Deep Ultrasound Simulation for Training Sonographers

A dissertation submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zürich)

presented by
Lin Zhang
Master of Science in Biomedical Engineering
ETH Zurich, Switzerland
born on 27. March 1994
citizen of China

accepted on the recommendation of
Prof. Dr. Orçun Göksel, examiner
Prof. Dr. Purang Abolmaesumi, co-examiner

2023
Dedicated to my fiancé Zhejun.
Abstract

Ultrasound (US) imaging is a low-cost and non-ionizing medical imaging modality, thus widely used in clinical diagnosis and therapy. However, due to the difficulties in image interpretation and probe manipulation, sonographers require extensive training. Current education relies mostly on books, tissue-mimicking phantoms, and volunteers. Books are not interactive, phantoms have limited realism, and training on real patients involve ethical concerns. These limitations motivate the development of computational simulation methods. These can enable real-life like training experience in an augmented-reality environment by using mock probes on mannequins. Various examination cases can then be simulated virtually and can be viewed interactively using different imaging settings. This is specifically important for the training of rare pathologies.

State-of-the-art methods in interactive ultrasound simulation for training employ computer-graphics based rendering of US images using ray-based representations of acoustic wavefronts. Such methods are already at a level of realism that allow for and currently utilized in some basic clinical training scenarios. However, these methods necessitate cumbersome modelling efforts and the manual fine tuning of simulation parameters in order to make the simulated images appear realistic. Even with such tedious efforts, simulated images often look artificial, with a clear discrepancy between them and in-vivo images. Minute anatomical detail cannot be captured in a computational model, both due to lack of such info and given reasonable human-resource effort. Furthermore, ultrasound physics is very difficult to fully replicate realistically in a computational environment and unknown post-processing steps that are manufacturer-specific and typically a trade-secret make generating machine-like images very challenging. Simulation quality is then often a trade-off between image realism and computational time, i.e. framerate, to satisfy interactivity constraints. To achieve high framerates, image quality needs to be compromised, e.g., by reducing essential, but time consuming simulation features. Recent advances in data-driven methods, specifically deep learning techniques, have shown promising results in generating photo-realistic images at interactive framerates. This thesis aims to investigate and design deep learning based approaches to improve the state-of-the-art ray-based simulators, enabling physics-motivated, realistic, and fast simulation frameworks for medical education.

First we investigate automatic estimation of tissue scatterer distributions, which is a nontrivial input parameter of modern ray-based simulation schemes and is vital for generating realistic ultrasound speckle textures. We propose a deep learning technique to learn a mapping between US images and the parameters of an underlying tissue scatterer distribution that would yield these images. Next we build a series of image-to-image translation techniques to increase the realism of simulated images, while keeping interactive simulation rates. To that end, we investigate generating high-quality simulated images, first from low-quality simulated images by recovering missing image information such that image quality can be increased without compromise on frame-rate. We next
extend this to generate simulated images directly from scene layouts, which obviates the need for an interactive complex simulation rendering process and any need to tune model-specific parameters. We also investigate generating photorealistic simulations with images resembling real-life ones. To that end, we present novel methods for unpaired training of simulated image to real image translation that enable generating real-life like images from simulated ones at inference time. Last but not the least, we introduce a novel bidirectional label-image translation framework, which enables realistic image generation directly from given scene layouts, obviating any computationally expensive rendering operation during real-time inference stage. With this proposed method, complex image simulations are learned during a non-time-critical offline learning stage, and machine-specific translations are learned using real image samples, also eliminating the cumbersome fine tuning of model parameters.

With the above contributions, our methods overcomes appearance shift between simulations and real world data, and enables realistic and real-time simulation for medical training with minimal modelling efforts.
Zusammenfassung


Mit den oben genannten Beiträgen überwindet unsere Methode die Verschiebung des Erscheinungsbildes zwischen Simulationen und realen Daten und ermöglicht eine realistische und Echtzeitsimulation für die medizinische Ausbildung mit minimalem Modellierungsaufwand.
Acknowledgements

The research presented in this thesis was supported by the Swiss Innovation Agency Innosuisse, collaborated with the industry partner VirtaMed.

First and foremost I am extremely grateful to my supervisor, Prof. Dr. Orçun Göksel, for providing me the opportunity to start my PhD in his research group, and continuous, unlimited support and advice during my PhD study, that will benefit me throughout my life. Thanks to his immense knowledge and plentiful academic experience, we together have overcame all the difficulties and challenges during this long journey. It would have been impossible to finish my studies without his unwavering support, guidance and outstanding feedback.

My special thanks go to Prof. Dr. Sebastian Kozerke for kindly being my second supervisor for administrative matters at ETH. I would like to extend my gratitude to my thesis co-examiner Prof. Dr. Purang Abolmaesemi, for contributing his time to review the thesis. I also thank all my collaborators: Dr. Valery Vishnevskiy, Dr. Tiziano Portenier, Dr. Christine Tanner, Dr. Michael Bajka, Dr. Christoph Paulus, Rastislav Starkov, Devavrat Tomar, Xenia Augustin, Deniz Bezek, and Mert Bilgin, for their inputs and time in producing scientific works together. Many thanks to Dr. Tobias Martin, Dr. Christoph Paulus and Rastislav Starkov, for their collaboration and efforts during data curation, and our clinical collaborator Dr. Michael Bajka for his valuable feedback on my work.

In particular, I would like to express my deepest appreciation to my technical cosupervisors Dr. Valery Vishnevskiy and Dr. Tiziano Portenier for countless scientific discussions. They were always there for discussing anything that I was unsure on. And their passion in research have inspired me a lot and positively affected my academic life.

My appreciation also goes to all the amazing group members at Computer Vision Lab for making my PhD life much more enjoyable: Dr. Firat Ozdemir, Dr. Alvaro Gomariz Carrillo, Dr. Fabien Pean, Dr. Baskhara Rao Chintada, Rastislav Starkov, Kevin Thandiackal, Dr. Pushpak Pati, Dieter Schweizer, Matija Ciganovic, Dr. Richard Rau, Dr. Baran Gözcü, Dr. Danda Paudel, Prof. Dr. Ender Konukoglu, Dr. Neerav Karani, Dr. Xiaoran Chen, Christina Krueger and many others; and to the new group members at Uppsala University: Deniz Can Bezek and Prof. Dr. Bozhi Liu.

Finally, but the most importantly, I would like to say big thank you to my lovely family, especially to my fiance Zhejun Zhang, for supporting me constantly over the past years. More than half of my PhD life was spent with Zhejun during the difficult pandemic time. No words can express my sincere appreciation for his support and accompanying. Lastly, I would like to mention my four-legs family members Panini and Pepino for filling my life with special meanings.
## Contents

1 Introduction
   1.1 Ultrasound Training .................................................. 3
   1.2 A Review of Related Methods in Ultrasound Simulation. ............ 4
      1.2.1 Convolution-based Simulation .................................. 4
      1.2.2 Ray-based Simulation ........................................... 6
      1.2.3 Challenges........................................................ 8
   1.3 Image-to-Image Translation. ......................................... 9
      1.3.1 Realistic Synthesis in Image-to-Image Translation ............. 9
      1.3.2 Content Preservation in Image-to-Image Translation .......... 12
      1.3.3 Problem Settings for Image-to-Image Translation .......... 13
   1.4 Thesis Goals .......................................................... 15
   1.5 Thesis Outline and Contributions .................................. 16

2 Deep Network for Scatterer Distribution Estimation for Ultrasound Image Simulation
   2.1 Introduction. .......................................................... 19
   2.2 Background .......................................................... 22
      2.2.1 Forward Problem of Ultrasound Simulation .................... 22
      2.2.2 Inverse Problem of Scatterer Reconstruction .................. 22
   2.3 Methods ............................................................... 23
      2.3.1 Statistical Model of Scatterer Distribution .................... 23
      2.3.2 Training Set Generation ....................................... 24
      2.3.3 Network Architecture and Training ............................ 25
   2.4 Experiments and Results ............................................. 26
      2.4.1 1D Example ....................................................... 26
      2.4.2 Evaluation Metrics .............................................. 27
      2.4.3 Synthetic Data .................................................... 29
      2.4.4 Gelatin Phantom ................................................ 32
      2.4.5 In Vivo Experiment. ............................................. 35
   2.5 Discussion ............................................................ 37
   2.6 Conclusion ............................................................ 40

3 Deep Image Translation for Enhancing Simulated Ultrasound Images  41
   3.1 Introduction. .......................................................... 41
   3.2 Materials and Methods ............................................... 42
   3.3 Experiments and Results ............................................. 45
   3.4 Discussion and Conclusions ......................................... 48
4 Learning Ultrasound Rendering from Cross-Sectional Model Slices for Simulated Training 49
   4.1 Introduction .............................................. 50
   4.2 Methods .................................................. 53
      4.2.1 Ray-based Ultrasound Simulation ....................... 53
      4.2.2 Ultrasound Simulation using a Generative Adversarial Network . 54
   4.3 Results and Discussion .................................. 55
   4.4 Conclusions .............................................. 61

5 Content-Preserving Unpaired Translation from Simulated to Realistic Ultrasound Images 63
   5.1 Introduction .............................................. 63
   5.2 Method .................................................. 65
   5.3 Experiments and Results .................................. 67
   5.4 Conclusions .............................................. 70
   5.5 Appendix ............................................... 70

6 Unpaired translation from semantic label maps to images by leveraging domain-specific simulations 75
   6.1 Introduction .............................................. 75
   6.2 Results .................................................. 78
      6.2.1 Compared methods ...................................... 78
      6.2.2 Evaluation ............................................ 79
      6.2.3 Experiments .......................................... 80
   6.3 Discussion ............................................... 86
   6.4 Methods ............................................... 88

7 Summary and Perspectives 97
   7.1 Summary of Contributions ................................. 97
   7.2 Future Perspectives ..................................... 98
# List of Figures

1.1 Introduction to convolution-based simulation .......................... 5
1.2 Introduction to ray-based simulation ................................... 6
1.3 Example simulated images compared to an in-vivo scan ............... 9
1.4 Schematic overview of sample image-to-image translation problems, demonstrated using ultrasound images and label-maps as examples. .......... 14
1.5 Schematic overview of the thesis organization ......................... 16

2.1 Illustration of the proposed method ................................. 23
2.2 Generation of training images ....................................... 25
2.3 1D signal example with simulated RF and envelope signals .......... 27
2.4 Simulated images of the numerical phantom for views rotated by 15°, 30°, and 45° .................................................. 30
2.5 Performance with respect to rotation angle .......................... 31
2.6 Simulated images of the numerical phantom for axial compressions of 10%, 30%, and 50% strain ...................................... 33
2.7 Performance evolution with increasing axial strain .................. 34
2.8 Results of the phantom experiment .................................. 35
2.9 Qualitative results of the in-vivo experiment ....................... 36
2.10 Speckle statistics comparison of the in-vivo experiment .......... 37

3.1 Example images of the employed data .............................. 43
3.2 Network Architecture .................................................. 45
3.3 Qualitative results ..................................................... 46
3.4 Inference on full field-of-view (FoV) images ......................... 48

4.1 Method overview ....................................................... 52
4.2 Inference on full field-of-view (FoV) images ......................... 58
4.3 Spatial pχ² error map for the examples shown in Fig. 4.2 ........... 59
4.4 Box plot of paired difference between SA2H and its ablated variants, LSA2H and NSA2H ......................................................... 60

5.1 (Left) Overview of our proposed framework. (Right) Illustrations of some of the loss functions used to train our model .................. 65
5.2 Examples of in-vivo images used to train our model ................. 67
5.3 Qualitative results, with images masked by foreground in segmentations . 72
5.4 Other sim-to-real image translation examples from our user study . 73
5.5 Example results of seg-to-real and seg-to-sim image translation .... 73
6.1 Overview of unpaired label-image translation by leveraging domain-specific simulations. .................................................. 77
6.2 Laparoscopy training experiment results. ............................... 81
6.3 Ultrasound training experiment results ................................. 83
6.4 Gaming experiment results ................................................. 85
6.5 Image-to-label translation results ........................................... 86
6.6 Overview of our proposed method SimIT ................................. 89
6.7 Additional qualitative label-to-image translation results ............... 94
6.8 Additional image-to-label translation results of SimIT ............... 95
List of Tables

2.1 Mean (mean), median (med), and maximum (max) errors for rotations between 0° and 45° with 1° increments ........................................... 30
2.2 Mean (mean), median (med) and maximum (max) errors across 10% to 50% strain with 1% increments ........................................... 32
2.3 Quantitative evaluation metrics for the gelatin phantom ................... 33

3.1 Simulation parameters ....................................................... 43
3.2 Quantitative results......................................................... 47

4.1 Simulation parameters ....................................................... 53
4.2 Quantitative results......................................................... 60

5.1 Quantitative metrics and ranking from the user study ...................... 69
5.2 Anatomical model and imaging parameters for the ultrasound simulation 70
5.3 Network architecture....................................................... 71

6.1 Quantitative evaluation..................................................... 82
Introduction

Medical ultrasound is well established for therapeutic as well as diagnostic purposes. For the former, US imaging is used to guide surgical procedures, such as needle insertion both in anaesthetic application [Gupta et al., 2011] and in prostate brachytherapy [Holm et al., 1983]. For the latter, US imaging is a safe diagnostic tool, especially for obstetrics and gynaecology, as US does not involve ionizing radiation. Therefore, ultrasound is commonly used during pregnancy for monitoring the development of fetus. With US, major structural anomalies of fetus, placenta and other pregnancy complications can be detected [Ewigman et al., 1993].

Need for training. Unlike other medical imaging modalities such as X-ray computer tomography and magnetic resonance imaging, US imaging exclusively relies on sonographer expertise in navigating a hand-held probe and visual inspection of the acquired images. Poor image quality, low signal-to-noise ratio, and various imaging artifacts make extensive sonographer training essential.

“See one, do one, teach one” [Vozenilek et al., 2004] has been a time-honored education concept for a long time. The medical education is mostly based on training on real patients, under the guidance of expert specialists. However, with increasing concern and focus on patient safety, this concept becomes less and less acceptable. Training on phantoms that emulate human tissue properties [Kim, 2016] can be an alternative practice tool, but such phantoms generally lack realism and are not capable of expressing real anatomical structures and all biological aspects.

A major limitation of the common medical education forms is the lack of ability to interactively train on rare pathologies. These can be vital to identify, but nevertheless often lack any opportunity to be observed in person during training. For example, fetal anomalies occur in only 2% to 5% of all pregnancies [Maul et al., 2004] thus an Ob/Gyn trainee would potentially only encounter a few such anomalies on actual patients. Indeed, after completing a typical one-year subject-specific education, students only see up to 80% of possible pathologies [Köhne et al., 2004]. Current limitations in the sonographer training have motivated the development of computational medical simulators.

Computational US simulation. Interpolative and generative methods are the two main computer-based techniques for simulating B-mode ultrasound images. The former approach generates realistic images by interpolating in 3D volumes reconstructed from pre-collected images. However, this approach can only replay prerecorded data and is thus highly restricted to the collected clinical images: The volumetric content cannot
be viewed from arbitrary directions and the method cannot generate novel content or cases. The latter method relies on solving complex acoustic wave equations, with relatively higher accuracy, but being very slow hence not suitable for real-time application. A linear approximation of wave interactions [Bamber et al., 1980] is the convolution model for US speckle generation for micro-level tissue interactions. This enables faster simulation but it does not capture macro-level acoustic interactions such as reflections, refraction, attenuation, and directional artefacts, with the images simulated using this method often appearing toy-like. In contrast to the above, advanced ray-based generative methods [Burger et al., 2012; Mattausch et al., 2018] allow for the simulation of new content, using 3D-modeled anatomical cases with different imaging parameters and conditions, in a real-time interactive virtual-reality (VR) environment. Such ray-based techniques model ultrasonic wave propagation using ray-tracing techniques on anatomical models, which are generated by computer-graphics artists with anatomical knowledge. Ray-based methods enable a new training paradigm based on a “See one, simulate more, teach more” ideology. Together with a mannequin and a mock probe providing realistic haptic feedback, medical students can be trained interactively for different pathologies with possible clinical scenarios. They can also practice without actual patients, whenever and wherever they want.

**Limitations of ray-based simulation.** There is a line of works on developing sophisticated ray tracing techniques to simulate realistic macroscopic wave interactions, such as reflection and refraction, which are essential for learning navigation skills. Despite several major advancements, the simulation realism in this area is bound by (1) the detail afforded by the underlying anatomical models and (2) the parametrization of local tissue acoustic properties, both of which are often non-trivial to provide: Large amounts of manual modelling efforts and tedious parameter tuning revisions are required, also necessitating close feedback from medical experts. Generating anatomical models from bottom-up with all the realistic and minute anatomical detail is nearly impossible. There are also unknown post-processing steps and image artefacts specific to US scanners, which cannot be fully modelled. Simulated images thus often look artificial, and a major discrepancy between simulated and in-vivo images prevails. Moreover, interactive computational constraints often lead to a compromise in image quality, e.g., the number of rays cast in a simulation determines the realism of images although this number is limited by the framerate for simulation to be interactive.

**Deep learning in image synthesis.** With the recent rise of deep learning, there has been a tremendous progress in image synthesis using data-driven approaches. Among these, generative adversarial network (GAN) [Goodfellow et al., 2014] is the leading technique in generating photorealistic images for content generation and image editing [Isola et al., 2017; Ling et al., 2021]. Recently some works have investigated the use of GANs in ultrasound simulation. A two-stage stack GAN is introduced in [Tom et al., 2018] for simulating intravascular US imagery given tissue echogenicity maps. In [Hu et al., 2017], freehand US fetal images are synthesized using a GAN conditioned on calibrated coordinates of the physical space. The first attempt to improve the realism of ray-traced US images is demonstrated in [Vitale et al., 2019], however, this method is prone to generating unrealistic deformations and hallucinated features.

In this thesis, a combination of ray-based simulation methods and deep learning approaches is studied i) for speeding up US image generation in simulations; and ii) for
1.1 Ultrasound Training

Medical ultrasound imaging uses high-frequency acoustic waves, which interact with different tissues in the body creating echoes recorded by US transducers. Since ultrasound often images a single cross-section within the tissue, the sonographer needs to successfully place the transducer to position the imaging plane on the intended anatomy, while ensuring an acoustic window, i.e., avoiding bones, gas, etc on the viewing path. Ultrasound being a real-time interactive modality, such navigational skill relies not only on a good anatomical knowledge but also on a good hand-eye coordination to infer the required probe motion from the displayed anatomical cross-section. Once the intended location is found, the sonographer needs to interpret the displayed image for the intended clinical need, which then requires image interpretation skills. During training, students hence need to learn to identify relevant anatomical structures (kidney, liver, heart, lungs, etc) and differentiate tissue types (soft tissue, bone, etc) based on image echogenicities, acoustic artefacts, and speckle textures. Image quality and visualization of certain structures highly depend on how the US probe is placed. Incorrect probe placement can lead to undesired distortions and image degradation. Probe manipulation (sliding, tilting, rotating, compressing, rocking) can both avoid such artifacts when they hinder the imaging, or help intentionally create some when they may help the diagnosis (e.g., to help infer inclusion content from acoustic shadowing, etc). All these above skills thus require extensive training.

An ideal simulator for sonographer training should mimic the real life screening experience as close as possible, leading to the following desiderata:

- **Interactivity:** Simulators need to be interactive, to mimic the real US examination experience where the user manipulates the imaging probe and sees the respective images in response, in real-time.

- **Anatomical accuracy:** The simulated images should contain all important and real anatomical structures, and ideally all structures in the region. For conventional model-based simulation approaches, for the image to contain a detail, the input 3D anatomical model should also have this modeled, which makes modeling very tedious.

- **Physics accuracy:** A correct simulation of ultrasound artefacts is crucial for distinguishing different anatomical structures and potential structural abnormalities. For example, acoustic shadow can help identify kidney stones via their strong reflection and attenuation.
• Range of content: To be realistic in representing a wide range of clinical scenarios and pathologies, a simulator should be able to operate on different anatomical models and cases.

• Visual realism: Realistic appearance contribute greatly to whether the simulator experience is close to real life counterpart. Realistic simulation creates a more immersive VR experience and also can facilitate the transfer of knowledge acquired from the simulation to real life applications.

• Low-cost: Training simulators with expensive hardware setups are challenging to popularize, especially considering middle to low income countries.

1.2 A Review of Related Methods in Ultrasound Simulation

Below we present a short review of some ultrasound simulation techniques relevant to this thesis, in particular leading to the techniques in [Mattausch et al., 2018] which was utilized in this thesis for the generation of simulated ultrasound images.

US simulation methods can be categorized into interpolative and generative approaches. The interpolative methods generate 2D US images by reconstructing from pre-collected 3D volumes [Rohling et al., 1999; San José-Estépar et al., 2003; Aiger et al., 1998; d’Aulignac et al., 2006; Goksel et al., 2009]. Being data-based, such techniques can generate highly realistic images, however, directional artefacts, e.g., acoustic shadowing, may not be represented correctly by interpolation. The variety of generated images highly depend on the collected data and it is difficult to simulate novel views and contents. In contrast, generative techniques consider ultrasound wave physics, thus can represent US directional artefacts more precisely. Wave-, convolution-, and ray-based methods are the three major generative approaches for US simulation. Wave-based approaches [Verweij et al., 2014; Treeby et al., 2012; Treeby et al., 2010] aim to simulate ultrasound propagation as exact as possible by solving complex wave equations. These approaches can thus take hours to simulate a single B-mode image and hence are not suitable for real-time applications. Both convolution and ray-based methods can operate in real-time and can thus be used for interactive training simulators. In the following, the concept of both methods are introduced and explained to motivate different parts and goals of this thesis that aims to improve ultrasound simulations for training.

1.2.1 Convolution-based Simulation

Speckle noise is an inherent property of ultrasound images. This granular noise texture is caused by sub-wavelength acoustic structures, which scatters the ultrasonic wave in all directions. Although speckle noise degrades the image quality, it can also help distinguish different tissues and a change in speckle textures can indicate potential pathologies. Therefore, simulating realistic speckle texture contributes not only to a realistic appearance of images for immersive training, but also is essential for training correct image interpretation, e.g., in image-based diagnosis.
1.2 A REVIEW OF RELATED METHODS IN ULTRASOUND SIMULATION

Figure 1.1: (a) Illustration of a point scatterer (top) and point spread function; (b) Example of simulated images with different speckle textures [Bamber et al., 1980] (c) Illustration of depth-wise varying PSFs [Mattausch et al., 2016]. PSF varies spatially, since the wave energy changes over the image domain.

With first-order Born approximation (weak scattering) for soft tissues [Jensen et al., 1993], US reflective intensity $I(l,a,e)$ can be defined as a convolution of a point spread function (PSF) $h(l,a,e)$ with a continuous distribution of point scatterers $g(l,a,e)$:

$$I(l,a,e) = g(l,a,e) * h(l,a,e) + \gamma(l,a,e)$$  \hspace{1cm} (1.1)

with $\gamma(l,a,e)$ representing additive noise and $(l,a,e)$ indicating the lateral, axial and elevational coordinates with respect to the probe origin.

Point spread function captures the ultrasound pulse and beam shape. It can be approximated with a three dimensional Gaussian envelope modulated by cosine function in the axial direction [Bamber et al., 1980], i.e. :

$$h(l,a,e) = e^{-\frac{l^2}{\sigma_l^2} - \frac{a^2}{\sigma_a^2} - \frac{e^2}{\sigma_e^2}} \cos(2\pi f_c a) = h_1(a) * h_2(l) * h_3(e),$$  \hspace{1cm} (1.2)

where $f_c$ is the acquisition center frequency; and $\sigma_a$, $\sigma_l$, and $\sigma_e$ denote the Gaussian envelope shape along the axial, lateral, and elevational directions, respectively. $h(l,a,e)$ can be separated into its axial component $h_1(a)$ representing the acoustic pulse, and $h_2(l)$ and $h_3(e)$ for the lateral and elevational beam profile. An illustration of a two-dimensional point spread function and simulated speckle textures is depicted in Figure 1.1(a-b). Due to focusing and diffraction effects, PSF in US imaging varies spatially, cf. Figure 1.1(c).

