Efficient Ultrasound Image Enhancement Using Lightweight CNNs

by

Farid Anjidani
B.Sc., University of Tehran, 2020

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF APPLIED SCIENCE

in the Department of Electrical and Computer Engineering

We acknowledge and respect the lekwungen peoples on whose traditional territory the university. stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day

Efficient Ultrasound Image Enhancement Using Lightweight CNNs

by

Farid Anjidani
B.Sc., University of Tehran, 2020

Supervisory Committee

Dr. Daler N. Rakhmatov, Supervisor
(Department of Electrical and Computer Engineering)

Dr. Mihai Sima, Departmental Member
(Department of Electrical and Computer Engineering)

ABSTRACT

Plane-wave ultrasound imaging allows for very high frame rates. During image reconstruction, conventional delay-and-sum beamforming can be replaced by the quicker Fourier-domain remapping method. Typically, after Fourier-domain reconstruction, postbeamforming interpolation is needed to increase the image grid resolution in the lateral dimension. To achieve this, we propose to use a fast lightweight superresolution convolutional neural network (CNN) operating on the Fourier-beamformed envelope data.

Specifically, we train different configurations of well-known *Efficient Sub-Pixel Convolutional Neural Network* (ESPCN) to perform both 1D and 2D upscaling. First, we pretrain a network using the diverse (non-ultrasound) dataset DIV2K. Then, we apply transfer learning on a small augmented dataset of public-domain experimental ultrasound images. Our results demonstrate that our approach is capable of producing enhanced ultrasound images having higher quality compared to non-CNN interpolation options and conventional delay-and-sum beamforming.

Contents

Supervisory Committee ii

Abstract iii

Contents iv

List of Tables vi

List of Figures vii

List of Acronyms ix

Acknowledgements xi

Dedication xii

1 Introduction 1

1.1 Ultrasound Basics 1

1.1.1 Ultrasound Imaging System 1

1.1.2 Ultrasound Imaging Modes 2

1.1.3 Image Quality Metrics 3

1.2 Plane-Wave Imaging 6

1.3 Thesis Contributions and Organization 12

2 Background 14

2.1 Frequency-Domain Beamforming 14

2.1.1 Temme-Mueller Migration 14

2.2 Deep Neural Networks 16

2.2.1 Convolution Layer 17

2.2.2 Fully-connected Layer 18

2.2.3 Activation Layer 18
2.2.4 Squeeze and Excitation Block 19
2.3 Single-Image Superresolution 19
2.3.1 SISR Problem Formulation 20
2.3.2 ESPCN Structure 21
2.4 Related Work 22

3 Efficient Ultrasound Image Enhancement Using Lightweight CNNs [1] 27
3.1 ESPCN Modifications 28
3.1.1 Two-Dimensional Upscaling 29
3.1.2 One-Dimensional Upscaling 29
3.2 Training 32
3.3 Experimental Results 32

4 Data Augmentation and Transfer Learning in Ultrasound Image Enhancement [2] 40
4.1 Data Augmentation for Ultrasound Imaging 40
4.1.1 Data Augmentation Methods 41
4.1.2 Applied Data Augmentation 42
4.2 Transfer Learning for Ultrasound Imaging 43
4.2.1 General-Domain Training 44
4.2.2 Target-Domain Training 44
4.3 Experimental Results 45

5 Conclusions and Future Work 51
5.1 Conclusions 51
5.2 Future Work 52

Bibliography 53

List of Tables

3.1 Complexity of Proposed ESPCN Configurations. 30
3.2 Image Quality Assessment: Single Plane Wave, $\alpha = 0°$. 35
3.3 Image Quality Assessment: 75 Plane Waves, $-16° \leq \alpha \leq +16°$. . . . 35

4.1 Image Quality Assessment: Single Plane Wave, $\alpha = 0°$. 48
4.2 Image Quality Assessment: 75 Plane Waves, $-16° \leq \alpha \leq +16°$. . . . 48

List of Figures

1.1 A-mode image example [3] 2
1.2 B-mode image example [4] 3
1.3 M-mode image example [3] 4
1.4 Full Width at Half Maximum (FWHM) illustration. 4
1.5 (a) Conventional focused beam transmission. (b) Plane-wave imaging [5] 6
1.6 Image reconstruction (beamforming) [6]. 7
1.7 Horizontal and vertical axes [7] 8
1.8 PW transmission [7] 8
1.9 Backscattered echoes [7] 9
1.10 Time delay for a plane wave [7] 9
1.11 Dynamic beamforming [7] 10
1.12 Time delay for a plane wave of angle α [7] 11

2.1 Convolution layer 17
2.2 Fully-connected layer 18
2.3 ReLU function 19
2.4 Squeeze and Excitation block [8] 20
2.5 ESPCN architecture featuring convolution layers for feature maps extraction, and a sub-pixel convolution layer that aggregates the feature maps from the LR space and builds the HR image in a single step. [9] 22

3.1 Proposed 1D-upscaling network with Squeeze and Excitation blocks. . 31
3.2 Baseline DAS beamforming, 75-PW compounded B-mode images (60-dB range) and evaluated regions: Type-1 (left), Type-2 (center), Type-3 (right). 34
3.3 Examples of Type-1 (left) and Type-2 (right) 75-PW compounded B-mode images: TI+BI. 36

3.4 Examples of TYPE-1 (left) and TYPE-2 (right) 75-PW compounded B-mode images: TM+N5*. 37
3.5 Examples of TYPE-1 (left) and TYPE-2 (right) 75-PW compounded B-mode images: TI+BI (top), TM+N5* (center), TM+BI+N7 (bottom). 38
3.6 Examples of TYPE-3 75-PW compounded envelope contours: TM+BI+N7. 39

4.1 Conventional DAS beamforming (75 plane waves, Tukey-25% window): TYPE-1 (left), TYPE-2 (center), and TYPE-3 (right) compounded B-mode images, 60-dB range. 47
4.2 Conventional DAS beamforming (zero-angle PW) with various apodizations: TYPE-1 B-mode images, 60-dB range. 49
4.3 TM migration (zero-angle PW) with various upscaling options: TYPE-1 B-mode images, 60-dB range. 50

List of Acronyms

BI	Bicubic Interpolation
CPWC	Coherent Plane-Wave Compounding
CT	Computed Tomography
CNN	Convolutional Neural Network
CNR	Contrast-to-Noise Ratio
DAS	Delay-and-Sum (Beamforming)
DSP	Digital Signal Processing
DNN	Deep Neural Network
ESPCN	Efficient Sub-pixel Convolutional Neural Network
FFT	Fast Fourier Transform
FI	Fourier Interpolation
FPS	Frame Per Second
FPGA	Field Programmable Gate Array
FWHM	Full Width at Half Maximum
GAN	Generative Adversarial Networks
IFFT	Inverse Fast Fourier Transform
MRI	Magnetic Resonance Imaging
MSE	Mean-Squared Error
MSLE	Mean Squared Logarithmic Error
MVDR	Minimal Variance Distortionless Response
PSNR	Peak Signal-to-Noise Ratio
PW	Plane Wave

PWI	Plane Wave Imaging
RoI	Region of Interest
ReLU	Rectified Linear Unit
SISR	Single-Image Super-Resolution
SSIM	Structural Similarity Index
TM	Temme-Mueller (Migration)
US	UltraSound

ACKNOWLEDGEMENTS

First and foremost, I would like to give my warmest and most sincere thanks to my supervisor, Dr. Daler N. Rakhmatov for his extensive support, enthusiasm, and encouragement throughout all stages of my study and research.

I would like to thank Dr. Mihai Sima (my supervisory committee member) for his support and insightful feedback, which have contributed to the improvement of this thesis.

I would like to express my sincere gratitude to Dr. Barbara D. Sawicki (my external examiner) for her valuable suggestions and crucial role in the successful completion of this thesis.

I would especially like to thank Sude and Pooria, my extraordinary friends and colleagues, for creating such a great and professional research environment.

Lastly, I would like to thank my amazing family and friends for their love and support along the way.

DEDICATION

To my Mom, Hamdam, and all the brave women in Iran

Chapter 1

Introduction

Ultrasound imaging is well-suited for obtaining subsurface images of the body's internal anatomy using high-frequency sound waves for diagnostic and therapeutic purposes, such as cancer detection, blood flow characterization, and fetal monitoring [10]. The fundamental idea behind ultrasonic imaging is echolocation, which utilizes sound reflections at acoustical interfaces. Compared to other commonly used imaging modalities such as CT and MRI for example, ultrasound imaging is non-ionizing and cost-effective, which gives rise to its popularity in the medical field [11]. There are many other important applications of ultrasound imaging, such as range detection, fluid flow monitoring, and non-destructive testing to detect flaws in mechanical structures such as failing welds [12].

Next, we present some basic background on biomedical ultrasound, which typically uses wave frequencies between 1 and 20 MHz. Then, we introduce a few important concepts related to ultrafast imaging with plane waves [13], which is the application focus of this thesis.

1.1 Ultrasound Basics

1.1.1 Ultrasound Imaging System

An array of piezoelectric elements makes up the transducer, which transforms electrical energy into ultrasound waves and vice versa during transmission and reception, respectively. The sound waves move through the medium and hit different targets that have different acoustic impedances. This causes reflections, and the transducer elements detect and turn the backscattered echoes into electrical impulses [14].

There are different kinds of transducer arrays: linear sequential array, linear phased array, curved sequential array, curved phased array, and annular array. Each transducer element works as a transmitter, emitting ultrasound waves towards the region of interest, and a receiver, acquiring echoes coming back from the insonified medium. Transducers can emit different kinds of waves (e.g., plane wave, diverging wave, focused wave) at different steering angles [15].

1.1.2 Ultrasound Imaging Modes

The images generated by an ultrasound device can be displayed in various ways, which are called ultrasound imaging modes.

A-mode, or amplitude mode, measures the arrival time of the echoes and displays their envelope versus propagation depth [6]. The one-dimensional format of the A-mode offers limited spatial information about the observed structure. A-mode is also used in therapeutic ultrasound to target a specific anomalous region so that the destructive wave energy may be precisely focused [3]. Fig. 1.1 shows an A-mode image example.

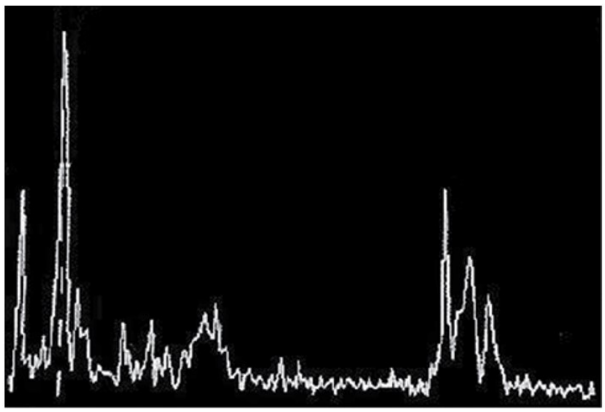


Figure 1.1: A-mode image example [3]

B-mode, or brightness mode, is the most common form of ultrasound imaging that uses high-bandwidth pulses to image a 2-D plane in the insonified target medium. In contrast to A-mode, B-mode produces two- or three-dimensional images in which the amplitude of the beamformed echoes is translated into pixel brightness. Each pixel's point location in the medium being imaged is given by its horizontal and vertical coordinates, and the echo strength is determined by the grayscale intensity of that pixel [6]. Fig. 1.2 shows an example of a B-mode image.

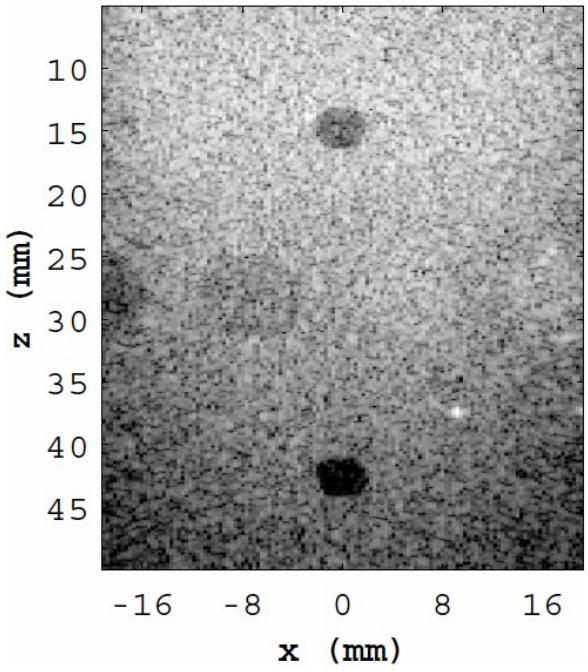


Figure 1.2: B-mode image example [4]

M-mode, or motion mode, is represented as a time-motion display of the acoustic wave along a selected image line. In contrast to a B-mode image, only one B-mode line is necessary, and information is repeatedly extracted from one line to assess the medium's motion. This makes it possible to show the movement of structures as a dynamic wave-like motion [16]. Fig. 1.3 is an example showing an M-mode image, placed below a B-mode image of the heart. The vertical axis corresponds to a selected depth section along some chosen A-scan line, while the horizontal axis represents time, capturing several heartbeats.

Doppler mode is based on the Doppler effect, which detects a change in a sound wave's frequency as a result of relative motion between a sound source and a receiver [10]. The primary application is monitoring and observing the blood flow throughout the body.

This work deals exclusively with two-dimensional B-mode imaging; therefore, all the discussions that follow fall into this category.

1.1.3 Image Quality Metrics

There are many different metrics to measure the quality of ultrasound images. Here, we briefly discuss different metrics that are most commonly used in the literature.

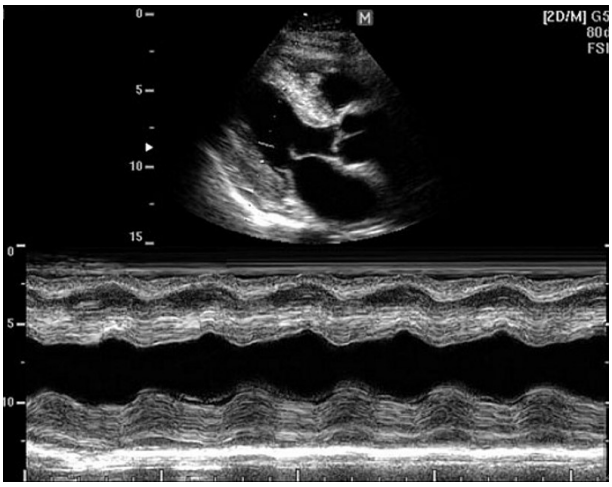


Figure 1.3: M-mode image example [3]

- **Full Width at Half Maximum(FWHM)**
  FWHM is commonly used to quantify a pulse width, as shown in Fig. 1.4. It is measured as the distance between coordinates x_1 and x_2 corresponding to the pulse curve points that are halfway away from the peak value f_{max}. This metric can be used to assess image resolution. For example, when a B-mode image contains a bright point-like target (i.e., a two-dimensional "pulse"), one can measure FWHM values in the lateral (horizontal) as well as axial (vertical) directions. Smaller FWHM indicates better resolution.

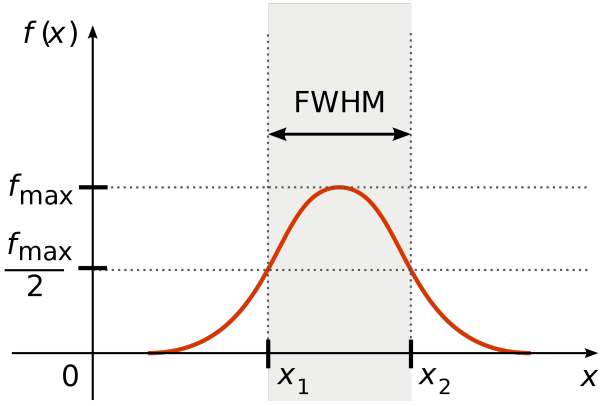


Figure 1.4: Full Width at Half Maximum (FWHM) illustration.

- **Contrast-to-Noise Ratio**
  In medical imaging, the desired information is often related to contrast rather

than the signal itself, therefore, one should use contrast-to-noise ratio (CNR) instead of signal-to-noise ratio (SNR). The former quantifies a system's ability to distinguish a certain region of interest (RoI) from its surrounding background. It can be expressed as [17]

$$CNR = 20\log_{10}(\frac{|\mu_{in} - \mu_{out}|}{\sqrt{(\sigma_{in}^2 + \sigma_{out}^2)/2}}), \tag{1.1}$$

where μ_{in} and μ_{out} are the mean signal levels inside and outside the RoI, and σ_{in} and σ_{out} are the corresponding standard deviations. The larger CNR value indicates better contrast.

- **Mean-Squared Logarithmic Error**

  The mean-squared error (MSE) is a very common criterion for indicating the degree of similarity between two images. It is always non-negative, and it is widely used as a loss function in machine learning. The MSE for two images I_1 and I_2 can be expressed as below:

$$MSE = \frac{\sum_{m,n}[I_1(m,n) - I_2(m,n)]^2}{MN}, \tag{1.2}$$

  where M and N are the height and width (in pixels) of the images, respectively.

