Content-Preserving Unpaired Translation from Simulated to Realistic Ultrasound Images

Devavrat Tomar[1], Lin Zhang[1(✉)], Tiziano Portenier[1], and Orcun Goksel[1,2]

[1] Computer-Assisted Applications in Medicine, ETH Zurich, Zurich, Switzerland
lin.zhang@vision.ee.ethz.ch
[2] Department of Information Technology, Uppsala University, Uppsala, Sweden

Abstract. Interactive simulation of ultrasound imaging greatly facilitates sonography training. Although ray-tracing based methods have shown promising results, obtaining realistic images requires substantial modeling effort and manual parameter tuning. In addition, current techniques still result in a significant appearance gap between simulated images and real clinical scans. Herein we introduce a novel *content-preserving* image translation framework (ConPres) to bridge this appearance gap, while maintaining the simulated anatomical layout. We achieve this goal by leveraging both simulated images with semantic segmentations and unpaired in-vivo ultrasound scans. Our framework is based on recent contrastive unpaired translation techniques and we propose a regularization approach by learning an auxiliary segmentation-to-real image translation task, which encourages the disentanglement of content and style. In addition, we extend the generator to be class-conditional, which enables the incorporation of additional losses, in particular a cyclic consistency loss, to further improve the translation quality. Qualitative and quantitative comparisons against state-of-the-art unpaired translation methods demonstrate the superiority of our proposed framework.

Keywords: Image translation · US simulation · Contrastive learning

1 Introduction

Ultrasound (US) is a commonly used medical imaging modality that supports real-time and safe clinical diagnosis, in particular in gynecology and obstetrics. However, the limited image quality and the hand-eye coordination required for probe manipulation necessitate extensive training of sonographers in image interpretation and navigation. Volunteer access and realism of phantoms being limited

D. Tomar and L. Zhang—Both authors contributed equally to this manuscript.

for training, especially of rare diseases, computational methods become essential as simulation-based training tools. To that end, interpolation of pre-acquired US volumes [6] provide only limited image diversity. Nevertheless, ray-tracing based methods have been demonstrated to successfully simulate images with realistic view-dependent ultrasonic artifacts, *e.g.* refraction and reflection [4]. Monte-Carlo ray-tracing [14] has further enabled realistic soft shadows and fuzzy reflections, while animated models and fusion of partial-frame simulations were also presented [20]. However, the simulation realism depends highly on the underlying anatomical models and the parametrization of tissue properties. Especially the noisy appearance of ultrasound images with typical speckle patterns are nontrivial to parameterize. Despite several approaches proposed to that end [13,21,27], images simulated from anatomical models still lack realism, with the generated images appearing synthetic compared to real US scans.

Learning-based image translation techniques have received increasing interest in solving ultrasound imaging tasks, *e.g.* cross-modality translation [11], image enhancement [10,25,26], and semantic image synthesis [2,22]. The aim of these techniques is to map images from a source domain to target domain, e.g. mapping low- to high-quality images. Generative adversarial networks (GANs) [7] have been widely used in image translation due to their superior performance in generating realistic images compared to supervised losses. In the paired setting, where images in the source domain have a corresponding ground truth image in the target domain, a combination of supervised per-pixel losses and a conditional GAN loss [15] has shown great success on various translation tasks [9]. In the absence of paired training samples, the translation problem becomes underconstrained and additional constraints are required to learn a successful translation. To tackle this issue, a cyclic consistency loss (cycleGAN) was proposed [28], where an inverse mapping from target to source domain is learned simultaneously, while a cycle consistency is ensured by minimizing a reconstruction loss between the output of the inverse mapping and the source image itself. Recent works have extended and applied cycle consistency on multi-domain translation [1,5,29]. Cycle consistency assumes a strong bijective relation between the domains. To relax the bijectivity assumption and reduce the training burden, Park et al. [17] proposed an alternative with a single-sided unpaired translation technique with contrastive learning. For US simulation, the standard cycleGAN was used in [24] to improve the realism of simulated US image frames, however, this method is prone to generate unrealistic deformations and hallucinated features.