US speckles can be simulated algebraically by executing convolution per scatterer, which is adopted in the simulation framework Field II [Jensen, 2004]. Bamber et al. [Bamber et al., 1980] proposed the discrete approximation of the PSF convolution in pixel space. Later by leveraging the axial-separability property of PSF speckle, convolution-based simulation was accelerated through fast separable 1D-convolutions on modern GPUs [Gao et al., 2009]. Burger et al. [Burger et al., 2012] combined the separable PSF convolution approach with ray-based simulations, which is later adopted in a state-of-the-art Monte-carlo based ray-tracing framework in [Mattausch et al., 2018].
1.2.2 Ray-based Simulation

The convolution model introduced above can capture microscopic wave interactions, but it is not capable of simulating macroscopic wave interactions that happen at the interfaces of large anatomical structures. The latter is the main cause of typical directional US artefacts, such as acoustic shadowing. Acoustic wavefront can be approximated as rays transporting energy in tissue, similarly to light rays. This assumption enables fast ultrasound image simulation using ray-based approaches to model macroscopic interactions of acoustic waves using well-known ray interaction physics [Burger et al., 2012]. Advanced ray-based methods combine ray-based and convolution models, taking wave interactions at all scales into account.

**Physics of ray-tracing:** Wave interactions at tissue interfaces can be modelled as a ray-surface intersection based on the concept of raytracing. Starting from an ultrasound transducer element, the wavefront (hereafter, ray) traveling through a medium loses its energy as a result of absorption and backscattering within the medium. This phenomenon is known as attenuation, which can be modelled using Beer-Lambert law

\[ I(d) = I_0 e^{-αf^nd}, \]

where \( I_0 \) is the initial intensity, \( d \) is the distance traveled in the medium with respect to the origin, \( α \) is the attenuation coefficient of the medium, \( f \) is the wave frequency, and \( n \) is a tissue dependent constant (often \( n=1 \) in soft tissues).

When a ray with a normalized direction \( \vec{V}_i \) arrives at the interface of two media with different acoustic impedances at an angle \( θ_i \), a part of the ray is reflected while the rest of it is transmitted to the other side potentially changing its direction, i.e. so-called refraction. Let the incident, reflection, and transmittance (refraction) angles with respect to the
interface normal $\vec{N}$ be $\theta_i$, $\theta_r$, and $\theta_t$, respectively. The normalized direction $\vec{V}_r$ and $\vec{V}_t$ of the reflected and transmitted (refracted) rays can then be computed as

$$\cos \theta_i = \vec{N}(-\vec{V}_i)$$

$$\cos \theta_t = \sqrt{1 - \left(\frac{Z_1}{Z_2}\right)^2(1 - \cos^2 \theta_i)}$$

$$\vec{V}_r = \vec{V}_i + 2(\cos \theta_i)\vec{N}$$

$$\vec{V}_t = \frac{Z_1}{Z_2}\vec{V}_i + \left(\frac{Z_1}{Z_2} \cos \theta_i - \cos \theta_t\right)$$

based on the acoustic impedances $Z_1$ and $Z_2$ of the incident medium1 and transmitted medium2. The transmittance (refraction) angle is given by Snell’s law

$$\frac{\sin \theta_i}{\sin \theta_t} = \frac{c_1}{c_2}$$

with the speed-of-sound $c_1$ and $c_2$ of medium1 and medium2, respectively. Reflection at an interface can be described based on interactions being at a macro- or a micro-level, as specular or diffuse, i.e. the former as a result of coherent reflections in the same direction, while the latter describing sub-wavelength level reflections causing incoherent distribution of some reflected acoustic energy. A simplified approximate model combining these two parts is used in [Burger et al., 2012], which adopts a modified Lambert’s cosine law with a parameter defining whether a medium reflects rays more specularly or diffusely.

The early ray-based simulation works [Burger et al., 2012; Computing et al., 2015] assume infinitely thin tissue surfaces, leading to unrealistic perfect reflection and refraction effects in simulated images. To generate more realistic wave interactions, stochastic Monte-Carlo path tracing method is proposed in [Mattausch et al., 2018]. Ray-surface interaction is simulated as a stochastic process and averaged over random ray paths cast from intersection points within the considered thickness of interacted surface. To better represent the actual tissue surfaces, a cosine-parametrized (rough) surface model is further proposed in [Mattausch et al., 2018]. Instead of a deterministic surface normal $\vec{N}$, perturbed normals sampled from a $\cos^s$ distribution are used. Then, $s = 0$ corresponds to a surface that reflects completely diffusely while $s = \infty$ reflects completely specularly.

A schematic comparison of the deterministic and stochastic surface models for ray-based simulation is illustrated in Figure 1.2(b). To avoid deterministic artifacts and to utilize the computation resources most efficiently, Monte-Carlo sampling techniques are then used to efficiently compute the above-mentioned interactions.

**Echo creation:** A recorded echo at tissue intersections is the sum of the reflection, refraction, and the backscattering term. While the refraction term is ignored in [Burger et al., 2012], a local illumination model is employed in [Mattausch et al., 2018] to take both reflection and refraction into account.

The backscattering is computed convolution-based using PSFs modelled by modulated Gaussian envelopes defined in Equation (1.2), which enables fast separable 1D convolutions. Depth-wise smoothly-varying PSFs are assumed for image simulation [Mattausch et al., 2017], as exemplified in Figure 1.1. Tissue scatterers are parametrized using a three-parameter stochastic model: a Gaussian distribution with $(\mu, \sigma)$ for scatterer amplitude and $d$ for scatterer density.
**Image generation:** A schematic illustration of a typical 2D ultrasound acquisition setup is shown in Figure 1.2(c). From each transducer element, a ray is cast and it travels through the tissues. Echos are recorded along each scanline as depicted in Figure 1.2(c). Out-of-plane effect is considered by instantiating rays at several elevational layers on the probe surface. Not to have deterministic discretization artifacts and to sample the analytical wavefront uniformly, several rays are emitted following a normal distribution around the center point of each intended scanline, i.e. to cover the physical transducer surface area. First, raytracing is used to find the ray intersections with triangulated anatomical surfaces and hence the ray segments in between each interface. Then, a ray marching phase is applied along each such segment, where the beamformed radio-frequency (RF) signal is simulated by convolving the instantaneous PSF (with the current ray amplitude) with the scatterer representation at the physical coordinates of each ray-marched point. Scatterers are referred in local coordinate frames attached to each anatomical structure such that motion and deformation of these objects are accurately represented [Starkov et al., 2019b; Starkov et al., 2019a]. To obtain final B-mode images, several post-processing steps are carried out including time-gain compensation, envelope detection, and dynamic compression.

The above-mentioned simulation techniques proposed in [Mattausch et al., 2018] are used in this thesis for the generation of simulated fetal ultrasound images.

**1.2.3 Challenges**

Example images simulated using the above stochastic raytracing model from a second-trimester abdominal scene are shown in Figure 1.3. Rendering images with an optimized simulation setting (high-quality images) herein takes 75 ms (13 fps), a low end of visually-acceptable US interactivity. Reducing the number of rays and elevational layers can further improve the simulation speed to 25 ms, however this would lead to a substantial degradation of image quality (low-quality images) unacceptable for actual training. The simulated scenes can also be more elaborate by increasing the detail and complexity of anatomical models but this would further increase computational costs via increased ray-model interactions to be computed. Accordingly, there is a trade-off between simulation quality and computation time for interactive frame-rates.

In Figure 1.3, we also observe that the simulated images look rather synthetic compared to the presented in-vivo example. This has multiple reasons: First, the realism for US speckle textures highly depends on the chosen tissue scatterer representation, which is difficult to parametrize due to lack of prior knowledge. Second, the simulated ray-based model is an approximation and it cannot capture all complex physical interactions. Furthermore, one even with a perfect interaction model, one cannot know what input parameters to use, e.g., what exact wave originates from which transducer elements, what exact anatomical detail is encountered, and even what post-processing methods (often trade-secret) the particular imaging device applies. Accordingly, even “perfectly” simulated ultrasound images will differ from actual US images.

To enable interactive and realistic simulation, automatic mapping from low-quality to high-quality simulated images as well as from simulated to real images are thus desired. Such mappings can be parametrized by neural networks with weights optimized
using many simulated and/or real images. Ideally, one might completely replace the computationally-intensive rendering step by generating images directly from label maps, which can be easily extracted from existing 3D anatomical models given the real-time interactive position of the transducer. Finding mapping functions as above is treated in this thesis as the problem of image translation.

1.3 Image-to-Image Translation

Image-to-image translation (I2I) is an active research field in computer vision, where the goal is to translate an image from one domain to another domain, such as from a sketch to an image, from day- to night-appearance, etc. The goal of I2I can be formally defined as learning a function $G: A \rightarrow B$, which maps source image $I_A \in A$ to the corresponding image $I_B$ in the target domain $B$. Thanks to recent advances in deep learning, there exists various learning based approaches for image translation. Among them, generative adversarial network (GAN), first introduced in [Goodfellow et al., 2014], has achieved great performance in generating photo-realistic images. This section includes the problem definition of I2I as well as a brief description of generative adversarial networks, which are commonly used techniques to address various I2I problems.

1.3.1 Realistic Synthesis in Image-to-Image Translation

Generative adversarial network (GAN) [Goodfellow et al., 2014] is the most common technique used to generate realistic-looking images, with the objective of generating image samples, which are indistinguishable to the training samples. Below the basic concept of GAN is introduced.

GAN composes of two deep neural networks, a generator $G(x; \theta_g)$ and a discriminator $D(x; \theta_d)$, parameterized by learnable parameters $\theta_g$ and $\theta_d$, respectively. $G$ learns to generate images from random noise input $z$, whereas $D$ acts as a binary classifier, which learns image features for distinguishing between real and fake images synthesized by
the generator. They are trained simultaneously in an adversarial manner, following the min-max game principle, i.e.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_{z}(z)}[\log(1 - D(G(z)))],$$  \hspace{1cm} (1.9)$$

where $p_{\text{data}}$ is the data distribution and $p_{z}$ is the noise distribution. In practice, the parameters of $G$ and $D$ are updated following an alternating training scheme. Given a generator $G$, the discriminator is trained using the following objective

$$\max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_{z}(z)}[\log(1 - D(G(z)))],$$  \hspace{1cm} (1.10)$$

which maximizes the probability of $D$ classifying $x$ as samples drawn from the distribution $p_{\text{data}}$ and minimizes the probability of classifying generated samples $G(z)$ as coming from $p_{\text{data}}$. Given a discriminator $D$, the generator is updated following

$$\min_G V(D, G) = \mathbb{E}_{z \sim p_{z}(z)}[\log(1 - D(G(z)))],$$  \hspace{1cm} (1.11)$$

which minimizes the probability of classifying generated samples $G(z)$ as not coming from $p_{\text{data}}$. In the ideal case, GAN training converges to a Nash equilibrium with $p_{x} = p_{\text{data}}$, where $p_{x}$ is the distribution of data generated by $G$.

This vanilla architecture is lack of control over generated data samples. To tackle this issue, a conditional version of GANs is introduced in [Mirza et al., 2014]. In the framework of conditional GANs, both $G$ and $D$ are conditioned on an additional input $y$, which is typically an image in the context of I2I tasks. The objective can then be written as

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x|y)] + \mathbb{E}_{z \sim p_{z}(z)}[\log(1 - D(G(z)|y))].$$  \hspace{1cm} (1.12)$$

**GAN loss functions:** In practice, it is very challenging to reach the Nash equilibrium and training GANs often suffers from the following typical problems:

- **Vanishing gradient:** If the discriminator dominates and is capable of classifying most of the samples correctly, the gradient for the generator vanishes. This often happens at the early training stage, when $D$ can easily find features to differentiate between real and fake samples while $G$ still performs relatively poorly.

- **Non-convergence:** An imbalance between generator and discriminator may lead to the training not converging, where the cost function then keeps oscillating.

- **Mode collapse:** When the generator fails to capture all the variety in the training samples, the generated images collapse to single modes.

There are many works investigating different loss functions for improving the stability of GAN training. Below, most commonly used alternative GAN losses in the context of I2I are listed and briefly explained:

- **Non-saturating GAN loss:** To tackle the vanishing gradient problem, a modification to the generator loss is proposed in the original GAN paper [Goodfellow et al., 2014]. Instead of minimizing the probability of predicting fake as real as in Equation (1.11),
the generator is then trained to maximize the probability of generated samples being predicted as real:

\[
\max_G V(D, G) = \mathbf{E}_{z \sim p_z}(\log D(G(z))).
\] (1.13)

This often leads to a more stable training.

- Wasserstein loss: This alternative GAN loss is introduced in [Arjovsky et al., 2017]. Training the vanilla GAN is equivalent to minimizing the Jensen-Shannon (JS) divergence between the real and predicted data distributions. The vanishing gradient problem caused by minimizing the JS divergence motivates the use of Wasserstein distance, which provides smooth and meaningful gradients throughout the network. Following the Kantorovich-Rubinstein duality, Wasserstein distance between two distributions \( P \) can be defined as

\[
W(P_r, P_g) = \frac{1}{K} \sup_{||f||_{L^1} < 1} \mathbf{E}_{x \sim P_r}[f(x)] - \mathbf{E}_{x \sim P_g}[f(x)],
\] (1.14)

where the function \( f \) is a 1-Lipschitz function, which can be parameterized by a neural network very similar to \( D \). This function \( f \) is often referred as critic, which scores images with real values instead of probabilities. The training objective is then defined as follows:

\[
\max_D V(D, G) = \mathbf{E}_{x \sim P_{\text{data}}}[D(x)] - \mathbf{E}_{z \sim p_z}[D(G(z))]
\] (1.15)

\[
\max_G V(D, G) = \mathbf{E}_{z \sim p_z}(\log D(G(z))).
\] (1.16)

To enforce the critic being a 1-Lipschitz function, the weights of \( D \) are constrained to be within a certain range \([-c, c]\) by a simple clipping. However, the choice of \( c \) can greatly affect the training convergence and quality of generated samples. To replace the weight clipping, gradient penalty (GP) is introduced in [Gulrajani et al., 2017] by penalizing the norm of gradient of \( D \) being away from its target norm 1 with respect to its input. Since a differentiable 1-Lipschitz function needs to have gradients with norm at most 1 everywhere, GP can then enforce \( D \) to be a 1-Lipschitz function. The training objective of Wasserstein GAN with GP (WGAN-GP) can be written as:

\[
\min_D V(D, G) = \mathbf{E}_{z \sim p_z}(D(G(z))) - \mathbf{E}_{x \sim P_{\text{data}}}[D(x)] + \lambda \mathbf{E}_{\delta \sim p_{\delta}}(||\nabla_D \delta||_2 - 1)^2,
\] (1.17)

where \( \delta \) is an data point randomly sampled along the line between real and generated data points. WGAN-GP was shown to be effective in stabilizing GAN training in complex image translation scenarios, e.g., involving multiple image domains [Choi et al., 2020].

- Least squares GAN (LSGAN) was proposed in [Mao et al., 2017] to tackle the vanishing gradient issue of GAN training, with the objective function defined as:

\[
\min_D V(D, G) = \mathbf{E}_{x \sim P_{\text{data}}}[((D(x) - 1)^2] + \mathbf{E}_{z \sim p_z}[D(G(z))^2]
\] (1.18)
\[
\min_G V(D, G) = E_{z\sim p_z(z)}[D(G(z))],
\]  
(1.19)

This least squares loss function penalizes generated samples based on their distance to the decision boundary, which addresses the problem of minimizing JS divergence, where the gradients for the samples far away from the decision boundary are often too small for the generator to learn anything useful. LSGAN is widely used for I2I problems [Liu et al., 2017; Park et al., 2020].

1.3.2 Content Preservation in Image-to-Image Translation

While GANs are capable of generating realistic images, preserving content of the source domain after translation becomes a major challenge in I2I tasks. Below, commonly used generator training losses for content preservation are listed, which can be mainly divided into pixel-level and feature-level losses:

Pixel-level losses:

- Reconstruction loss: pixel-wise image difference is typically used to enforce image fidelity, defined as:

\[
L(G) = E_{I_A,I_B}[||I_A - G(I_B)||_p],
\]  
(1.20)

where L1 and L2 norms are common choices for \(p\). Such pixel-wise loss incorporates image features other than scene structure, such as appearance, lighting, and texture, therefore can only be utilized when \(I_A\) and \(I_B\) are indeed paired [Isola et al., 2017].

- Cycle consistency loss: Without paired data, corresponding images in the source and target domains do not exist (or are not known) and the translation problem then becomes ill-posed. To reduce the solution space, cycle consistency loss was proposed in [Zhu et al., 2017b] assuming a bijective relation between source and target images. To that end, a reverse translation network \(F : B \rightarrow A\) is learned, while cycle consistency encourages forward and backward consistency via:

\[
L(G,F) = E_{I_A}[||F(G(I_A)) - I_A||_1] + E_{I_B}[||G(F(I_B)) - I_B||_1].
\]  
(1.21)

This idea has been widely leveraged in the I2I literature. Although the main purpose of cycle loss is to address unpaired data, it was also shown to be effective in content preservation, e.g., in combination with dual learning [Yi et al., 2017], stacked GANs [Li et al., 2018], and adaptive layer-instance normalization [Kim et al., 2019].

Feature-level losses:

- Perceptual loss: Instead of enforcing pixel-wise matching, perceptual loss [Johnson et al., 2016] compares high-level feature representations. It is computed as the Euclidean distance between the hidden layers of translated and target images, i.e.:

\[
L(G) = \frac{1}{M_i} ||F^{(i)}(I_B) - F^{(i)}(G(I_A))||^2_2,
\]  
(1.22)

where \(F^{(i)}\) denotes the \(i\)-th layer of a feature network \(F\) where \(M_i\) is the number of elements in \(F^{(i)}\). Pretrained deep networks, e.g., VGG [Simonyan et al., 2014],
is a common choice for $F$. Instead of using off-the-shelf pretrained networks, some works [Wang et al., 2018; Park et al., 2019; Portenier et al., 2020] propose to leverage the existing discriminator as feature extractor, as such discriminator is then trained explicitly on domain-specific images. Perceptual loss is capable of capturing high-level image content and texture features and invariant to small image shifts. However, it still cannot fully decouple the content and appearance features.

- **Constraints via shared latent space**: This idea [Liu et al., 2017] assumes functions $G_A$, $G_B$, $F_A$, and $F_B$ which map images from two separate domains $I_A$ and $I_B$ to a shared latent code $z = G_A(I_A) = G_B(I_B)$, using which the original images should then be possible to reconstruct, i.e., losses introduced to ensure $I_A = F_A(z)$ and $I_B = F_B(z)$. Note that these constraints implicitly imply a cycle consistency between the domains over the shared latent. This idea was later extended also for multi-modal image translation in [Huang et al., 2018].

- **Contrastive learning** [Park et al., 2020] is a method to bring in unsupervised information by contrasting feature projections of samples, often within a minibatch. It can be used image-level contrasting, e.g., for comparing images from within and across datasets for unsupervised classification [Chen et al., 2020] and segmentation [Gomariz et al., 2022]. It can also be used for patch-level contrasting which enforces structural similarity between images, e.g., the input and translated images in an unpaired translation setting. To that end, local image feature representations extracted from the generator encoder are contrasted in [Park et al., 2020], i.e.

$$
\mathcal{L}(z, z^+, z^-) = -\log \left( \frac{\exp(d(z, z^+)/\tau)}{\exp(d(z, z^+)/\tau) + \sum_{z^-} \exp(d(z, z^-)/\tau)} \right),
$$

(1.23)

where $z$ is the feature vector of patch sampled from $G(i_B)$; $z^+$ and $z^-$ are the feature vectors of patches within $i_A$, respectively, at the same location as $z$ and at different arbitrary locations; and $d(.)$ is some distance (e.g., cosine) between two latent vectors. The contrastive loss can be computed as

$$
\mathcal{L}(G) = \mathbb{E}_{i_A, i_B} \sum_{j=1}^{J} \sum_{s=1}^{S_l} \mathcal{L}(z, z^+, z^-)
$$

(1.24)

with $J$ number of encoder feature layers and $S_l$ number of sampled patches.

### 1.3.3 Problem Settings for Image-to-Image Translation

This section provides an overview of major I2I translation problems that have been investigated in the field. I2I can be mainly divided into two categories based on training data availability: **supervised** I2I trained with available pairs of source and target images, and **unsupervised** I2I trained with unpaired source and target image sets.

**Supervised I2I**: Image translation network can be learned from existing aligned source and target images. The problem can thus be better constrained by applying supervised losses [Isola et al., 2017; Wang et al., 2018]. Generating images from the corresponding
Two-domain I2I Multi-modal I2I Multi-domain I2I
Supervised I2I Unsupervised I2I
Training Testing Training Testing
,,,,…,…

Figure 1.4: Schematic overview of sample image-to-image translation problems, demonstrated using ultrasound images and label-maps as examples.

Semantic label maps is a type of I2I problem, often studied in the existence of label-image pairs. Most I2I methods adopt Unet [Ronneberger et al., 2015] or ResNet [He et al., 2016] as generator architecture, which are effective in processing the coarse-to-fine information in source images. However, layer normalizations typically used in those architectures tend to wash away semantic information when layer-maps are input. There are a line of works which target specifically to generate images from semantic labels, based on developing label-dependent layer normalization for preserving semantic information [Park et al., 2019; Liu et al., 2019; Zhu et al., 2020; Tan et al., 2020; Tan et al., 2021]. Among them, spatially-adaptive normalization (SPADE) [Park et al., 2019] is the most representative work, which proposes an effective layer normalization modulated by scale and bias learned from semantic label maps. Furthermore, dedicated discriminator designs leveraging semantic information have been shown to be effective in improving the output image synthesis quality, e.g., a segmentation-like discriminator [Sushko et al., 2020] or a classification-based feature learning module [Tang et al., 2020].

Unsupervised I2I: Source and target images are not paired, thus direct paired supervision is not possible. Especially in medical applications access to labelled data is often very limited or sometimes not even possible. Training such translation network is often very challenging, since with unpaired data it is not explicitly apparent to the network what information to preserve and what to translate. Although other uses can be envisioned, the typical goal of I2I in this setting is to preserve the content while translating appearance. Various approaches have been proposed to enforce content preservation, e.g., based on
cycle consistency constraint [Zhu et al., 2017b; Yi et al., 2017], shared latent space [Liu et al., 2017], layer normalization [Kim et al., 2019; Jiang et al., 2020], and contrastive learning [Park et al., 2020; Zheng et al., 2021], which were all discussed in the previous section.

Most unsupervised I2I methods focus on translations with the same representation on both ends, e.g., image-to-image. When source and target images coming with different representations, e.g., label-to-image, we refer this as representation gap, which is difficult to handle given unpaired data.

While the early I2I works [Isola et al., 2017; Zhu et al., 2017b; Liu et al., 2017; Kim et al., 2017] focus on I2I translation between two domains, several multi-modal [Zhu et al., 2017a; Huang et al., 2018], and multi-domain [Choi et al., 2018; Choi et al., 2020; Hui et al., 2018] translation methods have been proposed recently. Finding a deterministic one-to-one mapping in the image translation context is an ambiguous task, since a single input image may correspond to multiple possible outputs in the target domain. To tackle this problem, multi-modal I2I aims to model a conditional distribution that predicts multiple possible images in the target domain given an input image in the source domain. Multi-domain I2I methods take a step further and aim to handle translation of images from multiple domains with different attributes simultaneously using an unified framework. Different I2I problem settings are illustrated schematically in Figure 1.4, exemplified using ultrasound images.