  Another similar metric is Mean-Squared Logarithmic Error (MSLE), which is defined as an average of the pixel-wise cumulative squared logarithmic error:

$$MSLE = \frac{\sum_{m,n}[\log(I_1(m,n)) - \log(I_2(m,n))]^2}{MN}. \tag{1.3}$$

  The most important difference between the MSE and the MSLE is that the MSLE penalizes underestimation more than overestimation. This implies that if $I_2(m,n) < I_1(m,n)$, it will incur a bigger loss than in case of $I_2(m,n) > I_1(m,n)$ with the same absolute difference. Also, the MSLE does not depend on the image intensity scaling.

- **Peak Signal-to-Noise Ratio**

  Peak Signal-to-Noise Ratio (PSNR) is defined as the squared maximum value of the original image divided by the MSE error. Since many signals have a very wide dynamic range, the PSNR is usually expressed in decibels (dB):

$$PSNR = 10\log_{10}(\frac{(Max(I_1))^2}{MSE}), \tag{1.4}$$

where MSE is the mean-squared error which can be calculated using equation 1.2.

In this work, our image quality assessments rely primarily on FWHM and CNR measurements.

1.2 Plane-Wave Imaging

In conventional (focused) ultrasound imaging, the acoustic waves focus on a target at a certain depth, as illustrted in Fig. 1.5(a). This results in good image quality around the target area but low quality elsewhere within the same image. Therefore, one needs to perform multiple transmissions in sequence (scanning across the imaged section), which reduces the data acquisition rate given by:

$$r = \frac{v}{2z_{max}N_s}. \tag{1.5}$$

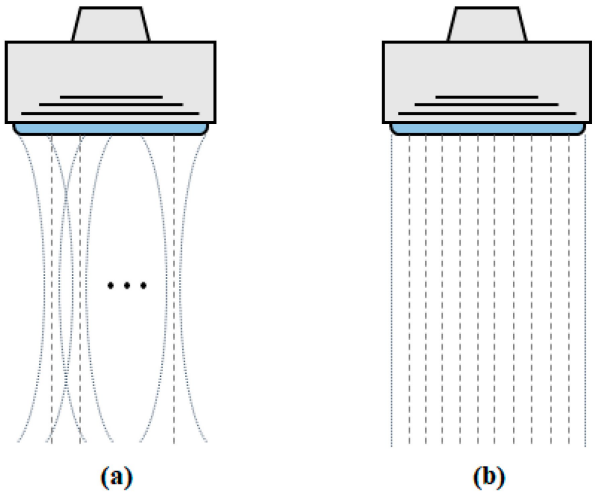


Figure 1.5: (a) Conventional focused beam transmission. (b) Plane-wave imaging [5]

In the above formula, N_s is the number of image scan lines, z_{max} is the maximum depth of interest, and v is the speed of sound. The latter is assumed to be constant, with a typical value of 1540 m/s for soft tissues. The resulting rate r is typically

under 100 frames per second (fps) which is inadequate for many applications such as blood flow analysis and tissue elastography [18].

With plane-wave imaging (PWI), the data can be acquired at a much higher rate of hundreds or even thousands of fps. Since the delay between consecutive image frames becomes smaller, it allows one to capture fast changes in the imaged medium. With this method, a sequence of focused transmissions is replaced with a single plane wave, as illustrated in Fig. 1.5(b).

Fig. 1.6 depicts a simplified view of ultrasound image reconstruction, or beamforming. After the transducer array emits an ultrasound wave, it receives the resulting backscattered echoes and stores their digitized values in memory. Such raw radio-frequency (RF) channel data is then processed by a beamformer to produce a reconstructed image. In the sequel, we present main ideas behind time-domain beamforming for PWI.

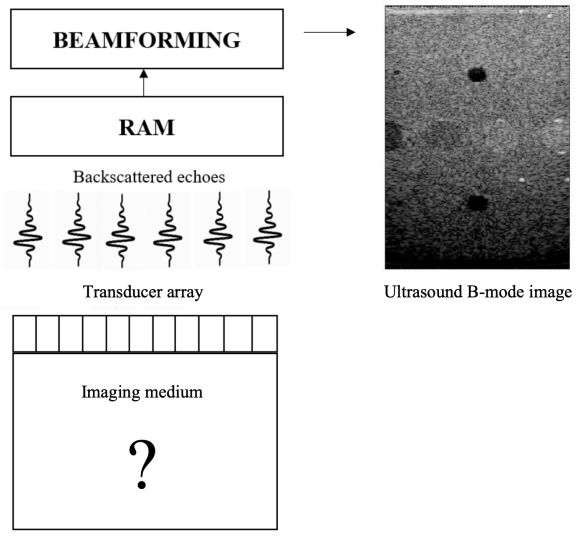


Figure 1.6: Image reconstruction (beamforming) [6].

As we mentioned, a faster alternative to sequential transmission of multiple focused beams is to insonify the region of interest at once using a single plane-wave (PW) emission. Fig. 1.7 shows the axis convention of a plane-wave imaging system. The ultrasound probe is made up of an array of transducer elements that are positioned along the x-axis on the surface of the imaging medium. The z-axis, perpendicular to the x-axis, represents the imaging depth.

A PW pulse is first sent into the medium by the transducers and gets backscattered when it encounters objects with different acoustic impedances, as seen in Fig. 1.8

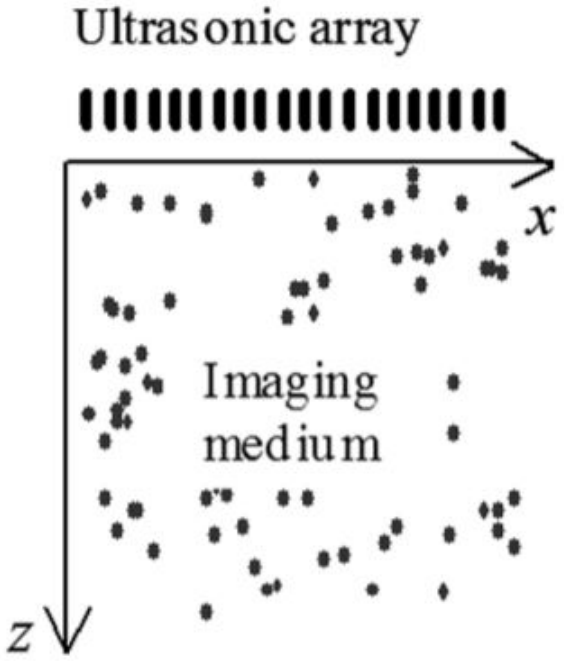


Figure 1.7: Horizontal and vertical axes [7]

and 1.9. The returning echoes are then recorded over time t by the transducer array and processed to form a PW image in two spatial dimensions (x, z).

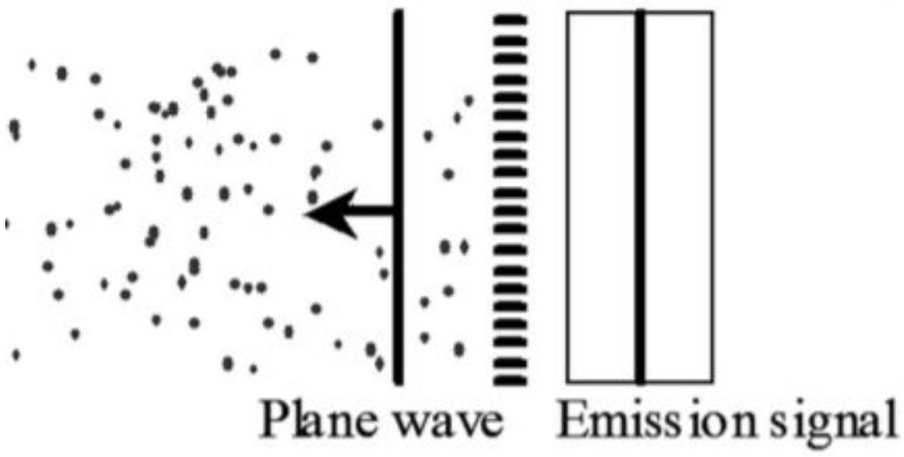


Figure 1.8: PW transmission [7]

As shown in Fig. 1.10, the travel time of a plane wave from the transducer array to point (x, z) and back to receiving element at x_1 can be expressed as:

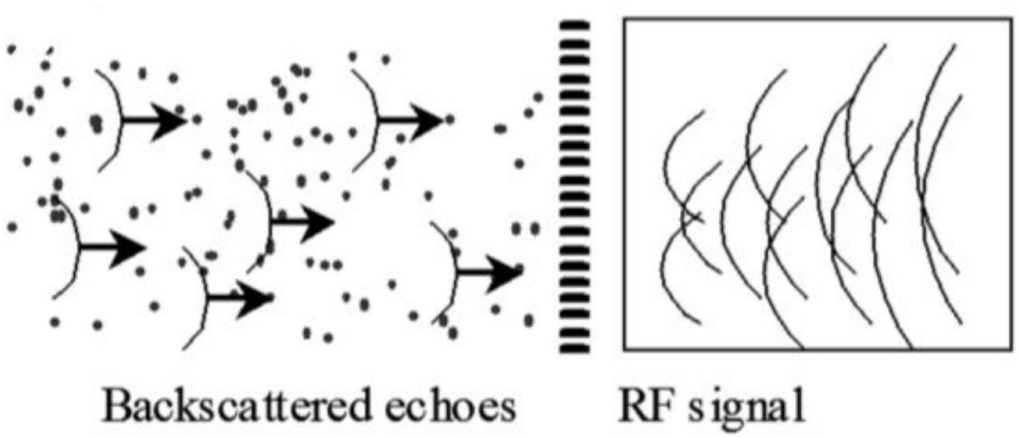


Figure 1.9: Backscattered echoes [7]

$$\tau(x_1, x, z) = \frac{z + \sqrt{z^2 + (x - x_1)^2}}{v}, \tag{1.6}$$

where v is the speed of sound, is assumed to be constant throughout the medium.

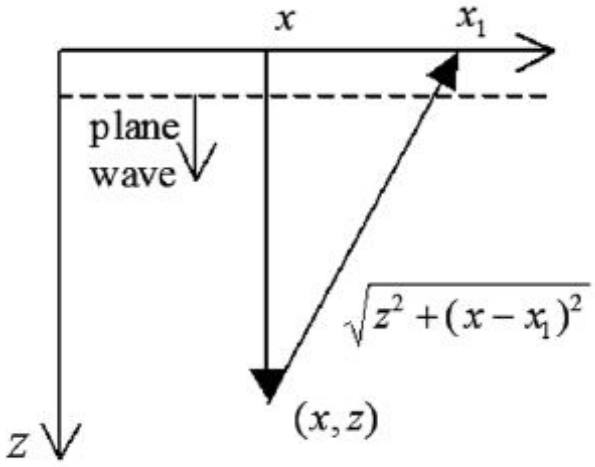


Figure 1.10: Time delay for a plane wave [7]

By coherently adding the contribution of each scatter, the imaged value at point (x, z) is given by:

$$D(x, z) = \sum_{i=n_{start}}^{n_{end}} RF(x_i, \tau(x_i, x, z))w_i, \tag{1.7}$$

where n_{start} and n_{end} represent the positional range of the transducer elements that contribute to the weighted sum of received echo signals $RF(x_i, t)$ after applying appropriate delays $\tau(x_i, x, z)$. Element-dependent delays τ are calculated and applied at each (x, z) point to produce the entire image, as shown in Fig. 1.11.

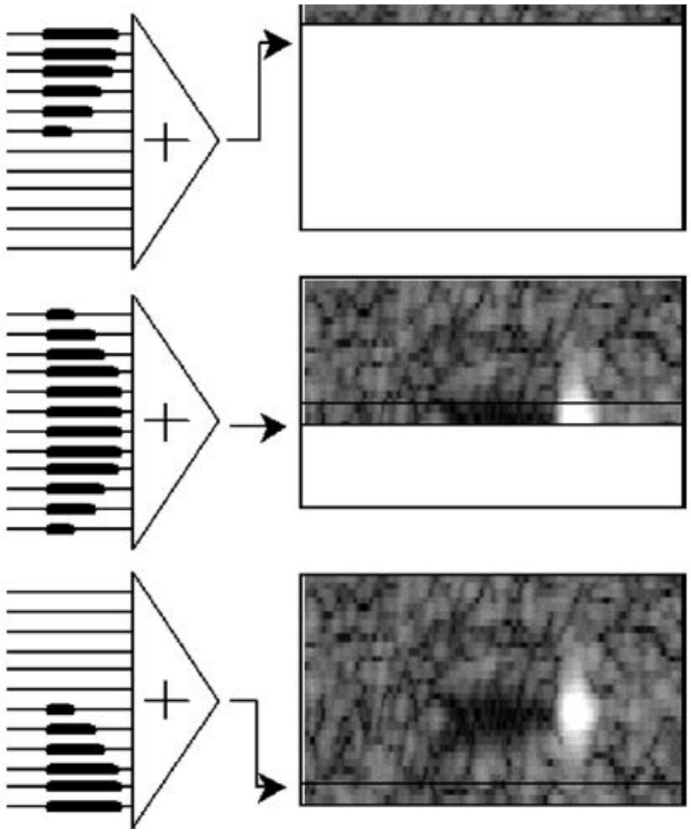


Figure 1.11: Dynamic beamforming [7]

Weights w_i in equation 1.7 can be either adaptive or non-adaptive when applied to delayed signals. In non-adaptive beamforming approaches, a fixed window function (such as Rectangular, Hamming, Hanning, Tukey, etc.) is chosen, which is independent of the input data [19]. In this work, we employ non-adaptive DAS beamforming with the Tukey apodization window and the 1/4 cosine fraction.

It has been shown that adaptive beamformers can enhance image quality by suppressing side lobes, but the weights must be dynamically determined based on the properties of the input data. For instance, the minimal variance distortionless response (MVDR) beamforming approach aims to reduce output power while letting the desired signal pass through undistorted (with unity gain)[20]. In some other works,

such as [21], [22], a convolutional neural network has been used to predict the weights for adaptive beamformers. Although adaptive beamformers produce higher-quality images, they are computationally more expensive than conventional non-adaptive beamformers.

Compared to a traditional ultrasound imaging system, a single PW image reconstruction has a much higher frame rate, but the lack of focus during transmission reduces the image quality. One approach to address the issue of PW image quality degradation is coherent plane-wave compounding (CPWC) [7]. CPWC yields better image quality by combining the beamformed data produced by successive transmission of tilted plane waves at various angles. For a PW tilted at an angle α as shown in Fig. 1.12, the travel time equation 1.6 for τ becomes modified as follows:

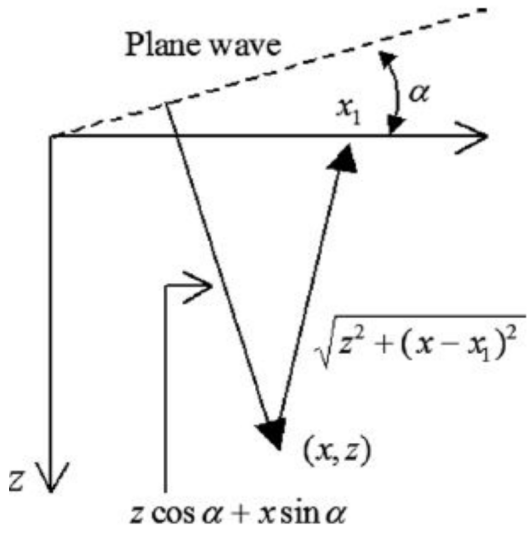


Figure 1.12: Time delay for a plane wave of angle α [7]

$$\tau(\alpha, x_1, x, z) = \frac{[(zcos\alpha + xsin\alpha) + \sqrt{z^2 + (x - x_1)^2}]}{v}. \tag{1.8}$$

Each plane wave of a specific tilted emission angle produces an image using equation 1.7, where the τ delays are now given by equation 1.8. The resulting angle-specific beamformed images are them coherently summed to produce the final compounded image. Compared to conventional sequential imaging employing focused transmission beams, CPWC can still achieve a comparatively high frame rate of over 1000 fps [7], despite the fact that generating multiple plane waves takes more time and thus reduces the frame rate.

1.3 Thesis Contributions and Organization

The image reconstruction technique described in the previous section is called *delay-and-sum* (DAS) beamforming. It is the most widely used technique in both PW and non-PW ultrasound imaging due to its flexibility, as it operates in the time domain. The main disadvantage of DAS beamforming is its relatively high computational complexity, dominated by applying τ delays to the received data (see equations 1.7 and 1.8) using interpolation. If the beamformed image size is $N_x \times N_z$, and the number of transducer elements is N_e, then DAS beamforming would take $O(N_x N_z N_e)$ time per PW angle α.

