ORIGINAL ARTICLE

Learning ultrasound rendering from cross-sectional model slices for simulated training

Lin Zhang[1] · **Tiziano Portenier**[1] · **Orcun Goksel**[1,2]

Abstract

Purpose Given the high level of expertise required for navigation and interpretation of ultrasound images, computational simulations can facilitate the training of such skills in virtual reality. With ray-tracing based simulations, realistic ultrasound images can be generated. However, due to computational constraints for interactivity, image quality typically needs to be compromised.

Methods We propose herein to bypass any rendering and simulation process at interactive time, by conducting such simulations during a non-time-critical offline stage and then learning image translation from cross-sectional model slices to such simulated frames. We use a generative adversarial framework with a dedicated generator architecture and input feeding scheme, which both substantially improve image quality without increase in network parameters. Integral attenuation maps derived from cross-sectional model slices, texture-friendly strided convolutions, providing stochastic noise and input maps to intermediate layers in order to preserve locality are all shown herein to greatly facilitate such translation task.

Results Given several quality metrics, the proposed method with only tissue maps as input is shown to provide comparable or superior results to a state-of-the-art that uses additional images of low-quality ultrasound renderings. An extensive ablation study shows the need and benefits from the individual contributions utilized in this work, based on qualitative examples and quantitative ultrasound similarity metrics. To that end, a local histogram statistics based error metric is proposed and demonstrated for visualization of local dissimilarities between ultrasound images.

Conclusion A deep-learning based direct transformation from interactive tissue slices to likeness of high quality renderings allow to obviate any complex rendering process in real-time, which could enable extremely realistic ultrasound simulations on consumer-hardware by moving the time-intensive processes to a one-time, offline, preprocessing data preparation stage that can be performed on dedicated high-end hardware.

Keywords Ultrasound simulation · Generative models · Attenuation

Introduction

Ultrasound (US) imaging is a real-time, non-invasive and radiation-free imaging modality, making it ideal for computer-assisted interventions. Nevertheless, the challenges in ultrasound probe manipulation and image interpretation necessitates extensive operator training. Computer-assisted ultrasound simulation can facilitate such training in a virtual-reality environment [2,17,19,25]. This does not require volunteer patients or on-site supervisors, while enabling students to learn and practice skills in a flexible, stress-free, and self-supervised manner. Furthermore, rare-to-encounter diseases and conditions can be presented and practiced in a realistic, interactive mode.

To provide training for US probe manipulation and image interpretation, computational simulators need to correctly represent ultrasonic physical effects, e.g. view-dependent artifacts and realistic tissue texture, and operate dynamically and interactively. In the literature US simulation methods can be categorized into interpolative and generative approaches. The former [4] generates 2D US images by interpolating pre-recorded 3D volumes. However, to generate a rich variety of images, a large amount of 3D volumes needs to be

✉ Lin Zhang
  lin.zhang@vision.ee.ethz.ch

1 Computer-assisted Applications in Medicine, ETH Zurich, Zürich, Switzerland

2 Department of Information Technology, Uppsala University, Uppsala, Sweden

acquired and stored. Furthermore, it is difficult to generate novel views and contents, since the physical model is in general not considered. In contrast to interpolative US simulation approaches, advanced generative methods [3,15,20,24] allow to generate variety of images with plausible view-dependent artifacts, e.g. for rare pathological cases. These techniques model ultrasonic wave propagation using ray-tracing techniques on anatomical models. New scenes of any given anatomical model can be simulated with different imaging parameters and conditions. There is a line of works on developing sophisticated ray tracing techniques to simulate realistic wave interactions. A deterministic surface-based ray tracing method is introduced in [3] to simulate ultrasonic directional wave interactions. The method is further extended in [20] to utilize patient-specific volumetric MRI or CT data. A more sophisticated stochastic wave-interaction and surface modeling is proposed in [15] to overcome the simplified assumption of the deterministic ray model. Given a 3D anatomical model, ray-based techniques using the state-of-the-art Monte-carlo ray-tracing framework manage to simulate US images with surprisingly high realism at interactive frame rates, as shown in [15,24] for fetal ultrasound imaging. Since ultrasound is a real-time, interactive imaging modality, the fundamental challenge for any training ultrasound simulation is to achieve real-time frame rates. Emulating ultrasound physics and appearance at an interactive rate has been a challenge tackled by most above-listed literature [3,4,15,21,23,24] and a long standing research question. In this paper, we propose to mimic the simulation with a deep learned model, such that the interactive image simulation only requires a quick inference of such model.

Recent advances in deep learning have enabled various learning based approaches for ultrasound simulation. Generative adversarial networks [5] (GANs) are the most promising models in this regard due to their outstanding performance in generating photo-realistic images. Conditional GANs are widely adopted for generating ultrasound images conditioned on physical input, such as calibrated physical coordinate [8] and echogenicity map [26]. A recent publication [12] has employed a generative autoencoder model learned with a large amount of tracked ultrasound data, to perform patient-specific image generation from transducer position and orientation. To simulate realistic ultrasound speckles, the authors in [1] have introduced a speckle layer to incorporate the physical model of speckle generation into a GAN-based data augmentation network. In [27], a cycle-GAN [29] model is employed for improving the realism of simulated ultrasound images. A constrained cycleGAN is proposed in [9] for unpaired translation from echocardiographic images acquired with point-of-care ultrasound to high-end devices. Recently, a GAN-based image translation framework has been proposed in [28] for recovering high quality US images equivalent to computationally expensive

ray-based simulations using low-quality and thus faster simulations, by also leveraging information from corresponding segmentation and integral attenuation maps. However, rendering such low-quality ultrasound images still requires additional computation time and sophisticated hardware resources.