1.4 Thesis Goals

The overarching goal of this thesis is to devise methods to enhance interactive and realistic ultrasound image simulation for sonographer training. Given the above-mentioned challenges, the following goals are identified to that end:

1. Automatic estimation and setting of existing simulation parameters from observed in-vivo images, in order to reduce modelling efforts and to enable more realistic tissue appearance from the simulations. A good example of such parameter is the tissue-specific scatter distribution input to convolution-based models that control the speckle texture of tissues.

2. To reduce the time required for image simulation, e.g., by replacing complex rendering steps with deep learning based image translation from simpler simulated images. Since I2I inference has a constant computational overhead, arbitrarily complex simulations can then be afforded, after being trained during a non-time-critical offline stage.

3. To make simulated images with much closer appearance to real world images, without requiring additional modeling and parameter finetuning overhead, i.e. using I2I techniques trained with real images.

4. To be able to use label-maps as input to such I2I translation such that any ray-based simulation/rendering at real time can be completely avoided.

The the goals above, several deep learning based techniques are developed and evaluated in this thesis, as summarized in the next section.
1.5 Thesis Outline and Contributions

The above-described goals are addressed in the different chapters of this thesis, as outline below. A schematic overview is depicted in Figure 1.5.

In ray-based simulation frameworks, US speckle textures are simulated via the convolution of a point spread function with scatterers representing tissue micro-structure. Such scatterer representations are difficult to parameterize and set, to obtain realistic image appearance from the simulation. Classical approaches formulate scatterer estimation as a deconvolution problem, which requires multiple acquisitions and/or intensive computation. In Chapter 2, a convolutional neural network based approach for probabilistic scatterer estimation is presented. Given a known statistical distribution of scatterers, the mapping between ultrasound image and a map of parameter distribution is learned by training a convolutional neural network, named ScatParam, on synthetic images. The proposed method is evaluated in numerical simulations and with in-vivo images, showing that the synthesized images from scatterer representations estimated with ScatParam closely match the observations with varying acquisition parameters. Such estimated scatterer parameters may also be used as a quantitative ultrasound biomarker. This chapter was published as a peer-reviewed journal article as [Zhang et al., 2020b].

Interactive simulation is required for training sonographers. To achieve sufficient frame rates, image quality as a tradeoff is often compromised, e.g., by disabling/reducing essential but time-consuming simulation features. In Chapter 3, a deep learning based approach is proposed, which leverages a deep generative model for recovering missing image features in low quality images simulated by reduced-computation algorithms. Since both low and high quality images can be simulated using the same simulation pipeline, this problem can be tackled as a paired supervised I2I translation setting. This work was published as a peer-reviewed conference workshop paper [Zhang et al., 2020a].

The above I2I method was extended in Chapter 4 for generating US images directly from label maps, which obviates any complex rendering process in real-time. With the available paired data, this label-to-image translation task is treated here also as a
supervised I2I problem. To assess the matching of US images, a local histogram statistics based error metric is proposed and demonstrated for visualization of local dissimilarities between ultrasound images. This chapter was published as a peer-reviewed journal article as [Zhang et al., 2021].

Although the ray-based US simulation can reproduce the major wave interactions, e.g., reflection, refraction, and scattering effects, the simulated images still look synthetic and distinguishable from clinical images. There are unknown post-processing steps and image artefacts specific to US scanners, which cannot be fully modelled. This work aims to translate simulated image representations to real-looking images, i.e. as an unpaired I2I problem. To reduce the appearance gap between simulated and real clinical images, a contrastive-learning based content-preserving image translation framework, named ConPres, is presented in Chapter 5, which bridges such appearance gap, while maintaining the simulated anatomical layout. This is achieved by leveraging both simulated images with semantic segmentations and unpaired in-vivo ultrasound scans. Qualitative and quantitative comparisons against state-of-the-art unpaired translation methods demonstrate the superiority of the proposed framework. This chapter was published as a peer-reviewed conference paper as [Tomar et al., 2021a].

The above works aim to improve the realism of existing simulated images, and thus simulation processes are still required during inference. To obviate such computationally-intensive operations in real-time, an unpaired label-to-image translation approach leveraging domain-specific simulations, named SimIT, is presented in Chapter 6. This method enables realistic image generation directly from simulated label maps, by learning from their unpaired sets, i.e. cast as an unpaired label-to-image translation problem. In addition, this approach affords an image segmenter as a by-product, which provides approximate segmentation of real images, without seeing any annotated real images. The proposed method is further evaluated on the laparoscopy training and gaming simulation, showing its applicability on various simulated scenes. This work is currently being reviewed for publication as a journal article.

This thesis is concluded in Chapter 7 with a brief summary of the proposed methods and contributions, and a discussion on potential future research directions.
Deep Network for Scatterer Distribution Estimation for Ultrasound Image Simulation

Simulation-based ultrasound training can be an essential educational tool. Realistic ultrasound image appearance with typical speckle texture can be modeled as convolution of a point spread function with point scatterers representing tissue microstructure. Such scatterer distribution, however, is in general not known and its estimation for a given tissue type is fundamentally an ill-posed inverse problem. In this paper, we demonstrate a convolutional neural network approach for probabilistic scatterer estimation from observed ultrasound data. We herein propose to impose a known statistical distribution on scatterers and learn the mapping between ultrasound image and distribution parameter map by training a convolutional neural network on synthetic images. In comparison with several existing approaches, we demonstrate in numerical simulations and with in-vivo images that the synthesized images from scatterer representations estimated with our approach closely match the observations with varying acquisition parameters such as compression and rotation of the imaged domain.

2.1 Introduction

Ultrasound (US) is a low-cost, non-invasive and portable imaging modality, and it has been widely used in the clinical routine, especially for obstetrics and gynaecology as it does not involve ionizing radiation. However, ultrasound scanners produce images suffering from limited spatial resolution, signal-to-noise ratio and tissue contrast, which make the interpretation of images very difficult. Clinical ultrasound practice relies exclusively on sonographer expertise in navigating a hand-held probe and visual inspection of the acquired images; and an extensive training is required for being able to conduct such clinical examinations. The current education is mostly based on training on real patients, under the guidance of expert specialists. This form of education is inefficient due to time

This chapter was published as a journal article: Lin Zhang, Valery Vishnevskiy, Orcun Goksel, “Deep network for scatterer distribution estimation for ultrasound image simulation”, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, Vol. 67, 2020, doi: 10.1109/TUFFC.2020.3018424
consumption of both specialists and patients and involve difficulties in finding volunteers especially when rare pathologies are concerned. Indeed, after completing the one year education students will only have seen up to 80% of possible pathologies [Köhn et al., 2004]. Learning from real ultrasound images offline for training is an alternative option that does not require volunteers. However, this approach does not allow interactive training and is highly restricted to the available clinical images. Computer-assisted ultrasound simulation can aid training, especially important for rare pathologies, which are still vital to identify. New scenarios of any given case can be simulated with transformations on pre-collected data, and different imaging parameters and conditions can be simulated from the same data. Furthermore, computerized simulations would allow for training in a real-time virtual-reality environment with complex anatomical scenarios and various pathologies, from which the medical students can obtain in-depth knowledge of possible clinical scenarios during training.

Wave propagation, convolution and ray-based methods are the three major techniques for US simulation. Wave-based approaches [Verweij et al., 2014; Treeby et al., 2012] model ultrasound propagation in tissue by solving complex acoustic wave equations. Therefore, these approaches are relatively slow and not suitable for real-time applications. Convolution-based [Jensen, 1996] methods approximate the ultrasound interactions as a spatial impulse response (point-spread function, PSF), which is then convolved with a representation of sub-wavelength tissue structures, known as scatterers, under the assumption that the acoustic field is linear. This can realistically reproduce typical ultrasound noisy texture, known as speckle, caused by the constructive and destructive inference of echoes scattered by countless tissue scatterers. On the one hand, speckle can be seen as noise that degrades tissue contrast [Tay et al., 2010], but on the other hand, it can help distinguish tissues and identify pathologies [Alessandrini et al., 2011a]. Restoration of this granular pattern is not only important for visual realism in a simulation scenario, but also for preserving structural and diagnostic information about the tissue. Ray-based methods [Burger et al., 2012; Mattausch et al., 2018] simulate the propagation of ultrasonic wavefront as rays using computed graphics techniques, which allows to simulate interactions such as refractions and reflections, while simulating speckle using a PSF convolution with a texture representing scatterers. Using stochastic and sophisticated interaction models with Monte-Carlo ray sampling, this was shown in [Mattausch et al., 2018] to lead to impressively realistic US images even at real-time framerates.

There are recent works that use deep learning, and in particular generative adversarial networks, for ultrasound image synthesis [Hu et al., 2017; Tom et al., 2018]. Nevertheless, these works aim to synthesize isolated individual images, without any intermediate (physical space) representation or parametrization, they are not designed for simulating images with speckle motion coherent with underlying physical motion, and therefore may not be applicable in a real-time, interactive simulation framework where temporal continuity of simulated frames is utmost important. In other words, for an infinitesimal movement of the probe to one side should result in the image moving infinitesimally to the other side. If every image is generated somewhat independently, this behaviour and the visual fluidity cannot be guaranteed. Whereas, using a fixed scatterer representation, the above condition can be satisfied. Such tissue representations can also allow scene editing operations such as copying and adding anatomy in the scatterer domain avoids any image artifacts, as shown in [Mattausch et al., 2017].
The problem with using scatterers in simulations is that such tissue representations are not known a priori. Assuming it can be modeled as a PSF convolution, finding a representation from observed US data can then be posed as a blind deconvolution problem. There has been several approaches and approximations to this problem, including inverse problem based solutions [Chen et al., 2015], variational methods [Zhao et al., 2016; Alessandrini et al., 2011b] and filtering techniques [Michailovich et al., 2007; Taxt, 1995]. Different from our motivation of ultrasound simulation, however, the above methods aim for image denoising / restoration to achieve higher contrast or quality for better diagnostic information. An iterative solution to the regularized deconvolution problem was proposed in [Mattausch et al., 2017] by solving the inverse problem jointly for multiple acquisitions of the same tissue, which are obtained efficiently with electronic beam steering. This allows to overconstrain the problem with observations from multiple PSFs. Discrete scatterer reconstructions were performed on a fine Cartesian grid in order to approximate sub-wavelength particles, where sparsity was enforced by using an \( \ell_1 \)-norm regularization. However, solving such inverse problem is computationally and memory-wise only possible for small patches, which were tiled in [Mattausch et al., 2017] over the imaging field, and still required hours to days to solve for a single image. Accordingly, such a method is not efficient and scalable for scans of many images or to apply in 3D.

Scatterers can alternatively be estimated as statistical distributions of random variables. A pipeline for generating synthetic echocardiographic ultrasound sequences was presented in [Alessandrini et al., 2015]. This was later extended to simulate vendor specific US images for speckle tracking algorithms [Alessandrini et al., 2017]. For this purpose, a scatterer space was populated with randomly sampled 3D cloud points, whose amplitudes were assigned according to the template B-Mode images after the compensation for the log-compression. Simulated speckle statistics were reported to be in good agreement with the known fitting distributions. This approach however does not take the point spread function, and therefore the constructive and destructive interference between scatterers into account, and sampling from B-mode (similarly from RF envelope) assumes scatterers all contributing non-negatively, which is not the case for modulated RF nature of typical US PSF. A simple Gaussian-parametrized model was fit to the inverse-problem reconstructed scatterers in [Mattausch et al., 2015]. This method was demonstrated for homogeneous tissues that other instances of the same tissue can be obtained with the found model for simulating new images, which were however reported as lack of visual variety of real tissues in [Mattausch et al., 2018].

For a given statistical model for scatterer distribution, we propose herein to estimate distribution parameter maps directly from observed ultrasound images. This can be used to instantiate new scatterer maps that would reproduce the original images when input to a convolution-based simulation. As the estimated parameter maps would represent a physical tissue space, they could be used to simulate new images faithfully with varying imaging conditions, viewing directions and other imaging parameter variations. Due to the power of deep neural networks in learning patterns of visual inputs, we propose to learn the mapping between simulated images and parameter maps by training a convolutional neural network.
2.2 Background

2.2.1 Forward Problem of Ultrasound Simulation

Based on the first order Born approximation (weak scattering) for soft tissues [Jensen et al., 1993], the interaction between ultrasound field and tissue scatterers can be formulated as a 2D convolution model in discrete domain:

$$I[l, a] = g[l, a] * h[l, a] + \gamma[l, a],$$

(2.1)

with radio-frequency (RF) US image intensity $I[l, a]$, scatterer intensity $g[l, a]$, spatial variant point spread function (PSF) $h[l, a]$ and noise term $\gamma[l, a]$. $[l, a]$ are the lateral and axial coordinates with respect to the probe origin, elevational thickness is ignored here. Ultrasound point spread function can be approximated with a two dimensional Gaussian kernel modulated by a cosine function in the axial direction [Burger et al., 2012], i.e.:

$$h[l, a] = e^{-\frac{l^2}{\sigma_l^2} - \frac{a^2}{\sigma_a^2}} \cos(2\pi f_c a),$$

(2.2)

where $f_c$ is the transducer center frequency, $\sigma_l$ and $\sigma_a$ determines the Gaussian shape along the lateral and axial direction. Due to nonuniform focusing and aperture, the PSF in US imaging is often assumed to be spatial-variant mainly along the axial direction [Nagy et al., 1998; Mattausch et al., 2017; Alessandrini et al., 2011a]. Some recent works model PSF as continuously non-stationary blurring, such as based on semigroup theory [Michailovich, 2017] and with the diffraction effects during wave propagation [Besson et al., 2019]. For deconvolution tasks, PSF has often been assumed to be patchwise invariant, an approach we also adopt in this paper.

2.2.2 Inverse Problem of Scatterer Reconstruction

Eq. (2.1) can be equivalently written in a matrix-vector form as $A x + n = b$ with the convolutional matrix $A \in \mathbb{R}^{M \times N}$ associated with PSF, a vector of scatterer amplitudes $x \in \mathbb{R}^N$, RF image intensities $b \in \mathbb{R}^M$ and the acquisition noise term $n \in \mathbb{R}^M$. Assuming that $N = M$ and imposing no constraints on $x$, a solution of this system of linear equations is referred to as the tissue reflectivity function (TRF) [Taxt, 1995; Zhao et al., 2016]. To mitigate ill-posed nature of this deconvolution problem, regularization is typically imposed to introduce a prior knowledge about the solution. Wiener filter [Hundt et al., 1980] is a common choice for efficient image deconvolution, which solves the inverse problem based on $\ell_2$-norm regularization of the solution magnitude. Other regularizations have been also widely explored, such as a $\ell_1$ or $\ell_p$-norm based on the assumption of Laplacian [Michailovich et al., 2007] and generalized Gaussian distribution [Alessandrini et al., 2011a] for the TRF $x$. Several computationally efficient methods have been proposed for TRF deconvolution with sophisticated forward models, such as axially varying kernels [Florea et al., 2018], physical model accounting for diffraction effects [Besson et al., 2019]. However, such TRF representation is difficult to attribute to a physical quantity and without any constraints on $x$, the Wiener filter solution may overfit to the observation $b$, e.g. a slight change in acquisition parameters may yield largely different TRF estimates.
2.3 Methods

For US simulation, we aim to find a tissue representation from observed image, which can be used to simulate the same tissue with varying imaging conditions. Rather than estimating a deterministic scatterer locations and amplitudes by solving a large-scale inverse problem, we propose to impose a statistical model on the scatterer distribution and infer corresponding parameters from the observation. We learn the mapping from a single US image to its parameter map in a supervised manner. The required paired data are generated by simulation, since no ground truth of tissue scatterer is available. An overview of our proposed pipeline, referred as ScatParam, can be seen in Fig. 2.1(a).

2.3.1 Statistical Model of Scatterer Distribution

We assume that each tissue type can be parametrized by a model with three parameters \((\rho_s, \mu_s, \sigma_s)\), the parameter \(\rho_s \in [0, 1]\) for scatterer density, the mean \(\mu_s\) and standard deviation \(\sigma_s\) for scatterer amplitude modeled to be normally distributed. Scatterer maps are sampled as follows: for each pixel, a Bernoulli distributed random variable is sampled, where the pixel takes the value one with probability \(\rho_s\) and zero with \(1 - \rho_s\). For non-zero...
pixels, their amplitudes are sampled from a Gaussian distribution $\mathcal{N}(\mu_s, \sigma_s)$. Tissue scattering strength is controlled by the mean, while $\sigma_s$ models random fluctuations around that mean. Since we only consider fully developed speckles, we set $\rho_s$ to a fixed value satisfying the Rayleigh criterion; herein set to a minimum of 100 scatterers per mm$^2$ for a fully-developed speckle pattern [Oosterveld et al., 1985].

The Rayleigh criterion and distribution statistics for estimating interference with isolated scatterers [Oosterveld et al., 1985] has only been studied for scatterers distributed randomly at continuum spatial locations. Typical convolution based simulation packages, such as Field II [Jensen, 1996] and SIMUS [Shahriari et al., 2018], accordingly use scatterer representations with floating-point locations in continuum domain. Reconstructing these on a discrete map therefore necessitate a sufficiently high grid resolution to approximate the continuum. Following [Mattausch et al., 2017] we choose to use an isotropic grid spacing (i.e. Cartesian grid) with the native axial resolution of the raw RF data, which inherently also satisfies the Nyquist criterion for lateral sampling. For instance, for a sampling frequency of 40 MHz and speed-of-sound of 1540 m/s, we use a scatterer map with a resolution of roughly 20 $\mu$m and with 5% of the pixels populated with scatterers, resulting in a scatterer density of 130 per mm$^2$.

### 2.3.2 Training Set Generation

Since scatterers are an abstract tissue representation, no point-wise ground truth exists, thus we create network training data by means of simulation. For parameter map generation, we use random synthetic shapes by overlapping irregular geometric shapes, with a procedure similar to [Vishnevskiy et al., 2019], where random coarse gray-scale patterns are interpolated at a finer resolution and finally thresholding them to create random shapes, as exemplified in Fig. 2.2(a). These aim to represent a rich variety of potential tissue structures without assuming particular anatomical priors, both to be invariant to any anatomical assumptions and region-of-interest as well as to allow the network for better generalization. We assume uniform distribution for Gaussian mean: $\mu_s \sim U(0, 1)$ and a fixed Gaussian standard deviation $\sigma_s = 0.05$. We assign one sampled mean value to each region, assuming the scatterers in each tissue region following the same distribution. The scatterer maps are sampled according to the procedure described in Section 2.3(A). RF images are generated by convolving a sampled scatterer map with a PSF. We assume spatially invariant PSFs for image patches, allowing very fast US image generation. PSFs are sampled from the analytic expression in Eq. (2.2), with the transducer center frequency $f_c = 6$ MHz, the sampling frequency $f_s = 40$ MHz and normalized by its $\ell_2$ norm. The vertical and axial spreading are uniformly sampled from $\sigma^2_l \in [0.2, 1]$ mm$^2$ and $\sigma^2_a \in [0.02, 0.05]$ mm$^2$. The images are then corrupted by additive Gaussian noise. The noise level is uniformly sampled in the interval $[2, 20]$% of the average signal value. The envelope images are used as the input to the neural network, which are taken as the absolute value of Hilbert transform of RF images, illustrated in Fig. 2.2(b). For training, 4000 parameter maps of $64 \times 128$ pixels with the random geometric shapes above were first generated offline. Then on-the-fly during each training batch, random scatterer maps were spatially sampled from a random subset of the parameter maps, following which PSF convolution and envelope detection (Hilbert transform) were also carried out on-the-fly for each training image.
2.3 METHODS

2.3.3 Network Architecture and Training

The general network architecture for ScatParam is illustrated in Fig. 2.1(b). An encoder-decoder network is used to extract features from the input US image and estimate its corresponding parameter map, used for scatterer map sampling. The model comprises an encoder and a decoder part along the axial direction, with skip connections between the corresponding layers. This design choice is due to very low lateral resolution in US imaging being 10 – 20 times lower than the axial. Parameter maps are estimated at a coarser axial resolution than input envelope images, assuming spatially smooth tissue content. This facilitates a more efficient utilization of network weights and hence inference power. Our preliminary experiments with equally high resolution in encoder and decoder did not indicate results substantially superior to our presented architecture. We use strided convolution to perform layer pooling and upsampling. Exponential linear unit activation [Clevert et al., 2015] is used at each layer except the output layer, which is linear. The network is trained using Adam optimizer [Kingma et al., 2014] with a learning rate of $10^{-4}$, minimizing the $\ell_1$-norm based loss function between the true $x$ and estimated parameter maps $\hat{x}$ given the acquisition $y$ as the input:

$$L(\Theta) = \mathbb{E}\|x - \hat{x}(y; \Theta)\|_1,$$  \hspace{1cm} (2.4)

with the network parameters $\Theta$, the empirical average $\mathbb{E}$ over the training sampling procedure. The batch size is set to 16. The network is trained for 20000 iterations.

The proposed pipeline shown in Fig. 2.1 is summarized as follows:

1. Data generation: simulation of synthetic ultrasound images involves scatterer map sampling and convolution with point spread functions;
2. Offline training: a convolutional neural network is trained with the simulated paired data for parameter map estimation;
3. Scatterer distribution estimation and sampling: for each observation, a scatterer map is sampled from the estimated distribution parameter map for synthesizing new images;

Figure 2.2: Generation of training images: (a) scatterer parameter maps, (b) envelope images simulated with scatterers sampled from the corresponding parameter maps, and (c) the corresponding parameter maps estimated by our method.
4. Convolution-based simulation: the sampled scatterer map is fed into the convolution-based simulator, which generates images with desired imaging parameters in real time.

2.4 Experiments and Results

We study our proposed method comparatively to its alternatives with experiments conducted on numerical simulations of synthetic phantoms, as well as on actual data acquired from a gelatin phantom and in vivo tissue. Given the background above, we consider three alternatives to compare our method against:

- Sampling scatterers from envelope image (SampleEnv): This is an adaptation of the method proposed in [Alessandrini et al., 2015] for sampling continuous scatterers from log-uncompressed B-mode images. We herein adapt this to sample on discrete scatterer maps and, for better accuracy, we use the original envelope images instead, as we have access to them.

- Tissue reflectivity function (TRF): This is the simple Wiener filter estimation to the deconvolution inverse problem [Jensen et al., 1994; Taxt, 1995]. We use a spatially constant filter kernel, as the PSF computed or estimated at the center of the imaged field of view.

- Iterative scatterer reconstruction (ScatRec) [Mattausch et al., 2017]: This is an inverse problem based approach is referred here as ScatRec [Mattausch et al., 2017], which reconstructs scatterer map based on a single observation.

- Deep Learning based estimation of parametric scatterer maps (ScatParam): This is our proposed method trained on simulated images, as detailed in the previous section.

For deconvolution based methods, TRF and ScatRec, in simulated data we used the known PSF from the simulations and for acquired data, we used a cepstrum-domain PSF estimation method described in [Mattausch et al., 2016], followed by least square fitting to the known parametric form in Eq. (2.2) to project them on our PSF model manifold, which our trained network is better conditioned on.

Since different scatterer representations cannot be compared directly and no ground truth is available, we evaluate the performance of scatterer estimation on the envelope images simulated from the estimated scatterer maps. For re-synthesizing images, we used the same forward simulation for all the methods, namely a discrete image-space convolution of the scatterers estimated by any particular method with the same depth-dependent PSFs estimated for deconvolution based methods. The convolution was implemented in Matlab to operate separately for each PSF on rows of image pixels.

2.4.1 1D Example

In contrast to US deconvolution (TRF and the regularized version ScatRec) that aims to reproduce the RF and hence speckle image precisely on a pixel level, the goal of
2.4 Experiments and Results

Figure 2.3: 1D signal example with simulated RF and envelope signals by TRF (left) and envelope signals by SampleEnv and ScatParam (right). Black star depicts the scatterer locations.

