09/09/2024

Enhancing Point-of-Care Ultrasound Image Quality Using a Conditional GAN and Paired Data

Hilde G.A. van der Pol

MASTER OF SCIENCE
TECHNICAL MEDICINE

Universiteit Leiden
TU Delft
Erasmus Universiteit Rotterdam
Netherlands Cancer Institute
Antoni van Leeuwenhoek

This page was intentionally left blank.

Enhancing Point-of-Care Ultrasound Image Quality Using a Conditional GAN and Paired Data

Hilde G.A., Pol van der

Student number: 4663209

09-09-2024

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Dept. of Biomechanical Engineering, TUDELFT

19-02-2024 – 23-09-2024

Supervisor(s):

Dr. Freija Geldof

Dr. Behdad Dashtbozorg

Prof. Dr. Jos. A. van der Hage

Thesis committee members:

Dr. Jifke F. Veenland, Erasmus MC (chair)

Dr. Freija Geldof, NKI-AvL

Dr. Behdad Dashtbozorg, NKI-AvL

Prof. Dr. Jos. A. van der Hage, LUMC

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus Universiteit Rotterdam

Acknowledgements

Before you lies my graduation thesis. This is the result of nine enriching months at the AvL-NKI, as well as the culmination of my academic journey over the past seven years. These years of studying and internships across Delft, Zurich, Rotterdam, and now Amsterdam have flown by, and I've enjoyed them greatly.

First, I want to express my gratitude to Freija and Behdad for their invaluable guidance. They were always helpful, supportive, and critical in the best possible way. Additionally, I truly appreciate the input of everyone else present at our weekly POCUS meetings—Mark, Lennard, Chrissy, and Lucie. We had many interesting discussions (always keeping it concise and sticking to the 7x7 format), along with plenty of laughs with our friends Mr. Beast and Little Bear. I'd also like to thank Jos for his medical insights and enthusiasm for the project. Lastly, I'd like to thank the KFI, who were always more than happy to help and for their excitement. Working with all of you was a pleasure!

To everyone else at the AVL—students, PhD students, and all others—thank you for being so open, kind, and always supportive and collaborative. I'm also grateful for the many coffee breaks, (party) lunch moments, and running clubs, which made my graduation internship all the more enjoyable.

Last, but certainly not least, I want to express my deep appreciation for my friends and family, who have always been there for me these past years. Thank you for your constant support throughout this journey!

As my time as a student comes to an end, I feel incredibly grateful for the valuable lessons I've learned professionally, but also on a personal level. I'm ready and excited for the next step—let's see what the future holds!

Hilde van der Pol
Amsterdam, September 2024

Abstract

Introduction. Point-of-care ultrasound (POCUS) devices are gaining popularity due to their portability and affordability, making ultrasound technology more accessible in various medical settings. However, these benefits of cost and portability come with a trade-off in imaging quality.

Aim. This thesis aims to enhance the image quality of POCUS devices using deep learning and a novel paired dataset of POCUS and high-end ultrasound images.

Method. First, an accurately paired dataset was created, comprising ex vivo and abdominal phantom images from both POCUS and high-end ultrasound systems. This was achieved by building an automated acquisition setup to ensure consistent capture locations, along with comprehensive image alignment and registration steps. Second, a deep learning network was developed to enhance POCUS image quality. A conditional Generative Adversarial Network (cGAN) with a U-Net generator, pretrained on simulation data, was trained and evaluated using this paired dataset.

Results. A paired dataset of 1064 images was collected. The proposed cGAN achieved significant improvements in image quality over low-quality input images, increasing the Structural Similarity Index Measure (SSIM) from 0.286 ± 0.062 to 0.540 ± 0.082 and Peak Signal-to-Noise Ratio (PSNR) from 19.155 ± 1.948 dB to 22.406 ± 2.189 dB. It also reduced the Natural Image Quality Evaluator (NIQE) from 7.948 ± 1.772 to 4.436 ± 0.528 and Perception-based Image Quality Evaluator (PIQE) from 31.116 ± 5.911 to 19.991 ± 5.722, where lower scores indicate higher quality. Additionally, qualitative assessment showed that end-users preferred the enhanced images over the original low-quality images.

Conclusions. This thesis presents the first accurately paired dataset of POCUS and high-end ultrasound images. Additionally, the developed cGAN successfully enhanced the quality of POCUS images, surpassing the results reported in similar studies. This work demonstrates the potential for reliable quality enhancement methods for POCUS while preserving its cost and portability benefits, ultimately increasing its value and impact in the medical field.

Contents

1

Introduction

The use of handheld devices suitable for point-of-care ultrasound (POCUS) has been on the rise in recent years. This increase in popularity can be attributed to several key characteristics of these devices. Firstly, their portability offers greater convenience compared to conventional cart-based devices. Moreover, these handheld POCUS devices are more affordable than traditional ultrasound machines [1–5], making ultrasound technology more accessible and expanding its application beyond the radiology department in hospitals. POCUS is particularly valuable in settings where larger, more expensive ultrasound equipment is impractical, such as in bedside emergency care, general practitioner offices, home care environments, and rural medicine facilities [6–13].

Despite these advantages, one of the primary drawbacks of ultrasound examination with a handheld POCUS device is their reduced imaging quality. This discrepancy in quality primarily arises from hardware constraints due to portability demands and cost limitations, which are reflected in the probe material, transducer elements, processor modules, and the absence of advanced post-processing algorithms [14, 15]. Compared to conventional high-end ultrasound systems, handheld POCUS devices typically exhibit reduced resolution and contrast, less distinct texture or edges of anatomical structures, and increased noise levels [6, 16, 17]. These limitations can potentially lead to less accurate diagnoses [2, 4]. Despite the advancements in POCUS technology in recent years, a trade-off remains between imaging quality and the benefits of cost and portability [14, 18, 19].

Efforts to enhance the POCUS image quality can be categorized into three main approaches. The first approach involves advancements in hardware. However, this approach is constrained by rising costs and compromised portability. Another option for quality improvement involves refinements in the ultrasound beamforming algorithm [20, 21]. Nevertheless, access to the raw radio frequency (RF) signals required for such improvements is limited in most commercial ultrasound systems. Therefore, this thesis centers its focus on a third alternative: modifications to the image post-processing methods, eliminating the need for hardware remodeling or operations on the raw RF signal.

Traditional post-processing techniques, such as filtering and deconvolution, have been employed for ultrasound image enhancement for some time [22]. However, in recent years

deep learning has emerged as a powerful tool, achieving state-of-the-art performance in various image processing tasks, including image quality enhancement [23–25]. Therefore, a literature review [26] was conducted to explore the current state-of-the-art progress in ultrasound image enhancement using deep learning for POCUS applications. This review highlights the potential of these methods, demonstrating substantial image quality improvements across various studies and datasets.