In this work, we propose to learn the rendering of ultrasound images given only cross-sectional model slice / segmentation and integral attenuation maps, the latter of which can be derived from former on-the-fly and helps distill global acoustic energy information locally. An overview of per-frame segmentation and integral attenuation map generation can be seen in Fig. 1a. Herein we adopt a Monte-carlo simulation presented in [15] for simulating the same abdominal scene as in previous work [28]. Since low-quality images as used in [28] provide information about wave interactions and speckle textures, omitting this modality in the network input renders the translation task much more challenging. We carefully inspect the generator architecture and propose several crucial architectural improvements to enable realistic quality images from merely the segmentation maps as input. Herein we present a local feature and texture preserving generator containing

- connections of network input to each intermediate layer for improved information flow;
- smaller receptive field to preserve spatial information;
- texture-friendly strided convolutions;
- dedicated noise images fed at each resolution.

With the above mentioned architectural changes in contrast to [28], our proposed network is able to simulate US images with realistic structural and textural content. To the best of our knowledge, this is the first work investigating deep neural networks to generate simulated ultrasound images using only structural tissue cross-sections.

Methods

Ray-based ultrasound simulation

In this work we adopt a Monte-Carlo ray tracing framework presented in [15] to simulate ultrasound B-mode images for an abdominal scene. Directional ultrasonic wave interactions such as reflection and refraction are simulated using a stochastic surface ray tracing model, which is able to create realistic looking soft shadows and tissue reflection boundaries. Ultrasound speckle formation is modeled by convolving 3D tissue scattering representation, parametrized by a Gaussian distribution [14], with spatially-variant point-spread-function. Each tissue type is assigned a set of pre-determined acoustic properties such as acous-
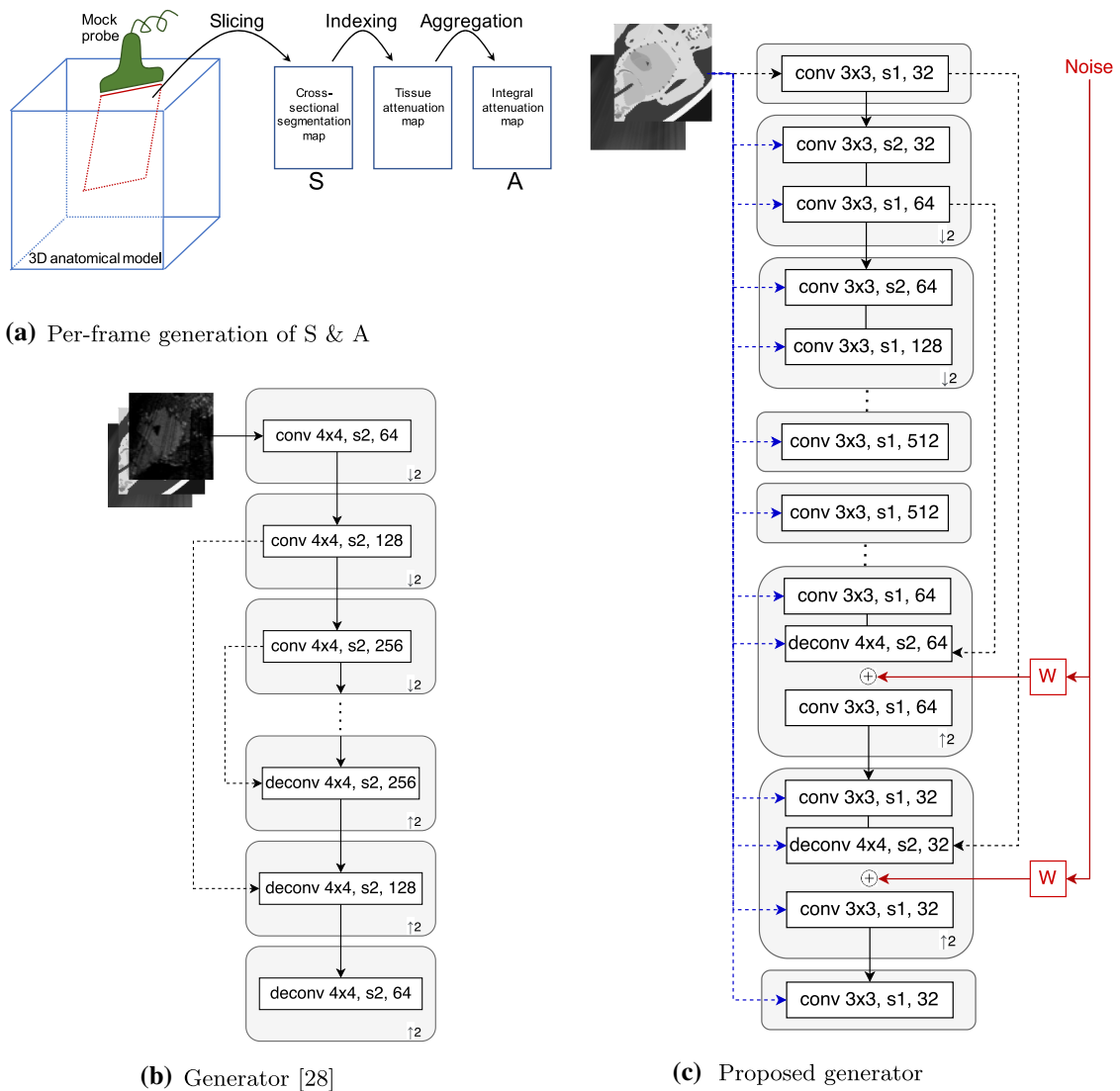
(a) Per-frame generation of S & A