SampleEnv and ScatParam are rather to generate realistic speckle textures that match the speckle distribution and tissue contrast of a real image. We herein illustrate their differences using a simplified 1D example depicted in Fig. 2.3. We first simulated a ground truth 1D ultrasound RF signal (RF GT) of 40 MHz sampling, by convolving a Gaussian-modulated sinusoidal PSF for a 5 MHz pulse with randomly placed scatterer locations within \([20, 180]\) \(\mu\)s (Scats GT). Scatterer amplitudes were randomly sampled from a Gaussian distribution with a mean of 1 and standard deviation of 0.05. For comparison, for each estimated representation we resynthesized the RF signal by convolving with the known PSF and then plot the envelope of these in comparison to the ground truth envelope (Env GT).

For TRF, we deconvolved RF GT using Wiener filtering with the known PSF, which yields almost a perfect match to the original data for this ideal, no-noise scenario, cf. RF TRF and Env TRF in Fig. 2.3(left). For SampleEnv and ScatParam, we resampled random scatterer locations shown in Fig. 2.3(right), the amplitudes of which were set for SampleEnv using the envelope intensities and for ScatParam from the original scatterer distribution, with the assumption that such distribution can be correctly estimated by our method. ScatParam and SampleEnv both do not reproduce the RF data exactly, but generate an envelope intensity distribution similar to the original envelope.

2.4.2 Evaluation Metrics

Several evaluation metrics are utilized to assess the simulation performance and compare our method ScatParam with SampleEnv, TRF and ScatRec. Three image-based metrics, mean image intensity (I), signal-to-noise ratio (SNR), contrast-to-noise ratio (CNR) and one histogram based metric, Kullback-Leibler (KL) divergence, are used to calculate the mismatch between ground truth and simulated envelope images.

**Mean Image Intensity (I)** can capture any global intensity shift in the simulated images, which can be caused by systematic biases in constructive or destructive interference when the scatterers are not distributed truly stochastically or in a view-dependent way. For
instance, if the scatterers estimated from one direction align in a structured way, when
the object is imaged from an oblique direction (e.g., with the half-wavelength projected
on the rotation angle aliased with the structure), these scatterers may then interfere with
each other mostly destructively, creating an artificial intensity drop in the image. The
change in mean image intensity is calculated as follows
\[
\Delta I = \left| \frac{I_t - I_s}{I_t} \right|, \quad \text{with} \quad I = \frac{1}{N} \sum_{j=1}^{N} s_j,
\]
where \(s_j\) denotes the image intensity value at the \(j\)-th pixel, and \(N\) is the number of pixels.
\(I_t\) and \(I_s\) are the mean image intensities of ground truth and simulated image, respectively.

**Signal-to-noise Ratio (SNR)** measures the global statistics of signal-and-noise ratio. For
the simulation purpose, we aim at reproducing the images at the same SNR level as the
ground truth. Any mismatch in SNR is then quantified as follows
\[
\Delta \text{SNR} = \left| \frac{\text{SNR}_t - \text{SNR}_s}{\text{SNR}_t} \right|, \quad \text{with} \quad \text{SNR} = \frac{\mu}{\sigma},
\]
where \(\mu\) and \(\sigma\) denote the mean and standard deviation of envelope image intensities.

**Contrast-to-noise ratio (CNR) mismatch** is defined similarly as
\[
\Delta \text{CNR} = \left| \frac{\text{CNR}_t - \text{CNR}_s}{\text{CNR}_t} \right|, \quad \text{with} \quad \text{CNR} = \frac{|\mu_{s1} - \mu_{s2}|}{\sigma_{s1} + \sigma_{s2}},
\]
where \(\mu_{s1}, \mu_{s2}, \sigma_{s1}, \text{ and } \sigma_{s2}\) denote the means and standard deviations of the image
intensities within two contrasting regions. This metric is clinically relevant, as incorrect
tissue contrast in the simulated images may lead to the learning of false diagnostic cues
during medical training.

**Kullback-Leibler (KL) divergence** compares the statistics between the histograms of two
images as follows:
\[
\text{KL}(h_s || h_t) = \sum_{l=1}^{D} h_s[l] \log \left( \frac{h_s[l]}{h_t[l]} \right),
\]
where \(h_t\) and \(h_s\) are the normalized histograms of the true and simulated images re-
spectively. The number of histogram bins \(D\) is set herein to 50. Histogram statistics are
widely explored for tissue characterization [Shankar et al., 1993; Tsui et al., 2008]. Hence,
a large discrepancy in the histograms could indicate a mismatch in the speckle pattern
appearance. Since computing histograms over the whole image could miss local speckle
texture information, we calculate a KL divergence metric locally within patches (herein
non-overlapping patches of \(3 \times 3 \text{mm}^2\) corresponding to \(10 \lambda\) per dimension) and report
herein the metric mean over all patches.

For calculating \(\Delta \text{SNR}, \Delta \text{CNR}\) and KL divergence, simulated image \((s)\) is normalized (or
brightness equalized) with respect to ground truth image \((t)\) by multiplying a factor \(\frac{\sum_{j} t_j}{\sum_{j} s_j}\)
similarly to [Mattausch et al., 2017], to eliminate effects in these metrics from any global
intensity shift, which is captured separately by \(\Delta I\).
2.4.3 Synthetic Data

This experiment evaluates the invariance of the reconstructed scatterer maps to various US imaging parameters. To that end, we used simulated images from Field II with controllable imaging conditions. A numerical phantom of $15 \times 15 \text{mm}^2$ with a $3 \text{mm}$ circular inclusion was simulated for imaging at $6.0 \text{MHz}$ center frequency, with a $128$ element $40 \text{mm}$ linear transducer, a single transmit focus at the center, and dynamic receive focusing. The phantom was placed $15 \text{mm}$ away from the transducer to avoid near field effects.

We evaluate the simulation results for 1) phantom rotation, which emulates imaging a region of interest from different viewing directions; 2) phantom compression, which emulates image plausibility under potential tissue deformation, e.g. induced by probe compression.

**Rotation Experiment**

In this experiment, we evaluate the invariance of the reconstructed scatterer maps to phantom rotation. For this, the box phantom were rotated around the phantom center with varying angles. Fig. 2.4 illustrates the results for the rotated views. The images were cropped to the region of interest. The top row shows the ground truth envelope images simulated by Field II for varying rotation angles ranging from $0^\circ$ to $45^\circ$ with a $15^\circ$ increment. For ScatRec, the image intensity drastically decreases with increasing rotation angle and the phantom almost disappears for $45^\circ$ due to destructive interference, as reported in [Mattausch et al., 2017]. The images simulated by SampleEnv appear dark, whereas the image mean intensity remains similar to the ground truth images for TRF and ScatParam during rotation. It can be well observed that the reconstruction of the proposed ScatParam is robust for different viewing angles.

We evaluate the simulation performance quantitatively for the above experiment by investigating the error for rotations with $1^\circ$ increments, with the results plotted in Fig. 2.5. ScatRec is close to the ground truth near $0^\circ$ given any metric, but it deviates largely from the ground truth at larger rotation angles, which corroborates the visual observations above. The errors for SampleEnv are consistently high irrespective of rotation, indicating not a successful scatterer reconstruction. In general, the error metrics of TRF vary without any pattern, remaining relatively similar across the angles, except a large variance in CNR. This is in agreement with the observation in Fig. 2.4 that the tissue contrast in the $15^\circ$ rotated view of TRF is diminished. The metrics for ScatParam exhibit minor fluctuations and remain overall relatively low and thus superior compared to the other methods.

Tab 2.1 summarizes the mean, median and maximum error across the $1^\circ$ increment results in Fig. 2.5. ScatParam is seen to achieve the lowest error overall, nearly $30\%$ lower than the second best method in $\Delta$CNR and KL divergence, demonstrating the representativeness and robustness of the estimated scatterer map with respect to probe rotation.

**Compression Experiment**

With this, we investigate the invariance of reconstructed scatterer maps with respect to physical deformation. An axial phantom compression was simulated by interpolating the estimated scatterer maps on grids deformed by varying levels of axial strain $e$. Simulated
**Figure 2.4:** Simulated images of the numerical phantom for views rotated by 15°, 30°, and 45°. Envelope images are shown.

**Table 2.1:** Mean (mean), median (med), and maximum (max) errors for rotations between 0° and 45° with 1° increments. Bold number indicates the smallest value per column.

<table>
<thead>
<tr>
<th>Metric</th>
<th>ΔI (%)</th>
<th>ΔSNR (%)</th>
<th>ΔCNR (%)</th>
<th>KL ($\times 10^{-2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>med</td>
<td>max</td>
<td>mean</td>
</tr>
<tr>
<td>SampleEnv</td>
<td>50.1</td>
<td>50.2</td>
<td>50.8</td>
<td>11.9</td>
</tr>
<tr>
<td>TRF</td>
<td>9.1</td>
<td>8.6</td>
<td>18.1</td>
<td>4.0</td>
</tr>
<tr>
<td>ScatRec</td>
<td>59.6</td>
<td>68.0</td>
<td>86.9</td>
<td>3.4</td>
</tr>
<tr>
<td>ScatParam</td>
<td>5.7</td>
<td>7.4</td>
<td>11.0</td>
<td>2.3</td>
</tr>
</tbody>
</table>
Figure 2.5: Performance with respect to rotation angle in terms of normalize mean image intensity difference ($\Delta I$ ($\%$)), signal-to-noise ratio difference ($\Delta SNR$ ($\%$)), contrast-to-noise ratio difference ($\Delta CNR$ ($\%$)) and histogram difference (KL divergence).
Table 2.2: Mean (mean), median (med) and maximum (max) errors across 10% to 50% strain with 1% increments. Bold number indicates the smallest value per column.

<table>
<thead>
<tr>
<th>Metric</th>
<th>ΔI (%)</th>
<th>ΔSNR (%)</th>
<th>ΔCNR (%)</th>
<th>KL ($\times 10^{-2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>med</td>
<td>max</td>
<td>mean</td>
</tr>
<tr>
<td>SampleEnv</td>
<td>38.5</td>
<td>39.4</td>
<td>46.3</td>
<td>12.7</td>
</tr>
<tr>
<td>TRF</td>
<td>24.4</td>
<td>24.1</td>
<td>36.1</td>
<td>8.3</td>
</tr>
<tr>
<td>ScatRec</td>
<td>25.6</td>
<td>27.1</td>
<td>38.3</td>
<td>12.9</td>
</tr>
<tr>
<td>ScatParam</td>
<td>26.0</td>
<td>25.1</td>
<td>41.7</td>
<td>1.8</td>
</tr>
</tbody>
</table>

images with all methods at $e = \{10,30,50\}$% compression are shown in Fig. 2.6. The images simulated by TRF are corrupted by aliasing artifacts. The simulation results of SampleEnv and ScatRec look similar with reduced intensity and degraded speckle pattern for large compression strain. The simulated images by ScatParam are visually closest to the ground truth images in terms of speckle appearance and contrast, but slightly hyperechoic for large compression.

We investigate the simulation errors for compression ranging from $e = 10\%$ to $e = 50\%$ with 1% increments in Fig. 2.7. Similarly to the rotation experiment, the error metrics of ScatRec increase proportionally with increasing compression, whereas our method performs consistently superior. SampleEnv and TRF yield large errors, especially in ΔCNR and KL divergence metrics.

The observations are supported by numerical results shown in Tab. 2.2, reporting the mean, minimum, and maximum errors for the above plots. Our proposed method ScatParam is seen to be superior in terms of ΔSNR, ΔCNR, and KL divergence. It achieves approximately 78% lower error in the mean ΔSNR, 68% in the mean ΔCNR and 28% in the mean KL divergence compared to the second best method. TRF achieves 6% lower error than ScatParam in the mean ΔI here.

2.4.4 Gelatin Phantom

Next we investigate the performance of our method for a real ultrasound scan of a gelatin phantom with corn starch as the scattering medium. A circular inclusion was made by adding twice as high as the concentration of starch in the background. The beamformed RF images were collected by a Fukuda Denshi UF-760AG ultrasound machine with a linear probe FUT-LA385-12P. The results in the top row of Fig. 2.8(a) show an excellent agreement of ScatRec with the ground truth image, as expected from an overconstrained optimization, while the SampleEnv and TRF results exhibiting speckle textures different than the ground truth. Our proposed method ScatParam generates images almost indistinguishable from the observation in terms of speckle texture and tissue contrast. The experiment is further evaluated with quantitative metrics in Tab. 2.3. ScatRec achieves the closest match to the ground truth image, as its scatterer estimation overfits to the observed phantom image. Comparing the remaining three, ScatParam yields CNR, SNR, and KL divergence metrics lower than TRF and SampleEnv, indicating that our method better preserves contrast and intensity distribution.
Figure 2.6: Simulated images of the numerical phantom for axial compressions of 10%, 30%, and 50% strain. Envelope images are shown.

Table 2.3: Quantitative evaluation metrics for the gelatin phantom. Bold number indicates the smallest error per column.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\Delta I$ (%)</th>
<th>$\Delta$SNR (%)</th>
<th>$\Delta$CNR (%)</th>
<th>KL ($\times 10^{-2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SampleEnv</td>
<td>44.0</td>
<td>19.2</td>
<td>13.3</td>
<td>25.5</td>
</tr>
<tr>
<td>TRF</td>
<td>5.0</td>
<td>10.0</td>
<td>24.0</td>
<td>21.2</td>
</tr>
<tr>
<td>ScatRec</td>
<td><strong>1.7</strong></td>
<td><strong>0.6</strong></td>
<td><strong>1.3</strong></td>
<td><strong>2.7</strong></td>
</tr>
<tr>
<td>ScatParam</td>
<td>7.4</td>
<td>0.9</td>
<td>8.1</td>
<td>17.0</td>
</tr>
</tbody>
</table>
Figure 2.7: Performance evolution with increasing axial strain $e$ in terms of difference in image mean intensity ($\Delta I (%)$), signal-to-noise ratio ($\Delta SNR (%)$), contrast-to-noise ratio ($\Delta CNR (%)$) and histogram difference (KL).
2.4 Experiments and Results

Figure 2.8: (a) Envelope images with scatterers estimated from the gelatin phantom, images simulated at 0° (top) and 45° (top). (b) Histogram of the ground truth and rotated views inside the area shown by a red rectangle in one sample image. Quantifying the difference between the ground truth and simulated image histograms using KL divergence indicates to the following errors: 0.101 for SampleEnv, 0.593 for TRF, 0.588 for ScatRec, and 0.032 for ScatParam.

The bottom row in Fig. 2.8(a) shows the simulated 45° rotated views of the phantom. For an isotropic scattering phantom, the mean echo and speckle texture statistics would be invariant to the viewing direction and hence, after rotation, a speckle appearance similar to the initial view is expected. Such coherence pattern may naturally not stay pixel-wise constant after, e.g., rotation, therefore no pixel-wise error metric was employed. Similarly to the rotation experiment of Field II, the image simulated by ScatRec appears hypoechoic for 45° angle, while the speckle pattern of TRF is severely distorted.

The ground truth rotated views are not obtained here. Nevertheless, we evaluate the results by comparing the histogram of the observed 0° envelope image to the histograms of the rotated envelope images (after brightness equalization) for the homogeneous region inside the red rectangle depicted in Fig. 2.8(a). Here Rayleigh statistics serve as an important criterion for assessing ultrasound speckle texture, as the envelope intensity should follow a Rayleigh distribution for fully developed speckles [Goodman, 1975]. The histogram of SampleEnv closely follows the observed ground truth histogram, i.e. an ideal Rayleigh distribution, consistent with the observation reported in [Alessandrini et al., 2015]. Nevertheless, our approach ScatParam is seen to have an even better agreement with the observed pattern; indeed, with an over 3-folds lower KL divergence score as listed in the figure caption.

2.4.5 In Vivo Experiment

For this purpose, beamformed RF data from an in-vivo scan of the liver was collected. Scatterer representations were estimated for the imaged region using all four methods presented. Fig. 2.9 depicts the results simulated in the acquisition configuration of 0° and simulating a rotation of 45° given a fixed dynamic range computed at min and max of the original 0° view, considering the fixed imaging settings during real-life probe and
scene manipulations. To highlight details imperceivable due to any intensity reduction, the last row in Fig. 2.9 shows the same simulated 45° views, dynamic range adjusted individually. At 0°, all methods perform somewhat similarly, with SampleEnv slightly hypoechoic, TRF misrepresenting speckle texture and ScatRec reconstructing an exact replica, as expected. Noticeably, given the dynamic compression for B-mode images presented for the in-vivo experiment, the loss of brightness in the image generated by SampleEnv is less prominent compared to the previous experiments showing envelope images. At 45°, however, one can observe a global intensity drop in SampleEnv and ScatRec and intensity increase in TRF with the distorted speckle patterns – indicating that these three all did not estimate robust scatterer representations. After compensating for the global intensity shift in the last figure row, the speckle textures for TRF and ScatRec still look unrealistic, whereas SampleEnv do not faithfully preserve structural details, e.g. at the bottom image region, with increased echogenicity in the upper region. In contrast,
2.5 Discussion

In this work, we demonstrated a learning based approach for probabilistic scatterer estimation in the context of realistic ultrasound image simulation. The proposed framework ScatParam involves sparse scatterer model with Gaussian distribution for scatterer amplitude and estimation of the Gaussian mean directly from US envelope images by neural network. Similar isotropic scatterer distributions were used in several earlier works, e.g. [Jensen, 1996; Alessandrini et al., 2011a; Burger et al., 2012; Mattausch et al., 2017; Alessandrini et al., 2015; Bamber et al., 1980]. In our preliminary experiments for estimating parameter maps of both mean and standard deviation, we have found that different such combinations may generate similar image outputs, making such double parameter estimation ill-posed. Therefore, in this work we fixed the standard deviation and estimated only the mean value. In addition, we assumed a fixed density for our model, since it is shown that in case of fully developed speckles the tissue characteristics are not affected by density [Thijssen et al., 1990]. Estimating scatterer density in addition to
the amplitude mean could allow us to distinguish between partially and fully developed speckles. Nevertheless, we observed comparable performance with only mean estimation for both numerical simulations and in-vivo experiments presented herein.

For complex in-vivo scatterer distributions more sophisticated models may be required. For example, for muscle fibers, anisotropic parametrizations such as with tensor, wavelet, or frequency-domain representations may be more suitable. Indeed, instead of hand-crafted parametric models, (arbitrary) distributions could potentially be parametrized using a neural network to be inferred from observed images.

In the paper, we have conducted several experiments evaluating the invariance of tissue properties with geometric transformations, which is of great importance for ensuring plausible simulation of the same tissue content under different imaging conditions. Rotation and axial compression are chosen, since they well represent the clinical examinations such as transducer in-plane tilting and compression. Furthermore, these experimental scenarios model potential variation in speckle appearance from directional changes (isotropy) and concentration changes (axial strain) in scatterer configurations.

In comparison to other deconvolution algorithms, our method does not require PSF estimation in advance. It was reported in [Mattausch et al., 2017] that accurate PSF estimation is needed as input to an inverse problem based method, analogously important for the Wiener filter and other deconvolution algorithms. Our trained CNN is able to capture the PSF information in the form of different speckle textures in input images, hence estimated parameter maps would be independent of imaging system. One can produce visually plausible images without considering PSF as in the method SampleEnv. However, the simulated speckle statistics with SampleEnv herein are not fully in agreement with the observations, since by dismissing the modulated nature of PSF and thereby any destructive interference, SampleEnv cannot fully model the interference between scatterers. Setting scatterer amplitudes directly using envelope intensities causes the incorrect translation of speckle variations into scatterer maps, i.e. higher amplitude scatterers lumped around the peaks of speckles. This leads to brighter hyperechoic and darker hypoechoic regions after PSF convolution. Such granular appearance is less visible after dynamic range compression in B-mode images, e.g. in Fig. 2.9, whereas these and resulting overall intensity reduction become apparent in the envelope images shown for the rest of the experimental results. Note that we present envelope images, as they allow easier interpretation of speckle patterns and imaging physics, where effects of methodological choices on the results are not masked by any graymap transform.

We herein chose Wiener filtering as a basic deconvolution baseline for TRF, with low computation and memory requirements. The results with artifacts and reduced contrast of TRF indicate that the use of the same low resolution of the input RF image in the output scatterers hinder interpolation in the scatterer domain after transformations, such as compressions. Although there are more recent forms of TRF, e.g. [Florea et al., 2018; Besson et al., 2019], ScatRec [Mattausch et al., 2017] was chosen herein as a state-of-the-art baseline performing a sophisticated deconvolution approach as an optimization of an inverse-problem definition; with a suitable model, i.e. norms and regularizers; with depth-dependent PSF; non-negative scatterer constraint; and with a higher scatterer resolution than the RF image domain. Regardless of baselines, our results indicate that our learning based solution ScatParam as a fast-implementable network solution
performs satisfactorily for image simulation and without requiring complex convolution modeling, PSF estimation, iterative optimization, or any other complex processing steps. Note that using only Gaussian noise model for the training samples, ScatParam is able to successfully estimate scatterer parametrizations for no-noise numerical experiments as well as unknown-noise phantom and in-vivo examples, potentially indicating the robustness of the proposed parametrization and the respectively trained NN to an assumed noise model.

Underconstrained inverse problem based approaches can be improved by using multiple measurements of the same tissue with different imaging parameters. A successful example was illustrated in [Mattausch et al., 2017], where ScatRec with multiple observations from beam steering was shown to be more robust to viewing angle changes. However, the aligned beam-steered images are in general not available from clinical scanners, let alone raw RF data. Our method takes envelope images as input and can thus accept clinical B-Mode images with slight modification of network training. For a fair comparison, we herein used the single view version of inverse problem, i.e. ScatRec1 in [Mattausch et al., 2017].

In this work any ultrasound image appearance is attributed solely to isotropic scattering, where attenuation variations and coherent reflections are not considered. For instance, all the methods compared in Fig. 2.8(a) attribute the slight attenuation in the original image behind the gelatin inclusion (after a depth of 40 mm) to some form of lower scattering amplitude, thereby resulting in the attenuation not correctly reproduced in the rotated images, e.g. extending diagonally rather than vertically. This demonstrates the need to take directional attenuation and reflections into account during any scatterer estimation process, which should be a focus of future studies. Furthermore, any potential reflections at anatomical boundaries, i.e. between supra-wavelength structures, would also be attributed to result from the reconstructed scatterers, which may then incorrectly reproduce the tissue from different viewing angles. In all our numerical examples, any such reflections were thus avoided. The gelatin phantom was made with isotropic scatterers and for the in-vivo example the liver was chosen for its relatively homogeneous speckle appearance. Additionally we conducted evaluations within small regions-of-interest selected in line with our assumptions. Accordingly, with the experiments we evaluated the methods in terms of speckle appearance, as opposed to realism of simulated attenuation or reflection effects.

Since the network is trained with isotropically distributed scatterers, any estimation for non-isotropic regions may be sub-optimal or likely be applicable only when viewed from a similar direction. Nevertheless, it may be possible to separate the direction-dependent image content prior to capturing only scattering effects. One can then simulate such directional wave interactions at a later time using ray tracing techniques [Burger et al., 2012; Mattausch et al., 2018]. For instance, reflection boundaries could be detected and removed using a simple phase symmetry (PS) algorithm [Hacihaliloglu et al., 2009], which is designed to estimate directional reflections at tissue boundaries such as bone surfaces. In a way similar to estimating and compensating for reflections, acoustic attenuation can indeed also be reconstructed a-priori [Rau et al., 2019] in order to spatially normalize incident acoustic energy to decouple its effect from our reconstructed scatterer amplitudes.
Our network training takes approximately 6 hours on Nvidia Titan XP GPU. Once trained, our network can estimate scatterer maps in milliseconds at inference time, for arbitrary input image size (as being a fully convolutional network architecture). In contrast, the inverse problem based approach takes approximately 2 hours for one image from a single view. For multiple (beam-steered) observations, the computation time would increase exponentially, quickly making this method infeasible for large images and 3D volumes.