Notably, many studies rely on paired low-quality POCUS images with high-quality images from high-end machines. While these image pairs are expected to contain the same locational information, they frequently suffer from spatial discrepancies due to acquisition challenges. These discrepancies, which can only be partially mitigated by registration methods, impact the training of deep learning networks. Moreover, they introduce uncertainties in the ground truth (i.e., the high-quality reference images), which makes it difficult to validate the quality enhancement achieved by such models. To the best of the authors' knowledge, this issue is present in all paired POCUS datasets reported in literature, including those used in the studies by Zhou et al. [6, 18] and the open-source dataset from the Ultrasound Image Enhancement Challenge (USenhance) [27] in collaboration with MICCAI 2023. This lack of accurately paired datasets represents a significant gap in the current literature, which this thesis aims to address. Although simulation data could offer an alternative solution, the review found that simulated datasets generally exhibit higher performance gains compared to in vivo and phantom datasets, suggesting that simulation results may not fully represent clinical scenarios. Therefore, the primary goal of this thesis is to develop a setup capable of acquiring an accurately paired dataset of POCUS images and high-end ultrasound images, addressing aspects such as consistent capture locations and precise image alignment and registration.

The second aim of this study is to leverage this paired dataset for the development of a deep-learning network aimed at enhancing POCUS images. The literature review [26] showed that Generative Adversarial Networks (GANs) are employed in the majority of studies focused on ultrasound enhancement. Originally proposed by Goodfellow et al. [28], GANs have gained significant popularity in recent years, particularly in image processing and generation. These deep learning frameworks have also made notable advancements in healthcare applications, demonstrating strong performance in medical image reconstruction and enhancement [25, 29–31]. In essence, a GAN consists of two competing deep learning networks—the generator and the discriminator—that engage in an adversarial process to train a model capable of generating realistic data. A particularly relevant variation of GANs is the conditional GAN (cGAN), with a well-known implementation called pix2pix, developed by Isola et al. [32]. This cGAN is specifically designed for image-to-image translation tasks. It is therefore well-suited for the objective of enhancing ultrasound images using a paired dataset, as the goal is to map low-quality handheld images to high-quality reference images.

In summary, this study has two main objectives; 1) to collect the first accurately paired dataset of POCUS and high-end ultrasound images, and 2) to use this paired dataset to

train a deep learning network for enhancing POCUS image quality. Thereby, this thesis aims to contribute to the quality enhancement of POCUS, while preserving its cost and portability benefits, ultimately increasing its value to the medical field.

2

Method

This thesis consists of two phases, which will be described in the following sections. Section 2.1 outlines the development of an image acquisition setup for the collection of a paired dataset of low-quality POCUS and high-quality, high-end ultrasound images. Section 2.2 presents the development of a deep learning network aimed at enhancing the quality of POCUS images.

2.1 Collection of paired ultrasound images

A measurement setup was designed to obtain a paired dataset consisting of images from both a low-cost POCUS system (low-quality) and a high-end ultrasound machine (high-quality). Additionally, paired images were collected using a mid-range POCUS device from a higher cost segment (intermediate-quality). In the following subsections, the ultrasound devices will first be specified. Next, the development of the image acquisition setup will be described, followed by an overview of the ex vivo specimens and abdominal phantom used. Then, the data acquisition along with a method for maximizing the locational agreement between paired images from different ultrasound devices will be outlined. Finally, the preprocessing steps for obtaining the paired dataset will be illustrated.

2.1.1 Ultrasound devices

Ultrasound imaging was performed using three devices of varying quality. These devices, in increasing order of quality, included a cost-effective handheld device (Telemed), a mid-range handheld device (Clarius), and a high-end cart-based device (Philips). The specifications for each of the ultrasound devices can be found in Table 2.1. Figure 2.1 shows a photo of each of the three ultrasound devices along with an image acquired by each device.

In this thesis, image acquisition was done using three probes. However, this study focused on enhancing the lowest-quality ultrasound images obtained with the low-cost probe (Telemed), utilizing the high-quality images acquired with the highest-quality ultrasound device (Philips). The primary goal of this study was to demonstrate the feasibility of such quality enhancement. Although images of intermediate-quality were also obtained with the mid-range handheld device (Clarius), there quality was not enhanced. Nevertheless, the same approach could be applied to these images in future research.

Table 2.1: Specifications of ultrasound devices used in this thesis.

	Telemed	**Clarius**	**Philips**
Description	Handheld, cost-effective, with wire	Handheld, mid-range, wireless (Bluetooth and Wi-Fi)	Cart-based, high-end, with wire
Device specifications	MicrUS EXT-1H; TELEMED, Vilnius, Lithuania	HD3; Clarius Mobile Health, Vancouver, British Columbia, Canada	CX50; Philips, Amsterdam, Netherlands
Probe specifications	TL15-6L25S-3	L20 HD3	L15-7io
Dimensions	106x105x21 mm (MicroUs box)	147x76x32 mm (wireless probe)	400x350x70 mm (device without trolley)
Weight	0.26 kg (MicroUs box)	0.29 kg (wireless probe)	7.3 kg (device without trolley)
Price estimation	≈€3.5K (MicrUs box + transducer)	≈€8K (wireless probe + software package)	>€50K (device + transducer)
Number of piezoelectric elements	64	192	128
Frequency range	6-15 MHz	8-20 MHz	7-15 MHz
Max depth	45 mm	40 mm	30 mm
Scan width	25 mm (scan width)	25 mm (Field of View)	23 mm (aperture)
Access to unprocessed raw images	Yes	No	No
Access to radiofrequency signals	No	No	No