In this work, we aim to improve the realism of computationally-simulated US images by converting their appearance to that of real in-vivo US scans, while preserving their anatomical content and view-dependent artefacts originating from the preceeding computational simulation. We build our framework on a recent contrastive unpaired translation framework [17] and introduce several contributions to improve translation quality. In particular, to encourage content preservation, we propose to (i) constrain the generator with the accompanying semantic labels of simulated images by learning an auxiliary segmentation-to-real image translation task; and (ii) apply a class-conditional generator, which in turn enables the incorporation of a cyclic loss.

2   Method

Given unpaired source images $X = \{x \in \mathbb{X}\}$ and target images $Y = \{y \in \mathbb{Y}\}$, we aim to learn a generator function $G : \mathbb{X} \mapsto \mathbb{Y}$, such that mapped images $G(x)$ have similar appearance (style) as images in Y, while preserving the structural content of the input image x. To achieve this goal, we divide G into an encoder G_{enc} and a decoder G_{dec}. G_{enc} is restricted to extract content-related features only, while G_{dec} learns to generate a desired target appearance using a patch contrastive loss. Combined with both cyclic and semantic regularizations, we design a multi-domain translation framework consisting of a single generator and discriminator (Fig. 1).
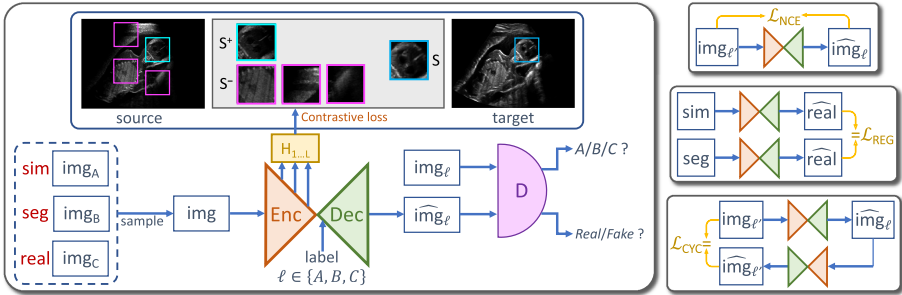

Fig. 1. (Left) Overview of our proposed framework. (Right) Illustrations of some of the loss functions used to train our model.

Adversarial Loss. We adopt the patchGAN discriminator [17] that discriminates real and fake images using a least squares GAN loss:

$$\mathcal{L}_{\text{GAN}}(X, Y) = \mathbb{E}_y \log[(D(y) - 1)^2] + \mathbb{E}_x \log[D(G(x))^2] . \tag{1}$$

Contrastive Loss. An unpaired contrastive translation framework (CUT) is presented in [17] that maximizes mutual information between image patches in the source and target domain to maintain the content of source images. The core of this approach is to enforce each translated patch to be (*i*) similar to the corresponding input patch, while (*ii*) different from any other input patches. For the similarity assessment, image patches are represented by hidden features of G_{enc}. A multi-layer perceptron (MLP) H_l with two hidden layers is then used to map the chosen encoder features h_l to an embedded representation $z_l = H_l(h_l) \in \mathbb{R}^{S_l \times C_l}$ with S_l spatial locations and C_l channels, where $h_l = G_{\text{enc}}^l(x)$ is the l-th hidden layer of G_{enc}. For each spatial location s in z_l, the corresponding patch feature vector $z_l^{s+} \in \mathbb{R}^{C_l}$ is then the positive sample and the features at any other locations are the negatives $z_l^{s-} \in \mathbb{R}^{(S_l-1) \times C_l}$. The corresponding patch

feature $\hat{z}_l^s = h_l(G_{enc}^l(\hat{y})) \in \mathbb{R}^{C_l}$ of the output image $\hat{y}$ acts as the query. The contrastive loss is defined as the cross-entropy loss