(b) Generator [28]

(c) Proposed generator

Fig. 1 **a** For each frame of ultrasound simulation, a cross-sectional tissue slice **S** is extracted given the mock probe position and a set of 3D triangulated anatomical surfaces. Referencing acoustic attenuation of each anatomical structure from a predefined indexed list, a spatial attenuation map is generated. Aggregating such maps along ultrasound propagation then yields the so-called integral attenuation maps **A** are derived for each frame in real-time. **b** Network architecture of the generator used in [28]. **c** Network architecture proposed herein to address the challenges of image translation from segmentation maps to complex B-mode image content. Here the white blocks (de)conv $k \times k$, ss, ch stands for a (de)convolutional layer with a filter size of $k \times k$, stride s, and ch filters. Gray block with ↓2 or ↑2 denotes the entire up- or downsampling block with a factor of 2. Dotted black lines stand for skip connections and blue lines for concatenation of the network input. Solid red line and block represents stochastic components of the network. The term W denotes learned per-channel weighting

tic impedance, attenuation, and speed-of-sound, which are modeled by hand according to literature and quality assessment by sonographers. To obtain gray-scale B-mode images, conventional post-processing steps are applied on simulated RF data, including envelope detection, time-gain compensation, log compression, and scan conversion. Herein we simulate the second trimester transabdominal scene of the entire abdomen. The US probe is placed at positions on a regular grid on the abdominal surface and different orientations at each position are collected to uniformly sample the fetus field-of-view. Detailed scene information and simulation parameters are provided in Table 1.

Model slice is rendered as the cross-section of the triangulated anatomical surfaces. It provides details about the anatomical layout, thus is also referred to segmentation map herein. The segmentation map is stored as an indexed image. Each tissue type has its unique intensity value, which is matched to a corresponding attenuation coefficient using a look-up table prior to axial aggregation for attenuation map generation.

Table 1 Simulation parameters

Parameter	Value	Parameter	Value
Triangles fetus	400k	Transducer frequency	8 MHz
Triangles mother	275k	Transducer field-of-view	70°
Image depth	15.0 cm	Axial samples	3072
Elevational layer	3	Ray per scanline	32

One of the most characteristic features of US images is acoustic shadow, which is important for image interpretation, since abnormal shadows may indicate the existence of tissue modification. This directional artifact is mainly caused by tissue attenuation, which leads to a decrease in sound wave intensity along the wave propagation path due to local tissue absorption and scattering. Therefore, acoustic shadow is a cumulative effect which requires a large network receptive field and significant network modeling ability. The authors in [28] propose to provide this cumulative information in a form of an integral attenuation map and they demonstrates that providing this additional information to the network greatly helps to reproduce acoustic shadows. Integral attenuation maps are computed as $a = e^{-\sum_{i=0}^{z} \mu[i]}$ with the image depth z and frequency-dependent and tissue specific attenuation coefficient μ. The acoustic shadow formation is thus approximated by accumulating the attenuation strength, quantified by μ, along the wave propagation path. For each image point, the acoustic attenuation is inferred from the segmentation value using a predefined look-up table. The integral attenuation map is generated separately for each pre-scan-converted image column, which corresponds to radial US propagation direction for the given convex US probe. The integral maps are then normalized by 98 %ile of image intensities, and scan-converted into Cartesian coordinates.

Ultrasound simulation using a generative adversarial network

In this work, we propose a generative adversarial network (GAN) for generating ultrasound images from segmentation and integral attenuation maps. Given segmentation maps and acoustic attenuation properties of each tissue, integral maps can be generated as in [28] in order to mimic wavefront traversal in tissue. This is known from [28] to improve rendered shadows and acoustic amplifications by bringing global echo amplitude modifications to a local context. In our ablation experiments we study the effect of attenuation maps, referred in the results with the acronym *att*.

In contrast to [28], herein we propose the following four major architectural improvements, as also illustrated in Fig. 1, which are later demonstrated to provide promising generation results in our experiments.

Input to all channels: To enable an effective information flow from the input images to each spatial resolution in the network, we provide the information at each intermediate layer by concatenating the maps along the channel dimension to the network activations, indicated by dotted blue connections in Fig. 1c. Although this was shown in [18] to be inferior to spatially-adaptive normalization (SPADE) layers in the context of natural images, we found this approach to outperform SPADE in our application setting, since it enables the network to generate location-specific features conditioned on both segmentation and integral attenuation map. In our ablation experiments, we refer to this concatenation approach with the acronym *concat*.

Preserving spatial information: To further preserve the local information from input maps to output images, we use a comparatively small receptive field in our network architecture. Since the integral attenuation map helps to transform global effects of acoustic shadows into local features, understanding and encoding the entire scene in a compact form is not required and would waste network capacity. We therefore use 4 downsampling blocks.

Texture-friendly decoder: Transposed convolutions are known to be prone to characteristic checkerboard artifacts due to uneven overlap when using odd kernel sizes [16]. Therefore they are not an ideal choice in texture generation tasks. To improve texture generation performance, we therefore enhance the decoding blocks by introducing additional stride-1 convolution layers, which helps to circumvent both low and high frequency artifacts. These changes in the convolutional architecture are referred collectively using the acronym *conv* in our ablation study.