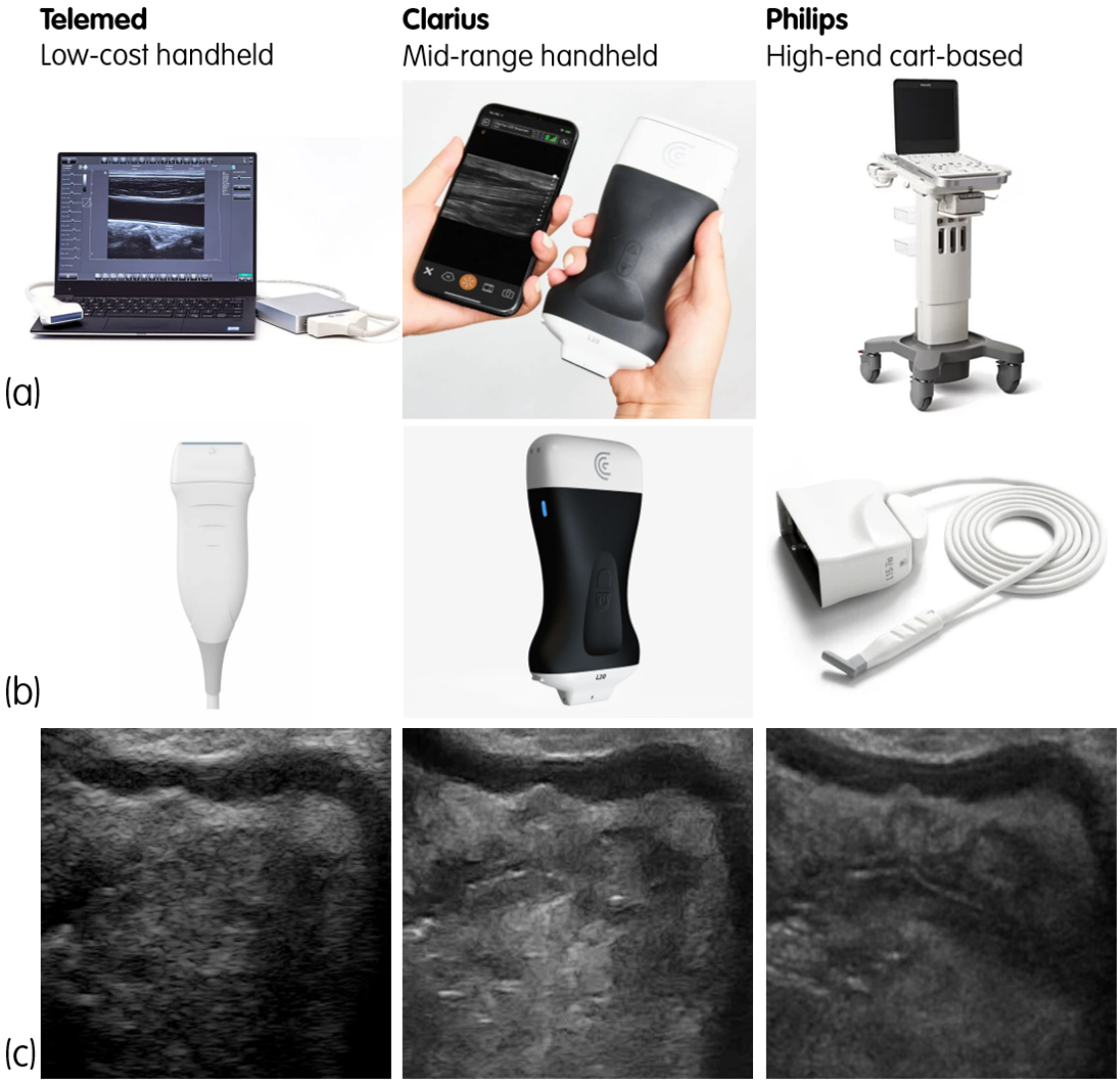

Figure 2.1: (a) Ultrasound devices used in this thesis, listed in increasing order of image quality: a cost-effective handheld device (Telemed), a mid-range handheld device (Clarius), and a high-end cart-based device (Philips). (b) Transducer used with each device. (c) Example ultrasound images acquired with each device.

2.1.2 Image acquisition setup

In order to develop a paired dataset, it was essential to ensure consistent image capture locations for each ultrasound probe. This was achieved using a motorized and automated setup, where the ultrasound probes were hovered over the image subject to acquire images at predetermined locations. This was accomplished with a two-axis gantry system—a framework allowing precise and automated movements along motorized rails—to which the ultrasound probes were attached. The NEJE 3 system, originally designed for laser engraving and capable of automatic and controllable movements along the x- and y-axes, served as the foundation. The system was modified by removing the laser engraver, elevating the structure, and extending the rails to accommodate larger image subjects. Additionally, a z-axis was added, allowing for manual height adjustment of the probes above the subject. A custom 3D-printed holder was designed to secure the ultrasound probes to the gantry system, ensuring they remained stationary throughout image acquisition sessions. This setup allowed each probe to capture images from consistent positions relative to the image subject. The gantry system, equipped with the probes and the 3D-printed holder, is shown in Figure 2.2-a.

Custom software was developed to automatically execute the image acquisition process according to a predefined grid of acquisition locations, which will be detailed in Section 2.1.4. The software, written in Python (version 3.8.18), controlled both the gantry system and the three ultrasound devices. The gantry system could be operated through G-code programming. For the Telemed ultrasound device, real-time image acquisition was possible using the TELEMED Real-time Imaging for the Research Dynamic Link Library (DLL) (version 1.0.1), which allows calling the high level programming library "Ultrasonography for Windows II" Software Development Kit (Usgfw2 SDK) (version 4.3.0). For the Clarius device, images could be streamed through Python using the Cast API (version 11.2.0), which requires the Clarius App (version 11.2.2) to be running simultaneously. For the Philips machine, images could be automatically acquired through Python using a VGA to USB Video Converter in combination with Open Broadcaster Software (OBS) Studio (version 30.1.2), enabling real-time recording of 1080p HD images. The developed software ensured fully automatic image acquisition, facilitating consistent and repeatable image capture for each ultrasound probe.

2.1.3 Ex vivo and abdominal phantom data

A diverse dataset was created using various ex vivo specimens and an abdominal phantom. For the ex vivo dataset, breast, colorectal, and sarcoma specimens were used. Patients were included if they were diagnosed with sarcoma or colorectal cancer and underwent a surgical resection, or were scheduled for a breast mastectomy at the Antoni van Leeuwenhoek Hospital - Netherlands Cancer Institute. Specimen with a maximum dimension smaller than 5 cm were excluded for practical reasons. All patients have given permission for the further utilization of their data and biological materials for scientific research. The specimens were collected and measured immediately after surgical removal. After the freshly excised specimen was collected, it was placed in a vacuum-sealed bag to facilitate stable