It is possible to perform PW image reconstruction in the Fourier (i.e., frequency) domain, which is usually several times faster than time-domain DAS beamforming. One example of such a method is *Temme-Mueller (TM) migration* originating from the geophysical literature [23]. Instead of point-by-point beamforming, it reconstructs an entire image at once as follows. First, the 2D Fourier transform is applied to the received data frame. Then, the resulting 2D spectrum is interpolated based on a special remapping formula, Finally, the 2D inverse Fourier transform is applied to the interpolated spectrum, which gives the beamformed image.

The computational complexity of TM migration is dominated by the Fourier transforms. If the received data frame $RF(x_i, t)$ is of size $N_e \times N_t$, then TM migration would take $O(N_e N_t \log(N_e N_t))$ time per PW angle α. Assuming for simplicity that $N_e = N_x$ and $N_t = N_z$, we get $O(N_z N_x^2)$ for DAS beamforming and $O(N_x N_z \log(N_x N_z))$ for TM migration. Since $N_x > \log(N_x N_z)$ in practice, TM migration turns out to be more efficient than DAS beamforming.

Due to the use of 2D Fourier transforms, TM migration outputs a reconstructed image whose size matches the size of the input data, i.e., $N_e \times N_t$. However, the desired image size $N_x \times N_z$ may be such that $N_x > N_e$, which means that the Fourier-beamformed image in question needs to be upscaled in the lateral dimension. Such upscaling can be done using interpolation, or alternatively, one can utilize a neural network. The latter approach is the subject of this thesis. In this work, we propose using a lightweight, fast convolutional neural network (CNN) to perform lateral upscaling, which also happens to improve image quality (namely, FWHM and in some cases CNR). We study different structures of superresolution CNNs, all derived from the well-known *Efficient Sub-Pixel Convolutional Neural Network* (ESPCN) [9].

The rest of this thesis is organized as follows: Chapter 2 presents theoretical

foundations of TM migration and provides background on convolutional neural networks (CNNs) in general and ESPCN in particular. It also summarizes related work. Chapter 3 describes our proposed approach and training methods for enhancing the quality of ultrasound images using lightweight superresolution CNNs. In chapter 4, we present the data augmentation methods utilized by the transfer learning technique to increase the image quality further. Chapter 5 provides concluding remarks and outlines some possible directions for future work.

Chapter 2

Background

2.1 Frequency-Domain Beamforming

One of the main methods for image reconstruction is Fourier-domain beamforming [24, 25, 26, 27, 28, 4, 29], that has proved to be advantageous over conventional time-domain DAS beamforming described in the previous chapter. Its advantages include computational efficiency, reduced sampling complexity, and improved image quality. In this work, we use Temme-Mueller (TM) migration [23] technique to convert the raw radio-frequency (RF) data in the (x, t) domain into beamformed in-phase/quadrature (IQ) data in the (x, z) domain, where t stands for time, and x and z are the lateral (along the transducer array) and axial (depth) coordinates, respectively.

2.1.1 Temme-Mueller Migration

Let $U = (x, z, t)$ denote the subsurface reflected wavefield resulting from a plane-wave emission at an angle α. We assume that the sound velocity v is constant throughout the medium. TM migration is based on the so-called imaging condition that can be expressed as below [23]:

$$D_\alpha(x, z) = U(x, z, t_I = \frac{x}{v}\sin\alpha + \frac{z}{v}\cos\alpha). \tag{2.1}$$

$D_\alpha(x, z)$ represents the migrated depth section. Note that the expression for the imaging condition t_I is the same as the expression for the PW transmits propagation delay portion (i.e., transducer-to-reflector travel time) in equation 1.8.

Considering the transducers elements record echoes at the surface $z = 0$ as a 2D

raw dataset $U(x, z = 0, t) = RF(x, t)$, the migration process aims to translate this raw dataset into a 2D beamformed data $D_\alpha(x, z)$. This process is driven by the use of the following equation in the frequency-wavenumber domain 2.2:

$$\frac{d^2\Psi}{dz^2} + (\frac{\omega^2}{v^2} - k_x^2)\Psi = 0, \tag{2.2}$$

where Ψ represents the Fourier-transformed reflected wave field:

$$\Psi(k_x, z, \omega) = \iint_{-\infty}^{\infty} U(x, z, t)e^{-j(\omega t + k_x x)}dtdx. \tag{2.3}$$

The general solution of Eq. 2.2 is as below:

$$\Psi(k_x, z, \omega) = C_{\text{up}}e^{j(\frac{\omega^2}{v^2} - k_x^2)^{\frac{1}{2}}z} + C_{\text{down}}e^{-j(\frac{\omega^2}{v^2} - k_x^2)^{\frac{1}{2}}z}. \tag{2.4}$$

The first term with the constant C_{up} represents waves going upwards in the negative z direction. On the other hand, the second term with the constant C_{down} represents waves going downward in the positive z direction. Following [23] we are only interested in the reflected waves; therefore, we set $C_{\text{down}} = 0$. Also, C_{up} can be determined from the boundary condition for $z = 0$. Indeed, setting $z = 0$ in 2.4 yields $C_{\text{up}} = \Psi(k_x, 0, \omega)$. Since $U(x, z = 0, t) = RF(x, t)$ is known (data recorded by the transducer array at the surface), its Fourier transform $\Psi(k_x, 0, \omega)$ is known as well.

Next, we replace $\Psi(k_x, z, \omega)$ in 2.3 by $\Psi(k_x, 0, \omega)\exp(j\sqrt{\omega^2/v^2 - k_x^2}z)$ from 2.4. To obtain $U(x, z, t)$, we simply need to apply the inverse Fourier transform to the left-hand side of 2.3:

$$U(x, z, t) = \frac{1}{4\pi^2}\iint_{-\infty}^{\infty} \Psi(k_x, 0, \omega).e^{j(\omega t + k_x x + (\frac{\omega^2}{v^2} - k_x^2)^{\frac{1}{2}}z)}d\omega dk_x \tag{2.5}$$

Finally, by letting $t = t_I$ from 2.1 we get the following equation,

$$D_\alpha(x, z) = \frac{1}{4\pi^2}\iint_{-\infty}^{\infty} \Psi(k_x, 0, \omega).e^{j(\omega\frac{x}{v}\sin\alpha + \omega\frac{z}{v}\cos\alpha + k_x x + (\frac{\omega^2}{v^2} - k_x^2)^{\frac{1}{2}}z)}d\omega dk_x. \tag{2.6}$$

This gives us $D_\alpha(x, z)$, but computing it directly is expensive because the frequency-wavenumber integral must be evaluated for each depth level z. We can resolve this issue by rewriting 2.6 as a Fourier integral after introducing two new variables A and

B as follows:

$$A(w, k_x) = \frac{w}{v}\sin\alpha + k_x \tag{2.7}$$

$$B(w, k_x) = \frac{w}{v}\cos\alpha + (\frac{\omega^2}{v^2} - k_x^2)^{\frac{1}{2}} \tag{2.8}$$

If we solve 2.7 and 2.8 for k_x and ω then insert them in2.6, we obtain:

$$k_x(A, B) = \frac{(A^2 - B^2)\sin\alpha + 2AB\cos\alpha}{2(A\sin\alpha + B\cos\alpha)}, \tag{2.9}$$

$$\omega(A, B) = \frac{v(A^2 + B^2)}{2(A\sin\alpha + B\cos\alpha)}. \tag{2.10}$$

After changing our integration variables k_x and ω to A and B respectively, we obtain the following expression [23]:

$$\begin{aligned} D_\alpha(x, z) = \frac{1}{4\pi^2}\iint_{-\infty}^{\infty} &\Psi(k_x(A, B), 0, \omega(A, B)) \\ &\times \frac{v|(A^2 - B^2)\cos\alpha - 2AB\sin\alpha|}{2(A\sin\alpha + B\cos\alpha)^2} e^{j(Ax+Bz)} dAdB. \end{aligned} \tag{2.11}$$

The above equation shows how one can get $D_\alpha(x, z)$. First, we compute the Fourier transform of the recorded data $U(x, z = 0, t) = RF(x, t)$, which gives $\Psi(k_x, 0, \omega)$. Second, we remap the latter into the migrated spectrum $\Psi(A, 0, B)$ via interpolation using equations for $k_x(A, B)$ and $\omega(A, B)$. Third, we multiply $\Psi(A, 0, B)$ pointwise by the factor $\frac{v|(A^2-B^2)\cos\alpha-2AB\sin\alpha|}{2(A\sin\alpha+B\cos\alpha)^2}$. Finally, compute the inverse Fourier transform of the resulting spectrum, which gives $D_\alpha(x, z)$.

2.2 Deep Neural Networks

In the past decade, Deep Neural Networks (DNN) have proved themselves as powerful tools for solving various real-world problems. In 2012, [30] made a breakthrough in Machine Learning by achieving a much lower top-5 error on the ImageNet challenge [31] compared to the runner-up. After that, some deep neural network models were proposed that surpassed human performance on the ImageNet large-scale visual recognition challenge [31]. Currently, DNNs have numerous applications, such as image analysis, speech recognition, natural language processing, and recommendation

systems. For each task, many models have been proposed in recent years. In the sequel, we will see the basic building blocks of these models.

A DNN layer consists of small individual units called neurons. Each neuron has a coefficient called weight, which is to be learned. Neurons are the building blocks of many layers, but not all layers have neurons. Different layers are used in DNN models, such as the convolution layer, fully-connected layer, activation layer, and pooling layer. There are many other layers, but we will not talk about them in this chapter.

2.2.1 Convolution Layer

This layer is one of the main reasons behind the breakthrough of deep neural networks in 2012. Mainly, it is used for feature detection. Not all input neurons are connected to output neurons in the convolution layer. The two-dimensional convolution is primarily used; in this scenario, there is an image and a filter, as illustrated in Fig. 2.1. The filter slides on the image, and by multiplying the corresponding pixels and summing them up, an output pixel is generated. In the convolution layer, the local relationship between the pixels is extracted, which makes the convolution layer an excellent building block for image processing models. A hyper-parameter in neural networks is stride value, which determines the amount of filter movement over the image when sliding.

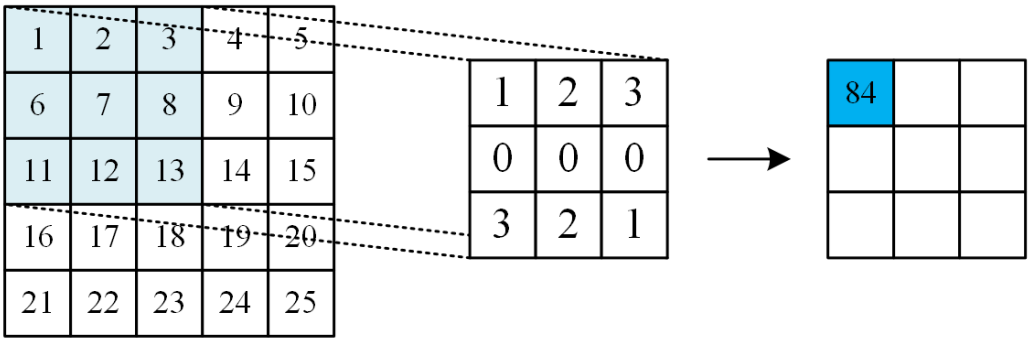


Figure 2.1: Convolution layer

Separable convolution is a variant of the convolution layer. The separable convolution was introduced to decrease the computations and parameters of the neural network, and it is mainly used for mobile devices. It consists of two different layers: depthwise and pointwise convolutions. For example, MobileNet [32] uses a com-

bination of depthwise and pointwise convolutions to lower the number of network parameters and operations, targeting resource-constrained applications.

2.2.2 Fully-connected Layer

In the fully-connected (FC) layer, in contrast with the convolution layer, all the preceding layer neurons are connected to the succeeding layer neurons. Mainly, these layers are placed last for classification tasks. The FC layer helps map the representation between the input and the output. We can see a fully connected layer structure in Fig. 2.2.

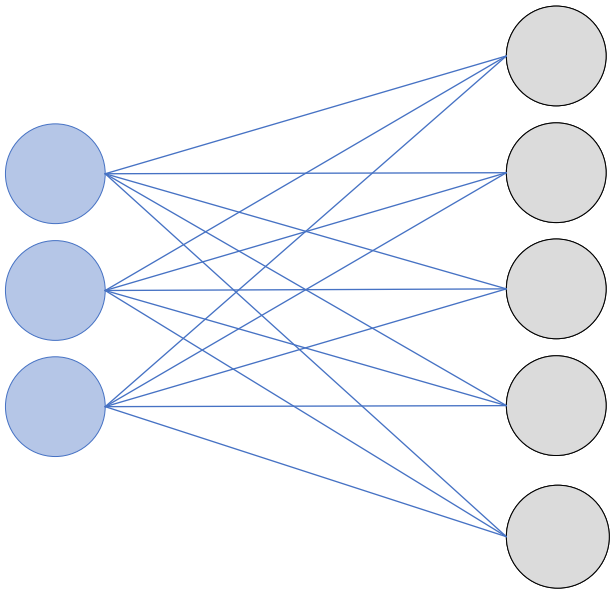


Figure 2.2: Fully-connected layer

2.2.3 Activation Layer

Inspired by biological neurons, there is a layer that decides whether neurons are activated. A non-linear transformation is applied to the input data by this layer. In neural networks, various activation functions exist, each with its unique advantages

and limitations. Among them, the Rectified Linear Unit (ReLU) is a well-known example that is widely used due to its computational efficiency and demonstrated effectiveness in numerous applications. The ReLU function is shown in 2.3. The choice of activation function can have a significant impact on the performance of the neural network.

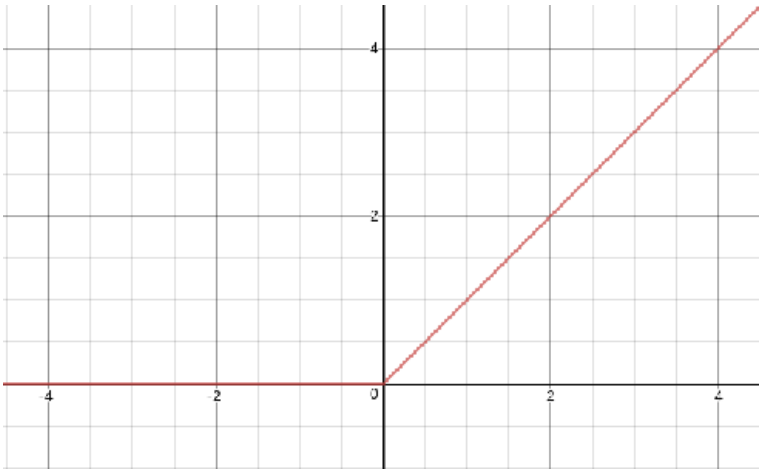


Figure 2.3: ReLU function

2.2.4 Squeeze and Excitation Block

The Squeeze and Excitation block [8], illustrated in Fig. 2.4, enhances channel interdependencies at almost no computational cost. CNNs extract hierarchical data from images using their convolutional filters. Upper layers may recognize faces, letters, or other complicated geometrical forms, while lower layers locate basic context elements like edges. Networks give each channel the same weight when making output feature maps, but the Squeeze and Excitation block implements an adaptive channel weighting method that takes into account the content of each channel. This can be achieved by adding a single parameter to each channel and giving it a linear scalar reflecting how relevant that channel information is.

2.3 Single-Image Superresolution

Image superresolution (SR) is one of the areas of machine learning that has received increasing attention in recent years. A superresolution task is to convert a low-resolution image to a corresponding high-resolution image with better visual quality

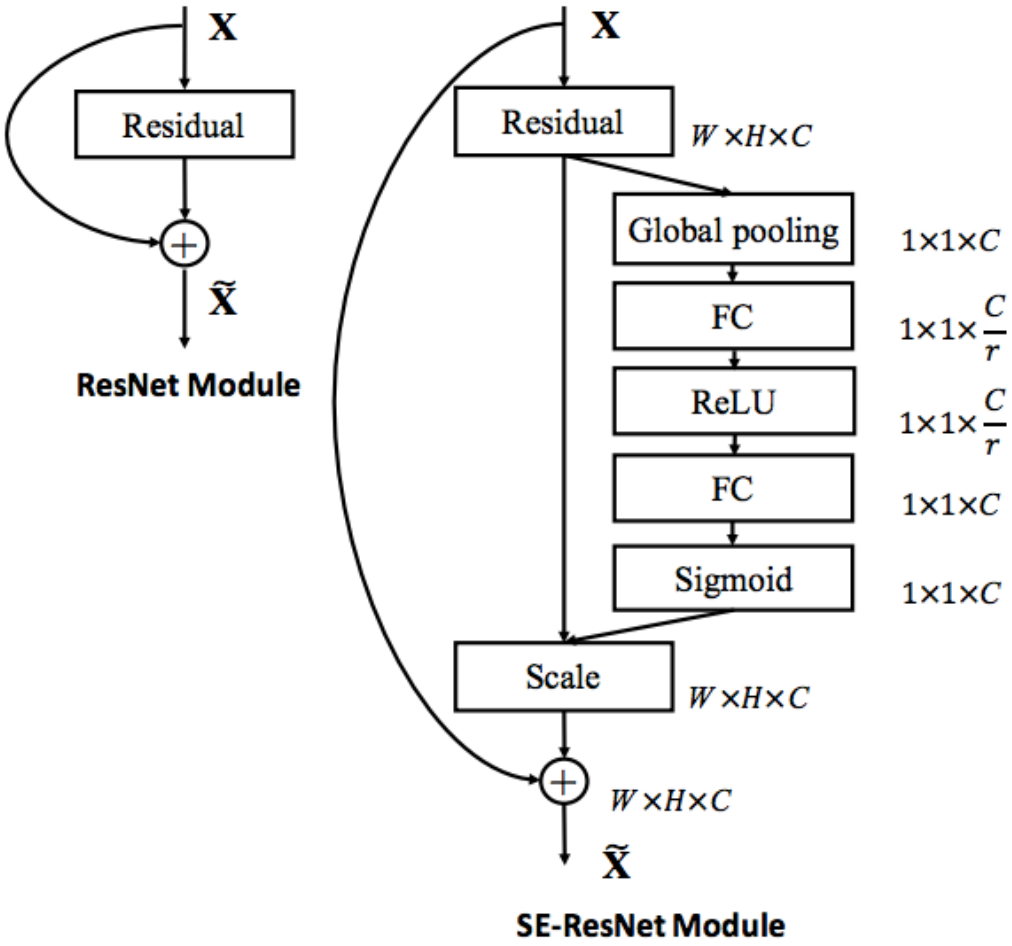


Figure 2.4: Squeeze and Excitation block [8]

and more refined details. The high-resolution image can also be referred to as an upsampled or interpolated image. Superresolution can be achieved using a single image or multiple images as the input, but in this section, we focus on single-image superresolution (SISR), which has attracted extensive interest in recent years.