$$l(\hat{z}_l^s, z_l^{s+}, z_l^{s-}) = -\log \left[\frac{\exp(\hat{z}_l^s \cdot z_l^{s+}/\tau)}{\exp(\hat{z}_l^s \cdot z_l^{s+}/\tau) + \sum_{k=1}^{S_l-1} \exp(\hat{z}_l^s \cdot z_{l,k}^{s-}/\tau)} \right], \quad (2)$$

with the temperature parameter τ set to 0.07, following [17]. Using features from multiple encoder depths allows us to enforce patch similarity on multiple scales, leading to the following noise contrastive estimation (NCE) loss

$$\mathcal{L}_{\mathrm{NCE}}(X) = \mathbb{E}_x \sum_{l=1}^{L} \sum_{s=1}^{S_l} l(\hat{z}_l^s, z_l^{s+}, z_l^{s-}), \quad (3)$$

where L is the number of layers used for computing the loss. To encourage the generator to translate the domain-specific image appearance only, $\mathcal{L}_{\mathrm{NCE}}$ is also evaluated on the target domain $\mathbb{Y}$, which acts as an identity loss, similarly to the cyclic consistency loss in [28]. The final objective in CUT [17] is defined as

$$\mathcal{L}_{\mathrm{CUT}}(X,Y) = \mathcal{L}_{\mathrm{GAN}}(X,Y) + \mathcal{L}_{\mathrm{NCE}}(X) + \mathcal{L}_{\mathrm{NCE}}(Y). \quad (4)$$

Semantic-Consistent Regularization. To encourage the disentanglement of content and style, we leverage available surrogate segmentation maps $S = \{s \in \mathbb{S}\}$ of the simulated images (sim). In addition to sim-to-real translation, our generator then learns to also synthesize real images from segmentation maps (seg), *i.e.* seg-to-real translation. Since segmentation maps contain only content and no style, it is ensured that, after passing G_{enc}, there is no style left in the features, therefore G_{dec} has to introduce styles entirely from scratch. Learning this auxiliary task thus helps to prevent style leakage from G_{enc}, enforcing G_{enc} to extract only content-relevant features. In this modified CUT framework with semantic input (CUT+S), we minimize

$$\mathcal{L}_{\mathrm{CUT+S}} = \mathcal{L}_{\mathrm{CUT}}(X,Y) + \mathcal{L}_{\mathrm{GAN}}(S,Y) + \mathcal{L}_{\mathrm{NCE}}(S). \quad (5)$$

In addition, we regularize G to generate the same output for paired seg and sim, thus explicitly incorporating the semantic information of simulated images into the generator. We achieve this by minimizing the following semantic-consistent regularization loss: $\mathcal{L}_{\mathrm{REG}}(X,S) = \mathbb{E}_{x,s}||G(x) - G(s)||_1$. Our consistency-based training objective then becomes:

$$\mathcal{L}_{\mathrm{CUT+SC}} = \mathcal{L}_{\mathrm{CUT+S}} + \lambda_{\mathrm{REG}}\mathcal{L}_{\mathrm{REG}}(X,S). \quad (6)$$

Multi-Domain Translation. In preliminary experiments, we observed that despite the identity contrastive loss and semantic inputs, the generator still alters the image content, since the above losses do not explicitly enforce the structural consistency between input and translated images. To mitigate this issue, we require a cyclic consistency loss similar to [28]. For this purpose, we extend the

so-far single-direction translation to a multi-domain translation framework, while keeping a unified (now conditional) generator and discriminator, inspired by StarGAN [5]. Here, G_{dec} is trained to transfer the target appearance, conditioned by the target class label $\ell \in \{\mathbb{A}, \mathbb{B}, \mathbb{S}\}$ given the classes $\mathbb{A}$ simulated image, $\mathbb{B}$ real image, and $\mathbb{S}$ semantic map. The class label is encoded as a one-hot vector and concatenated to the input of the decoder. The cyclic consistency loss is then defined as

$$\mathcal{L}_{\text{CYC}}(X) = \mathbb{E}_{x,\ell,\ell'} ||x - G(G(x,\ell),\ell')||_1, \tag{7}$$

