Simplifying Interpretation of Ultrasound Imaging: Deep Learning Approaches for Phase Aberration Correction and Automatic Segmentation

Mostafa Sharifzadeh

A Thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Electrical and Computer Engineering) at
Concordia University
Montréal, Québec, Canada

October 2024

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **Mostafa Sharifzadeh**

Entitled: **Simplifying Interpretation of Ultrasound Imaging: Deep Learning Approaches for Phase Aberration Correction and Automatic Segmentation**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Farah Hafeez

_____ External Examiner
Dr. Roger Zemp

_____ Arms-Length Examiner
Dr. William Lynch

_____ Examiner
Dr. Arash Mohammadi

_____ Examiner
Dr. Wei-Ping Zhu

_____ Thesis Co-supervisor
Dr. Hassan Rivaz

_____ Thesis Co-supervisor
Dr. Habib Benali

Approved by _____
Dr. Jun Cai, Graduate Program Director

November 29, 2024 _____
Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Simplifying Interpretation of Ultrasound Imaging: Deep Learning Approaches for Phase Aberration Correction and Automatic Segmentation

Mostafa Sharifzadeh, Ph.D.
Concordia University, 2024

Medical ultrasound imaging is a widely used diagnostic tool in clinical practice, offering several advantages, including high temporal resolution, non-invasiveness, cost-effectiveness, and portability. Despite these benefits, ultrasound modality often suffers from lower image quality compared to other modalities, such as magnetic resonance imaging, which complicates image interpretation and poses diagnostic challenges, even for experienced clinicians. Given its unique advantages, simplifying the interpretation of ultrasound images can profoundly impact the accessibility and affordability of healthcare. This thesis aims to enhance the interpretability of ultrasound images using deep learning (DL)-based approaches on two parallel fronts.

The first front focuses on improving image quality by addressing the phase aberration effect, a primary contributor to the degradation of medical ultrasound images. Phase aberration arises from spatial variations in sound speed within heterogeneous media, introducing artifacts such as blurring and geometric distortions. This effect hinders the accurate representation of tissue structures and complicates clinical interpretation. To tackle this, we propose two novel methods. The first involves training a convolutional neural network (CNN) to estimate the aberration profile from the B-mode image and employing it to compensate for the aberration effects. The second introduces an aberration-to-aberration approach combined with an innovative loss function to train a CNN that directly predicts corrected radio frequency data without requiring ground truth.

The second front focuses on the automatic segmentation of ultrasound images and explores the challenges associated with employing DL-based approaches. Manual segmentation, typically performed by expert clinicians, is time-consuming and prone to human error, and automating this process can simplify the interpretation of ultrasound images. While DL methods have demonstrated considerable potential, ultrasound image segmentation poses unique challenges due to artifacts such as shadowing, reverberation, refraction,

phase aberration, and speckle noise. The scarcity of medical data further complicates these challenges, limiting the generalizability and robustness of models in clinical settings. To address these limitations, we investigate the shift-variance problem in CNNs and propose pyramidal blur-pooling layers to mitigate this issue. Furthermore, we tackle domain shift and data scarcity by employing a domain adaptation method and introducing an ultra-fast ultrasound image simulation technique based on frequency domain analysis.

Acknowledgments

If you have a great supervisor during your Ph.D., you are a lucky Ph.D. student. If you have two, you are fortunate. And if you have a loving family by your side as well, you stand among the truly blessed.

I would like to dedicate this thesis to my supervisors, Pr. Hassan Rivaz and Pr. Habib Benali, who are not only great mentors but also kind-hearted individuals. And, of course, to my dear parents and my dear sisters.

Contents

List of Figures

List of Tables

List of Abbreviations

AR autoregressive

CNR contrast-to-noise ratio

CNN convolutional neural network

DAS delay-and-sum

DL deep learning

DNN deep neural network

DSC Dice similarity coefficient

FDA Fourier domain adaptation

FFT fast Fourier transform

FLAIR fluid-attenuated inversion recovery

FWHM full width at half maximum

FXPF frequency-space prediction filtering

GAN generative adversarial network

gCNR generalized contrast-to-noise ratio

IFFT inverse fast Fourier transform

IQ in-phase and quadrature

IVUS intravascular ultrasound

MAE mean absolute error

MAIN-AAA mitigating aberration-induced noise: aberration-to-aberration approach

MSE mean square error

MRI magnetic resonance imaging

MTL multi-task learning

NCC normalized cross-correlation

NNCC nearest-neighbor cross-correlation

PBP pyramidal blur-pooling

PICMUS plane-wave imaging challenge in medical ultrasound

PSF point spread function

RF radio frequency

RMS root mean square

SNR signal-to-noise ratio

SOTA state-of-the-art

Chapter 1

Introduction

1.1 Ultrasound Imaging

Medical ultrasound is one of the most widely used tools in clinical practice across various applications, including therapy [1, 2], diagnostics [3, 4], and image-guided interventions [5]. Additionally, it is often integrated with other modalities, such as photoacoustic techniques [6]. The widespread adoption of ultrasound originates from several advantages, including real-time imaging, non-invasiveness, cost-effectiveness, and portability. The ultrasound modality operates based on acoustic oscillation, with frequencies exceeding 20 kHz, which is the upper limit of human hearing and is classified as ultrasound. In ultrasound imaging, a piezoelectric crystal is excited electrically by a voltage pulse, causing it to vibrate and generate acoustic waves. These waves propagate through a medium, and when they encounter any discontinuities in mechanical characteristics along their path, a portion of their energy is reflected back. These reflections, known as echoes, cause the piezoelectric crystal to vibrate again and generate electronic signals, which are recorded. These recorded signals, known as radio frequency (RF) data, are processed to reconstruct a representation of the medium. The recorded data can be displayed in various formats, including A-mode, B-mode, C-scan, and M-mode [7]:

In A-mode, the amplitude of the echo signal is plotted along the vertical axis versus time on the horizontal axis.

In B-mode, which is also known as brightness modulation and is the most common method for displaying ultrasound images, the amplitude of the echo signal is displayed as brightness along a line that represents time. Converting RF data to B-mode involves envelope detection, typically followed by applying log compression.

In C-scan, a 2D planar representation from a constant depth is obtained by moving the

transducer mechanically over the scanning area. C-scan is primarily used as part of three-dimensional ultrasound imaging [8].

In M-mode, where M stands for motion, the speed and displacement of moving structures are primarily captured. In this mode, B-mode scans are obtained in a constant direction and stacked along the vertical axis over time to construct a 2D image.

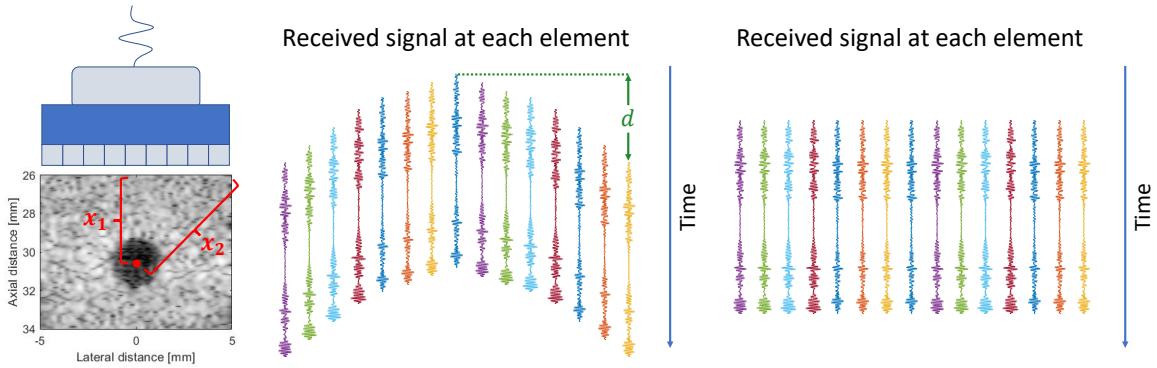


Figure 1.1: Delay-and-sum beamforming in ultrasound imaging. (a) Ultrasound image of a cyst phantom, with the red dot indicating the pixel intended for reconstruction from the reflected waves. Echo signals received at each transducer element are shown (b) before and (c) after applying delay compensation.

In practice, a transducer, which includes a batch of piezoelectric crystals, transmits ultrasound waves of desired shapes and records the received echo signals. Considering the B-mode case as an example, the goal is to translate the echo data received by the transducer into brightness and reconstruct an image from the reflected waves. Given a transducer imaging a medium, consider the red dot in Fig. 1.1 (a) as the pixel whose reconstruction is desired. The reflected waves must travel different distances from that point to reach each transducer element, depending on the transducer's geometry. Therefore, as shown in Fig. 1.1 (b), each element experiences a different delay for the received signal, which can be calculated based on the geometrical distance and the speed of sound. For instance, according to Fig. 1.1 (a), the rightmost element experiences a delay d_1:

$$d_1 = \frac{x_2 - x_1}{c} \tag{1.1}$$

where x_1 is the distance from the reference element, x_2 is the distance from the rightmost element, and c is the speed of sound. As shown in Fig. 1.1 (c), the received signals can be compensated for the delays d_1 to d_n. Finally, the amplitude of the echo signals corresponding to the indicated red dot in Fig. 1.1 (a) can be determined by summing the corresponding

amplitudes from the delayed signals received at all elements in Fig. 1.1 (c). This method is known as delay-and-sum (DAS) beamforming.

1.2 Phase Aberration Correction

While medical ultrasound is one of the most widely used tools in clinical practice, it often suffers from artifacts such as the phase aberration effect, which complicates clinical interpretation and compromises the diagnostic accuracy of examinations. Phase aberration, one of the main sources of image quality degradation [9], is caused by spatially varying sound speed while traveling through a heterogeneous medium [10]. In a perfectly homogeneous medium, like the one shown in Fig. 1.1, the sound speed is known and remains constant. Consequently, the travel time of a pulse from any transducer element to any point in the medium can be calculated using basic geometric principles, as outlined in Eq. (1.1). Therefore, the required time delays that need to be applied to each element can be determined accurately to compensate for traveling path length differences and form the desired beam in transmit beamforming. Similarly, in receive beamforming, time delays can be calculated and applied to received echo signals in order to sum them coherently. In practice, however, the human body is a heterogeneous medium, where, for instance, the sound speed in fat and muscle is approximately 1460 m/s and 1610 m/s, respectively, which indicates a variation of almost 10% [11]. The variation is even higher in applications such as transcranial imaging [12], where the average sound speed in the skull is nearly 2740 m/s [13]. The phase aberration effect in a heterogeneous medium alters the focal point in focused imaging and perturbs the flat wavefront propagation in plane-wave imaging during the transmission, and prevents coherent summation of echo signals in both imaging techniques during the reception, all of which cause suboptimal image quality.

Aberration correction has been studied for years in the medical ultrasound community, as it can enhance the interpretability of images by improving anatomical fidelity and spatial localization, both of which lead to improved diagnostic accuracy and precision in image-guided interventions. Several techniques attempted to estimate delay errors by maximizing the cross-correlation [10] or minimizing the absolute differences between RF signals received at adjacent array elements [14], maximizing mean speckle brightness in a region of interest [15], or incorporating a virtual point reflector generated by iterative time reversal focusing [16]. Li *et al.* utilized the generalized coherence factor for reducing focusing errors, especially the ones caused by sound speed inhomogeneities [17]. Napolitano *et al.* analyzed lateral spatial frequency content in reconstructed images to find the optimal

sound speed for subsequent imaging that maximizes the focus quality [18]. Shin *et al.* employed a technique called frequency-space prediction filtering (FXPF), which presupposes the existence of an autoregressive (AR) model across the echo signals received at the transducer elements and removes any components that do not conform to the established model [19, 20].

As opposed to methods that model the phase aberration effect by a fixed near-field phase screen in front of the transducer, the locally adaptive phase aberration correction technique [21] assumed a spatially varying near-field phase screen and employed multistatic synthetic aperture data to perform the correction at each point adaptively. Lambert *et al.* suggested compensating for the spatially-distributed aberrations by decoupling aberrations undergone by the outgoing and incoming waves utilizing the distortion matrix built from the focused reflection matrix, which contains the responses between virtual transducers synthesized from the transmitted and received focal spots [22, 23, 24].

A different category of techniques utilizes echo signals as input and returns an estimation of the spatial distribution of sound speed in a given medium [25, 26]. Although these methods are not an immediate approach for aberration correction, the estimated distribution can be subsequently employed to compensate for the phase aberration effect, for instance, by reconstructing the image by computing beamforming delays assuming that sound travels on straight line paths [27] or using a set of refraction-corrected delays based on the Eikonal equation [28], which can be efficiently solved using the fast marching method [29]. The computed ultrasound tomography in echo mode (CUTE) method correlated the phase shifts across a sequence of beamformed plane-wave images obtained with different steering angles and exploited that to estimate the distribution of sound speed [30]. Jakovljevic *et al.* proposed and solved a model via gradient descent that establishes a connection between the local speed of sound along a wave propagation path and the average speed of sound over that path [31], where the latter is measured using the method proposed in [32]. Although the efficacy of this model was demonstrated in layered heterogeneous media, the performance often drops when the variations of sound speed are not insignificant along the lateral axis. Rehman *et al.* introduced a tomography-based method that directly accounts for propagation paths between the scattering volume and each transducer element to mitigate that issue [33]. They also proposed an inverse-modeled phase aberration computed tomography (IMPACT) framework, which utilizes multistatic synthetic aperture data, estimates the global average sound speed [34] by maximizing coherence for each point, applies an inversion to compute the local sound speed, and finally exploits them in two different Eikonal equation-based and wavefield correlation-based distributed aberration correction

techniques [35]. In addition to approaches aimed at rectifying the aberrated image, certain beamformers are specifically designed to remain robust against this artifact. These beamformers leverage singular value decomposition applied to matrices constructed either from aperture data in focused imaging [36] or from a matrix of backscattered data derived from multiple transmissions in plane-wave imaging [37].

Recently, utilizing deep learning (DL)-based techniques for phase aberration correction has attracted growing interest. Feigin *et al.* simulated a dataset using the k-Wave toolbox [38], wherein the organs in tissue were modeled as uniform ellipses over a homogeneous background with different sound speeds. They trained a convolutional neural network (CNN) on the dataset to estimate sound speed distribution, taking raw RF channel data of three plane-wave transmissions as inputs [39]. In a similar approach, demodulated in-phase and quadrature (IQ) data were provided to the network as the inputs [40]. Additional comparable methodologies have been proposed in the literature [41, 42] for the same purpose. Koike *et al.* trained a network by mapping aberrated RF inputs to their corresponding aberration-free RF target outputs [43]. Shen *et al.* utilized a CNN to estimate the aberrated point spread function (PSF) from beamformed IQ data and subsequently applied the inverse filter to rectify the data [44]. Additionally, there are DL-based beamformers designed to exhibit robustness to the aberration by suppressing off-axis scattering [45] or by mapping images beamformed with randomly perturbed sound speed values to clean images beamformed with a reference sound speed value [46].

1.3 Automatic Segmentation

In various applications, segmenting regions of interest within an ultrasound image enhances its interpretability compared to analyzing the raw image by isolating key anatomical structures or pathological areas. Image segmentation entails pixel-level labelling of images to obtain a representation of data that is more meaningful and easier to analyze for a specific purpose and is a crucial task in numerous applications such as registration [47], image-guided biopsy and therapy [48], automatic staging of stenoses in intravascular ultrasound (IVUS) diagnosis [49], detection of vessel boundaries to monitor cardiovascular diseases [50], and delineation of the cardiac structures [51, 52].

Manual segmentation, often performed by experienced clinicians, is labor-intensive, time-consuming, and prone to human error. Due to these challenges, along with the promising results that CNNs have demonstrated in segmentation tasks, automatic segmentation using CNNs has been extensively investigated in the medical ultrasound community. Yap *et*

al. [53] compared the performance of three CNN-based approaches against four traditional state-of-the-art (SOTA) algorithms for breast lesion detection: Radial Gradient Index [54], Multifractal Filtering [55], Rule-based Region Ranking [56], and Deformable Part Models [57]. Besides, to overcome the lack of public datasets in this domain, they made a breast lesion ultrasound dataset available for research purposes. In 2D echocardiographic images, Smistad *et al.* developed a multi-view network to segment the left ventricular in different apical views [58], and Leclerc *et al.* evaluated several encoder-decoder CNNs for segmenting cardiac structures and estimating clinical indices [51]. They also developed the Refining U-Net (RU-Net) and a multi-task Localization U-Net (LU-Net) to refine and improve the robustness of segmentation [59, 60, 61]. Abraham *et al.* addressed the data imbalance issue [62] in lesion segmentation by proposing a generalized focal loss function based on the Tversky index and combining it with an improved version of an attention U-Net [63]. Karimi *et al.* proposed three different methods to estimate Hausdorff distance from the segmentation probability map produced by a CNN, and suggested three loss functions for training CNNs that lead to a reduction in Hausdorff distance without degrading other segmentation metrics such as the Dice similarity coefficient (DSC) [64]. Gu *et al.* introduced a comprehensive attention-based CNN (CA-Net) by making extensive use of multiple attentions in a CNN architecture for more accurate and explainable medical image segmentation. They claimed that the network is aware of the most important spatial positions, as well as channels and scales at the same time [65]. For prostate segmentation in 2D and 3D transrectal ultrasound images, van Sloun *et al.* employed a variant of U-net [66, 67, 68], and Wang *et al.* developed a 3D deep neural network equipped with attention modules by harnessing the deep attentive features [69]. Li *et al.* employed three modified U-Nets combined with the concept of cascaded networks in IVUS images [70]. Looney *et al.* presented a multi-class CNN for real-time segmentation of the placenta, amniotic fluid, and fetus in 3D ultrasound [71]. For a similar purpose, Zimmer *et al.* used an auxiliary task to improve the performance and introduced a method to extract the whole placenta at late gestation using multi-view images [72, 73]. Zhou *et al.* proposed a fully automated solution to segment the myotendinous junction region in successive ultrasound images in a single shot using a region-adaptive network (RAN), which learns about the salient information of the myotendinous junction [74]. They also introduced an approach that combined a voxel-based fully convolution network (Voxel-FCN) and a continuous max-flow post-processing module to automatically segment the carotid media-adventitia (MAB) and lumen-intima boundaries (LIB) and to generate the vessel-wall-volume (VWV) measurement from three-dimensional ultrasound images [75]. Park *et al.* proposed a technique

to improve the measurement accuracy of the flow velocity in arteries, especially in the near-wall region, by introducing a U-Net-based architecture called USUNet followed by compensation for the effect of wall motion [50]. Amiri *et al.* exploited test time augmentation to improve the accuracy of segmentation of ultrasound images [76]. They also investigated several different transfer learning schemes in ultrasound segmentation [77].

1.4 Thesis Statement

While ultrasound imaging is widely used as a diagnostic tool, its full potential is often constrained by artifacts such as phase aberration, which complicate image interpretation and pose diagnostic challenges. This thesis aims to improve the interpretability of ultrasound images using DL-based approaches on two parallel fronts: enhancing image quality by mitigating the effects of phase aberration and automating the segmentation of ultrasound images while addressing the challenges associated with segmentation.

1.5 Objectives and Contributions

The first part of this thesis focuses on enhancing image quality by mitigating the phase aberration effect, which is a significant factor in the deterioration of medical ultrasound images. In Chapter 2, we propose a novel method to estimate the aberration profile from an ultrasound B-mode image using a deep CNN to compensate for the phase aberration effect. Unlike traditional methods, which typically rely on time-consuming processing techniques applied to RF channel data and require multiple iterations for acceptable accuracy, the proposed approach uses only the B-mode image to predict the aberration profile in a single step with high accuracy. However, the proposed method requires ground-truth aberration profiles for training the CNN, and obtaining such profiles in real-world scenarios is challenging. To address this issue, in Chapter 3, we present a DL-based method that, for the first time in the literature, does not require ground truth to correct the phase aberration effect, enabling direct training on real data. We train a network where both the input and target output are randomly aberrated RF data. Moreover, we show that a conventional loss function, such as mean square error (MSE), is insufficient for training such a network to achieve optimal performance. Instead, we propose an adaptive mixed loss function that incorporates both B-mode and RF data, leading to more efficient convergence and improved performance. Regarding RF data normalization, we demonstrate in Appendix B that conventional min-max scaling for normalizing RF data reduces the efficiency of the training set and propose

that the individual standardization of each image substantially improves the performance of neural networks by utilizing the data more efficiently. Additionally, we publicly release our dataset, consisting of over 180,000 aberrated single plane-wave images (RF data). Although not utilized in the proposed method, each aberrated image is paired with its corresponding aberration profile and a non-aberrated version, aiming to alleviate data scarcity in developing DL-based techniques for phase aberration correction. The source code is also released along with the dataset at `http://code.sonography.ai/main-aaa`.

The second part of this thesis focuses on the automatic segmentation of ultrasound images and explores the challenges associated with employing DL-based approaches for this task. In Chapter 4, we investigate the shift-variance problem in CNNs. Despite their increasing popularity in automatic ultrasound image segmentation, CNNs are not shift-equivariant. This means that if the input is translated, such as laterally, by one pixel, the resulting output segmentation can change drastically. While accuracy is an evident criterion for ultrasound image segmentation, output consistency across different tests is equally crucial for tracking changes in regions of interest in applications such as monitoring the patient's response to treatment, measuring the progression or regression of the disease, reaching a diagnosis, or treatment planning. To the best of our knowledge, this problem has not been studied in ultrasound image segmentation or even more broadly in ultrasound images. Herein, in addition to investigating and quantifying the shift-variance problem, we evaluate the performance of a recently published technique, called BlurPooling [78], for addressing the problem. Additionally, we propose the Pyramidal BlurPooling method, which outperforms BlurPooling in terms of output consistency and segmentation accuracy. We also demonstrate that data augmentation is not a replacement for the proposed method. In Appendix C, we benchmark this method on 3D intraoperative ultrasound images. The source code is also released at `http://code.sonography.ai`.

Another challenge associated with DL-based approaches is that although they outperform classical methods in segmentation tasks, there is a trade-off between their performance and data availability. CNNs with high learning capacities may suffer from overfitting, particularly in the medical domain, where data is often limited. To mitigate this issue, augmenting datasets with synthetic data is a widely adopted strategy; however, generating a large number of images using packages such as Field II [79, 80] is time-consuming. In Chapter 5, we introduce a novel ultra-fast ultrasound image simulation method based on the Fourier transform and evaluate its efficacy in a lesion segmentation task. We demonstrate that data augmentation employing images generated by the proposed method substantially outperforms Field II regarding the DSC, while the simulation process is nearly 36,000

times faster (using CPU). On the other hand, a common challenge in utilizing simulated ultrasound data for training neural networks is the domain shift problem. This issue arises when the distribution of simulated images differs from that of real images, resulting in models trained on synthetic data that are not generalizable to clinical data. The domain shift problem can occur even when sufficient real data is available, but the data were acquired using different scanners or settings. In Chapter 6, we evaluate the effectiveness of the Fourier domain adaptation method and demonstrate its efficacy in enhancing the performance of a breast lesion segmentation task.

1.6 List of Publications

Journal Papers

- M. Sharifzadeh, S. Goudarzi, A. Tang, H. Benali, and H. Rivaz, "Mitigating Aberration-Induced Noise: A Deep Learning-Based Aberration-To-Aberration Approach," IEEE Transactions on Medical Imaging, 2024.
- M. Sharifzadeh, H. Benali, and H. Rivaz, "Investigating Shift Variance of Convolutional Neural Networks in Ultrasound Image Segmentation," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 69, no. 5, pp. 1703–1713, 2022.
- M. Sharifzadeh, H. Benali, and H. Rivaz, "Phase Aberration Correction: A Convolutional Neural Network Approach," IEEE Access, vol. 8, pp. 162252–162260, 2020.

Conference Papers

- M. Sharifzadeh, H. Benali, and H. Rivaz, "Frequency-Space Prediction Filtering for Phase Aberration Correction in Plane-Wave Ultrasound," in 2023 IEEE International Ultrasonics Symposium (IUS), pp. 1–4, IEEE, 2023.
- M. Sharifzadeh, H. Benali, and H. Rivaz, "Robust RF Data Normalization for Deep Learning," in 2023 IEEE International Ultrasonics Symposium (IUS), pp. 1–4, IEEE, 2023.
- M. Sharifzadeh, H. Benali, and H. Rivaz, "Segmentation of Intraoperative 3D Ultrasound Images Using a Pyramidal Blur-Pooled 2D U-Net," in MICCAI Challenge on Correction of Brainshift with Intra-Operative Ultrasound, pp. 69–75, Springer, 2022.

- M. Sharifzadeh, A. K. Tehrani, H. Benali, and H. Rivaz, "Ultrasound Domain Adaptation Using Frequency Domain Analysis," in 2021 IEEE International Ultrasonics Symposium (IUS), pp. 1–4, IEEE, 2021.
- M. Sharifzadeh, H. Benali, and H. Rivaz, "An Ultra-Fast Method for Simulation of Realistic Ultrasound Images," in 2021 IEEE International Ultrasonics Symposium (IUS), pp. 1–4, IEEE, 2021.
- M. Sharifzadeh, H. Benali, and H. Rivaz, "Shift-Invariant Segmentation in Breast Ultrasound Images," in 2021 IEEE International Ultrasonics Symposium (IUS), pp. 1–4, IEEE, 2021.

Journal Papers as a Co-author (not included in this thesis)

- H. Asgariandehkordi, S. Goudarzi*, M. Sharifzadeh*, A. Basarab, and H. Rivaz, "Denoising Plane Wave Ultrasound Images Using Diffusion Probabilistic Models," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 2024.
- A. Tehrani, M. Sharifzadeh, E. Boctor, and H. Rivaz, "Bi-Directional Semi-Supervised Training of Convolutional Neural Networks for Ultrasound Elastography Displacement Estimation," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 69, no. 4, pp. 1181–1190, 2022.

Chapter 2

Phase Aberration Correction: A Convolutional Neural Network Approach

This chapter is based on our published paper [81].

This chapter proposes a novel method for estimating the aberration profile from an ultrasound B-mode image using a CNN to compensate for the phase aberration effect. To the best of our knowledge, this is the first study that employs CNNs for phase aberration correction in ultrasound images. In most cases, CNNs are regarded as a black box, where the output is a corrected version of the input. However, our objective here is not to directly generate a corrected image from the aberrated input. Instead, we aim to estimate the aberration profile, represented as a vector with a limited number of elements. The advantages of our approach are twofold. First, image fidelity is critical in medical applications, especially when it comes to accurately reconstructing hypoechoic targets such as blood vessels and heart chambers. Estimating the aberration profile enables image formation using an open system. Second, it allows us to guide the network to focus on one task, aberration correction, by estimating a small vector instead of a large image. In contrast to traditional methods, which mostly apply time-consuming processing techniques on RF channel data and need several iterations for reasonable accuracy, the proposed approach utilizes only the B-mode image to estimate the aberration profile in one shot with high accuracy. We experimentally investigate the main characteristics of the proposed approach and present a quantitative evaluation of the estimated aberration profile.

The proposed method is compared with the conventional DAS method and a method based on nearest-neighbor cross-correlation (NNCC) [82, 83]. In the NNCC method, one

element is set as the reference, and pairs of adjacent RF channel signals are selected. The normalized cross-correlation (NCC) between one signal and time-shifted versions of the other is then calculated to determine the time delay that maximizes the NCC. Repeating this procedure for all $N-1$ pairs of adjacent RF channel signals in a sub-aperture with N elements yields the aberration profile. However, the NNCC is only able to estimate the relative delays between the probe elements, not the true mean delay error across the aperture. To address this problem, Monjazebi *et al.* [84] presented an optimization-based algorithm to maximize the brightness and variance over a region of interest of the reconstructed image to estimate the absolute mean delay error. The results demonstrate that the proposed CNN method substantially outperforms other approaches based on both quantitative metrics and qualitative assessments.

2.1 Methodology

2.1.1 Aberration profiles

Phase aberration is generally parameterized by a combination of its strength and correlation length. The strength is the root mean square (RMS) of the aberrator function in nanoseconds, and the correlation length, which represents spatial frequency content, is referred to as full width at half maximum (FWHM) of the aberrator autocorrelation function in millimeters [85]. An aberration profile becomes stronger and induces more degradation effect as its strength is increased and its correlation length is decreased, which means higher amplitude with more fluctuations. Several studies in the literature have reported parameters of aberration profiles based on experimental measurements. For instance, the strength and correlation length for *in-vivo* and *ex-vivo* breast tissue are reported as 28.0 ns, 3.48 mm, and 66.8 ns, 4.3 mm respectively [86, 87]. We generated 35,000 one-dimensional aberration profiles by convolving a Gaussian function with Gaussian random numbers similar to [85]. Aberration profiles were varied uniformly in strength and correlation length ranging from 20 to 70 ns, and from 3 to 9 mm, respectively, to cover an extended set of tissues.

2.1.2 Training Dataset

The publicly available Field II simulation package [79, 80] was used to simulated 35,000 aberrated ultrasound images containing 15 scatterers per resolution cell and uniformly distributed inside a phantom of the size of $20\times20\times10$ mm, which was centered at the focal

point and located at an axial depth of 20 mm from the face of the transducer with 64 elements. The phase aberration effect was simulated by applying delays of previously generated aberration profiles to the transducer elements for both transmission and reception to reflect a realistic aberration effect. Simulation parameters are tabulated in Table 2.1.

Table 2.1: Field II parameters for data simulation.

Parameter	**Value**
Center Frequency	5 MHz
Number of Elements	64
Element Height	5 mm
Element Width	Equals to wavelength
Kerf	0.05 mm
Transmit Focus	30 mm

In each simulation, we randomly created 1 to 4 inclusions with a random diameter ranging from 2 to 5 mm and positioned randomly inside the phantom. Each region was either an anechoic region or a hypoechoic region with an equal probability. For the latter, the amplitudes of all inside scatterers were multiplied by a random constant ranging from 0 to 0.5.

2.1.3 Deep Convolutional Neural Networks

In all simulations, the RF channel data were utilized to generate B-mode images, and the results were resized to 256×256 pixels using third-order spline interpolation. Before being fed into the network, mean subtraction was performed to center the dataset around the origin along each dimension. Furthermore, each dimension was divided by its standard deviation to normalize the data. For both pre-processing steps, statistics were computed solely on the training set and then applied to all training, validation, and test sets.

The aberration profiles were set as the output of the network. This output was a vector of size 64, with each number representing the time delay experienced by the corresponding transducer element. The aberration profiles were also normalized to range from -1 to 1 using the maximum absolute value in the training set. This coefficient was saved to convert the network estimations back to the original scale. Fig. 2.1 shows the outline of the proposed method.

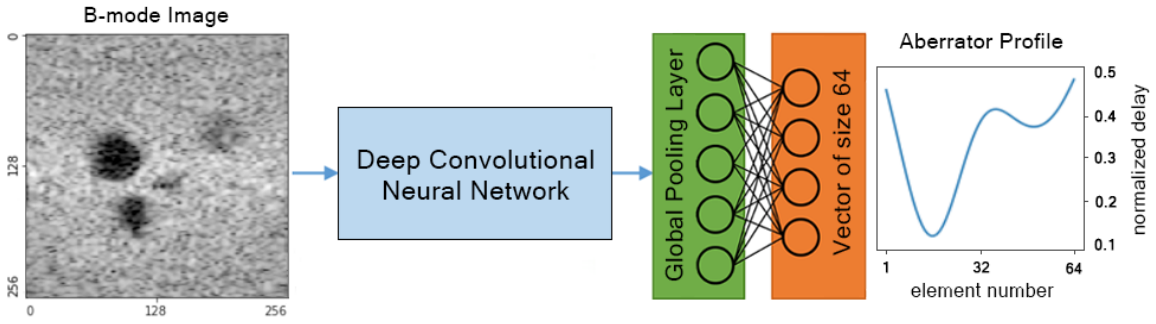


Figure 2.1: Outline of the proposed method. B-mode image is the input, and the estimated aberration profile (vector of size 64) is the output.

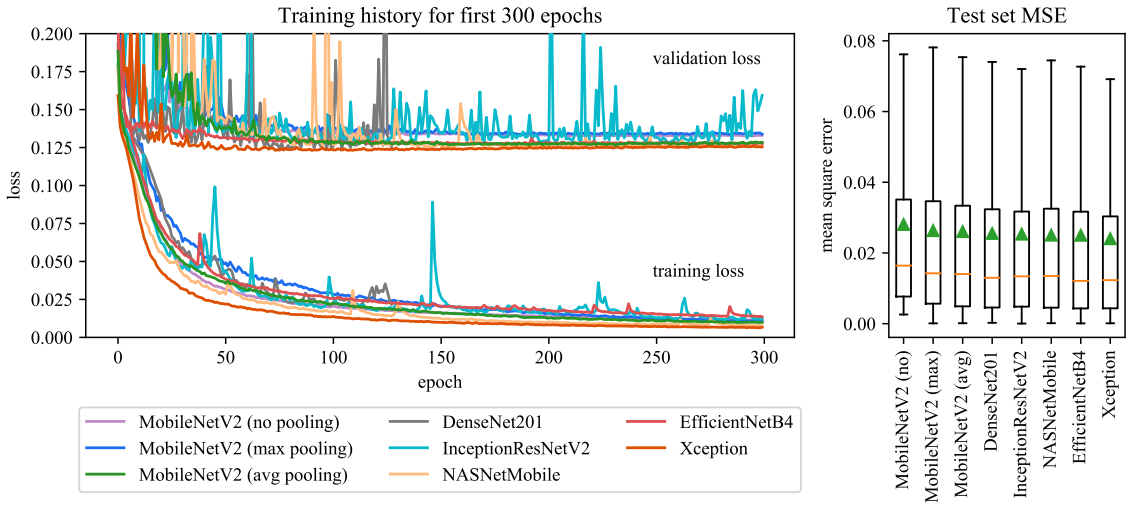


Figure 2.2: Training history and MSE calculated over the test set for several SOTA CNNs. The networks were modified by replacing the final layer with a fully connected layer using a linear activation function to address the regression problem.

Evaluation of Networks

As the first experiment, we investigated the performance of several SOTA CNNs including MobileNetV2 [88], DenseNet [89], InceptionResNetV2 [90], NASNetMobile [91], EfficientNet [92], and Xception [93]. As these networks were originally designed for classification problems, we removed the last layer and replaced it with a fully connected layer with a linear activation function in order to solve the regression problem. The AMSGrad optimizer [94], a variant of ADAM [95], was utilized for training the networks. The learning rate was heuristically set to 0.001. While MSE is a common loss function for solving

regression problems using DL, it is more sensitive to outliers and may result in a less general model in their presence. In this chapter, the mean absolute error (MAE) loss function was employed as we found it more stable during the training process, similar to what is reported in [96], and [97]. All trainings were performed using a single NVIDIA TITAN Xp with 12 GB of memory.

The 15,000 images from the simulated dataset were divided into three training, validation, and test sets composed of 10,000, 3000, and 2000 samples, respectively. An early stopping strategy was implemented to stop training when the validation loss stops improving after 50 epochs, with a minimum of 300 epochs required, regardless of the early stopping rule. For each training epoch, we saved the weights only if the validation loss had been improved and finally used the best weights for the test data. Although we only employed the early stopping strategy to avoid overfitting, applying weight regularization techniques such as L1, L2, or L1L2 regularization, or adding dropout, could also have helped mitigate overfitting. Fig. 2.2 shows training history and the test set MSE for each network. In addition, a sample aberration profile from the test set and its corresponding estimations using each network is illustrated in Fig. 2.3 (top). Although the MobileNetV2 did not attain the best MSE, it achieved comparable results to other networks despite being much less computationally expensive. As such, we chose this network for the rest of this chapter.

Global Pooling

To investigate the effect of global pooling, we trained MobileNetV2 under three different conditions: without a global pooling layer (adding a flatten layer before the fully connected layer), with a global average pooling layer, and with a global max pooling layer. Results for a test sample are presented in Fig. 2.3. As anticipated, including a global pooling layer resulted in a lower error, as it functions as a regularizer [98]. Although some experiments reported that max pooling provides a higher performance because of its nonlinearity [99], we chose the global average pooling, as in addition to its slightly lower MSE in our experiment, it provides a smoother estimation. The jagged estimation obtained by max pooling, regardless of its error, highly decreases the correlation length of the estimated aberration profile, which is not desirable in this application and induces a stronger aberration. We believe this is due to aliasing issues caused by max pooling [78]. Fig. 2.3 shows how the estimation of a sample aberration profile is smoother by employing a global average pooling.

