user study.

Data availability statement

The data that support this study was made publicly available by the authors as a Kaggle dataset [5]

Conflict of interest statement

The authors declare that they have no conflict of interest.

References

[1] S. Vitale, J. I. Orlando, E. Iarussi, and I. Larrabide, Improving realism in patient-specific abdominal ultrasound simulation using CycleGANs, International journal of computer assisted radiology and surgery **15**, 183–192 (2020).

[2] T. Kameda and N. Taniguchi, Overview of point-of-care abdominal ultrasound in emergency and critical care, Journal of Intensive Care **4**, 53 (2016).

[3] J. Urbina, S. M. Monks, and S. B. Crawford, Simulation in Ultrasound Training for Obstetrics and Gynecology: A Literature Review, Simulation **15** (2021).

[4] V. A. Dinh, J. Y. Fu, S. Lu, A. Chiem, J. C. Fox, and M. Blaivas, Integration of ultrasound in medical education at United States medical schools: a national survey of directors' experiences, Journal of ultrasound in medicine **35**, 413–419 (2016).

[5] M. Østergaard, C. Ewertsen, L. Konge, E. Albrecht-Beste, and M. B. Nielsen, Simulation-based abdominal ultrasound training–a systematic review, Ultraschall in der Medizin-European Journal of Ultrasound **37**, 253–261 (2016).

[5] https://www.kaggle.com/datasets/ignaciorlando/ussimandsegm

[6] D. J. Canty, J. A. Hayes, D. A. Story, and C. F. Royse, Ultrasound simulator-assisted teaching of cardiac anatomy to preclinical anatomy students: A pilot randomized trial of a three-hour learning exposure, Anatomical sciences education **8**, 21–30 (2015).

[7] B. P. Dromey, D. M. Peebles, and D. V. Stoyanov, A systematic review and meta-analysis of the use of high-fidelity simulation in obstetric ultrasound, Simulation in Healthcare **16**, 52–59 (2021).

[8] M. Donnez, F.-X. Carton, F. Le Lann, E. De Schlichting, and M. Chabanas, Realistic synthesis of brain tumor resection ultrasound images with a generative adversarial network, in *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 11598, pages 637–642, SPIE, 2021.

[9] L. Bargsten and A. Schlaefer, SpeckleGAN: a generative adversarial network with an adaptive speckle layer to augment limited training data for ultrasound image processing, International journal of computer assisted radiology and surgery **15**, 1427–1436 (2020).

[10] R. Shams, R. Hartley, and N. Navab, Real-time simulation of medical ultrasound from CT images, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 734–741, Springer, 2008.

[11] B. Burger, S. Bettinghausen, M. Radle, and J. Hesser, Real-time GPU-based ultrasound simulation using deformable mesh models, IEEE transactions on medical imaging **32**, 609–618 (2012).

[12] O. Mattausch and O. Goksel, Monte-carlo ray-tracing for realistic interactive ultrasound simulation, in *Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine*, pages 173–181, 2016.

[13] D. Tomar, L. Zhang, T. Portenier, and O. Goksel, Content-preserving unpaired translation from simulated to realistic ultrasound images, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 659–669, Springer, 2021.

[14] L. Ruthotto and E. Haber, An introduction to deep generative modeling, GAMM-Mitteilungen **44**, e202100008 (2021).

[15] F. Tom and D. Sheet, Simulating patho-realistic ultrasound images using deep generative networks with adversarial learning, in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 1174–1177, IEEE, 2018.

[16] L. Zhang, T. Portenier, and O. Goksel, Learning ultrasound rendering from cross-sectional model slices for simulated training, International Journal of Computer Assisted Radiology and Surgery **16**, 721–730 (2021).

[17] J. Liang, X. Yang, Y. Huang, H. Li, S. He, X. Hu, Z. Chen, W. Xue, J. Cheng, and D. Ni, Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis, Medical Image Analysis **79**, 102461 (2022).

[18] G. Pigeau, L. Elbatarny, V. Wu, A. Schonewille, G. Fichtinger, and T. Ungi, Ultrasound image simulation with generative adversarial network, in *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 11315, pages 54–60, SPIE, 2020.

[19] N. J. Cronin, T. Finni, and O. Seynnes, Using deep learning to generate synthetic B-mode musculoskeletal ultrasound images, Computer methods and programs in biomedicine **196**, 105583 (2020).

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

[21] J. P. Cohen, M. Luck, and S. Honari, Distribution matching losses can hallucinate features in medical image translation, in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 529–536, Springer, 2018.

[22] P. Rubi, E. F. Vera, J. Larrabide, M. Calvo, J. D'Amato, and I. Larrabide, Comparison of real-time ultrasound simulation models using abdominal CT images, in *12th international symposium on medical information processing and analysis*, volume 10160, pages 55–63, SPIE, 2017.