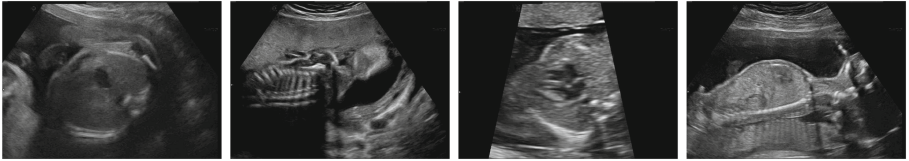where ℓ' is the class label of the input image and ℓ is label of the target class.

Fig. 2. Examples of in-vivo images used to train our model.

Classification Loss. To enable class-dependent classification (CLS) with the discriminator [5], D tries to predict the correct domain class label ℓ' for a given *real* image x as an auxiliary task, *i.e.*

$$\mathcal{L}_{\text{CLS,r}}(X) = \mathbb{E}_{x,\ell'}[-\log D(\ell'|x)], \tag{8}$$

while G tries to fool D with *fake* images to be classified as target domain ℓ by minimizing

$$\mathcal{L}_{\text{CLS,f}}(X) = \mathbb{E}_{x,\ell}[-\log D(\ell|G(x,\ell))]. \tag{9}$$

Final Objective. For our final model (ConPres), the training objective is evaluated by randomly sampling two pairs of domains $(X_i, Y_i) \in \{(\mathbb{A}, \mathbb{B}, \mathbb{S}) \backslash X_i \neq Y_i\}$ for $i = [1, 2]$, given the following discriminator and generator losses

$$\mathcal{L}_{\text{ConPres}}^{\text{D}} = \sum_{i=1}^{2} -\mathcal{L}_{\text{GAN}}(X_i, Y_i) + \lambda_{\text{CLS,r}} \mathcal{L}_{\text{CLS,r}}(X_i), \tag{10}$$

$$\mathcal{L}_{\text{ConPres}}^{\text{G}} = \sum_{i=1}^{2} \mathcal{L}_{\text{CUT}}(X_i, Y_i) + \lambda_{\text{CLS,f}} \mathcal{L}_{\text{CLS,f}}(X_i) + \lambda_{\text{CYC}} \mathcal{L}_{\text{CYC}}(X_i)$$
$$+ \mathbb{1}_{[(X_1 = \mathbb{A} \wedge X_2 = \mathbb{S}) \vee (X_1 = \mathbb{S} \wedge X_2 = \mathbb{A})]} \lambda_{\text{REG}} \mathcal{L}_{\text{REG}}(X_1, X_2) \tag{11}$$

with the indicator function $\mathbb{1}_{[.]}$ and the hyperparameters $\lambda_{\{.\}}$ for weighting loss components. We set $\lambda_{\text{REG}}=0$ when the two source domains are not $\mathbb{A}$ and $\mathbb{S}$.

3   Experiments and Results

Real In-vivo Images. 22 ultrasound sequences were collected using a GE Voluson E8 machine during standard fetal screening exams of 8 patients. Each

sequence is several seconds long. We extracted all 4427 frames and resize them to 256×354, see Fig. 2 for some examples. The resulting image set was randomly split into training-validation-test sets by a 80–10–10% ratio.

US Simulation. We used a ray-tracing framework to render B-mode images from a geometric fetal model, by simulating a convex probe placed at multiple locations and orientations on the abdominal surface, with imaging settings listed in the supplement. At each location, simply rasterizing a cross-section through the triangulated anatomical surfaces at the ultrasound center imaging plane provided corresponding semantic maps. Figure 3 shows example B-mode images with corresponding semantic maps. A total of 6669 simulated frames were resized to 256×354 and randomly split into training-validation-test sets by 80–10–10%.