Stochastic texture generation: To generate random speckle textures, the network needs access to a stochastic source, especially when omitting low-quality B-mode image input. A straightforward approach to ensure a stochastic process is to feed an explicit noise image as an additional input channel. Recently, Karras et al. [10] proposed an alternative method, by perturbing feature channels using additive Gaussian noise with learned per-channel weighting. This warrants the network to disentangle global and local stochastic variations. Motivated by its astonishing performance in generating fine stochastic details such as hair, we study this approach for the generation of ultrasound textures. Accordingly, Gaussian

noise images with different sizes are generated and weighted by a learned per-channel scaling factor. The weighted noise channels are then added to the decoding feature layers after skip connections, shown as solid red lines in Fig. 1c. In the ablation study we referred to this technique with the acronym *noise*. The more conventional way of adding noise as an input layer is equivalent to replacing the low-quality image in [28] with noise, and therefore is referred as NSA2H in our comparative study.

We adopt the patchGAN discriminator and the training objective from [28]. The loss function consists of a GAN training objective L_{GAN} and an L_1-based data fidelity term L_F as follows:

$$L = L_{GAN}(G, D) + \lambda L_F(G),$$

$$L_{GAN} = \mathbf{E}_{s,a,y}[\log D(y|s, a)] \tag{1}$$
$$+ \mathbf{E}_{s,a}[\log(1 - D(m \circ G(s, a)|s, a)], \tag{2}$$

$$L_F = \mathbf{E}_{s,a,y}[||y - m \circ G(s, a)||_1], \tag{3}$$

where $s \in S$ is segmentation map, $a \in A$ is integral attenuation map, and m is the binary mask indicating the convex imaging region. $\mathbf{E}_{s,a,y}$ and $\mathbf{E}_{s,a}$ are the expected values, respectively, over all samples of $\{s, a, y\}$ triples and $\{s, a\}$ pairs. The generator G maps the source segmentation/attenuation map to the target US image, whereas the discriminator D discriminates real and generated images conditioned on the generator input. Before computing the loss, the generator output is pixel-wise multiplied with the binary mask m, which is denoted by the operator $\circ$.

Results and discussion

Implementation and network details. All our models are trained using the Adam optimizer [11] with a learning rate of 0.0002 and exponential decay rates $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The batch size is set to 4 and the loss weighting parameter λ set to 100. The leaky rectified linear unit is used in the encoder and the rectified linear unit throughout the decoder, except its output layer, which is activated using tanh. Nonlinear activations are followed by instance normalization. We use the same dataset as in [28] consisting of 6669 3-tuples (s, a, y) with an image size of 1000×1386. The network is trained using on-the-fly cropped image patches with a size of 512×512 to make the training more efficient. Randomly selected 6000 images were used for training and the rest for evaluation, following the same dataset split as in [28].

Compared methods. We refer our proposed network architecture as SA2H, which translates from segmentation map S and integral attenuation map A to a high-quality image H. To evaluate individual architectural proposals of SA2H, we ablate each component separately and refer to as "SA2H-component" in the ablation studies below. We accordingly compare SA2H against the following alternatives:

- **SA2H-att** omitting integral attenuation maps as input to the network;
- **SA2H-concat** providing segmentation and attenuation maps at the input layer only without concatenating in the hidden layers (removing the blue connections in Fig. 1 (c));
- **SA2H-conv** omitting texture-friendly convolutional components in the decoder and falling back to even sized transposed convolution kernels;
- **SA2H-noise** omitting Gaussian noise images as stochastic input and provide an input noise layer instead (removing the red connections in Fig. 1 (c));
- **LSA2H** the recent low-to-high-quality image translation network presented in [28], which has a low-quality rendered ultrasound image L as an additional network input. Given the additional ultrasound physics and texture information encoded in the low-quality image input, LSA2H represents an upper-bound (or, considering our architectural improvements, rather a *silver-standard*).
- **NSA2H** a lower-bound baseline by replacing the low-quality image L in LSA2H with an uncorrelated Gaussian noise image N.

Qualitative evaluation. Fig. 2 depicts the qualitative results for all the models mentioned above, with arrows pointing at structures relevant to discussion points below. The visual results of the ablated variants of SA2H show substantial quality degradation compared to the full SA2H model, demonstrating the importance of each proposed architectural contribution. Given only segmentation map in the network input, SA2H-att fails to generate acoustic shadows, e.g. those cast by the ribs. Detailed structures such as the cervical vertebrae are blurred out in the SA2H-concat results, which also contain hallucinated structures mainly due to insufficient preservation of input information along the encoding-decoding path. With SA2H-conv, checkerboard artefacts are observed due to the lack of proposed additional stride-1 convolutional layers. SA2H-noise without any explicit noise input is seen to be sub-optimal at generating textural details. The baseline method NSA2H fails to preserve anatomical structures and acoustic shadows in all cases, while the simulated textures also show significant artefacts such as checkerboard patterns. Realism of different simulation aspects may become relevant given different clinical applications and scenarios. For instance, improved structural preservation, e.g. with the hyperechoic bony structures such as the skull and the ribs, of the final model over its ablated variants and NSA2H may prove relevant in fetal head measurements, while the textural improvements facilitating screening fetal organ maturity, e.g. lungs. Compared