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

[24] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, Least squares generative adversarial networks, in *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[25] O. Jimenez-del Toro et al., Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks, IEEE transactions on medical imaging **35**, 2459–2475 (2016).

[26] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[27] S. Dangi and C. Linte, DenseUNet-K: A simplified Densely Connected Fully Convolutional Network for Image-to-Image Translation, (2019).

[28] X. Sun, H. Li, and W.-N. Lee, Constrained CycleGAN for effective generation of ultrasound sector images of improved spatial resolution, Physics in Medicine and Biology (2023).

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

[30] D. Ulyanov, A. Vedaldi, and V. Lempitsky, Instance normalization: The missing ingredient for fast stylization, arXiv preprint arXiv:1607.08022 (2016).

[31] D. Ulyanov, A. Vedaldi, and V. Lempitsky, Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6924–6932, 2017.

[32] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[33] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, Contrastive learning for unpaired image-to-image translation, in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345, Springer, 2020.

[34] D. Tomar, L. Zhang, T. Portenier, and O. Goksel, Content-preserving unpaired translation from simulated to realistic ultrasound images, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 659–669, Springer, 2021.

[35] D. Torbunov, Y. Huang, H.-H. Tseng, H. Yu, J. Huang, S. Yoo, M. Lin, B. Viren, and Y. Ren, Rethinking CycleGAN: Improving Quality of GANs for Unpaired Image-to-Image Translation, arXiv preprint arXiv:2303.16280 (2023).

[36] B. Kim, G. Kwon, K. Kim, and J. C. Ye, Unpaired Image-to-Image Translation via Neural Schr\" odinger Bridge, arXiv preprint arXiv:2305.15086 (2023).

[37] X. Ma, N. Anantrasirichai, S. Bolomytis, and A. Achim, PMT: Partial-Modality Translation Based on Diffusion Models for Prostate Magnetic Resonance and Ultrasound Image Registration, in *Annual Conference on Medical Image Understanding and Analysis*, pages 285–297, Springer, 2024.

[38] H. Alqahtani, M. Kavakli-Thorne, G. Kumar, and F. SBSSTC, An analysis of evaluation metrics of gans, in *International Conference on Information Technology and Applications (ICITA)*, volume 7, 2019.

[39] A. Borji, Pros and cons of GAN evaluation measures: New developments, Computer Vision and Image Understanding **215**, 103329 (2022).

[40] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, in *Advances in Neural Information Processing Systems*, volume 30, Curran Associates, Inc., 2017.

[41] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, Demystifying mmd gans, arXiv preprint arXiv:1801.01401 (2018).

[42] J. Liang et al., Weakly-supervised high-fidelity ultrasound video synthesis with feature decoupling, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 310–319, Springer, 2022.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the inception architecture for computer vision, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[44] A. Obukhov, M. Seitzer, P. Wu, S. Zhydenko, J. Kyl, and E. Lin, High-fidelity performance metrics for generative models in PyTorch, 2020.

[45] C. Bonferroni, Statistic theory of classes and calculation of probabilities, Volume in Honor of Riccardo della Volta. Florence: University of Florence , 1–62 (1937).

[46] J. Cohen, *Statistical power analysis for the behavioral sciences*, routledge, 2013.

[47] G. E. Mailloux, M. Bertrand, R. Stampfler, and S. Ethier, Local histogram information content of ultrasound B-mode echographic texture, Ultrasound in medicine & biology **11**, 743–750 (1985).

[48] D. China, F. Tom, S. Nandamuri, A. Kar, M. Srinivasan, P. Mitra, and D. Sheet, Ultra-compression: framework for high density compression of ultrasound volumes using physics modeling deep neural networks, in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 798–801, IEEE, 2019.

[49] A. K. Tripathi, S. Mukhopadhyay, and A. K. Dhara, Performance metrics for image contrast, in *2011 International Conference on Image Information Processing*, pages 1–4, IEEE, 2011.

[50] J. R. De Leeuw, jsPsych: A JavaScript library for creating behavioral experiments in a Web browser, Behavior research methods **47**, 1–12 (2015).

[51] J. Wasserthal et al., TotalSegmentator: robust segmentation of 104 anatomic structures in CT images, Radiology: Artificial Intelligence **5** (2023).

[52] A. Myronenko, D. Yang, Y. He, and D. Xu, Automated 3D Segmentation of Kidneys and Tumors in MICCAI KiTS 2023 Challenge, in *International Challenge on Kidney and Kidney Tumor Segmentation*, pages 1–7, Springer, 2023.

[53] L. Chen, H. Liao, W. Kong, D. Zhang, and F. Chen, Anatomy preserving GAN for realistic simulation of intraoperative liver ultrasound images, Computer Methods and Programs in Biomedicine **240**, 107642 (2023).