2.3.1 SISR Problem Formulation

Let us consider a low-resolution (LR) image denoted by y and the corresponding high-resolution (HR) image denoted by x, We can view y as an outcome of some degradation process applied to x [33]:

$$y = \Omega(x; \delta_\eta), \tag{2.12}$$

where Ω is the image degradation function, and δ_η is the degradation parameter, which can be affected by noise, compression, and other factors. In many cases, only y is available without any precise information about δ_η or Ω. Superresolution aims to decrease degradation as much as possible and recover an approximation $\hat{x}$ of the ground-truth image x. Effectively, we have an inverse problem that can be formulated as:

$$\hat{x} = \Omega^{-1}(y; \delta_\zeta), \tag{2.13}$$

where δ_ζ is a parameter for the inverse of the degradation function Ω^{-1}. Given that the degradation process can be quite complex, most works prefer to approximate 2.12 as:

$$y = (x \otimes k) \downarrow_s + n, \tag{2.14}$$

where k is the blurring kernel, and $x \otimes k$ is the convolution operation involving the ground-truth image and the blurring kernel. Symbol $\downarrow_s$ refers to a downsampling operation with the factor of s after convolution, while n represents the noise commonly assumed to be white Gaussian. Under such modelling, the inverse function is essentially a deconvolution operation.

2.3.2 ESPCN Structure

Efficient sub-pixel convolutional neural network (ESPCN) [9] is a well-known architecture that transforms a low-resolution input image $y = I_{LR}$ of size $H \times W \times C$ into a high-resolution output image $\hat{x} = I_{HR}$ of size $sH \times sW \times C$, where s is the scaling factor. As this network needs a single image as an input, it belongs to the class of SISR architectures. In ESPCN, in contrast to some other networks such as [34, 35, 36], increasing the resolution of the input image happens at the end of the network, which greatly reduces the computational complexity. Therefore, the key advantage of the EPSCN architecture over other alternatives is its low computational cost, which meets our need for highly efficient ultrasound image formation.

Given I_{LR} as an input image, an L-layer ESPCN uses the first $L-1$ convolution layers to perform feature extraction in order to do upscaling. The first $L-1$ layers produce a tensor of size $H \times W \times s^2C$. Based on [9], the nonlinear mapping of each layer denoted as f^l can be expressed as follows:

$$f^l(I_{LR}) = \Phi(w_1 \times f^{l-1}(I_{LR}) + b_l),\ 1 \leq l \leq L-1, \tag{2.15}$$

where Φ is the activation function, w_1 is a weight tensor of size $n_{l-1} \times n_l \times k_l \times k_l$, and b_l is the bias vector of length n_l. Symbols k_l and n_l represent the convolution filter size and number of features at layer l, respectively. We also have $f^0 = I_{LR}$ and $n_0 = C$.

As illustrated in Fig. 2.5, the final layer plays a crucial role in upscaling. This layer converts the generated feature maps into a final high-resolution image I_{HR} using fast subpixel shuffling [9].

$$I_{HR} = f^L(I_{LR}) = S(w_L \times f^{L-1}(I_{IR}) + b_L), \tag{2.16}$$

where S represents the subpixel shuffling operation which reshapes $H \times W \times s^2C$ tensor (low-resolution feature maps) into the $sH \times sW \times C$ tensor (high-resolution image).

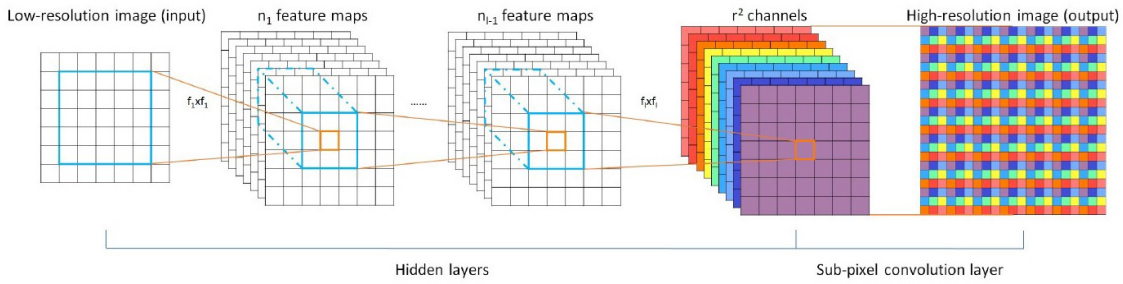


Figure 2.5: ESPCN architecture featuring convolution layers for feature maps extraction, and a sub-pixel convolution layer that aggregates the feature maps from the LR space and builds the HR image in a single step. [9]

2.4 Related Work

Using deep learning in ultrasound imaging has been extensively studied, targetting not only pre-beamforming and post-beamforming[37, 38, 39, 40, 41, 42], but also the beamforming process itself [21, 43, 44, 45, 46, 47].

In [40], the authors propose a two-step CNN-based image reconstruction approach that is compatible with real-time imaging to address the growing demand for high-quality imaging from single unfocused acquisitions. A residual CNN with multiscale

and multichannel filtering, trained to remove the diffraction artifacts inherent to ultrafast US imaging, is used to restore a high-quality image from a low-quality estimate obtained through a back projection-based operation similar to conventional delay-and-sum beamforming. They use the mean signed logarithmic absolute error (MSLAE) as a training loss function to take into consideration both the high dynamic range and the oscillatory characteristics of radio frequency US images.

Reference [47] describes a convolutional neural network beamformer based on the combination of GoogleNet and U-Net. This network's input is RF data, and the output is beamformed IQ data. In training datasets, the authors used both simulation and experimental data for generating pre-training datasets and transfer learning datasets. After training, the CNN is employed as a beamformer that applies specific weights to the echoes received from different angles to reconstruct the image. This CNN-based beamformer is designed to learn the best weights for each angle to enhance image quality, and a backpropagation algorithm updates the weights during the training process.

In order to improve the image quality of handheld or portable ultrasound devices, a two-stage generative adversarial network (GAN) structure is developed in [37]. The generator's front-end tool at stage 1 is a U-Net model. In the reconstructed images, it retrieves structural characteristics at low frequencies. A GAN network is used in stage 2 to identify the latent space between images of low and high quality. The output image of the U-Net model and a low-quality image are both inputs to the generator. Reconstructed generator images and high-quality real images are the two inputs the discriminator requires to be trained to distinguish between real and generated images. This technique exhibits encouraging outcomes and could potentially enhance the availability of high-quality medical imaging in resource-constrained areas.

In [38], the authors propose neural networks to enhance raw RF signals in the frequency domain and then perform the inverse Fourier transform to reconstruct the time-domain signals subject to beamforming. In contrast, we do not modify raw RF data in any way, we perform image reconstruction in the Fourier domain instead, and we apply a neural network only after the beamforming process is completed.

Another approach targets coherent compounding. In [48], the authors describe a CNN that takes 3-PW compounded images and produces output images whose quality is comparable to those using a larger number of PWs for compounding. In order to improve the resolution of images acquired from a single plane wave insonification, Perdios et al. [39] introduced an image enhancement approach based on three different

U-Net-like convolutional neural networks which use simulated data for training.

Reference [41] uses a U-Net-like GAN structure and attention mechanism for PW image reconstruction. It focuses on an image-to-image enhancement that directly applies the proposed CNN to the beamformed image in order to get a higher-quality image.

In [42], the authors employ an unsupervised CycleGAN technique, which has shown effective for numerous image domain quality enhancements, including ultrasound. This network is used as a denoising filter to reduce speckles in US images. In this approach, high-quality ground-truth images are not needed, as they are difficult to obtain.

Traditional interpolation methods, such as bicubic and Lanczos, use mathematical algorithms to interpolate the missing high-frequency details. On the other hand, deep learning-based methods, such as CNNs and GANs, use large amounts of training data to learn the mapping between low-resolution and high-resolution images. Single-image super-resolution (SISR) which has been broadly studied in the past few years, aims to reconstruct a high-resolution image from a single low-resolution image. The end goal of SISR is to upscale a low-resolution image to a high-resolution image while preserving its visual quality and content.

The upscaling process can be applied in early stages, such as in SRCNN [35]. Such upscaling can be computationally expensive, so some networks perform upscaling in the later stages [49], [50]. In other words, most of the data processing in such networks takes place in low-resolution space. ESPCN belongs to this category as well.

In [51], the authors have used separable convolutions within the ESPCN structure and demonstrated their benefit in terms of both reduced network complexity and improved PSNR performance. Additionally, they introduced a supersampling approach: by oversampling the output image and then downsampling it adaptively, they believe more plentiful high-resolution information may be learned. The authors also describe recursive depth-wise separable convolutions to reuse the shared model parameters and achieve a higher level of image quality through recursive iterations without increasing the number of model parameters.

Designing very deep networks is made possible by residual learning, which employs skip connections to prevent gradients from vanishing. In residual super-resolution, one example is the original ResNet [52] architecture for image classification modified by the Enhanced Deep Super-Resolution (EDSR) [49]. To be more specific, the authors

showed notable benefits by removing batch normalization layers (from each residual block) and ReLU activation (outside residual blocks). They have also expanded their technique to work on multiple scales. In our work, we are also not using any batch normalization layer because it can interfere with the information that is necessary to perform super-resolution.

Ultrasound datasets are difficult and expensive to acquire, and data augmentation techniques can be helpful in addressing this problem. In [53], by training eight different CNNs to predict mammography masses, the authors have evaluated the performance of eight different image-augmenting transformations (such as flips, rotations, Gaussian filtering, etc.) and discovered that applying a combination of multiple augmentation techniques considerably enhanced the classification performance. Another approach relies on generative adversarial networks to generate new data. In [54], the authors investigate CycleGAN's potential for augmenting data in CT segmentation tasks. They trained a CycleGAN to convert contrast CT images into non-contrast images using a large image collection. The trained CycleGAN was then utilized to augment training with these synthetic non-contrast images. Finally, they report the benefit of their model by evaluating how well the segmentation performance of a U-Net trained on the original dataset compares with a U-Net trained on a combined dataset of synthetic non-contrast images.

Reference [55] discusses how different CNN models are used to classify abdominal images, and a transfer learning approach is used to increase the accuracy of the results. Two convolutional neural networks based on CaffeNet [56] and VGGNet [57] that had previously been trained on the 2012 Large Scale Visual Recognition Challenge dataset were retrained on the training set using their fully connected layers. Each network's convolutional layers had their weights frozen, so that they could behave as fixed feature extractors.

In the context of the related work described above, our work belongs to the class of post-beamforming techniques for image enhancement. We utilize the Fourier-domain beamforming technique followed by a lightweight upscaling CNN, where most of the data processing occurs in low-resolution space. Such coupling of a Fourier-domain beamforming with a superresolution CNN in order to compensate for the undersampling of the lateral direction is something that has not been studied to the best of our knowledge. Our approach is distinguished by its remarkably high computational efficiency in comparison to similar methods. Furthermore, we not only evaluate multiple configurations of low-cost superresolution CNNs (chapter 3), but also demonstrate

the viability of applying non-GAN data augmentation and transfer learning to help solve our upscaling problem (chapter 4).

Chapter 3

Efficient Ultrasound Image Enhancement Using Lightweight CNNs [1]

In this chapter, we present our approach to ultrasound image reconstruction using a combination of TM migration and various lightweight CNNs derived from ESPCN.

The first task is to obtain a beamformed and compounded IQ data frame, denoted by $D(z, x)$ in the sequel. This is achieved by performing TM migration in four steps [23, 58] applied to 2D raw RF channel data $\widetilde{D}_\alpha(t, x) = U(x, 0, t)^T$ acquired for multiple PW angles α (see chapter 2):

1. Compute the input spectrum $\widetilde{S}_\alpha(\omega, k_x) = \Psi(k_x, 0, \omega)^T$ from $\widetilde{D}_\alpha(t, x)$;
2. Remap $\widetilde{S}_\alpha(\omega, k_x)$ into the migrated spectrum $S_(k_z, k_x) = \Psi(k_x(A, B), 0, \omega(A, B))^T$ multiplied by the scaling factor $\frac{v|(A^2-B^2)\cos\alpha-2AB\sin\alpha|}{2(A\sin\alpha+B\cos\alpha)^2}$;
3. Sum individual $S_\alpha(k_z, k_x)$ over all α angles to obtain the compounded spectrum $S(k_z, k_x)$;
4. Compute the final IQ data frame $D(z, x)$ after taking the inverse Fourier transform of $S(k_z, k_x)$.

Taking the absolute value of $D(z, x)$ gives us the 2D envelope $E(z, x)$. We then perform max-normalization and log-compression on $E(z, x)$ to get the corresponding B-mode image.

If we assume that each $\widetilde{D}_\alpha(t,x)$ has a size of $N_t \times N_x$, then $D(z,x)$ will have the same size of $N_z \times N_x$, where $N_z = N_t$. Because the temporal sampling frequency is usually 4–10 times greater than the transducer central frequency [28], $\widetilde{D}_\alpha(t,x)$ is oversampled along the t-axis. This results in $D(z,x)$ also being oversampled along the z-axis. To put it differently, the spacing between adjacent depth levels (Δz) is already small enough, so there is no need to upscale in this dimension. However, the spacing between adjacent transducer elements (Δx) is often larger than Δz due to spatial undersampling. Therefore $D(z,x)$ or its envelope $E(z,x)$ should be upscaled in the lateral dimension.[1]

In this chapter, we will explore three straightforward approaches to derive an upscaled $N_z \times sN_x$ envelope $E'(z,x_\uparrow)$ from $E(z,x)$, where s denotes the upscaling factor such that $\Delta x_\uparrow = \Delta x/s$:

1. Apply bicubic interpolation (BI) to $D_\alpha(z,x)$, thus producing upscaled $D'(z,x_\uparrow)$ that gives $E' = |D'|$;

2. Apply Fourier interpolation (FI) to $S(k_z,k_x)$, thus producing upscaled $D'(z,x_\uparrow)$ that gives $E' = |D'|$;

3. Use a superresolution CNN to produce $E'(z,x_\uparrow)$ directly from $E(z,x)$.

For the last option, we propose a set of lightweight CNN models based on ESPCN architecture [9](see chapter 2). These CNN models outperform both bicubic and frequency interpolation methods, as well as a conventional DAS beamformer that operates on the same $N_z \times sN_x$ grid. Our CNN models not only upscale ultrasound envelope data, but also enhance the resulting image quality. We have examined multiple CNN configurations that have low parameter counts ranging from 32K down to 5K.

3.1 ESPCN Modifications

As discussed earlier, our objective is to upscale the envelope in the lateral dimension. Given the beamformed envelope $E(z,x)$ of size $N_z \times N_x$, we want the laterally upscaled envelope $E'(z,x_\uparrow)$ of size $N_z \times sN_x$. In other words, our goal is to reduce the lateral grid spacing by s. In this work, s is equal to 3.

[1] In contrast, DAS beamforming can calculate the 2D IQ data for any (z,x) location on a grid, with the option to specify the desired lateral or depth spacing.

3.1.1 Two-Dimensional Upscaling

Since ESPCN upscales images in both H (axial) and W (lateral) dimensions, we first downsample envelope $E(z, x)$ by s along the z-axis. Then, the resulting envelope is fed into the neural network as a gray-scale input I_{LR} with a single channel. It produces the corresponding high-resolution output image $I_{HR} = E'$. Another option is to downsample output images along the z-axis after upscaling, but in this case, the number of computations increases significantly because the CNN would operate on bigger input images. Our first proposed network has the following configuration: $L = 4$, $(n_1, k_1) = (64, 5)$, $(n_2, k_2) = (32, 3)$, $(n_3, k_3) = (32, 3)$, $(n_4, k_4) = (9, 3)$, with ReLU as ϕ.