Fig. 2 Inference on full field-of-view (FoV) images. Segmentation and integral attenuation maps are shown, respectively, at the top left and bottom right corners of the simulated B-mode images
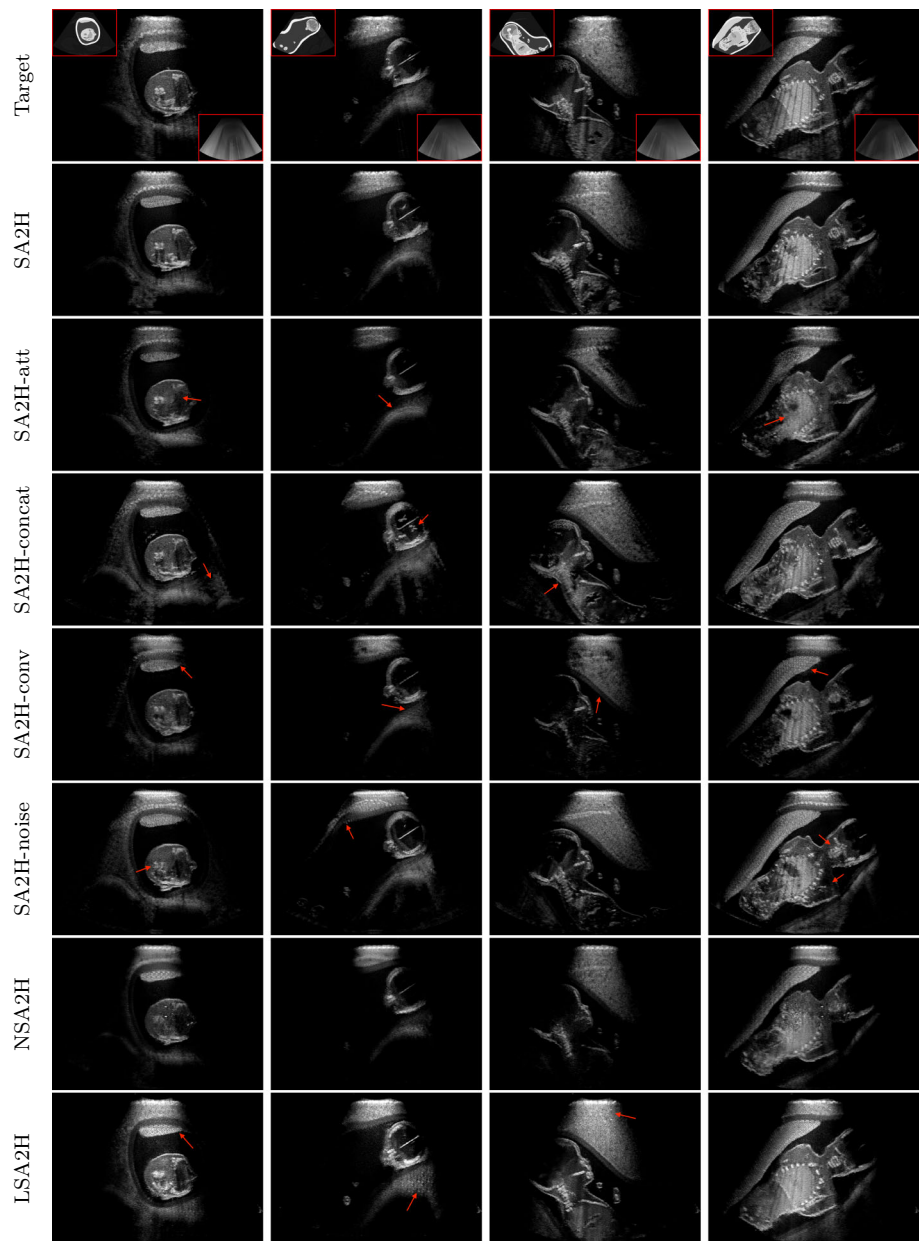
to the silver-standard model LSA2H with a low-quality rendered image as additional input, SA2H is seen to be on par in structural preservation. Note that shadowing on homogenous regions (e.g. the rib shadowing on the homogenous lung region on the 4th column of Fig. 2) with our proposed method SA2H is represented more faithfully compared to LSA2H, whereas shadows on structurally complex regions (e.g. the skull shadowing around the heart and surrounding tissues on the 3rd column of Fig. 2) are suboptimal with our SA2H. Therefore, one may have to evaluate our method given particular simulation tasks, e.g. its clinical validity for fetal heart exams. However, even with low quality rendered images, LSA2H leads to artificial enhancements of intensities, lack of acoustic shadows, and low-quality textures especially near

the probe, for which SA2H yields satisfactory results as illustrated in Fig. 2.

Quantitative evaluation. Metrics for assessment of US image quality and realism is an open research topic. Pixel-wise difference and SNR metrics assuming paired ground-truth images are often suboptimal and potentially inconclusive in judging US image realism, given the noisy speckle appearance and the inherent features and artifacts of B-mode images. To assess such local image matching, we herein also utilize a sliding path-based histogram comparison metric. Additionally, we utilize a typical metric for generative models that quantify distributions given features of a pretrained