Metrics.   We use the following metrics to quantitatively evaluate our method:

- **Structural similarity index** (SSIM) measures the structural similarity between simulated and translated images, quantifying content preservation. We evaluate SSIM within regions having content in simulated images.
- **Fréchet inception distance** (FID) [8] measures the feature distribution difference between two sets of images, herein real and translated, using feature vectors of Inception network. Since a large number of samples is required to reduce estimation bias, we use the *pre-aux* layer features, which has a smaller dimensionality than the default pooling layer features.
- **Kernel inception distance** (KID) [3] is an alternative unbiased metric to evaluate GAN performance. KID is computed as the squared maximum mean-discrepancy between the features of Inception network. We use the default pooling layer features of Inception, to compute this score.

Implementation Details.   We use a least-squares GAN loss with patchGAN discriminator as in [5]. The generator follows an encoder-decoder architecture, where the encoder consists of two stride-2 convolution layers followed by 4 residual blocks, while the decoder consists of 4 residual blocks followed by two convolution layers with bilinear upsampling. For architectural details, please see the supplementary material. To compute the contrastive loss, we extract features from the input layer, the stride-2 convolution layers, and the outputs of the first three residual blocks of the encoder. For CUT and its variants CUT+S and CUT+SC, we used the default layers in [17]. To compute λ_{REG}, the sampled simulated and segmentation images in each batch are paired. We used Adam [12] optimizer to train our model for 100 epochs with an l_2 regularization of 10^{-4} on model parameters with gradient clipping and $\beta = (0.5, 0.999)$. We set $\lambda_{\text{CLS},*} = 0.1$, $\lambda_{\text{REG}} = 1$ and $\lambda_{\text{CYC}} = 10$. We set the hyper-parameters based on similar losses in the compared implementations, for comparability; while we grid-searched the others, *e.g.* λ_{REG}, for stable GAN training. We implemented our model in PyTorch [19]. For KID and FID computations, we used the implementation of [16].

Comparative Study.   We compare our proposed ConPres to several state-of-the-art unpaired image translation methods:

- **CycleGAN** [28]: A conventional approach with cyclic consistency loss.
- **SASAN** [23]: CycleGAN extension with self-attentive spatial adaptive normalization, leveraging semantic information to retain anatomical structures, while translating using spatial attention modules and SPADE layers [18].
- **CUT** [17]: Unpaired contrastive framework for image translation.
- **StarGAN** [5]: A unified GAN framework for multi-domain translation.

We used the official implementations and default hyperparameters for training all the baselines. To assess the effectiveness of the proposed architecture and losses, we also compare with the models CUT+S (CUT plus the seg-to-real translation) and CUT+SC (CUT+S plus $\mathcal{L}_{\mathrm{REG}}$).

Fig. 3. Qualitative results, with images masked by foreground in segmentations.

In Fig. 3 we show that only learning an auxiliary seg-to-real translation, i.e. CUT+S, cannot guide the network to learn the semantics of simulated images.

CUT+SC with the loss term $\mathcal{L}_{\text{REG}}$ largely reduces hallucinated image content, although it still fails to generate fine anatomical details. With the multi-domain conditional generator and additional losses of ConPres, translated images preserve content and feature a realistic appearance. Training without $\mathcal{L}_{\text{NCE}}$ leads to training instability.

Comparison to State-of-the-Art. As seen qualitatively from the examples in Fig. 3, our method substantially outperforms the alternatives in terms of content preservation, while translating realistic US appearance. CycleGAN, SASAN, and CUT hallucinate inexistent tissue regions fail to generate fine anatomical structures, *e.g.* the ribs. StarGAN fails to generate faithful ultrasound speckle appearance, which leads to highly unrealistic images. Our method ConPres preserves anatomical structures, while enhancing the images with a realistic appearance. It further faithfully preserves acoustic shadows, even without explicit enforcement. However, as seen from the last column, the refraction artefact appears artificial in the images translated by all the methods. Note that although the imaging field-of-view (FoV) and probe opening in the simulation is significantly different from the real in-vivo images (Fig. 2) used for training, our ConPres maintains the input FoV closely compared to previous state-of-the-art. The results in Table 1 quantitatively confirm the superiority of our method. Note that SSIM and FID/KID are used to measure translation performance from two different and sometimes competing aspects, with the former metric for quantifying structure preservation and the latter metrics for image realism.