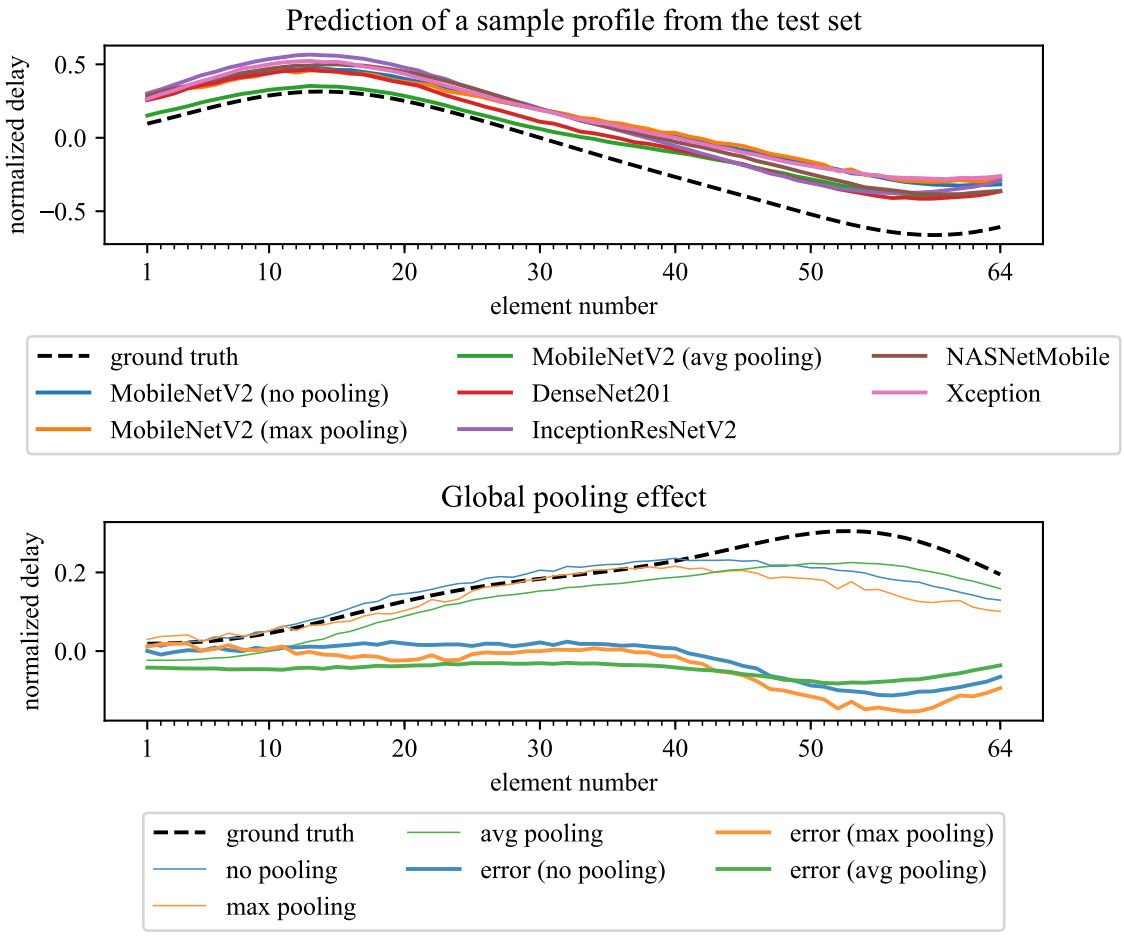


Figure 2.3: (Top) A sample aberration profile from the test set and its corresponding estimated profiles using different CNNs. (Bottom) Results of estimating an aberration profile under three different conditions: without a global pooling layer, with a global average pooling layer, and with a global max pooling layer.

MSE per Element

It is interesting to study the accuracy of delay estimation for different elements. The solid line in Fig. 2.4 (left) shows that MSE of estimated delays is not the same for all elements. Considering all scan lines at the same time, the mean power received by each element from echo signals decreases by moving away from the middle element. Therefore, their corresponding delays have less contribution to the information which supposed to be captured by the network. However, MSE decreases for the very first and very last elements. We believe this is due to the significant contribution of these elements in generating boundary artifacts, which provides the network with a strong texture for the estimation. To demonstrate that

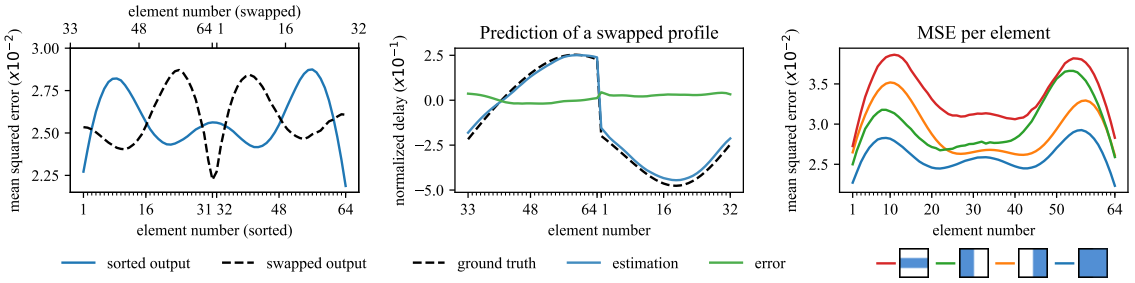


Figure 2.4: Accuracy of delay estimation for each element. (Left) The solid line illustrates that the MSE of the estimated delays varies across different elements. To investigate whether this variation originates from the network, we swapped the first and second halves of the output vectors across the dataset and retrained the network. The dashed black line represents the predicted output for this swapped experiment. (Middle) A sample profile from the swapped dataset and the corresponding predicted output. (Right) Sample predicted outputs for trained networks when only a portion of the image is provided as input.

the origin of this MSE variation is not the network, we swapped the first and second halves of the output vectors across the dataset and trained the network again. The dashed line in Fig. 2.4 (middle) shows a sample swapped profile. Figure 2.4 (left) demonstrates that the MSE for each element remains almost the same before and after the swapping process. This finding indicates that the observed results are not artifacts of the network architecture or the boundary effects of the CNN.

In addition to using the whole image as the input, we also trained networks by feeding only a portion of the image. Fig. 2.4 (right) shows the results for four different cases. The legend represents feeding the network with the whole, right half, left half, and vertically middle part of the image, respectively. We can see that, for instance, the delay estimation error increases for elements at the right side of the probe when the network is fed with merely the left half of the image. It demonstrates that the left half of the image contains less information for estimating element delays on the right side of the probe, as those elements had less contribution in generating the left half of the image. The fourth case shows the worst MSE because of both discarding a portion of axial information, as well as a smaller sample size.

Training Set Size

We trained the network with training sets of sizes 7500, 10,000, 20,000, and 30,000 images. As shown in Fig. 2.5, the MSE over the test set was 0.0293, 0.0260, 0.0254, and 0.0240, respectively, which indicated 11.21%, 2.65%, and 5.52% improvement from the first to the

last.

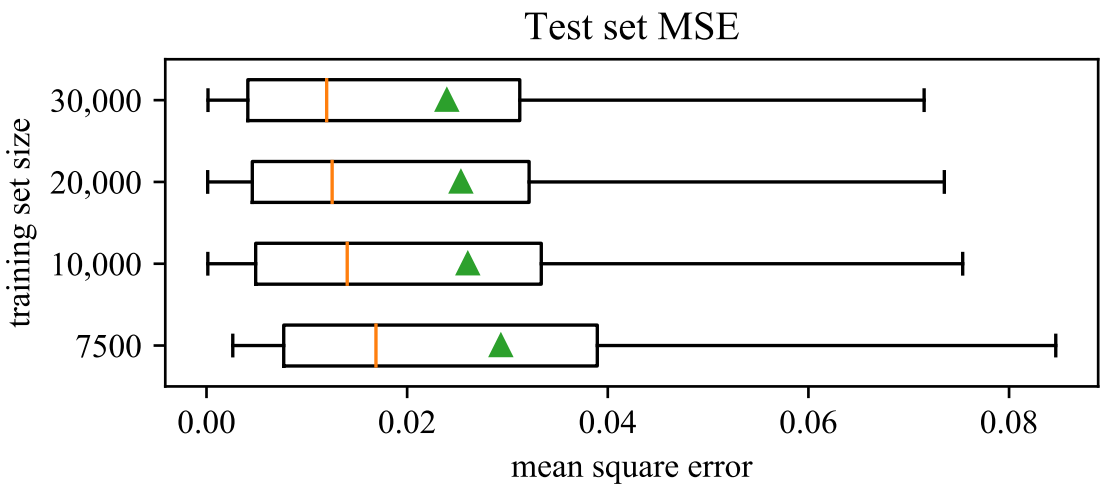


Figure 2.5: The effect of training set size on error. The green triangle and orange line represent the mean and median, respectively.

Output Size

In all previous experiments, the output was a vector of size 64 in which each vector element was the delay of the corresponding transducer element. To study the effect of the output size, we conducted three different new experiments: (1) Instead of estimating delays of all 64 elements together using a single network, we modified the network to estimate the delay of each element individually to see if we can achieve higher accuracy by training 64 separate networks. To that goal, we replaced the output layer with a new layer consisting of only a single neuron. (2) We reduced the output size by a factor of 4 by replacing the output layer with a new layer consisting of 16 neurons to see if higher performance is achievable by reducing the number of parameters. In this case, aberration profiles were downsampled to a vector of size 16 during training the network. For the test, we upsampled outputs again to their original size to have the estimated delays for all 64 elements. (3) We increased the output size by a factor of 4 by replacing the output layer with a new layer consisting of 256 neurons. For training, aberration profiles were upsampled to a vector of size 256, and during the test, we downsampled estimated delays to their original size to make the results comparable.

We trained networks for three aforementioned cases with 30,000 images and used the trained model to estimate aberration profiles of the test set composed of 2000 images. For

the first case, we only trained one network for element #32. Fig. 2.6 (top) shows the MSE for element #32 when we tried to estimate the corresponding delay alone or with other elements together. We can see that MSE decreases by increasing the number of tasks that are asked from the network. Assuming that learning each delay is a separate task, we considerably reduce the risk of overfitting by sharing the hidden layers between all tasks and estimating all delays together. This approach aligns with the multi-task learning (MTL) framework. For example, Baxter [100] demonstrated that the risk of overfitting shared parameters is an order of magnitude N smaller than the risk of overfitting task-specific parameters, where N is the number of tasks. This is why involving more auxiliary tasks, i.e. estimating delays of more elements, leads to smaller errors. Our setup in estimating all delays is analogous to hard parameter-sharing in MTL [101]. Intuitively, the larger number of delays we are estimating simultaneously leads to a more general representation and less chance of overfitting. Fig. 2.6 (bottom) shows the MSE of all elements for different output sizes, except for the output size of 1 which is estimated only for element #32.

2.1.4 Quality Metrics

We evaluate the proposed method for different levels of aberration from weak to strong using contrast, contrast-to-noise ratio (CNR), and signal-to-noise ratio (SNR) to quantitatively measure the quality of reconstructed images:

$$Contrast = -20\ log_{10}(\frac{\mu_{\text{target}}}{\mu_{\text{background}}}) \tag{2.1}$$

$$CNR = 20log_{10}(\frac{|\mu_{\text{background}} - \mu_{\text{target}}|}{\sqrt{\sigma^2_{\text{background}} + \sigma^2_{\text{target}}}}) \tag{2.2}$$

$$SNR = \frac{\mu_{\text{background}}}{\sigma_{\text{background}}} \tag{2.3}$$

where μ is the mean and σ is the standard deviation. For contrast and CNR, the target refers to inside a circular region with a radius of 0.8 times the cyst radius, and the background refers to a region between two concentric circles with radii of 1.1 and 1.8 times the cyst radius. For SNR, the background refers to a square region far from the cyst. These metrics were calculated on the envelope-detected image in the linear domain and prior to applying log-compression.

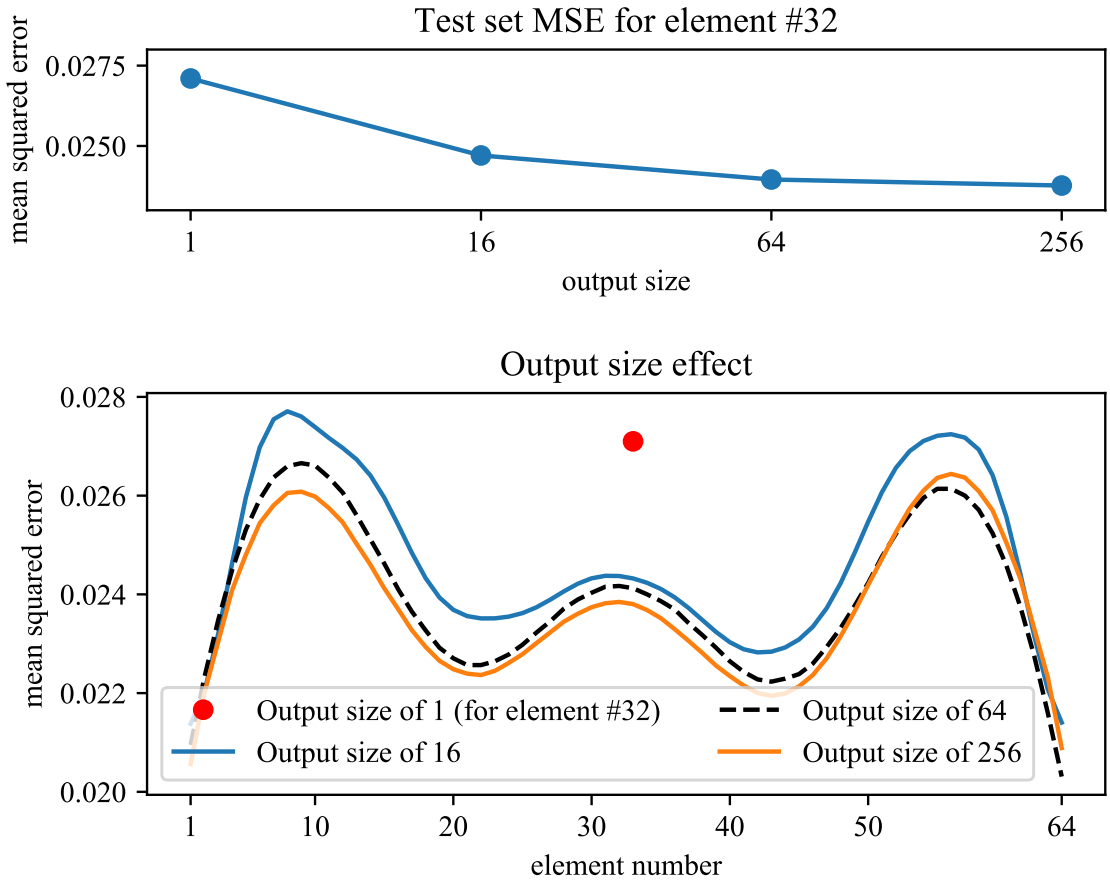


Figure 2.6: MTL versus learning a few tasks. Framing the estimation of each delay as an individual task, we achieve higher accuracy by estimating all delays together rather than estimating them separately.

2.2 Results

The best weights for the network trained with 30,000 images were used to estimate 2,000 aberration profiles of the test set, and the outputs were subsequently subtracted from the corresponding ground truths. The resulting errors can be considered as new aberration profiles, which still induce aberration in the corrected images after compensating for the phase aberration effect.

To quantitatively evaluate the proposed method regarding parameters of aberration profiles, their strength and correlation length were calculated before and after the correction of the test set and shown in Fig. 2.7.

To be able to compare the image quality improvement achieved by the proposed CNN

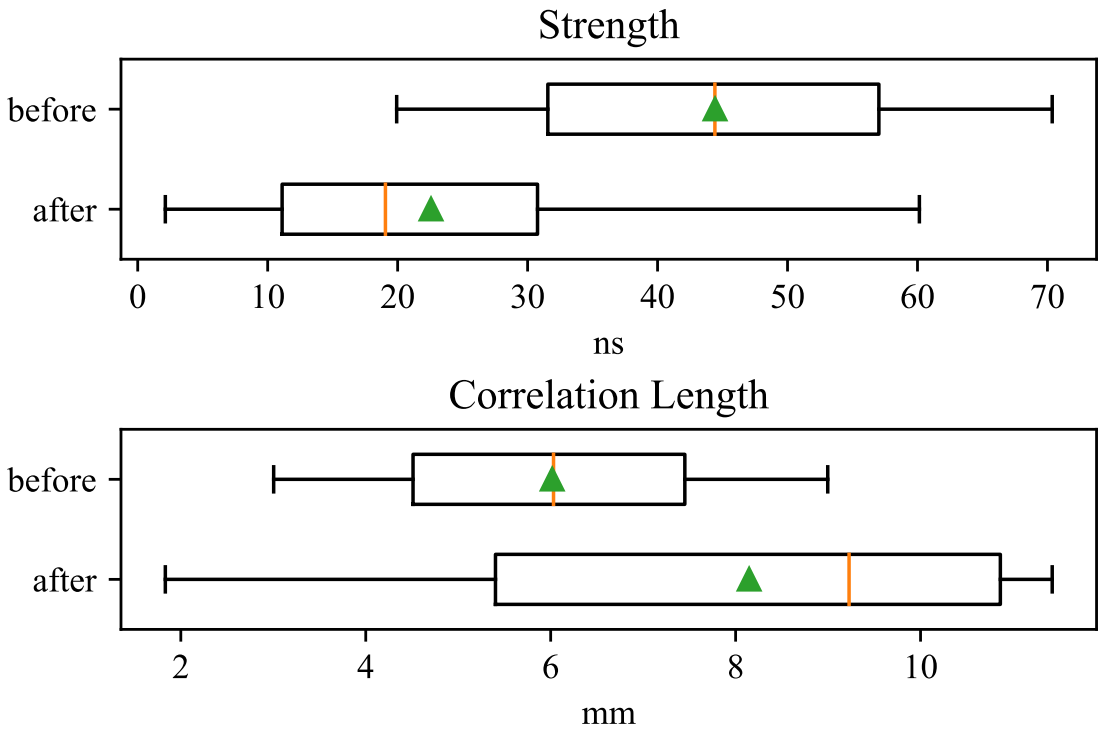


Figure 2.7: Evaluation of aberration profiles strength and correlation length before and after compensating for the phase aberration effect using the proposed method. The green triangle and orange line represent the mean and median, respectively.

method, we simulated another test set composed of 200 aberrated ultrasound images containing an anechoic cyst with a diameter of 4 mm centered at the phantom. Aberration profiles were varied uniformly from weak (20 ns in RMS and 6 mm in correlation length for the first image) to strong (70 ns in RMS and 3 mm in correlation length for the last image). Every other setting was the same as the previous dataset. The best weights for the network trained with 30,000 images and the output size of 64 were used to estimate aberration profiles and corrected images. Three sample images reconstructed using DAS, NNCC, and the proposed CNN methods are pictured in Fig. 2.8, along with their corresponding ground truths and estimated aberration profiles. Top, middle, and bottom rows show samples with weak, moderate, and strong aberration levels, respectively.

Fig. 2.9 presents image quality metrics, encompassing contrast, CNR, and SNR, which were calculated for each method over the test set. The mean and standard deviation values for each comparison are also summarized in Table 2.2.

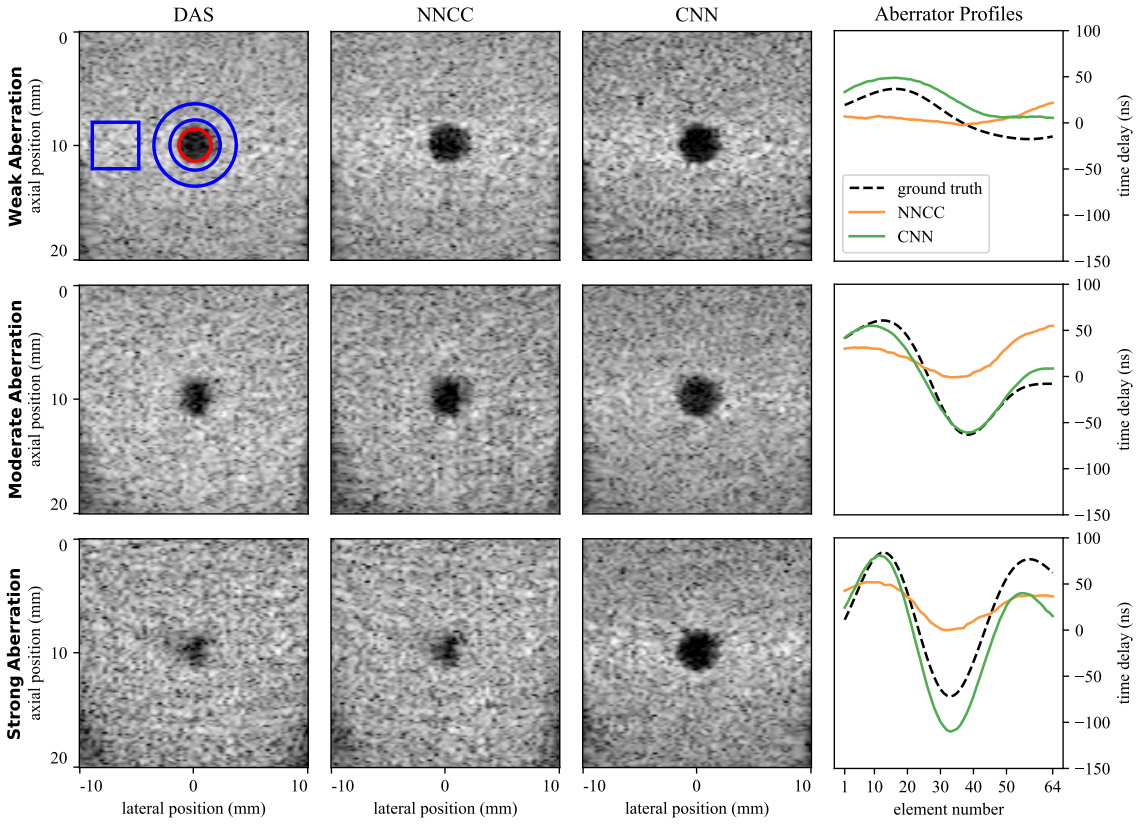


Figure 2.8: Three sample images reconstructed using DAS, NNCC, and the proposed CNN methods. The last column shows the corresponding ground truth and estimated aberration profiles. The top, middle, and bottom rows show samples with weak, moderate, and strong aberration, respectively.

Table 2.2: Results for anechoic cyst phantoms

Metric	DAS	NNCC	**CNN**
Contrast (dB)	22.49 ± 5.95	23.52 ± 5.19	29.71 ± 2.24
CNR (dB)	2.91 ± 1.04	2.95 ± 0.94	3.35 ± 0.61
SNR	1.67 ± 0.14	1.64 ± 0.13	1.56 ± 0.1

2.3 Discussion

We subtracted estimated aberration profiles from corresponding ground truths for every 2000 samples of the test set. Assessing the resulted errors regarding both strength and correlation length parameters is informative; because they are new aberration profiles, which

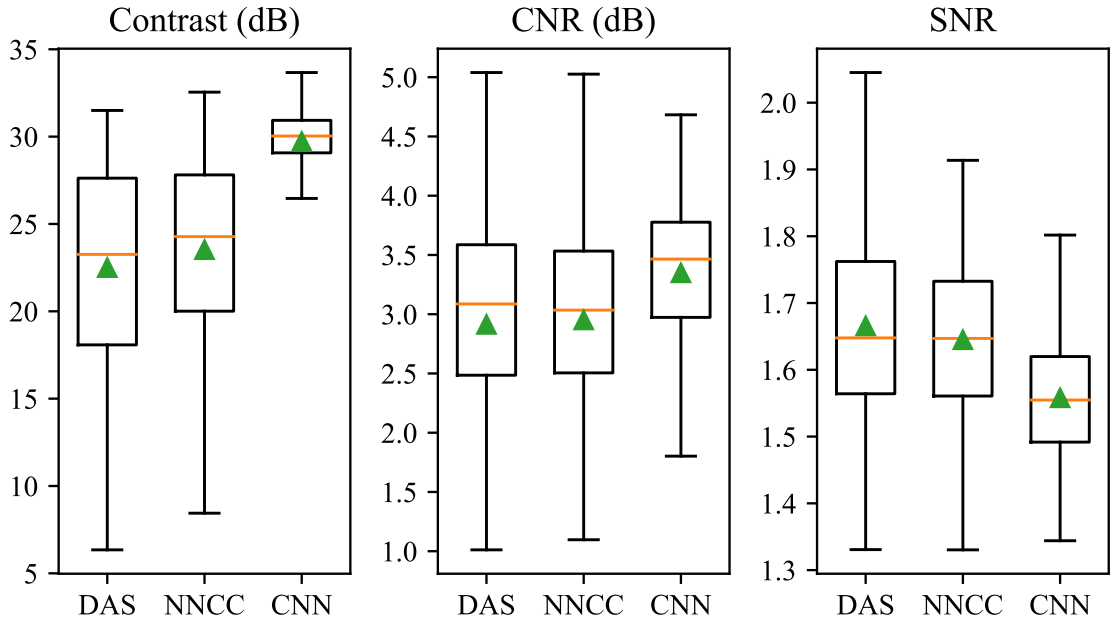


Figure 2.9: Comparison of contrast (dB), CNR (dB), and SNR image quality metrics for DAS, NNCC, and proposed CNN method. Quality metrics are computed from anechoic cysts centered at phantoms and aberrated with different levels of phase aberration. The green triangle and orange line represent the mean and median, respectively.

yet induced aberration to the corrected images. Fig. 2.7 shows that how the proposed method weakened the induced aberration by reducing mean strength from 44.42 ns to 22.57 ns, and increasing mean correlation length from 5.98 mm to 8.15 mm, which indicates improvement by 49.19% and 36.29%, respectively.

Fig. 2.8 shows that the proposed CNN method successfully reconstructed the cyst lesion for all three aberration levels. However, although the NNCC method managed to follow the trend of the ground truth in moderate and strong aberration cases, it barely showed any improvement in the contrast of the cyst lesion.

As we can see in Fig. 2.9, the proposed CNN method dramatically outperformed other methods in terms of contrast, achieving a 7.22 dB improvement. The phase aberration effect tends to smear the anechoic cyst and decrease the contrast; however, this effect can be mitigated through a more accurate estimation of the aberration profile. In contrast, as we expected, the SNR is decreased by the proposed method, as it recovered the speckle pattern, which had been reduced because of the blurring effect induced by aberration. For the same reason, i.e. increasing the background variance, the improvement of CNR is not as substantial as the contrast. However, it's not a drawback for the proposed method as it

was not supposed to blur the speckle pattern in the first place.

The simulation results proved this concept that a CNN can estimate the aberration profile from an ultrasound B-mode image. To be able to successfully apply the proposed method on experimental phantom data or real clinical data, additional considerations may need to be taken into account alongside training the network to minimize the domain shift problem. First, the simulation parameters, such as the center frequency and the number of transducer elements, need to be similar to the experimental transducer. Finally, applying a domain adaptation method may help to achieve a lower error.

2.4 Conclusion

For the first time, we proposed a method to compensate for phase aberration in ultrasound images using neural networks. We generated aberration profiles with a variety of strengths and correlation lengths and employed them to simulate aberrated B-mode images to mimic this artifact for an extended set of tissues, according to experimental measurements reported in the literature. Deep CNNs were trained to take the B-mode image and estimate the aberration profile. Several SOTA CNNs were modified to solve the regression problem and evaluated for this task. In addition to the effect of global pooling and training set size, we explored the effect of boundary artifacts, and how extracting features from different parts of the image can affect the estimated delay per transducer element. We also showed how estimating all delays together, instead of estimating each delay alone, leads to a better performance due to hard parameter sharing in MTL. The proposed method was evaluated in terms of aberration strength and correlation length, as well as image quality metrics, including contrast, CNR, and SNR. The results demonstrated that our method dramatically outperforms both the DAS and NNCC methods.

Chapter 3

Mitigating Aberration-Induced Noise: A Deep Learning-Based Aberration-to-Aberration Approach

This chapter is based on our published paper [102].

In the previous chapter, we introduced an aberration correction method and demonstrated that a CNN can estimate delay errors, or the aberration profile, directly from B-mode images. Although this approach offers a more explainable solution than directly estimating the corrected image, it requires ground truth data for network training. Obtaining ground truths in real-world scenarios can be challenging, if not impossible. As a result, the methods requiring ground truths have to rely solely on simulated data for training, leading to a drop in performance when testing on experimental data due to the domain shift problem. Recent aberration correction studies have recognized the need to eliminate the requirement of ground truths; however, even in such efforts, reconstructed images with a fixed sound speed value of 1540 m/s were still considered clean images [46].

In this chapter, for the first time, we propose a novel DL-based method that does not require ground truth to correct the phase aberration problem and, as such, can be directly trained on real data. We train a network wherein both the input and target output are randomly aberrated RF data. Moreover, we demonstrate that a conventional loss function such as MSE is inadequate for training such a network to achieve optimal performance. Instead, we propose an adaptive mixed loss function that employs both B-mode and RF data, resulting in more efficient convergence and enhanced performance. In addition, we publicly release our dataset, comprising over 180,000 aberrated single plane-wave images (RF data). Although not utilized in the proposed method, each aberrated image is paired with

its corresponding aberration profile and the non-aberrated version, aiming to mitigate the data scarcity problem in developing DL-based techniques for phase aberration correction. Contributions of this chapter can be summarized as follows:

1. We propose the first DL-based aberration correction method that eliminates the need for ground truth in the training phase. Both input and target output are randomly aberrated RF data, which enables us to use real experimental data for training, fine-tuning, or both without any explicit assumption regarding the presence or absence of phase aberrations.

2. Our training setup presents a significant challenge as both the input and desired output of the network contain aberrations that randomly differ in each frame and epoch. Adding to the complexity is the fact that RF data includes high-frequency components. We demonstrate that a conventional loss function such as MSE is inadequate for training such a network. To address this challenge, we introduce a loss function that incorporates both B-mode and RF data and evaluate its performance.

3. We publicly release a dataset comprising 1,802 sets of single plane-wave images (RF data). Each set includes 100 aberrated versions of the same realization. Although not utilized in the proposed method of this chapter, corresponding aberration profiles, and non-aberrated versions are also included for comprehensiveness. To the best of our knowledge, this is the first dataset practically suitable for developing DL-based techniques in this domain, given its size and structure. The source code is also available along with the dataset at `http://code.sonography.ai/main-aaa`.

The proposed method **m**itigates **a**berration-**i**nduced **n**oise using an **a**berration-to-**a**berration **a**pproach, which we name MAIN-AAA, and show that it substantially improves aberrated images in simulation and phantom experiments. Collaborating with an expert radiologist, we could also visually corroborate improvements in *in-vivo* images.

3.1 Methodology

3.1.1 Aberration-to-Aberration Approach

Let us consider a set of noisy scalar measurements, denoted by $a = (a_1, a_2, ..., a_N)$, representing the recorded signal amplitude corresponding to reflection from a particular point within a medium. To estimate the true amplitude, a common approach involves finding

a value $\hat{a}$ that minimizes the expected deviation from measurements according to a loss function L:

$$\arg\min_{\hat{a}} \mathbb{E}_a\{L(\hat{a}, a)\}. \tag{3.1}$$

For $L(\hat{a}, a) = (\hat{a} - a)^2$, it is straightforward to demonstrate that this minimum occurs at the arithmetic mean of the measurements. Training neural networks as regression models is a generalization of this point estimation approach, which means that training a network with infinite samples utilizing an MSE loss function estimates the expectation of the target samples [103].

In the context of ultrasound images, tissue response can be represented as a single point within a high-dimensional manifold. Phase aberrations and artifacts, such as those caused by sidelobes and multiple scattering, can shift this point, deviating from its original position. However, these artifacts tend to be inconsistent across different images, whereas the tissue response remains consistent. Consequently, when training a network using randomly aberrated images, the objective is to disentangle these artifacts from the tissue response by interrogating different aberrated instances.

3.1.2 Phase Aberration Model

We modeled the phase aberration effect by assuming a near-field phase screen in front of the transducer, which introduces different delay errors to each transducer element during both transmission and reception. Although this model does not make any assumptions regarding the spatial distribution of sound speed within the medium, it proves particularly useful in scenarios where an aberrator layer in front of the transducer is so dominant that other sources of aberration in the remainder of the medium are negligible. An example is imaging overweight subjects, where the wave must propagate through a thick layer of fat found in the near-field [20]. In these cases, slight lateral variations in the thickness of the strong aberrator layer may impose strong aberrations, often impeding the optimal performance of methods designed to estimate the distribution of sound speed [33].

The aberration profile in this model is represented as an array, where each element of the array corresponds to a delay error value assigned to a specific transducer element. Aberration profiles are characterized by their strength and correlation length. The strength is defined as its RMS amplitude in nanoseconds, and the correlation length, which represents the spatial frequency content, is defined as the FWHM of its autocorrelation in millimeters [85]. An aberration profile becomes stronger and induces more degradation effect as its

strength is increased, and its correlation length is decreased, which means higher amplitude with more fluctuation across the aperture [81]. Experiments were conducted in some literature to estimate the parameters of aberration profiles. For instance, the strength and correlation length for *in-vivo* and *ex-vivo* breast tissue are reported as 28.0 ns, 3.48 mm, and 66.8 ns, 4.3 mm respectively [86, 87]. We generated random aberration profiles by convolving a Gaussian function with Gaussian random numbers [85], where they were varied uniformly in strength and correlation length ranging from 20 to 80 ns, and from 4 to 9 mm, respectively, to encompass a broad range of tissues.

3.1.3 Phase Aberration Implementation

We first explain the methodology used for introducing phase aberration into simulated and experimental phantom data. Details regarding the simulation or acquisition of data, as well as where each approach was employed, will be discussed later. The Supplementary Video, at *http://code.sonography.ai/main-aaa*, also provides an overview.

Simulated Aberration

To introduce the aberration effect into simulated data, we utilized full synthetic aperture data and synthesized aberrated plane-wave images under linear and steady conditions. Fig. 3.1 demonstrates a typical configuration of ultrasonic imaging systems in the (a) absence and (b) presence of a near-field phase screen, which can be defined by an aberration profile τ_a. A linear array transducer consisting of N elements is positioned in direct contact with the imaging medium of interest. The array is oriented such that the x-axis is parallel to its length, while the depth direction within the imaging medium is represented by the z-axis. After a single plane-wave transmission, the received echo signal at time t by element n located at x_n can be calculated using full synthetic aperture data as follows:

$$RF\left(x_n, t\right) = \sum_{m=1}^{N} RF_{fsa}\left(x_m, x_n, t + \tau_a\left(x_m\right)\right), \tag{3.2}$$

where $RF_{fsa}(x_m, x_n, t)$ is the received echo signal at time t by element n located at x_n solely due to excitation of element m located at x_m and $\tau_a\left(x_m\right)$ is the delay error that element m experiences according to the aberration profile τ_a. In the absence of aberration, as shown in Fig. 3.1(a), we can assume synchronous excitation times for all piezoelectric elements during synthesizing, equivalent to transmitting a flat wavefront. In this case, the delay error τ_a equals zero for all transducer elements. However, to simulate the phase

aberration effect during transmission, as shown in Fig. 3.1(b), we assumed asynchronous excitation times for piezoelectric elements by applying delay errors imposed by the aberration profile. In the absence of phase aberration, the required time for the acoustic wave to travel to point (x, z) and return to the transducer element n located at x_n is

$$\tau(x_n, x, z) = (z + \sqrt{z^2 + (x - x_n)^2})/c, \tag{3.3}$$

where c is the sound speed. The phase aberration effect in reception was implemented as a set of time delay errors corresponding to backscattered signals and according to the aberration profile τ_a. To this end, and given the calculated time delay, each point (x, z) within the region of interest can be reconstructed as

$$s(x, z) = \sum_{n=k-[a/2]}^{k+[a/2]} RF(x_n, \tau(x_n, x, z) + \tau_a(x_n)), \tag{3.4}$$

where k is the nearest transducer element to x, and [.] represents rounding to the nearest integer. Aperture size a determines the number of elements that contribute to the signal and can be expressed using the f-number, which was set to 1.75 in this chapter and is defined as $F = z/a$. In summary, delay errors $\tau_a(x_m)$ and $\tau_a(x_n)$ in Eqs. (3.2) and (3.4) contribute to the aberrations that occur during transmission and reception, respectively, where the former simulates asynchronous excitation of piezoelectric elements during synthesizing the plane-wave and the latter disorders time delays corresponding to received echo signals.

Quasi-physical Aberration

Our approach for introducing a quasi-physical aberration to an experimental phantom required programming a Vantage 256 research scanner. We programmed the scanner to excite transducer elements asynchronously according to a given aberration profile. To this end, delay errors corresponding to each element were calculated in wavelengths of the transducer center frequency and written to the *TX.Delay* array, provided by the scanner programming interface, resulting in the generation of an aberrated wavefront during single plane-wave imaging. Moreover, delay errors introduced by the aberration profile were taken into account during the reception process for reconstructing the image.

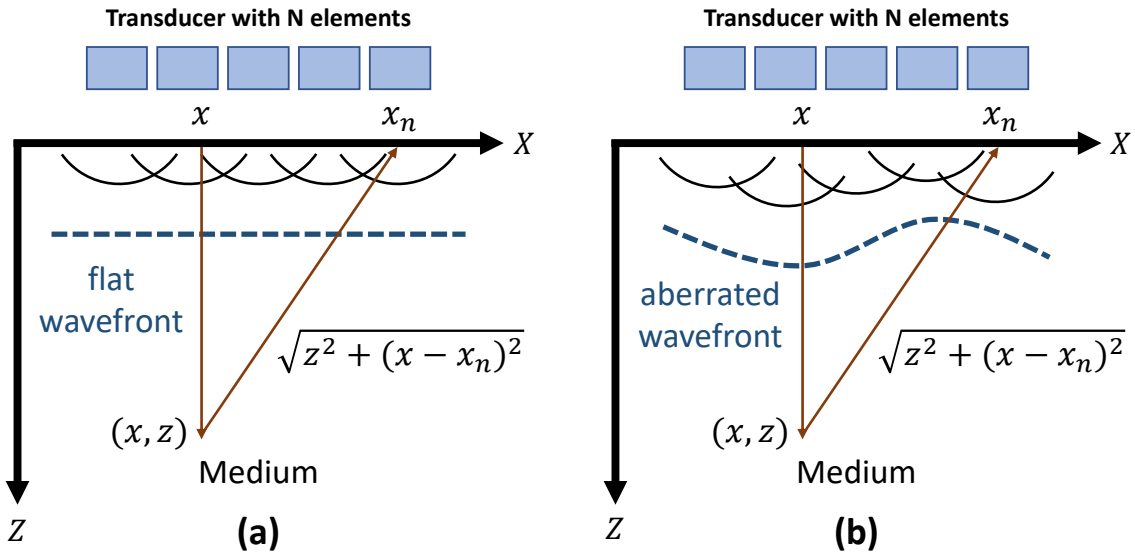


Figure 3.1: A typical configuration of ultrasonic imaging systems in the (a) absence and (b) presence of a near-field phase screen.

Physical Aberration

To introduce a physical aberration to the experimental phantom, we placed an uneven layer of chicken bologna between the probe and the phantom, where the thickness of the left and right halves was approximately 3 mm and 6 mm, respectively. Although the precise value of the sound speed within this layer was unknown, we could be confident that it introduced the aberration effect due to its uneven thickness and observing the effect in the resulting image. To ensure proper contact, we filled the gap between the thinner half and the probe's surface with conductive gel and positioned the center of the probe at the discontinuity.