This is our baseline model with $31,977 \approx 32K$ parameters and is referred to as model N32. Its computational footprint is $2,086M$ MACs, which corresponds to processing the inputs of size 512×128 to produce the output of size 1536×384.

As discussed in the previous chapter, one of the most effective approaches to reduce the CNNs' complexity is replacing the standard 2D convolution layers with separable ones [32]. We take our initial network configuration N32 and replace all standard convolutions with separable ones. The resulting model has $4,738 \approx 5K$ parameters and is denoted by N5. Given the critical importance of the last layer, we also consider keeping its convolution operations standard, which yields the model N7 with $6,754 \approx 7K$ parameters. As expected, introducing separable convolutions at a given layer l significantly impacts both parameter and MAC counts. It can be shown (based on the analysis presented in [32]) that the latter is reduced by the factor of $1/(1/n_l + 1/k_l^2)$. In our case, the overall computational footprints of N5 and N7 are $301M$ and $434M$ MACs, respectively.

3.1.2 One-Dimensional Upscaling

In the context of envelope upscaling, all of the SISR networks would upscale images in both dimensions. In other words, 2D upscaling is applied to axial and lateral dimensions, even though we only need to upscale along the x-axis. To perform strictly 1D upscaling, we propose the CNN structure similar to N5, with the following two adjustments illustrated in Fig. 3.1:

1. The final layer has $n_4 = s$ (as opposed to $n_4 = s^2$ in the 2D upscaling case), where $s = 3$. This layer is followed by a 1D subpixel shuffling that reshapes the $H \times W \times sC$ input into the $H \times sW \times C$ output, where $C = 1$.

2. The second and third layers are combined with the Squeeze and Excitation blocks that help capture inter-channel dependencies [8]. As mentioned earlier, such blocks allow for adaptive recalibration of features, often leading to better performance.

This new network, referred to as model N5* has $4,604 \approx 5K$ parameters. Its computational footprint is 829M MACs, which corresponds to 1D upscaling of a 1536×128 input into a 1536×384 output. It should be noted that N5* has a larger input compared to N5 that performs 2D upscaling of a 512×128 input into a 1536×384 output. This explains why N5 requires fewer computations than N5*, i.e., 2D upscaling turns out to be cheaper than 1D upscaling. However, the latter can still help enhance image quality.

Putting it all together, Table 3.1 summarizes the number of parameters of the proposed networks and their computational costs in terms of MACs. In the next section, we discuss our training approach.

Table 3.1: Complexity of Proposed ESPCN Configurations.

Model	Parameters	MACs(M)	Upscaling
N32	$31,977$	$2,086$	$2D$
N7	$6,754$	434	$2D$
N5	$4,738$	301	$2D$
N5*	$4,604$	829	$1D$

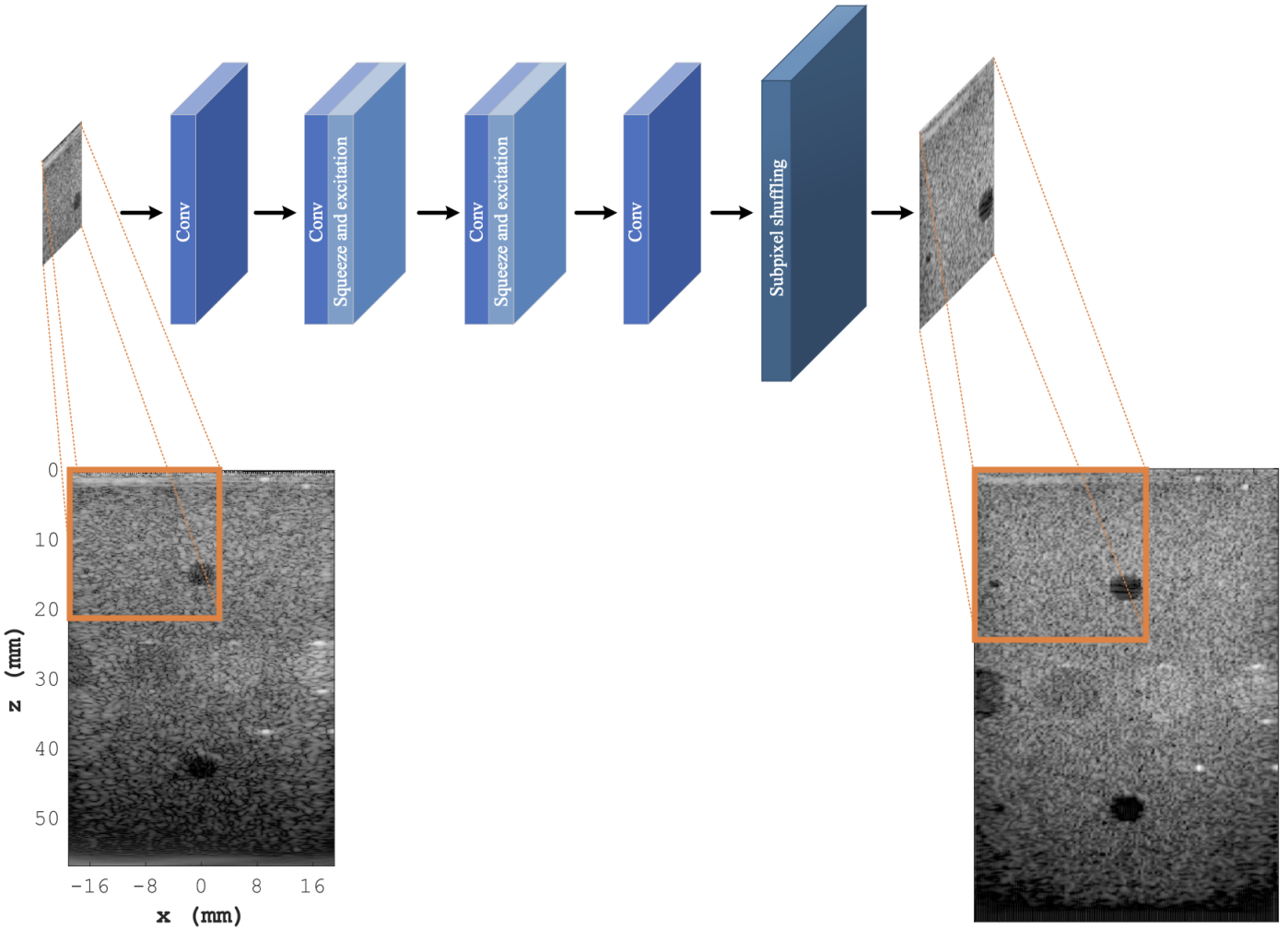


Figure 3.1: Proposed 1D-upscaling network with Squeeze and Excitation blocks.

3.2 Training

For training, we used a subset of 400 grayscale-converted images from the DIV2K dataset [59]. We used the Adam optimizer [60] with a learning rate of 10^{-3}, a loss function of mean squared logarithmic error (MSLE), and a batch size of four. The MSLE choice is essential because CNN-upscaled envelopes will subsequently undergo log-compression to form B-mode ultrasound images.

As the DIV2K dataset does not have any ultrasound-domain images, we set an early-stopping threshold for the loss function to avoid overfitting. Specifically, the network weights were taken from a training epoch at which the validation MSLE value dropped below 0.061. Also, we used the Glorot uniform initializer [61] for the first three layers and the orthogonal initializer [62] for the last layer.

3.3 Experimental Results

Our experimental results are based on three different datasets provided by PICMUS [63, 17] which also includes a library of MATLAB scripts for image quality measurements and baseline DAS beamforming. We have a collection of 75 raw data frames $\widetilde{D}_\alpha(t, x)$ acquired over 75 PW emissions at different angles α within the interval $[-16°, +16°]$.

The aforementioned datasets denoted by TYPE-1, TYPE-2, and TYPE-3 are derived from three imaged regions of an ultrasound phantom that mimics tissue. A DAS beamformer provided by PICMUS yields their respective compounded B-mode images displayed in Fig. 3.2. Tukey apodization window with a 1/4 cosine fraction has been used during DAS beamforming. The following quality indicators are measured in the evaluation regions highlighted in Fig. 3.2.

- TYPE-1: Contrast-to-noise (CNR) ratios, associated with the top and bottom cysts-like targets;
- TYPE-2: Mean full-width at half-maximum ($\overline{\text{FWHM}}$), averaged over seven wire targets and measured both axially (along the z-axis) and laterally (along the x-axis);
- TYPE-3: Minimum axial and lateral distances (MINIMUM) between vertically and horizontally adjacent pairs of wire targets that are still distinguishable via measurements. As shown in Fig. 3.2, these wires are placed in a decreasing axial (i.e., vertically towards the region's bottom left corner) and laterally (i.e., horizontally

towards the region's bottom right corner) distance. A smaller measured distance means that a closer pair of adjacent wire targets could be distinguished during evaluation.

All acquired raw data frames $\widetilde{D}_\alpha(t,x)$ contain 1536 rows and 128 columns; therefore, after beamforming and compounding, we get a 1536×128 envelope $E(z,x)$ for each image type. As mentioned earlier, we are using $s = 3$, which yields upscaled envelopes $E'(z,x_\uparrow)$ of size 1536×384.

All of the metrics we used to evaluate the quality of our images are listed in Tables 3.2 and 3.3. Our discussion will concentrate mostly on the CNR and lateral $\overline{\text{FWHM}}$ values.

Table 3.2 covers the scenario of a single PW emission at $\alpha = 0°$, which corresponds to the least amount of raw data acquisition with no compounding. The best $\overline{\text{FWHM}}_{\text{Lateral}} = 0.64$ mm is given by TM+N5*. This value is 18% better than that of TM+BI (i.e., without using a CNN), and it is also only 3% worse than $\overline{\text{FWHM}}_{\text{Lateral}} = 0.61$ mm given by conventional DAS beamforming in the case of 75 PW emissions (see Table 3.3). It should be noted that the latter costs substantially more to acquire data and process data. The best $\text{CNR}_{\text{Top}} = 10.3$ dB is due to TM+BI+N7, where the final upscaled envelope is obtained by averaging the envelopes computed by both BI and N7 (used as an illustrative example). The best $\text{CNR}_{\text{Bottom}} = 9.6$ dB is given by TM+N7.

Table 3.3 presents the case of 75 multi-angle PW emissions corresponding to better-quality compounded images. The best $\overline{\text{FWHM}}_{\text{Lateral}} = 0.35$ mm is given by TM+N5*, which is 32% and 43% better than those given by TM+BI and DAS, respectively. The best $\text{CNR}_{\text{Top}} = 14.2$ dB and $\text{CNR}_{\text{Bottom}} = 13.6$ dB are due to TM+BI+N7 and TM+BI+N5*, respectively. These CNR values are better than the ones produced by TM+BI and DAS by at least 1 dB. This example illustrates that combining the CNN and BI outputs can lead to CNR improvements at the expense of worsened FWHM values compared to those given by the network output alone. Still, TM+BI+N5* yields $\overline{\text{FWHM}}_{\text{Lateral}}$ that is 18% better in comparison to TM+BI. It is also important to mention that only TM migration coupled with a CNN has managed to attain sub-millimeter $\text{MinD}_{\text{Lateral}}$, in contrast to DAS, TM+BI, and TM+FI. In other words, our proposed approach allows one to resolve a closer-spaced pair of laterally adjacent wire targets. The best $\text{MinD}_{\text{Lateral}} = 0.66$ mm is given by TM+N5*. The ground-truth distance between the corresponding wire targets is 0.5 mm [64].

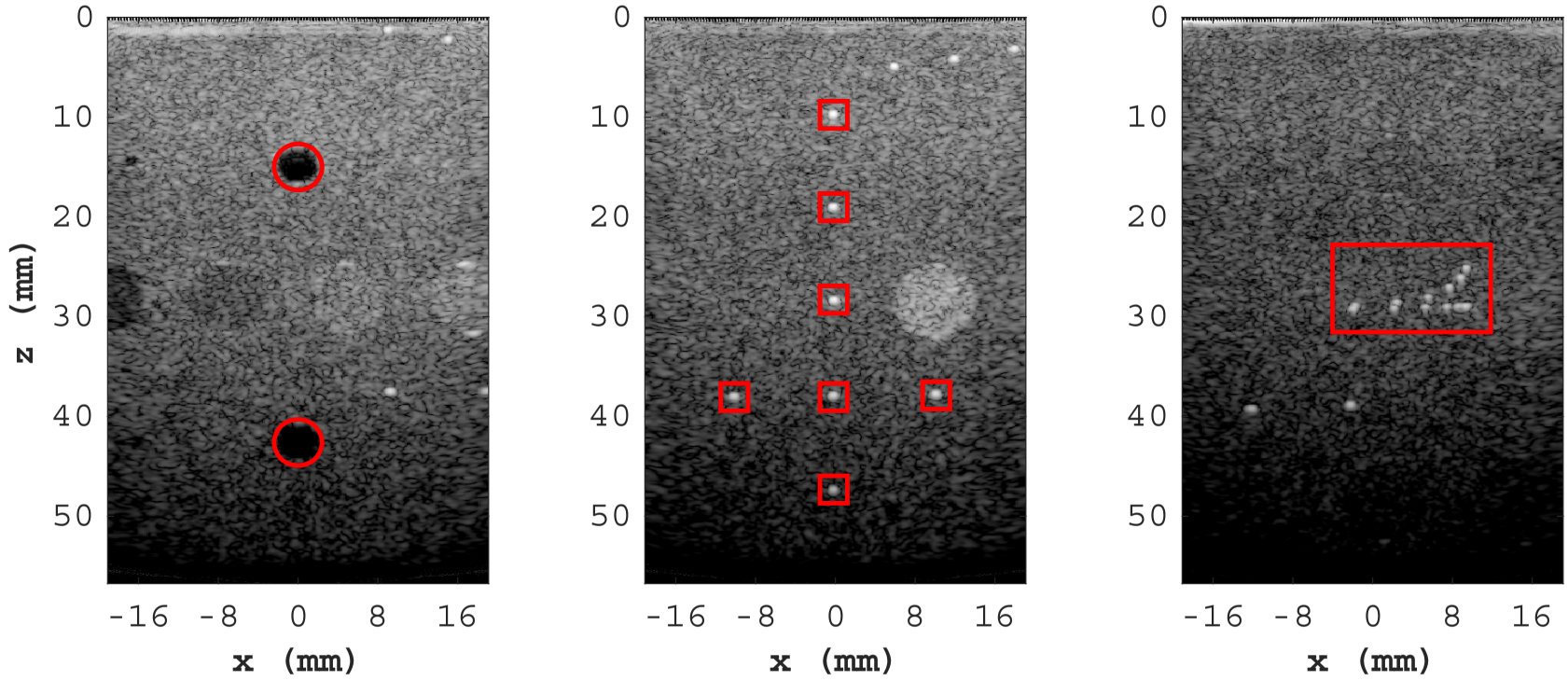


Figure 3.2: Baseline DAS beamforming, 75-PW compounded B-mode images (60-dB range) and evaluated regions: TYPE-1 (left), TYPE-2 (center), TYPE-3 (right).

Table 3.2: Image Quality Assessment: Single Plane Wave, $\alpha = 0°$.

Beam-forming	Top/Bottom CNR (dB)	Axial/Lateral $\overline{\text{FWHM}}$ (mm)	Axial/Lateral MinD (mm)
DAS	8.4/7.1	0.49/0.96	0.64/**1.62**
TM+BI	8.7/7.8	0.49/0.77	0.62/1.80
TM+FI	8.3/7.4	0.49/0.82	0.62/1.78
TM+N32	8.5/7.2	**0.42**/0.66	0.63/1.79
TM+N7	9.7/**9.6**	0.44/0.69	**0.58**/1.82
TM+N5	8.8/7.9	0.46/0.67	0.65/1.69
TM+N5*	8.9/8.1	0.50/**0.64**	0.63/1.77
TM+BI+N7	**10.3**/9.4	0.46/0.74	0.60/1.80
TM+BI+N5*	9.8/8.8	0.49/0.72	0.60/1.80

Table 3.3: Image Quality Assessment: 75 Plane Waves, $-16° \leq \alpha \leq +16°$.