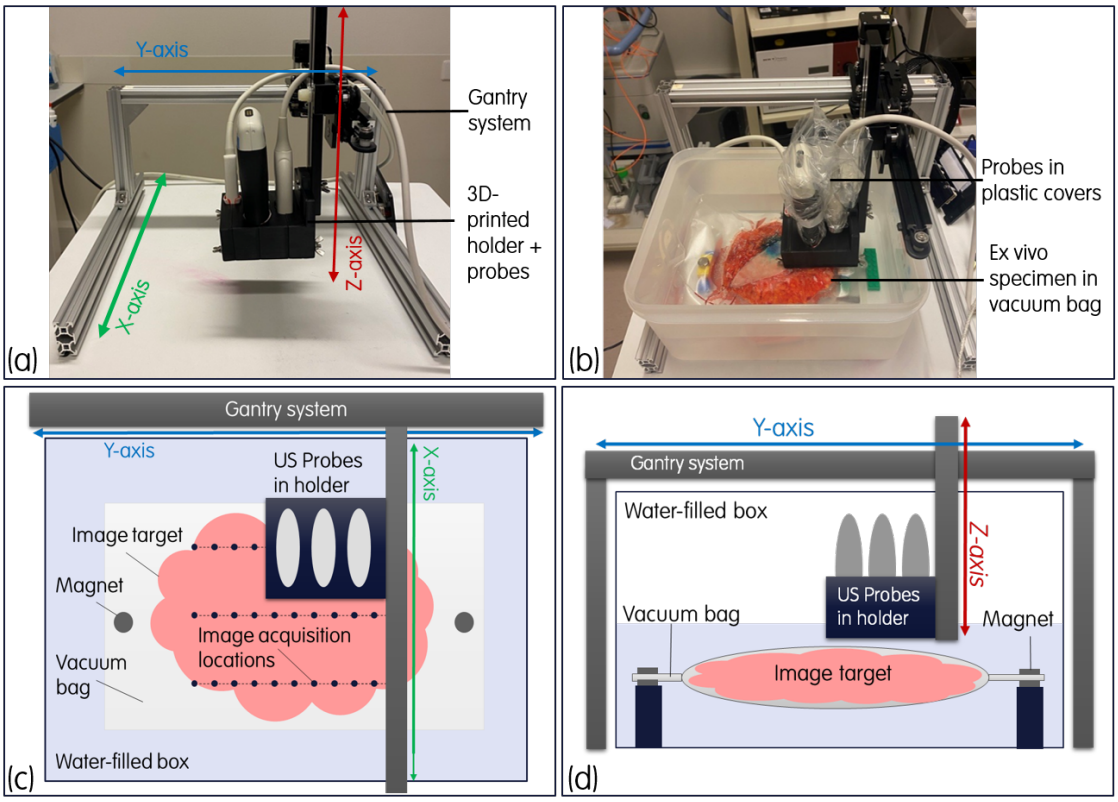

Figure 2.2: Overview of the image acquisition setup: (a) The gantry system equipped with three ultrasound probes secured in a 3D-printed holder; (b) The complete setup during image acquisition of an ex vivo specimen inside a vacuum-sealed bag, with the probes protected from water by plastic sterile covers; (c) Schematic top view of the image acquisition setup, showing the ex vivo specimen inside the vacuum-sealed bag. Additionally, the grid of acquisition locations is visualized, with each dot representing a position for image capture; (d) Schematic side view of the image acquisition setup.

placement in a water-filled container. Placing the specimen inside the water ensured optimal acoustic contact of the ultrasound devices without air interference. This was crucial, as the ultrasound probes hovered over the specimens rather than making direct contact in the developed image acquisition setup. The vacuum-sealed bag ensured preservation of the specimen when placed in the water. Additionally, it prevented the specimen from floating, as the bag could be secured within the container using magnets. An example of an ex vivo specimen inside a vacuum-sealed bag is depicted in Figure 2.2-b.

Additionally, images were acquired using the Abdominal Intraoperative and Laparoscopic Ultrasound Phantom "IOUSFAN" (Kyoto Kagaku Co., Ltd, Kyoto, Japan). This abdominal phantom simulates the anatomical structures of abdominal organs and is designed as a training tool for abdominal ultrasound examinations during surgery. This phantom was filled with water for optimal acoustic contact, which was necessary as the probes were not in direct contact with the phantom. An illustration of this abdominal phantom is shown in Figure 2.3.

Figure 2.3: Abdominal Intraoperative and Laparoscopic Ultrasound Phantom.

2.1.4 Data acquisition

For each device, a scan depth of 3 centimeters was used with a corresponding frequency of 14 MHz. The maximum number of focal zones was used, resulting in 4 zones for the Telemed and Philips devices, and 1 focal zone for the Clarius device. The gain and dynamic range settings were adjusted as needed, avoiding underexposure, oversaturation, and loss of contrast or resolution. This was monitored by checking the pixel value histograms so that no more than 0.1% of the pixels exceeded the maximum value of 255, ensuring minimal saturation and thereby preventing loss of information.

For each specimen and the abdominal phantom, the aim was to capture as many images as possible, while minimizing the overlap between images. This was achieved by acquiring images at locations according to a predefined grid with fixed step sizes: 20 mm in the x-direction, corresponding to the approximate image width captured by the ultrasound probes, and 5 mm in the y-direction, corresponding to the approximate slice thickness. The dimensions of the grid were adjusted based on the size of each specimen in both the x- and y-directions. The covered areas of the specimens ranged from a minimum of 5 cm to a maximum of 20 cm. Consequently, between 40 and 140 images were acquired per specimen. An example grid of acquisition locations can be seen in Figure 2.2-c. An image of the complete image acquisition setup, as well as a schematic overview, is shown in Figure 2.2.

2.1.5 Maximizing locational agreement between paired images

Maximizing the locational agreement between paired images acquired with different probes was crucial for developing an accurately paired dataset. Consistent image capture locations for each probe were enabled by the automated image acquisition setup with fixed probe positions, as explained in Section 2.1.2. However, further refinement was needed to achieve more precise alignment between images from each probe. Locational differences primarily arose from the use of stick-on sterile plastic probe covers, which were necessary to protect the probes from water. These probe covers can be seen in Figure 2.2-b. These covers could introduce variations in the probes' positioning relative to each other due to the accumulation of excessive plastic, creating offsets. These variations occurred only between different

imaging sessions, when the covers were re-applied and the probes were re-secured in the holder. Once the covers were in place and the probes were fixed, their relative positions remained constant within each session.

The offset between images acquired by each probe was minimized in two steps: an alignment step and a registration step. These steps were facilitated by an in-house developed calibration phantom. The calibration phantom consisted of hyperechoic targets—nylon monofilament wires with a diameter of 0.30 mm—arranged in a specific pattern within a 3D-printed tool. An image of the calibration phantom is shown in Figure 2.4. Placing the calibration phantom in a water-filled box made the hyperechoic wire targets clearly visible against the anechoic background.

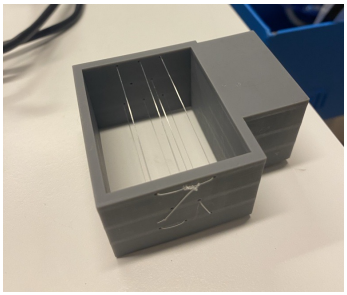

Figure 2.4: Photo of the in-house developed calibration phantom, consisting of a 3D-printed tool and hyperechoic nylon monofilament wires.

Image alignment

First, an alignment step was performed before each image acquisition session. The goal of this alignment was to ensure that each probe captured an image slice as similar as possible to the others. To achieve this, the distance between the probes needed to be precisely established, accounting for offsets introduced by variations in probe and cover placement. In this study, this distance between the probes is defined in the y-direction. This y-alignment was critical to perform before each measurement, as it cannot be corrected post-acquisition due to the loss of spatial information. During alignment, the calibration phantom was positioned so that the first probe was precisely aligned above a hyperechoic wire parallel to the direction of the image slice, as illustrated in Figure 2.5-a. The probe holder with the fixed probes was then moved along the y-axis to align the second and third probes with the wire. This was done by testing a range of different possible distances between the probes (Figure 2.5-b). The images that best depicted the wire were identified (Figure 2.5-c). Because the wire was very thin, precise positioning was ensured; even a small offset resulted in an image that did not capture the wire. This information was then used to determine the correct distance between the probes for that specific imaging session. These established distances were then taken into account by the software, ensuring that each probe captured the same image slices. This process is illustrated in Figure 2.5.