3.1.4 Datasets

Simulated

We simulated a synthetic dataset consisting of 1802 image sets using the publicly available Field II simulation package [80, 79], containing an average scatterer density of 60 per resolution cell (fully developed speckle pattern). The scatterers were uniformly distributed inside a phantom measuring 45 mm in the lateral and 40 mm in the axial direction, positioned at an axial depth of 10 mm from the face of the transducer. We introduced contrast to the images by incorporating five different types of echogenicities: anechoic regions, hypoechoic regions, hyperechoic regions, diverse echogenicities, and point targets.

To generate the first three types, we took 600 samples (200 samples per type) from a publicly available dataset, known as XPIE [104], which included segmented natural images. We then disregarded natural images and resampled only their corresponding segmentation masks to match the phantom's dimensions. Finally, the amplitude of scatterers located inside the mask was multiplied by a weight, which was zero for anechoic regions, a uniform random number $\in [0.063, 0.501]$ for -12 dB to -3 dB hypoechoic regions, and a uniform random number $\in [2, 15.8]$ for +3 dB to +12 dB hyperechoic regions. To enrich the range of echogenicity, we obtained an additional 1000 samples from the XPIE dataset, but this time we discarded the segmentation masks and instead resampled only the natural images with the same dimensions as the phantom. These images were then converted to grayscale, and similar to [105], the pixel intensities were utilized to weight the scatterers' amplitude according to their respective positions via bilinear interpolation. To enhance the contrast of the ultrasound images, we preprocessed natural images by performing histogram equalization and thresholding pixel values below 0.1 to zero and those above 0.9 to 1. Leveraging natural images and masks for simulation, as described in this subsection, offers the advantage of providing the network with a broader range of features compared to images containing only cysts or selectively chosen shaped regions. To simulate the remaining 200 sets, we introduced multiple randomly positioned point targets to each, where the number of them was determined by a uniform random number $\in [10, 20]$, and their amplitudes were set randomly by drawing from a uniform distribution between 12 dB to 16 dB higher than the mean amplitude of other scatterers. In addition, two test sets were simulated for evaluation purposes: a contrast test set and a resolution test set. The former comprised two anechoic cysts with diameters of 10 mm and 15 mm at central lateral positions and depths of 10 mm and 28 mm, respectively. The latter included a total of 19 point targets arranged in a vertical line at the central lateral position and two horizontal lines at depths of 10 mm and 30 mm. The transducer settings used for simulation were similar to those of the 128-element linear array L11-5v (Verasonics, Kirkland, WA) and are summarized in Table 3.1.4. The center and sampling frequencies were set to 5.208 MHz and 20.832 MHz, respectively. It should be noted that due to the numerical precision of simulations in Field II, the initial sampling frequency was set to 104.16 MHz, and the simulated data was later downsampled by a factor of 5. All images were simulated using a full synthetic aperture scan, followed by synthesizing plane-wave images [106] with 384 columns from the acquired data and saved as RF data. We synthesized 100 randomly aberrated versions of each image according to the procedure elaborated in subsection 3.1.3. Although the non-aberrated version of images was not required for the proposed method, we opted to

Table 3.1: The settings of linear array transducer L11-5v

Parameter	**Value**	**Unit**
Number of Elements	128	elements
Elevation Focus	20	mm
Element Height	5	mm
Element Width	0.27	mm
Kerf	0.03	mm

include them in the published dataset to enhance its comprehensiveness and facilitate the utilization of our data in a broader range of applications. This is because other methods may rely on non-aberrated images as a reference or ground truth. Fig. 3.2 shows samples from the simulated dataset.

Experimental Phantom

An L11-5v linear array transducer was operated using a Vantage 256 system (Verasonics, Kirkland, WA) to scan a multi-purpose multi-tissue ultrasound phantom (Model 040GSE, CIRS, Norfolk, VA). We acquired one scan of anechoic cylinders for evaluations and an additional 30 scans from other regions of the phantom for fine-tuning. In each acquisition, 51 single plane-wave images were captured, including one non-aberrated image (not for training and solely for visualization) and 50 randomly aberrated images utilizing pre-generated aberration profiles as elaborated in subsection 3.1.3. To increase the frame rate, all 1550 required aberration profiles were randomly generated in advance and saved on the disk. Given the fixed position of both the probe and phantom and the sufficiently high frame rate, we assured that all 51 images belonged to the exact same region.

In-vivo

Two *in-vivo* images acquired from the carotid artery of a volunteer with cross-sectional and longitudinal views were employed from the publicly available dataset provided by the plane-wave imaging challenge in medical ultrasound (PICMUS) [107]. Although this dataset was not explicitly designed for assessing aberration correction techniques, it was utilized due to the unavailability of other *in-vivo* plane-wave images specifically acquired with aberrations as testing on a publicly available dataset allows other researchers to compare to our results.

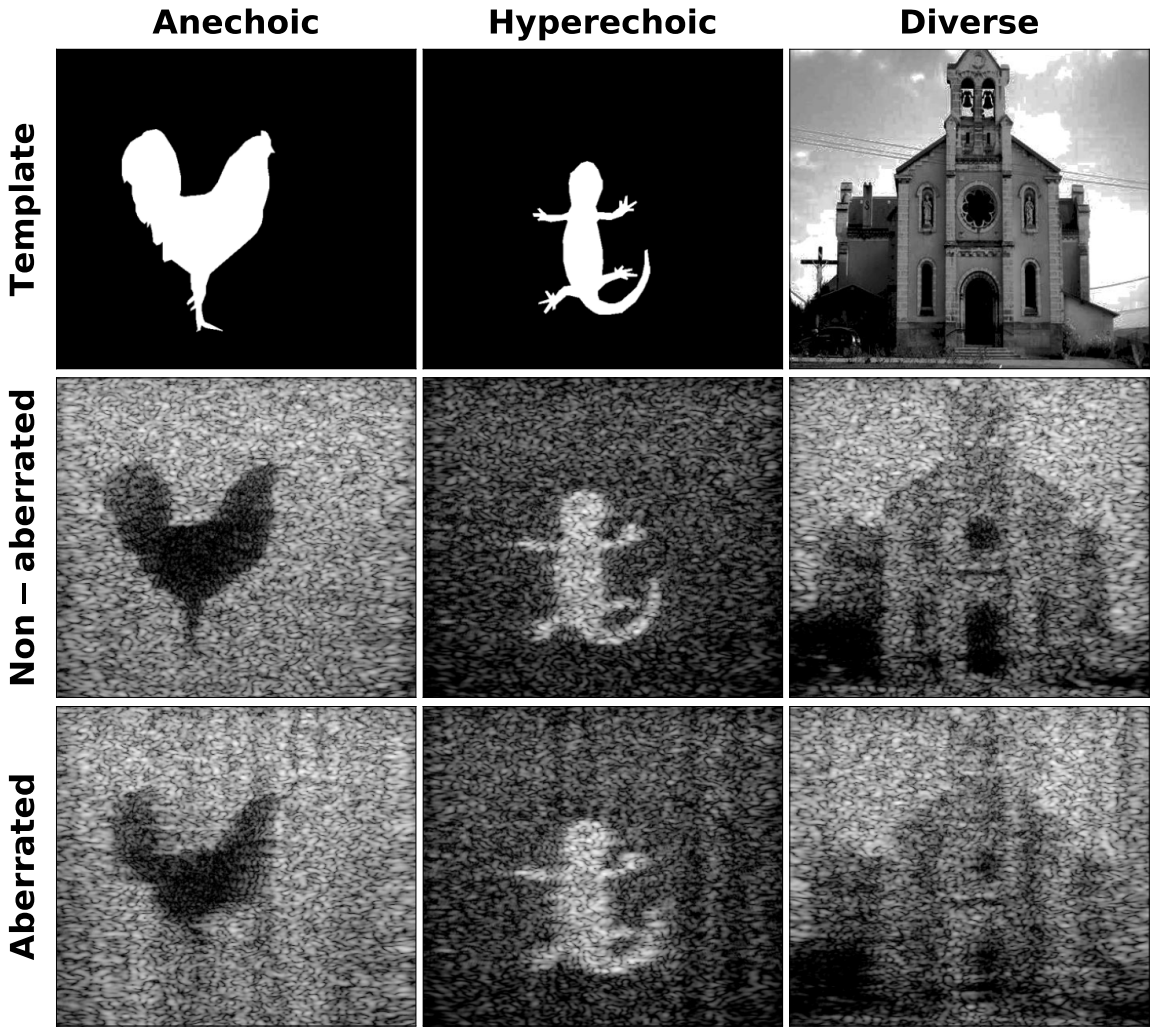


Figure 3.2: Samples from the simulated dataset. The left and middle columns showcase examples of anechoic and hyperechoic regions generated using arbitrary segmentation masks. The right column presents a diverse echogenicity example generated based on a natural image. For each case, the template, non-aberrated, and a sample aberrated version are presented in the first to third rows, respectively. Templates and non-aberrated images are included solely for visualization purposes and were not utilized in the proposed method.

3.1.5 Training

Inspired by Lehtinen *et al.* [103], the U-Net encoder-decoder CNN architecture [108] was employed to map beamformed RF data input to beamformed RF data target output, where both input and target output were distinct randomly aberrated versions of the same realization. The network was trained on 1800 simulated image sets for 5000 epochs, each set comprising 100 aberrated versions. In each epoch, a random pair of aberrated versions

were mapped to each other. To optimize memory usage and accelerate the training process, we downsampled images laterally by a factor of 2, resulting in 192 columns for each image. Moreover, normalization was performed as outlined in Appendix B. A linear activation function was employed in the last layer, and the batch size was set to 32. We utilized Adam [95] with a zero weight decay as the optimizer. The learning rate was initially set to 10^{-3} and halved at epochs 500, 1000, 1500, and 4000. Fine-tuning on experimental images was performed with the same configurations by extending the training by an additional 20% of the original epochs while utilizing a constant and substantially lower learning rate of 5×10^{-5}. To mitigate the impact of non-stationarity and attenuation in RF data training, we partitioned experimental images into three axial sections, each with a 3% overlap, and fine-tuned a distinct network for each depth. When testing experimental images, we fed each image depth to its corresponding network and patched the outputs by blending the envelope of overlapping margins using weighted spatial averaging before displaying the final image. We implemented the method using PyTorch and trained all the models on two NVIDIA A100 GPUs in parallel.

3.1.6 Loss Function

Let $\boldsymbol{S}, \boldsymbol{S'}, \hat{\boldsymbol{S}} \in \mathbb{R}^{p \times q}$ represent input aberrated RF data, target output aberrated RF data, and network output, respectively. The aberration-to-aberration problem can be formulated as

$$\hat{\boldsymbol{S}} = f_{cnn}(\boldsymbol{S}, \boldsymbol{\theta}), \tag{3.5}$$

$$\boldsymbol{\theta^*} = \underset{\theta}{argmin}\ L(\boldsymbol{S'}, \hat{\boldsymbol{S}}), \tag{3.6}$$

where $f_{cnn} : \mathbb{R}^{p \times q} \rightarrow \mathbb{R}^{p \times q}$ is the U-Net, $\boldsymbol{\theta}$ are the network's parameters, and during the training phase, an optimizer is utilized to find optimal parameters $\boldsymbol{\theta^*}$ that minimize the error, measured by a loss function L, between network's output $\hat{\boldsymbol{S}}$ and target output $\boldsymbol{S'}$. In this problem, input and target output were highly fluctuating aberrated RF data, which were randomly substituted at each epoch. We demonstrated in a pilot study in Section 3.2.1 that the network encounters challenges in mapping pairs when comparing RF data directly using a conventional MSE loss defined as

$$L_{mse}(\boldsymbol{S'}, \hat{\boldsymbol{S}}) = \frac{1}{p \times q} ||\boldsymbol{S'} - \hat{\boldsymbol{S}}||^2. \tag{3.7}$$

On the other hand, we illustrated that training the network using the same loss function but on B-mode data leads to improved convergence, which can be attributed to the smoother loss landscape associated with B-mode data. Nonetheless, this improved convergence comes at the expense of discarding valuable information present in RF data. To leverage the benefits of both data types, we proposed an adaptive mixed loss function that gradually shifts from B-mode data to RF data as the training progresses,

$$\begin{aligned} L_{\text{adaptive\_mixed}}(\boldsymbol{S'}, \hat{\boldsymbol{S}}) = (1-\alpha) L_{mse}(\mathcal{B}\{\boldsymbol{S'}\}, \mathcal{B}\{\hat{\boldsymbol{S}}\}) \\ + \alpha L_{mse}(\boldsymbol{S'}, \hat{\boldsymbol{S}}), \end{aligned} \tag{3.8}$$

$$\alpha = \frac{\text{current epoch number}}{\text{total number of epochs}}, \tag{3.9}$$

where $\mathcal{B}\{.\}$ denotes the log-compressed envelope data standardized by mean subtraction and division by its standard deviation.

Our interpretation suggests that the proposed loss function guides the optimizer towards a correct solution by initially utilizing simpler data, before gradually incorporating more complex, fluctuating RF data to take full advantage of the richer information, like curriculum learning [109]. This helps to avoid getting stuck in local minima during the initial stages of the optimization.

3.1.7 Methods for Comparison

Among recent aberration correction methods, many either require multiple plane-wave transmissions [30, 37] or utilize multistatic aperture data for synthetic focusing across all points [35]. We compared the proposed method with two approaches applicable to single plane-wave images, enabling a fair comparison.

Beamsum

The beamsum method, which has recently demonstrated promising results in cardiac imaging [110], estimates delay errors by maximizing NCC between individual channel signals and a common reference signal, known as the beamsum [83]. In this method, after applying beamforming time delays, all channel signals are summed to form the reference signal. Subsequently, each channel signal is aligned with the reference one by maximizing their normalized cross-correlation. A potential limitation arises from a relatively low correlation between individual channel signals and the beamsum, especially in plane-wave images with

limited steering angles. To mitigate this, we averaged each channel signal with those from its n adjacent channels before being compared with the beamsum. While this averaging might theoretically reduce the accuracy of the aberration profile estimation, it enhances overall performance in practical applications when the correlation is low. In this chapter, we heuristically set n to 4 to achieve optimal performance. Additionally, to ensure a fair comparison, corrections using the beamsum method were only applied during reception, without iterative corrections during subsequent transmissions.

FXPF

The FXPF method has proven effective in filtering out acoustic clutter and random noise [19], and its application has expanded to include mitigating noise induced by phase aberration [20]. Let us consider the received RF signal at time t by element n located at x_n and denote its Fourier transform as $RF_n(f) = \mathscr{F}\{RF(x_n, t)\}$. The FXPF method establishes an AR model of order d across the RF channel signals received at transducer elements. Specifically, in the frequency domain and for each temporal frequency f_0, the method predicts a signal as a linear combination of the signals received by the preceding channels:

$$\begin{aligned} RF_{n+1}(f_0) = b_1 RF_n(f_0) + b_2 RF_{n-1}(f_0) + \\ ... + b_d RF_{n+1-d}(f_0). \end{aligned} \tag{3.10}$$

Estimating coefficients b from noisy data filters out non-conforming components based on the established model. Further details can be found in [19, 20]. Although FXPF had been previously employed for focused images, we adapted this method for plane-wave images. The key adjustment involved applying apodization before using the method on the data to avoid image deterioration at shallow depths. This alteration was necessary due to significant variation in channel data across different elements at these depths, where signals from more distant elements are inaccurate and negatively affect the AR model.

In all experiments, the FXPF method was employed with an AR model of order 2 and 3 iterations, determined to yield the optimal performance through a 6×6 grid search, with each parameter ranging from 1 to 6. Consistent with the original study, we set a stability factor of 0.01 and a kernel size equivalent to one wavelength. As the implementation of this method was not publicly available, we took the initiative to publicly release our own implementation, to enhance the reproducibility of the reported results.

3.1.8 Quality Metrics

To quantitatively measure the quality of reconstructed images, we calculated contrast, generalized contrast-to-noise ratio (gCNR) [111], speckle SNR, and FWHM metrics for the test images:

$$Contrast = -20\log_{10}(\frac{\mu_t}{\mu_b}), \tag{3.11}$$

$$SNR = \frac{\mu_b}{\sigma_b}, \tag{3.12}$$

$$gCNR = 1 - \int_{-\infty}^{+\infty} \min_x\{p_t(x),\ p_b(x)\}dx, \tag{3.13}$$

where t and b stand for target and background regions, respectively, μ is the mean, and σ is the standard deviation. In Eq. (3.13), x denotes the image value at any given pixel, and $p(x)$ is the probability density function of the values taken by pixels of a region. The gCNR ranges from 0 to 1, with a higher value indicating better contrast. To provide a fair comparison, all metrics were calculated on the envelope-detected image in the linear domain before applying the log-compression and its subsequent changes to the dynamic range.

3.2 Results

3.2.1 Pilot Study

To assess the performance of the proposed adaptive mixed loss function, we conducted an isolated pilot study using solely the simulated contrast test set described in Section 3.1.4. That set consisted of 100 aberrated versions of the same realization, where 99 versions served as the training set, and the remaining one version was used for testing in this pilot study. The network was trained using the configuration specified in Section 3.1.5, in which during each epoch, each of the 99 versions was randomly mapped to another one. We trained three distinct networks, fed them with the test version, shown in Fig. 3.3(b), and compared their outputs. The first network was trained using B-mode data, both as input and output, utilizing MSE loss. The resulting output is depicted in (c), where cyst boundaries were mostly recovered, but the image appears to be blurry compared to the non-aberrated (a) and aberrated (b) images. This blurring effect is consistent with the findings reported in [112], where the objective was speckle filtering. To further illustrate the principles outlined in Section 3.1.1 regarding the aberration-to-aberration approach, we averaged the 99

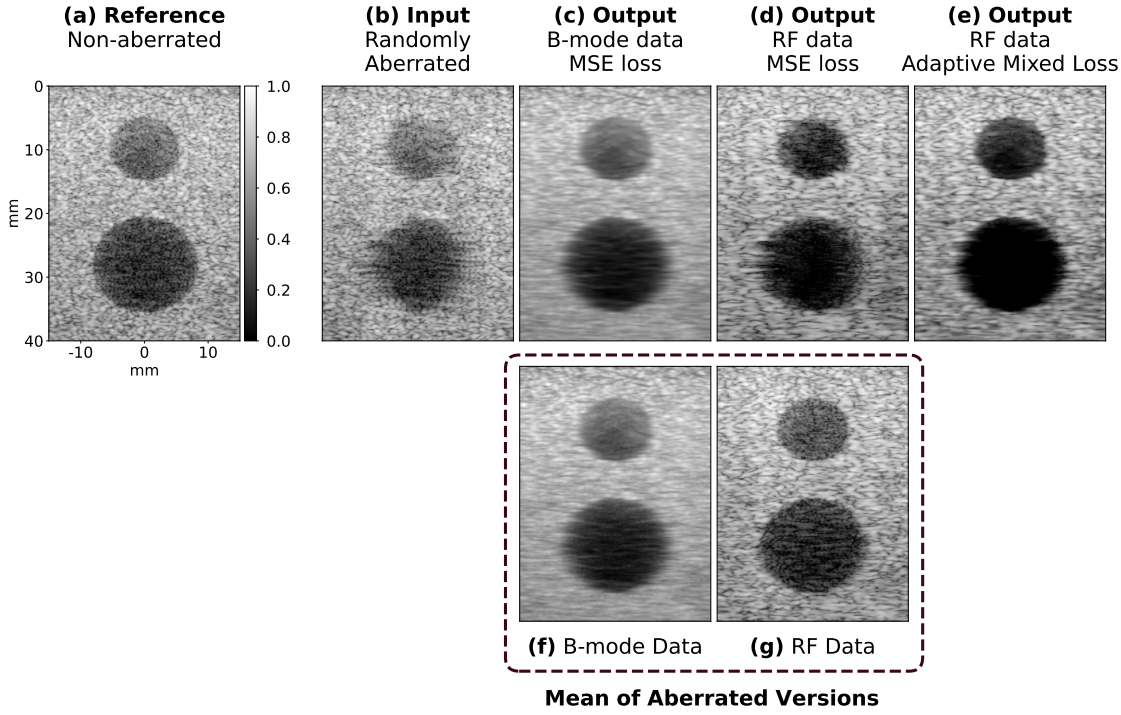


Figure 3.3: Training with different data types and loss functions in a pilot study. (a) The non-aberrated image, shown merely as a reference and not used for training. (b) The aberrated input image. The output of the network when it is trained on (c) B-mode data using the MSE loss function, (d) RF data using the MSE loss function, and (e) RF data using the proposed adaptive mixed loss function. Moreover, (f) and (g) show the mean of 99 aberrated versions that served as the training set in this pilot study, separately for B-mode and RF data. All images were normalized to their maximum intensity value and displayed on a 50 dB dynamic range.

aberrated versions that served as the training set in this pilot study, separately for B-mode and RF data. The results are showcased in (f) and (g), which are aligned with findings in [113]. Interestingly, the network output in (c) closely resembles that of (f), indicating that the first network, trained with MSE loss, attempted to average the aberrated B-mode targets. However, the speckle pattern contains valuable information that can be utilized in applications such as elastography [114, 115, 116]. Motivated by the richer information content present in RF data and aiming for a sharper output similar to the one shown in (g), we trained the second network using RF data as both input and output. As shown in (d), the network encountered challenges in mapping pairs of highly fluctuating aberrated RF data, which were randomly substituted at each epoch, leading to limitations in recovering cyst boundaries compared to the B-mode data scenario.

Inspired by the results obtained from training with B-mode and RF data, we combined

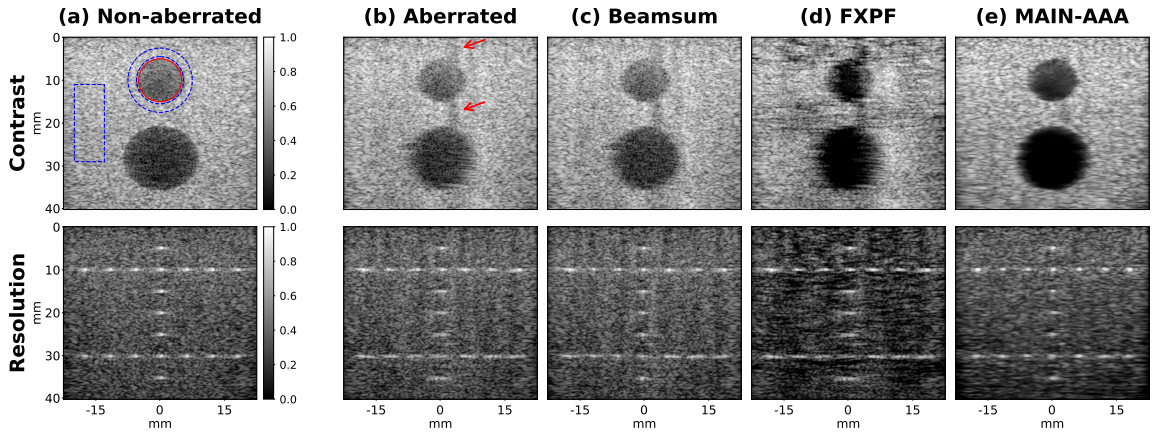


Figure 3.4: Simulated contrast and resolution test images. (a) The non-aberrated image, shown merely as a reference and not used for training. (b) A sample aberrated image reconstructed using DAS. (c) Beamsum output. (d) FXPF output. (e) MAIN-AAA output. All images were normalized to their maximum intensity value and displayed on a 50 dB dynamic range.

both approaches by training the third network using RF data as both input and output but employing the proposed adaptive mixed loss function. The proposed loss function gradually shifts from B-mode data to RF data as the training progresses toward convergence. As shown in (e), this approach exploited the advantages of the rich information within RF data and produced a sharper image compared to (c) while still retaining the ability to recover boundaries more efficiently compared to (d). The enhanced contrast compared to (a) and (g) is also elaborated upon in the Discussion section. Although further investigations are required, we believe that the advantages of the proposed loss function extend beyond the aberration correction task and can potentially improve the performance of other networks working with RF data in various tasks.

3.2.2 Main Study

Based on the findings from the pilot study, we chose the adaptive mixed loss function and utilized it for the subsequent experiments presented in this chapter. In the main study, we trained the network using 1800 simulated image sets and evaluated its performance on two contrast and resolution test sets, each including 100 aberrated versions of the corresponding image. One such aberrated version, reconstructed using the conventional DAS, is shown in Fig. 3.4(b) for each test image, followed by the resulting outputs of the beamsum, FXPF, and the proposed method. Note that in contrast to the pilot study, the network, in this case, was trained on images similar to those depicted in Fig. 3.2 and had never seen, for instance,

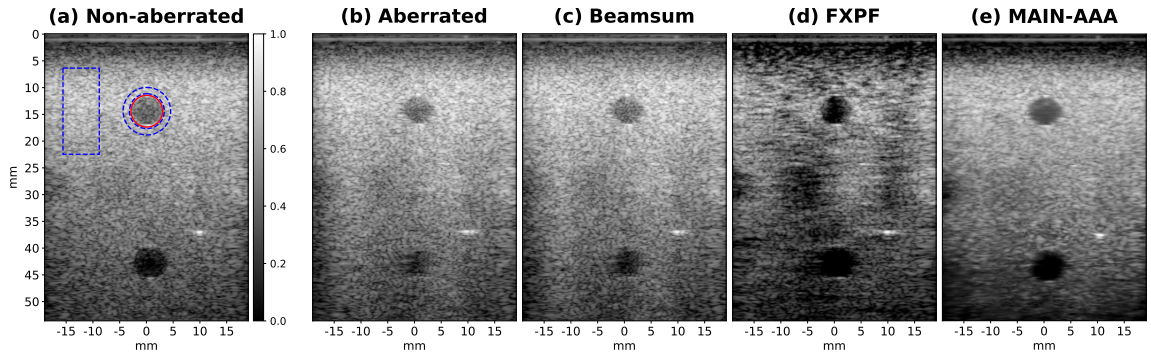


Figure 3.5: Experimental phantom results with quasi-physical aberrations. (a) The non-aberrated image, shown merely as a reference, and not used for training. (b) A sample aberrated image reconstructed using DAS. (c) Beamsum output. (d) FXPF output. (e) MAIN-AAA output. All images were normalized to their maximum intensity value and displayed on a 60 dB dynamic range.

a perfectly circular cyst during the training phase.

To perform a quantitative evaluation, we calculated the average values of contrast and gCNR across the top and bottom anechoic cysts, as well as the speckle SNR in the contrast test image. Although target and background regions, used for calculating metrics, were chosen similarly for both cysts, they are depicted only for the top cyst in the non-aberrated image in Fig. 3.4(a) for brevity. For these metrics, the target region was inside a concentric circle with the same radius as that of the cyst (solid red circle). For contrast and gCNR, the background was the region between two concentric circles with radii of 1.1 and 1.5 times the cyst radius (dashed blue circles), while for speckle SNR, it was inside a rectangle far from the cyst (dashed blue rectangle). Additionally, to evaluate resolution, FWHM was measured for 19 point targets within the resolution test image in the lateral direction. To isolate the FWHM values of each point target from its adjacent ones, we confined the lateral profile to a 4 mm span on either side. The results were obtained for 100 aberrated versions of each test image and are shown in Fig. 3.6.

Fig. 3.5 presents the results for one of the aberrated versions of the experimental phantom test image, which was acquired with quasi-physical aberrations as explained in Section 3.1.3. The quality metrics were calculated similar to those for the simulated test sets. The top and bottom anechoic cysts were utilized for calculating contrast metrics, while the point target at a depth of 37 mm was employed for resolution metrics. These metrics were obtained for all 50 aberrated versions of the test set and presented in Fig. 3.7.

Fig. 3.8 presents the results for the experimental phantom aberrated with a physical aberrator. The first column displays the image reconstructed using conventional DAS,

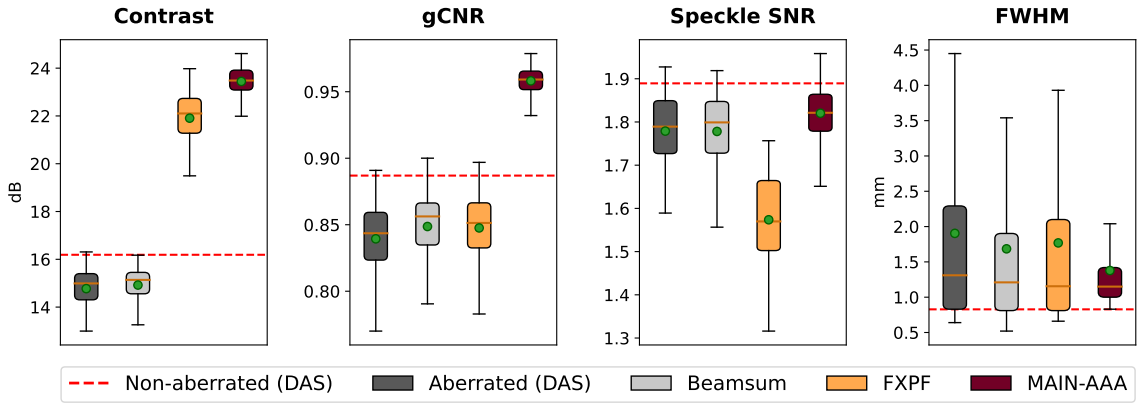


Figure 3.6: Quality metrics computed across the simulated test images. Contrast (dB), gCNR, and speckle SNR metrics computed across the contrast test set, with the FWHM metric obtained across the resolution test set. The green circle and orange horizontal line represent the mean and median, respectively.

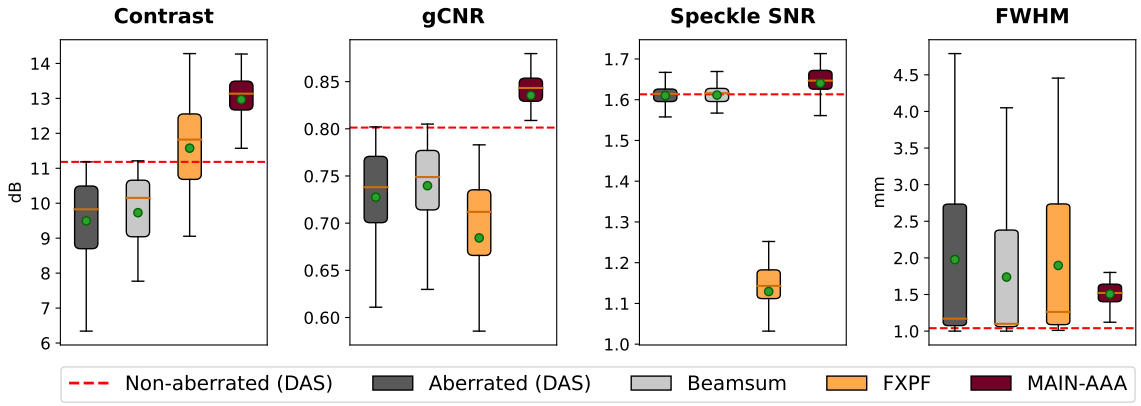


Figure 3.7: Quality metrics computed across the test images from the experimental phantom with quasi-physical aberrations. Contrast metrics were determined using the top and bottom cysts, while the resolution metric was based on the point target positioned at a depth of 37 mm. The green circle and orange horizontal line represent the mean and median, respectively.

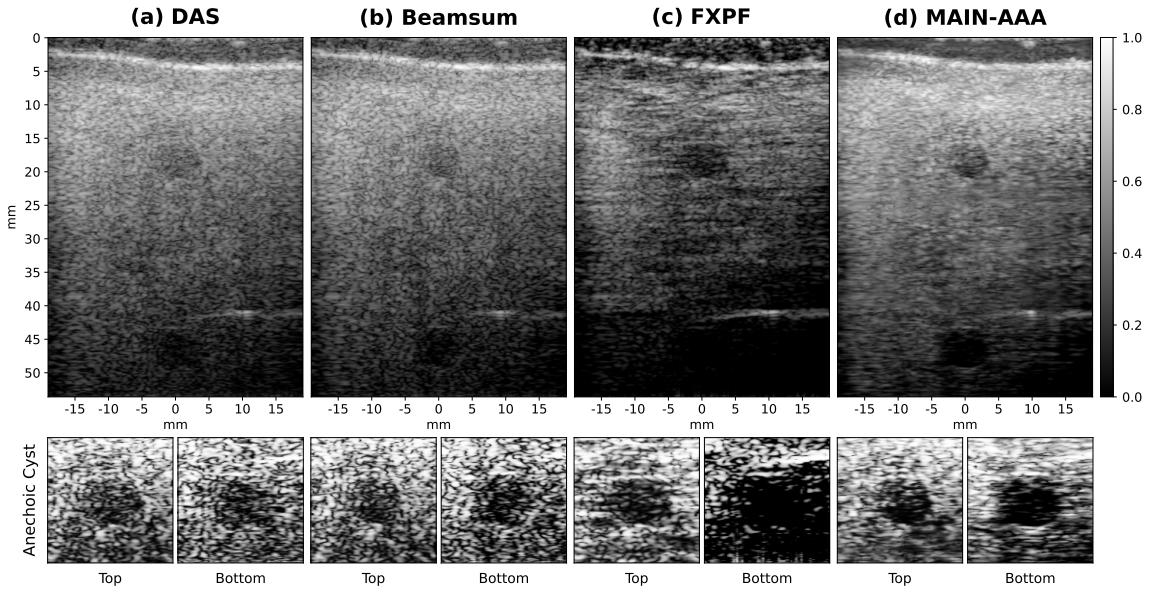


Figure 3.8: Experimental phantom aberrated using a physical aberrator layer. (a) DAS reconstruction. (b) Beamsum output. (c) FXPF output. (d) MAIN-AAA output. All images were normalized to their maximum intensity value and displayed on a 60 dB dynamic range. The second row shows cropped regions of interest (top and bottom anechoic cysts) corresponding to each image, where they were histogram-equalized to enhance visual comparability.

while the subsequent columns show the output images of the beamsum, FXPF, and proposed methods. To enhance visual comparability, the top and bottom cysts were cropped, then histogram-equalized, and displayed under their respective images. The results indicate that the proposed method outperformed the others in recovering cyst boundaries, especially the bottom one. Finally, we applied the methods to *in-vivo* cross-sectional and longitudinal carotid artery images obtained from the PICMUS dataset. The results are shown in Fig. 3.9, including annotations highlighting specific features, which will be further explained in the subsequent section.

3.3 Discussion

In the pilot study, we conducted an experiment where we mapped different aberrated versions of the same image to each other to demonstrate the effectiveness of the proposed adaptive mixed loss function in correcting the phase aberration effect without over-smoothing the RF data and without requiring a non-aberrated ground truth. Notably, the results of this

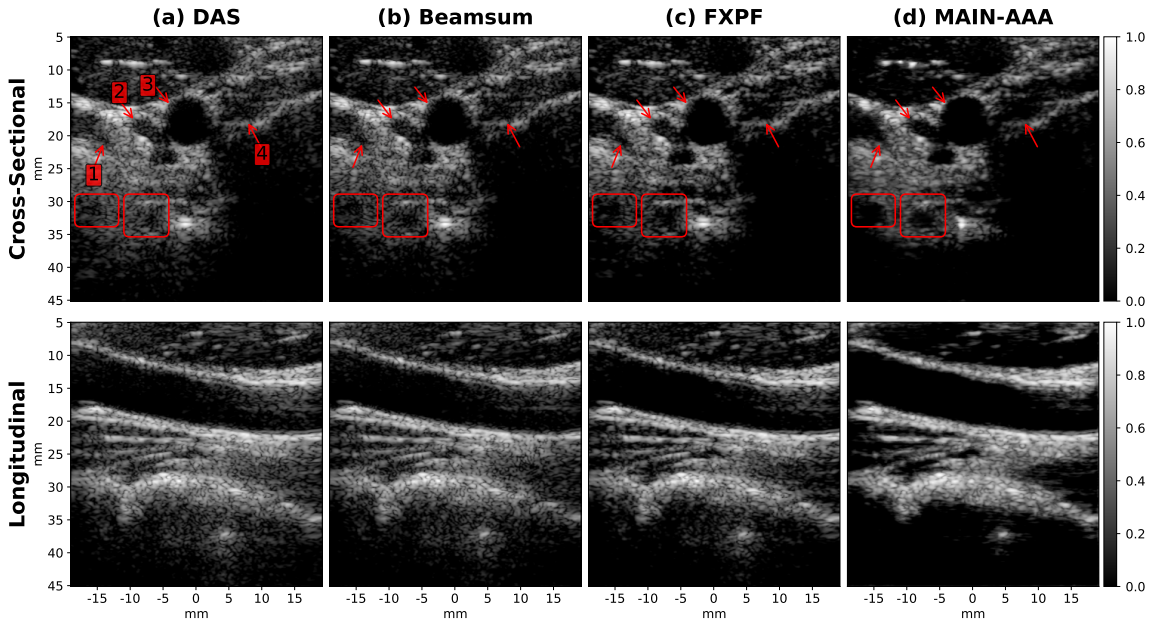


Figure 3.9: *In-vivo* cross-sectional and longitudinal carotid artery images from the PICMUS dataset. (a) DAS reconstruction. (b) Beamsum output. (c) FXPF output. (d) MAIN-AAA output. All images were normalized to their maximum intensity value and displayed on a 50 dB dynamic range.

experiment also revealed an interesting finding. Specifically, it is evident in Fig. 3.3 that not only the phase aberration effect introduced in the input (b) is corrected in the output (e), but a higher contrast is achieved even compared to the non-aberrated reference image (a) and the averaged image (g). One possible explanation for this superior performance could be attributed to the ability of the network to leverage the RF data across the entire image and to average across all plausible explanations in order to output each region of the corrected version. By taking into account all the data points collectively, the network can make more informed decisions regarding each individual value during the reconstruction process. This can be analogized to the non-local means denoising algorithm [117] in traditional image processing, which has been shown to outperform local filters in achieving higher performance. As a result, the network does not rely solely on RF data from local areas to correct the aberration but instead takes advantage of information from the entire dataset, resulting in an improved image compared to the reference image reconstructed using DAS. Another possible explanation is that the network develops the ability to effectively eliminate noise and clutter while randomly mapping one aberrated version to another. Since the noise and clutter are inconsistent across different aberrated versions, the network learns to disentangle them from the consistent tissue response by averaging plausible explanations. Interestingly,

this finding aligns with the study by Jing *et al.* [118], in which they proposed enhancing the spatial resolution of plane-wave images by introducing weak aberration into received data. They calculated the pixel-wise standard deviation of multiple aberrated versions and subtracted the result from the original image. Although our approach differs entirely from theirs, the concept of obtaining an enhanced image from its aberrated versions is similar and can explain the improvement over the reference image reconstructed using DAS.