Beam-forming	Top/Bottom CNR (dB)	Axial/Lateral $\overline{\text{FWHM}}$ (mm)	Axial/Lateral MinD (mm)
DAS	13.2/12.2	0.49/0.61	0.62/1.11
TM+BI	12.5/12.2	0.49/0.52	0.57/1.17
TM+FI	12.4/12.1	0.49/0.52	0.62/1.07
TM+N32	12.3/10.8	**0.42**/0.38	0.62/0.74
TM+N7	14.0/9.8	0.45/0.41	0.58/0.73
TM+N5	13.3/9.3	0.47/0.40	0.66/0.82
TM+N5*	12.8/10.5	0.50/**0.35**	**0.57**/**0.66**
TM+BI+N7	**14.2**/13.5	0.47/0.45	0.61/0.71
TM+BI+N5*	13.4/**13.6**	0.50/0.43	0.57/0.67

Fig. 3.3, 3.4, and 3.5 shows illustrative examples of the Type-1 and Type-2 75-PW compounded B-mode images produced by TM+BI, TM+N5*, and TM+BI+N7, respectively. Note that the CNR differences from Table 3.3 do not necessarily translate into clear visual differences in the Type-1 images, suggesting that this metric should be treated with caution. On the other hand, superior $\overline{\text{FWHM}}_{\text{Lateral}}$ given by TM+N5* is indicative of the sharper appearance of the wire targets in the Type-2 images. Since it is difficult to see closely spaced wire targets in the Type-3 B-mode images, Fig. 3.6 instead shows the contours of the upscaled and max-normalized envelopes in question. For the two wire targets located near the bottom right corner (laterally spaced by 0.5 mm [64]), using TM+N5* results in relatively clearer separation compared to TM+BI and TM+BI+N7.

(a) TYPE-1, TM+BI (b) TYPE-2, TM+BI

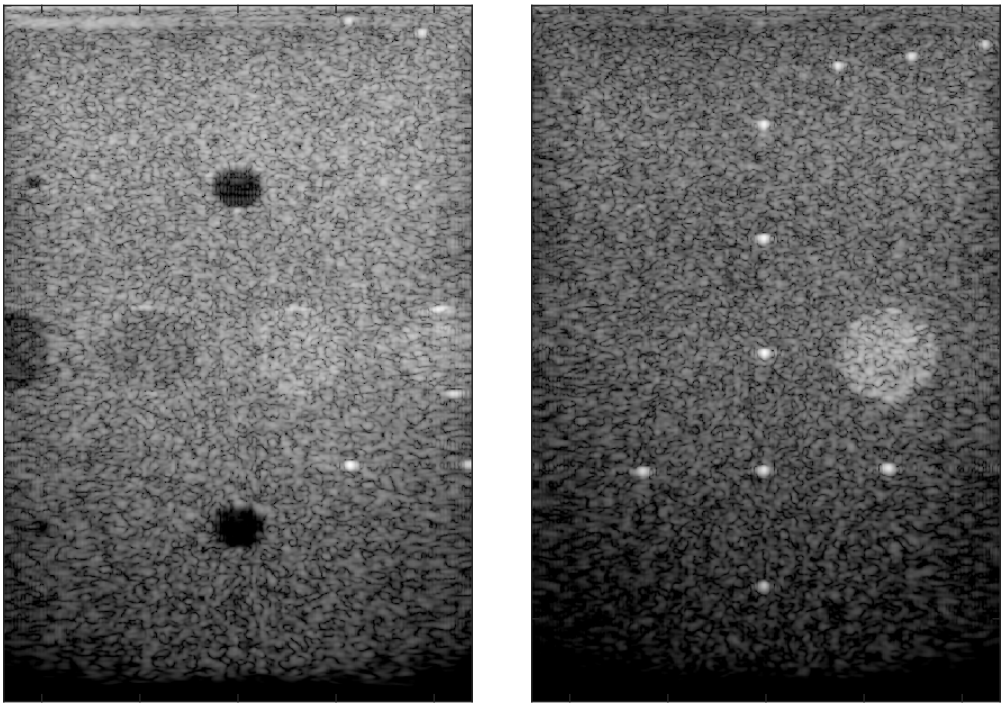


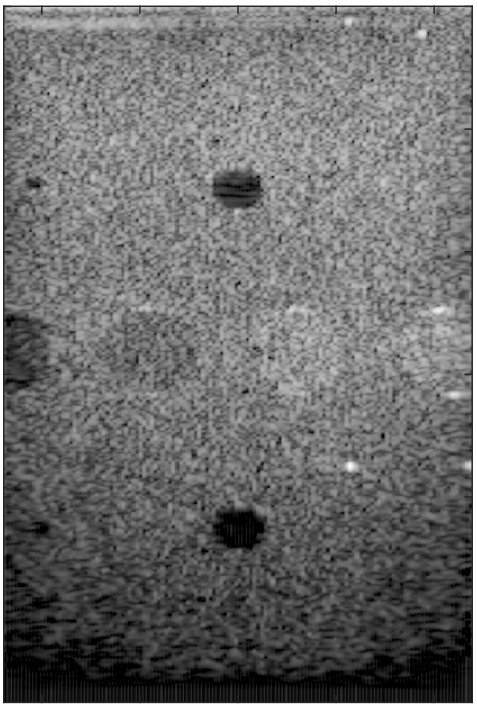
Figure 3.3: Examples of TYPE-1 (left) and TYPE-2 (right) 75-PW compounded B-mode images: TI+BI.

One can argue that the overall winner is TM+BI+N7 because it consistently outperforms DAS, as well as non-CNN methods TM+BI and TM+FI, in terms of the CNR and FWHM values shown in Tables 3.2 and 3.3. While the other methods that utilize a lightweight CNN (most notably TM+N5*) may offer even better $\overline{\text{FWHM}}_{\text{Lateral}}$ and $\text{MIND}_{\text{Lateral}}$, they often result in lower CNR_{Top} and $\text{CNR}_{\text{Bottom}}$.

It should be noted that all of the reported TM-based methods are much faster than DAS beamforming. For example, the latter takes approximately 2.10 s in the 1-PW case and 154.0 s in the 75-PW case (2.6-GHz i7 CPU, 16-GB RAM, Windows 10, MATLAB 2021b), whereas TM+BI+N7 takes only 0.58 s and 5.6 s, respectively.

In this chapter, we have not used any ultrasound images during training and validation, which makes it challenging to identify an appropriate epoch for model weight selection. This motivates our work presented in the next chapter.

(a) Type-1, TM+N5*



(b) Type-2, TM+N5*

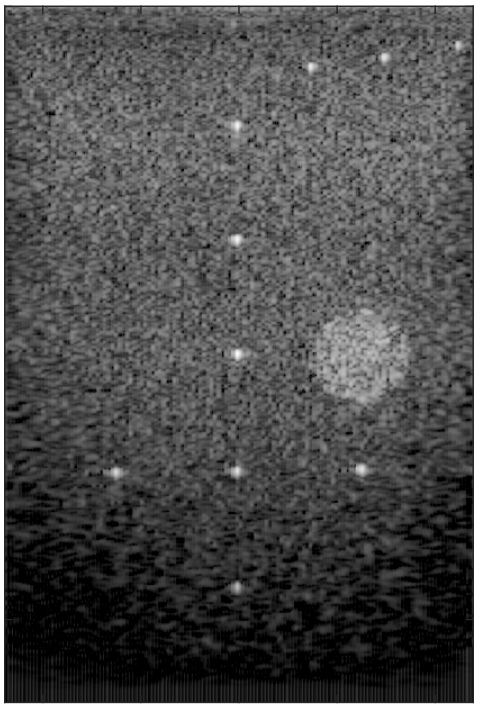


Figure 3.4: Examples of Type-1 (left) and Type-2 (right) 75-PW compounded B-mode images: TM+N5*.

(a) Type-1, TM+BI+N7 (b) Type-2, TM+BI+N7

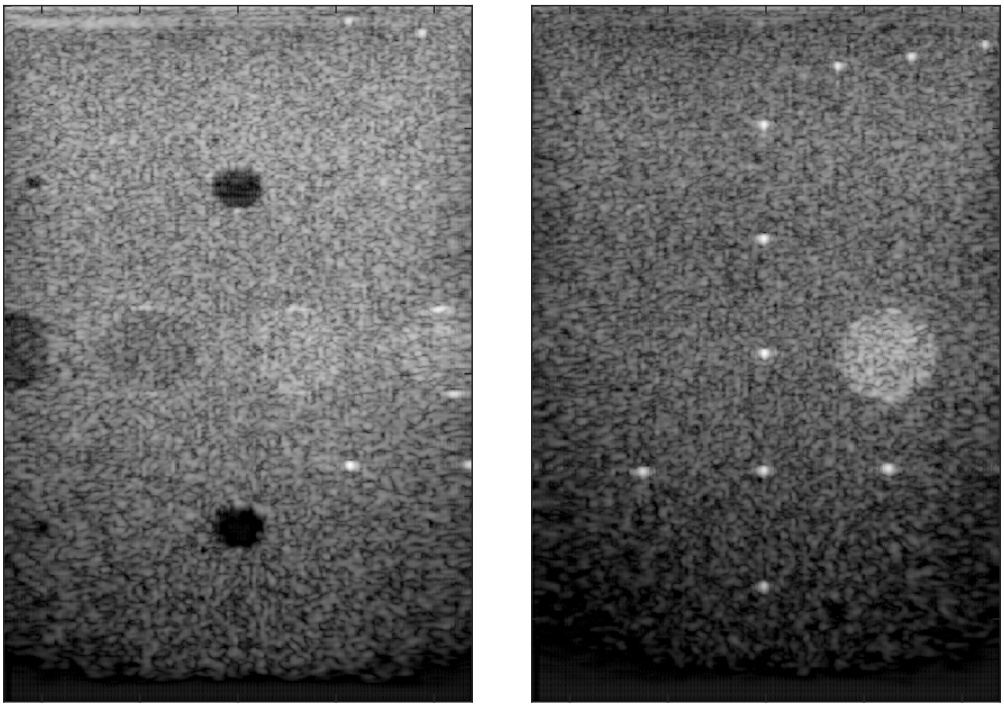


Figure 3.5: Examples of Type-1 (left) and Type-2 (right) 75-PW compounded B-mode images: TI+BI (top), TM+N5* (center), TM+BI+N7 (bottom).

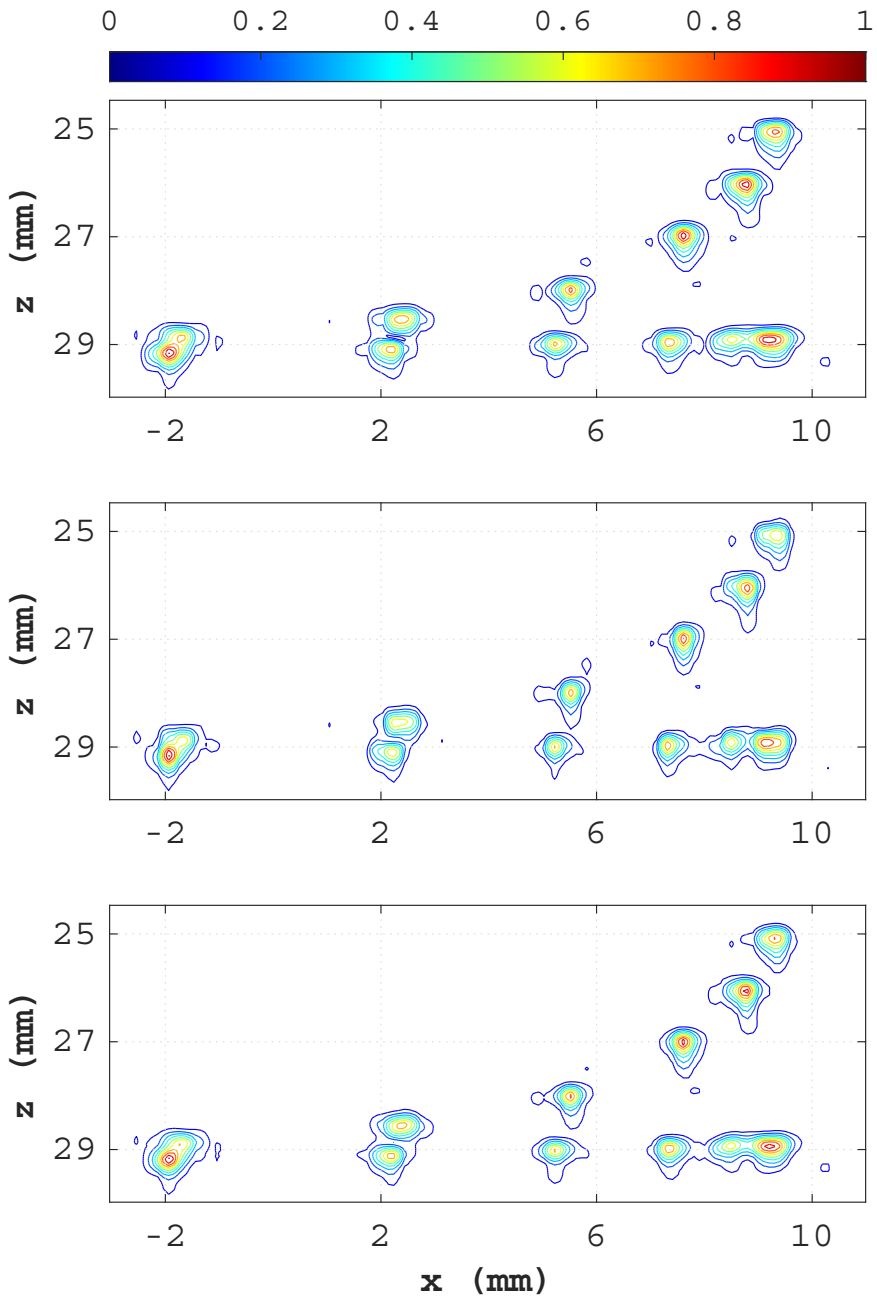


Figure 3.6: Examples of Type-3 75-PW compounded envelope contours: TM+BI+N7.

Chapter 4

Data Augmentation and Transfer Learning in Ultrasound Image Enhancement [2]

As the ultrasound domain lacks access to big data, we utilize data augmentation techniques to increase the size of our dataset for training purposes. Data augmentation encompasses a suite of techniques that enhance the size and quality of training datasets [65]. In this chapter, we present our approach to transfer learning on augmented ultrasound data.

4.1 Data Augmentation for Ultrasound Imaging

Deep learning models have made great progress in many tasks due to network architecture advancements, the power of computational processing units, and access to big data. Most deep learning models are data-hungry, but in many situations, big data is not accessible because it involves a process that is expensive or time-consuming [66]. When we do not have access to big data, a network model may suffer from poor generalizability. Generalizability refers to the performance difference of a model when evaluated on previously seen data (training data) versus data it has never seen before (testing data) [65]. When the generalization performance is not good enough, overfitting to the training data has occurred. Data augmentation is a powerful technique to produce more data in order to avoid overfitting. Ultrasound imaging is one of the domains in which we do not have access to big data. Therefore, we take advantage

of data augmentation to increase the size of the dataset.

Transfer learning is another interesting paradigm to prevent overfitting when we do not have access to big data [67]. The idea behind transfer learning is to train the network in a general domain with a big dataset and then use those trained weights as initial weights in the target-domain training process in which the dataset is limited.

In this work, we use data augmentation to generate additional ultrasound-domain data, which is then utilized for transfer learning to improve our network output based on the discussed metrics.

4.1.1 Data Augmentation Methods

There are different methods for data augmentation. In this work, we were inspired by [68] and [69] and implemented our own random augmentation applied to ultrasound images. We did not reuse the original code from [68] because it was not designed for ultrasound images, and it did not work well for superresolution of images that have only one channel in grayscale color space. In what follows, we briefly summarize various image transformations in use:

- **Flipping**
  We have utilized flipping with respect to both horizontal and vertical axes. Such augmentation is very easy to implement and is also very useful on superresolution ultrasound images.

- **Color space**
  This transformation can be done in different ways, but all of them operate in the color space. For example, one can perform inversion of the intensity of pixels. Another example is to adjust image brightness randomly. A more advanced approach is to work on the histograms of the images, such as histogram equalization, which tends to change the image contrast. All of the methods mentioned have been used in our implementation.

- **Cropping**
  We randomly crop a central patch of the image and then resize the cropped image to its original size.

- **Shear X**
  In shearing, the image pixels are randomly shifted along the horizontal (X) axis

so that the image shape changes from a rectangle to a parallelogram. In our case, however, we pad the image to have a rectangular shape at the end.

- **Rotation**
  We rotate an image by a random angle within the range from $-20°$ to $+20°$, which typically gives better results [65]. We pad the rotated image so that it has the same size as the original one.

- **Translation**
  Translation shifts the image in one of four directions: left, right, up, or down. This can be very useful for ultrasound superresolution tasks.

- **Patch swapping**
  Inspired by [69] we randomly select a box in the original image and swap that patch with a patch from another available image. The position of the boxes in both images is exactly the same.

- **Noise injection**
  The noise injection process entails adding a matrix of random values, often selected using a Gaussian distribution.

4.1.2 Applied Data Augmentation

Using two out of three PICMUS raw datasets (see chapter 3), we first obtain low-resolution envelopes of size $N_z/s \times N_x = 512 \times 128$ (where $N_z = N_t$) by Fourier beamforming for angle $\alpha = 0°$ and then downsampling in the z-axis direction by the factor of $s = 3$. We utilized these two low-resolution (LR) envelopes as our input images for data augmentations. We produced the target high-resolution (HR) envelopes by coherently compounding all 75 DAS-beamformed data frames of size $N_z \times sN_x = 1536 \times 384$ corresponding to $N_\alpha = 75$ emission angles. Then, we sliced the LR envelopes horizontally into four patches of size of 128×128, and similarly, we sliced the HR envelopes into four 384×384 patches. Thus, we acquired a total of 8 input/output pairs. Inspired by [68] and [69], we augmented the resulting 8 patch pairs to obtain 160 pairs of images as follows.