Table 1. Quantitative metrics and ranking from the user study (mean $\pm$std). Best results are marked bold. "Seg" gives if semantic maps are used as network input.

	Seg	Method	SSIM ↑	FID ↓	KID ↓	Ranking $\in [1,6]$↓
	—	Simulation	—	2.37	0.41	3.98 ± 1.35
Others	✗	CycleGAN [28]	71.73 ± 5.18	1.78	0.32	2.86 ± 1.27
	✓	SASAN [23]	68.20 ± 4.00	2.36	0.39	3.59 ± 1.55
	✗	CUT [17]	67.28 ± 4.62	1.77	0.31	2.92 ± 1.20
	✓	StarGAN [5]	63.62 ± 4.82	1.93	0.47	5.76 ± 0.61
Ours	✓	CUT+S	68.88 ± 4.63	2.25	0.41	
	✓	CUT+SC	$\mathbf{80.56 \pm 2.11}$	1.87	0.38	
	✓	ConPres	72.13 ± 4.58	**1.51**	**0.24**	**1.89±1.07**

A user study was performed with 18 participants (14 technical and 4 clinical ultrasound experts) to evaluate the realism of translated images for 20 US frames. For each frame, a separate questionaire window opened in a web interface, presenting the participants with six candidate images including the input simulated

famre and its translated versions using CUT, CycleGAN, SASAN, StarGAN, and ConPres. As a reference for the given ultrasound machine appearance, we also showed a fixed set of 10 real in-vivo images. The participants were asked to rank the candidate images based on "their likelihood for being an image from this machine". The average rank score is reported in Table 1. Based on a paired Wilcoxon signed rank test, our method is significantly superior to any competing method (all p-values $< 10^{-18}$).

Discussion. Note that, despite both being fetal images, the simulated and the real images have substantially different anatomical contents, which makes the translation task extremely challenging. Nevertheless, our proposed framework is able to generate images with appearance strikingly close to real images, with far superior realism than its competitors. Besides sim-to-real translation, given its multi-domain conditional nature, our proposed framework without any further training can also translate images between the other domains, *e.g.* seg-to-real or seg-to-sim, with examples presented in the supplementary material.

4   Conclusions

We have introduced a contrastive unpaired translation framework with a class-conditional generator, for improving ultrasound simulation realism. By applying cyclic and semantic consistency constraints, our proposed method can translate domain-specific appearance, while preserving the original content. This is shown to outperform state-of-the-art unpaired translation methods. With the proposed methods, we largely close the appearance gap between simulated and real images. Future works may include an evaluation of the effects of translated images on US training as well as an investigation of seg-to-real image translation, which can enable to completely dispense with any expensive rendering.

References

1. Almahairi, A., Rajeshwar, S., Sordoni, A., Bachman, P., Courville, A.: Augmented CycleGAN: learning many-to-many mappings from unpaired data. In: International Conference on Machine Learning (ICML), pp. 195–204 (2018)
2. Bargsten, L., Schlaefer, A.: SpeckleGAN: a generative adversarial network with an adaptive speckle layer to augment limited training data for ultrasound image processing. Int. J. Comp. Asst. Radiol. Surg. **15**(9), 1427–1436 (2020)
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. preprint arXiv:1801.01401 (2018)
4. Burger, B., Bettinghausen, S., Radle, M., Hesser, J.: Real-time GPU-based ultrasound simulation using deformable mesh models. IEEE Trans. Med. Imaging **32**(3), 609–618 (2013)