The main study involved training a general model on a dataset containing images similar to those presented in Fig. 3.2 and evaluating its performance using contrast and resolution test sets, with sample images depicted in Fig. 3.4. Red arrows in the aberrated contrast image (b) highlight the shadowing effect of the perturbed wavefront during transmission. The proposed method outperformed both beamsum and FXPF, which failed to detect and correct this effect. As previously mentioned, the beamsum method was applied only during the reception, unable to compensate for this effect without iterative corrections during subsequent transmissions. Similarly, the FXPF method relies solely on the local signal information of a single image and eliminates components that do not conform to the AR model across the echo signals received at the transducer elements. Nevertheless, in cases where all the echo signals experience a decrease in amplitude, the algorithm is unable to estimate a corrected signal with a higher amplitude. Instead, it tends to amplify the darkness of already dark regions, which may not necessarily correspond to anatomically relevant tissues, such as an anechoic cyst. The findings are consistent with the metrics reported in Fig. 3.6, indicating that while the FXPF algorithm improved contrast, it slightly impacted gCNR. Conversely, the proposed MAIN-AAA method enhanced contrast and achieved a higher gCNR of 0.96, which is substantially closer to the maximum value of 1.

Similarly, we can observe in Fig. 3.5 that the proposed method recovered the size of the anechoic cyst at the bottom of the image more accurately, contrasting with the beamsum and FXPF methods which respectively led to an underestimation and overestimation of its size. As reported in Fig. 3.7, although the FXPF method improved the mean contrast of cysts, the mean gCNR actually decreased due to its aforementioned limitation. Conversely, the beamsum and proposed methods consistently enhanced both the contrast and gCNR metrics, with MAIN-AAA outperforming the other in both metrics. In addition to the top and bottom anechoic cysts, this image contained four additional cysts positioned at a depth of 30 mm, arranged from left to right with contrast levels of -6 dB, -3 dB, +3 dB, and +6 dB relative to the background. It can be observed that, for instance, the -3 dB target was recovered with higher accuracy in terms of both its shape and contrast level, aligning with our prior knowledge that it was a hypoechoic cyst with a contrast level of -3 dB.

In both simulation and phantom experiments, the FXPF method reduced speckle SNR, which aligns with the findings reported in [20]. As illustrated in Figs. 3.4 and 3.5, this method increased the variance of the values within the blue rectangle while it preserved or even reduced their mean (darker region), thereby leading to a reduction in speckle SNR. In contrast, the proposed method consistently preserved the speckle SNR at a level approximately comparable to that of the aberrated image reconstructed using conventional DAS and the beamsum method. This preservation is deemed positive, given that the proposed method inherently operates by averaging all plausible explanations, thereby tending to smooth images. One of the objectives of introducing the adaptive mixed loss function was to prevent the network from over-smoothing images. While smoothing the speckle pattern could enhance the SNR, preserving it was desired for the proposed method.

Both simulation and phantom experiments demonstrated a relatively similar trend in resolution metrics. As shown in Fig. 3.6 and Fig. 3.7, MAIN-AAA achieved the best mean FWHM, followed by the beamsum method. According to the mean values, the distributions of FWHM seemed more skewed in the simulation experiment compared to the phantom experiment because the experimental phantom images featured only one point target, whereas the simulated images contained 19 point targets distributed across different depths, with a less pronounced impact of aberration on shallower targets. While the mean FWHM across multiple data points can serve as a reliable measure for assessing resolution, it is worth noting that the phase aberration effect can also lead to artificially lower FWHM values. Such instances may occur when calculating the metric at the edge of the target or on a noisy profile, often due to a displaced or entirely missing target. If a method mitigates the issue by partially recovering such a missing target, this recovery may contribute to increasing the FWHM value. This partially explains why the proposed method consistently yielded higher minimum FWHM values due to recovering more erroneous values. Another reason is its inherent averaging nature, which limits its ability to achieve resolutions equivalent to non-aberrated versions.

Fig. 3.8 shows a similar trend as observed in Figs. 3.4 and 3.5, where the improvement in the bottom cyst is more prominent than in the top cyst. This observation can be attributed to the fact that the severity of the phase aberration effect, which needs to be corrected, is lower at shallower depths compared to deeper depths for two reasons. Firstly, perturbations in the wavefront escalate with propagation, leading to an increase in the aberration effect during transmission as the wavefront moves forward. Secondly, in plane-wave imaging, the reconstruction process uses a smaller aperture size a at shallower depths and gradually

increases it for deeper depths according to the f-number. Employing fewer neighboring elements during the reconstruction of lower depths can limit the aberration effect, especially when the aberration profile lacks abrupt changes, as a smaller segment of the aberration profile directly affects the reconstruction.

In Fig. 3.9, we compared the methods on cross-sectional and longitudinal views of the *in-vivo* carotid artery image obtained from the PICMUS dataset. Despite the absence of intentional aberrations and any ground truth, we observed interesting results in these images. Several parts of the images were modified after applying the aberration correction methods, and we highlighted some of the notable alterations. In the cross-sectional view, the reconstructed image obtained by the DAS method exhibited indistinct boundaries for the right subclavian vein (arrow #1). Although the FXPF method mitigated the issue, the proposed MAIN-AAA method achieved a higher level of boundary contrast, allowing differentiation of the vessel wall and its anechoic lumen. A similar trend was observed for the right jugular vein (arrow #2) and right common carotid artery (arrow #3), wherein the proposed method reconstructed images with superior tissue differentiation as evidenced by sharper boundaries and more distinct anatomical structures compared to other methods. Similarly, the proposed method demonstrated a higher contrast for the posterior edge of the right thyroid lobe (arrow #4), thereby further enhancing the visual quality of the reconstructed image. Additionally, the MAIN-AAA method revealed ovoid-shaped structures within the two rectangles, which may represent the right subclavian artery (small rectangle) and right vertebral artery (large rectangle) seen cross-sectionally. These areas were barely visible using the DAS method, potentially because of their depth and size, as well as potential aberration induced by the highly anisotropic sternocleidomastoid muscle (the large oval at the top left of the figure). Although an aberration-free ground truth for this image is unavailable, this representation is supported by the similarity of the pattern within these rectangles in the DAS image to the aberrated anechoic cysts seen in the phantom images and the fact that both the beamsum and FXPF methods, which are not learning-based, identified them by attempting to enhance their contrast, while the proposed method produced substantially sharper boundaries for these tissues. In the longitudinal image, while both the beamsum and FXPF methods aimed to mitigate reverberation within the lumen of the common carotid artery at a depth of 17 mm, the proposed method outperformed them, resulting in a markedly sharper vessel wall contrast. A similar trend was observed for the muscle structure at depths ranging from 20 to 28 mm. Interestingly, regardless of the contrast improvement, both the beamsum and proposed methods slightly refined the bright point at a depth of 37 mm, reducing its dispersion. Although the absence of a ground

truth complicates evaluating the accuracy of this modification, considerations such as the position of the bright point suggest that these methods have contributed to a more precise depiction.

The capability of a network trained based on the near-field phase screen model to mitigate the noise induced by phase aberration in images such as the presented *in-vivo* ones, which might be affected by distributed aberrations and do not necessarily adhere to the model, could be a subject of doubt. To address this concern, we can assume that the aberration at any given point within a medium is a consequence of variations in sound speed along the trajectory linking said point to each element of the aperture. Thus, regardless of the distribution of the aberrator, the variations in sound speed that contribute to the aberration of a particular point can still be approximated by a near-field phase screen. In other words, distributed aberrations throughout a heterogeneous medium can be characterized by multiple aberration profiles, each corresponding to a specific point. Chau *et al.* [21] developed a locally adaptive phase aberration correction method based on this assumption, estimating a local aberration profile at each point in the discretized image domain. While, theoretically, each point within the propagation medium may require a dedicated aberration profile, adopting the concept of finite-sized isoplanatic patches allows a single aberration profile to effectively model aberrations for all adjacent points within the same patch, which dramatically reduces the number of profiles required to characterize distributed aberrations across the medium. Consequently, we speculate that a network, trained and fine-tuned on over 180,000 unique aberration profiles, could learn from the presented variations and become capable of correcting distributed aberrations locally, even if they cannot be modeled by a single near-field phase screen.

While the proposed method eliminates the need for ground truth data, a drawback is its reliance on averaging aberrated patterns to approximate the original tissue response. We investigated the incorporation of both low and high-frequency data, introducing an adaptive mixed loss function to facilitate the network's utilization of RF data and produce sharper outputs. Nevertheless, the underlying averaging principle inherently results in a smoother speckle pattern compared to methods that directly compensate for delay errors, such as beamsum. Furthermore, although modeling phase aberrations using near-field phase screens effectively mitigated noise resulting from distributed aberrations by addressing them locally, the performance is ultimately restricted by the aberration model within the training dataset. Future studies could explore the utilization of more complex models for phase aberrations, potentially leading to improved method efficacy. Additionally, we

demonstrated the effectiveness of an aberration-to-aberration approach using single plane-wave images. Expanding on this work by incorporating compounded plane-wave images with additional steering angles [119] or exploring other imaging modes presents potential avenues for future research.

3.4 Conclusions

We proposed a novel DL-based approach for correcting phase aberration that eliminates the requirement for ground truths. We illustrated that a conventional loss function, such as MSE, is inadequate to achieve optimal performance and introduced an adaptive mixed loss function to train a network capable of mapping aberrated RF data to aberrated RF data. This approach permits training or fine-tuning using experimental images without prior assumptions about the presence or absence of aberration. Furthermore, we demonstrated the feasibility of obtaining the required data for this method using a programmable transducer and acquiring multiple aberrated versions of the same scene during a single scan. Apart from releasing the code for the proposed and the FXPF method, we also made available to the public a dataset containing more than a thousand sets of single plane-wave images stored as RF data, where each set comprises 100 aberrated versions of the same realization along with corresponding aberration profiles and the non-aberrated version, aiming to facilitate the advancement of DL-based methods for correcting phase aberration in ultrasound images.

Chapter 4

Investigating Shift-Variance of Convolutional Neural Networks in Ultrasound Image Segmentation

This chapter is based on our published papers [120, 121].

The primary focus of the previous two chapters was to enhance the interpretability of ultrasound images by improving their quality. This chapter pursues the same objective through a parallel approach, focusing on the automatic segmentation of ultrasound images, which can contribute to simplifying their interpretation. Manual segmentation is regarded as the gold standard in many applications, but it is tedious and impossible to perform fast enough for real-time applications. To overcome these issues, automatic segmentation has been a highly sought-after research area in medical image processing, with approaches based on CNNs gaining increasing attention recently. However, while CNNs had been considered shift-invariant for many years, it has been recently reported that they are subject to the shift-variance problem [122, 78, 123, 124, 125, 126]. In other words, if the input translates by only one pixel, the segmentation result may change. Although this drawback may be tolerated in applications such as natural image classification, it hinders CNNs' performance in sensitive applications, such as medical image segmentation, where the reliance of the networks on main features, as well as reproducibility of the results, are essential. Since in many scenarios, changes in images are being tracked to monitor the progression or regression of the disease or the patients' response to the therapy. In addition, in image-guided interventions wherein the target moves with breathing or other physiological motions, shift-equivariance is paramount. All of this brings us to the conclusion that in addition to accuracy, consistency is also a crucial metric that needs to be considered

while evaluating CNN-based methods for medical applications such as ultrasound image segmentation.

To shed light on the shift-variance problem in CNNs, Fig. 4.1 shows the effect of input translations on output segmentation masks. The left column shows an identical input image from the test set, which was translated diagonally from top-left to bottom-right for -3, -1, 1, and 3 pixels. The next column shows the corresponding segmentation masks generated by a baseline method without addressing the shift-variance problem. The baseline method was affected by small translations and generated substantially different output masks.

While the output robustness against input translations can be crucial in medical applications, in the literature, the main focus has primarily been on improving segmentation accuracy metrics such as the mean boundary distance, Hausdorff distance, DSC, and volume difference or overlap. Since CNN-based methods are prone to the shift-variance problem, the effect of input translations on the output should not be overlooked. This chapter represents the first study, to our knowledge, that investigates the shift-variance problem of CNNs in the context of ultrasound image segmentation or even more broadly in ultrasound imaging. The key contributions of this chapter are summarized as follows:

- We investigate the shift-variance problem of CNNs in ultrasound image segmentation using synthetic and *in-vivo* datasets.
- A recently published technique, called BlurPooling [78], is applied to mitigate the shift-variance problem, and its performance is evaluated on all datasets.
- We propose a new approach, called Pyramidal BlurPooling, which takes into account the nature of the ultrasound segmentation task, and outperforms BlurPooling in both shift-equivariance and segmentation accuracy.
- We demonstrate that data augmentation by random translations, a common practice to aid training, is not a replacement for the proposed method.

4.1 Background

4.1.1 Shift-variance in CNNs

One of the motivations for proposing CNN and applying pooling layers as a part of their architectures was making networks robust to irrelevant changes, such as different scales of the same object or image translations [127]. This idea had been considered valid for many

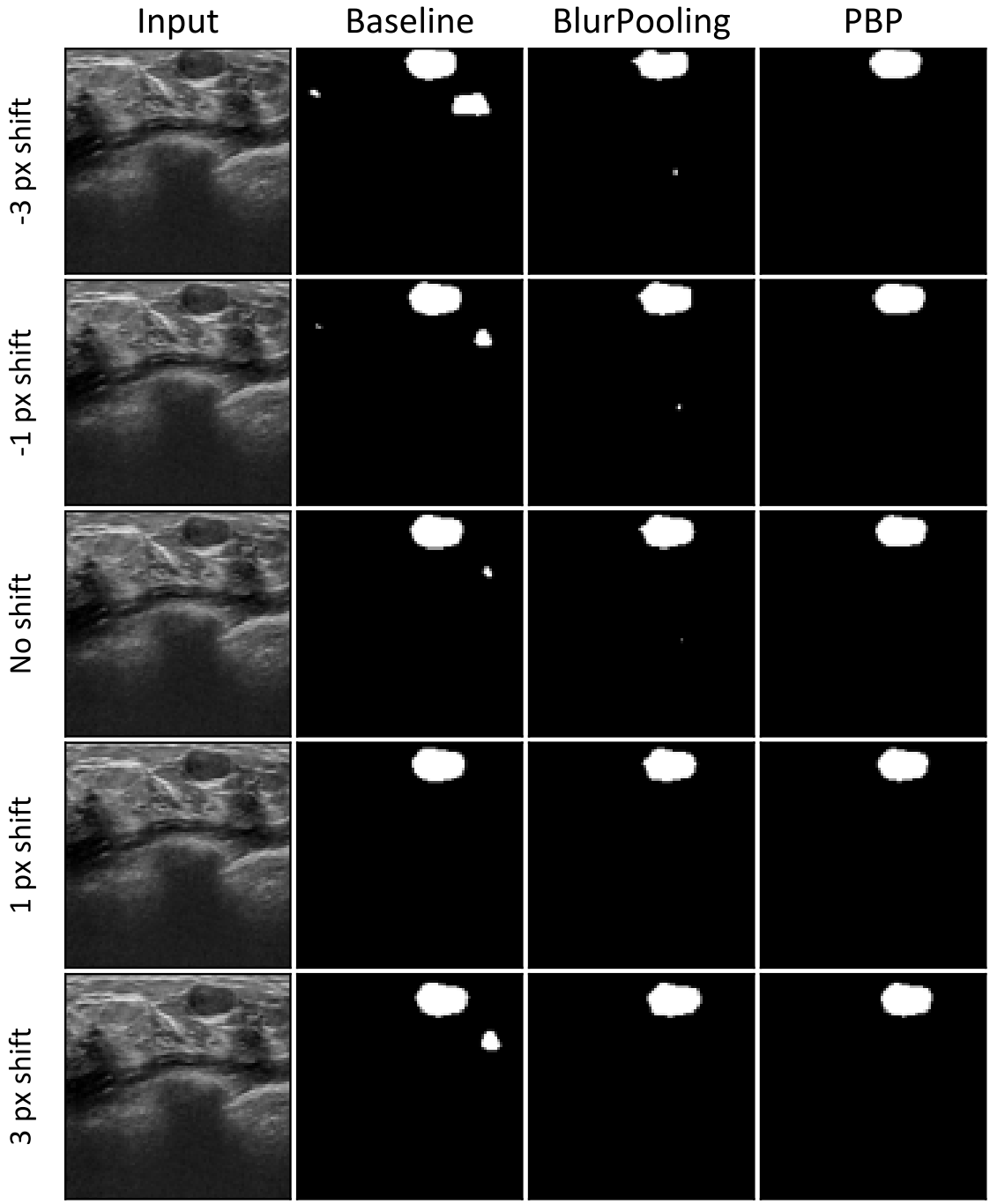


Figure 4.1: Illustration of the effect of input translations on generating output segmentation masks. The left column shows an identical input image from the test set translated diagonally for k pixels, where $k \in \{-3, -1, 0, +1, +3\}$. The following columns show corresponding segmentation masks generated by the baseline method, BlurPooling (7×7), and the proposed method (Pyramidal BlurPooling), respectively.

years based on two reasons: the convolutional nature of layers is shift-equivariance, and data augmentation can be employed by feeding the network with variant versions of the same input. Utilizing convolutional layers in today's CNNs is inspired by Neocognitron architecture proposed by Fukushima *et al.* [128] and popularized by LeCun *et al.* [129]. In the Neocognitron architecture, the authors assumed that because all layers are convolutional, the output of the final layer will not be affected by input translations. The data augmentation is motivated by the fact that, for instance, if we randomly crop the input image, the network will see translated versions of the same object during training. Consequently, the trained network will be invariant to both input translations and the absolute spatial location of objects. Nevertheless, it has been reported in the literature that most of the SOTA CNNs are not robust to input translations [122, 78, 123, 124]. Azulay *et al.* showed that the chance of generating a different output by a CNN after translating its input by only a single pixel could be as high as 30% [124].

Neither data augmentation nor convolutional layers guarantee shift-equivariance. If all layers of a CNN were purely convolutional, input translations would be preserved through all layers then, and the network would be shift-equivariance. However, most modern CNNs contain other layers as well, among which are downsampling layers. It has been suggested that ignoring the Nyquist–Shannon sampling theorem by these downsampling layers is an origin of the shift-variance problem in modern CNNs [124, 78]. The downsampling operation is mostly performed using pooling or strided-convolutional layers with a stride of more than one. These layers have been employed frequently in commonly-used CNNs such as ResNet [130], VGG [131], MobileNetV2[88], U-Net [108]. Similarly, showing translated versions of inputs to the network as a data augmentation method may help the network to learn input translations, but shift-equivariance will be learned merely for similar samples that have been seen before during the training phase. However, the distribution of the training set can be highly biased, and those samples that do not necessarily follow that bias will not take advantage of the learned shift-equivariance. The problem can even be amplified in most medical applications with limited acquired data, where the bias is higher due to the limited number of samples in training sets.

4.1.2 BlurPooling

Consider a band-limited signal, which contains no frequencies higher than $\boldsymbol{B}$ Hz. In order to sample the signal without losing any information, the well-known Nyquist–Shannon sampling theorem in the digital signal processing domain indicates that the sampling rate

must be higher than ***2B*** Hz, or otherwise, the reconstructed signal will suffer from aliasing artifact. As mentioned previously, downsampling layers such as max-pooling and strided-convolutional layers do not necessarily respect the Nyquist–Shannon sampling theorem, which leads to sensitivity to input translations and consequently the shift-variance problem [78, 124].

A well-known signal processing approach for anti-aliasing is applying a low-pass filter before downsampling. Influenced by this idea, Zhang *et al.* [78] proposed merging a low-pass filter with pooling or strided convolutional layers to mitigate the aliasing effect, with minimal additional computation. They referred to the proposed approach as BlurPooling. Although it has been thought that there is a trade-off between blurred-downsampling and max-pooling [132], they demonstrated that they are compatible. This method decomposes pooling and strided-convolutional layers into two separate operations:

1. Reducing the operation's stride to 1. It is equivalent to apply the same operation densely. For instance, for a max-pooling layer with a 2×2 kernel and stride 2, the max() operator will be applied as previously, but with stride 1 instead of 2. This dense operation preserves shift-equivariance.

2. Applying an anti-aliasing filter with an $m\times m$ kernel and finally subsampling with the desired factor. Unlike the previous step, which does not change the input dimension, this step reduces the input dimensions as expected.

Fig. 4.2 illustrates the difference between the implementation of a conventional max-pooling layer and its equivalent BlurPooling (anti-aliased max-pooling) layer. The top path shows a conventional max-pooling layer, which does not respect the Nyquist–Shannon sampling theorem during downsampling leading to aliasing artifact and, consequently, lack of shift-equivariance. The bottom path decomposes this procedure into two steps: 1) A densely-evaluated max-pooling. 2) Applying an anti-aliasing filter followed by a subsampling operation. In the second step, applying the anti-aliasing filter before the subsampling step mitigates the aliasing effect without compromising the advantages of the max() operation. The same concept is also applicable to other downsampling layers, such as strided convolutional layers. This approach suggests augmenting pooling and strided-convolutional layers by low-pass filtering instead of replacing them with averaging operators, which provides shift-equivariance while preserving the advantages of those layers.

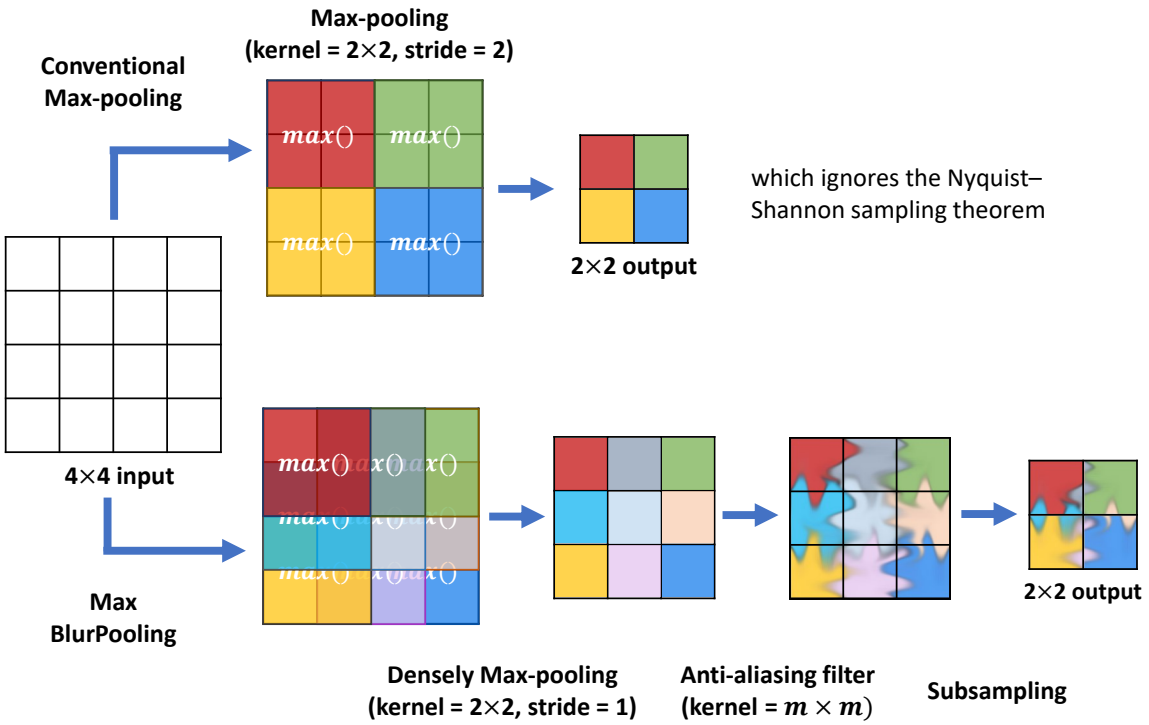


Figure 4.2: Illustration of the difference between a conventional max-pooling layer and its equivalent BlurPooling layer. (Top path) A conventional max-pooling layer. It does not respect the Nyquist–Shannon sampling theorem during downsampling leading to aliasing artifact and, consequently, lack of shift-equivariance. (Bottom path) A blurpooling layer. It decomposes the max-pool operator into two steps: 1) A densely-evaluated max-pooling. 2) Applying an anti-aliasing filter followed by a subsampling operation.

4.2 Methodology

Let $\boldsymbol{I} \in \mathbb{R}^d$ and $\hat{\boldsymbol{S}} \in \{0,1\}^d$ denote a sample input image and the corresponding output segmentation mask, respectively. The segmentation problem can be formulated as

$$\hat{\boldsymbol{S}} = f_{seg}(\boldsymbol{I}, \boldsymbol{\theta}) \tag{4.1}$$

where $f_{seg} : \mathbb{R}^d \rightarrow \{0,1\}^d$ is the segmentation CNN, and $\boldsymbol{\theta}$ are the network's parameters. By training the CNN, an optimizer is utilized to find optimal parameters $\boldsymbol{\theta}^*$ that minimize the error, measured by a loss function L, between predicted mask $\hat{\boldsymbol{S}}$ and ground truth $\boldsymbol{S}$

$$\boldsymbol{\theta}^* = \underset{\theta}{argmin}\ L(\boldsymbol{S}, \hat{\boldsymbol{S}}) \tag{4.2}$$

4.2.1 Network architecture

The U-Net [108] was designed for biomedical segmentation applications to work with a limited number of training samples more efficiently. Showing promising results, it is no wonder that the network attracted growing interest in the ultrasound image segmentation domain, and an extended range of networks has been built upon that.

As the primary goal of this chapter was to investigate the effect of input translations in ultrasound images, we chose the vanilla U-Net as the baseline method to cover an extensive range of previous works. It allows generalizing conclusions of this chapter and makes them applicable to other work that utilized the original or extended versions of the U-Net without considering the shift-variance problem. Findings can also be valid for studies employing any other networks with conventional downsampling layers, such as max-pooling or strided-convolution, as the source of the problem.

The U-Net consists of a contracting path (encoder) followed by an expansive path (decoder), in which the former extracts locality features, and the latter resamples the image maps with contextual information. Skip connections are also employed to produce more semantically meaningful outputs by concatenating low- and high-level features. Without losing generality, we replaced the transposed convolutions in the original U-Net with bilinear upsampling layers in favor of memory efficiency. For investigating the shift-variance problem, we employed three variants of U-Net by altering its downsampling layers as the source of shift-variance. Fig. 4.3 illustrates the network architectures with three different types of downsampling layers in the contracting path. In the first version, referred to as the baseline network, we utilized max-pooling layers similar to the original U-Net. In the second version, referred to as the BlurPooling network, we replaced the original max-pooling layers with BlurPooling layers, where anti-aliasing filters in all four layers have an identical size of $m \times m$. The third version is described in the following subsection.

4.2.2 Pyramidal BlurPooling (PBP)

BlurPooling [78] was originally introduced for mitigating the shift-variance problem and mainly tested in classification applications and on datasets of natural images, such as CIFAR10 [133], and ImageNet [134]. In such cases, the predicted probability of the correct class may drop by applying an anti-aliasing filter to feature maps while it climbs, for instance, from the top-3 values to the top-1 value, which leads to preserving or improving the top-N accuracy. However, a concern with adding anti-aliasing filtering is undermining the accuracy of generated segmentation masks, which is critical in sensitive applications, such

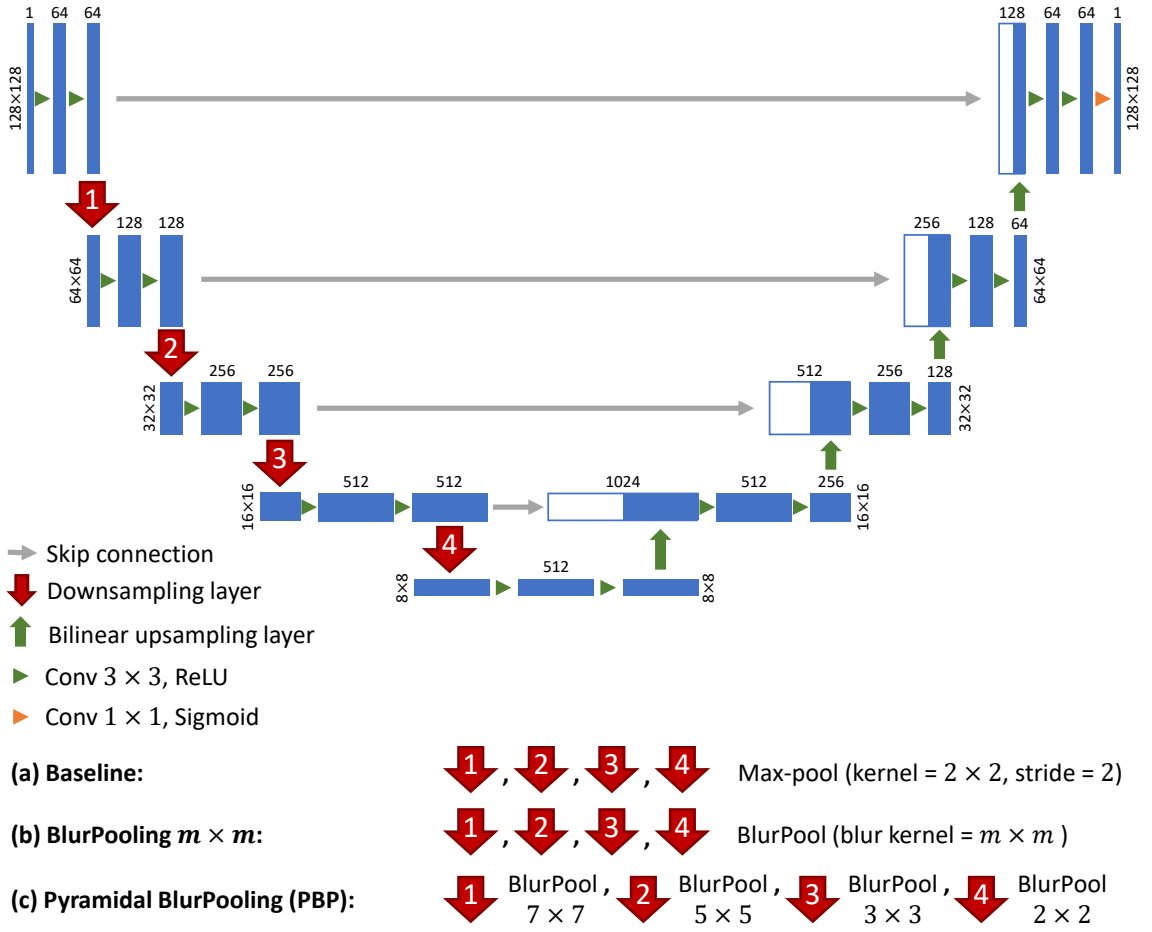


Figure 4.3: Networks architectures. (a) Similar to the vanilla U-Net, max-pooling layers (kernel=2×2, stride=2) were utilized as downsampling layers. (b) Max-pooling layers were altered with their corresponding BlurPooling layers, in which the size of anti-aliasing filters was identical ($m \times m$) for all four layers. (c) Similar to the previous case, BlurPooling layers were utilized instead of max-pooling layers; however, the size of anti-aliasing filters gradually decreased at each downsampling layer from the first to the fourth one.

as medical image segmentation. It has been reported that employing BlurPooling layers in image-to-image translation networks compromises the accuracy [78]. It is shown that while the quality of the generated image holds for small anti-aliasing filters, there is a trade-off between quality and shift-invariance for larger ones.

To address this trade-off, we proposed using a pyramidal stack of anti-aliasing filters. As shown in Fig. 4.3, filters were applied to densely max-pooled feature maps of sizes $128\times128\times c$, $64\times64\times c$, $32\times32\times c$, and $16\times16\times c$ from the first to the fourth downsampling layer, respectively, where c is the number of channels. In Fig. 4.3 (c), labeled as pyramidal blur-pooling (PBP), instead of using anti-aliasing filters of the same size for all

downsampling layers, we started with a filter of size 7×7 and made it smaller at each downsampling layer. Moving from the first layer to the last one, more energy of the feature maps concentrates at lower frequencies. Therefore, the shift-equivariance effect can be preserved with a smaller anti-aliasing filter at deeper layers without compromising accuracy.

4.2.3 Anti-aliasing filters

In this chapter, we utilized similar smoothing filters used in Laplacian pyramids [78, 135, 136]. Let $\boldsymbol{H_s}$ and $\boldsymbol{w_s}$ denote the $m \times m$ filter, and a vector of size $m \times 1$, respectively

$$\boldsymbol{H_s} = \boldsymbol{w_s} \otimes \boldsymbol{w_s} \tag{4.3}$$

where operator $\otimes$ represents the outer product. For the filter $m = 2$, we simply chose $\boldsymbol{w_s} = \frac{1}{2}[1, 1]$. For filters $m = 2k + 1$, $\boldsymbol{w_s}$ was chosen subject to the following constraints [135]:

$$\text{Normalization: } \sum_{n=-(m-1)/2}^{(m-1)/2} \boldsymbol{w_s}[n] = 1 \tag{4.4}$$

$$\text{Symmetry: } \boldsymbol{w_s}[n] = \boldsymbol{w_s}[-n] \quad for\ all\ n \tag{4.5}$$

$$\text{Unimodal: } \boldsymbol{w_s}[n_1] \geq \boldsymbol{w_s}[n_2] \geq 0 \quad for\ 0 \leq n_1 \leq n_2 \tag{4.6}$$

$$\text{Equal Contribution: } \sum_{|n|\ even} \boldsymbol{w_s}[n] = \sum_{|n|\ odd} \boldsymbol{w_s}[n] \tag{4.7}$$

In summary, we obtained anti-aliasing filters by taking the outer product of the following vectors with themselves:

$$\boldsymbol{w_s} = \begin{cases} \frac{1}{2}[1, 1], & m = 2 \\ \frac{1}{4}[1, 2, 1], & m = 3 \\ \frac{1}{16}[1, 4, 6, 4, 1], & m = 5 \\ \frac{1}{64}[1, 6, 15, 20, 15, 6, 1], & m = 7 \end{cases} \tag{4.8}$$

4.2.4 Datasets

Synthetic dataset

We simulated a synthetic dataset as a reference wherein ground truths are error-free and independent of radiologists' bias to quantify the shift-variance problem in such a scenario. A total of 163 ultrasound images were simulated using the publicly available Field II simulation package [79, 80] containing 100,000 scatterers uniformly distributed inside a phantom of size 50 mm $\times$ 10 mm $\times$ 50 mm in x, y, and z directions, respectively. All phantoms were positioned at an axial depth of 30 mm from the face of the transducer, and each contained an anechoic region with a different shape. To generate those anechoic regions, instead of using arbitrary shapes, we took corresponding ground truth masks of 163 images of the UDIAT ultrasound breast dataset [53] and resampled them with the same size as the phantom. Then we assigned a zero amplitude to scatterers that were located inside the mask. In each simulation, we set the transmit focus at the center of mass of the mask. Finally, we considered masks as the exact ground truths of the simulated images. We split simulated images into three training, validation, and test subsets, each containing 100, 30, and 33 samples, respectively. The simulation parameters are summarized in Table 4.1.

Table 4.1: Field II parameters for the synthetic dataset.

Parameter	**Value**
Speed of Sound	1540 m/s
Center Frequency	8.5 MHz
Subaperture Size	64 elements
Number of Scan Lines	100
Element Height	5 mm
Element Width	Equals to wavelength
Kerf	0.05 mm

UDIAT dataset

We used a publicly available ultrasound breast images dataset, collected from the UDIAT Diagnostic Centre in 2012 with a Siemens ACUSON Sequoia C512 system and a 17L5 HD linear array transducer (8.5 MHz) [53]. It is comprised of 163 breast B-mode ultrasound

images, with a mean image size of 760×570 pixels, containing lesions of different sizes at different locations. Lesions were categorized into two classes: benign lesions and cancerous masses, with 110 and 53 samples, respectively. Most of the lesions in the dataset were hypoechoic regions, in which the intensity of the lesion was lower than its background. The dataset also contained respective delineations of the breast lesions as ground truths in separate files, which had been obtained manually by experienced radiologists. We split images into three training, validation, and test subsets, each containing 100, 30, and 33 samples, respectively. Besides, in each subset, we approximately preserved the original ratio of samples per class.

Baheya dataset

To investigate how increasing the number of training samples may affect the shift-variance problem in different scenarios such as with or without augmentation, we employed a larger publicly available dataset collected at Baheya hospital with a LOGIQ E9 ultrasound system equipped with an ML6-15-D Matrix linear probe (1-5 MHz) [137]. It contained 780 breast ultrasound images with an average image size of 500×500 pixels, collected from 600 female patients with ages ranging from 25 to 75 years old. The dataset was categorized into three classes: normal, benign, and malignant each with 133, 437, 210 cases, respectively. The ground truths, delineated manually by expert radiologists, were presented along with original images. In the case of images containing more than one lesion, each lesion's ground truth had been stored in a separate file. Therefore as a preprocessing step, in those cases, we merged ground truths to have one ground truth file per image. Besides, as the main focus of this chapter was to investigate the effect of input translations on the predicted segmentation mask, we did not utilize images of the normal class and split the remaining 647 images into three subsets: training, validation, and test, each containing 387, 130, and 130 samples, respectively. Although we reproduced the original classes' ratio in the subsets, we could not separate samples at the subject level because the required information was not available. However, since we fixed the same subsets for all experiments, it did not affect the comparison purposes.