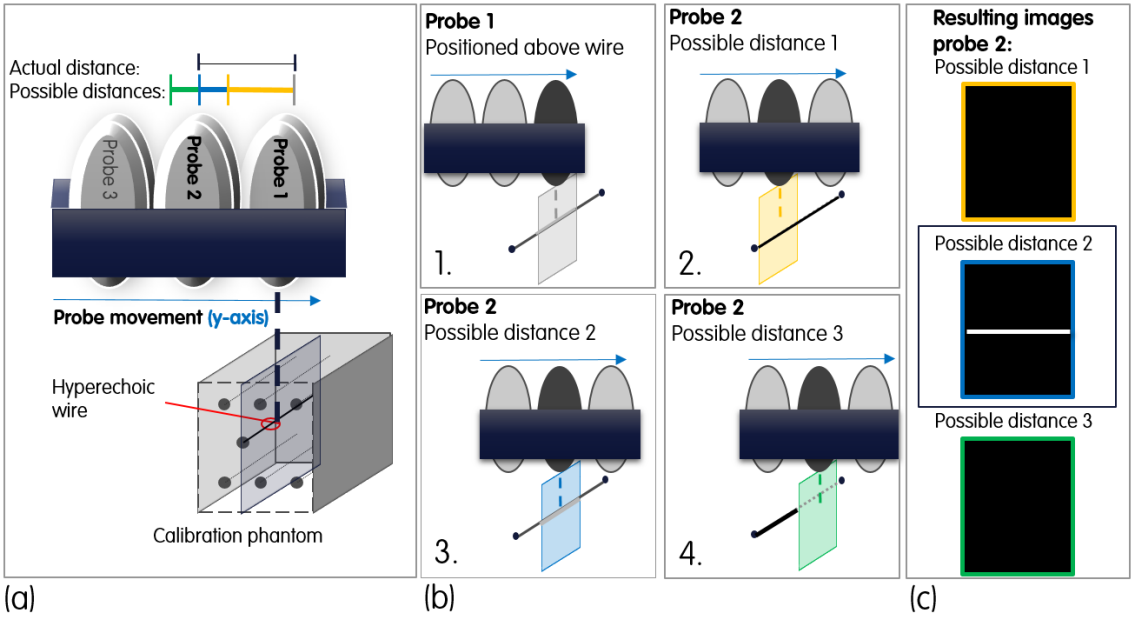

Figure 2.5: (a) Side view of Y-axis alignment process to determine the correct distance between probes, which was done to ensure each image captured the same image slice. This example shows how the correct distance between probes 1 and 2 was obtained by moving the probe holder according to different possible distances between the two probes. (b) Probe 1 was placed precisely above the wire (b.1). Then probe 2 was moved towards the wire according to the different possible distances, acquiring an image at each position (b.2-b.4). (c) Resulting images acquired for probe 2 in position b.2-b.4. The wire was only visible when probe 2 was exactly above it, indicating the correct distance between probe 1 and 2 (distance 2 in this example). The same process was repeated for probes 1 and 3.

Image registration

Second, a registration step was performed before each imaging session. While the previous step ensured that each probe captured the same image slice, this registration step focused on aligning the image plane, ensuring that each probe captured consistent information across the image width and height. For this registration step, the same calibration phantom with hyperechoic wires was used, but now 90 degrees rotated. When acquiring an ultrasound image of the calibration phantom in this orientation, it resulted in a dot pattern. Before each image acquisition session, a registration image depicting these dots was acquired for each probe, resulting in three registration images. This process is shown in Figure 2.6.

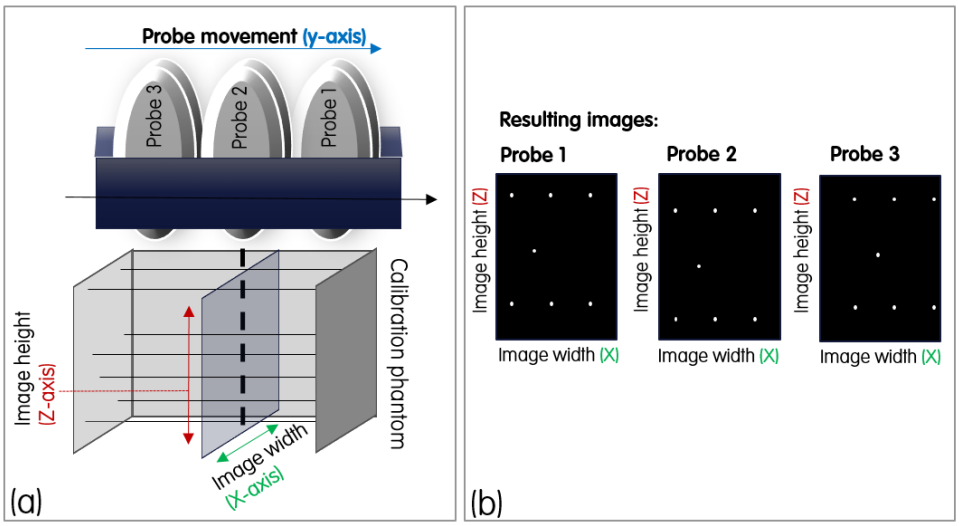

Figure 2.6: (a) Side view of the acquisition of registration images using the calibration phantom. Each probe was positioned at the same location above the calibration phantom, to capture a registration image. (b) Resulting images acquired for each probe, showing a dot pattern. These dot-pattern images were used in the next step of the registration process.

These calibration phantom images, depicting the dot patterns, were used to accurately register the images from each probe, aligning them along the image width (X-axis) and height (Z-axis). In this study, the registration was performed only for the low-quality images from the low-cost handheld ultrasound device and the high-quality images, while disregarding the intermediate-quality images from the mid-range handheld device. First, three corresponding dots were selected from the low- and high-quality registration images of the same imaging session. These dots were used to calculate a 2x3 transformation matrix, which applies an affine transformation involving translation, rotation, scaling, and skewing, while preserving collinearity (see Figure 2.7-a). While this served as an important initial registration step, it was not sufficient for training the deep learning network, for which it is crucial that the images are registered as accurately as possible. Poorly registered images could cause the network to focus on correcting misalignments rather than enhancing the low-quality images to more closely resemble the high-quality reference.

Therefore, a more comprehensive registration step was implemented using Elastix [33, 34], a well-established open-source toolbox offering a wide range of algorithms for image registration. Since the images do not show complete overlap, structures at the edges of one image that are not present in the other can hinder registration. To address this, the Elastix registration was constrained to the areas present in both images by using the regions of interest (ROIs) from the initial registration step as masks. For each imaging session, these ROI masks were applied to one pair of images—one low-quality and one high-quality—acquired at the same location (see Figure 2.7-b). Elastix then determined the spatial transformation needed to best align each low-quality image with its corresponding high-quality image by optimizing an image similarity metric (see Figure 2.7-c). This registration method in Elastix was applied to each image pair. An affine registration, as defined in the default Affine parameter map of Elastix, was chosen as a starting point. Affine registration was deemed the most suitable for this dataset since it provides a good balance between flexibility for correcting small global misalignments and preservation of anatomical structures, while avoiding unnecessary complexity. Given the challenges of image registration and the lack of a one-size-fits-all solution, the parameter settings were fine-tuned through trial and error to achieve satisfactory results for this specific dataset. The final parameter settings used are detailed in Section 3.1.2.

2.1.6 Image preprocessing

Before using the images as training input for the deep learning network, several preprocessing steps were performed. This involved cropping, resampling, and normalizing the pixel value range. After the paired low- and high-quality images were registered as described in 2.1.5, the paired images were cropped to retain only the overlapping area, ensuring that both images were rectangular and of identical size, as shown in Figure 2.7-d. To meet the input requirements of the deep learning network, images were resampled and, if necessary, further cropped to achieve consistent dimensions across all images. Both image width and height were adjusted to be divisible by 32 pixels, a standard requirement for deep learning networks. The aspect ratio was preserved, and cropping was minimized to retain as much information as possible. Finally, the images were normalized to a range of -1 to 1 to prevent large gradients, accelerate convergence and enable high precision weights of the deep learning network.