5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8789–8797 (2018)
6. Goksel, O., Salcudean, S.E.: B-Mode ultrasound image simulation in deformable 3-D medium. IEEE Trans. Med. Imaging **28**(11), 1657–1669 (2009)
7. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 2672–2680 (2014)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 6626–6637 (2017)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1125–1134 (2017)
10. Jafari, M.H., et al.: Cardiac point-of-care to cart-based ultrasound translation using constrained CycleGAN. Int. J. Comp. Asst. Radiol. Surg. **15**, 1–10 (2020)
11. Jiao, J., Namburete, A.I., Papageorghiou, A.T., Noble, J.A.: Self-supervised ultrasound to MRI fetal brain image synthesis. IEEE Trans. Med. Imaging **39**(12), 4413–4424 (2020)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
13. Mattausch, O., Goksel, O.: Image-based reconstruction of tissue scatterers using beam steering for ultrasound simulation. IEEE Trans. Med. Imaging **37**(3), 767–780 (2017)
14. Mattausch, O., Makhinya, M., Goksel, O.: Realistic ultrasound simulation of complex surface models using interactive Monte-Carlo path tracing. Comput. Graph. Forum **37**, 202–213 (2018)
15. Mirza, M., Osindero, S.: Conditional generative adversarial nets. preprint arXiv:1411.1784 (2014)
16. Obukhov, A., Seitzer, M., Wu, P.W., Zhydenko, S., Kyl, J., Lin, E.Y.J.: High-fidelity performance metrics for generative models in pytorch (2020). https://github.com/toshas/torch-fidelity. Accessed 26 Feb 2021
17. Park, T., Efros, A.A., Zhang, R., Zhu, J.-Y.: Contrastive learning for unpaired image-to-image translation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 319–345. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_19
18. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2337–2346 (2019)
19. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. preprint arXiv:1912.01703 (2019)
20. Starkov, R., Tanner, C., Bajka, M., Goksel, O.: Ultrasound simulation with animated anatomical models and on-the-fly fusion with real images via path-tracing. Comput. Graph. **82**, 44–52 (2019)
21. Starkov, R., Zhang, L., Bajka, M., Tanner, C., Goksel, O.: Ultrasound simulation with deformable and patient-specific scatterer maps. Int. J. Comp. Asst. Radiol. Surg. **14**(9), 1589–1599 (2019)
22. Tom, F., Sheet, D.: Simulating patho-realistic ultrasound images using deep generative networks with adversarial learning. In: IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1174–1177 (2018)

23. Tomar, D., Lortkipanidze, M., Vray, G., Bozorgtabar, B., Thiran, J.P.: Self-attentive spatial adaptive normalization for cross-modality domain adaptation. IEEE Trans. Med. Imaging (2021)
24. Vitale, S., Orlando, J.I., Iarussi, E., Larrabide, I.: Improving realism in patient-specific abdominal ultrasound simulation using CycleGANs. Int. J. Comp. Asst. Radiol. Surg. **15**, 1–10 (2019)
25. Zhang, L., Portenier, T., Goksel, O.: Learning ultrasound rendering from cross-sectional model slices for simulated training. Int. J. Comp. Asst. Radiol. Surg. **16**(5), 721–730 (2021)
26. Zhang, L., Portenier, T., Paulus, C., Goksel, O.: Deep image translation for enhancing simulated ultrasound images. In: Hu, Y., et al. (eds.) ASMUS/PIPPI -2020. LNCS, vol. 12437, pp. 85–94. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60334-2_9
27. Zhang, L., Vishnevskiy, V., Goksel, O.: Deep network for scatterer distribution estimation for ultrasound image simulation. IEEE Trans. Ultrason. Ferroelectr. Freq. Control (TUFFC) **67**(12), 2553–2564 (2020)
28. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (CVPR), pp. 2223–2232 (2017)
29. Zhu, J.Y., et al.: Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)