2.6 Conclusion

We have demonstrated a learning-based technique to efficiently estimate the distribution of tissue scatterer representation, which can then be fed directly into convolution- or ray-tracing-based simulation techniques [Burger et al., 2012; Mattausch et al., 2018] to simulate realistic images for sonographer training. The proposed network is trained only with synthetic images generated with random shapes and spatial invariant convolution. In comparison to the state-of-the-art methods, we demonstrate with numerical simulations the proposed estimation pipeline being robust for simulating images at different viewing angles and tissue deformations. The method is further evaluated on a tissue-mimicking gelatin phantom and an in-vivo liver image, demonstrating the generalization ability of our network to real ultrasound scans.
Deep Image Translation for Enhancing Simulated Ultrasound Images

Ultrasound simulation based on ray tracing enables the synthesis of highly realistic images. It can provide an interactive environment for training sonographers as an educational tool. However, due to high computational demand, there is a trade-off between image quality and interactivity, potentially leading to sub-optimal results at interactive rates. In this work we introduce a deep learning approach based on adversarial training that mitigates this trade-off by improving the quality of simulated images with constant computation time. An image-to-image translation framework is utilized to translate low quality images into high quality versions. To incorporate anatomical information potentially lost in low quality images, we additionally provide segmentation maps to image translation. Furthermore, we propose to leverage information from acoustic attenuation maps to better preserve acoustic shadows and directional artifacts, an invaluable feature for ultrasound image interpretation. The proposed method yields an improvement of 7.2% in Fréchet Inception Distance and 8.9% in patch-based Kullback-Leibler divergence.

3.1 Introduction

Ultrasound (US) is a low-cost, real-time, and portable diagnostic imaging technique without ionizing radiation, hence widely used in gynecology and obstetrics. Since its interpretation can be nontrivial due to ultrasound-specific artifacts such as acoustic shadows and tissue-specific speckle texture, sonographer training is crucial. For an education tool, ray tracing can be used for US simulation [Burger et al., 2013; Computing et al., 2015], where US wavefront is represented with rays on the GPU to simulate interaction with tissue layers, whereas speckle patterns are simulated with a convolutional model of tissue speckle noise. With stochastic Monte-Carlo sampling of rays [Mattausch et al., 2018], this can produce realistic looking images. However, interactive computational constraints often necessitate a compromise in image quality, e.g. with limited number of rays or by disabling or reducing essential simulation features.

This chapter was published as a conference workshop article: Lin Zhang, Tiziano Portenier, Christoph Paulus, Orcun Goksel, “Deep image translation for enhancing simulated ultrasound images”, MICCAI Workshop on Advances in Simplifying Medical Ultrasound, doi: 10.1007/978-3-030-60334-2_9
Deep learning has achieved great success in various computer vision and graphics tasks. In particular, generative adversarial networks (GANs) [Goodfellow et al., 2014] have been demonstrated as a powerful tool for image synthesis and translation [Isola et al., 2017; Zhu et al., 2017b]. GANs have been widely adapted for various medical image synthesis tasks, such as image inpainting [Armanious et al., 2019] and cross modality translation in both supervised [Armanious et al., 2020; Nie et al., 2018] and unsupervised [Wolterink et al., 2017; Zhang et al., 2018] settings. In US image synthesis, a two-stage stack GAN was introduced in [Tom et al., 2018] for simulating intravascular US imagery conditioned on tissue echogenicity map. In [Hu et al., 2017], freehand US images are generated conditioned on calibrated physical coordinates. Recently in [Vitale et al., 2019], feasibility of improving the realism of ray-traced US images has been demonstrated using cycleGAN [Zhu et al., 2017b].

In this work we propose a deep learning based approach for improving the quality of simulated US images that are obtained using a ray tracing algorithm, such that computationally simpler (low quality) images can be used to generate higher quality images mimicking a computationally sophisticated simulation that may not be feasible at interactive frame rates. Access to a simulation framework together with comprehensive anatomical models allows us to obtain realistic paired images of differing quality aligned with anatomical models. Therefore, we tackle this problem in an image-to-image translation setting with paired low and high quality images. Our framework leverages conditional GANs [Mirza et al., 2014] to recover image features that are missing in the low quality images. Since low quality images may have missing anatomical structures, which introduces ambiguities in the image translation process, we propose to additionally leverage information that is readily available from the underlying simulation algorithm. For this purpose, we use 2D segmentation map slices at given transducer locations, to provide any anatomical information missing from low quality images. Since major acoustic effects such as shadows are integral along wave path and hence global in nature, they would require large network receptive fields to model. Thus, we further propose to incorporate integral attenuation maps as additional input to the network. Such segmentation and attenuation maps can be easily obtained as by-products of ray-based simulation frameworks [Burger et al., 2013; Mattausch et al., 2018; Computing et al., 2015].

3.2 Materials and Methods

Data Generation. Simulated B-mode US images are generated using a Monte-Carlo ray tracing framework on a custom geometric fetal model for obstetric training [Mattausch et al., 2018]. US wave interactions are simulated using a surface ray tracing model to find the ray segments between tissue boundaries. Tissue properties such as acoustic impedance, attenuation and speed-of-sound are assigned to each tissue type from literature and based on sonographers' visual inspection. Along each extracted ray segment, a ray-marching algorithm is applied on the GPU to emulate US scatterer texture by convolving a locally changing point-spread-function with an underlying tissue scatterer representation generated randomly using Gaussian distributions per tissue type [Mattausch et al., 2015]. Simulated RF data is post-processed with envelope detection, time-gain compensation,
3.2 Materials and Methods

(a) LQ image  (b) HQ image  (c) Segmentation  (d) Attenuation

Figure 3.1: Low quality (a) and high quality (b) simulation outputs, with corresponding segmentation map (c) and integral attenuation map (d).

log compression and scan-conversion into Cartesian coordinates, yielding a gray-scale B-mode image.

US images. For each regularly-sampled key frame of a simulated US fetal exam, paired low and high quality images are generated using two simulation passes: low quality images using one primary ray per US scanline and one elevational layer; and high quality images using 32 primary rays per scanline and three elevational layers [Mattausch et al., 2018]. Other simulation parameters are kept identical for both simulation passes, cf. Table 5.2. Example B-mode images are shown in Fig. 3.1(a-b).

Image Mask. A fixed binary image mask demarcating the imaging region after scan-conversion for the convex probe is also provided as input to the network, in order to constrain the meaningful image translation region and help to save generator capacity.

Segmentation Maps. As additional input for our method, segmentation maps as the cross-section of input triangulated anatomical surfaces are also output by the simulation, corresponding to each low-/high-quality image, cf. Fig. 3.1(c).

Attenuation Maps. A characteristic feature in real US images is the presence of directional artifacts, which is also valuable for the interpretation of images, for instance in diagnosis of pathology. It is therefore important to accurately simulate such artifacts for training purposes. Besides reflection and refraction effects, a major source of directional US artifacts is attenuation, which is caused by a reduction in acoustic intensity along

<table>
<thead>
<tr>
<th>parameter</th>
<th>value</th>
<th>parameter</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangles fetus</td>
<td>400k</td>
<td>Transducer frequency</td>
<td>8 MHz</td>
</tr>
<tr>
<td>Triangles mother</td>
<td>275k</td>
<td>Transducer field-of-view</td>
<td>70°</td>
</tr>
<tr>
<td>Image depth</td>
<td>15.0 cm</td>
<td>Axial samples</td>
<td>3072</td>
</tr>
</tbody>
</table>

Table 3.1: Simulation parameters
the wave travel path due to local tissue effects such as absorption, scattering, and mode conversion. Since such artifacts are not only a function of local tissue properties but an integral function along the viewing direction, we propose to directly provide this integrated information to the translation network, hypothesized to improve the quality of translation.

Acoustic intensity arriving at a depth $z$ can be modeled as $I(z) = I_0 e^{-\mu z}$, where $\mu$ is the attenuation constant at a given imaging frequency and $I_0$ is the initial intensity. Given that the waves travel through different tissue layers with varying attenuation constants $\mu(z)$, the total intensity arriving at a point $z$ can be approximated by

$$I(z, \mu|_{0}) = I_0 \prod_{i=0}^{z} e^{-\mu[i]} = I_0 e^{-\sum_{i=0}^{z} \mu[i]}.$$  \hspace{1cm} (3.1)

To approximate such attenuation effect, we create attenuation integral maps $a = e^{-\sum_{i=0}^{z} \mu[i]}$, accumulated for each image point along the respective ultrasound propagation path. For better dynamic range and to avoid outliers, these maps are normalized by the $98\%$ile of image intensities and then scan-converted into the same Cartesian coordinate frame as the simulated B-mode images. Fig. 3.1(d) shows sample integral attenuation maps.

**Image Translation Network.** Our image-to-image translation framework is based on the pix2pix network proposed in [Isola et al., 2017]. Simulated low and high quality US images are considered as source and target domain, respectively, where a translation network $G$ learns a mapping from the source to the target domain. Specifically, $G$ maps the low quality US image $x$, the binary mask $m$, the segmentation map $s$, and the attenuation integral map $a$ to the high quality US image $y$, i.e.: $G : \{x, m, s, a\} \rightarrow \{y\}$. The discriminator $D$ is trained to distinguish between real and fake high quality images conditioned on the corresponding inputs to the generator. The objective function of the conditional GAN consists of a weighted sum between a GAN loss $L_{GAN}$ and a data fidelity term $L_F$, i.e.,

$$L = L_{GAN}(G, D) + \lambda L_F(G),$$  \hspace{1cm} (3.2)

$$L_{GAN} = E_{\hat{x},y}[\log D(y|\hat{x})] + E_{\hat{x}}[\log(1 - D(G(\hat{x})|\hat{x})]],$$  \hspace{1cm} (3.3)

$$L_F = E_{\hat{x},y}[||y - G(\hat{x})||_1],$$  \hspace{1cm} (3.4)

where $\hat{x} = (x, m, s, a)$. Before computing the losses, the output is element-wise multiplied with the binary mask to restrict the loss to the relevant output regions.

Similarly to [Isola et al., 2017], we use a deterministic $G$ parametrized using a 8-layer Unet with skip connections and $D$ using a 4-layer convolutional network, i.e. a patchGAN discriminator. Instance normalization is applied before nonlinear activation. The full field-of-view B-mode images from the simulation are of size $1000 \times 1386$ pixels. Applying pix2pix directly at such high resolution may lead to unsatisfactory results, as reported in [Wang et al., 2018]. We therefore use randomly cropped patches of a smaller size. A patch size of $512 \times 512$ pixels is found empirically to provide sufficient anatomical context, without degradation in image quality. Fig. 4.1 shows an overview of our network architecture.
3.3 Experiments and Results

Implementation Details and Network Training. We use the Adam optimizer [Kingma et al., 2014] with a learning rate of 0.0002 and exponential decay rates $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Since GANs in general underfit [Wu et al., 2016] and the Nash equilibrium is often not reached in practice, we early stop training at 50k iterations, by when FID of a randomly-sampled training subset saturates. We use a batch size of 16 and set $\lambda = 100$. Our dataset consists of 6669 4-tuples $(x, y, s, a)$ and a constant binary mask $m$ covering the beam shape for all samples. We use randomly-selected 6000 images for training and the rest for evaluation. To quantitatively evaluate our models, from each test image we randomly crop four patches of size $512 \times 512$, yielding an evaluation set of 2676 image patches that are not seen during training. Note that our original dataset consists of images that are temporally far apart, thus the test images cannot be temporally consecutive and thus inherently similar to any training images.

Comparative Evaluation. To demonstrate the effectiveness of the proposed additional inputs from the image formation process, we conduct an ablation study by considering different combinations of network inputs. We refer the pix2pix network with low quality image and binary mask in the input channel as our baseline $L2H_M$. We compare this baseline with the following variants: 1) $L2H_{MS}$: $L2H_M$ with segmentation map $s$ as additional input; 2) $L2H_{MSA}$: $L2H_{MS}$ with attenuation integral map $a$ as additional input.

Qualitative Results. Fig. 3.3 shows a visual comparison of the three model variants on four examples. The baseline $L2H_M$ fails to preserve anatomical structures due to missing structural information in the input images. Resulting ambiguities in the network prediction cause artifacts such as blur in regions that feature fine details such as bones. Providing segmentation maps as additional input ($L2H_{MS}$) greatly reduces such artifacts as shown in Fig. 3.3(c). However, $L2H_{MS}$ still struggles in modeling complex non-local
features such as directional occlusion artifacts, note the lack of acoustic shadows in Fig. 3.3(c). In contrast, our final model $L2H_{MSA}$ is able to accurately synthesize these features and produces translations significantly closer to the target, as demonstrated in Fig. 3.3(d). In particular, our proposed model with segmentation and attenuation integral maps is able to recover both missing anatomical structures and directional artefacts.

Quantitative Results. The effectiveness of the proposed model is further evaluated using the following quantitative metrics:

1) PSNR: Peak signal-to-noise ratio between two images $A$ and $B$ is defined by $\text{PSNR} = 10 \log_{10} \left( \frac{255}{\text{MSE}} \right)$ with mean squared error $\text{MSE}$ between $A$ and $B$.

2) SSIM: Structural similarity index quantifies the visual changes in structural information as $\text{SSIM}(A, B) = \frac{(2\mu_{A}B + c_1)(2\sigma_{AB} + c_2)}{(\mu_{A}^2 + \mu_{B}^2 + c_1)(\sigma_{A}^2 + \sigma_{B}^2 + c_2)}$, with regularization constants $c_1$ and $c_2$, local means $\mu_A$ and $\mu_B$, local standard deviations $\sigma_A$ and $\sigma_B$, and cross covariance $\sigma_{AB}$. We use the default parameters of the MATLAB implementation to compute the metric.

3) pKL: Speckle appearance, relevant for tissue characterization in US images [Shankar et al., 1993], affects image histogram statistics. Hence, discrepancy in histogram statistics
can quantify differences in tissue-specific speckle patterns. Kullback-Leibler divergence compares normalized histograms $h_A$ and $h_B$ of two images $A$ and $B$ as: $KL(h_A || h_B) = \sum_{l=1..d} h_A[l] \log \left( \frac{h_A[l]}{h_B[l]} \right)$. We set the number of histogram bins $d$ to 50. To emphasize structural differences, we calculate KL divergence locally within $32 \times 32$ sized non-overlapping patches and report the metric mean, called patch KL (pKL) herein.

4) FID: Fréchet Inception Distance compares the distributions of generated samples and real samples by computing the distance between two multivariate Gaussians fitted to hidden activations of Inception network v3. This is a widely used metric to evaluate GAN performance, capturing both perceptual image quality and mode diversity. For this purpose, center crops of test images are sub-divided into four pieces of $299 \times 299$, to match Inception v3 input size.

Tab. 5.1 summarizes quantitative results for all models and all metrics, with the additional comparison to the discrepancy between low quality and high quality images as reference. A preliminary baseline experiment without GAN loss resulted in very blurry images with an FID score of 184.71. The results in Tab. 5.1 demonstrate that $L2H_{\text{MSA}}$ achieves the best translation performance in terms of all proposed metrics. The effectiveness of providing informative inputs to the network is well demonstrated in the gradual improvement in PSNR, SSIM and pKL, showing higher fidelity in anatomical structures and directional shadow artifacts. The metric pKL gives further indication of closer speckle appearance achieved by $L2H_{\text{MSA}}$. Based on Wilcoxon signed-rank tests, improvements of $L2H_{\text{MSA}}$ over $L2H_{\text{MS}}$ and those two over the baseline $L2H_M$ are statistically significant ($p < 10^{-5}$) for all evaluation metrics. Moreover, FID score indicates higher statistical similarity between the target and generated images using the proposed final model, with an improvement of 7.2% compared to $L2H_M$.

**Full Field-of-view Images.** Above image translation has been demonstrated on patches. For the entire field-of-view (FoV) US images, patch fusion from image translation of non-overlapping patches would cause artifacts at image seams. Averaging overlapping patches, on the other hand, would blur the essential US texture. Although seamless tiling of US images is possible using graphical models [Flach et al., 2016], this requires prohibitively long computation time. Herein, we instead directly apply our trained generator on full FoV low-quality images, since the generator is fully convolutional and thus can operate on images of arbitrary size. Fig. 4.2 shows two examples of translated images by $L2H_{\text{MS}}$ and $L2H_{\text{MSA}}$, demonstrating direct inference on full FoV images. While

<table>
<thead>
<tr>
<th></th>
<th>PSNR</th>
<th>SSIM [%]</th>
<th>pKL ($\times 10^2$)</th>
<th>FID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std</td>
<td>%ile</td>
<td>mean</td>
</tr>
<tr>
<td>Low quality</td>
<td>25.31</td>
<td>4.07</td>
<td>20.18</td>
<td>64.05</td>
</tr>
<tr>
<td>$L2H_M$</td>
<td>29.07</td>
<td>3.71</td>
<td>24.62</td>
<td>70.75</td>
</tr>
<tr>
<td>$L2H_{\text{MS}}$</td>
<td>29.26</td>
<td>3.71</td>
<td>24.78</td>
<td>71.22</td>
</tr>
<tr>
<td>$L2H_{\text{MSA}}$</td>
<td><strong>29.40</strong></td>
<td>3.71</td>
<td><strong>24.89</strong></td>
<td><strong>71.47</strong></td>
</tr>
</tbody>
</table>
3.4 Discussion and Conclusions

We have proposed a patch-based generative adversarial network for improving the quality of simulated US images, via image translation of computationally low-cost images to high quality simulation outputs. Providing segmentation and attenuation integral maps to the translation framework greatly improves preservation of anatomical structures and synthesis of important acoustic shadows. Continuous simulation parameters, such as transmit focus and depth-dependent lateral resolution, are implicitly captured by our framework thanks to training on image patches. For discrete simulation parameters such as imaging mode and transducer frequency that can take a handful of different values in typical clinical imaging, it is feasible to train a separate GAN for each such setting.

Image rendering time highly depends on chosen simulation parameters and 3D mesh model complexity. For instance, high framerates are reported for a simpler model in [Starkov et al., 2019b]. Rendering high and low quality images herein takes 75 ms and 40 ms, respectively. Our network inference time with a non-optimized code is 12.6 ms on average for full FoV images on a GTX 2080 Ti using TensorRT. This timing improvement is rather a lower-bound, since network inference can be further accelerated, e.g. with FPGAs [Guo et al., 2017]. Furthermore, since a pass through the network runs in constant time, potential time gain can be arbitrarily high depending on the desired complexity of the target simulation. With our proposed framework a trade-off between image quality and computational speed is obviated, thus enabling interactive framerates even with sophisticated anatomical scenes and computationally-taxing simulation settings. Although the convolutional network can process arbitrary sized image, translating full FoV images without any artifacts is still a challenge.
Learning Ultrasound Rendering from Cross-Sectional Model Slices for Simulated Training

**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.

4.1 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 [Blum et al., 2013; Østergaard et al., 2019; Ramaiah et al., 2020; Sun et al., 2011]. 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 [Goksel et al., 2009] 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 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 [Burger et al., 2013; Mattausch et al., 2018; Computing et al., 2015; Starkov et al., 2019b] 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 [Burger et al., 2013] to simulate ultrasonic directional wave interactions. The method is further extended in [Computing et al., 2015] to utilize patient-specific volumetric MRI or CT data. A more sophisticated stochastic wave-interaction and surface modeling is proposed in [Mattausch et al., 2018] 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 [Mattausch et al., 2018; Starkov et al., 2019b] 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 [Shams et al., 2008; Goksel et al., 2009; Burger et al., 2013; Mattausch et al., 2018; Starkov et al., 2019a; Starkov et al., 2019b] 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 [Goodfellow et al., 2014] (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 [Hu et al., 2017] and echogenicity map [Tom et al., 2018]. A recent publication [Magnetti et al., 2020] 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 [Bargsten et al., 2020] have introduced a speckle layer to incorporate the physical model of speckle generation into a GAN-based data augmentation network. In [Vitale et al., 2019], a cycleGAN [Zhu et al., 2017b] model is employed for improving the realism of simulated ultrasound images. A constrained cycleGAN is proposed in [Jafari et al., 2020] 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 [Zhang et al., 2020a] 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. 4.1(a). Herein we adopt a Monte-carlo simulation presented in [Mattausch et al., 2018] for simulating the same abdominal scene as in previous work [Zhang et al., 2020a]. Since low-quality images as used in [Zhang et al., 2020a] 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 [Zhang et al., 2020a], 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.
Figure 4.1: (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 [Zhang et al., 2020a]. (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.
4.2 Methods

4.2.1 Ray-based Ultrasound Simulation

In this work we adopt a Monte-Carlo ray tracing framework presented in [Mattausch et al., 2018] 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 [Mattausch et al., 2015], with spatially-variant point-spread-function. Each tissue type is assigned a set of pre-determined acoustic properties such as acoustic 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 5.2.

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.

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 [Zhang et al., 2020a] 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 $\mu$. The acoustic shadow formation is thus approximated by accumulating the attenuation strength, quantified by $\mu$, 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.

4.2.2 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 [Zhang et al., 2020a] in order to mimic wavefront traversal in tissue. This is known from [Zhang et al., 2020a] 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 \textit{att}.

In contrast to [Zhang et al., 2020a], herein we propose the following four major architectural improvements, as also illustrated in Fig. 4.1, which are later demonstrated to provide promising generation results in our experiments.

\textbf{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. 4.1(c). Although this was shown in [Park et al., 2019] 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 \textit{concat}.

\textbf{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.

\textbf{Texture-friendly decoder}: Transposed convolutions are known to be prone to characteristic checkerboard artifacts due to uneven overlap when using odd kernel sizes [Odena et al., 2016]. 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 \textit{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. [Karras et al., 2019] 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. 4.1(c). 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 [Zhang et al., 2020a] with noise, and therefore is referred as NSA2H in our comparative study.

We adopt the patchGAN discriminator and the training objective from [Zhang et al., 2020a]. The loss function consists of a GAN training objective $L_{\text{GAN}}$ and an $L_1$-based data fidelity term $L_F$ as follows:

$$L = L_{\text{GAN}}(G, D) + \lambda L_F(G), \quad (4.1)$$

$$L_{\text{GAN}} = \mathbb{E}_{s,a,y} [\log D(y|s,a)] + \mathbb{E}_{s,a} [\log (1 - D(m \circ G(s,a)|s,a))], \quad (4.2)$$

$$L_F = \mathbb{E}_{s,a,y} [||y - m \circ G(s,a)||_1], \quad (4.3)$$

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. $\mathbb{E}_{s,a,y}$ and $\mathbb{E}_{s,a}$ 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 $\circ$.

4.3 Results and Discussion

Implementation and network details. All our models are trained using the Adam optimizer [Kingma et al., 2014] 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 $\lambda$ 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 [Zhang et al., 2020a] consisting of 6669 3-tuples $(s, a, y)$ with an image size of $1000 \times 1386$. The network is trained using on-the-fly cropped image patches with a size of $512 \times 512$ 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 [Zhang et al., 2020a].
**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. 4.1 (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. 4.1 (c));
- **LSA2H** the recent low-to-high-quality image translation network presented in [Zhang et al., 2020a], 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. 4.2 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 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. 4.2) 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. 4.2) 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. 4.2.

**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 visual model. Accordingly, we utilize the following complementary quantitative metrics:

- **Peak signal-to-noise ratio (PSNR)** is computed as $\text{PSNR} = 10 \log_{10}(\frac{255}{\text{MSE}})$ 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.

- $\chi^2$: We use $\chi^2$ distance to measure the difference in image histogram statistics, which are commonly used for tissue characterization [Mailloux et al., 1985; Shankar et al., 1993]. This metric indicates potential mismatch in tissue speckle patterns, which affects image histogram statistics. $\chi^2$ distance computes the difference between histograms $h_A$ and $h_B$ as: $\chi^2(h_A || h_B) = \frac{1}{2} \sum_{l=1..d} (\frac{h_A[l] - h_B[l]}{h_A[l] + h_B[l]})^2$. The number of histogram bins $d$ is set to 50. We compute this metric locally within non-overlapping sliding patches with a size of $20 \times 20$ to capture local textural information of US speckle patterns. Herein we compute the root mean squared error, referred to patch $\chi^2$ ($\chi^2$).