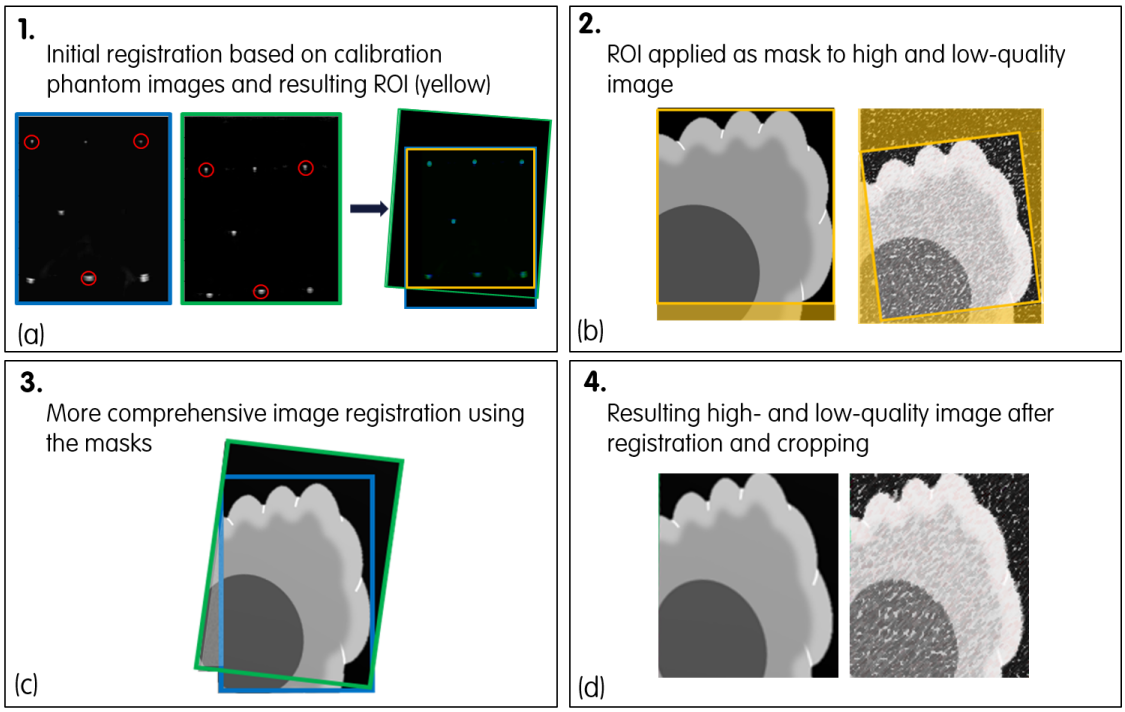

Figure 2.7: Overview of the registration steps applied to the dataset of low- and high-quality paired images. (a) Initial registration of images using calibration phantom images from both the low- and high-quality probes. Three corresponding dots from each image were selected, as indicated by the red circles, to calculate the initial registration (b) The initial registration is used to define regions of interest (ROIs) for both low- and high-quality images of an ex vivo subject, which are then applied as masks. (c) More comprehensive registration for this ex vivo image pair is performed using the Elastix library with the defined masks. (d) Final images after registration and cropping for both low- and high-quality datasets.

2.2 Development of deep learning network

Using the acquired paired dataset, a deep learning network was developed to enhance POCUS image quality. The paired images were used as input to train the network to map the low-quality images to the corresponding high-quality ones. The following subsections will provide an overview of the network architecture and its loss function. Next, additional steps to enhance the model, including pretraining and image augmentation, will be discussed. Then, the implementation details will be outlined, followed by a description of the experiments conducted in the ablation study. Finally, the quantitative and qualitative evaluation methods employed will be discussed.

2.2.1 Network architecture

This study employed a conditional GAN (cGAN). A GAN is a framework for estimating generative models via an adversarial process [28]. It involves the simultaneous training of two models: a generative model and a discriminative model, which compete against each other in a two-player game. The generator's objective is to produce outputs that are indistinguishable from real images, while the discriminator's role is to accurately distinguish between real and generated images. Unlike traditional GANs, which learn a generative model of data, cGANs learn a conditional generative model. This makes cGANs particularly suitable for image-to-image translation tasks, where the network is conditioned on an input image to generate a corresponding output image[32]. A schematic overview of the cGAN framework is shown in Figure 2.8.

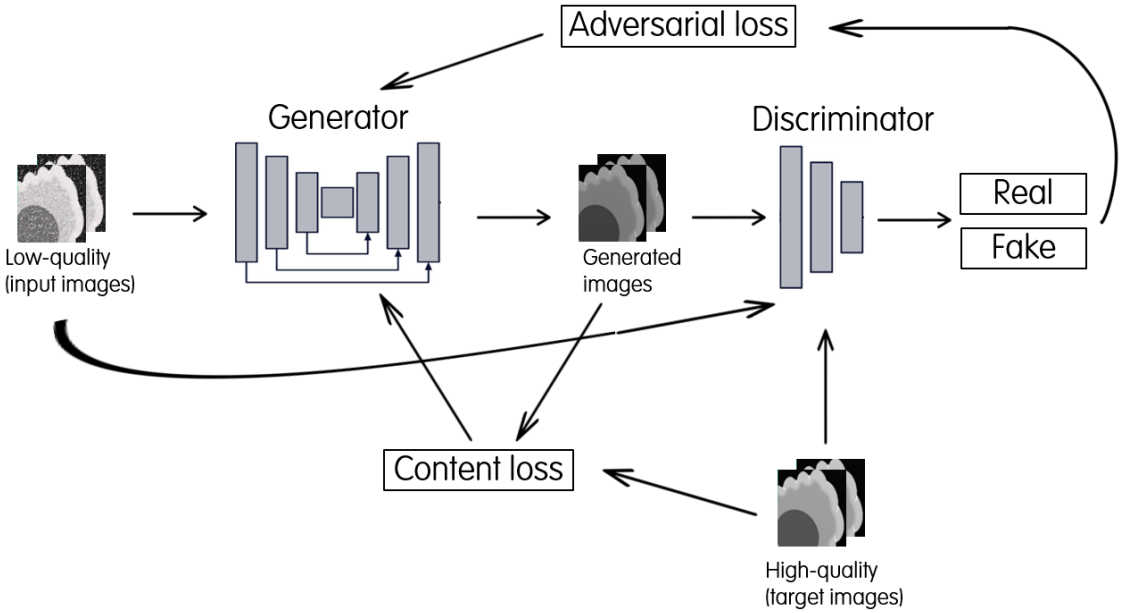

Figure 2.8: Schematic overview of the cGAN framework.

The open-source cGAN model by Isola et al. [32], referred to as pix2pix, was used in this study. It was developed as a general-purpose solution to image-to-image translation problems. The generator in this model employs a U-Net[35], which is an encoder-decoder architecture where the input undergoes progressive downsampling through a series of layers until reaching a bottleneck layer. The process is then reversed to upsample the data back to its original resolution. Additionally, U-Net incorporates skip connections into the encoder-decoder network. These connections enable the network to bypass the bottleneck by directly transferring spatial information across the network, helping to recover details lost during the downsampling process.