Mixed ultrasound dataset

To explore the effect of combining two datasets on output consistency, we concatenated training, validation, and test subsets of the UDIAT dataset with the corresponding ones of the Baheya dataset to create the new dataset with training, validation, and test subsets, each

containing 487, 160, and 163 samples, respectively.

Brain magnetic resonance imaging (MRI) dataset

Ultrasound scans are among the most challenging images for CNNs due to several complexities such as speckle noise, non-uniform resolution, and ambiguous boundaries. To evaluate the generalizability of the study and severity of the shift-variance problem on less challenging and even larger datasets, we employed the low-grade glioma segmentation dataset from TCIA (The Cancer Imaging Archive) [138, 139], which is comprised of magnetic resonance images from 110 patients' brain. Image sizes were 256×256 pixels, and the number of slices ranged from 20 to 88. We divided patients into three subsets: training, validation, and test composed of 66, 22, and 22 patients, respectively. Although pre-contrast and post-contrast sequences were also provided, we utilized only fluid-attenuated inversion recovery (FLAIR) with their corresponding manual abnormality segmentation masks. Besides, we only used slices containing at least one abnormality and their corresponding ground truths as two-dimensional images. Finally, our training, validation, and test sets were comprised of 833, 268, 272 samples, respectively.

For all datasets, we resampled the images and their corresponding masks to an identical size of 138×138 pixels. Sample images from each dataset are shown in Fig. 4.4

4.2.5 Training Strategy

In total, we trained 500 networks from scratch. For each dataset, we trained 5 different networks without data augmentation: a baseline network, three BlurPooling networks with anti-aliasing filters of sizes 3×3, 5×5, and 7×7, and finally a PBP network. Then we trained another 5 networks corresponding to the previous ones and with the same configurations, while this time, data augmentation was applied. Finally, due to small training datasets, we repeated each training 10 times with different random initializations to mitigate the randomness out of the interest of this study. Although those 10 initializations were randomly generated using LeCun method [140], we used 10 fixed initializations and trained the corresponding networks with identical initializations. For instance, if the first repeat of the baseline network (without data augmentation) trained with initialization #1, the first repeat of the rest 9 networks, i.e., baseline with data augmentation, BlurPooling $m \times m$, and PBP networks (with and without data augmentation) were also initialized with initialization #1. The same 10 initializations were similarly used across all datasets.

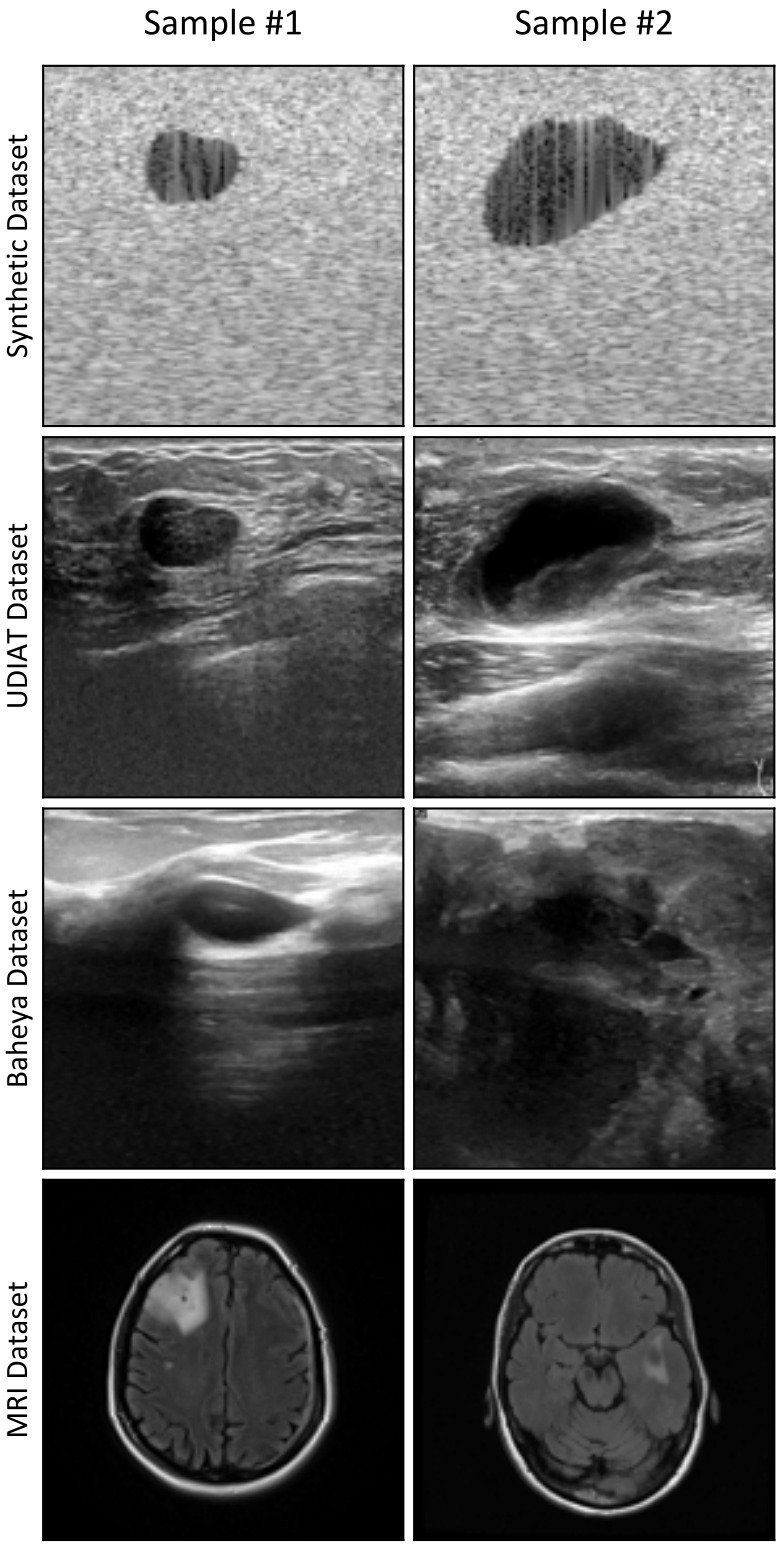
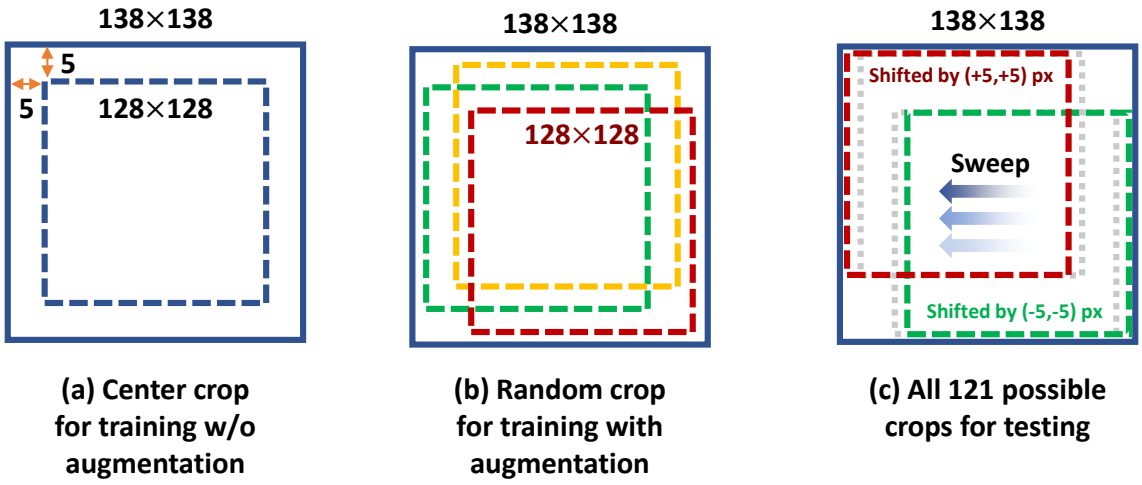


Figure 4.4: Sample images from datasets employed in experiments. For UDIAT and Baheya datasets, samples #1 and #2 belong to benign and malignant categories, respectively.



(a) Center crop for training w/o augmentation

(b) Random crop for training with augmentation

(c) All 121 possible crops for testing

Figure 4.5: To mimic input translations, we resampled all images to a size of 138×138 in the first stage. (a) For training without data augmentation, where translations were not required, we always center cropped a 128×128 square. (b) For training with data augmentation, at each iteration, a 128×128 square was cropped at a random location. (c) For testing, we mimicked all 121 possible translations. For instance, cropping the top-left region mimics a translated version with respect to the center cropped version where it is shifted by +5 pixels in both horizontal and vertical directions.

Since DSC is one of the most common metrics for evaluating medical image segmentation, we chose the loss function based on this metric that quantifies the area overlap between the predicted and ground truth masks

$$DSC(\boldsymbol{S}, \hat{\boldsymbol{S}}) = \frac{2\left|\boldsymbol{S} \cap \hat{\boldsymbol{S}}\right| + \varepsilon}{|\boldsymbol{S}| + \left|\hat{\boldsymbol{S}}\right| + \varepsilon} \tag{4.9}$$

$$L(\boldsymbol{S}, \hat{\boldsymbol{S}}) = 1 - DSC(\boldsymbol{S}, \hat{\boldsymbol{S}}) \tag{4.10}$$

Moreover, the sigmoid function was employed as the activation function of the last layer, and the learning rate and batch size were 2×10^{-4} and 24, respectively. We utilized AdamW [141], a variant of Adam [95] as the optimizer, and set the weight decay parameter to 10^{-2}. AdamW yields a better regularization by decoupling the weight decay from the optimization steps taken with regard to the loss function. To avoid overfitting, in addition to applying a weight regularization strategy, an early stopping strategy was also pursued to stop training when the validation loss stops improving after 100 epochs. For each training

epoch, we saved model weights only if the validation loss had been improved and finally used the best weights for testing. We used the same configuration for all experiments. They were programmed using the PyTorch package [142], and training was performed on a single NVIDIA TITAN Xp GPU with 12 GB of memory.

4.2.6 Augmentation

One may wonder that instead of modifying the downsampling layers, the shift-equivariance in CNNs can be achieved by showing the networks the translated versions of images in the training set. To evaluate the effectiveness of this solution and compare it with the other approach, we conducted each experiment both with and without data augmentation. For the former case, as shown in Fig. 4.5 (a), we trained networks with center cropped images of size 128×128 pixels. For the latter case, since the focus was on achieving shift-equivariance, we augmented the data merely by input translations and isolated experiments from other transformations. To this end, we applied an on-the-fly augmentation by randomly choosing a different square with a size of 128×128 within the original image at each training epoch (Fig. 4.5 (b)). Since the stopping strategy was based on the validation set, we always used center cropped images for validation, even for augmented training cases.

4.2.7 Evaluation Metrics

Accuracy

To report the accuracy, we used the segmentation error defined in Eq. (4.10). The error range is [0, 1], and it is closer to 0 for more accurate predictions.

Consistency

To quantify shift-equivariance, we evaluated the output consistency across input translations. To this aim, as shown in Fig. 4.5 (c), we mimicked input translations by translating a sample image $\boldsymbol{I}$ from the test set by i pixels horizontally and j pixels vertically with respect to its center cropped version, where $\{i, j \in \mathbb{N} \mid -5 \leq i, j \leq 5\}$. We labeled such a translation as (i, j) and the translated input as $\boldsymbol{I}_{ij}$. Then we obtained corresponding output segmentation masks $\hat{\boldsymbol{S}}_{ij}$ for all 121 possible translated versions. Finally, the variance over segmentation errors was calculated for the test sample $\boldsymbol{I}$, where a lower variance denotes

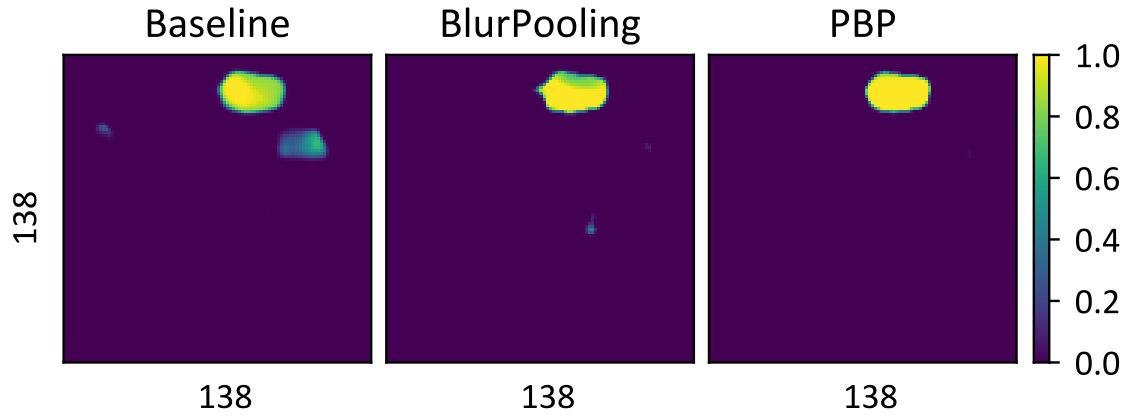


Figure 4.6: Illustration of the effect of input translations on generating segmentation masks. An identical input image from the test set was translated by (i, j) pixels, where $\{i, j \in \mathbb{N} \mid -5 \leq i, j \leq 5\}$. Each network generated output segmentation masks corresponding to those 121 translated inputs. Outputs were compensated for translations with respect to the reference and finally averaged over all translations. (Left) Baseline network. (Middle) BlurPooling with an anti-aliasing filter of size 7×7. (Right) PBP U-Net.

more consistency and higher shift-equivariance.

$$L_{i,j} = L(\boldsymbol{S}_{ij}, \hat{\boldsymbol{S}}_{ij}) \tag{4.11}$$

$$Error\ Mean\ = \overline{L_{i,j}} = \frac{1}{121} \sum_{i=-5}^{i=5} \sum_{j=-5}^{j=5} L_{i,j} \tag{4.12}$$

$$Error\ Variance\ = \frac{1}{120} \sum_{i=-5}^{i=5} \sum_{j=-5}^{j=5} (L_{i,j} - \overline{L_{i,j}})^2 \tag{4.13}$$

4.3 Results

As mentioned before, we trained three types of networks using each dataset: 1) a vanilla U-Net with conventional max-pooling layers referred to as the baseline, 2) three U-Nets, in which max-pooling layers had been replaced with corresponding BlurPooling layers with anti-aliasing filters of sizes 3×3, 5×5, and 7×7, and 3) a U-Net with PBP layers as downsampling layers.

Fig. 4.1 illustrates output segmentation masks corresponding to an identical sample

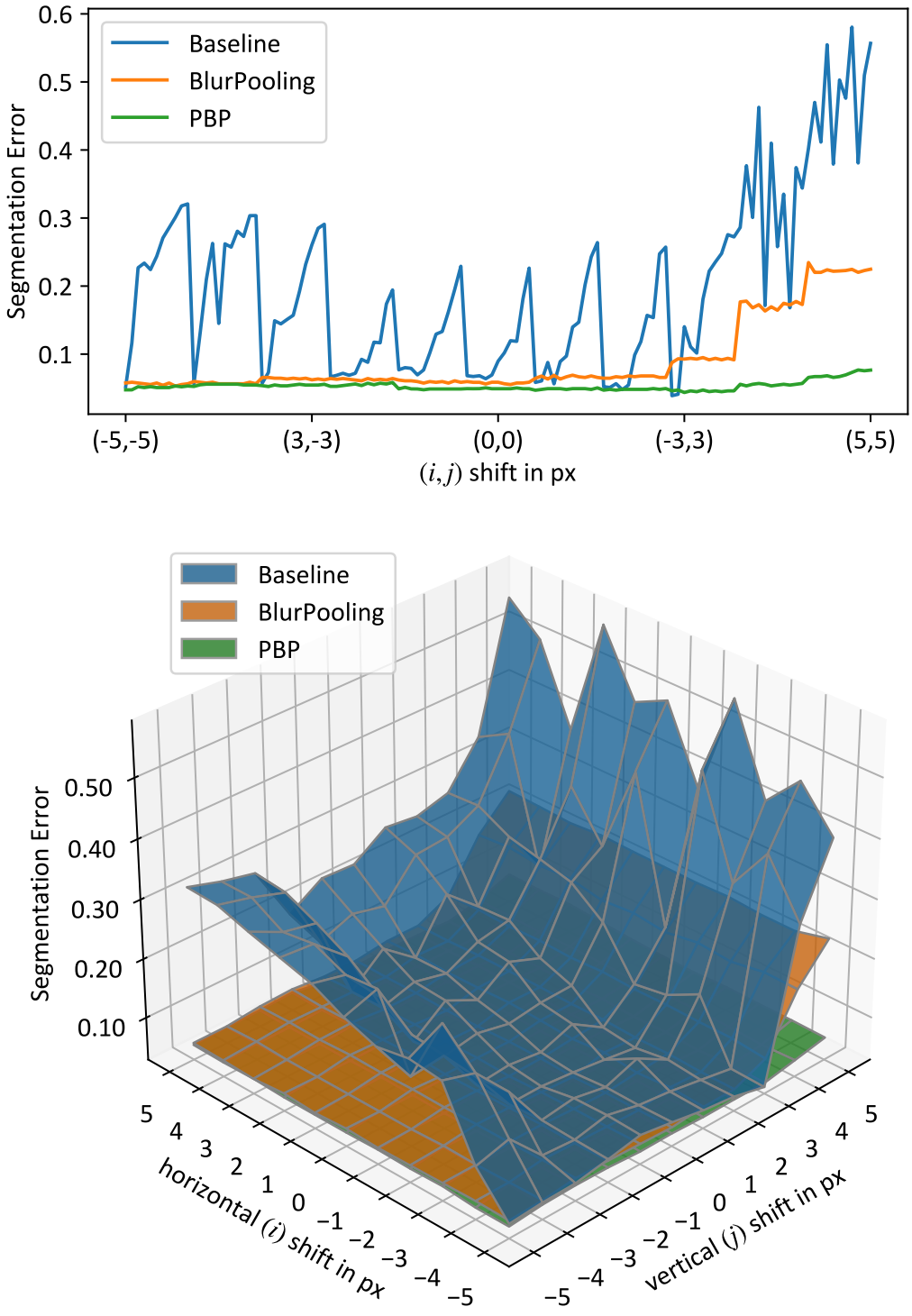


Figure 4.7: Segmentation errors for 121 translated versions of an identical input image from the test set. The input image was translated for (i, j) pixels, where $\{i, j \in \mathbb{N} \mid -5 \leq i, j \leq 5\}$. (Top) The 2D view wherein i changes faster than j from -5 to $+5$. (Bottom) Same values in a 3D view.

from the test set of UDUAT dataset as an input. The input was translated diagonally for (k, k) pixels, where $k \in \{-3, -1, 0, +1, +3\}$. The left column displays the inputs, and the following columns show the output segmentation masks generated by the baseline method with conventional max-pooling layers, the BlurPooling method with an anti-aliasing filter of size 7×7, and the proposed method with PBP downsampling layers, respectively. In this figure, we limited the results up to 5 translated versions for brevity. For better visualization and to provide a more trustworthy comparison, we generated corresponding outputs for all 121 translated versions of the same test sample. Then we compensated for translations by zero-padding in the output segmentation masks regarding the reference and finally averaged across all translated versions. The results are presented in Fig. 4.6, where the left, middle, and right images show averaged outputs for baseline, BlurPooling, and PBP methods, respectively. Moreover, the values of segmentation errors corresponding to those 121 translated versions are plotted in Fig. 4.7 in both 2D (top) and 3D (bottom) views.

To quantify the output accuracy and output consistency for each method, as mentioned in Section 4.2.7, we calculated the mean and variance of the segmentation errors for each test sample $\boldsymbol{I}$ over its all translated versions using Eqs. (4.12) and (4.13). Then we averaged the results over the 10 training repeats with random initializations, and finally over the test set. The whole procedure was completed for each method with and without data augmentation during training, and the results for each dataset are illustrated in Fig. 4.8.

4.4 Discussion

4.4.1 Consistency

Fig. 4.1 sheds light on the shift-variance problem in CNN-based methods and provides a visual perception of the importance of output segmentation consistency. While diagonally translated inputs in the left column look similar, it can be seen that the baseline method generated substantially different output segmentations in the second column. It demonstrates that translating the input even by one pixel may drastically alter the output segmentation mask. In the third column, applying the BlurPooling method alleviated the problem and improved the output consistency; however, it came at the cost of losing accuracy at the lesion's boundaries. The fourth column demonstrates how the proposed method mitigated the problem and improved the output consistency without compromising accuracy by replacing conventional max-pooling layers with pyramidal BlurPooling layers. Note that shift values are merely a convention during the test phase, and it is not expected to achieve a lower error

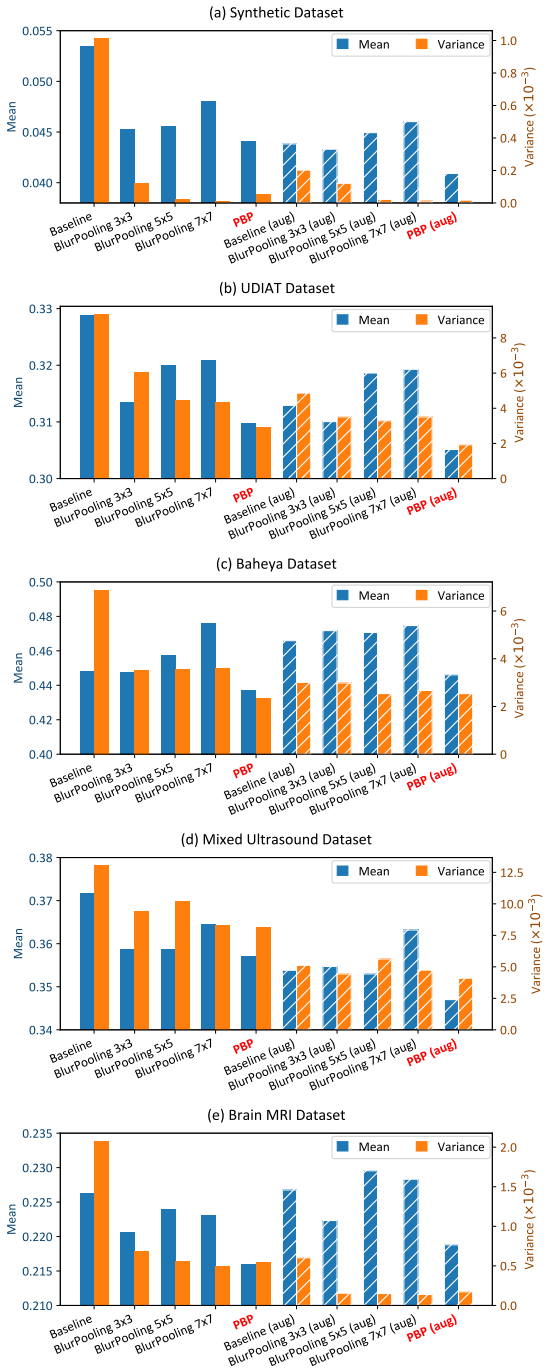


Figure 4.8: Comparison of segmentation accuracies, as well as output consistencies. Hatched and solid bars represent training networks with and without data augmentation, respectively. The results are demonstrated for (a) the synthetic dataset, (b) the UDIAT dataset, (c) the Baheya dataset, (d) the mixed ultrasound dataset, and (e) the brain MRI dataset. Lower error mean and error variance are better and indicate higher accuracy and consistency, respectively.

for the no-shift case since networks are not aware of the origin. Fig. 4.6 gives a broader view of the problem, where output segmentation masks are averaged over 121 translated versions. In addition to detecting wrong regions as the lesion for some translated versions, the baseline method failed to detect the lesion consistently, and the predicted mask at the right side of the lesion is very blurry. Conversely, the BlurPooling and proposed methods offered a more robust prediction across translated inputs, where the proposed method achieved a higher accuracy.

Fig. 4.7 confirms the same concept as well, where the error generated by the baseline method fluctuates more intensively over input translations compared to the BlurPooling and the proposed method.

In Fig. 4.8, it can be observed that applying BlurPooling $m \times m$ layers always reduced error variance (improved consistency) compared to the baseline method. As expected, the general trend is that applying progressively stronger low-pass filters yields higher output consistency. Results for the synthetic, UDIAT, and MRI datasets (without augmentation) confirm that BlurPooling layers with filters of sizes, for instance, 5×5 and 7×7 led to lower error variances in comparison with a 3×3, where we can see that larger filters almost achieved a zero error variance for the simple and ideal synthetic dataset. However, in some cases, such as the Baheya dataset (without augmentation), increasing the filter size did not make a considerable difference in the consistency. Because this dataset was more challenging and the output generated by the network had an intrinsic level of uncertainty. Therefore, some degree of error variance was inevitable, and further improvements were hindered by the saturated consistency, even by increasing the low-pass filter size. Results for the mixed ultrasound dataset suggest that combining two datasets from different distributions may lead to a lower output consistency compared to each one of the datasets separately.

4.4.2 Accuracy

Fig. 4.8 shows that for all datasets, except for the Baheya one, utilizing BlurPooling $m \times m$ layers yielded a better accuracy compared to the baseline for non-augmented cases across all m values. It may seem surprising as we generally expect to see degradation in accuracy after applying a low-pass filter. However, as the results demonstrated, it is not always the case, and BlurPooling layers may improve accuracy as well for three reasons: 1) Considering the fact that applying low-pass filters did not add any learnable parameters to the network, they might act as a regularization method and control the capacity of CNN to prevent overfitting. This improvement has been observed in classification applications as

well [78]. 2) It is worth noting that low-pass filters were not applied on feature maps directly but on densely max-pooled maps instead. As shown in Fig. 4.2 (bottom path), this approach enables the network to take advantage of information that was supposed to be entirely ignored. Therefore, exploiting that information might enhance the accuracy. 3) Although low-pass filters were originally applied to mitigate the shift-variance problem, they may be effective in suppressing spurious noise sources, such as speckle noise or other artifacts in the signal that may make the learning process more challenging. In Fig. 4.8 (a), the results demonstrate that the baseline method with the explained configurations managed to achieve a DSC as high as 94.6% for the synthetic dataset. Interestingly, applying BlurPooling layers improved the accuracy even in this case because of the aforementioned reasons. For the Baheya dataset in Fig. 4.8 (c), utilizing BlurPooling 3×3 layers could not improve the accuracy in comparison with the baseline; however, it managed to preserve the accuracy while decreased the error variance from 6.9×10^{-3} to 3.5×10^{-3}, i.e., improved consistency by almost a factor of 2. Comparing the results across BlurPooling $m \times m$ networks shows that accuracy generally decreased by employing a larger filter. It was expected because, at some point, the smoothing effect of the filter compromises its potential advantages, meaning that larger filters may improve consistency at the expense of accuracy.

4.4.3 Pyramidal BlurPooling

Fig. 4.8 also provides a comparison of the proposed method (labeled as PBP) with the baseline and BlurPooling $m \times m$ methods. We demonstrated that the overall trend with increasing filter size was better output consistency and the general trend with decreasing filter size was higher segmentation accuracy. As mentioned in Section 4.2.2, the proposed method employs a pyramidal stack of low-pass filters to combine the best of two worlds, in which the filter size is larger at the first downsampling layer and gradually decreases by moving toward the fourth one, aiming to enhance consistency without compromising the accuracy. As is evident from Fig. 4.8 (b), (c), and (d), for the UDIAT and Baheya, and mixed ultrasound datasets, the proposed method consistently outperformed all baseline and BlurPooling $m \times m$ methods with or without data augmentation, whether in terms of accuracy or consistency. Interestingly, in these cases, the proposed method achieved higher accuracy in comparison with BlurPooling 3×3, as well as better consistency compared to BlurPooling 7×7. As for the former, in the last downsampling layer, low-pass filters

were applied on feature maps with a size of 16×16, where most of the energy of the signal concentrated around zero. Therefore, in the pyramidal BlurPooling, we utilized the smallest possible filter size (2×2) to avoid the unnecessary smoothing effect. As for the latter, regardless of the accuracy, one might expect that the BlurPooling method with an anti-aliasing filter of size 7×7 must always obtain a higher consistency compared to the PBP method. However, accuracy and consistency are not completely decoupled, meaning that changing one of them can affect the other one. For instance, a BlurPooling 7×7 network may hold a high output consistency across most of the input translations; however, in specific circumstances, such as where the lesion is too close to the image borders, the network can fail to detect the lesion for translations toward the borders due to lower accuracy, and consequently, these outliers cause a higher error variance in the final results. Meaning that compare to the BlurPooling 7×7, PBP achieves even more consistency by providing higher accuracy and avoiding those outliers. Fig. 4.1, 4.6, and 4.7 illustrate this case for a sample image where the BlurPooling 7×7 method ended up with a higher error variance (more fluctuations in segmentation error in Fig. 4.7) because of failing to detect the lesion accurately for vertical translations, where the lesion was too close to the image border. This is also confirmed in Fig. 4.8 (a), and (e), where instead of challenging ultrasound datasets containing lesions with vague boundaries and speckle noise, we applied methods on less challenging synthetic and brain MRI datasets. In these cases, while the proposed method provided higher accuracy compared to BlurPooling 7×7, the consistency is slightly worse in the absence of outliers generated by the inaccuracy of BlurPooling 7×7 in challenging translations.

It is worth noting that although using different sets of filter sizes in PBP can lead to different results, we noticed that choosing a set of low-pass filters with slightly different sizes does not dramatically affect the results, and the concept of not using filters with the same size for all feature maps plays the main role. However, as a rule of thumb, to find the size of the first filter, we visually investigated input images and found the largest possible low-pass filter that can be applied to the input image without noticeably changing the boundaries or visibility of the lesions. In our experiments, the inputs had a size of 128×128, and we chose a 7×7 low-pass filter for the first downsampling layer and decreased the filter size gradually moving toward deeper layers. The same rule can be applied for larger or smaller inputs. If the input size is too small or the network has more downsampling layers leading to very small feature maps at deeper layers, we also have this option to apply low-pass filters only to more shallow layers and leave the deeper layers with conventional max-pooling to achieve better results.

4.4.4 Data Augmentation

We reported the results of every experiment with and without data augmentation in Fig. 4.8, represented by hatched and solid bars, respectively. In most cases, such as the UDIAT, Baheya, and brain MRI datasets, the proposed method without data augmentation outperformed the baseline method even with data augmentation in terms of both accuracy and consistency. However, even in the remaining cases, the proposed method still outperformed the baseline method of the same category (i.e., with or without augmentation). For data augmentation, both its type and parameters need to be chosen carefully concerning the specifications of the dataset. For more illustration, take the classification of images containing handwritten digits as an example, where augmenting the dataset with rotating or flipping may lead to a wrong label for digit "nine" by changing it to digit "six" or vice versa. Despite this evident example, less obvious scenarios can happen as well. In Fig. 4.8 (c) and (e), we can see that utilizing data augmentation increased error means. In the brain MRI dataset, for instance, skull boundaries might be powerful features for the network. Therefore, randomly cropping a portion of boundaries results in a more challenging learning process leading to lower performance. In this case, the crop size can be considered a hyperparameter that needs to be tuned carefully. It is an advantage for the PBP method, which provides a built-in data augmentation without the aforementioned limitation. Moreover, applying BlurPooling layers on top of the data augmentation improved the consistency even further. It demonstrates that data augmentation cannot be considered as a replacement for the proposed method, and they are not interchangeable concepts.

4.5 Conclusions

The study presented in this chapter was encouraged by the fact that although accuracy is a principal measure in ultrasound image segmentation using CNNs, the consistency of outputs across different tests must not be overlooked due to multiple clinical motivations. According to the rapidly growing exploitation of CNNs in this domain, and based on what has been reported in the literature, we challenged the common assumption that CNNs are shift-invariant. For the first time, we investigated the shift-variance problem in ultrasound image segmentation or even more broadly in ultrasound images. To cover an extensive range of previous studies, we chose U-Net as the baseline network due to widespread utilization of either its vanilla version or its variants by the community. Moreover, we discussed the origin of the shift-variance problem that enables us to generalize the concept of the study

to other networks with different architectures from the U-Net, which still use conventional downsampling layers such as max-pooling or strided-convolutional layers without respecting the Nyquist–Shannon sampling theorem. Demonstrating the existence of the shift-variance problem in the ultrasound image segmentation task, we applied a recently published technique referred to as BlurPooling to mitigate the problem and evaluated its performance with different configurations. For evaluation, we quantified the shift-variance problem using a metric based on error variance and conducted all experiments with and without data augmentation to illustrate that augmentation techniques are not a replacement for modifying downsampling layers. Finally, we presented the Pyramidal BlurPooling method specifically for medical image segmentation, in which the size of blurring kernels decreases gradually at deeper downsampling layers, where more energy of feature maps is concentrated at lower frequencies. Testing on *in-vivo* ultrasound datasets, we demonstrated that the proposed method outperformed the baseline and BlurPooling methods, where it drastically improved the output consistency and, to a lesser extent, segmentation accuracy.

Chapter 5

An Ultra-Fast Method for Simulation of Realistic Ultrasound Images

This chapter is based on our published paper [143].

As mentioned in Chapter 4, CNNs have attracted a rapidly growing interest in a variety of different processing tasks in the medical ultrasound community, including segmentation. However, the performance of CNNs is highly reliant on both the amount and fidelity of the training data. Therefore, scarce data is almost always a concern, particularly in the medical field, where clinical data is not easily accessible. The utilization of synthetic data is a popular approach to address this challenge. However, simulating a large number of images using packages such as Field II is time-consuming, and the distribution of simulated images is far from that of the real images. Herein, we introduce a novel ultra-fast ultrasound image simulation method based on the Fourier transform and evaluate its performance in a lesion segmentation task. We demonstrate that data augmentation using the images generated by the proposed method substantially outperforms Field II in terms of the DSC, while the simulation is almost 36000 times faster (both on CPU).

Simulating ultrasound images have been extensively investigated in the medical context, and several publicly available packages have been released for this purpose [144, 38, 79]. Solving acoustic wave equations in the medium is one of the most well-known approaches to that aim [145], where complex equations make it computationally expensive and relatively slow. Treeby *et al.* used the k-space pseudospectral method to reduce the complexity for modeling nonlinear ultrasound propagation in heterogeneous media with power law absorption [146]. Jensen *et al.* suggested calculating pulsed pressure fields based on the Tupholme-Stepanishen method, wherein shape, excitation, and apodization of the transducer could be set as parameters. They divided the surface into small rectangular

patches and calculated the field in each one to obtain the final field by summing their responses. Besides, they used a far-field approximation instead of the geometric one in favor of a faster calculation compared to the older methods [80].

Another approach for ultrasound simulation is based on ray-tracing methods, where the graphics processing unit (GPU) is employed to simulate the propagation of the ultrasound wavefront as rays. Given scatterers' distribution, this approach is capable of generating the speckle pattern by convolving a PSF with scatterers, while simulating complex ultrasound interactions, such as refractions and reflections. Bürger *et al.* suggested a simulation method based on a convolution-enhanced ray-tracing approach and employed a deformable mesh model. They demonstrated that a better simulation of artifacts is achievable by following the path of the ultrasound pulse [147]. Mattausch *et al.* proposed using interactive Monte-Carlo path tracing for simulation of complex surface interactions, which enables more realistic simulation of tissue interactions, such as soft shadows and fuzzy reflections [148].

Recently, DL-based approaches are also exploited for synthesizing ultrasound images. Zhang *et al.* demonstrated an approach to estimate the probabilistic scatterer from observed ultrasound data by imposing a known statistical distribution on scatterers and learn the mapping between ultrasound image and distribution parameter map by training a CNN on synthetic images [149]. Hu *et al.* proposed a method based on a conditional generative adversarial network (GAN) to simulate ultrasound images at given 3D spatial locations relative to the patient anatomy [150]. Cronin *et al.* investigated a framework that accepts synthetic masks and real images as inputs of a GAN and generates realistic B-mode musculoskeletal ultrasound images that are statistically similar to real images [151]. Liang *et al.* introduced an end-to-end framework for enhancing the structure fidelity and resolution of simulated images by employing a sketch GAN and a progressive training strategy and validated that on the follicle and ovary ultrasound image synthesis [152]. GANs were also adopted for simulating intraoperative ultrasound images of the brain after tumor resection surgery [153], intravascular [154], and kidney ultrasound images [155].

In this chapter, we propose a paradigm shift for ultra-fast simulation of B-mode ultrasound images, which is entirely different from the abovementioned methods and is based on the Fourier transform.

5.1 Ultra-fast simulation

To simulate a new ultrasound image containing lesion(s) with known ground truth, we propose taking a real ultrasound image and an arbitrary mask (as the ground truth) and substituting the phase information of the low-frequency spectrum of the real image with the corresponding information of the mask.

Let I_r, I_m, and $I_s \in \mathbb{R}^{W \times H}$ represent a real ultrasound image, an arbitrary mask, and the new simulated output image, respectively. In addition, let $\mathcal{F}_M(I) : \mathbb{R}^{W \times H} \rightarrow \mathbb{R}^{W \times H}$ and $\mathcal{F}_P(I) : \mathbb{R}^{W \times H} \rightarrow \mathbb{R}^{W \times H}$ denote the magnitude and phase of the Fourier transform $\mathcal{F}$ of the image I:

$$\mathcal{F}(I)(m,n) = \sum_{w=0}^{W-1} \sum_{h=0}^{H-1} I(w,h) e^{-j2\pi(\frac{h}{H}n + \frac{w}{W}m)} \tag{5.1}$$

Accordingly, given $\mathcal{F}_M(I)$ and $\mathcal{F}_P(I)$, $\mathcal{F}^{-1}$ is the inverse Fourier transform that converts back the signal from the frequency domain to the image domain.

$$I = \mathcal{F}^{-1}(\mathcal{F}_M(I), \mathcal{F}_P(I)) \tag{5.2}$$

where $j^2 = -1$, and (5.1) and (5.2) can be implemented using fast Fourier transform (FFT) [156] and inverse fast Fourier transform (IFFT) algorithms, respectively.
Further, let denote with M_α a matrix of size $W \times H$:

$$M_\alpha(w,h) = \begin{cases} 1, & \frac{(w-\frac{W}{2})^2}{(\alpha\frac{W}{2})^2} + \frac{(h-\frac{H}{2})^2}{(\alpha\frac{H}{2})^2} \leq 1 \\ 0, & otherwise \end{cases} \tag{5.3}$$

Finally, given a pair of a real image and an arbitrary mask, the proposed method for simulating a new image can be formulated as:

$$I_s = \mathcal{F}^{-1}(\mathcal{F}_M(I_r), M_\alpha \cdot \mathcal{F}_P(I_m) + (1 - M_\alpha) \cdot \mathcal{F}_P(I_r)) \tag{5.4}$$

where $\alpha \in \mathbb{R}$ is a parameter that specifies the amount of phase information that needs to be replaced, and in this chapter, we set $\alpha = 0.11$. Fig. 5.1 illustrates the proposed method, where (a) is an arbitrary mask I_m, and (b) is a real ultrasound image I_r. After taking the FFT of both images, we replaced the phase information of the low-frequency spectrum of the real image with the corresponding information of the mask. Finally, by taking the IFFT of the modified real image I_r, the simulated image I_s was generated.