Fig. 3 Spatial pχ^2 error map for the examples shown in Fig. 2. Error map is displayed within the range [0,1]. The corresponding ultrasound image is shown below the map. Target image is shown in the right most column
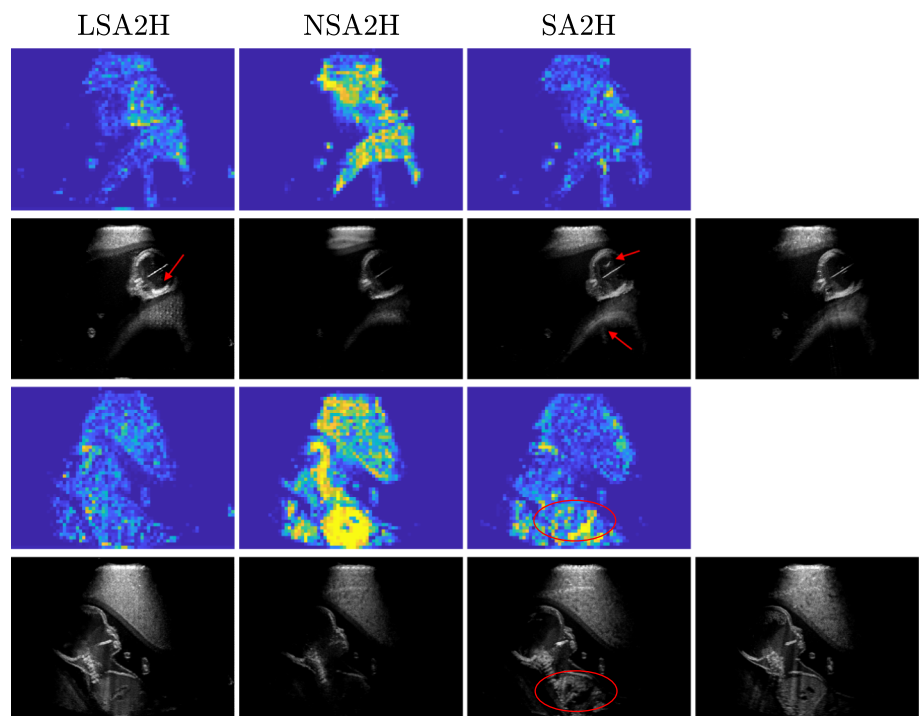

Table 2 Quantitative results. Bold number indicates the best performance. Mean and standard deviation are reported for PSNR, MAE and pχ^2. Network capacity is given as number of trainable parameters

	PSNR mean	std	MAE mean	std	pχ^2 ($\times 10^{-2}$) mean	std	FID	#params
LSA2H	27.38	0.52	**6.21**	1.56	**13.61**	1.39	64.42	57.2M
NSA2H	25.59	1.81	8.22	2.77	31.39	5.51	67.28	57.2M
SA2H-att	25.04	1.61	8.75	2.75	29.40	5.07	92.49	14.4M
SA2H-concat	26.23	1.22	9.08	2.93	29.36	5.87	76.40	14.4M
SA2H-conv	24.85	1.31	8.87	2.55	32.70	4.33	93.18	16.2M
SA2H-noise	26.45	1.34	8.14	2.66	25.95	4.64	97.85	14.5M
SA2H	**28.19**	1.23	6.37	2.08	18.60	5.23	**32.34**	14.5M

visual model. Accorindingly, we utilize the following complementary quantiative metrics:

- *Peak signal-to-noise ratio (PSNR)* is computed as PSNR $= 10 \log_{10}(\frac{255}{\text{MSE}})$ with mean squared error MSE between two images.
- *Mean absolute error (MAE)* measures pixel-wise difference between two images. High MAE value may indicate large intensity shift or structural mismatch. However, it cannot provide information about texture difference.
- pχ^2: We use χ^2 distance to measure the difference in image histogram statistics, which are commonly used for tissue characterization [13,22]. This metric indicates potential mismatch in tissue speckle patterns, which affects image histogram statistics. χ^2 distance computes the difference between histograms h_A and h_B as: $\chi^2(h_A||h_B) = \frac{1}{2} \sum_{l=1..d} \frac{(h_A[l]-h_B[l])^2}{h_A[l]+h_B[l]}$. The number of histogram bins d is set to 50. We compute this met-

ric locally within non-overlapping sliding patches with a size of 20×20 to capture local textural information of US speckle patterns. Herein we compute the root mean squared error, referred to *patch χ^2 (pχ^2)*.