The discriminator used is termed PATCHGan, a convolutional neural network consisting of layers with increasing depth that attempts to classify each 70×70 patch in an image as real or fake. Using a patch size smaller than the full size of the image is advantageous because a smaller PATCHGan has fewer parameters, runs faster, and can be applied to arbitrarily large images while still producing high-quality results [32]. The discriminator operates convolutionally across the image, averaging all patch-level responses to produce the final output. It focuses solely on local structures at the scale of these image patches, thus capturing high-frequency, detailed information. Therefore, the PATCHGAN can be understood as a form of texture/style loss [36, 37]. This design choice for the discriminator, restricting it to model only local and detailed structures, is motivated by the use of an L1 term in the generator. This L1 term ensures that low-frequency global information is accurately captured, as will be explained in subsection 2.2.2.

2.2.2 Loss function

The loss function of GAN models is a critical aspect of their design and is known as the adversarial loss function [28]. This adversarial loss function, used for training both the generator and the discriminator networks, aims to create a model capable of generating realistic data. An interesting extension of the cGAN architecture by Isola et al. [32] involves modifying the loss function by adding a content loss to the generator loss. This addition ensures that the generated images are not only indistinguishable from real ones but also closely resemble the target images. These two key components—the adversarial loss and content loss—that make up the cGAN loss function will be further explained below. Additionally, they are illustrated in the schematic overview of the cGAN framework in Figure 2.8.

Adversarial Loss

GANs are generative models that learn to map a random noise vector z to a target image y, represented as $G : z \rightarrow y$. In contrast, cGANs learn to map an observed image x and a random noise vector z to y, expressed as $G : \{x, z\} \rightarrow y$. In this context, x is the low-quality input image and y is the high-quality target image. The generator G is trained to produce outputs that are indistinguishable from real images. More specifically, in this context, the generator of the cGAN tries to generate an image using the low-quality input image that is indistinguishable from the high-quality reference image. The discriminator D receives pairs consisting of the input image with the high-quality target image and the input

image with the generated image. It then tries to discriminate between the pair containing the real image (high-quality reference image) and the fake image (image generated by the generator). This adversarial training framework is central to GANs, in which the two networks are pitted against each other and thereby improve themselves. The objective of this adversarial training is captured by the adversarial loss function $\mathcal{L}_{\text{cGAN}}(G, D)$. The adversarial loss is implemented as binary cross-entropy (BCE) and is formulated as:

$$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))], \tag{2.1}$$

where G is the generator, D is the discriminator, $\mathbb{E}$ denotes the expectation operator, x is the low-quality input, and y is the high-quality target image. The generator G aims to minimize this loss function. To do so, it must maximize the probability that the discriminator classifies the generated images $G(x, z)$ as real, thereby minimizing the second term in the equation. Note that the generator has no control over the first term so it will only try to minimize the second term. In contrast, the discriminator D aims to maximize the loss function by learning to distinguish as accurately as possible between fake and real images. It does this by increasing the probability that it correctly identifies the pair with the input image and the high-quality reference image $D(x, y)$ as real, thereby maximizing the first term in the loss function. At the same time, it tries to minimize the probability that it mistakenly classifies the generated image $D(x, G(x, z)))$ as real, thereby also maximizing the second term. In summary, the discriminator's parameters are adjusted to maximize the loss function, while the generator's parameters are adjusted to minimize it. This creates the classic min-max game central to GANs, ultimately leading to the optimal generator G^*, as expressed by the equation:

$$G^* = \arg \min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D). \tag{2.2}$$

Content Loss

To further improve the ability of the generator to generate images that are indistinguishable from the high-quality reference images, the generator loss doesn't only comprise the adversarial loss, but also includes a content loss. The aim of adding this content loss is to ensure that the generator not only fools the discriminator but also produces images that are close to the target images, in this context, the high-quality reference images. The content loss is usually a traditional pixel-based loss function. In pix2pix, the content loss is represented by the L1 loss. The L1 loss measures the pixel-wise difference between the ground truth image y and the generated image $G(x, z)$, encouraging the generator to produce outputs that are close to the target images, defined as:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}\left[\|y - G(x, z)\|_1\right]. \tag{2.3}$$

Thus, combining both adversarial and content (L1) loss, gives the following objective for the optimal generator:

$$G^* = \arg \min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) + \lambda \mathcal{L}_{L1}(G), \tag{2.4}$$

where λ is a weight that balances the contribution of the content loss and adversarial loss, and is set to 100.

However, the conducted literature review on deep learning models for POCUS quality enhancement [26] showed that multiple content loss functions, such as pixel, structure, and sometimes feature-based information, are often combined. While the L1 loss serves as a pixel loss, Structural Similarity Index (SSIM) was added in this study as a structure loss. The SSIM loss function enhances the visual similarity between the generated and real images, as demonstrated by Zhou et al. [6], who found that incorporating SSIM loss helped their GAN model to capture more detailed information in ultrasound images, especially tissue structure and speckle. The SSIM loss models the distortion of the reconstructed image based on three factors: brightness, contrast, and structure, and is defined as follows:

$$\mathcal{L}_{\text{SSIM}}(G) = \mathbb{E}_{x,y,z}\left[1 - \frac{(2\mu_y\mu_{G(x,z)} + C_1)(2\sigma_{yG(x,z)} + C_2)}{(\mu_y^2 + \mu_{G(x,z)}^2 + C_1)(\sigma_y^2 + \sigma_{G(x,z)}^2 + C_2)}\right], \tag{2.5}$$

where μ_y and $\mu_{G(x,z)}$ are the means of the real and generated images, σ_y and $\sigma_{G(x,z)}$ are the variances, $\sigma_{yG(x,z)}$ is the covariance, and C_1 and C_2 are constants used to stabilize the division. The final objective for the generator, including the adversarial loss, L1 loss, and SSIM loss, becomes:

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) + \lambda_1 \mathcal{L}_{L1}(G) + \lambda_2 \mathcal{L}_{\text{SSIM}}(G), \tag{2.6}$$

with λ_1 and λ_2 balancing the contributions of L1 and SSIM losses, set at 70 and 30, respectively.

2.2.3 Pretraining with simulation data

In addition to the cGAN network architecture described in section 2.2.1, a pretraining step was added using a simulation dataset. Pretraining was implemented to initialize the network's weights with useful features learned from a large and similar dataset— in this case, a simulation dataset— which could potentially enhance its performance on the POCUS dataset in the resulting network.

To this end, a simulation dataset was created using ultrasound images from open-source breast datasets [38–44], which were generally of high-quality. The datasets included images both with and without lesions, with the lesions being either benign or malignant. These high-quality images were artificially degraded to simulate low-quality input images, resulting in pairs of high- and low-quality images. Differences between low-quality handheld images and high-quality images typically include reduced resolution and contrast, less distinct texture and edges, and increased noise [6, 16, 17], which can also be seen in Figure 2.1-c. This makes the image-to-image translation for quality enhancement in this study a style transfer problem rather than just noise removal.