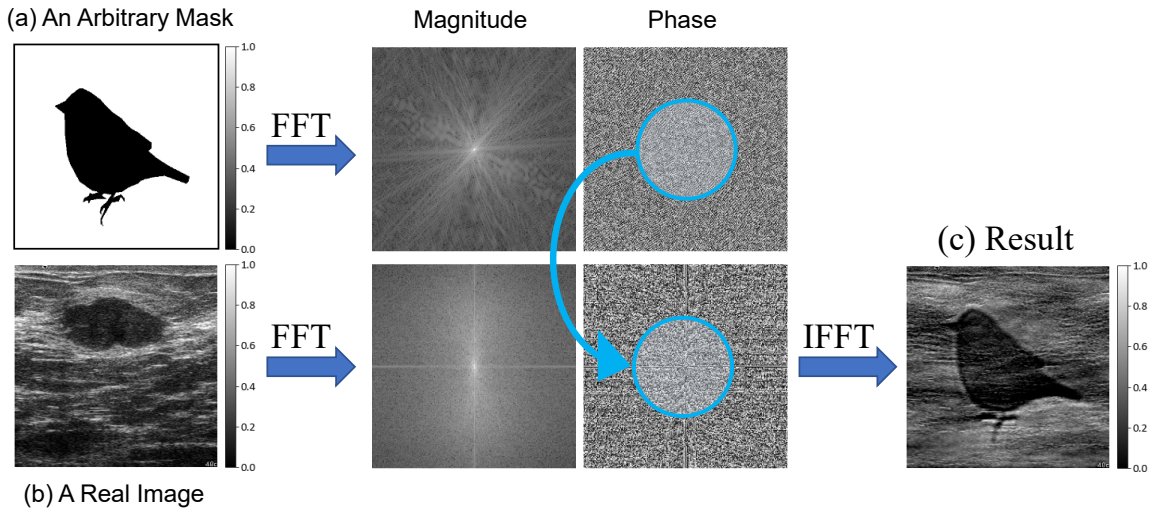


Figure 5.1: Given an arbitrary mask and a real ultrasound image, the proposed method takes the FFT of both inputs and replaces the phase information of the low-frequency spectrum of the real image with the corresponding information of the arbitrary mask to generate the output. (a) An arbitrary mask. (b) A real ultrasound image. (c) The simulated output image.

5.2 Segmentation Task

Given a sample input image $I \in \mathbb{R}^{W \times H}$ and its corresponding output segmentation mask $\hat{S} \in \{0,1\}^{W \times H}$, the segmentation problem can be formulated as:

$$\hat{S} = f_{seg}(I, \boldsymbol{\theta}) \tag{5.5}$$

where W and H are width and height of the image, respectively, $f_{seg} : \mathbb{R}^{W \times H} \rightarrow \{0,1\}^{W \times H}$ is the segmentation CNN, and $\boldsymbol{\theta}$ are the network's parameters. By training the model, an optimizer tries to find optimal parameters $\boldsymbol{\theta}^*$ that minimize the error, measured by a loss function L, between predicted mask $\hat{S}$ and ground truth S

$$\boldsymbol{\theta}^* = \underset{\theta}{argmin}\ L(S, \hat{S}) \tag{5.6}$$

5.2.1 Datasets

In-vivo Dataset

We utilized a publicly available ultrasound breast images dataset, known as Dataset B [53], which was collected in 2012 from the UDIAT Diagnostic Centre with a Siemens ACUSON Sequoia C512 system and a 17L5 HD linear array transducer with a frequency of 8.5 MHz.

The dataset consisted of 163 breasts B-mode ultrasound images from different women with a mean image size of 760×570 pixels, where each one included lesions of different sizes at different locations. Lesions were categorized into two classes of benign and cancerous, with 110 and 53 images in each class, respectively. Corresponding lesion masks were also delineated by experienced radiologists and provided along with the dataset as ground truth masks. We resampled all images to a size of 256×256 pixels and split the dataset into three training, validation, and test sets, each containing 20, 20, and 123 images, respectively.

Field II Dataset

To compare the proposed method with Field II, we simulated 1000 focused images using this publicly available simulation package [80, 79]. Each image contained 100,000 scatterers uniformly distributed inside a phantom of size 50 mm × 10 mm × 50 mm in x, y, and z directions, respectively. Phantoms were positioned at an axial depth of 20 mm from the face of the transducer and centered at the focal point. Besides, we added an anechoic region with an arbitrary shape to each one. To generate those anechoic regions, we took 1000 samples with only one salient object from a publicly available dataset, known as XPIE [104], which contained segmented natural images. Then we discarded natural images and resampled only their ground truth masks with the same size as the phantom. Finally, we assigned a zero weight to the amplitude of those scatterers which were located inside the mask. The advantages of this method were twofold: First, it enabled us to consider the masks as the ground truths of simulated images. Second, we provided the network with a wider range of features compared to regions with limited shapes. Finally, we resampled all images to 256×256 pixels, and split them into two training and validation sets, each containing 800 and 200 images, respectively. Note that this data was merely used for training and validation and did not contain a test set. The parameters of the Field II simulation are summarized in Table 5.1.

Ultra-Fast Dataset

For simulating an image using the proposed method, an arbitrary mask and a real image are required. We took the same 1000 masks from the XPIE dataset, which were used for simulating the Field II dataset, and randomly paired each one with a real image from the training set of the *in-vivo* dataset to simulate 1000 new images. Similar to the Field II dataset, we resampled all images to 256×256 pixels and split them into two training and validation sets, each containing 800 and 200 images, respectively, where there was no need

Table 5.1: Field II parameters for data simulation.

Parameter	**Value**
Sound Speed	1540 m/s
Number of Lines	50
Number of Elements	192
Number of Active Elements	64
Elevation Element Height	5 mm
Element Width	Equals to wavelength
Kerf	0.05 mm

for a test set.

5.2.2 Network Architecture and Training Strategy

We used a vanilla U-Net [108] to evaluate the performance of the proposed method in a segmentation task. U-Net was proposed particularly for biomedical image segmentation, where the size of the training set is small. Its architecture comprises an encoder followed by a decoder, and skip connections are also employed to concatenate low-level features of the encoder with high-level ones in the decoder.

For the training process, we set the learning rate and batch size to 1×10^{-4} and 16, respectively. The AdamW [141], a variant of Adam [95], with a weight decay of 10^{-2} was exploited as the optimizer, and a sigmoid activation function was used for the output layer. The loss function was defined based on the DSC. This metric quantifies the area overlap between the ground truth and predicted masks and was also used for evaluating the segmentation performance:

$$DSC(S, \hat{S}) = \frac{2\left|S \cap \hat{S}\right| + \varepsilon}{|S| + \left|\hat{S}\right| + \varepsilon} \tag{5.7}$$

where ε is a small number that prevents numerical instability for small masks. After each training or fine-tuning epoch, model weights were saved only if the validation loss had been improved, and finally, the best model was used for testing. Experiments were implemented using the PyTorch package [142] and run on an NVIDIA TITAN Xp GPU with 12 GB of memory.

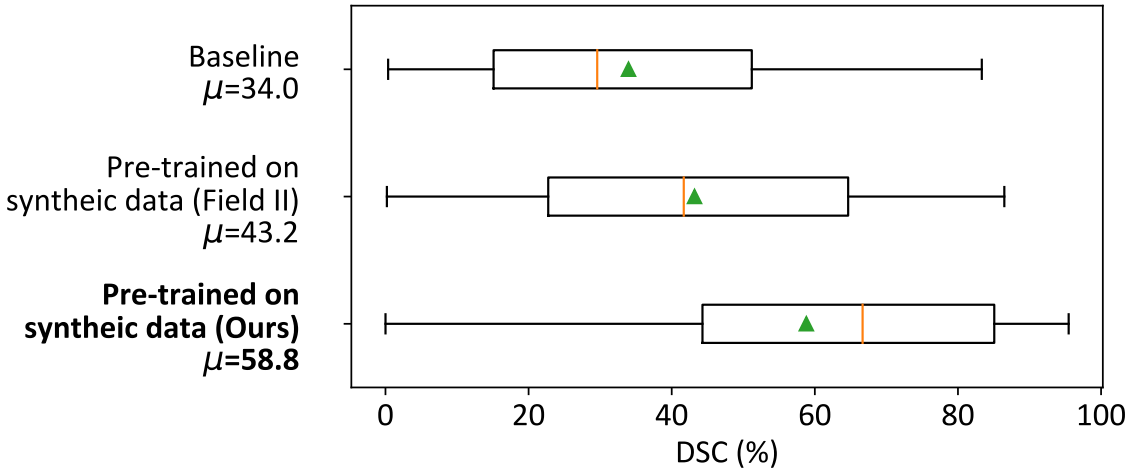


Figure 5.2: Comparison of DSC over the *in-vivo* test set achieved by three conducted experiments. (Top) Training the network from scratch merely using *in-vivo* training set. (Middle) Pre-training the network using synthetic data simulated by Field II and then fine-tuning on the *in-vivo* training set. (Bottom) Pre-training the network using synthetic data simulated by the ultra-fast proposed method and then fine-tuning on the *in-vivo* training set. The triangle and vertical line represent the mean and median, respectively.

5.3 Results

To assess the proposed method for improving the performance of the segmentation task, we conducted three separate experiments. In the first one, labeled as the baseline, a network was trained merely using 20 training images of the *in-vivo* dataset for 200 epochs. In the next experiment, first, the network was trained using 800 training images of the Field II dataset for 150 epochs, and then it was fined-tined using 20 training images of the *in-vivo* dataset for 50 more epochs. Finally, as the last experiment, we repeated the second one, except that instead of the Field II dataset, the ultrafast dataset was employed for pre-training the network. As mentioned before, to avoid data leakage, only the training set of the *in-vivo* dataset had been used for simulating the ultrafast dataset.

Fig. 5.2 shows the DSC results over the test set of the *in-vivo* dataset for all three experiments. As expected, the baseline method achieved the lowest mean DSC due to training on 20 real ultrasound images and without pre-training on simulated images. The second experiment obtained a better performance by taking advantage of pre-training on synthetic images simulated by Field II and then fine-tuning on real ultrasound images. Finally, the third experiment demonstrated that pre-training the network on synthetic data simulated by the proposed method achieved 24.8% higher mean DSC than the baseline

experiment and outperformed Field II simulations by 15.6% improvement in mean DSC. Another advantage of the proposed method is that most of its computational cost is devoted to taking FFT and IFFT, which made simulating the ultra-fast dataset almost 36000 times faster than the Field II dataset using the same CPU.

5.4 Conclusion

We introduced a novel ultra-fast approach based on the Fourier transform for simulating ultrasound images. In this approach, in contrast with the existing methods such as solving acoustic wave equations, employing ray-tracing, or using GANs, we proposed replacing the phase information of the low-frequency spectrum of a real ultrasound image with the corresponding information of an arbitrary mask to simulate a new image containing lesion(s) with known ground truth. We assessed the utility of this method in a lesion segmentation task, where a U-Net was pre-trained using synthetic data. We demonstrated that images simulated by the proposed method outperformed Field II simulations in terms of improving the mean DSC by 15.6%, while the simulation was almost 36000 times faster (both on CPU).

Chapter 6

Ultrasound Domain Adaptation Using Frequency Domain Analysis

This chapter is based on our published paper [157].

As mentioned in the previous chapter, utilizing synthetic data to train networks is a popular approach to address the challenge of limited data availability in the medical field. Although this method can alleviate the issue, models trained on synthetic data often face difficulties in generalizing effectively to real-world applications, which involve handling images obtained from various scanners and different protocols [158]. This issue originally comes from the fact that deep neural networks typically assume that both training and test sets have been drawn from the same distribution [159], which is not necessarily true, especially regarding the recent trend of using synthetic data for training. This problem is usually referred to as the domain shift problem, which induces a dramatic performance drop [160].

Domain adaptation methods are a well-known solution to address the domain shift problem and have been investigated in the medical ultrasound domain. Tierney *et al.* proposed a scheme that incorporates both simulated and unlabeled *in-vivo* data to train a beamformer. They employed cycle-consistent generative adversarial networks to map between simulated and *in-vivo* data in both the input and ground truth target domains [161, 162]. Ying *et al.* introduced a multi-scale self-attention unsupervised network for domain adaptation between labeled thyroid ultrasound images and unlabeled ones in a different domain [163]. Meng *et al.* introduced mutual information-based disentangled neural networks for classifying unseen categories of fetal ultrasound images in different domains. They extracted generalizable categorical features by explicitly disentangling categorical and domain features via mutual information minimization to transfer knowledge to unseen categories in

a target domain [164]. Zhang *et al.* proposed a deep-stacked transformation approach for generalizing medical image segmentation models to unseen domains and evaluated it on segmentation tasks involving MRI and ultrasound modalities [165].

Recently, Yang *et al.* [166] introduced the Fourier domain adaptation (FDA) method in the field of computer vision. They proposed that moving sample A from the source distribution to the distribution of sample B in the target dataset can be achieved by computing the FFT of both samples and substituting the magnitude of the low-frequency spectrum of the source sample with the target sample and finally reconstructing the modified source sample using IFFT. This method is much faster than DL-based methods, and they demonstrated promising results to adapt synthetic dataset GTA5 [167] to the real domain dataset CityScapes [168], which both contain urban street scenes.

We believe this method can perform even better on ultrasound images than urban street scenes because two common differences between synthetic and real ultrasound data are caused by unknown values of attenuation and speed of sound (SOS) in real tissues. Attenuation leads to slow variations in the amplitude of the B-mode image, and a mismatch between the nominal and true values of the SOS creates aberration and subsequent blurring. As such, both of these domain shifts are low-frequency in nature and can be compensated by swapping the low-frequency spectrum of the synthetic and real image.

In this chapter, we exploit the FDA method to mitigate the domain shift problem of synthetic ultrasound images and evaluate its performance in a breast lesion segmentation task.

6.1 Methodology

6.1.1 Datasets

Synthetic Dataset

We simulated 1000 ultrasound images using the publicly available Field II simulation package [79, 80] containing 100,000 scatterers uniformly distributed inside a phantom of size 50 mm $\times$ 10 mm $\times$ 50 mm in x, y, and z directions, respectively. All phantoms were centered at the focal point, positioned at an axial depth of 20 mm from the face of the transducer, and each contained an anechoic region with a different shape. To generate those anechoic regions, we took 1000 samples with only one salient object from a publicly available dataset, denoted as XPIE [104], which contained segmented natural images. Then we discarded natural images and resampled only their ground truth masks with the

same size as the phantom. Finally, we assigned a zero weight to the amplitude of scatterers which were located inside the mask. The advantages of this method were twofold: First, the mask could be considered as the ground truth of the simulated images. Second, we provided the network with an extended range of features as opposed to regions with limited shapes. Finally, we resampled all images to a size of 256×256 and split the dataset into two training and validation sets, each containing 800 and 200 images, respectively. Note that this data was only used for training and validation and did not contain a test set. The simulation parameters were the same as those outlined in Table 5.1.

In-vivo Dataset

We exploited an ultrasound breast images dataset, known as Dataset B [53]. The dataset was publicly available and collected in 2012 from the UDIAT Diagnostic Centre with a Siemens ACUSON Sequoia C512 system and a 17L5 HD linear array transducer. It included 163 breast B-mode ultrasound images containing lesions of different sizes at different locations, with a mean image size of 760×570 pixels. Lesions were categorized into benign and cancerous classes, with 110 and 53 samples in each class, respectively. The dataset also contained respective ground truth masks of the breast lesions, manually obtained by experienced radiologists. We resampled all images to a size of 256×256, and split the dataset into three training, validation, and test sets, each containing 20, 20, and 123 images, respectively.

6.1.2 Fourier Domain Adaptation

To mitigate the domain shift problem, the FDA method [166] suggests replacing the magnitude of the low-frequency spectrum of source samples with target samples. Let $I_s \in \mathbb{R}^{W\times H}$, and $I_t \in \mathbb{R}^{W\times H}$ represent a simulated image using Field II (source) and an *in-vivo* image (target), respectively. Besides, let $\mathcal{F}_M(I) : \mathbb{R}^{W\times H} \rightarrow \mathbb{R}^{W\times H}$ and $\mathcal{F}_P(I) : \mathbb{R}^{W\times H} \rightarrow \mathbb{R}^{W\times H}$ be the magnitude and phase of the Fourier transform $\mathcal{F}$ of the image I, as defined in Eq. (5.1). Accordingly, given $\mathcal{F}_M(I)$ and $\mathcal{F}_P(I)$, $\mathcal{F}^{-1}$ is the inverse Fourier transform that converts back the signal from the frequency domain to the image domain, as indicated in Eq. (5.2). Further, let denote with M_α a mask matrix of size $W \times H$:

$$M_\alpha(w,h) = \begin{cases} 1, & -\alpha < \frac{2w}{W} - 1 < \alpha, -\alpha < \frac{2h}{H} - 1 < \alpha \\ 0, & otherwise \end{cases} \tag{6.1}$$

Finally, given a pair of simulated and *in-vivo* images, the FDA method can be formalized as:

$$I_{s \to t} = \mathcal{F}^{-1}(M_\alpha \cdot \mathcal{F}_M(I_t) + (1 - M_\alpha) \cdot \mathcal{F}_M(I_s), \mathcal{F}_P(I_s)) \tag{6.2}$$

where $\alpha \in (0, 1)$. In this chapter, we set $\alpha = 0.014$. Fig. 6.1 illustrates the FDA method. It shows (a) a simulated image I_s, and (b) a real ultrasound image I_t from the *in-vivo* dataset. After taking the FFT of both images, the magnitude of the low-frequency spectrum of the simulated image has been replaced with the real one. Finally, by taking the IFFT, the output $I_{s \to t}$ has been obtained.

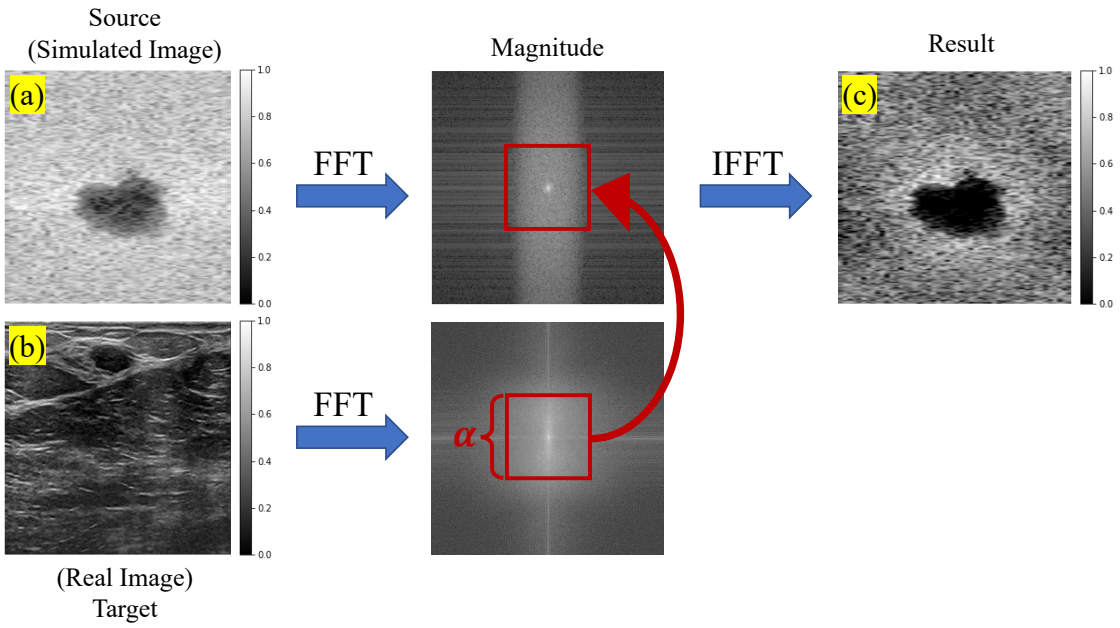


Figure 6.1: The FDA method takes the FFT of simulated and real images, which belong to source and target distributions, respectively. Then it replaces the magnitude of the low-frequency spectrum of the simulated image with the real one. Finally, it obtains the output by taking the IFFT from the modified simulated image. (a) A synthetic ultrasound image, which is simulated using Field II and belongs to the source distribution. (b) A real ultrasound image, which belongs to the target distribution. (c) The output, which seems closer to the target distribution.

6.1.3 Network Architecture and Training Strategy

We used a vanilla U-Net [108] to evaluate the performance of the FDA method on ultrasound images for a segmentation task. The sigmoid function was chosen as the activation

function of the output layer, and the learning rate and batch size were 1×10^{-4} and 16, respectively. The AdamW [141], a variant of Adam [95], with a weight decay of 10^{-2} was exploited as the optimizer. Additionally, we used the DSC, as defined in Eq. 5.7, to evaluate segmentation performance. The loss function was also based on this metric, which quantifies the area of overlap between the ground truth and predicted masks. For each epoch of training or fine-tuning, the model weights were stored only when the validation loss had been improved, and finally, the best weights were used for testing. Experiments were implemented using the PyTorch package [142], and training was performed on an NVIDIA TITAN Xp GPU with 12 GB of memory.

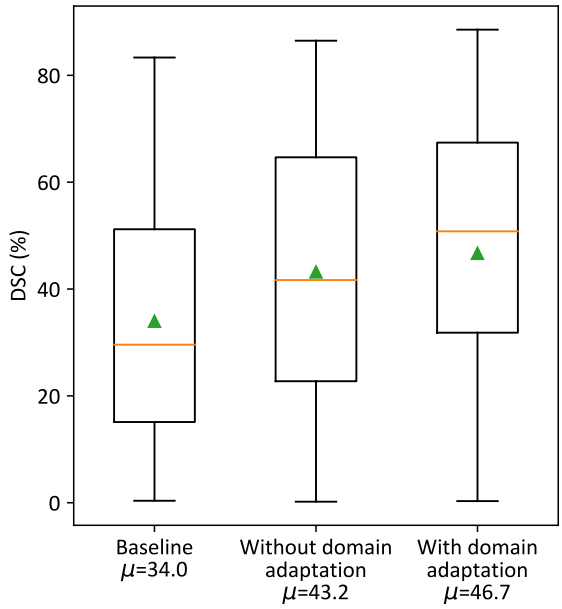


Figure 6.2: Quantitative comparison of DSC over the *in-vivo* test set. (Left) Training the network from scratch merely using *in-vivo* images. (Middle) Use pre-trained weights obtained from training on simulated images without applying the FDA method. (Right) Use pre-trained weights obtained from training on simulated images with applying the FDA method. The triangle and horizontal line represent the mean and median, respectively.

6.2 Results

To assess the effect of the FDA method for mitigating the domain shift problem, we conducted three different experiments. In the first experiment, labeled as the baseline, a network was trained merely using 20 training samples of the *in-vivo* dataset for 200 epochs. In

the next experiment, first, the network was trained using 800 training samples of the simulated dataset without applying the FDA method for 150 epochs. Then it was fined-tined using 20 training samples of the *in-vivo* dataset for 50 epochs. Finally, as the third experiment, we applied FDA to the simulated images and repeated the previous experiment. To apply the FDA method, for each training simulated image and at each iteration, the target image was randomly chosen from 40 images in the training and validation sets of the *in-vivo* dataset. However, we used a fixed set of target images for applying the FDA method on the validation samples. Since the main purpose of using the validation set was to find and save the best model across different epochs, injecting randomness caused by choosing random target images was not desired.

Fig. 6.2 illustrates the DSC results for 123 test set images of the *in-vivo* dataset for the aforementioned three experiments. As we expected, the baseline method led to the lowest DSC due to training on only 20 real ultrasound images without taking advantage of pre-training on simulated images. The second experiment achieved a better performance by pre-training on simulated data and using *in-vivo* images for fine-tuning. However, it suffered from the domain shift problem, where there was a high discrepancy between the distribution of the simulated data and the *in-vivo* data. The third experiment showed a 3.5% improvement in mean DSC, obtained by applying the FDA method.

6.3 Conclusion

In conclusion, we claimed that important differences between simulated and real ultrasound data are low-frequency in nature. We employed the FDA method, which replaces the magnitude of the low-frequency spectrum of a synthetic image with a real one to tackle the issue of domain shift. For the first time, we exploited the FDA method in segmentation of ultrasound images, and more generally in ultrasound imaging. We demonstrated that applying this fast and simple method on simulated ultrasound data can improve the mean DSC as high as 3.5% compared to using simulated data without applying any domain adaptation method.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

To enhance the interpretability of ultrasound images by tackling the phase aberration problem, we proposed two novel phase aberration correction techniques. In the first approach, we trained a CNN to estimate the aberration profile from an ultrasound B-mode image. This estimated profile was subsequently used to compensate for the phase aberration. We experimentally examined the main characteristics of the proposed method and provided a quantitative evaluation of the predicted aberration profile. Results demonstrated that the CNN-based method substantially outperforms the conventional DAS method and another technique based on NNCC. In the second approach, we introduced MAIN-AAA, a DL-based aberration correction technique that did not require ground truth data during the training phase. In this approach, we proposed training a network where both the input and target outputs were randomly aberrated RF data. Additionally, we introduced an adaptive mixed loss function that gradually shifts from B-mode data to RF data as training progresses toward convergence. This proposed loss function achieved superior performance by utilizing smoother B-mode data in the beginning, guiding the optimizer toward a correct solution, and gradually incorporating more fluctuating RF data to fully leverage its richer information. This strategy helped to avoid getting stuck in local minima during the initial stages of optimization. Each of these aberration correction methods presents distinct advantages and limitations. In contrast to MAIN-AAA, which estimates the corrected image using an end-to-end network, the first method offers a more explainable solution by estimating the aberration law before correcting the phase aberration effect. However, MAIN-AAA does not require ground truth data, allowing it to be trained on real data rather than relying on synthetic data. Furthermore, the first method assumes that the aberration

is caused by a near-field phase screen, whereas MAIN-AAA can also mitigate distributed aberrations.

In order to simplify the interpretation of ultrasound images through automatic segmentation, we investigated the shift-variance problem in CNNs and evaluated the effectiveness of replacing max-pooling layers in the encoder of the U-Net with BlurPooling layers to address this issue. Furthermore, instead of using anti-aliasing filters of the same size for all downsampling layers, we proposed the Pyramidal BlurPooling method, which incorporates a pyramidal stack of anti-aliasing filters at each BlurPooling layer to improve the output consistency without compromising accuracy. We demonstrated that the proposed method outperforms BlurPooling in terms of output consistency and segmentation accuracy. To address the scarcity of data in DL-based approaches for ultrasound image segmentation, we introduced a novel ultra-fast image simulation method and evaluated its effectiveness in a lesion segmentation task. Our findings demonstrated that pretraining a U-Net with images generated by this method significantly outperforms the traditional Field II simulation, as measured by the DSC. Notably, the proposed simulation process is nearly 36,000 times faster than Field II when both methods are executed on a CPU. A limitation of the proposed method is that it is only applicable to segmentation tasks and may not be as effective for other tasks, such as classification. In scenarios where data is available but from different domains, we employed a domain adaptation method based on frequency domain analysis, which replaces the magnitude of the low-frequency spectrum of an image from the source distribution with an image from the target distribution. Results demonstrated the effectiveness of this approach in enhancing the performance of a breast lesion segmentation task.

7.2 Future Work

Phase aberration is a significant factor contributing to the degradation of ultrasound image quality, and mitigating its effects remains an active area of research. In Chapter 2, we demonstrated that a CNN can effectively estimate aberrator profiles from B-mode images, providing an explainable approach to mitigate aberration by compensating for the estimated delays. Estimating the aberration profile is a crucial step in transcranial ultrasound, not only for imaging but also for therapeutic applications, a class of non-invasive techniques that utilize ultrasound energy to alter brain structure or function, for instance, to modulate brain activity [169]. In transcranial ultrasound, the skull is often modeled as a near-field phase screen, where auxiliary imaging modalities such as MRI or computed tomography

are employed to estimate the aberration profile according to the skull thickness and density [170, 171]. This profile is then used to correct the strong phase aberration induced by the skull, both for image reconstruction and for focusing energy beams during therapy. An interesting avenue for future research involves investigating the application of the proposed CNN-based method in transcranial ultrasound by estimating the aberration profile directly from ultrasound images, potentially eliminating the need for auxiliary modalities.

In transcranial ultrasound, the near-field phase screen model can be considered an effective aberration model due to the presence of the skull. However, there are scenarios where the aberration is spatially distributed throughout the medium, making a near-field phase screen model insufficient. As previously mentioned, the MAIN-AAA method introduced in Chapter 3 can effectively address distributed aberrations. However, this method estimates the corrected data directly, bypassing the estimation of the aberration law, which results in a less explainable approach. In future work, the explainable method proposed in Chapter 2 can be extended to address distributed aberrations by training a CNN to generate a three-dimensional aberration law rather than estimating a single one-dimensional aberration profile for the entire medium. In this framework, the output would represent the aberration profiles along one dimension, while the other two dimensions correspond to the axial and lateral positions of different regions within the medium. Based on the physical principles of ultrasound, abrupt changes in delays between regions are unlikely, provided the regions are sufficiently small. To account for this, regularization terms can be incorporated into the loss function to enforce continuity both element-wise and region-wise, thereby improving the network's performance.

Regarding the automatic segmentation of ultrasound images, we investigated the shift-variance problem and introduced PBP layers to mitigate this issue. A limitation of this segmentation method is its training on only two relatively small ultrasound datasets due to data scarcity, which restricts its generalizability when applied to images obtained from diverse organs, devices, or settings. To address this challenge, we proposed an ultrafast method for simulating ultrasound images and utilized domain adaptation techniques. Recently, foundation models such as SAM [172] and SAM2 [173] have gained considerable interest as highly generalized models for segmentation tasks. These transformer-based models are pre-trained on an excessive number of images, including over 1 billion masks. While foundation models have shown promising results in zero-shot natural image segmentation tasks, their performance on ultrasound images tends to be relatively lower due to the inherent differences between natural and ultrasound images, compounded by the general scarcity of medical ultrasound data. In parallel, stable diffusion models have recently been

employed to address data scarcity in medical imaging by generating an arbitrary number of synthetic images with diverse characteristics [174]. Several avenues for future research include evaluating the performance of foundation models in the segmentation of ultrasound images, investigating the shift-variance problem within these transformer-based models, and exploring potential solutions to alleviate this issue. Furthermore, fine-tuning these foundation models using datasets of real ultrasound images, augmented with synthetic images generated by stable diffusion models, may offer enhanced performance.

In ultrasound segmentation, we aim to delineate anatomical targets or lesions within organs such as the liver, kidneys, or brain, which are often obscured by aberrating layers like fat or the skull. Another direction for future research involves assessing the impact of phase aberration correction methods on segmentation accuracy within a sequential processing framework, followed by exploring techniques for simultaneous phase aberration correction and image segmentation. While aberration correction enhances image quality, we anticipate that any errors in correction will propagate to the segmentation task. Instead of employing a pipeline that treats these tasks sequentially, a MTL framework can be used to jointly optimize both processes. By sharing representations between the two tasks, the network enables mutual learning, where phase aberration correction is guided by segmentation information and vice versa. This allows the aberration correction task to directly benefit from the segmentation masks, which capture high-level structural information about the target, something that is impossible in a sequential pipeline. The segmentation process, in turn, operates on corrected features, resulting in a synergistic improvement of both tasks. This joint optimization can minimize error propagation and enhance the overall accuracy and robustness of both phase correction and segmentation.

Bibliography

[1] G. Ter Haar, "Therapeutic applications of ultrasound," *Progress in biophysics and molecular biology*, vol. 93, no. 1-3, pp. 111–129, 2007.

[2] R. J. Paproski, A. Forbrich, M. Hitt, and R. Zemp, "Rna biomarker release with ultrasound and phase-change nanodroplets," *Ultrasound in Medicine & Biology*, vol. 40, no. 8, pp. 1847–1856, 2014.

[3] D. Karimi, Q. Zeng, P. Mathur, A. Avinash, S. Mahdavi, I. Spadinger, P. Abolmaesumi, and S. E. Salcudean, "Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images," *Medical image analysis*, vol. 57, pp. 186–196, 2019.

[4] S. Azizi, P. Yan, A. Tahmasebi, P. Pinto, B. Wood, J. Tae Kwak, S. Xu, B. Turkbey, P. Choyke, P. Mousavi, *et al.*, "Learning from noisy label statistics: detecting high grade prostate cancer in ultrasound guided biopsy," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, pp. 21–29, Springer, 2018.

[5] I. Hacihaliloglu, E. C. Chen, P. Mousavi, P. Abolmaesumi, E. Boctor, and C. A. Linte, "Interventional imaging: Ultrasound," in *Handbook of Medical Image Computing and Computer Assisted Intervention*, pp. 701–720, Elsevier, 2020.

[6] R. J. Zemp, L. Song, R. Bitton, K. K. Shung, and L. V. Wang, "Realtime photoacoustic microscopy in vivo with a 30-mhz ultrasound array transducer," *Optics express*, vol. 16, no. 11, pp. 7915–7928, 2008.

[7] M. Halliwell, "A tutorial on ultrasonic physics and imaging techniques," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 224, no. 2, pp. 127–142, 2010.

[8] R. W. Prager, U. Z. Ijaz, A. Gee, and G. M. Treece, "Three-dimensional ultrasound imaging," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 224, no. 2, pp. 193–223, 2010.

[9] G. F. Pinton, G. E. Trahey, and J. J. Dahl, "Sources of image degradation in fundamental and harmonic ultrasound imaging using nonlinear, full-wave simulations,"

IEEE transactions on ultrasonics, ferroelectrics, and frequency control, vol. 58, no. 4, pp. 754–765, 2011.

[10] S. Flax and M. O'Donnell, "Phase-aberration correction using signals from point reflectors and diffuse scatterers: Basic principles," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 35, no. 6, pp. 758–767, 1988.

[11] S. Goss, R. Johnston, and F. Dunn, "Compilation of empirical ultrasonic properties of mammalian tissues. ii.," *The Journal of the Acoustical Society of America*, vol. 68, no. 1, pp. 93–108, 1980.

[12] M. A. OŔeilly and K. Hynynen, "A super-resolution ultrasound method for brain vascular mapping," *Medical physics*, vol. 40, no. 11, p. 110701, 2013.

[13] S. A. Goss, R. L. Johnston, and F. Dunn, "Comprehensive compilation of empirical ultrasonic properties of mammalian tissues.," *The Journal of the Acoustical Society of America*, vol. 64, no. 2, pp. 423–457, 1978.

[14] M. Karaman, A. Atalar, H. Koymen, and M. O'Donnell, "A phase aberration correction method for ultrasound imaging," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 40, no. 4, pp. 275–282, 1993.

[15] L. Nock, G. E. Trahey, and S. W. Smith, "Phase aberration correction in medical ultrasound using speckle brightness as a quality factor," *The Journal of the Acoustical Society of America*, vol. 85, no. 5, pp. 1819–1833, 1989.

[16] B.-F. Osmanski, G. Montaldo, M. Tanter, and M. Fink, "Aberration correction by time reversal of moving speckle noise," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 59, no. 7, pp. 1575–1583, 2012.

[17] P.-C. Li and M.-L. Li, "Adaptive imaging using the generalized coherence factor," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 50, no. 2, pp. 128–141, 2003.

[18] D. Napolitano, C.-H. Chou, G. McLaughlin, T.-L. Ji, L. Mo, D. DeBusschere, and R. Steins, "Sound speed correction in ultrasound imaging," *Ultrasonics*, vol. 44, pp. e43–e46, 2006.

[19] J. Shin and L. Huang, "Spatial prediction filtering of acoustic clutter and random noise in medical ultrasound imaging," *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 396–406, 2016.

[20] J. Shin, L. Huang, and J. T. Yen, "Spatial prediction filtering for medical ultrasound in aberration and random noise," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 65, no. 10, pp. 1845–1856, 2018.

[21] G. Chau, M. Jakovljevic, R. Lavarello, and J. Dahl, "A locally adaptive phase aberration correction (lapac) method for synthetic aperture sequences," *Ultrasonic imaging*, vol. 41, no. 1, pp. 3–16, 2019.

[22] W. Lambert, L. A. Cobus, T. Frappart, M. Fink, and A. Aubry, "Distortion matrix approach for ultrasound imaging of random scattering media," *Proceedings of the National Academy of Sciences*, vol. 117, no. 26, pp. 14645–14656, 2020.

[23] W. Lambert, J. Robin, L. A. Cobus, M. Fink, and A. Aubry, "Ultrasound matrix imaging—part i: The focused reflection matrix, the f-factor and the role of multiple scattering," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3907–3920, 2022.

[24] W. Lambert, L. A. Cobus, J. Robin, M. Fink, and A. Aubry, "Ultrasound matrix imaging—part ii: The distortion matrix for aberration correction over multiple isoplanatic patches," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3921–3938, 2022.