- *Fréchet Inception Distance (FID)* [7] quantifies the similarity of generated samples to real samples by computing the distance between the feature vectors of the classification network Inception v3. The feature vectors are fitted to multivariate Gaussian and the difference is computed using Fréchet distance. FID score is a widely used metric for assessing GAN performance. Lower FID indicates better image quality. We compute FID on 512×512 sized center crops of generated full field-of-view images, which are further divided into four sub-crops of 299×299, to increase the number of samples and match Inception v3 input size.
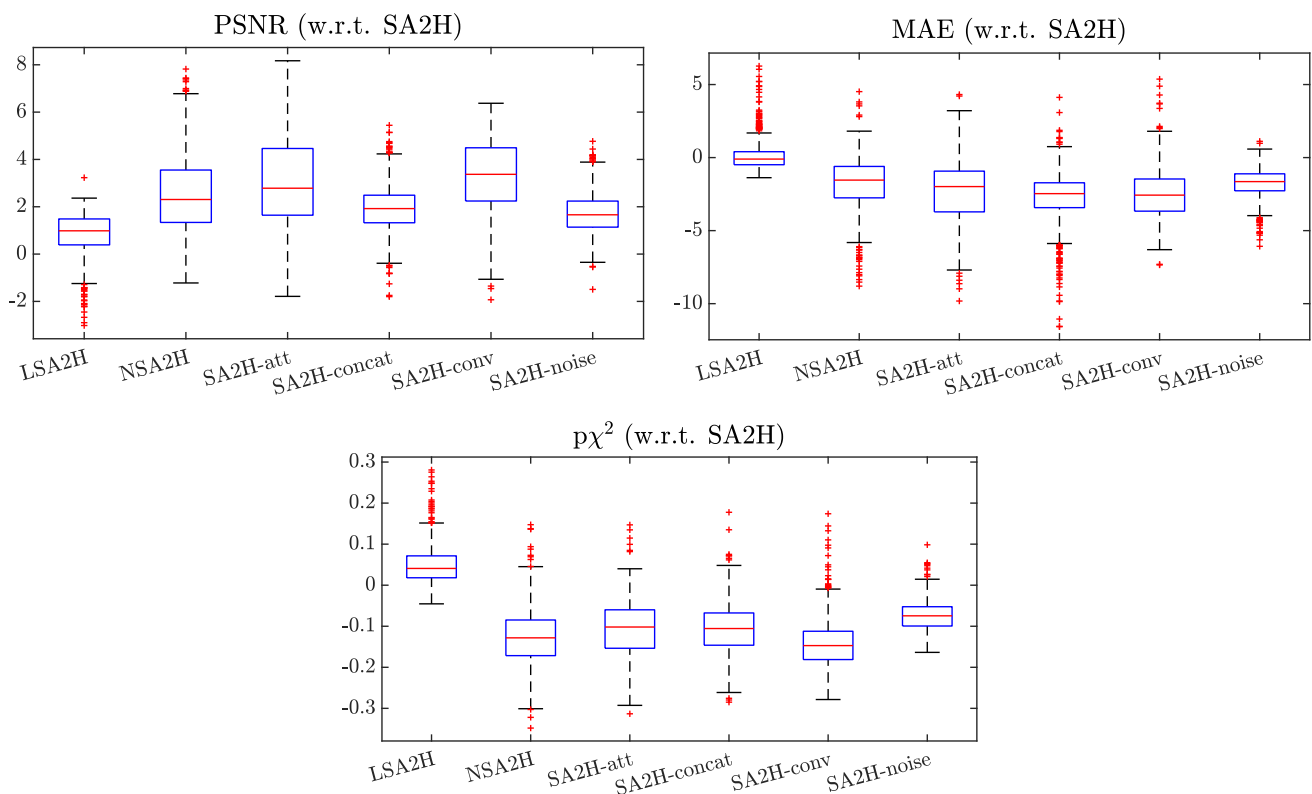
Fig. 4 Box plot of paired difference between SA2H and its ablated variants, LSA2H and NSA2H. Note that higher PSNR, lower MAE and $p\chi^2$ indicates better performance. For FID, no image-wise metric can be computed

For the interpretation of the local errors, sample spatial $p\chi^2$ error maps are depicted in Fig. 3 for LSA2H, NSA2H, and SA2H for the middle two examples shown in Fig. 2. Both images generated by NSA2H have a lot of missing structures and accordingly have high error almost all over the map. Artificial skull enhancement with LSA2H is seem to evoke large $p\chi^2$ error, as shown in the corresponding error map, whereas the bright spots in the error map of SA2H reflect some hallucinated shadows and structure in the brain. All of above mentioned regions of interest are marked by red arrows. In the bottom example, SA2H fails to generate faithful content at the bottom region marked by red circles, which is well indicated by the error map as well.

The quantitative results are summarized in Table 2. The effectiveness of all the proposed architectural improvements is well demonstrated by the significant performance drop of the ablated variants in all the metrics, which also corroborate with our qualitative results shown above. The proposed model SA2H has achieves an improvement of over 50% in FID and 40% in $p\chi^2$ over the baseline NSA2H, indicating a significantly higher fidelity in generated images using the proposed method. Surprisingly, for PSNR and FID metrics, SA2H outperforms the LSA2H model, which has extra ultrasound information provided as a low-quality image input.

In Fig. 4 we show the distribution of paired differences in PSNR, MAE, and $p\chi^2$ between each tested model and the proposed SA2H one. As seen, for all ablated variants and the baseline NSA2H, our proposed method provides a significant improvement for all metrics.

The presented method is aimed to replace the computationally expensive online rendering process, such that image quality is not restricted by the interactivity constraint of a simulation complexity and runtime. Using training images from a wide range of positions and orientations on a given abdominal model, our generator learns to simulate B-mode images from arbitrary slices of this model, as assessed with unseen image locations in our evaluation. For other anatomical models or locations, one can retrain the network with their corresponding simulated images and cross-sections.

Note that paired segmentation/attenuation maps and US images are required to train our network. The choice of training data is therefore restricted to synthetic images. Hence, the realism of the generated images by our proposed approach cannot go beyond the realism of the underlying simulation. To incorporate real ultrasound images, one may utilize unpaired deep learning models, e.g. cycleGAN [27,29] , as a potential future research direction.

Our work similarly to many other GAN-based techniques for US simulation [8,26,27] deal with single image translation or generation. However, for sonographer training purposes, interactive and dynamic simulators are required to generate temporally consistent image sequences. However, without explicitly enforcing temporal consistency, these methods are not guaranteed to generate images free from temporal artifacts, which is a potential research direction for future work.

Conclusions

In this work, we demonstrate a GAN-based framework with a local feature preserving generator architecture for learning ultrasound rendering from cross-sectional segmentation and subsequent integral attenuation maps, the latter of which greatly facilitates the generation of directional acoustic shadows. Combining with texture-friendly decoder blocks and a proposed noise feeding strategy, the presented network improves quality of translated images in terms of structure and texture compared to the state-of-the-art and any baseline method with an improvement of over 50% in FID score. An extensive ablation study has been carried out to demonstrate the effectiveness of each proposed architectural contribution. Compared to an earlier approach of simulated ultrasound generation using image translation from low-quality rendered images, the current approach does not require sophisticated online rendering step, while still able to generate ultrasound images with good anatomical structural correspondence and superior texture appearance compared to that state-of-the-art. For evaluation, besides traditional metrics we also propose a local histogram statistics based metric, while demonstrating on examples how it captures visually-perceived local differences between images, which is a typically challenging task due to the inherent speckle noise in ultrasound images.