- **Fréchet Inception Distance (FID)** [Heusel et al., 2017] 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 $512 \times 512$ sized center crops of generated full field-of-view images, which are further divided into four sub-crops of $299 \times 299$, to increase the number of samples and match Inception v3 input size.

For the interpretation of the local errors, sample spatial $\chi^2$ error maps are depicted in Fig. 4.3 for LSA2H, NSA2H, and SA2H for the middle two examples shown in Fig. 4.2. 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 $\chi^2$ 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
Figure 4.2: 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.
4.3 Results and Discussion

Figure 4.3: Spatial $\chi^2$ error map for the examples shown in Fig. 4.2. 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.

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 Tab 5.1. 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 $\chi^2$ 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. 4.4 we show the distribution of paired differences in PSNR, MAE, and $\chi^2$ 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
Table 4.2: Quantitative results. Bold number indicates the best performance. Mean and standard deviation are reported for PSNR, MAE and $p\chi^2$. Network capacity is given as number of trainable parameters.

<table>
<thead>
<tr>
<th></th>
<th>PSNR</th>
<th>MAE</th>
<th>$p\chi^2$ ($\times 10^{-2}$)</th>
<th>FID</th>
<th>#params</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std</td>
<td>mean</td>
<td>std</td>
<td></td>
</tr>
<tr>
<td>LSA2H</td>
<td>27.38</td>
<td>0.52</td>
<td>6.21</td>
<td>1.56</td>
<td>13.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.39</td>
<td></td>
<td>64.42</td>
</tr>
<tr>
<td>NSA2H</td>
<td>25.59</td>
<td>1.81</td>
<td>8.22</td>
<td>2.77</td>
<td>31.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>67.28</td>
<td></td>
<td>67.28</td>
</tr>
<tr>
<td>SA2H-att</td>
<td>25.04</td>
<td>1.61</td>
<td>8.75</td>
<td>2.75</td>
<td>29.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>92.49</td>
<td></td>
<td>92.49</td>
</tr>
<tr>
<td>SA2H-concat</td>
<td>26.23</td>
<td>1.22</td>
<td>9.08</td>
<td>2.93</td>
<td>29.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>76.40</td>
<td></td>
<td>76.40</td>
</tr>
<tr>
<td>SA2H-conv</td>
<td>24.85</td>
<td>1.31</td>
<td>8.87</td>
<td>2.55</td>
<td>32.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>93.18</td>
<td></td>
<td>93.18</td>
</tr>
<tr>
<td>SA2H-noise</td>
<td>26.45</td>
<td>1.34</td>
<td>8.14</td>
<td>2.66</td>
<td>25.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>97.85</td>
<td></td>
<td>97.85</td>
</tr>
<tr>
<td>SA2H</td>
<td>28.19</td>
<td>1.23</td>
<td>6.37</td>
<td>2.08</td>
<td>18.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>32.34</td>
<td></td>
<td>32.34</td>
</tr>
</tbody>
</table>

Figure 4.4: 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.
4.4 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 [Guo et al., 2017]. 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.
Content-Preserving Unpaired Translation from Simulated to Realistic Ultrasound Images

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.

5.1 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 for training, especially of rare diseases, computational methods become essential as simulation-based training tools. To that end, interpolation of pre-acquired US volumes [Goksel et al., 2009] provide only

This chapter was published as a conference article: Devavrat Tomar, Lin Zhang, Tiziano Portenier, Orcun Goksel, “Content-preserving unpaired translation from simulated to realistic ultrasound images”, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 659–669, 2021, doi: 10.1007/978-3-030-87237-3_3
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 [Burger et al., 2013]. Monte-Carlo ray-tracing [Mattausch et al., 2018] has further enabled realistic soft shadows and fuzzy reflections, while animated models and fusion of partial-frame simulations were also presented [Starkov et al., 2019a]. 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 [Mattausch et al., 2017; Starkov et al., 2019b; Zhang et al., 2020b], 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 [Jiao et al., 2020], image enhancement [Jafari et al., 2020; Zhang et al., 2020a; Zhang et al., 2021], and semantic image synthesis [Bärgsten et al., 2020; Tom et al., 2018]. 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) [Goodfellow et al., 2014] 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 [Mirza et al., 2014] has shown great success on various translation tasks [Isola et al., 2017]. In the absence of paired training samples, the translation problem becomes under-constrained and additional constraints are required to learn a successful translation. To tackle this issue, a cyclic consistency loss (cycleGAN) was proposed [Zhu et al., 2017b], 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 [Almahairi et al., 2018; Choi et al., 2018; Zhu et al., 2017a]. Cycle consistency assumes a strong bijective relation between the domains. To relax the bijectivity assumption and reduce the training burden, Park et al. [Park et al., 2020] proposed an alternative with a single-sided unpaired translation technique with contrastive learning. For US simulation, the standard cycleGAN was used in [Vitale et al., 2019] 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 [Park et al., 2020] 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.
5.2 Method

Given unpaired source images $X = \{x \in X\}$ and target images $Y = \{y \in Y\}$, we aim to learn a generator function $G : X \mapsto Y$, 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 $G_{\text{enc}}$ and a decoder $G_{\text{dec}}$. $G_{\text{enc}}$ is restricted to extract content-related features only, while $G_{\text{dec}}$ 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. 5.1).

**Adversarial loss.** We adopt the patchGAN discriminator [Park et al., 2020] that discriminates real and fake images using a least squares GAN loss:

$$\mathcal{L}_{\text{GAN}}(X, Y) = \mathbb{E}_y \log[(D(y) - 1)^2] + \mathbb{E}_x \log[D(G(x))^2].$$  \hspace{1cm} (5.1)

**Contrastive loss.** An unpaired contrastive translation framework (CUT) is presented in [Park et al., 2020] 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 $G_{\text{enc}}$. 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 $G_{\text{enc}}$. 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 $z_{l,s}^{i} = h_l(G_{\text{enc}}^l(\hat{y})) \in \mathbb{R}^{C_l}$ of the output image $\hat{y}$ acts as the query. The contrastive loss is defined as the cross-entropy loss

$$I(z_{l,s}^{i}, z_{l,s}^{+}, z_{l,s}^{-}) = -\log \left[ \frac{\exp(z_{l,s}^{i} \cdot z_{l,s}^{+}/\tau)}{\exp(z_{l,s}^{i} \cdot z_{l,s}^{+}/\tau) + \sum_{k=1}^{S_l-1} \exp(z_{l,s}^{i} \cdot z_{l,s}^{k}/\tau)} \right],$$  \hspace{1cm} (5.2)
with the temperature parameter $\tau$ set to 0.07, following [Park et al., 2020]. Using features from multiple encoder depths allows us to enforce patch similarity on multiple scales, leading to the following noise contrastive estimation (NCE) loss

$$L_{\text{NCE}}(X) = \mathbb{E}_x \sum_{l=1}^L \sum_{s=1}^{S_l} l(z^s_l, z^{s+}_l, z^{s-}_l),$$

(5.3)

where $L$ is the number of layers used for computing the loss. To encourage the generator to translate the domain-specific image appearance only, $L_{\text{NCE}}$ is also evaluated on the target domain $Y$, which acts as an identity loss, similarly to the cyclic consistency loss in [Zhu et al., 2017b]. The final objective in CUT [Park et al., 2020] is defined as

$$L_{\text{CUT}}(X, Y) = L_{\text{GAN}}(X, Y) + L_{\text{NCE}}(X) + L_{\text{NCE}}(Y).$$

(5.4)

Semantic-consistent regularization. To encourage the disentanglement of content and style, we leverage available surrogate segmentation maps $S = \{s \in S\}$ 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 $G_{\text{enc}}$, there is no style left in the features, therefore $G_{\text{dec}}$ has to introduce styles entirely from scratch. Learning this auxiliary task thus prevents any (style) leakage from $G_{\text{enc}}$, enforcing $G_{\text{enc}}$ to extract only content-relevant features. In this modified CUT framework with semantic input (CUT+S), we minimize

$$L_{\text{CUT+S}} = L_{\text{CUT}}(X, Y) + L_{\text{GAN}}(S, Y) + L_{\text{NCE}}(S).$$

(5.5)

In addition, we regularize $G$ to generate the ground truth 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: $L_{\text{REG}}(X, S) = \mathbb{E}_{x,s} ||G(x) - G(s)||_1$. Our consistency-based training objective then becomes:

$$L_{\text{CUT+SC}} = L_{\text{CUT+S}} + \lambda_{\text{REG}} L_{\text{REG}}(X, S).$$

(5.6)

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 [Zhu et al., 2017b]. 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 [Choi et al., 2018]. Here, $G_{\text{dec}}$ is trained to transfer the target appearance, conditioned by the target class label $\ell \in \{A, B, S\}$ given the classes $A$ simulated image, $B$ real image, and $S$ 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

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

(5.7)

where $\ell'$ is the class label of the input image and $\ell$ is label of the target class.
5.3 Experiments and Results

Classification loss. To enable class-dependent classification (CLS) with the discriminator [Choi et al., 2018], $D$ tries to predict the correct domain class label $\ell'$ for a given real image $x$ as an auxiliary task, i.e.

$$L_{CLS,r}(X) = \mathbb{E}_{x,\ell'}[-\log D(\ell'|x)], \quad (5.8)$$

while $G$ tries to fool $D$ with fake images to be classified as target domain $\ell$ by minimizing

$$L_{CLS,f}(X) = \mathbb{E}_{x,\ell}[-\log D(\ell|G(x, \ell))]. \quad (5.9)$$

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

$$L_{ConPres}^D = \sum_{i=1}^{2} -L_{GAN}(X_i, Y_i) + \lambda_{CLS,r} L_{CLS,r}(X_i), \quad (5.10)$$

$$L_{ConPres}^G = \sum_{i=1}^{2} L_{CUT}(X_i, Y_i) + \lambda_{CLS,f} L_{CLS,f}(X_i) + \lambda_{CYC} L_{CYC}(X_i) + \mathbb{1}_{[(X_1=A \land X_2=S) \lor (X_1=S \land X_2=A)]} \lambda_{REG} L_{REG}(X_1, X_2) \quad (5.11)$$

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

5.3 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 $256 \times 354$, see Fig. 5.2 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. Fig. 5.3 shows example B-mode images with corresponding semantic maps. A total of 6669 simulated frames were resized to $256 \times 354$ 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) [Heusel et al., 2017] 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) [Bińkowski et al., 2018] 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 [Choi et al., 2018]. 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 [Park et al., 2020]. To compute $\lambda_{\text{REG}}$, the sampled simulated and segmentation images in each batch are paired. We used Adam [Kingma et al., 2014] optimizer to train our model for 100 epochs with an $l_2$ regularization of $10^{-4}$ with gradient clipping on model parameters 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., $\lambda_{\text{REG}}$, for stable GAN training. We implemented our model in PyTorch [Paszke et al., 2019]. For KID and FID computations, we used the implementation of [Obukhov et al., 2020].

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

• **CycleGAN** [Zhu et al., 2017b]: A conventional approach with cyclic consistency loss.

• **SASAN** [Tomar et al., 2021b]: CycleGAN extension with self-attentive spatial adaptive normalization, leveraging semantic information to retain anatomical structures, while translating using spatial attention modules and SPADE layers [Park et al., 2019].

• **CUT** [Park et al., 2020]: Unpaired contrastive framework for image translation.

• **StarGAN** [Choi et al., 2018]: 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 $L_{\text{REG}}$).
Table 5.1: Quantitative metrics and ranking from the user study (mean±std). Best results are marked bold. “Seg” gives if semantic maps are used as network input.

<table>
<thead>
<tr>
<th>Seg</th>
<th>Method</th>
<th>SSIM ↑</th>
<th>FID ↓</th>
<th>KID ↓</th>
<th>Ranking ∈ [1, 6]↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>Simulation</td>
<td>—</td>
<td>2.37</td>
<td>0.41</td>
<td>3.98±1.35</td>
</tr>
<tr>
<td>✗</td>
<td>CycleGAN [Zhu et al., 2017b]</td>
<td>71.73±1.78</td>
<td>1.78</td>
<td>0.32</td>
<td>2.86±1.27</td>
</tr>
<tr>
<td>✓</td>
<td>SASAN [Tomar et al., 2021b]</td>
<td>68.20±4.00</td>
<td>2.36</td>
<td>0.39</td>
<td>3.59±1.55</td>
</tr>
<tr>
<td>✗</td>
<td>CUT [Park et al., 2020]</td>
<td>67.28±4.62</td>
<td>1.77</td>
<td>0.31</td>
<td>2.92±1.20</td>
</tr>
<tr>
<td>✓</td>
<td>StarGAN [Choi et al., 2018]</td>
<td>63.62±4.82</td>
<td>1.93</td>
<td>0.47</td>
<td>5.76±0.61</td>
</tr>
<tr>
<td>✓</td>
<td>CUT+S</td>
<td>68.88±4.63</td>
<td>2.25</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>CUT+SC</td>
<td>80.56±2.11</td>
<td>1.87</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>ConPres</td>
<td>72.13±4.58</td>
<td>1.51</td>
<td>0.24</td>
<td>1.89±1.07</td>
</tr>
</tbody>
</table>

In Fig. 5.3 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 $\mathcal{L}_{\text{NCE}}$ is unstable.

Comparison to state-of-the-art. As seen qualitatively from the examples in Fig. 5.3, 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. 5.2) used for training, our ConPres maintains the input FoV closely compared to previous state-of-the-art. The results in Tab 5.1 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.

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 questionnaire window opened in a web interface, presenting the participants with six candidate images including the input simulated frame 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 Tab 5.1. 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.

### 5.4 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.

### 5.5 Appendix

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangles fetus</td>
<td>400k</td>
<td>Transducer frequency</td>
<td>8 MHz</td>
</tr>
<tr>
<td>Triangles mother</td>
<td>275k</td>
<td>Field-of-view</td>
<td>70°</td>
</tr>
<tr>
<td>Image depth</td>
<td>15.0 cm</td>
<td>Axial samples</td>
<td>3072</td>
</tr>
<tr>
<td>Elevational layer</td>
<td>3</td>
<td>Ray/scanline</td>
<td>32</td>
</tr>
</tbody>
</table>
Table 5.3: Generator Architecture. IN: instance normalization, K: kernel size of convolution layers, shape $h \times w \times ch$: height × weight × number of channels. We use bilinear interpolation for Upsample layers. Note that we use instance normalization only in the encoder. Discriminator Architecture. K: kernel size of the convolution layers, shape $h \times w \times ch$: height × weight × number of channels.

<table>
<thead>
<tr>
<th>Layer Name</th>
<th>Input Shape</th>
<th>Output Shape</th>
<th>Layer Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Down-sampling</td>
<td>$h \times w \times 1$</td>
<td>$h \times w \times 64$</td>
<td>CONV-(K7 × 7), IN, ReLU</td>
</tr>
<tr>
<td></td>
<td>$h \times w \times 64$</td>
<td>$\frac{h}{2} \times \frac{w}{2} \times 128$</td>
<td>CONV-(K4 × 4), IN, ReLU</td>
</tr>
<tr>
<td></td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 256$</td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 256$</td>
<td>Residual Block: CONV-(K3 × 3), IN, ReLU</td>
</tr>
<tr>
<td>Bottleneck</td>
<td>$\frac{h}{2} \times \frac{w}{2} \times 256$</td>
<td>$\frac{h}{2} \times \frac{w}{2} \times 256$</td>
<td>Residual Block: CONV-(K3 × 3), ReLU</td>
</tr>
<tr>
<td></td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 256$</td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 256$</td>
<td>Residual Block: CONV-(K3 × 3), ReLU</td>
</tr>
<tr>
<td></td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 256$</td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 256$</td>
<td>Residual Block: CONV-(K3 × 3), ReLU</td>
</tr>
<tr>
<td></td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 256$</td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 256$</td>
<td>Residual Block: CONV-(K3 × 3), ReLU</td>
</tr>
<tr>
<td>Decoder</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bottleneck</td>
<td>$\frac{h}{2} \times \frac{w}{2} \times (256 + n_c)$</td>
<td>$\frac{h}{2} \times \frac{w}{2} \times (256 + n_c)$</td>
<td>Residual Block: CONV-(K3 × 3), ReLU</td>
</tr>
<tr>
<td></td>
<td>$\frac{h}{4} \times \frac{w}{4} \times (256 + n_c)$</td>
<td>$\frac{h}{4} \times \frac{w}{4} \times (256 + n_c)$</td>
<td>Residual Block: CONV-(K3 × 3), ReLU</td>
</tr>
<tr>
<td></td>
<td>$\frac{h}{4} \times \frac{w}{4} \times (256 + n_c)$</td>
<td>$\frac{h}{4} \times \frac{w}{4} \times (256 + n_c)$</td>
<td>Residual Block: CONV-(K3 × 3), ReLU</td>
</tr>
<tr>
<td></td>
<td>$\frac{h}{4} \times \frac{w}{4} \times (256 + n_c)$</td>
<td>$\frac{h}{4} \times \frac{w}{4} \times (256 + n_c)$</td>
<td>Residual Block: CONV-(K3 × 3), ReLU</td>
</tr>
<tr>
<td></td>
<td>$\frac{h}{4} \times \frac{w}{4} \times (256 + n_c)$</td>
<td>$\frac{h}{4} \times \frac{w}{4} \times (256 + n_c)$</td>
<td>Residual Block: CONV-(K3 × 3), ReLU</td>
</tr>
<tr>
<td>Up-sampling</td>
<td>$\frac{h}{2} \times \frac{w}{2} \times 128$</td>
<td>$\frac{h}{2} \times \frac{w}{2} \times 128$</td>
<td>CONV-(K3 × 3), Upsample, ReLU</td>
</tr>
<tr>
<td></td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 256$</td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 256$</td>
<td>CONV-(K3 × 3), Upsample, ReLU</td>
</tr>
<tr>
<td>Final</td>
<td>$h \times w \times 64$</td>
<td>$h \times w \times 1$</td>
<td>CONV-(K3 × 3), Tanh</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer Name</th>
<th>Input Shape</th>
<th>Output Shape</th>
<th>Layer Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Layer</td>
<td>$h \times w \times 1$</td>
<td>$\frac{h}{2} \times \frac{w}{2} \times 64$</td>
<td>CONV-(K4 × 4), Leaky ReLU</td>
</tr>
<tr>
<td>Hidden Layers</td>
<td>$\frac{h}{2} \times \frac{w}{2} \times 64$</td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 128$</td>
<td>CONV-(K4 × 4), Leaky ReLU</td>
</tr>
<tr>
<td></td>
<td>$\frac{h}{4} \times \frac{w}{4} \times 128$</td>
<td>$\frac{h}{8} \times \frac{w}{8} \times 256$</td>
<td>CONV-(K4 × 4), Leaky ReLU</td>
</tr>
<tr>
<td></td>
<td>$\frac{h}{8} \times \frac{w}{8} \times 256$</td>
<td>$\frac{h}{16} \times \frac{w}{16} \times 512$</td>
<td>CONV-(K4 × 4), Leaky ReLU</td>
</tr>
<tr>
<td>Output Layer: $D_{fake/real}$</td>
<td>$\frac{h}{16} \times \frac{w}{16} \times 512$</td>
<td>$\frac{h}{16} \times \frac{w}{16} \times 1$</td>
<td>CONV-(K3 × 3)</td>
</tr>
<tr>
<td>Output Layer: $D_{classifier}$</td>
<td>$\frac{h}{16} \times \frac{w}{16} \times 512$</td>
<td>$1 \times 1 \times n_c$</td>
<td>CONV-(K $\frac{h}{16} \times \frac{w}{16}$)</td>
</tr>
</tbody>
</table>
Figure 5.3: Qualitative results, with images masked by foreground in segmentations.
Figure 5.4: Other sim-to-real image translation examples from our user study. Our method is seen to preserve the original content while translating the appearance. Top: input simulated images, middle: translated real images, bottom: semantic maps for reference.

Figure 5.5: Example results of seg-to-real and seg-to-sim image translation: Top: input semantic maps, middle: translated real images, bottom: translated simulation images.
6

Unpaired translation from semantic label maps to images by leveraging domain-specific simulations

Photorealistic image generation from simulated label maps are necessitated in several contexts, such as for medical training in virtual reality. With conventional deep learning methods, this task requires images that are paired with semantic annotations, which typically are unavailable. We introduce a contrastive learning framework for generating photorealistic images from simulated label maps, by learning from unpaired sets of both. Due to potentially large scene differences between real images and label maps, existing unpaired image translation methods lead to artifacts of scene modification in synthesized images. We utilize simulated images as surrogate targets for a contrastive loss, while ensuring consistency by utilizing features from a reverse translation network. Our method enables bidirectional label-image translations, which is demonstrated in a variety of scenarios and datasets, including laparoscopy, ultrasound, and driving scenes. By comparing with state-of-the-art unpaired translation methods, our proposed method is shown to generate realistic and scene-accurate translations.

6.1 Introduction

Photorealistic image simulation has been an active research area for decades with a wide range of applications from movie- and game-industries [Masuch et al., 2004; Tewari et al., 2020] to medical imaging for surgical training [Frangi et al., 2018; ElHelw et al., 2004; Burger et al., 2013; Mattausch et al., 2018]. Extensive research in modelling imaging physics [Pharr et al., 2016; Mattausch et al., 2018; Burger et al., 2013] and material representations [Weyrich et al., 2009; Zhang et al., 2020a] has substantially improved simulation realism in different applications, but there is still a very perceivable visual difference between the state-of-the-art simulators and real world images. Recent progress in deep learning has paved the way for synthesizing photorealistic images by learning image features from large-scale real data. Among them, generative adversarial networks [Good-
fellow et al., 2014] have shown promising results in generating photorealistic images. Methods were shown to successfully generate images from noise inputs with controlled styles at different level of details learned from given domains [Karras et al., 2019; Karras et al., 2020a; Karras et al., 2021] as well as to achieve semantic image synthesis [Isola et al., 2017; Park et al., 2019; Wang et al., 2018; Liu et al., 2019; Jiang et al., 2020] given paired label and target images. In the absence of paired training samples, methods based on cycle consistency loss [Zhu et al., 2017b; Yi et al., 2017], shared latent space [Huang et al., 2018; Liu et al., 2017], layer normalizations [Kim et al., 2019] and more recently contrastive loss [Park et al., 2020; Zheng et al., 2021] have been investigated with varying levels of success.

As opposed to the widely used cyclic approaches, contrastive learning (CL) based unpaired image translation methods focus on translating in single direction by relaxing the strong bijective assumption, which has achieved impressive results in various unpaired translation settings [Park et al., 2020]. Compared to image-level feature contrasting for unsupervised classification and segmentation, patch-level contrasting is employed in [Park et al., 2020] which enforces structural similarity between images. Various approaches have been accordingly proposed for bridging the appearance gap between synthetic and real images [Shrivastava et al., 2017; Vitale et al., 2020; Hoffman et al., 2018; Sharan et al., 2021], also further improved by leveraging auxiliary simulation information such as simulated semantic maps [Tomar et al., 2021a] and geometry buffers (G-buffers) generated during 3D computer graphics rendering [Richter et al., 2022]. A parallel line of work investigate photographic transfer [Li et al., 2017; Luan et al., 2017; Huang et al., 2017; Li et al., 2019], aiming at translating the appearance of reference images to simulated contents; however, such methods require lengthy and difficult-to-parametrize optimizations for each single target image. All above work aim to improve the realism of existing, sub-realistic (e.g., simulated) images and hence require the existence of preceding, complex simulation and rendering methods.