[25] S. J. Sanabria, E. Ozkan, M. Rominger, and O. Goksel, "Spatial domain reconstruction for imaging speed-of-sound with pulse-echo ultrasound: simulation and in vivo study," *Physics in Medicine & Biology*, vol. 63, no. 21, p. 215015, 2018.

[26] R. Rau, D. Schweizer, V. Vishnevskiy, and O. Goksel, "Ultrasound aberration correction based on local speed-of-sound map estimation," in *2019 IEEE International Ultrasonics Symposium (IUS)*, pp. 2003–2006, IEEE, 2019.

[27] M. Jaeger, E. Robinson, H. G. Akarçay, and M. Frenz, "Full correction for spatially distributed speed-of-sound in echo ultrasound based on measuring aberration delays via transmit beam steering," *Physics in medicine & biology*, vol. 60, no. 11, p. 4497, 2015.

[28] R. Ali and J. J. Dahl, "Distributed Phase Aberration Correction Techniques Based on Local Sound Speed Estimates," in *IEEE International Ultrasonics Symposium, IUS*, vol. 2018-October, pp. 1–4, IEEE, oct 2018.

[29] J. A. Sethian and A. M. Popovici, "3-d traveltime computation using the fast marching method," *Geophysics*, vol. 64, no. 2, pp. 516–523, 1999.

[30] P. Stähli, M. Kuriakose, M. Frenz, and M. Jaeger, "Improved forward model for quantitative pulse-echo speed-of-sound imaging," *Ultrasonics*, vol. 108, p. 106168, 2020.

[31] M. Jakovljevic, S. Hsieh, R. Ali, G. Chau Loo Kung, D. Hyun, and J. J. Dahl, "Local speed of sound estimation in tissue using pulse-echo ultrasound: Model-based approach," *The Journal of the Acoustical Society of America*, vol. 144, no. 1, pp. 254–266, 2018.

[32] M. E. Anderson and G. E. Trahey, "The direct estimation of sound speed using pulse–echo ultrasound," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3099–3106, 1998.

[33] R. Ali and J. J. Dahl, "Travel-time tomography for local sound speed reconstruction using average sound speeds," in *2019 IEEE International Ultrasonics Symposium (IUS)*, pp. 2007–2010, IEEE, 2019.

[34] T. Brevett, S. J. Sanabria, R. Ali, and J. Dahl, "Speed of sound estimation at multiple angles from common midpoint gathers of non-beamformed data," in *2022 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2022.

[35] R. Ali, T. Brevett, D. Hyun, L. L. Brickson, and J. J. Dahl, "Distributed aberration correction techniques based on tomographic sound speed estimates," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 5, pp. 1714–1726, 2022.

[36] M.-L. Li, "Adaptive imaging using principal-component-synthesized aperture data," in *2008 IEEE Ultrasonics Symposium*, pp. 1076–1079, IEEE, 2008.

[37] H. Bendjador, T. Deffieux, and M. Tanter, "The svd beamformer: Physical principles and application to ultrafast adaptive ultrasound," *IEEE transactions on medical imaging*, vol. 39, no. 10, pp. 3100–3112, 2020.

[38] B. E. Treeby and B. T. Cox, "k-wave: Matlab toolbox for the simulation and reconstruction of photoacoustic wave fields," *Journal of biomedical optics*, vol. 15, no. 2, pp. 021314–021314, 2010.

[39] M. Feigin, D. Freedman, and B. W. Anthony, "A deep learning framework for single-sided sound speed inversion in medical ultrasound," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 4, pp. 1142–1151, 2019.

[40] F. Khun Jush, P. M. Dueppenbecker, and A. Maier, "Data-driven speed-of-sound reconstruction for medical ultrasound: impacts of training data format and imperfections on convergence," in *Medical Image Understanding and Analysis: 25th Annual Conference, MIUA 2021, Oxford, United Kingdom, July 12–14, 2021, Proceedings 25*, pp. 140–150, Springer, 2021.

[41] F. K. Jush, M. Biele, P. M. Dueppenbecker, O. Schmidt, and A. Maier, "Dnn-based speed-of-sound reconstruction for automated breast ultrasound," in *2020 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–7, IEEE, 2020.

[42] J. R. Young, S. Schoen, V. Kumar, K. Thomenius, and A. E. Samir, "Soundai: improved imaging with learned sound speed maps," in *2022 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2022.

[43] T. Koike, N. Tomii, Y. Watanabe, T. Azuma, and S. Takagi, "Deep learning for hetero–homo conversion in channel-domain for phase aberration correction in ultrasound imaging," *Ultrasonics*, vol. 129, p. 106890, 2023.

[44] W.-H. Shen and M.-L. Li, "A novel adaptive imaging technique using point spread function reshaping," in *2022 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–3, IEEE, 2022.

[45] A. C. Luchies and B. C. Byram, "Assessing the robustness of frequency-domain ultrasound beamforming using deep neural networks," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 67, no. 11, pp. 2321–2335, 2020.

[46] S. Khan, J. Huh, and J. C. Ye, "Phase aberration robust beamformer for planewave us using self-supervised learning," *arXiv preprint arXiv:2202.08262*, 2022.

[47] Y. Hu, H. U. Ahmed, Z. Taylor, C. Allen, M. Emberton, D. Hawkes, and D. Barratt, "Mr to ultrasound registration for image-guided prostate interventions," *Medical image analysis*, vol. 16, no. 3, pp. 687–703, 2012.

[48] H. Hricak, P. L. Choyke, S. C. Eberhardt, S. A. Leibel, and P. T. Scardino, "Imaging prostate cancer: a multidisciplinary perspective," *Radiology*, vol. 243, no. 1, pp. 28–53, 2007.

[49] A. Katouzian, E. D. Angelini, S. G. Carlier, J. S. Suri, N. Navab, and A. F. Laine, "A state-of-the-art review on segmentation algorithms in intravascular ultrasound (ivus) images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 5, pp. 823–834, 2012.

[50] J. H. Park and S. J. Lee, "Ultrasound deep learning for wall segmentation and near-wall blood flow measurement," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 10, pp. 2022–2032, 2020.

[51] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2d echocardiography," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[52] N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande, and P.-M. Jodoin, "Cardiac segmentation with strong anatomical guarantees," *IEEE transactions on medical imaging*, vol. 39, no. 11, pp. 3703–3713, 2020.

[53] M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, and R. Marti, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1218–1226, 2017.

[54] K. Drukker, M. L. Giger, K. Horsch, M. A. Kupinski, C. J. Vyborny, and E. B. Mendelson, "Computerized lesion detection on breast ultrasound," *Medical physics*, vol. 29, no. 7, pp. 1438–1446, 2002.

[55] M. H. Yap, E. A. Edirisinghe, and H. E. Bez, "A novel algorithm for initial lesion detection in ultrasound breast images," *Journal of Applied Clinical Medical Physics*, vol. 9, no. 4, pp. 181–199, 2008.

[56] J. Shan, H. Cheng, and Y. Wang, "Completely automated segmentation approach for breast ultrasound images using multiple-domain features," *Ultrasound in medicine & biology*, vol. 38, no. 2, pp. 262–275, 2012.

[57] G. Pons, R. Martí, S. Ganau, M. Sentís, and J. Martí, "Feasibility study of lesion detection using deformable part models in breast ultrasound images," in *Pattern Recognition and Image Analysis: 6th Iberian Conference, IbPRIA 2013, Funchal, Madeira, Portugal, June 5-7, 2013. Proceedings 6*, pp. 269–276, Springer, 2013.

[58] E. Smistad, I. M. Salte, A. Østvik, S. Leclerc, O. Bernard, and L. Lovstakken, "Segmentation of apical long axis, four-and two-chamber views using deep neural networks," in *2019 IEEE International Ultrasonics Symposium (IUS)*, pp. 8–11, IEEE, 2019.

[59] S. Leclerc, E. Smistad, T. Grenier, C. Lartizien, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, *et al.*, "Ru-net: A refining segmentation network for 2d echocardiography," in *2019 IEEE International Ultrasonics Symposium (IUS)*, pp. 1160–1163, IEEE, 2019.

[60] S. Leclerc, E. Smistad, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, M. Belhamissi, S. Israilov, T. Grenier, *et al.*, "Lu-net: a multistage attention network to improve the robustness of segmentation of left ventricular structures in 2-d echocardiography," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2519–2530, 2020.

[61] S. Leclerc, E. Smistad, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, *et al.*, "Deep learning segmentation in 2d echocardiography using the camus dataset: Automatic assessment of the anatomical shape validity," *arXiv preprint arXiv:1908.02994*, 2019.

[62] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pp. 240–248, Springer, 2017.

[63] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pp. 683–687, IEEE, 2019.

[64] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Transactions on medical imaging*, vol. 39, no. 2, pp. 499–513, 2019.

[65] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 2, pp. 699–711, 2020.

[66] R. Van Sloun, R. Wildeboer, A. Postema, C. Mannaerts, M. Gayer, H. Wijkstra, and M. Mischi, "Zonal segmentation in transrectal ultrasound images of the prostate through deep learning," in *2018 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2018.

[67] R. J. van Sloun, R. R. Wildeboer, C. K. Mannaerts, A. W. Postema, M. Gayet, H. P. Beerlage, G. Salomon, H. Wijkstra, and M. Mischi, "Deep learning for real-time, automatic, and scanner-adapted prostate (zone) segmentation of transrectal ultrasound, for example, magnetic resonance imaging–transrectal ultrasound fusion prostate biopsy," *European urology focus*, vol. 7, no. 1, pp. 78–85, 2021.

[68] R. J. van Sloun, R. R. Wildeboer, C. K. Mannaerts, A. W. Postema, M. Gayet, H. P. Beerlage, G. Salomon, H. Wijkstra, and M. Mischi, "Deep learning for real-time, automatic, and scanner-adapted prostate (zone) segmentation of transrectal ultrasound, for example, magnetic resonance imaging–transrectal ultrasound fusion prostate biopsy," *European urology focus*, vol. 7, no. 1, pp. 78–85, 2021.

[69] Y. Wang, H. Dou, X. Hu, L. Zhu, X. Yang, M. Xu, J. Qin, P.-A. Heng, T. Wang, and D. Ni, "Deep attentive features for prostate segmentation in 3d transrectal ultrasound," *IEEE transactions on medical imaging*, vol. 38, no. 12, pp. 2768–2778, 2019.

[70] Y.-C. Li, T.-Y. Shen, C.-C. Chen, W.-T. Chang, P.-Y. Lee, and C.-C. J. Huang, "Automatic detection of atherosclerotic plaque and calcification from intravascular ultrasound images by using deep convolutional neural networks," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 68, no. 5, pp. 1762–1772, 2021.

[71] P. Looney, Y. Yin, S. L. Collins, K. H. Nicolaides, W. Plasencia, M. Molloholli, S. Natsis, and G. N. Stevenson, "Fully automated 3-d ultrasound segmentation of the placenta, amniotic fluid, and fetus for early pregnancy assessment," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 68, no. 6, pp. 2038–2047, 2021.

[72] V. A. Zimmer, A. Gomez, E. Skelton, N. Toussaint, T. Zhang, B. Khanal, R. Wright, Y. Noh, A. Ho, J. Matthew, *et al.*, "Towards whole placenta segmentation at late gestation using multi-view ultrasound images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*, pp. 628–636, Springer, 2019.

[73] V. A. Zimmer, A. Gomez, E. Skelton, N. Ghavami, R. Wright, L. Li, J. Matthew, J. V. Hajnal, and J. A. Schnabel, "A multi-task approach using positional information for ultrasound placenta segmentation," in *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4-8, 2020, Proceedings 1*, pp. 264–273, Springer, 2020.

[74] G.-Q. Zhou, E.-Z. Huo, M. Yuan, P. Zhou, R.-L. Wang, K.-N. Wang, Y. Chen, and X.-P. He, "A single-shot region-adaptive network for myotendinous junction segmentation in muscular ultrasound images," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 67, no. 12, pp. 2531–2542, 2020.

[75] R. Zhou, F. Guo, M. R. Azarpazhooh, J. D. Spence, E. Ukwatta, M. Ding, and A. Fenster, "A voxel-based fully convolution network and continuous max-flow for carotid vessel-wall-volume segmentation from 3d ultrasound images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2844–2855, 2020.

[76] M. Amiri, R. Brooks, B. Behboodi, and H. Rivaz, "Two-stage ultrasound image segmentation using u-net and test time augmentation," *International journal of computer assisted radiology and surgery*, vol. 15, pp. 981–988, 2020.

[77] M. Amiri, R. Brooks, and H. Rivaz, "Fine-tuning u-net for ultrasound image segmentation: different layers, different outcomes," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2510–2518, 2020.

[78] R. Zhang, "Making convolutional networks shift-invariant again," in *International conference on machine learning*, pp. 7324–7334, PMLR, 2019.

[79] J. A. Jensen, "Field: A program for simulating ultrasound systems," *Medical & Biological Engineering & Computing*, vol. 34, no. sup. 1, pp. 351–353, 1997.

[80] J. A. Jensen and N. B. Svendsen, "Calculation of pressure fields from arbitrarily shaped, apodized, and excited ultrasound transducers," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 39, no. 2, pp. 262–267, 1992.

[81] M. Sharifzadeh, H. Benali, and H. Rivaz, "Phase aberration correction: A convolutional neural network approach," *IEEE Access*, vol. 8, pp. 162252–162260, 2020.

[82] M. O'donnell and S. Flax, "Phase-aberration correction using signals from point reflectors and diffuse scatterers: Measurements," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 35, no. 6, pp. 768–774, 1988.

[83] M. O'Donnell and W. E. Engeler, "Correlation-based aberration correction in the presence of inoperable elements," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 39, no. 6, pp. 700–707, 1992.

[84] D. Monjazebi and Y. Xu, "Phase-aberration delay estimation in synthetic transmit aperture diagnostic ultrasound," in *2019 IEEE International Ultrasonics Symposium (IUS)*, pp. 2011–2014, IEEE, 2019.

[85] J. J. Dahl, D. A. Guenther, and G. E. Trahey, "Adaptive imaging and spatial compounding in the presence of aberration," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 52, no. 7, pp. 1131–1144, 2005.

[86] L. M. Hinkelman, D.-L. Liu, R. C. Waag, Q. Zhu, and B. D. Steinberg, "Measurement and correction of ultrasonic pulse distortion produced by the human breast," *The Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1958–1969, 1995.

[87] A. Fernandez, J. J. Dahl, D. M. Dumont, and G. E. Trahey, "Aberration measurement and correction with a high resolution 1.75 d array," in *2001 IEEE Ultrasonics Symposium. Proceedings. An International Symposium (Cat. No. 01CH37263)*, vol. 2, pp. 1489–1494, IEEE, 2001.

[88] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[89] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[90] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[91] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697–8710, 2018.

[92] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[93] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[94] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," *arXiv preprint arXiv:1904.09237*, 2019.

[95] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[96] R. Girshick, "Fast r-cnn," *arXiv preprint arXiv:1504.08083*, 2015.

[97] M. Rashid, X. Gu, and Y. Jae Lee, "Interspecies knowledge transfer for facial keypoint detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6894–6903, 2017.

[98] M. Lin, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[99] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *International conference on artificial neural networks*, pp. 92–101, Springer, 2010.

[100] J. Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine learning*, vol. 28, pp. 7–39, 1997.

[101] S. Ruder, "An overview of multi-task learning for deep learning," *Sebastian Rude r*, 2017.

[102] M. Sharifzadeh, S. Goudarzi, A. Tang, H. Benali, and H. Rivaz, "Mitigating aberration-induced noise: A deep learning-based aberration-to-aberration approach," *IEEE Transactions on Medical Imaging*, 2024.

[103] J. Lehtinen, "Noise2noise: Learning image restoration without clean data," *arXiv preprint arXiv:1803.04189*, 2018.

[104] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4142–4150, 2017.

[105] D. Hyun, L. L. Brickson, K. T. Looby, and J. J. Dahl, "Beamforming and speckle reduction using neural networks," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 66, no. 5, pp. 898–910, 2019.

[106] A. Rodriguez-Molares, H. Torp, B. Denarie, and L. Løvstakken, "The angular apodization in coherent plane-wave compounding [correspondence]," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 62, no. 11, pp. 2018–2023, 2015.

[107] H. Liebgott, A. Rodriguez-Molares, F. Cervenansky, J. A. Jensen, and O. Bernard, "Plane-wave imaging challenge in medical ultrasound," in *2016 IEEE International ultrasonics symposium (IUS)*, pp. 1–4, IEEE, 2016.

[108] H. Su, F. Xing, X. Kong, Y. Xie, S. Zhang, and L. Yang, "Robust cell detection and segmentation in histopathological images using sparse reconstruction and stacked denoising autoencoders," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 383–390, Springer, 2015.

[109] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

[110] S.-E. Måsøy, B. Dénarié, A. Sørnes, E. Holte, B. Grenne, T. Espeland, E. A. R. Berg, O. M. H. Rindal, W. Rigby, and T. Bjåstad, "Aberration correction in 2d echocardiography," *Quantitative Imaging in Medicine and Surgery*, vol. 13, no. 7, p. 4603, 2023.

[111] A. Rodriguez-Molares, O. M. H. Rindal, J. D'hooge, S.-E. Måsøy, A. Austeng, M. A. L. Bell, and H. Torp, "The generalized contrast-to-noise ratio: A formal definition for lesion detectability," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 4, pp. 745–759, 2019.

[112] R. Göbl, C. Hennersperger, and N. Navab, "Speckle2speckle: Unsupervised learning of ultrasound speckle filtering without clean data," *arXiv preprint arXiv:2208.00402*, 2022.

[113] M. Fink, R. Mallart, and F. Cancre, "The random phase transducer: A new technique for incoherent processing-basic principles and theory," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 37, no. 2, pp. 54–69, 1990.

[114] A. K. Tehrani, M. Sharifzadeh, E. Boctor, and H. Rivaz, "Bi-directional semi-supervised training of convolutional neural networks for ultrasound elastography displacement estimation," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 4, pp. 1181–1190, 2022.

[115] M. Ashikuzzaman, A. Sadeghi-Naini, A. Samani, and H. Rivaz, "Combining first- and second-order continuity constraints in ultrasound elastography," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 7, pp. 2407–2418, 2021.

[116] M. Mirzaei, A. Asif, M. Fortin, and H. Rivaz, "3d normalized cross-correlation for estimation of the displacement field in ultrasound elastography," *Ultrasonics*, vol. 102, p. 106053, 2020.

[117] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 2, pp. 60–65, Ieee, 2005.

[118] B. Jing and B. D. Lindsey, "Phase modulation beamforming for ultrafast plane-wave imaging," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 67, no. 10, pp. 2003–2011, 2020.

[119] G. Montaldo, M. Tanter, J. Bercoff, N. Benech, and M. Fink, "Coherent plane-wave compounding for very high frame rate ultrasonography and transient elastography," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 56, no. 3, pp. 489–506, 2009.

[120] M. Sharifzadeh, H. Benali, and H. Rivaz, "Shift-invariant segmentation in breast ultrasound images," in *2021 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2021.

[121] M. Sharifzadeh, H. Benali, and H. Rivaz, "Investigating shift variance of convolutional neural networks in ultrasound image segmentation," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 5, pp. 1703–1713, 2022.

[122] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *International conference on machine learning*, pp. 1802–1811, PMLR, 2019.

[123] O. S. Kayhan and J. C. v. Gemert, "On translation invariance in cnns: Convolutional layers can exploit absolute spatial location," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14274–14285, 2020.

[124] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?," *Journal of Machine Learning Research*, vol. 20, no. 184, pp. 1–25, 2019.

[125] J. Lee, J. Yang, and Z. Wang, "What does cnn shift invariance look like? a visualization study," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 196–210, Springer, 2020.

[126] A. Chaman and I. Dokmanic, "Truly shift-invariant convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3773–3783, 2021.

[127] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833, Springer, 2014.

[128] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[129] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

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

[131] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[132] K. Lenc and A. Vedaldi, "Understanding image representations by measuring their equivariance and equivalence," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

[133] L. Vinet and A. Zhedanov, "A 'missing'family of classical orthogonal polynomials," *Journal of Physics A: Mathematical and Theoretical*, vol. 44, no. 8, p. 085201, 2011.

[134] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[135] P. J. Burt, "Fast filter transform for image processing," *Computer graphics and image processing*, vol. 16, no. 1, pp. 20–51, 1981.

[136] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," in *Readings in computer vision*, pp. 671–679, Elsevier, 1987.

[137] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[138] M. Buda, A. Saha, and M. A. Mazurowski, "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm," *Computers in biology and medicine*, vol. 109, pp. 218–225, 2019.

[139] M. A. Mazurowski, K. Clark, N. M. Czarnek, P. Shamsesfandabadi, K. B. Peters, and A. Saha, "Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the cancer genome atlas data," *Journal of neuro-oncology*, vol. 133, pp. 27–35, 2017.

[140] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*, pp. 9–50, Springer, 2002.

[141] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[142] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[143] M. Sharifzadeh, H. Benali, and H. Rivaz, "An ultra-fast method for simulation of realistic ultrasound images," in *2021 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2021.

[144] R. J. McGough, "Focus: Fast object-oriented c++ ultrasound simulator," *see http://www. egr. msu. edu/fultras-web*, 2015.

[145] M. D. Verweij, B. E. Treeby, K. W. van Dongen, and L. Demi, "Simulation of ultrasound fields," in *Comprehensive biomedical physics*, pp. 465–499, Elsevier, 2014.

[146] B. E. Treeby, J. Jaros, A. P. Rendell, and B. T. Cox, "Modeling nonlinear ultrasound propagation in heterogeneous media with power law absorption using a k-space pseudospectral method," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4324–4336, 2012.

[147] B. Burger, S. Bettinghausen, M. Radle, and J. Hesser, "Real-time gpu-based ultrasound simulation using deformable mesh models," *IEEE transactions on medical imaging*, vol. 32, no. 3, pp. 609–618, 2012.

[148] O. Mattausch, M. Makhinya, and O. Goksel, "Realistic ultrasound simulation of complex surface models using interactive monte-carlo path tracing," in *Computer Graphics Forum*, vol. 37, pp. 202–213, Wiley Online Library, 2018.

[149] L. Zhang, V. Vishnevskiy, and O. Goksel, "Deep network for scatterer distribution estimation for ultrasound image simulation," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2553–2564, 2020.

[150] Y. Hu, E. Gibson, L.-L. Lee, W. Xie, D. C. Barratt, T. Vercauteren, and J. A. Noble, "Freehand ultrasound image simulation with spatially-conditioned generative adversarial networks," in *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment: Fifth International Workshop, CMMI 2017, Second International Workshop, RAMBO 2017, and First International Workshop, SWITCH 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings 5*, pp. 105–115, Springer, 2017.

[151] N. J. Cronin, T. Finni, and O. Seynnes, "Using deep learning to generate synthetic b-mode musculoskeletal ultrasound images," *Computer methods and programs in biomedicine*, vol. 196, p. 105583, 2020.

[152] J. Liang *et al.*, "Synthesis and edition of ultrasound images via sketch guided progressive growing gans". ieee 17th international symposium on biomedical imaging (isbi)," 2020.

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

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

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

[156] M. Frigo and S. G. Johnson, "Fftw: An adaptive software architecture for the fft," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 3, pp. 1381–1384, IEEE, 1998.

[157] M. Sharifzadeh, A. K. Tehrani, H. Benali, and H. Rivaz, "Ultrasound domain adaptation using frequency domain analysis," in *2021 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2021.

[158] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.-E. de Leeuw, C. M. Tempany, B. Van Ginneken, *et al.*, "Transfer learning for domain adaptation in mri: Application in brain lesion segmentation," in *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pp. 516–524, Springer, 2017.

[159] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3339–3348, 2018.

[160] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *2011 international conference on computer vision*, pp. 999–1006, IEEE, 2011.

[161] J. Tierney, A. Luchies, C. Khan, B. Byram, and M. Berger, "Accounting for domain shift in neural network ultrasound beamforming," in *2020 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–3, IEEE, 2020.

[162] J. Tierney, A. Luchies, C. Khan, J. Baker, D. Brown, B. Byram, and M. Berger, "Training deep network ultrasound beamformers with unlabeled in vivo data," *IEEE transactions on medical imaging*, vol. 41, no. 1, pp. 158–171, 2021.

[163] X. Ying, Y. Zhang, X. Wei, M. Yu, J. Zhu, J. Gao, Z. Liu, X. Li, and R. Yu, "Ms-dan: multi-scale self-attention unsupervised domain adaptation network for thyroid ultrasound images," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 871–876, IEEE, 2020.

[164] Q. Meng, J. Matthew, V. A. Zimmer, A. Gomez, D. F. Lloyd, D. Rueckert, and B. Kainz, "Mutual information-based disentangled neural networks for classifying unseen categories in different domains: Application to fetal ultrasound imaging," *IEEE transactions on medical imaging*, vol. 40, no. 2, pp. 722–734, 2020.

[165] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu, *et al.*, "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE transactions on medical imaging*, vol. 39, no. 7, pp. 2531–2540, 2020.

[166] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4085–4095, 2020.

[167] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 102–118, Springer, 2016.

[168] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[169] W. Legon, T. F. Sato, A. Opitz, J. Mueller, A. Barbour, A. Williams, and W. J. Tyler, "Transcranial focused ultrasound modulates the activity of primary somatosensory cortex in humans," *Nature neuroscience*, vol. 17, no. 2, pp. 322–329, 2014.

[170] K. Hynynen and J. Sun, "Trans-skull ultrasound therapy: The feasibility of using image-derived skull thickness information to correct the phase distortion," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 46, no. 3, pp. 752–755, 1999.

[171] R. M. Jones and K. Hynynen, "Comparison of analytical and numerical approaches for ct-based aberration correction in transcranial passive acoustic imaging," *Physics in Medicine & Biology*, vol. 61, no. 1, p. 23, 2015.

[172] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

[173] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[174] Z. Zhang, L. Yao, B. Wang, D. Jha, E. Keles, A. Medetalibeyoglu, and U. Bagci, "Emit-diff: Enhancing medical image segmentation via text-guided diffusion model," *arXiv preprint arXiv:2310.12868*, 2023.

[175] M. Sharifzadeh, H. Benali, and H. Rivaz, "Frequency-space prediction filtering for phase aberration correction in plane-wave ultrasound," in *2023 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2023.

[176] L. L. Canales, "Random noise reduction," in *SEG Technical Program Expanded Abstracts 1984*, pp. 525–527, Society of Exploration Geophysicists, 1984.

[177] M. Sharifzadeh, H. Benali, and H. Rivaz, "Robust rf data normalization for deep learning," in *2023 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2023.

[178] J. Lu, F. Millioz, D. Garcia, S. Salles, D. Ye, and D. Friboulet, "Complex convolutional neural networks for ultrafast ultrasound imaging reconstruction from in-phase/quadrature signal," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 69, no. 2, pp. 592–603, 2021.

[179] S. Goudarzi and H. Rivaz, "Deep reconstruction of high-quality ultrasound images from raw plane-wave data: A simulation and in vivo study," *Ultrasonics*, vol. 125, p. 106778, 2022.

[180] Z. Lei, S. Gao, H. Hasegawa, Z. Zhang, M. Zhou, and K. Sedraoui, "Fully complex-valued gated recurrent neural network for ultrasound imaging," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[181] H. Asgariandehkordi, S. Goudarzi, A. Basarab, and H. Rivaz, "Deep ultrasound denoising using diffusion probabilistic models," in *2023 IEEE International Ultrasonics Symposium (IUS)*, pp. 1–4, IEEE, 2023.

[182] S. Salari, A. Rasoulian, H. Rivaz, and Y. Xiao, "Focalerrornet: Uncertainty-aware focal modulation network for inter-modal registration error estimation in ultrasound-guided neurosurgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 689–698, Springer, 2023.

[183] A. K. Tehrani, M. Ashikuzzaman, and H. Rivaz, "Lateral strain imaging using self-supervised and physically inspired constraints in unsupervised regularized elastography," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1462–1471, 2022.

[184] A. Tehrani, I. Rosado-Mendez, and H. Rivaz, "Deep estimation of viscoelastic and backscatter quantitative ultrasound," *The Journal of the Acoustical Society of America*, vol. 152, no. 4_Supplement, pp. A74–A74, 2022.

[185] U. Soylu and M. L. Oelze, "Calibrating data mismatches in deep learning-based quantitative ultrasound using setting transfer functions," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 70, no. 6, pp. 510–520, 2023.

[186] U. Soylu and M. L. Oelze, "A data-efficient deep learning strategy for tissue characterization via quantitative ultrasound: Zone training," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 70, no. 5, pp. 368–377, 2023.

[187] M. Sharifzadeh, H. Benali, and H. Rivaz, "Segmentation of intraoperative 3d ultrasound images using a pyramidal blur-pooled 2d u-net," in *MICCAI Challenge on Correction of Brainshift with Intra-Operative Ultrasound*, pp. 69–75, Springer, 2022.

[188] L. Dixon, A. Lim, M. Grech-Sollars, D. Nandi, and S. Camp, "Intraoperative ultrasound in brain tumor surgery: A review and implementation guide," *Neurosurgical Review*, vol. 45, no. 4, pp. 2503–2515, 2022.

[189] H. Rivaz, S. J.-S. Chen, and D. L. Collins, "Automatic deformable mr-ultrasound registration for image-guided neurosurgery," *IEEE transactions on medical imaging*, vol. 34, no. 2, pp. 366–380, 2014.

[190] Y. Xiao, M. Fortin, G. Unsgård, H. Rivaz, and I. Reinertsen, "Re trospective evaluation of cerebral tumors (resect): A clinical database of pre-operative mri and intra-operative ultrasound in low-grade glioma surgeries," *Medical physics*, vol. 44, no. 7, pp. 3875–3882, 2017.

[191] B. Behboodi, F.-X. Carton, M. Chabanas, S. De Ribaupierre, O. Solheim, B. K. Munkvold, H. Rivaz, Y. Xiao, and I. Reinertsen, "Resect-seg: Open access annotations of intra-operative brain tumor ultrasound images," *arXiv preprint arXiv:2207.07494*, 2022.

[192] F.-X. Carton, M. Chabanas, F. Le Lann, and J. H. Noble, "Automatic segmentation of brain tumor resections in intraoperative ultrasound images using u-net," *Journal of Medical Imaging*, vol. 7, no. 3, pp. 031503–031503, 2020.

[193] F.-X. Carton, J. H. Noble, and M. Chabanas, "Automatic segmentation of brain tumor resections in intraoperative ultrasound images," in *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 10951, pp. 211–217, SPIE, 2019.

Appendix A

Frequency-Space Prediction Filtering for Phase Aberration Correction in Plane-Wave Ultrasound

This chapter is based on our published paper [175].

In Chapter 3, we compared the proposed method with the frequency-space (F-X) prediction filtering technique or FXPF. Initially developed for random noise suppression in seismic imaging [176], this method has recently been applied to correct phase aberrations in focused ultrasound imaging [20]. The FXPF method assumes an AR model of order p across the signals received by the transducer elements, systematically eliminating any components that deviate from the established model.

This appendix highlights the challenges of applying this technique to plane-wave imaging. At shallower depths, signals from more distant elements become less relevant, resulting in fewer elements contributing to image reconstruction. Since the number of contributing signals varies with depth, utilizing a fixed-order AR model across all depths leads to suboptimal performance. To address this issue, we propose an AR model with an adaptive order and quantify its effectiveness using contrast and gCNR metrics.

A.1 Methodology

A.1.1 Adaptive FXPF

Let us consider a transducer with N elements and denote the Fourier transform of the received RF signal at time t by element $n \in [1, N]$ located at x_n as $RF_n(f) = \mathcal{F}\{RF(x_n, t)\}$.

The FXPF establishes an AR model of order p across the RF channel signals received at transducer elements. Specifically, in the frequency domain and for each temporal frequency f_k, the method predicts a signal as a linear combination of the signals received by the p preceding channels:

$$\begin{aligned} RF_{n+1}(f_k) = a_1(f_k)RF_n(f_k) + a_2(f_k)RF_{n-1}(f_k) + \\ a_3(f_k)RF_{n-2}(f_k) + ... + a_p(f_k)RF_{n+1-p}(f_k), \end{aligned} \tag{A.1}$$

where coefficients denoted by a need to be estimated. Given that Eq. (A.1) represents a convolution, it can be expressed as

$$RF_{n+1} = \begin{bmatrix} RF_n & RF_{n-1} & \dots & RF_{n+1-p} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix}, \tag{A.2}$$

where the f_k was left out for simplicity of notation, but the equation pertains to a specific temporal frequency, denoted as f_k. Equation (A.2) can be written in a more general form as the product of a matrix and a vector. For example, when $p = 4$, it can be written as

$$\begin{bmatrix} RF_2 \\ RF_3 \\ RF_4 \\ RF_5 \\ \vdots \\ RF_{n+1} \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} RF_1 & 0 & 0 & 0 \\ RF_2 & RF_1 & 0 & 0 \\ RF_3 & RF_2 & RF_1 & 0 \\ RF_4 & RF_3 & RF_2 & RF_1 \\ \vdots & \vdots & \vdots & \vdots \\ RF_n & RF_{n-1} & RF_{n-2} & RF_{n-3} \\ 0 & RF_n & RF_{n-1} & RF_{n-2} \\ 0 & 0 & RF_n & RF_{n-1} \\ 0 & 0 & 0 & RF_n \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}. \tag{A.3}$$

Let us express Eq. (A.3) as

$$\mathbf{d} = \mathbf{M}\mathbf{a}, \tag{A.4}$$

where $\mathbf{d}$ represents the vector comprising values associated with the current elements, $\mathbf{M}$ denotes the convolution matrix consisting of values corresponding to the preceding elements, and $\mathbf{a}$ is the prediction error filter with a length of p. In practice, RF channel data are inevitably contaminated with random noise from various sources. Therefore, the prediction error filter $\mathbf{a}$ in Eq. (A.4) must be estimated from the noisy data $\mathbf{d}$. Achieving this

requires minimizing the energy associated with the prediction error:

$$\mathbf{L} = \|\mathbf{Ma} - \mathbf{d}\|_2^2, \tag{A.5}$$

where $\|.\|_2^2$ is the square of the Euclidean norm. To minimize the cost function $\mathbf{L}$, it is required to set $\frac{\partial \mathbf{L}}{\partial \mathbf{a}} = 0$, which results in

$$\mathbf{M^T d} = \mathbf{M^T Ma}. \tag{A.6}$$

We can obtain an estimate $\hat{\mathbf{a}}$ of the prediction error filter $\mathbf{a}$ as

$$\hat{\mathbf{a}} = (\mathbf{M^T M} + \mu \mathbf{I})^{-1} \mathbf{M^T d} \tag{A.7}$$

where a stability factor μ is added into the diagonal components of $\mathbf{M^T M}$ to enhance the stability of the matrix inversion. In this chapter, μ was set to 0.01, and the results exhibit minimal sensitivity to its value. After obtaining the estimated prediction error filter $\hat{\mathbf{a}}$, an estimate $\hat{\mathbf{d}}$ of the noise-free signal $\mathbf{d}$ can be acquired by applying it to the noisy data $\mathbf{M}$:

$$\hat{\mathbf{d}} = \mathbf{M}\hat{\mathbf{a}}, \tag{A.8}$$

where components of noisy data that do not conform to the established AR model are filtered out. Finally, the filtered RF signals can be obtained by applying the inverse Fourier transform.

While FXPF has been utilized for phase aberration correction in focused images [20], employing this method for plane-wave images poses a challenge. This challenge primarily arises from the substantial variation in channel data across elements at shallower depths, where signals from more distant elements become irrelevant and may negatively impact the performance of the AR model. Even after applying apodization, using a high-order AR model for shallow depths with only a few echo signals may lead to over-smoothing during prediction filtering. In such scenarios, adopting a fixed-order AR model across all depths would result in suboptimal performance. To address this issue, we propose the utilization of an AR model with an adaptive order, defined as follows:

$$p(z) = \min(p_{max}, \lceil p_{max} \times \left(\frac{z}{f \times L} \right)^{\beta} \rceil), \tag{A.9}$$

where f represents the f-number, z is the depth, p_{max} is the maximum order used at depths where all elements are utilized for reconstruction, $\lceil . \rceil$ denotes rounding up to the nearest

integer, and β is the non-linearity coefficient that controls the speed of transition from lower orders to higher orders. In summary, as per the formulation given by Eq. (A.9), the AR model featuring an adaptive order always commences with a lower order (e.g., $p = 1$) for shallower depths, progressively increasing the order until it reaches p_{max}, a value we established for the deepest depths.

Although the technique was presented based on a forward AR model, a backward AR model can also be established by reversing the sequence of transducer elements [20]. To minimize potential directional biases and enhance the performance of the technique, the data underwent two independent filtering processes using both forward and backward AR models. The final output was then determined by averaging the results of these dual filtering paths. In practice, we used a moving axial kernel to compute the fast Fourier transform. Rather than processing the entire image all at once, we progressively shifted the kernel along the axial direction until the full depth was covered. Furthermore, once the method has been applied to the image for an initial iteration, it can undergo subsequent iterations, as long as it continues to yield improved outcomes. In our experimental setup, we set the f-number to 1.75, employed an axial kernel size equivalent to one wavelength, and applied the FXPF method for 2 iterations. For the adaptive FXPF, we configured p_{max} to be 4, while β was set to $1/3$. These specific parameters were selected due to their production of the most optimal results in our cases.

A.1.2 Tissue-Mimicking Phantom Data

An L11-5v linear array transducer was operated using a Vantage 256 system (Verasonics, Kirkland, WA) to acquire a single plane-wave image from a multi-purpose multi-tissue ultrasound phantom (Model 040GSE, CIRS, Norfolk, VA). The center and sampling frequencies were set at 5.208 MHz and 20.832 MHz, respectively, with the sound speed assumed to be 1540 m/s. The transducer settings are the same as those summarized in Table 3.1.4. We introduced a quasi-physical aberration to the image by programming the scanner to excite transducer elements asynchronously according to a randomly generated aberration profile, as explained in Section 3.1.3. Moreover, delay errors introduced by the aberration profile were taken into account during the reception process for reconstructing the image. Received signals were stored as RF channel data after applying beamforming delays, serving as the input for the proposed method.