Let T represent a set of image transformations that will be used throughout the augmentation process. Both horizontal and vertical flips are always included in T.

We add a random number of transformations between 2 and 5 from our list of transformations. Therefore, we have a random array of four to seven transformations. Next, we shuffle this array to randomize the sequence of augmentations. We apply this array to a pair of LR/HR patches. The process is repeated 20 times per pair, which produces a total of 160 input/output training samples.

As an illustrative example, let's assume that the random number between 2 and 5 is 4. This means we add 4 random transformations to vertical and horizontal flipping. One possible set T in this case is shown below:
(cropping, shear X, rotation, patch swapping, vertical flipping, horizontal flipping).
Next, we shuffle the elements of T to create a randomized sequence of the corresponding transformations, such as:
(horizontal flipping, patch swapping, rotation, vertical flipping, cropping, shear X).

After applying this sequence of transformations to a pair of patches, we use the PSNR threshold of 20 to either accept or reject such augmentation. If rejected, the process will be repeated for the pair in question.

Finally, we reload the pre-trained model to train it with the new dataset. We keep the original model without changing its structure and adjust the learning rate to a smaller value. We discuss the details of training in the next section.

4.2 Transfer Learning for Ultrasound Imaging

In transfer learning, we aim to improve the performance of the target domain by transferring the knowledge we have from a different but related source domain. Researchers utilize transfer learning when they do not have access to big data in the target domain. Transfer learning involves three main steps outlined below.

1. **Source-domain training**
   Firstly, we train the model in the source domain, which is related to the target domain.

2. **Model and Parameters Modification**
   After we have the pre-trained model on the source dataset, we can modify the model or the parameters to prepare it for the training on the target domain.

3. **Target-domain training**
   We train the model on the target dataset. The weights from the source-domain

training are used as the initial weights for the target-domain training.

There are different types of transfer learning [70]; however, in this work, we focus on inductive transfer learning. The source and target domains in inductive transfer learning are the same, but the source and target tasks are distinct from one another.

In this work, we utilize transfer learning for the task of single-image superresolution. At first, we train the model on DIV2K [59], which is a superresolution dataset but completely distinct from the ultrasound imaging domain. In the second phase of training, we load the DIV2K-trained model and train it further in the ultrasound imaging domain using two out of three PICMUS datasets.

4.2.1 General-Domain Training

Numerous image datasets, including DIV2K [59, 71], have been used to support and expand the research on the application of deep learning to single-image superresolution tasks [72, 73]. DIV2K consists of 1,000 high-resolution RGB images split into training, validation, and testing subsets having 800, 100, and 100 images, respectively. We converted all of the images to grayscale color space prior to training because our target data is in grayscale format. We chose the option of bicubic downsampling by a factor of $s = 3$ to produce the low-resolution images. We also used the Gaussian filter in order to prevent aliasing artifacts due to downsampling.

During our general-domain training on the DIV2K dataset, we used a batch size of four and employed the Adam optimizer [60], with the MSLE (mean squared logarithmic error) loss function. The rationale for selecting the MSLE criterion is that the CNN-predicted output (i.e., upscaled envelope data) is subsequently log-compressed to obtain B-mode ultrasound images. The learning rate was set to 10^{-3}, and the chosen model was `N7` introduced in the previous chapter. We used the Glorot uniform initializer [61] for the first three feature-extracting layers (with separable convolutions) and an orthogonal matrix initializer [62] for the final subpixel layer (with standard convolutions).

4.2.2 Target-Domain Training

The DIV2K dataset does not have any ultrasound images, which further motivates subsequent transfer learning to adjust pre-trained CNN weights. In our case, we

need raw channel data recorded by an ultrasound transducer array to perform receive beamforming and generate both low- and high-resolution envelope data (i.e., upscaling input and output samples). One source of such raw ultrasound data is the well-known PICMUS evaluation framework [17]. It includes three experimental 3D data volumes, referred to as TYPE-1, TYPE-2, and TYPE-3. We applied data augmentation to TYPE-2 and TYPE-3 beamformed envelopes to create a small dataset of 160 images (see section 4.1.2) for ultrasound-domain training.

After reloading the model from the general-domain phase of training, we used the same settings as in the DIV2K case (except for the learning rate) and trained the model with an ultrasound-specific dataset. The learning rate was set to 10^{-4} for the first 15 epochs, and then scheduled to be reduced by the factor of $e^{-0.1}$ every two epochs.

4.3 Experimental Results

In this section, we present our evaluation result based on the PICMUS TYPE-1 experimental data, as TYPE-2 and TYPE-3 have already been used for target-domain training. Fig. 4.1 illustrates a DAS-beamformed TYPE-1 B-mode image with its three evaluated areas denoted as A, B, and C. For the anechoic cyst-like targets A and B, we report measured CNR values in dB. For the hyperechoic point-like target C, we report measured axial and lateral FWHM values in mm. These image quality indicators have been computed by the PICMUS software [17] and are listed in Tables 4.1 and 4.2.

As our baseline (DAS), we have used a PICMUS-provided reference implementation of DAS beamforming and considered four choices of receive apodization windows: none, boxcar, Hamming, and Tukey-25% (i.e., 1/4 cosine fraction). We evaluated three potential solutions for Temme-Mueller (TM) migration combined with regular interpolation (labeled TM+I): linear, bicubic, and Fourier [74]. We investigated two options for TM migration coupled with our upscaling CNN (labeled TM+N7): using the model weights acquired immediately after initial training (DIV2K) and then taken after transfer learning (DIV2K+TL).

Our numerical assessment results are summarized in Table 4.1 for the case of a single PW emission, which corresponds to the least expensive data acquisition. It is clear that the generated CNR and FWHM values for the DAS are greatly affected by the apodization window selection. On the other hand, various TM+I interpolation

settings result in rather consistent values. The best $\mathrm{CNR}_A = 11.8$ dB is obtained using DAS beamforming with Hamming apodization; however, it yields the worst lateral $\mathrm{FWHM}_C = 1.32$ mm. Coupling TM migration with $\mathsf{N7}_{\mathrm{DIV2K+TL}}$ produces the second largest CNR_A, as well as the best $\mathrm{CNR}_B = 11.7$ dB and the best lateral $\mathrm{FWHM}_C = 0.61$ mm. The best axial $\mathrm{FWHM}_C = 0.43$ mm is given by $\mathsf{TM}{+}\mathsf{N7}_{\mathrm{DIV2K}}$.

Table 4.2 presents our results for the coherent compounding case of $N_\alpha = 75$ PW emissions. The best $\mathrm{CNR}_A = 15.1$ dB and the best lateral $\mathrm{FWHM}_C = 0.41$ mm are due to $\mathsf{TM}{+}\mathsf{N7}_{\mathrm{DIV2K+TL}}$, while the best $\mathrm{FWHM}_C = 0.45$ mm is again given by $\mathsf{TM}{+}\mathsf{N7}_{\mathrm{DIV2K}}$. DAS beamforming with boxcar apodization yields the best $\mathrm{CNR}_B = 12.5$ dB and the second largest CNR_A.

Surprisingly, the output of $\mathsf{TM}{+}\mathsf{N7}_{\mathrm{DIV2K}}$ has undesirable $\mathrm{CNR}_B = 6.2$ dB in Table 4.2, which is worse than 8.2 dB in the case of a single PW emission. It is also surprising that, unlike $\mathsf{TM}{+}\mathsf{I}$ or DAS, $\mathsf{TM}{+}\mathsf{N7}_{\mathrm{DIV2K+TL}}$ has failed to noticeably increase CNR_B with the use of coherent compounding (11.8 versus 11.7 dB in Table 4.1). These observations suggest that further investigations are needed to obtain a better understanding of CNN behaviour.

DAS-beamformed B-mode images for the case of a single PW emission at $\alpha = 0°$ are shown in Fig. 4.2. The non-interpolated data of TM migration can be seen in Fig. 4.3, where undesirable pixelation is visible. Fig. 4.3 also shows the outputs of our image reconstruction scenarios: $\mathsf{TM}{+}\mathsf{I}_{\mathrm{Bucibic}}$, $\mathsf{TM}{+}\mathsf{N7}_{\mathrm{DIV2K}}$, and $\mathsf{TM}{+}\mathsf{N7}_{\mathrm{DIV2K+TL}}$. Since $\mathsf{N7}$ is a superresolution CNN, it naturally improves FWHM values (in comparison to regular interpolation). Additionally, it has successfully enhanced CNR values, particularly after transfer learning. The latter appears to introduce some despeckling effect into $\mathsf{N7}$ operation. However, in diagnostic cases that involve speckle tracking, such an effect is not desirable.

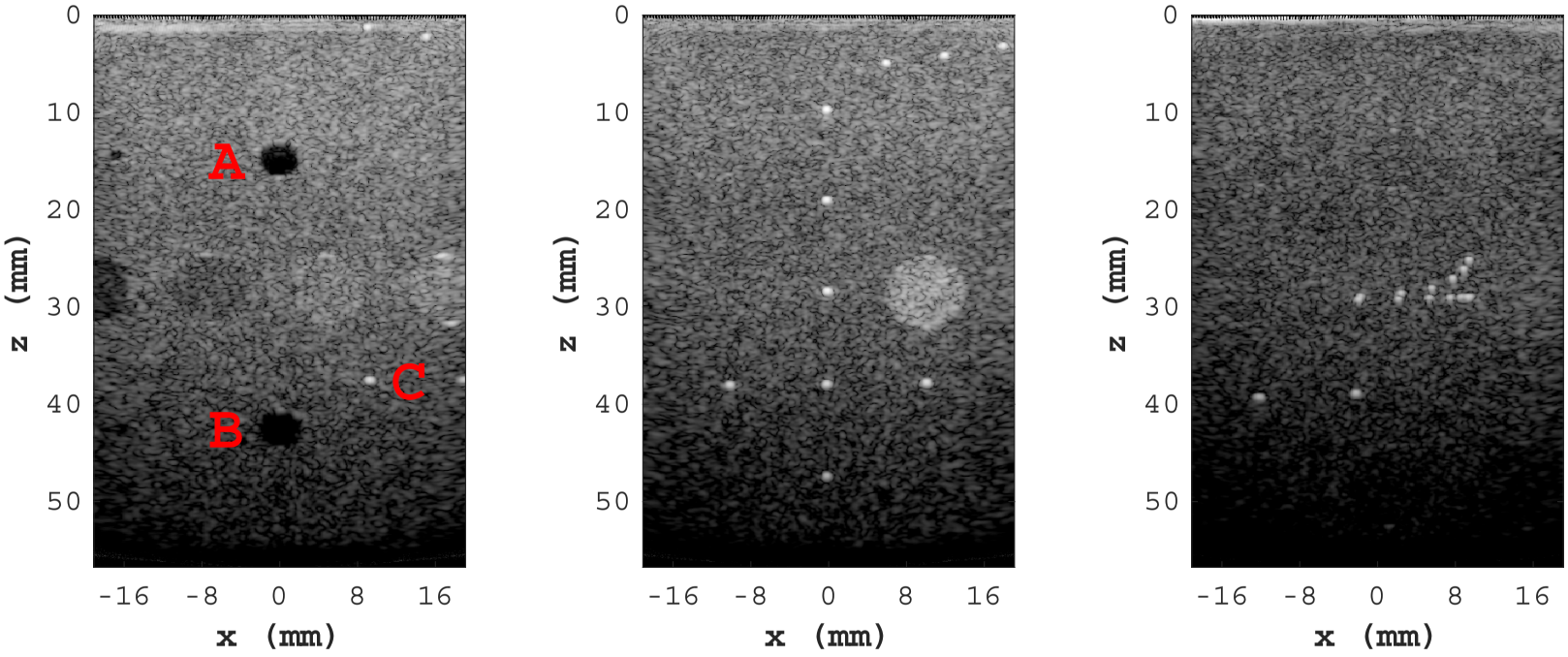


Figure 4.1: Conventional DAS beamforming (75 plane waves, Tukey-25% window): TYPE-1 (left), TYPE-2 (center), and TYPE-3 (right) compounded B-mode images, 60-dB range.

Table 4.1: Image Quality Assessment: Single Plane Wave, $\alpha = 0°$.

Beamforming Method		CNR, dB		FWHM, mm	
		A	B	C Axial	C Lateral
DAS	—	3.8	7.1	0.48	0.84
	Boxcar	7.6	6.1	0.49	0.90
	Hamming	**11.8**	8.4	0.48	1.32
	Tukey-25%	8.4	7.1	0.49	0.97
TM+I	Linear	8.8	7.9	0.49	0.87
	Bicubic	8.7	7.8	0.49	0.84
	Fourier	8.3	7.4	0.48	0.85
TM+N7	DIV2K	10.1	8.2	**0.43**	0.68
	DIV2K+TL	11.3	**11.7**	0.44	**0.61**

Table 4.2: Image Quality Assessment: 75 Plane Waves, $-16° \leq \alpha \leq +16°$.

Beamforming Method		CNR, dB		FWHM, mm	
		A	B	C Axial	C Lateral
DAS	—	12.9	**12.5**	0.49	0.58
	Boxcar	13.5	**12.5**	0.49	0.60
	Hamming	12.5	12.4	0.49	0.73
	Tukey-25%	13.2	12.2	0.49	0.63
TM+I	Linear	12.3	12.1	0.49	0.59
	Bicubic	12.5	12.2	0.49	0.54
	Fourier	12.4	12.1	0.49	0.55
TM+N7	DIV2K	12.7	6.2	**0.45**	0.42
	DIV2K+TL	**15.1**	11.8	0.46	**0.41**

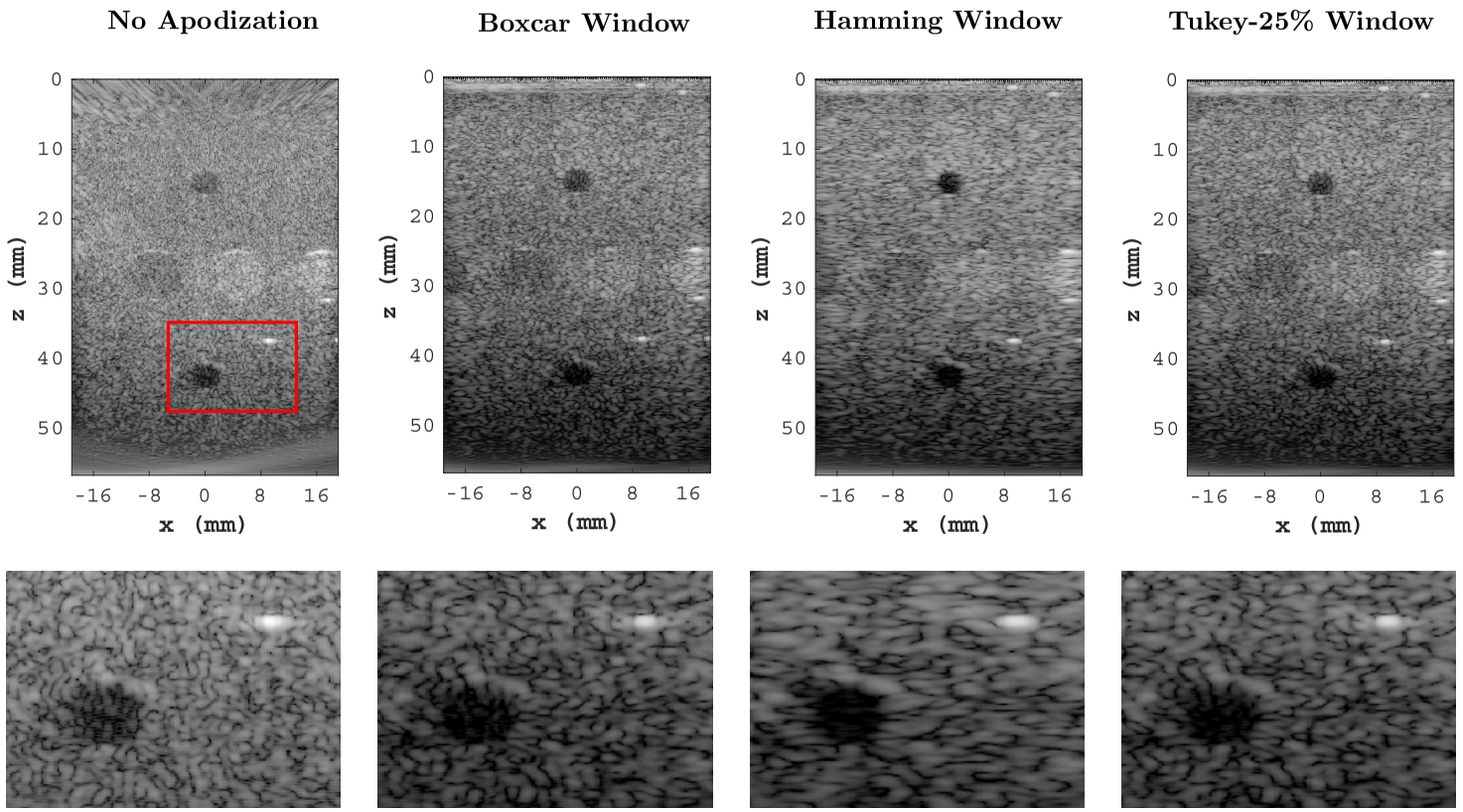


Figure 4.2: Conventional DAS beamforming (zero-angle PW) with various apodizations: TYPE-1 B-mode images, 60-dB range.