Learning rendering with GANs warrants sophisticated rendering for ultrasound simulation to be carried out on consumer-hardware. The current rendering simulation settings lead to a frame time of 75 ms, a low-end of visually-acceptable US interactivity. Our GAN requires a constant computation time of 40 ms per frame using TensorRT, hence nearly doubling the frame rate. Moreover, pre-trained network can further be efficiently transferred and implemented on low-end devices, such as with FPGAs [6]. With a wider availability of simulated training of ultrasound with realistic imagery, future sonographers can be trained more effectively and also for rare-cases both for diagnostic and interventional imaging applications.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethics approval and Informed Consent Being based on synthetic models and simulated data, no ethical approval nor consent was required for this work.

References

1. Bargsten L, Schlaefer A (2020) 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(9):1427–1436
2. Blum T, Rieger A, Navab N, Friess H, Martignoni M (2013) A review of computer-based simulators for ultrasound training. Simulat Healthcare 8(2):98–108
3. Burger B, Bettinghausen S, Radle M, Hesser J (2013) Real-time GPU-based ultrasound simulation using deformable mesh models. IEEE Trans Med Imag 32(3):609–618
4. Goksel O, Salcudean SE (2009) B-mode ultrasound image simulation in deformable 3-D medium. IEEE Trans med imag 28(11):1657–1669
5. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio, Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680
6. Guo K, Zeng S, Yu J, Wang Y, Yang H (2017) A survey of fpga-based neural network accelerator. ACM Trans. Reconfig. Technol. Syst. **9**(4), 11:1–26
7. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems, pp. 6626–6637
8. Hu Y, Gibson E, Lee LL, Xie W, Barratt DC, Vercauteren T, Noble JA (2017) Freehand ultrasound image simulation with spatially-conditioned generative adversarial networks. In: Molecular imaging, reconstruction and analysis of moving body organs, and stroke imaging and treatment, pp. 105–115
9. Jafari MH, Girgis H, Van Woudenberg N, Moulson N, Luong C, Fung A, Balthazaar S, Jue J, Tsang M, Nair P, Gin K, Rohling R, Abolmaesumi P, Tsang T (2020) Cardiac point-of-care to cart-based ultrasound translation using constrained CycleGAN. International Journal of Computer Assisted Radiology and Surgery pp. 1–10
10. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4401–4410

11. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: International Conference on Learning Representations, pp. 1–15

12. Magnetti C, Zimmer V, Ghavami N, Skelton E, Matthew J, Lloyd K, Hajnal J, Schnabel JA, Gomez A (2020) Deep generative models to simulate 2D patient-specific ultrasound images in real time. In: Annual Conference on Medical Image Understanding and Analysis, pp. 423–435

13. Mailloux GE, Bertrand M, Stampfler R, Ethier S (1985) Local histogram information content of ultrasound B-mode echographic texture. Ultrasound med & biology 11(5):743–750

14. Mattausch O, Goksel O (2015) Scatterer reconstruction and parametrization of homogeneous tissue for ultrasound image simulation. In: IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6350–6353

15. Mattausch O, Makhinya M, Goksel O (2018) Realistic ultrasound simulation of complex surface models using interactive monte-carlo path tracing. Comput Graph Forum 37:202–213

16. Odena A, Dumoulin V, Olah C (2016) Deconvolution and checker-board artifacts. Distill 1(10):e3

17. Østergaard ML, Nielsen KR, Albrecht-Beste E, Ersbøll AK, Konge L, Nielsen MB (2019) Simulator training improves ultrasound scanning performance on patients: A randomized controlled trial. Eur radiol 29(6):3210–3218

18. Park T, Liu MY, Wang TC, Zhu JY (2019) Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2337–2346

19. Ramaiah SD, Ray KA, Reindollar RH (2020) Simulation training for embryo transfer: findings from the american society for reproductive medicine embryo transfer certificate course. Fertility and Sterility

20. Salehi M, Ahmadi SA, Prevost R, Navab N, Wein W (2015) Patient-specific 3D ultrasound simulation based on convolutional ray-tracing and appearance optimization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 510–518

21. Shams R, Hartley R, Navab N (2008) Real-time simulation of medical ultrasound from CT images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 734–741

22. Shankar PM, Reid JM, Ortega H, Piccoli CW, Goldberg BB (1993) Use of non-rayleigh statistics for the identification of tumors in ultrasonic B-scans of the breast. IEEE Trans Med Imag 12(4):687–692

23. Starkov R, Tanner C, Bajka M, Goksel O (2019) Ultrasound simulation with animated anatomical models and on-the-fly fusion with real images via path-tracing. Comput & Graph 82:44–52

24. Starkov R, Zhang L, Bajka M, Tanner C, Goksel O (2019) Ultrasound simulation with deformable and patient-specific scatterer maps. International Journal of Computer Assisted Radiology and Surgery 14(9):1589–1599

25. Sun B, McKenzie F (2011) Real-time sonography simulation for medical training. Int J Edu Info Technol 5(3):328–335

26. Tom F, Sheet D (2018) Simulating patho-realistic ultrasound images using deep generative networks with adversarial learning. In: International Symposium on Biomedical Imaging (ISBI), pp. 1174–1177

27. Vitale S, Orlando JI, Iarussi E, Larrabide I (2019) Improving realism in patient-specific abdominal ultrasound simulation using CycleGANs. Int J Comput Assisted Radiol Surg. 1–10

28. Zhang L, Portenier T, Paulus C, Goksel O (2020) Deep image translation for enhancing simulated ultrasound images. In: Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis, pp. 85–94

29. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232