A.1.3 Quality Metrics

To quantitatively measure the quality of the reconstructed images, we calculated contrast and gCNR [111] metrics, as defined in Eqs. (3.11) and (3.13) for the top and bottom anechoic cysts using the target and background regions shown in Fig. A.1(a).

A.2 Results and Discussion

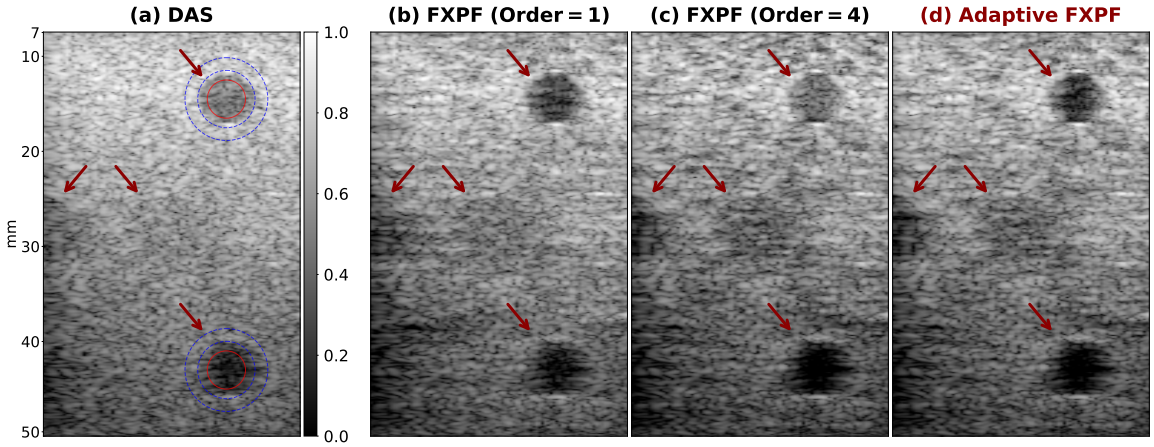


Figure A.1: Qualitative comparison between FXPF methods with fixed orders and an adaptive order. (a) An aberrated single plane-wave image reconstructed using DAS. (b) The FXPF output with a fixed order of 1. (c) The FXPF output with a fixed order of 4. (d) The FXPF output with an adaptive order.

Fig. A.1(a) shows an aberrated single plane-wave image reconstructed using the DAS method. To mitigate the phase aberration effect, we applied the FXPF method with three distinct configurations. These include two AR models with fixed orders of 1 and 4, as well as an additional AR model incorporating the proposed adaptive order. The outputs obtained using fixed orders of 1 and 4 are illustrated in Fig. A.2(b) and (c), respectively. While the model with a fixed order of 1 effectively enhanced the contrast of the anechoic cyst at shallow depths, it was nearly ineffective for the -6 dB and -3 dB hypoechoic cysts at the middle, as well as for the anechoic cyst at the bottom of the image. Conversely, the model with a fixed order of 4 improved the quality of the deeper cysts but degraded the contrast of the top cyst. The output of the adaptive FXPF is shown in (d), highlighting a solution that effectively combines the advantages of both previous settings. This achievement was made possible by adaptively adjusting the order, utilizing a lower-order model for shallower depths, and progressively increasing the order for deeper depths. Note that we

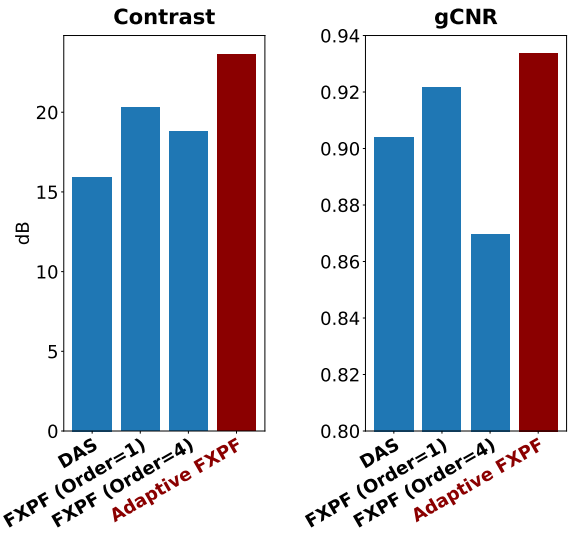


Figure A.2: Quantitative comparison between FXPF methods with fixed orders and an adaptive order.

generally observe more improvement for the bottom cyst when compared to the top one in all the images. This observation can be explained by the fact that the severity of the phase aberration effect, which requires correction, tends to be lower at shallower depths in contrast to deeper depths for two reasons. Firstly, perturbations in the wavefront become more pronounced as it propagates, resulting in an increased aberration effect during transmission as the wavefront advances. Secondly, as mentioned earlier, the aperture size is smaller at shallower depths, which mitigates the issue of incoherent summation at lower depths, as only a smaller number of neighboring elements are involved in the process of image reconstruction. The contrast and gCNR metrics were calculated on the envelope-detected image in the linear domain before applying the log-compression, where the target region was inside the solid red circle and the background was the region between two dashed blue concentric circles. The average values across two cysts are reported in Fig. A.2.

A.3 Conclusion

We demonstrated a challenge associated with phase aberration correction in plane-wave images using the FXPF method. This challenge arises due to substantial variations in channel data across elements at shallower depths, where signals from more distant elements lose relevance and can adversely affect the performance of the AR model. To address this challenge, we proposed the adaptive FXPF, which adjusts the order of the AR model

by employing a lower order for shallower depths and progressively increases the order for deeper depths. Both qualitative and quantitative results indicated that the adaptive approach provides higher performance in correcting the phase aberration effect.

Appendix B

RF Data Normalization for Deep Learning

This chapter is based on our published paper [177].

In Chapter 3, we introduced a method for normalizing RF data, distinct from conventional min-max scaling, to enable more efficient utilization of the generated dataset. This method is particularly relevant as DL-based techniques have recently gained considerable interest in medical ultrasound image processing, often demonstrating superior performance compared to traditional approaches in various tasks. These techniques leverage the power of neural networks to enhance image quality and aid in diagnostics. They have been applied not only to phase aberration correction tasks [81, 43] but also to tasks such as beamforming [178, 179, 180], speckle reduction [181], image segmentation, image registration [182], elastography [183], quantitative ultrasound [184, 185, 186], and more.

Additionally, there has been a growing adoption of RF data in DL-based approaches due to the fact that these methods prove more effective with more data, and RF data inherently contains richer information compared to envelope or B-mode data. RF data has a higher fidelity stemming from its raw and unprocessed form and contains complex details about the interaction between ultrasound waves and tissue structures. This makes it particularly well-suited for DL techniques, where these methods can leverage the complexity of RF data to detect subtle tissue differences, texture variations, and acoustic properties that might be missed in envelope or B-mode data. However, the highly fluctuating nature of RF data poses a challenge for neural networks to learn effectively during training, given that regions exhibiting comparable patterns might not appear very similar in RF data representation from a network's perspective. This challenge is exacerbated due to the high dynamic range of signal amplitudes. Substantial differences in amplitudes of raw ultrasound signals can

arise from variations in tissue density, acoustic impedance, the presence of bright specular reflectors, and other factors.

RF data may be acquired under varying power settings and from diverse sources, including various simulation packages and ultrasound machines, and needs to be normalized. Many researchers choose to utilize min-max scaling to align the RF data within a predefined range, such as [-1, 1], or less conventional ranges like [0, 1]. However, it is important to recognize the susceptibility of these techniques to RF data. While min-max scaling exhibits efficacy in the context of natural images, it may not be as effective for fluctuating RF signals with a very high dynamic range.

In this chapter, we demonstrate the inadequacy of the conventional min-max scaling techniques for normalizing RF data and illustrate how large amplitudes generated by a typical structure, such as a bright specular reflector, introduce challenges to the learning process of a neural network by preventing it from utilizing the data effectively. Additionally, we propose that employing a robust normalization method substantially improves the network's performance.

B.1 Methodology

B.1.1 Robust Normalization

Let us denote the RF data of the ultrasound image by $RF(x, y)$. A conventional normalization technique involves dividing the RF data by its maximum absolute value, resulting in $RF_{MaxAbs}(x, y)$ as given by

$$RF_{MaxAbs}(x, y) = \frac{RF(x, y)}{\max |RF(x, y)|}. \tag{B.1}$$

This step plays a critical role in transforming the RF data, acquired from various simulation packages and ultrasound machines, to a consistent range of [-1, 1]. We propose that following the prior step, applying individual standardization to the image substantially enhances the performance of deep neural networks by utilizing the data more efficiently:

$$RF_{Robust}(x, y) = \frac{RF_{MaxAbs}(x, y)}{\sigma} \tag{B.2}$$

where σ is the standard deviation of values across $RF_{MaxAbs}(x, y)$.

By individually dividing each image by its corresponding standard deviation, the RF

data is efficiently normalized with regard to its variability, which mitigates the impact of large echoes in the process of comparing different images. It is worth noting that this approach yields a distinct outcome compared to the well-known standardization technique applied within DL frameworks, which relies on dataset-wide statistics. In this particular context, the term "robust" indicates that the RF data from regions with similar patterns undergo a transformation that results in an increased similarity between the scale of their amplitude values. This interpretation of "robust" should not be confused with the notion that the standard deviation value is insensitive to the larger amplitude of bright echoes.

B.1.2 Dataset

We synthesized 100 single plane-wave images using a full synthetic aperture scan, each representing a randomly aberrated version of an identical phantom measuring 45 mm laterally and 40 mm axially containing two anechoic cysts. The full synthetic aperture scan was simulated using Field II [79], and images were synthesized as elaborated in Section 3.1.3. Additionally, among the aberrated images, one of them was selected and subsequently replicated. However, for the replicated version, we added a point target into the phantom before running simulations, introducing bright echoes into the RF data. Furthermore, we created non-aberrated versions of both phantoms, with and without the point target, for visualization purposes. Transducer settings used for simulation were similar to those of the 128-element linear array L11-5v (Verasonics, Kirkland, WA). The central and sampling frequencies were set to 5.208 MHz and 20.832 MHz, respectively. To ensure accurate numerical results in Field II simulations, we initially set the sampling frequency to 104.16 MHz and then downscaled the simulated data by a factor of 5.

B.1.3 Phase Aberration Correction Task

We evaluated the effectiveness of the robust RF data normalization in a phase aberration correction task by conducting a simple yet enlightening experiment. To correct the phase aberration effect, the aberration-to-aberration approach proposed in Chapter 3 was employed. In this approach, the network maps distinct randomly aberrated versions of the same realization to each other during the training phase and is expected to output a corrected version in the inference phase. Out of the 100 aberrated versions of the phantom without the point target, 99 versions served as a training set, and during each epoch, each of the 99 versions was randomly mapped to another one. The remaining version, along with its replica containing a point target, was reserved for evaluation purposes.

Two U-Nets [108] were trained using identical settings on the same training set with no point target. The only difference lay in the preprocessing steps: for the first network, each image in the dataset was normalized by a division with its maximum value, whereas for the second network, a robust normalization approach was employed. In the inference phase, the test images were also subjected to normalization, consistent with the method employed during the network's training.

The networks were trained for 1000 epochs with a batch size of 32, employing a linear activation function for the last layer. In both cases, the dataset was standardized by subtracting the mean and dividing it by its standard deviation. We utilized the adaptive mixed loss, proposed in Section 3.1.6, as the loss function and Adam [95] with a zero weight decay as the optimizer. The learning rate was initially set to 10^{-3} and halved at epoch 500.

B.2 Results and Discussion

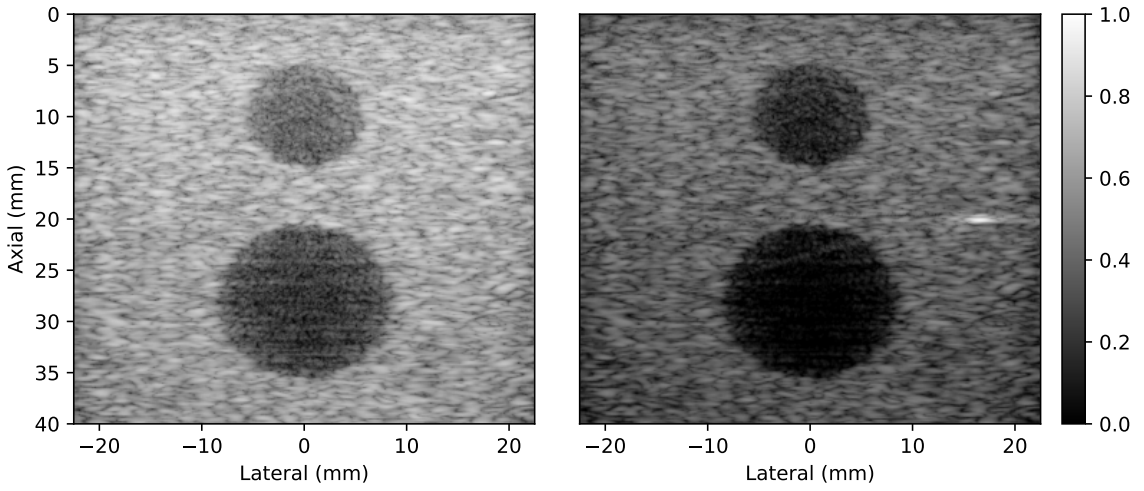


Figure B.1: A pair of ultrasound images simulated to be exactly identical using the Field II simulation package, where the only distinction between them lay in the presence of a point target within the second one. Both images are normalized in the same range and shown on the same dynamic range.

Consider a pair of ultrasound images meticulously simulated to be identical using the Field II simulation package. The only distinction between these images lay in the presence of a point target exclusively within the second one. The RF data of each simulated image was normalized by dividing it by its maximum absolute value. The B-mode images are shown in Fig. B.1. To isolate specific structures and analyze their amplitude variations,

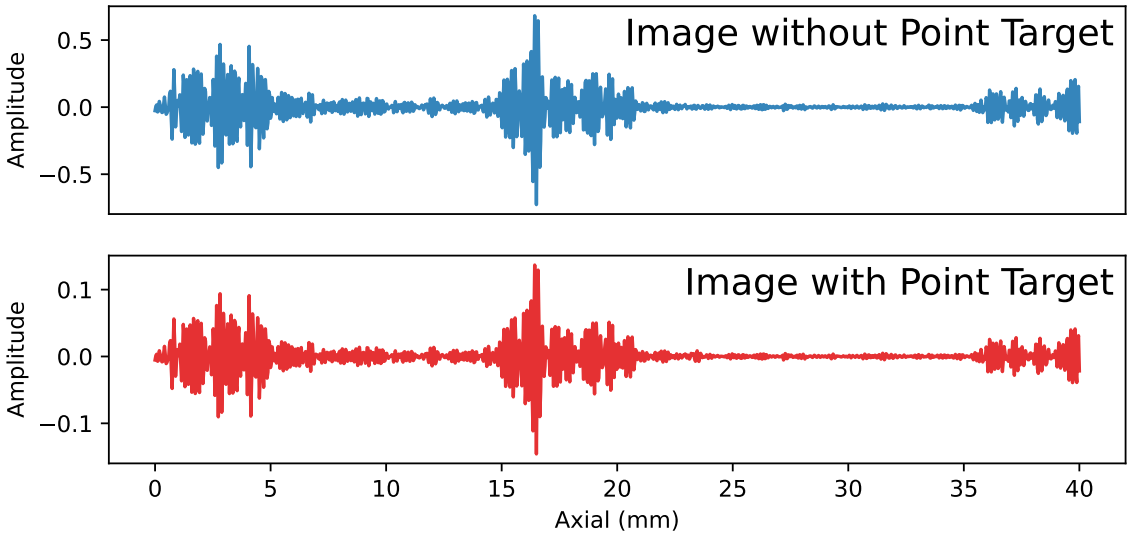


Figure B.2: The RF data associated with the middle column of images with and without the point target shown in Fig. B.1. The RF data were normalized by dividing them by their maximum absolute values across the entire image. In the top signal, the range of amplitude is roughly 5 times larger.

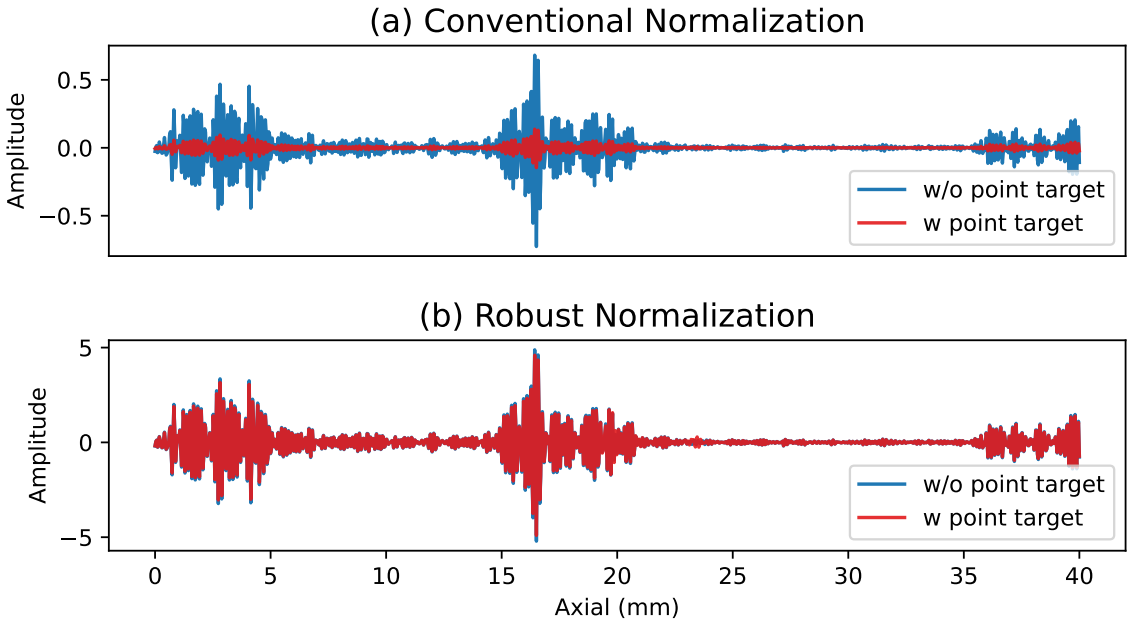


Figure B.3: The RF data associated with the middle column of images with and without the point target shown in Fig. B.1. In the bottom figure, the two signals almost overlap.

the RF data corresponding to the middle column of each image is plotted in Fig. B.2. It is evident that while the fundamental pattern of the RF data remains consistent, there is

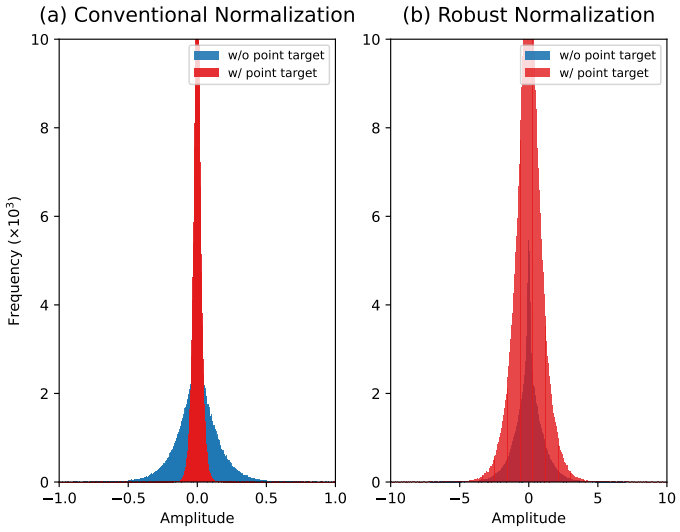


Figure B.4: Histograms of RF data associated with the images shown in Fig. B.1, where the data was normalized by (a) the conventional method and (b) the robust method.

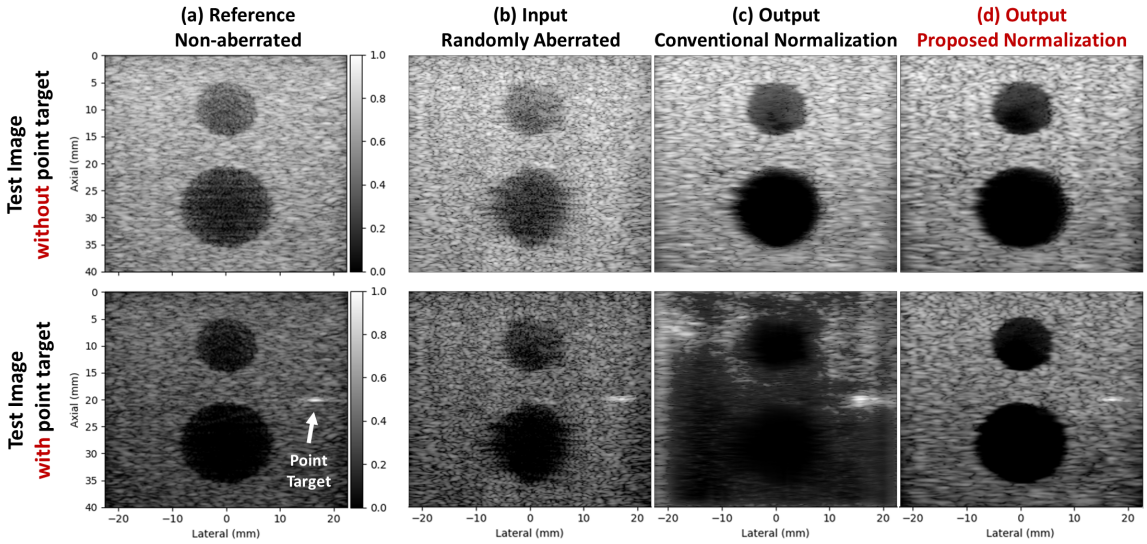


Figure B.5: Evaluating the efficacy of the robust normalization technique in a phase aberration correction task. (a) Non-aberrated reference images with and without point targets, reconstructed using DAS. (b) Randomly aberrated inputs with and without point target. (c) Output from the network trained on conventionally normalized data, utilizing similarly normalized inputs. (d) Output from the network trained on robustly normalized data, also with inputs normalized in a similar manner. All images are displayed in the same scale on a 50 dB dynamic range.

a substantial difference between their amplitude ranges. As expected, the image containing the point target exhibits markedly lower amplitudes compared to the image lacking the

said point target. The RF data for both images are superimposed in Fig. B.3(a), where the amplitude corresponding to the image with the point target became almost negligible compared to the other image. Therefore, from a network's perspective, an identical pattern in one image could be construed as a constant signal with a zero amplitude (e.g., an anechoic region) in the other image. Consequently, the network is unable to transfer its learned knowledge from one signal to the other.

By employing the proposed robust normalization to preprocess the RF data of images, a noticeable improvement was observed in addressing the previously mentioned issue. As illustrated in Fig. B.3(b), the amplitudes of RF signals, which were initially expected to be identical, subsequently fell within the same range. Note that in the context of conventional normalization, the RF data was normalized by dividing each image by its maximum absolute value present within the whole image. Consequently, as shown in Fig. B.3, the expected outcome was that the amplitude of the RF signal would be confined within the range of [-1, 1]. However, by applying the robust normalization, the amplitude values expanded to cover a broader range beyond [-1, 1]. This expansion in range does not raise any concerns as long as the same robust normalization approach is consistently applied during both the training and testing phases. The histograms shown in Fig. B.4 represent the full RF data for both images, which were normalized using conventional and robust techniques.

To evaluate the effectiveness of the proposed robust normalization technique in a phase aberration correction task, we trained two different networks utilizing an identical dataset devoid of any point targets. Therefore, neither of the networks had been exposed to any point targets during the training phase. Nonetheless, during the training process, the initial network's dataset images underwent normalization through the conventional method, while the second network's images were normalized using the robust approach. We provided both networks with two aberrated test images: one containing a point target, and the other without, as illustrated in Fig. B.5(b), with all images displayed in the same scale on a 50 dB dynamic range. As we can see in the outputs shown in (c), the inclusion of a point target within the RF data introduced large echoes, and applying a conventional normalization resulted in the compression of all other values towards proximity to zero, producing an inferior output. In contrast, as shown in (d), applying robust normalization effectively mitigated the impact of those large echoes on the remaining data. It enabled the network to leverage its learned knowledge from the other image and output a corrected image, even in the case where its training set did not include any point targets.

It is essential to recognize that adding bright specular reflectors to the training dataset can indeed help the network in enhancing its ability to handle these features. However, it is

crucial to emphasize that such augmentation does not serve as a replacement for the robust normalization technique. Even by adding those reflectors with different random intensities to the dataset, the network still may fail to establish a coherent relationship between similar structures when their amplitudes occur on different scales. This effect is comparable to dividing the dataset into multiple smaller subsets based on the presence of large echoes and their amplitudes, which subsequently results in a suboptimal efficiency.

B.3 Conclusion

We investigated the importance of normalizing RF data on the performance of DL-based approaches and demonstrated the inadequacy of conventional min-max scaling techniques, particularly in a phase aberration correction task. We showed that standardizing RF data individually, considering the variability within each image, leads to a more consistent range of amplitude values for similar regions across different images. This process alleviates the impact of large echoes and helps the network seamlessly transfer its learned knowledge across images, resulting in higher performance.

Appendix C

Segmentation of Intraoperative 3D Ultrasound Images Using a Pyramidal Blur-Pooled 2D U-Net

This chapter is based on our published paper [187].

In this appendix, we benchmark the PBP U-Net proposed in Chapter 4 to perform two tasks requested by the CuRIOUS 2022 - Segmentation Challenge organizers: segmentation of the brain tumor in pre-resection 3D ultrasound images (Task 1) and segmentation of the resection cavity in post-resection 3D ultrasound images (Task 2). The success rate of safely resecting a brain tumor during neurosurgery highly depends on an accurate and reliable intraoperative neuronavigation [188]. Preoperative imaging methods such as MRI play a pivotal role in neurosurgery; however, distortions, deformations, and brain shifts make those images less valuable during the operation. Intraoperative ultrasound is an affordable, safe, and real-time imaging technique that, due to its high temporal resolution, can be easily incorporated into the surgical workflow and provides live imaging during surgery. Although ultrasound images are more difficult to interpret than those from other modalities, such as MRI, automatic segmentation of intraoperative ultrasound images provides an effective solution to this issue by, for instance, facilitating the registration of preoperative MRI and intraoperative ultrasound images [189]. We employ the PBP U-Net to segment the tumor and resection cavity before, during, and after resection in 3D intraoperative ultrasound images. Slicing the 3D image along three transverse, sagittal, and coronal axes, we train a different model corresponding to each axis and average three predicted masks to obtain the final prediction. It is demonstrated that the averaged mask consistently achieves a DSC greater than or equal to each individual mask predicted by only one model along one axis.

C.1 Methodology

C.1.1 Dataset

We used RESECT, the publicly available dataset, including preoperative contrast-enhanced T1-weighted and T2 FLAIR MRI scans alongside three 3D volumes of intraoperative ultrasound scans acquired from 23 clinical patients with low-grade gliomas before, during, and after undergoing tumor resection surgeries [190]. In this appendix, only the ultrasound volumes were employed for the segmentation tasks, where delineations of the brain tumors and resection cavities had been provided as ground truths in addition to the original database [191].

We split the dataset into training and validation sets containing 19 and 4 cases, respectively. Seven additional cases, provided by the challenge organizers, were used as the test set. All volumes were normalized between 0 and 1, individually, then zero-padded symmetrically to the maximum size existing in the dataset along each axis and finally resampled to the size of $150\times150\times150$.

C.1.2 Network Architecture

We employed the PBP U-Net, proposed in Chapter 4, a variant of U-Net that is more robust to the shift-variance problem and provides higher output consistency. Compared to the vanilla U-Net, the max-pooling layers are replaced with blur-pooling layers in PBP U-Net. The anti-aliasing filters in pyramidal blur-pooled were of sizes 7×7, 5×5, 3×3, and 2×2, from the first to the forth downsampling layer, respectively.

C.1.3 Training Strategy

For each task, we trained three different models using the 2D images acquired by slicing 3D volumes along three transverse, sagittal, and coronal axes. By slicing 3D volumes, we obtained a highly imbalanced dataset wherein many images had a completely black mask (no foreground). To mitigate this issue and to achieve a faster training time, we trained the models merely using images with a non-zero mask (including at least one pixel as the foreground) and discarded the rest. However, even in this case, the dataset was still imbalanced as a large majority of the pixels were background in the remaining masks. To alleviate this problem, we employed the focal Tversky loss function which is a generalized focal loss function based on the Tversky index and was proposed to address the issue of

data imbalance in medical image segmentation [63].

For Task 1 experiments, we merely used before-resection volumes; however, for Task 2 experiments, both during, and after-resection volumes were combined and considered as one dataset. Task 2 experiments were trained from scratch; whereas Task 1 experiments were initialized using pre-trained weights obtained from Task 2.

The sigmoid function was employed as the activation function of the last layer, and the batch size was 32. We utilized AdamW [141] as the optimizer, and set the weight decay parameter to 10^{-2}. We trained each network for 500 epochs, and saved model weights only if the validation loss had been improved and finally used the best weights for testing. The learning rate was set to 2×10^{-4} initially and was lowered by 2 times at epochs 300 and 400. The same configuration was used for all experiments. They were implemented using the PyTorch package [142], and training was performed on two NVIDIA A100 GPUs utilizing the DataParallel class, which parallelizes the training by splitting the input across the two GPUs by chunking in the batch dimension.

C.1.4 Augmentation

During the training, six on-the-fly augmentation techniques were applied. We randomly scaled the 2D images by $s\%$ along both axes, where $s \in [-7, +7]$. We also applied a Gaussian smoothing filter with a kernel of size $k \times k$ and standard deviation σ, where $k \in \{0, 2, 3, 5\}$ and $\sigma \in [0, 0.6]$ were chosen randomly. Besides, we randomly altered images' brightness and contrast and flipped (horizontally) and rotated ($\theta \ degrees$) them, where the chance of flipping was 50%, and $\theta \in [-15, 15]$. As the dominant noise source in ultrasound images, we modeled the speckle noise as a multiplicative noise and randomly applied it to the images:

$$\boldsymbol{I_{noisy}} = \boldsymbol{I} + \boldsymbol{NI} \tag{C.1}$$

where $\boldsymbol{N}$ is a matrix with the same size of the image consisting of normally distributed values with zero mean and standard deviation $\sigma \in [0, 0.01]$.

Finally, we randomly cropped a patch of size 128×128 from images of the original size 150×150, which is equivalent to the translation augmentation. Since choosing and storing the best model was based on the validation set, we always used center-copped images without any augmentations during the validation phase.

Table C.1: DSC of the validation set, where B, D, and A stand for before, during, and after resection, respectively, with the highest values shown in bold.

	Case #24			Case #25			Case #26			Case #27		
Stage	B	D	A	B	D	A	B	D	A	B	D	A
Axis 0	0.25	0.13	0.87	0.66	0.83	0.88	0.87	0.89	0.95	0.70	0.11	0.93
Axis 1	0.27	0.1	0.83	0.63	0.78	0.84	0.91	0.85	0.96	0.76	0.16	0.94
Axis 2	0.27	0.05	0.86	0.64	0.78	0.87	0.90	0.88	0.90	0.78	0.11	0.84
Final	**0.47**	**0.23**	**0.88**	**0.73**	**0.89**	**0.92**	**0.92**	**0.93**	**0.96**	**0.85**	**0.21**	**0.94**

C.2 Results

To predict the segmentation mask of each 3D volume, we followed the pre-processing procedure same as for the training. The volume was normalized between 0 and 1, zero-padded symmetrically to the maximum size existing in the dataset along each axis, resampled to the size of $150\times150\times150$, and center-cropped to $128\times128\times128$. Then it was sliced along the three axes, and 2D slices along each axis were fed into the corresponding network to predict 2D masks. Afterward, 2D masks were stacked together to form three 3D volumes, each corresponding to one of the axes. Finally, we averaged three volumes and thresholded the resulting volume at 0.5. To make sure that the mask size matches the original image size, it was zero-padded symmetrically to the size of $150\times150\times150$ and resampled to the size of the original image volume to obtain the final predicted mask.

Applying the method to before, during, and after resection volumes of all cases in the validation set, Table C.1 shows the resulted DSC according to the predicted masks and ground truths. Although the dataset contained 23 cases, note that the cases in the table are labeled based on the original labels in the publicly available dataset, wherein they were not necessarily numbered consecutively. According to Table C.1, an easy (#26) and a difficult (#24) case are chosen, and the qualitative results are demonstrated in Fig. C.1. It shows the results for Task 1, where the tumor is segmented in a pre-resection image, and Task 2, where the resection cavity is segmented in a post-resection image. Finally, the performance across the test set is summarized in Table C.2 based on the averaged DSC, HD95, recall, and precision metrics provided by the challenge organizer after submitting the results.

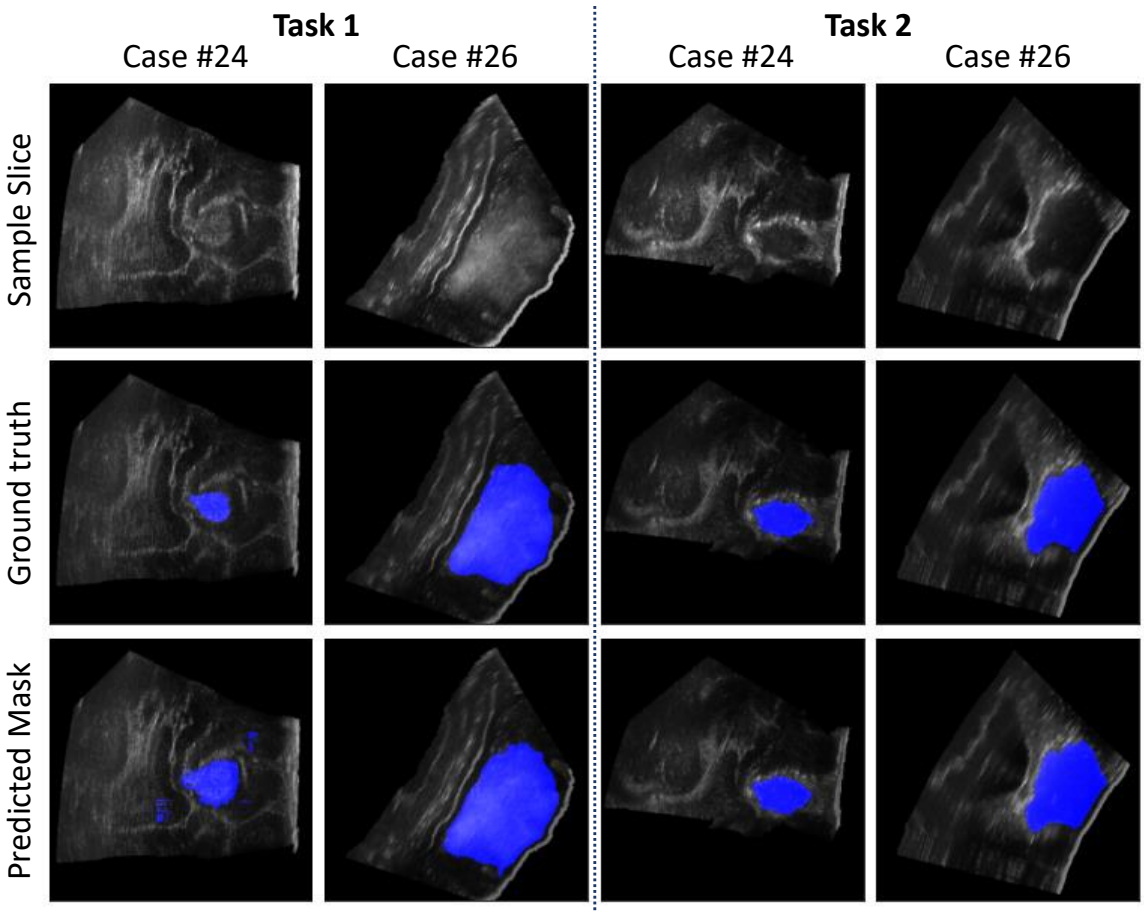


Figure C.1: Sample slices from two cases of the validation set. Cases #24 and #26 represent the lowest and highest DSC, respectively. The first and second columns show samples of Task 1, where the tumor is segmented in a pre-resection image, and the third and fourth columns correspond to Task 2, where the resection cavity is segmented in a post-resection image.

C.3 Discussion and Conclusions

We benchmarked the PBP U-Net as a baseline method to segment the tumor and resection cavity in 3D intraoperative ultrasound images. To this end, we predicted three different masks for each volume based on the slicing axis and averaged them before thresholding to obtain the final mask. In Table C.1, we can see that the final mask consistently achieved a DSC greater than or equal to each individual mask predicted by only one model along one axis. In Table C.2, a drop in performance can be observed for the test set, compared to the validation set. To further improvement of the performance, an n-fold cross-validation approach could be followed.

Table C.2: Performance of the method across the test set.

	DSC	HD95	Recall	Precision
Task 1	0.53	71.57	0.64	0.54
Task 2	0.62	36.08	0.54	0.80

One of the drawbacks of this appendix was that we separated 4 cases as the validation set and those cases were never taken into account for the training, which means losing a large portion of data in a small dataset of only 23 cases. Another limitation was utilizing only one slice as the input; however, similar to [192, 193], adjacent slices also could be fed into the network as the input channels to improve the predictions.

In this appendix, we performed training and testing on the same small dataset. However, higher performance is expected by augmenting the small training set with more annotated data. For instance, the approach described in Chapter 5 for ultra-fast simulation of realistic ultrasound images could be utilized to augment the dataset with simulated data. Another approach would be augmenting the training set with other annotated ultrasound datasets. In these approaches, employing a domain adaptation method [157] to mitigate the domain shift problem between two datasets plays a pivotal role in achieving higher performances.