Photorealistic image generation directly from simulated scene layouts, so-called label-maps, would obviate any complex and computationally-intensive simulation/rendering process in real-time, by learning the internals of such rendering into a generative model during a non-time-critical offline learning stage. Such label-maps can typically be extracted easily from existing simulation pipelines, only given 3D objects models and a vantage point (thus a scene), i.e. without a need to tune model-specific parameters nor to compute complex physical interactions. To illustrate this further, a generic simulation pipeline is given in Figure 6.1(a). Given the above motivation, we aim to generate images $I^S$ with realistic appearance but simulation-controlled content, i.e. $I^S_R$ as appearance and content of a representation are hereafter indicated in sub- and super-script, respectively. With this convention, the methods from the literature mentioned earlier mostly target image-to-image translation of $I^S_S \rightarrow I^S_S$. In comparison, label-to-image translation as we intend is often more challenging due to the large domain shift between these two representations. Generating simulated images from labels, i.e. $L^S_S \rightarrow I^S_S$ translation, was studied in [Zhang et al., 2020a] for accelerating simulated image generation in real-time, for which a conditional GAN with supervised per-pixel loss was shown to provide promising results. This, however, relatively simpler compared to our goal of generating $I^S_R$, since the former can be cast as a paired translation problem where the paired data $(L^S_S, I^S_S)$ is typically available from conventional simulation pipelines. In contrast for our
6.1 Introduction

Physics of contact, deformation, ... 

Geometric processing: Projective transform, cross-sectional slice, ...

Contrastive loss $Z$ $M$!
$M$!
$G$ $L$!!
$I$!!
$I$""

Figure 6.1: Overview of unpaired label-image translation by leveraging domain-specific simulations. 
(a) An illustration of the simulation/generation pipeline from 3D computer graphics (CG) model to label-maps $L_{S}$ and images $I_{S}$ with the subscript indicating the style domain and superscript indicating the content domain. $R$ and $S$ denote the real and simulated domains, respectively. Note that the goal is to generate images $I_{S}$ with realistic appearance for simulated content, based on (i.e. consistent with) label-maps $L_{S}$ of simulated scenes. To that end, one can collect and use many real-life images $I_{R}$, but these will not one-to-one match existing simulated content therefore preventing classical supervised training. 
(b) A schematic summary of existing unpaired image translation approaches CycleGAN [Zhu et al., 2017b], CUT [Park et al., 2020], and ConPres [Tomar et al., 2021a], as well as our proposed methods SimIT and SimIT-C. We define the label-to-image mapping function as $G : L \rightarrow I$ and the image-to-label mapping function as $F : I \rightarrow L$, with the label domain $L$ and image domain $I$. Both $G$ and $F$ are parameterized with a deep neural network consisting of an encoder and a decoder, i.e. $G = [G^{E}, G^{D}]$ and $F = [F^{E}, F^{D}]$. The function $H : Z' \rightarrow Z$ maps between two latent domains. In the contrastive learning framework, contrastive loss is computed on features obtained from the mappings $M^{G}(.) = G^{E}(H(.))$ or $M^{F}(.) = F^{E}(H(.))$. 

R and S denote the real and simulated domains, respectively.
desired target of \(I_R^S\), there exists no such paired label data. A large domain gap together with the lack of paired data make our intended label-to-realistic-image translation very challenging, and, to the best of our knowledge, without any working solution so far.

In this work we target the above problem of photorealistic image generation directly from label-maps. To facilitate the learning of appearance information from real images \(I_R^R\), we propose to utilize any available physics-based simulation to generate intermediate image representations \(I_S^S\). We utilize these as a stepping stone to help bridge the domain gap between the labels \(I_S^S\) and their real-image counterparts \(I_R^R\) as desired. To that end, we introduce a contrastive learning based image translation framework that leverages physics-based simulations/rendering in the training of unpaired label-to-image translation, but without needing such simulations during the real-time inference stage. Compared to the existing works [Richter et al., 2022; Tomar et al., 2021a; Zhang et al., 2020a; Shrivastava et al., 2017], our proposed method performs image generation and realism enhancement simultaneously in a single step. We demonstrate our method on enhancing medical image simulators for training, as well as car driving simulators for entertainment.

Our proposed solution builds on a bidirectional (cyclic) translation idea. As a by-product of this design, it can also perform the inverse operation of image-to-label translation, i.e. semantic image segmentation is also learned meanwhile in an unsupervised fashion without seeing any annotations of real images. We also evaluate such segmentation outcomes in this work, as they opens future possibilities of alleviating annotation efforts.

6.2 Results

6.2.1 Compared methods

**Proposed method** We call our proposed method for generating realistic images from simulated semantic label-maps, with the target style learned from real images while retaining overall content matching the simulated scene, as Simulation-based Image Translation framework (SimIT). Realistic and scene-accurate translation given unpaired data is herein enabled by two major contributions:

1. To address missing label-image pair information, we leverage existing physics-based simulations by using the simulated images (that are inherently paired with corresponding label-maps) as surrogate targets for contrastive learning.

2. To enforce content/structure preservation, we devise a method that contrasts domain-specific image features extracted from a translation network that is trained using a cycle consistency loss. This further enables bidirectional translation, i.e. in both label-to-image and image-to-label directions.

**Compared methods** We evaluate SimIT comparatively to the following three state-of-the-art unpaired image translation methods: CycleGAN [Zhu et al., 2017b] is a conventional approach with cyclic consistency loss by employing separate generators and discriminators in each direction. CUT [Park et al., 2020] is a unidirectional translation framework based on patch-based multi-scale contrastive loss computed on generator features.
ConPres [Tomar et al., 2021a] is a multi-domain translation framework that leverages simulated label-image pairs to retain structural content. Together with cycle-consistency and contrastive losses, ConPres proposes a regularization loss for semantic consistency, which enforces a generator to create the same output for paired images and label-maps. Consequently, ConPres can be used for both image-to-image and label-to-image translation. The latter being the focus herein, we employ that use-case of ConPres in our comparisons. High-level conceptual schematics of above-mentioned three approaches are illustrated in Figure 6.1(b).

Ablations To further evaluate our two major contributions listed above, we ablated them cumulatively from SimIT, resulting in the following reduced models for comparison: SimIT-C (SimIT without cycle loss) is a unidirectional version of SimIT, i.e. without learning an inverse translation from image to labels, where the contrastive loss is then computed using features from the label-to-image generator, c.f. SimIT-C in Figure 6.1(b). SimIT-CS (SimIT-C without leveraging simulations) does not utilize any simulation information, where the contrastive loss is then computed between semantic label-maps and translated images, similarly to CUT in Figure 6.1(b).

6.2.2 Evaluation

All results are evaluated on unseen test data. We employ the non-parametric two-sided Wilcoxon signed-rank test to assess differences between paired test results and report statistical significance with p-value. Methodological and implementation details are given later in the Methods. We compared the methods on three different applications (more details given in the Methods):

Laparoscopy training As physics-based simulation, computer-graphics rendering techniques were employed [Harders, 2008; Tuchschmid, 2010], to simulate synthetic laparoscopic images from a 3D abdominal model. During the rendering of each frame, a camera projection of anatomical labels provided the corresponding semantic label-maps. This simulated dataset with paired image and label-maps are herein called as LaparoSim.

For the laparoscopy application, we employed two different datasets of real images, thus evaluating two different target styles, called herein: Style-C, represented by the public Cholec dataset containing 80 videos of cholecystectomy surgeries [Twinanda et al., 2016]; and Style-H, represented by a single, in-house laparoscopic surgery video clip with a length of 13 minutes. Sample images can be seen in Figure 6.2.

Ultrasound training The simulated images were generated using a ray-tracing framework [Mattausch et al., 2018] from a second trimester fetal model, by emulating a convex ultrasound probe at multiple locations and orientations on the abdominal surface, with imaging settings following [Tomar et al., 2021a]. Semantic label-map is rendered as a cross-section through the anatomical surfaces at the ultrasound center imaging plane. We refer this simulated dataset as USSim.

For the targeted real-image style, sample ultrasound images were collected using a GE Voluson E10 machine during standard fetal screening exams of 24 patients. We refer this as GE-E10 style.
**Gaming**  As the gaming simulation, we used the GTA dataset [Richter et al., 2016] containing image-label pairs from a car-driving game. For the real image style, we used the Cityscapes dataset [Cordts et al., 2016] containing images of street scenes from German cities.

6.2.3 Experiments

**Label-to-image translation**

For the laparoscopy training application, we present the results in Figure 6.2 for separately training two different styles. As seen qualitatively in Figure 6.2(b), CycleGAN and CUT hallucinate inexistent tissue regions, e.g., fat tissues. ConPres achieves structural preservation by leveraging information from LaparoSim label-image pairs, but fails completely in generating tissue textures, which leads to highly unrealistic image style. Going from label-to-image, our method SimIT is seen to outperform the state-of-the-art in terms of anatomical content preservation as well as in achieving a realistic image appearance. This observation is substantiated by the quantitative evaluation in Table 6.1(a), where image realism is empirically measured using Frechet and Kernel Inception Distances (FID and KID, respectively) between translated and real image sets, and the content preservation is measured via the structural similarity index measure (SSIM) between translated and corresponding simulated images. Note that SimIT also achieves the lowest SSIM standard deviation, indicating its consistent content preservation over the test samples. A test-image wise paired comparison of all methods with respect to SimIT is presented in Figure 6.2(c), which shows ConPres as the closest contender in terms of content preservation (SSIM) but with largely unrealistic image translation, as also demonstrated qualitatively (Figure 6.2(b)) and tabulated empirically (Table 6.1(a)).

Compared to the proposed method SimIT, its ablated variants, SimIT-C and SimIT-CS perform substantially poorer as seen quantitatively in Table 6.1(a) and Figure 6.2(c), and qualitatively in Figure 6.2(d). This demonstrates the importance of our proposed method components. SimIT-CS lacks our proposed component for utilizing simulations with a contrastive loss in learning the label-to-image translation, and as such it can be seen as a variant of CUT implemented in our framework. With no explicit label-to-image pairs provided, SimIT-CS then learns to simply emulate all structures seen in the real examples, hence erroneously changing the image content as seen in the presented examples. Using simulated images as surrogate targets for contrastive loss (SimIT-C in Figure 6.2(d)) largely prevents such superfluous content generation. Still SimIT-C only uses the features from a label domain for contrasting, and such features cannot be well aligned with image features. With the proposed method SimIT, the addition of a custom cycle loss allows for training a bidirectional translation, where the features from an encoder operating on images can then instead be used for contrasting. With such domain-consistent features, content preservation is further enhanced, as seen both quantitatively given the error of SimIT-C in Figure 6.2(c), and qualitatively by visually comparing these variants in Figure 6.2(d).

Evaluation on ultrasound training and gaming applications further confirms the superior performance of our proposed method on label-to-image translation task (Figures 6.3
Figure 6.2: (a) Examples of real laparoscopic images with two different appearances: Style-C for the public Cholec80 dataset and Style-H for an in-house single-video dataset. (b) Qualitative comparison of images translated from input LaparSim label-maps, using the proposed SimIT and alternative methods. For reference purposes, conventionally simulated/rendered LaparoSim images are shown on the right. (c) Quantitative evaluation of structural preservation via Structure Similarity Index Metric (SSIM). Using a paired test, distributions of pair-wise differences over the test set are shown by comparing SimIT to each alternative method and ablated variant, i.e. the larger the positive difference is, the more superior SimIT is with respect to another method. Significance is indicated with respect to SimIT (represented with the dotted lines) or between different models (—), with P-values of ≤ 0.0001 (***) (d) Qualitative comparison of our proposed method SimIT to its ablated variants, with translated images zoomed in on the white field-of-view shown in the simulated image as reference.
Table 6.1: Quantitative metrics reported as mean±std. Arrows indicate direction of superiority; ↑ meaning the higher the better, and ↓ the lower. KID is reported in 10^{-2} unit. Best results are marked bold.

(a) Laparoscopy

<table>
<thead>
<tr>
<th>Method</th>
<th>Style-C</th>
<th></th>
<th>Realism</th>
<th></th>
<th>Style-H</th>
<th></th>
<th>Realism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Content</td>
<td>SSIM [%] ↑</td>
<td>KID ↓</td>
<td>FID ↓</td>
<td>Content</td>
<td>SSIM [%] ↑</td>
<td>KID ↓</td>
</tr>
<tr>
<td>Simulation</td>
<td>—</td>
<td>257.39</td>
<td>17.76</td>
<td>—</td>
<td>—</td>
<td>201.32</td>
<td>12.42</td>
</tr>
<tr>
<td>CUT [Park et al., 2020]</td>
<td>49.79±13.75</td>
<td>234.65</td>
<td>12.85</td>
<td>—</td>
<td>58.74±6.77</td>
<td>222.81</td>
<td>13.42</td>
</tr>
<tr>
<td>ConPres [Tomar et al., 2021a]</td>
<td>71.12±3.86</td>
<td>380.70</td>
<td>36.72</td>
<td>—</td>
<td>75.76±5.56</td>
<td>379.82</td>
<td>36.80</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SimIT-CS</td>
<td>41.77±7.98</td>
<td>202.24</td>
<td>10.40</td>
<td>56.15±5.23</td>
<td>147.06</td>
<td>7.03</td>
<td></td>
</tr>
<tr>
<td>SimIT-C</td>
<td>58.05±7.34</td>
<td>210.94</td>
<td>12.65</td>
<td>72.87±2.12</td>
<td>175.38</td>
<td>11.61</td>
<td></td>
</tr>
<tr>
<td>SimIT</td>
<td>75.56±2.42</td>
<td>214.22</td>
<td>11.97</td>
<td>83.69±1.63</td>
<td>161.29</td>
<td>7.13</td>
<td></td>
</tr>
</tbody>
</table>

(b) Ultrasound

<table>
<thead>
<tr>
<th>Method</th>
<th>Content</th>
<th>IoU [%] ↑</th>
<th>FID ↓</th>
<th>KID ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation</td>
<td>—</td>
<td>297.24</td>
<td>38.14</td>
<td></td>
</tr>
<tr>
<td>CycleGAN [Zhu et al., 2017b]</td>
<td>2.33±1.10</td>
<td>46.67</td>
<td>2.02</td>
<td></td>
</tr>
<tr>
<td>CUT [Park et al., 2020]</td>
<td>3.39±1.47</td>
<td>46.62</td>
<td>1.63</td>
<td></td>
</tr>
<tr>
<td>ConPres [Tomar et al., 2021a]</td>
<td>5.01±1.95</td>
<td>95.31</td>
<td>7.74</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SimIT-CS</td>
<td>2.97±1.30</td>
<td>56.70</td>
<td>3.57</td>
<td></td>
</tr>
<tr>
<td>SimIT-C</td>
<td>9.50±4.42</td>
<td>46.10</td>
<td>2.33</td>
<td></td>
</tr>
<tr>
<td>SimIT</td>
<td>20.54±6.80</td>
<td>79.02</td>
<td>6.14</td>
<td></td>
</tr>
</tbody>
</table>

(c) Gaming

<table>
<thead>
<tr>
<th>Method</th>
<th>Content</th>
<th>Realism</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mAP ↑</td>
<td>pixAcc ↑</td>
</tr>
<tr>
<td>Simulation</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CycleGAN [Zhu et al., 2017b]</td>
<td>14.12</td>
<td>55.54</td>
</tr>
<tr>
<td>CUT [Park et al., 2020]</td>
<td>12.42</td>
<td>53.10</td>
</tr>
<tr>
<td>ConPres [Tomar et al., 2021a]</td>
<td>15.44</td>
<td>60.33</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SimIT-CS</td>
<td>13.74</td>
<td>56.22</td>
</tr>
<tr>
<td>SimIT-C</td>
<td>20.76</td>
<td>63.12</td>
</tr>
<tr>
<td>SimIT</td>
<td>22.34</td>
<td>68.83</td>
</tr>
</tbody>
</table>
6.2 Results

Figure 6.3: Ultrasound training experiment results: (a) Examples of real ultrasound images with the style of GE-E10 ultrasound machine. (b) Visual examples of images translated using SimIT compared to alternative methods. (c) Qualitative results of SimIT compared to its ablated variants, with bone surfaces visualized in purple and green respectively for those from translated and simulated (reference) images. (d) Probability maps of containing content at each location, averaged over the training images. A big scene distribution difference can be observed between simulated and target images. (e) Quantitative evaluation of structural preservation. To this end, a paired test is employed to compare the IoU scores between the bone maps extracted from the simulated and translated images. The difference is computed by subtracting the score of other models from SimIT, i.e. the larger the positive difference is, the more superior SimIT is with respect to another method. Significance is indicated with respect to our proposed model SimIT (marked with dotted lines) or between different models (|---|), with P-values of $\leq 0.0001$ (****).
and 6.4). Translated ultrasound image examples in Figure 6.3(b) demonstrate that the alternative methods are not always correct with the echogenicity (brightness) profile of different anatomical regions, e.g., outside the uterus is sometimes black and sometimes white, and the same for the amniotic fluid. ConPres preserves the echogenicity better than CycleGAN and CUT by leveraging simulated label-image pairs, however, it is biased towards interpreting input labels as pixel intensities. In comparison, SimIT can retain correct echogenicity of each region, which can be seen by comparing to the reference simulated images, while translating into a realistic appearance in the style of GE-E10 images. Furthermore, our method successfully preserves fine anatomical structures, e.g., the ribs in the top row example, whereas the other compared methods fail with such detail. We herein assess content preservation for ultrasound images based on the alignment of bone surfaces, delineated by a phase-symmetry based bone-surface estimation method [Hacihaliloglu et al., 2009] as exemplified in Figure 6.3(c). Alignment is then quantified by the intersection over union (IoU) score of bone surface pixels extracted from translated and simulated (reference) images (cf. Figure 6.3(e) and Table 6.1(b)). When comparing image realism via FID/KID scores, SimIT does not yield the best values, which is hypothetically caused by the large scene difference between real and simulated training images, as illustrated in Figure 6.3(d). Since the alternative methods do not enforce strict restrictions on content preservation, they hallucinate content in an unconstrained manner, which helps lower their FID/KID scores; nevertheless, such arbitrary content does not match the input label-maps, hence not fit for our purposes, as also quantified by the structural preservation scores, i.e., IoU for the ultrasound training experiment.

For the gaming experiment, similarly to ultrasound experiment above, CUT and CycleGAN largely hallucinate content, ignoring the input GTA label-maps and hence do not satisfy the desired content preservation criterion. For example, as seen in Figure 6.4(b), the sky is erroneously filled with trees, since the real Cityscapes images contain more trees and less sky compared to the simulated GTA dataset [Richter et al., 2022]. Then, the discriminator can easily differentiate between real and fake images by looking at the image top region, which in return encourages the generator to hallucinate more trees in the sky. In comparison to CycleGAN and CUT, the domain-consistent deep features used in our proposed SimIT explicitly encourages content preservation. To evaluate structural consistency, we apply on the translated images the pretrained semantic segmentation network DRN [Yu et al., 2017] following [Park et al., 2020] and report the resulting segmentation metrics: mean average precision (mAP), pixel accuracy (pixAcc), class accuracy (classAcc) – see the Methods for details. SimIT achieves the best scores among all the methods for these content preservation metrics. Evaluating image realism using FID/KID scores (Table 6.1(c)), SimIT outperforms the state-of-the-art, while faring suboptimal compared to its ablated variants, which however in turn fail at successfully retaining content. This well indicates the conflicting nature between content preservation and image realism, especially in the presence of substantial layout differences between simulated and real image sets.

**Image-to-label translation** By introducing the cyclic loss to train our framework with domain-consistent features, at the end of the training we also obtain a generator that can translate real images to label-maps, i.e., a semantic segmenter for real images. Note that such segmenter is trained truly unsupervised, i.e., without requiring any annotation of any real image. To evaluate the segmentation outcome of image-to-label translations
of SimIT, we compare resulting label-maps to semantic segmentations for the datasets where such annotations are available, \textit{i.e.} the CholecSeg8k dataset as the Style-C target for our laparoscopy application and the Cityscapes dataset for our gaming application.

For laparoscopy comparison, we report upper-bound segmentation results from a ResNet50 network trained on annotated images from the CholecSeg8k dataset introduced by [Hong et al., 2020], which is a subset of Cholec data [Twinanda et al., 2016]. In Figure 6.5(a) a sample input image is shown together with its supervised ResNet50 segmentation as upper-bound; the semantic map predicted by SimIT used as a segmenter; and the ground-truth annotation for this test image. Average segmentation scores are reported in Figure 6.5(c). For the gaming application, we compare SimIT with the segmentation network DRN [Yu et al., 2017], a standard technique for the Cityscapes dataset. DRN was trained on the labelled training set, acting as a supervised upper-bound herein. A qualitative sample comparison is shown in Figure 6.5(b) with quantitative results tabulated in Figure 6.5(d). SimIT presents a fair performance despite not having seen any labeled real images and while not specifically targeting such segmentation problem.

For the ultrasound application, SimIT was trained with gray-scale label-maps, as this performed well for the main focus of label-to-image translation and also using one-hot label encoding (performed for the other two applications) was less stable in the parametrization of ultrasound training. Without one-hot labels, our trained network fails to estimate meaningful label-maps from real ultrasound images. This is mainly due to having nearly 80 different tissue classes and some classes with similar ultrasound texture.

---

**Figure 6.4:** Gaming experiment results: (a) Examples of images from Cityscapes dataset. (b) Visual examples of images translated using SimIT compared to alternative methods. (c) Qualitative results of SimIT compared to its ablated variants.
and appearance, which makes the segmentation problem very challenging and, given no one-hot labels, also ill-posed as then a grayscale regression problem.

### 6.3 Discussion

In this work we present a contrastive learning based unpaired image-label translation framework, by leveraging domain-specific simulations to generate photorealistic images from synthetic semantic label-maps. We demonstrate the superior content-preservation performance of our proposed method across several datasets. Our bidirectional framework as a by-product affords an image segmenter, which is demonstrated herein to provide approximate segmentations of real images. Note that such segmentation by our method requires no annotations of real images in training, and it utilizes merely existing computational simulation outputs. As demonstrated in Figures 6.2 to 6.4(b), the unsupervised losses between label and image representations, i.e. the cycle consistency loss [Zhu et al., 2017b] and the contrastive loss [Park et al., 2020], may lead to scene modification and semantic flipping, i.e. the content not being preserved semantically consistently.

To mitigate these, we leverage simulated images as intermediate image representation to bridge the gap between the source and target domains. Among compared methods,
ConPres is the closest to our work, as it also leverages simulated pairs to enforce content preservation. To encourage the generator encoder to extract content-related features, ConPres uses a unified generator for three translation tasks: label-to-image, image-to-image, and image-to-label. This, however, complicates the generator’s task, leading to sub-optimal results in the label-to-image direction. In comparison, we suggest to utilize task-specific generators, relaxing the constraints of each generator. We accordingly leverage simulated images as surrogate targets only for loss computation, not as an additional generation task. During our preliminary experiments, we found out that using pixel-level supervised losses, e.g., L1/L2 loss, to assess content similarity between simulated and translated images is problematic due to the intrinsic appearance shift, lighting variations, and texture differences between two domains. We also experimented with employing the discriminator as a feature extractor for computing feature-level losses, e.g., the perceptual loss [Johnson et al., 2016], but the results were again not satisfactory. In comparison to the above, the currently utilized patch-based contrastive loss is less affected by any appearance shift, and can therefore successfully discern and focus on content dissimilarity. Together with our utilization of adaptive discriminator augmentation and limited discriminator receptive field, the contrasting of image features from the proposed addition of an image-to-label network has endowed results with substantial content preservation without degrading image realism.