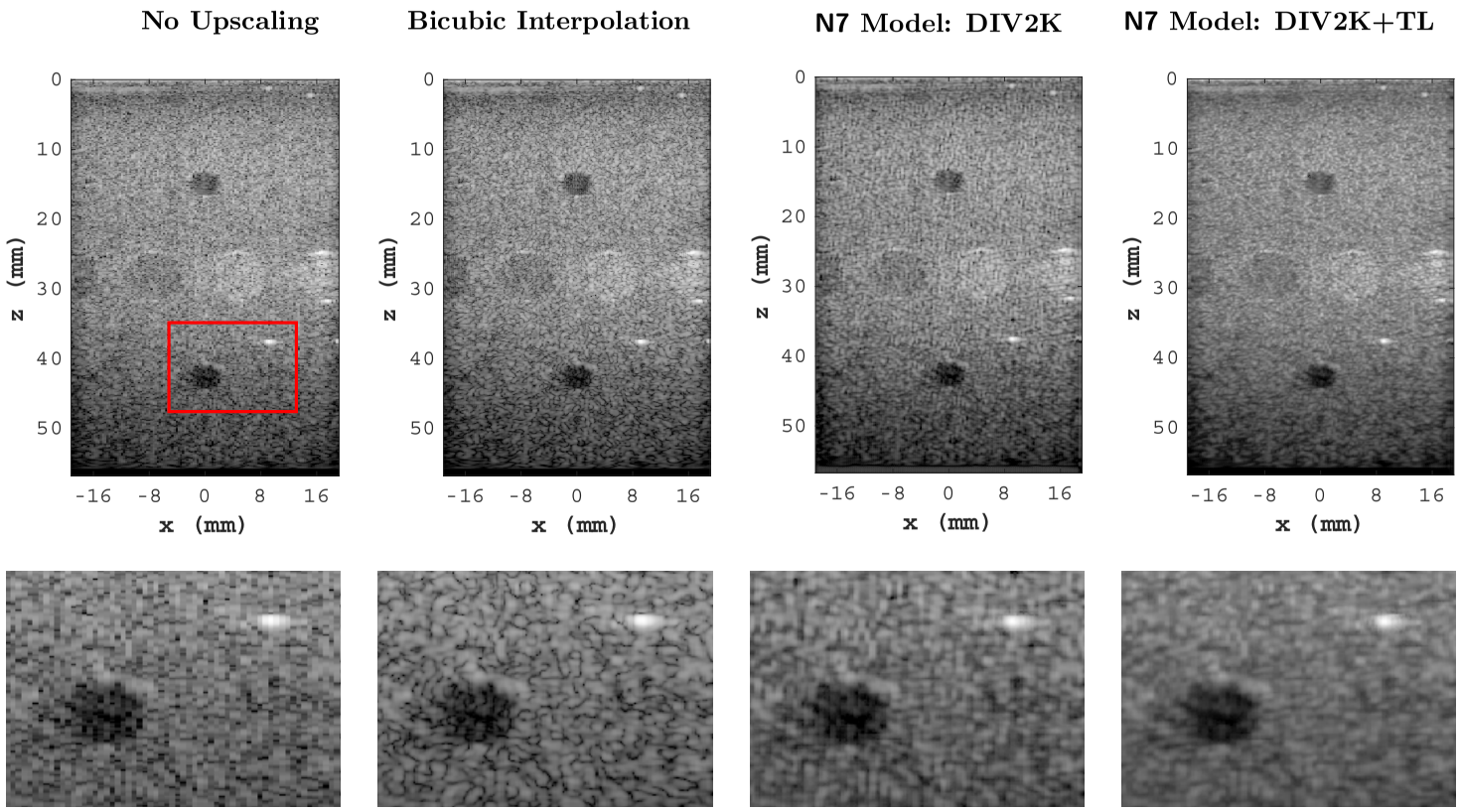


Figure 4.3: TM migration (zero-angle PW) with various upscaling options: TYPE-1 B-mode images, 60-dB range.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In our work, we proposed two approaches to enhancing 2D Fourier-beamformed envelope data of ultrasound images using lightweight superresolution CNNs. The **first** approach (described in chapter 3) uses non-ultrasound domain images for CNN training. We investigated several ESPCN-based network configurations and showed that, based on the FWHM and CNR measurements on three beamformed PICMUS datasets (labelled Type-1, Type-2, and Type-3), the image quality improved. In the **second** approach (described in chapter 4), we applied transfer learning that involves two-phase training. The first phase, general-domain training, was similar to the one used in chapter 3. The second phase, target-domain training, utilized augmented ultrasound images. We used data augmentation to generate 160 images from two PICMUS experimental datasets (Type-2 and Type-3). The third available dataset (Type-1) was used to evaluate the reconstructed image quality via CNR and FWHM measurements. In this case, the image quality significantly improved.

Another advantage of transfer learning is that it helps stabilize results. The network without transfer learning is sensitive to different training runs and initializations during general-domain training because of the difference between the training and testing datasets. Another observed benefit pertains to the selection of suitable model weights. The end results with transfer learning did not depend on whether we selected our model weights based on the early stopping we mentioned in chapter 3 or based on the last epoch of general-domain training.

5.2 Future Work

One potential area for future work aiming to enhance upscaled image quality is to explore more advanced CNN architectures and training methods [40]. This could involve investigating multi-scale networks, GANs, and self-supervised learning. A deeper study on the initialization of the network related to ultrasound imaging should also be considered. Network architecture search [75] is another promising technique that could be used here, as it uses reinforcement learning, genetic algorithms, or gradient-based optimization to find the architecture that performs well on the target task.

The placement of the CNN in different stages of the beamforming process can also be worth exploring in depth. An interesting study would be the use of a CNN in the frequency domain. For example, the remapping function can be replaced by a CNN which remaps and upscales the image at the same time. For this, the HR images must be transformed into the frequency domain. With this approach, we might get better image quality at the end of the data processing pipeline.

Another area for further research would be implementing the aforementioned lightweight networks in FPGA targeting real-time embedded imaging systems. This would allow for the addition of highly efficient beamforming capabilities to portable ultrasound imaging devices.

Bibliography

[1] F. Anjidani and D. Rakhmatov, "Efficient ultrasound image enhancement using lightweight CNNs," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, Oct 2022, pp. 684–688.

[2] ——, "Coupling fast superresolution CNN with fast plane-wave Fourier-domain beamforming," in *IEEE International Ultrasonics Symposium (IUS)*. IEEE, Oct 2022, pp. 1–5.

[3] A. Hadzic, *Hadzic's Peripheral Nerve Blocks and Anatomy for Ultrasound-guided Regional Anesthesia*. McGraw Hill Professional, 2011.

[4] M. Albulayli and D. Rakhmatov, "Fourier domain depth migration for plane-wave ultrasound imaging," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 65, no. 8, pp. 1321–1333, 2018.

[5] D. Go, J. Kang, I. Song, and Y. Yoo, "Efficient transmit delay calculation in ultrasound coherent plane-wave compound imaging for curved array transducers," *Applied Sciences*, vol. 9, no. 13, p. 2752, 2019.

[6] S. Musti, "Plane-wave Fourier-domain beamforming with CNN-assisted resolution enhancement," MEng report, University of Victoria, Victoria, 2022.

[7] 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.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 7132–7141.

[9] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 1874–1883.

[10] T. L. Szabo, *Diagnostic Ultrasound Imaging: Inside Out*. Academic Press, 2004.

[11] M. Postema, *Fundamentals of Medical Ultrasonics*. CRC Press, 2011.

[12] S. K. Dwivedi, M. Vishwakarma, and A. Soni, "Advances and researches on non destructive testing: A review," *Materials Today: Proceedings*, vol. 5, no. 2, pp. 3690–3698, 2018.

[13] M. Tanter and M. Fink, "Ultrafast imaging in biomedical ultrasound," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 61, no. 1, pp. 102–119, 2014.

[14] J. Park, J. B. Kang, J. H. Chang, and Y. Yoo, "Speckle reduction techniques in medical ultrasound imaging," *Biomedical Engineering Letters*, vol. 4, no. 1, pp. 32–40, 2014.

[15] M. Ali, D. Magee, and U. Dasgupta, "Signal processing overview of ultrasound systems for medical imaging," *SPRAB12, Texas Instruments, Texas*, vol. 55, 2008.

[16] H. T. Lutz and H. A. Gharbi, *Manual of Diagnostic Ultrasound in Infectious Tropical Diseases*. Springer, 2006.

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

[18] R. Cohen, Y. Sde-Chen, T. Chernyakova, C. Fraschini, J. Bercoff, and Y. C. Eldar, "Fourier domain beamforming for coherent plane-wave compounding," in *IEEE International Ultrasonics Symposium (IUS)*. IEEE, Oct 2015, pp. 1–4.

[19] H. L. Van Trees, "Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory," *Wiley*, 2002.

[20] T. S. Kiong, S. B. Salem, J. K. S. Paw, K. P. Sankar, and S. Darzi, "Minimum variance distortionless response beamformer with enhanced nulling level control via dynamic mutated artificial immune system," *The Scientific World Journal*, vol. 2014, 2014.

[21] S. Khan, J. Huh, and J. C. Ye, "Adaptive and compressive beamforming using deep learning for medical ultrasound," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 8, pp. 1558–1572, 2020.

[22] B. Luijten, R. Cohen, F. J. de Bruijn, H. A. Schmeitz, M. Mischi, Y. C. Eldar, and R. J. van Sloun, "Deep learning for fast adaptive beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2019, pp. 1333–1337.

[23] P. Temme and G. Müller, "Fast plane-wave and single-shot migration by Fourier transform," *Journal of Geophysics*, vol. 60, no. 1, pp. 19–27, 1986.

[24] J. Cheng and J.-y. Lu, "Extended high-frame rate imaging method with limited-diffraction beams," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 53, no. 5, pp. 880–899, 2006.

[25] A. Besson, M. Zhang, F. Varray, H. Liebgott, D. Friboulet, Y. Wiaux, J.-P. Thiran, R. E. Carrillo, and O. Bernard, "A sparse reconstruction framework for Fourier-based plane-wave imaging," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 63, no. 12, pp. 2092–2106, 2016.

[26] D. D. Liu and T.-L. Ji, "Plane wave image formation in spatial-temporal frequency domain," in *IEEE International Ultrasonics Symposium (IUS)*. IEEE, Sep 2016, pp. 1–5.

[27] C. Chen, G. A. Hendriks, R. J. van Sloun, H. H. Hansen, and C. L. de Korte, "Improved plane-wave ultrasound beamforming by incorporating angular weighting and coherent compounding in Fourier domain," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 65, no. 5, pp. 749–765, 2018.

[28] T. Chernyakova, R. Cohen, R. Mulayoff, Y. Sde-Chen, C. Fraschini, J. Bercoff, and Y. C. Eldar, "Fourier-domain beamforming and structure-based reconstruction for plane-wave imaging," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 65, no. 10, pp. 1810–1821, 2018.

[29] D. Rakhmatov, "Slant-stack migration applied to plane-wave ultrasound imaging," in *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, Nov 2021, pp. 4027–4030.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2009, pp. 248–255.

[32] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[33] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[34] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1256–1272, 2016.

[35] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[36] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deeply improved sparse coding for image super-resolution," *arXiv preprint arXiv:1507.08905*, vol. 2, no. 3, p. 4, 2015.

[37] Z. Zhou, Y. Wang, Y. Guo, Y. Qi, and J. Yu, "Image quality improvement of hand-held ultrasound devices with a two-stage generative adversarial network," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 1, pp. 298–311, 2019.

[38] Y. Qi, Y. Guo, and Y. Wang, "Image quality enhancement using a deep neural network for plane wave medical ultrasound imaging," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 4, pp. 926–934, 2020.

[39] D. Perdios, M. Vonlanthen, A. Besson, F. Martinez, M. Arditi, and J.-P. Thiran, "Deep convolutional neural network for ultrasound image enhancement," in *IEEE International Ultrasonics Symposium (IUS)*. IEEE, Oct 2018, pp. 1–4.

[40] D. Perdios, M. Vonlanthen, F. Martinez, M. Arditi, and J.-P. Thiran, "CNN-based image reconstruction method for ultrafast ultrasound imaging," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 4, pp. 1154–1168, 2021.

[41] J. Tang, B. Zou, C. Li, S. Feng, and H. Peng, "Plane-wave image reconstruction via generative adversarial network and attention mechanism," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–15, 2021.

[42] S. Khan, J. Huh, and J. C. Ye, "Variational formulation of unsupervised deep learning for ultrasound image artifact removal," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 6, pp. 2086–2100, 2021.

[43] B. Luijten, R. Cohen, F. J. de Bruijn, H. A. Schmeitz, M. Mischi, Y. C. Eldar, and R. J. van Sloun, "Adaptive ultrasound beamforming using deep learning," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3967–3978, 2020.

[44] A. A. Nair, K. N. Washington, T. D. Tran, A. Reiter, and M. A. L. Bell, "Deep learning to obtain simultaneous image and segmentation outputs from a single input of raw ultrasound channel data," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 12, pp. 2493–2509, 2020.

[45] J. Zhang, Q. He, Y. Xiao, H. Zheng, C. Wang, and J. Luo, "Ultrasound image reconstruction from plane wave radio-frequency data by self-supervised deep neural network," *Medical Image Analysis*, vol. 70, p. 102018, 2021.

[46] D. Hyun, A. Wiacek, S. Goudarzi, S. Rothlübbers, A. Asif, K. Eickel, Y. C. Eldar, J. Huang, M. Mischi, H. Rivaz *et al.*, "Deep learning for ultrasound image formation: Cubdl evaluation framework and open datasets," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 12, pp. 3466–3483, 2021.

[47] J.-Y. Lu, P.-Y. Lee, and C.-C. Huang, "Improving image quality for single-angle plane wave ultrasound imaging with convolutional neural network beam-

former," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 4, pp. 1326–1336, 2022.

[48] M. Gasse, F. Millioz, E. Roux, D. Garcia, H. Liebgott, and D. Friboulet, "High-quality plane wave compounding using convolutional neural networks," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 64, no. 10, pp. 1637–1639, 2017.

[49] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CPVRW)*. IEEE, 2017, pp. 136–144.

[50] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision (ECCV)*. Springer, Oct 2016, pp. 391–407.

[51] K.-W. Hung, Z. Zhang, and J. Jiang, "Real-time image super-resolution using recursive depthwise separable convolution network," *IEEE Access*, vol. 7, pp. 99 804–99 816, 2019.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 770–778.

[53] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential data augmentation techniques for medical imaging classification tasks," in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 979.

[54] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, "Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks," *Scientific Reports*, vol. 9, no. 1, pp. 1–9, 2019.

[55] P. M. Cheng and H. S. Malhi, "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," *Journal of Digital Imaging*, vol. 30, no. 2, pp. 234–243, 2017.

[56] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embed-

ding," in *ACM International Conference on Multimedia (ACM-MM)*. ACM, Nov 2014, pp. 675–678.

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

[58] S. Musti, F. Anjidani, and D. Rakhmatov, "Plane-wave Fourier-domain beamforming with CNN-assisted resolution enhancement," in *IEEE International Ultrasonics Symposium (IUS)*. IEEE, Sep 2021, pp. 1–4.

[59] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 126–135.

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

[61] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*. JMLR Workshop and Conference Proceedings, Mar 2010, pp. 249–256.

[62] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

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

[64] C. T. S. . P. Technology, "Multi-purpose multi-tissue ultrasound phantom—model 040gse," *Computerized Imaging Reference Systems, Inc, website*.

[65] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[66] A. Adadi, "A survey on data-efficient algorithms in big data era," *Journal of Big Data*, vol. 8, no. 1, pp. 1–54, 2021.

[67] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1019–1034, 2014.

[68] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE/CVF, 2020, pp. 702–703.

[69] L. H. Lee, Y. Gao, and J. A. Noble, "Principled ultrasound data augmentation for classification of standard planes," in *International Conference on Information Processing in Medical Imaging (IPMI)*. Springer, Jun 2021, pp. 729–741.

[70] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[71] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 114–125.

[72] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.

[73] S. Anwar, S. Khan, and N. Barnes, "A deep journey into super-resolution: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[74] L. Marple, "Computing the discrete-time "analytic" signal via FFT," *IEEE Transactions on Signal Processing*, vol. 47, no. 9, pp. 2600–2603, 1999.

[75] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.