The above-mentioned challenge of measuring image content similarity during training also reflects on the evaluation of inference results for structural preservation. For the gaming application experiment, we have employed a pretrained segmentation network to assess the content difference between simulated and translated images. This approach is only feasible when a large amount of annotated images from the target domain are available to train a segmentation network. When such segmentation method is not available, choosing surrogate metrics for quantifying structural similarity is a non-trivial task. Compared to metrics based on pixel-wise difference, SSIM is relatively less sensitive to appearance shifts. Thus, we used SSIM to capture structural differences between laparoscopic images. However, SSIM is less suitable for ultrasound images due to highly directional artifacts and the inherent speckle noise. In ultrasound, bone surfaces are major anatomical landmarks which appear consistently as hyperechoic bands due to their relatively higher acoustic impedance [Ozdemir et al., 2020]. We thus use bone surfaces extracted from simulated and translated ultrasound images using a phase-symmetry based method (described further in the Methods) for assessing structural preservation in ultrasound.

As noted from the qualitative results shown in Figure 6.2-6.4, SimIT is not capable of recovering image features, which are not encoded in label-maps, such as lighting and depth information in laparoscopy and gaming applications, and the acoustic shadows in the ultrasound application. Auxiliary scene information, e.g., geometry and material information from the simulation, can be potentially integrated into SimIT as additional inputs, which can be a potential future direction to further improve content preservation.

The proposed method can have further uses than simulated training, such as synthetic image generation with paired annotations for other applications [Pinaya et al., 2022; Sandfort et al., 2019].
6.4 Methods

Herein we use the notation $X^z_y$ to represent the domain of any sample, where $X$ is the representation from \{L: label-map, I: image\}, and $y$ and $z$ are, respectively, the style (appearance) and content of such representation from \{S: simulated, R: real\}. We aim to learn a generator $G : L^S \rightarrow I^R$ which maps a simulated label-map $L^S \in L^S$ to a real-appearing image $I^R \in I^R$ while preserving the simulation-consistent semantic content, i.e. $(\cdot)^S$.

Generator $G$ is divided into an encoder $G_E$ for extracting content-related features and a decoder $G_D$ for generating target appearance. It is possible to collect many real image examples $\{I^R \in I^R\}$ and also to simulate label-image pairs $(L^S, I^S)$, but paired data of the intended source-target translation, i.e. $(L^S, I^R)$, is inexistent and very challenging to procure. The unpaired data described above does not allow for direct supervision in learning $G$. Existing unpaired methods often change both content and style together, and the ones that aim content preservation only targets image-to-image translation, with methods we show herein not to simply extend to label-to-image translation targeted herein. An overview of the methods can be followed in Figure 6.6.

Generative adversarial training For learning a generator $G$ and its discriminator $D_I$ differentiating images $I$ as real or fake, a non-saturating GAN loss with the R1 regularization [Mescheder et al., 2018] is used, i.e.:

$$
L^{\text{GAN}}(\{L^S\}, \{I^R\}) = \mathbb{E}_I[\log(D_I(I^R) - 1)] + \mathbb{E}_L[\log D_I(G(L^S))] + \frac{\gamma_I}{2} \mathbb{E}_I[\|\nabla D_I(I^R)\|^2]
$$

(6.1)

with the regularization parameter $\gamma_I$.

Label-to-image translation guided by simulation Herein we propose to leverage information from simulations to achieve semantic preservation while learning $G$. To that end, we utilize simulated (synthetic) images $\{I^S \in I^S\}$ during training, which can have paired input label-maps $\{L^S \in L^S\}$ generated from the existing simulation framework (Figure 6.1(a)) and available for training. We encourage scene-consistent translations using a contrastive loss [Park et al., 2020] on image patches, where corresponding patches from the source and translated images (positive pairs) are brought closer in a learned feature space. This space is defined as a projection of the manifold learned by the generator encoder, as illustrated in Figure 6.6(b). Meanwhile non-corresponding (arbitrary) patches are treated as negative samples and hence pushed farther apart in that feature space. Compared to the pixel-wise supervised losses, contrastive loss is known to be less affected by image appearance. It was utilized in [Park et al., 2020] for unpaired image-to-image translation, i.e. when both source and target are of the same representation being in image domain $I$. However, for label-to-image translation, the source and target representations differ, i.e. while each pixel in $L$ denotes a label, each pixel in $I$ denotes a color. Thus, directly contrasting label and image features cannot successfully guide the network for the intended task; as also seen with the suboptimal performance of our ablation variant SimIT-CS. To alleviate this problem, herein we leverage available simulated images $I^S \in I^S$ as “surrogate” source images. This implicitly enables the use of existing simulated images $I^S$.) Note that these images that require complex rendering
6.4 METHODS

Figure 6.6: (a) Schematic overview of our proposed method SimIT with (b) an illustration of contrastive loss. (c) Schematic overview of the ablated version SimIT-C, and (d) SimIT-CS. (e) Schematics of the generator architecture. The number below each convolutional block indicates the channel number. For $G$ the input channel number is the number of classes (#class) and the output channel number is the number of image channels (3 for RGB images), for $F$ vice versa.
operations are used for loss computation during the training of our method, so they are not needed during inference. This is in contrast to the earlier works [Tomar et al., 2021a; Richter et al., 2022] where the rendered images are used as an input to the translation network and are thus complex rendering operations are still required during inference in real-time.

**Bidirectional label-image translation framework** To extract domain-consistent feature representations, we propose to employ an additional generator \( F : \mathbb{R}^{R} \rightarrow \mathbb{R}^{S} \) with \( F = [F^E, F^D] \) consisting of an encoder \( F^E \) and a decoder \( F^D \), acting in the opposite direction for translating \( I \rightarrow L \), i.e. mapping an image back to a label-map. Unlike [Park et al., 2020] which contrasts features from \( G^E \) operating on the source domain \( \mathbb{I} \) with labels, we propose to contrast the features of the segmenter encoder \( F^E \) trained to extract features for inferring semantic content from images, and it is thus more suited for comparing image similarity.

**Patch-based contrastive loss** For an input image \( x \) and its translated image \( \hat{x} = G^D(G^E(x)) \), we contrast feature maps \( z^F_{j} = H^F(F^E(x)) \) and \( z^F_{\hat{j}} = H^F(F^E(\hat{x})) \), with a light-weight projection head \( H^F \) mapping the features of \( j \)-th hidden layer in \( F^E \). Given \( z^F_{j,s} \) as a query feature at a given spatial location \( s \) within the feature map \( z^F_{j} \) of the translated image, the corresponding input feature \( z^F_{j,s+} \) at the same location (denoted by +) is considered as a positive sample, while any other input feature \( z^F_{j,s-} \) at arbitrary locations act as negative samples. Noise contrastive loss (NCE) [Park et al., 2020] for the \( j \)-th layer feature can then be computed as the sum of contrastive loss for several \( S_j \) at randomly sampled spatial locations as follows:

\[
L_{\text{NCE},j}(z^F_{j,s}, z^F_{\hat{j}}) = \sum_{s=1}^{S_j} L_{\text{CE}}(z^F_{j,s}, z^F_{j,s+}, z^F_{j,s-})
\]

with

\[
L_{\text{CE}}(z, z^+, z^-) = -\log \left[ \frac{\exp(d(z, z^+)/\tau)}{\exp(d(z, z^+)/\tau) + \sum_{-} \exp(d(z, z^-)/\tau)} \right],
\]

where \( z \) is the feature vector of query; \( z^+ \) and \( z^- \) are the feature vectors of positive and negative samples, respectively; \( d(z_1, z_2) \) is a distance metric between two latent vectors (herein the cosine distance); and \( \tau \) is a temperature parameter controlling the smoothing of joint likelihoods.

We herein propose to leverage the simulated image domain as surrogate source domain by computing contrastive loss between the translated image \( I^S_R = G^D(G^E(I^S_S)) \) and simulated image \( I^S_S \) paired to \( I^S_S \), i.e. :

\[
L_{CL}^F(\{I^S_S\}) = \mathbb{E}_L \sum_{j=1}^{J} L_{\text{NCE},j}(H^F_j(F^E(I^S_S)), H^F_j(F^E(I^R_S)))
\]

computed over \( J \) feature layers to contrast information at different resolutions.

For a more expressive and distinctive label space, we herein encode label-maps as one-hot representations, which prevents misinterpretation of categorical labels suitable for class separation as pixel intensities for regression.
Learning image-to-label translation using cycle consistency loss  In this work, we treat and train the image-to-label translator $F$ to perform pixel-wise semantic labeling task, known as image segmentation. Based on our cyclic framework, we use the existing label-to-image mapping network $G$ to assess such segmentation accuracy, hence also obviating a need for pixel-wise annotations of real images, which are difficult to procure. With that, we compute a cycle reconstruction loss between a real image $I^R$ and the image reconstructed from the predicted label-map $\tilde{I}^R = G(F(I^R))$. Image similarity is measured using the NCE loss, as it is less sensitive to appearance shifts, as follows

$$L_{\text{CYC}}(\{I^R\}) = E_L \sum_{j=1}^J L_{\text{NCE},j}(H^F_j(F(I^R)), H^F_j(F(\tilde{I}^R)))$$  \hspace{1cm} (6.5)$$

with the projection head $H^F_j$ for the $j$-th layer feature of encoder $F$. For $F$ and its discriminator $D_L$ for the label representation direction, we employ a GAN training objective $L_{\text{GAN}}^F(\{I^R\}, \{L^S\})$ similar to the original direction $G$, but with a different regularization parameter $\gamma_L$.

Training objective  A schematic illustration of the proposed method SimIT summarizing the above components is shown in Figure 6.6(a). Network training is performed by alternately optimizing the following two losses:

$$L_G(\{L^S\}, \{I^R\}) = L_{\text{GAN}}^G(L^S, I^R) + \lambda_G \cdot L_{\text{CL}}^G(L^S)$$  \hspace{1cm} (6.6)$$

$$L_F(\{L^S\}, \{I^R\}) = L_{\text{GAN}}^F(L^S, I^R) + \lambda_F \cdot L_{\text{CYC}}(I^R)$$  \hspace{1cm} (6.7)$$

with the loss weighting parameters $\lambda_G$ and $\lambda_F$.

We compare our full model against the ablated variant SimIT-C by excluding the inverse mapping $L_F$ with the cyclic loss, as seen in Figure 6.6(c). Further ablating the paired simulation images yield the variant SimIT-CS that instead uses the labels for contrast- ing (Figure 6.6(d)). For both ablated variants, encoder $G^E$ is used for computing the contrastive loss $L_{\text{CL}}^G$.

Network architecture  We build our method on the StyleGAN2 framework [Karras et al., 2020a] for adversarial training. We accordingly use a ResNet-based generator architecture [Park et al., 2020] with four down- and up-sampling layers and 6 residual blocks (Figure 6.6(d)). We use skip connections between the down- and upsampling layers to avoid information loss. For the image synthesis decoder $G^E$ we use weight demodulation [Karras et al., 2020a]

$$w''_{ijk} = \frac{w'_{ijk}}{\sqrt{\sum_{i,k} w'_{ijk}^2 + \epsilon}} \text{ with } w'_{ijk} = s_i \cdot w_{ijk}$$  \hspace{1cm} (6.8)$$

where $w_{ijk}$ is the convolution weight from the $i$-th input feature map to the $j$-th output feature map; $k$ denotes the spatial footprint of the convolution; and the multiplier $s_i$ is set to 1. To provide stochastic means for texture synthesis, especially important to generate the noisy speckle patterns of ultrasound images, we perturb each feature layer with an additive Gaussian (noise) image scaled by learned weights following [Karras et al., 2019]. The output layer for $G$ and $F$ is linear and sigmoid, respectively. We use ReLU activation
for all intermediate layers. For both $D_I$ and $D_L$, we adopt the feedforward discriminator architecture in [Karras et al., 2020a]. In training we use randomly cropped image patches, which enables the discriminator to ignore global scene differences between simulation and real domains.

**Experimental data utilization** *LaparoSim* consists of 1850 synthetic laparoscopic image-label pairs simulated from a 3D abdominal model. We randomly split this data into train-validation-test sets with 80-10-10% ratio. *Style-C* consist of 2100 images from the Cholec dataset. We excluded all frames with surgical tools, since surgical tools are not handled in our simulation. For Style-C testing, we used the 213 frames that has ground-truth labels provided in [Hong et al., 2020], and the remaining frames were used as Style-C training data. *Style-H* consists of 2262 frames in total, which was randomly split in 80-20% ratio, respectively, for training and testing. Some Style-H images had major blurring artifacts due to camera motion, so we manually removed any blurry frames, since we treat the frames separately without any temporal information and such temporal effect is also not represented in the label-maps. All the images were resized to $256 \times 432$.

*USSim* consists of 6669 simulated image-label pairs, which we resized to $256 \times 354$ and randomly split into training-validation-test sets with 80-10-10% ratio. *GE-E10 style* consists of 2328 ultrasound images from 24 patients. We randomly selected images from 20 patients for training and 4 for testing, resulting in 1902 training images and 426 test images.

*GTA* dataset [Richter et al., 2016] contains 24966 image-label pairs. We followed its original train-validation-test split. *Cityscapes* dataset [Cordts et al., 2016] contains 3475 image-label pairs of street scenes from German cities. We used its original training set for our network training, and its validation set with ground-truth labels for testing our semantic segmentation output. As in [Park et al., 2020], we resized all the images to $256 \times 256$.

**Implementation** We implemented our method in PyTorch [Paszke et al., 2019]. We used Adam [Kingma et al., 2014] optimizer with parameters $\beta = (0, 0.99)$ and a learning rate of $10^{-3}$ for $G$ and $10^{-4}$ for $F$. We applied adaptive discriminator augmentation using its default hyperparameters [Karras et al., 2020b]. The generator is trained on image patches of size $256 \times 256$ while the discriminator receptive field is $64 \times 64$. Our network training involves alternating updates of $G$ and $F$. We trained our models for 400 epochs. To compute the contrastive loss, we extract features from the four (stride-2) convolution layers in the encoder at 256 locations randomly selected for each mini-batch. We use a two-layer MLP with 256 units at each layer and ReLU activation for $H^G$, and the identity mapping for $H^F$. NCE temperature parameter $\tau$ is set to 0.07 following [Park et al., 2020]. Generator loss weighting $\lambda_G$ is set to 5 for all the experiments. $\lambda_F$ is set to 1 for the laparoscopy and ultrasound, and 0.5 for the gaming experiment. R1 regularization parameters $\gamma_I$ and $\gamma_L$ are set to 0.01 and 1.0, respectively, for all the experiments. For all compared methods we used their public implementations provided by the corresponding authors with their default hyperparameters.

**Evaluation metrics** We use the following for quantitative evaluation:

- **Image realism.** Fréchet inception distance (FID) [Heusel et al., 2017] is common for assessing the quality of images generated by GANs, by comparing the feature
distribution between two sets of images, herein real and translated, using feature vectors of an ImageNet-pretrained Inception network. Kernel inception distance (KID) [Binkowski et al., 2018] is an alternative metric to evaluate GAN performance. KID is computed as the squared maximum mean-discrepancy between the features of Inception network. KID is then not biased by the number of samples used, unlike FID.

- **Content preservation.** For laparoscopy images, content preservation is assessed using structural similarity between simulated and translated images, quantified via *Structural similarity index* (SSIM) computed as

\[
\text{SSIM}(A, B) = \frac{(2\mu_A \mu_B + c_1)(2\sigma_{AB} + c_2)}{(\mu_A^2 + \mu_B^2 + c_1)(\sigma_A^2 + \sigma_B^2 + c_2)}
\]

with regularization constants \(c_1\) and \(c_2\), local intensity means \(\mu_A\) and \(\mu_B\), local standard deviations \(\sigma_A\) and \(\sigma_B\), and cross covariance \(\sigma_{AB}\). To compute this metric, we used the python package *scikit-image* with its default parameters. For ultrasound images, due to potential artifacts, typical speckle noise, and a lack of sharp edges, we instead used the similarity of bone surfaces for assessing structure preservation. To that end, we extracted bone surfaces from each image using [Hacihaliloglu et al., 2009]. This method is based on local phase symmetry in B-mode images, and operates by aggregating images filtered by log-Gabor kernels with different orientations \(r\) and scales \(m\) defined as

\[
G_{r,m}(\omega, \phi) = \exp \left( -\frac{\log(\omega / \omega_0)^2}{2 \log(\kappa_m / \omega_0)^2} - \frac{(\phi - \phi_r)^2}{2 \sigma_{\phi}^2} \right),
\]

where parameters \(\phi_r, \omega_0, \kappa_m, \) and \(\sigma_{\phi}\) define the filter orientation, center frequency, scaling factor, and angular bandwidth of the employed filters, respectively. Following [Ozdemir et al., 2020], we set \(\kappa_m / \omega_0 = 0.25\) and \(\phi_r = \left[ \frac{1}{6} \pi, \frac{3}{6} \pi, \frac{5}{6} \pi \right]\). To assess preservation, we report *intersection over union* (IoU) of pixels belonging to bone surfaces extracted from corresponding simulated and translated images. We exclude from computations the top 25 pixels of images corresponding to skin reflections.

- **Segmentation.** For the laparoscopic CholecSeg8k dataset, we trained a semantic segmentation network with ResNet50 architecture initialized with weights pretrained on the ImageNet using the pytorch segmentation library [Iakubovskii, 2019], following the training settings from a public implementation on the Kaggle repository of this dataset. We randomly picked video24 for validation, video{09,17,26,28,43} for testing, and the rest for training. We report F1 score for six classes that are also in our simulation. For the Cityscapes dataset in the gaming application, we trained a segmentation network suggested for this dataset in [Park et al., 2020], with the DRN-D22 architecture [Yu et al., 2017] at 256 × 128 resolution with the default parameters from its public implementation. Following [Park et al., 2020], we report Cityscapes semantic segmentation results using *mean average precision* (mAP) over the classes; *pixel-wise accuracy* (pixAcc) as the percentage of correctly classified pixels; and *average class accuracy* (classAcc) over given classes.

**Supplementary information.** Additional example images from image-to-label translation are presented in Figure 6.7 and from label-to-image translation in Figure 6.8.
**Figure 6.7:** Additional qualitative label-to-image translation results for the laparoscopic simulation application Style-C (top three rows) and Style-H (next three rows), the ultrasound simulation application (next three rows), and the gaming application (last three rows).
Figure 6.8: Additional image-to-label translation results of SimIT with the input real image $I^R_R$, translated result $F(I^R_R)$, reconstructed image $G(F(I^R_R))$, and ground truth label-maps for reference; for (a) in-house laparoscopic, (b) CholecSeg8k, and (c) Cityscapes images. No ground-truth annotation is available for the dataset in (a). Note that although the reconstructed images are generated from translated images, they are ordered in-reverse to place the same representations with the input and GT next to each other for easier visual comparison.
7

Summary and Perspectives

In this thesis, several deep learning approaches have been proposed for improving ray-based ultrasound simulation, with the discussion of each proposed method already included in each chapter. This final chapter provides an overview of key contributions and important findings, and highlights potential future research directions.

7.1 Summary of Contributions

In Chapter 2, a learning-based method for the estimation of parameters for tissue scatterer distributions is demonstrated. We estimated the Gaussian distribution parameters, making the assumption that scatterers follow a normal distribution in the tissues. In comparison to classical methods, the scatterer representations estimated by the proposed method with varying acquisition configurations have been shown to be robust for simulating images, such as being insensitive to compression and rotation of the imaged domain. Despite being trained on synthetic images, the proposed method can generalized well to unseen in-vivo images, and enables fast scatterer estimation in milliseconds. Being fully convolutional, it can be run on any image size, and can be easily extended to 3D volumes. Such estimated scatterer parameters may also be used as a diagnostic ultrasound biomarker for tissue characterization.

A GAN-based approach for simulation quality enhancement is presented in Chapter 3. An image-to-image translation framework based on adversarial training is utilized to translate computationally low-cost simulated images to high-quality simulated images. This helps avoid runtime computational cost for high-quality simulation. Leveraging supportive contextual information is found to be beneficial for such translation task: We have utilized label maps for better semantic control, and acoustic shadow maps, which are obtained by integrating tissue attenuation coefficients along the propagation path, for facilitating the generation of global directional artefacts.

Chapter 4 extends the above idea to generating high-quality simulated images directly from semantic label and attenuation maps, which obviates any complex rendering process in real-time. An image translation framework has been devised similarly to Chapter 3 but with a more dedicated generator design with texture-friendly convolution blocks, stochastic noise, and label-map input to intermediate layers. We have shown this to perform superior in preserving local semantics and in successfully generating US speckle textures. To evaluate the similarity between two US images, metrics based on pixel-wise
differences are sub-optimal given the inherent noisy speckle appearance of B-mode images. To assess local texture matching, an error metric based on local histogram statistics has been proposed herein and demonstrated to be capable of capturing visually-perceived local differences between US images.

While Chapters 3 and 4 demonstrate the feasibility of using GAN-based image translation approaches for accelerating US simulations, Chapter 5 explores image translation techniques for enhancing simulation realism, to generate in-vivo-like images. Unlike the above settings, this is an unpaired problem since real images with matching simulated counterparts are not available, which makes the translation problem under-constrained. For generating realistic-looking images from simulated images, regularizing the generator with contrastive learning and multi-domain translation have been proposed in Chapter 5 as being effective for content preservation while translating target appearance successfully. Given the multi-domain nature of the proposed method, image translation between the other domains can be performed without any further training, e.g., label-to-real, label-to-simulated, simulated-to-real image translation.

To tackle the challenging label-to-image translation, a bidirectional translation framework is presented in Chapter 6. Existing unpaired image translation methods focus on image-to-image translation, and perform sub-optimally for label-to-image translation due to the large representation gap between the source (label maps) and target (images) domain. To mitigate this, we leverage simulated images as intermediate representation during training for loss computation. Such simulated images nevertheless are not needed during the real-time inference stage, thus obviating any complex rendering in simulation runtime. Contrastive loss is utilized to compare similarity between images from two different domains, which is less affected by the appearance shift due to domain gap. This enables realistic image generation directly from simulated label-maps. Employing features extracted from an extended inverse translation network endows substantial improvements in content preservation. Furthermore, the inverse network learns semantic image segmentation without seeing any annotations of real images. Further experiments on other simulation applications show the broad applicability of the proposed method.

7.2 Future Perspectives

While addressing the issues of ray-based US simulations with deep learning approaches, the findings in this thesis have inspired a number of interesting future research directions:

- The learning-based scatterer estimation demonstrated in Chapter 2 can be extended with more sophisticated statistical models, which may better represent actual scatterer distributions.
- The proposed image translation methods in Chapter 3-6 aim to translate individual images. Even if consecutive frames may look visually coherent, without understanding temporal dynamics and explicitly enforcing this, these methods are not guaranteed to generate coherent speckle motion forgiven probe motion. Temporal continuity of simulated frames is important in an interactive and dynamic simulation framework, but has not been studied in the literature. In the future, our proposed solutions can be analyzed whether they satisfy temporal continuity,
and if not, this problem can be tackled by exploring different techniques such as spatio-temporal adversarial objectives to leverage temporal information from image sequences.

- FID/KID metrics have several shortcomings in evaluating image realism. In the future, better metrics for assessing realism are direly needed, in particular for ultrasound images. Classical GAN evaluation metrics rely on image features typically extracted from off-the-shelf feature networks trained for classifying natural images. Such feature networks are sub-optimal for extracting US-related features. Developing US specific feature extractors is an important stepping stone in the research areas related to deep learning.

- The proposed method in Chapter 6 can segment real images as a free by-product without seeing any annotations of real images. This opens future research directions to obviate annotation efforts on US imaging datasets.

- It is possible to adopt the simulation-to-real image translation methods for unsupervised parametric ultrasound image reconstruction, where annotated data is rarely available. Classical end-to-end learning models, trained in a supervised manner using synthetic data, rarely translate successfully to unseen clinical data, due to the large difference between the training and test domain. This has thus a good potential to improve the generalization ability by removing the appearance shift based on the unsupervised image translation concepts proposed in this thesis.
Bibliography


List of Publications

Articles in peer-reviewed journals:


Conference contributions:


Preprints:


* These authors contributed equally