To resemble these multi-faceted differences, the quality of the open-source images was

downgraded using a variety of methods, including the addition of speckle and multiplicative noise, blurring, random gamma adjustment, JPEG compression, and down-sampling. Additionally, minor affine offsets were added between the image pairs to simulate inevitable minor registration inaccuracies present in the dataset acquired in this study. These degradation methods were applied to each image pair in a randomized manner, with parameters set within a specified range for each method, to introduce variability into the dataset. An example image pair from the created simulation dataset is shown in Figure 2.9. The simulation dataset contained 5564 image pairs. Pretraining was conducted over 200 epochs, after which the pretrained generator was used in the standard training process.

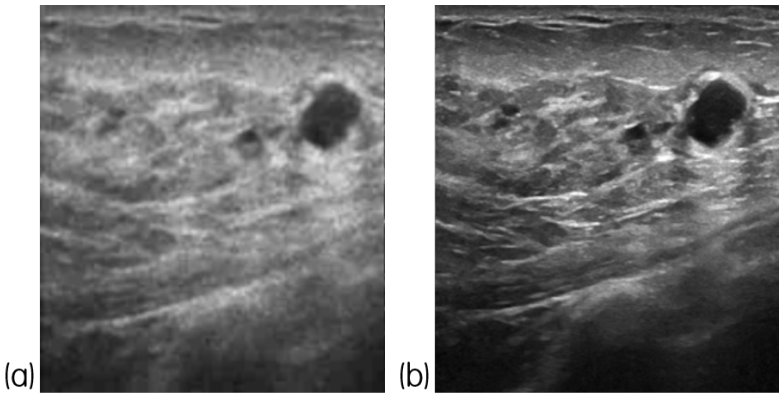

Figure 2.9: Image pair from the simulation dataset used for pretraining, showing (a) the artificially degraded low-quality input image and (b) the corresponding high-quality target image.

2.2.4 Data augmentation

As neural networks generally require substantial training data to achieve good performance and avoid overfitting, a data augmentation step was incorporated to artificially expand the dataset size and variability. The original pix2pix paper employed geometric augmentations such as random jittering (slightly resizing and then randomly cropping) and mirroring. To further enhance the augmentation in this study, additional illumination adjustments were applied using the Albumentations library (v.1.4.14) [45]. This library facilitates the creation of augmentation pipelines by defining a sequence of transformations with specified probabilities and magnitudes, effectively transforming both the input and target images. The additional transforms included changes in brightness and contrast, as well as luminance adjustments through gamma correction. These transforms were applied according to set probabilities, with the strength of each transform randomly determined within a specified range of magnitudes for each transformation. The effects of randomly applying these augmentations to both images in a low- and high-quality pair are shown in Figure 2.10.

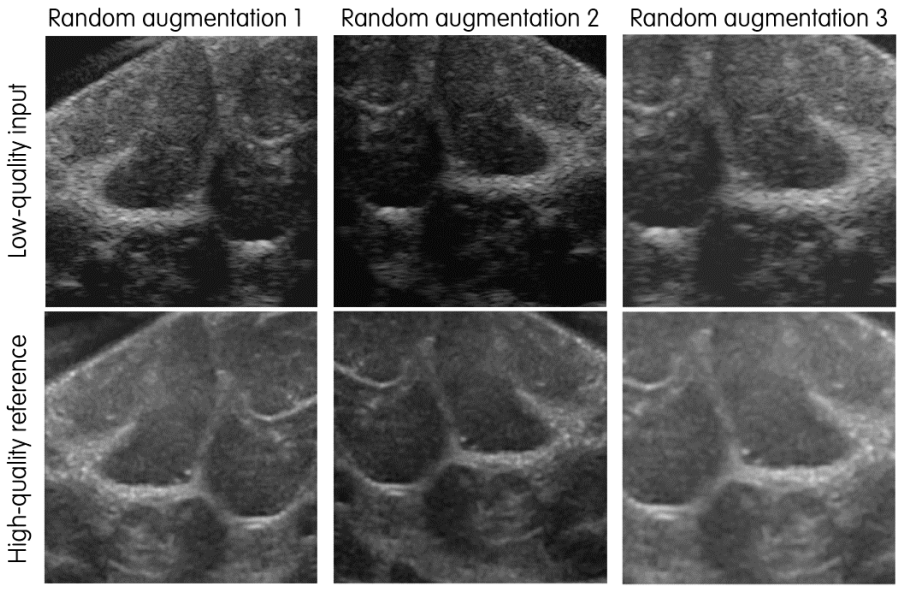

Figure 2.10: The effects of randomly applied augmentations. The top row displays low-quality images, and the bottom row shows the corresponding high-quality reference images. The columns represent three different augmentation instances.

2.2.5 Implementation details

For the development of the cGAN, the complete dataset, including both phantom and ex vivo data, was randomly split into training and test sets with a 90:10 ratio, without stratification by patient. From the test set, 10 samples were reserved for validation to assess model convergence. The cGAN is implemented using PyTorch (version 2.3.1) on a single Nvidia GeForce GTX 1080 GPU.

Following the pix2pix approach [32], the network was optimized by alternating gradient descent steps between the discriminator (D) and generator (G). Instead of training G to minimize $log(1 - D(x, G(x, z)))$, the approach maximizes $logD(x, G(x, z))$, as suggested in the original GAN paper [28]. To balance the training process, the objective function for the discriminator is halved, resulting in more balanced learning rates and slowing down the discriminator's learning compared to the generator's. The networks are trained using minibatch SGD with the Adam optimizer, with a learning rate of 0.0002 and momentum parameters $\beta 1 = 0.5$, $\beta 2 = 0.999$. Given the relatively small size of the training dataset, the number of filters in the generator was reduced. Specifically, the number of filters in each layer was decreased, with the last convolutional layer reducing from 64 to 32 filters. This change led to a reduction in the total number of parameters from approximately 55,000 to 15,000. This reduction was deemed sufficient for effective learning while enhancing training speed and reducing memory usage. A batch size of 32 was used to improve network stability, unlike the original pix2pix approach, which experimented with batch sizes ranging from 1 to 10.

The model was trained for 300 epochs. Unlike other deep learning models that are trained

with a loss function until convergence, a GAN generator model is trained using a second discriminator model. As such, there is no objective loss function and no straightforward way to monitor training progress or to model quality based on loss alone [46]. In this study, L1 and SSIM loss were used alongside adversarial loss. Unlike adversarial loss, these losses provide some insight into the training progress. However, these losses are only indicative, as they represent just a part of the total loss function. For example, a low L1 loss can still result in generated images of poor quality due to blurring or missing important details, making L1 loss insufficient on its own for determining when to stop training. Therefore, it is common practice to generate and visually inspect a few sample images during training to check convergence by human judgment [46]. This approach was also applied in this study, using the 10 validation images. Although manual inspection is the simplest method of model evaluation, it is subjective, and results can be closely related, making conclusive decisions difficult. To address this, SSIM and PSNR metrics were calculated on the validation dataset as well. Consequently, training was stopped based on a combination of L1 and SSIM loss, visual inspection of the validation samples, and performance metrics for these validation samples.These aspects were monitored for signs of convergence, including when improvements in L1 and SSIM loss plateaued, when validation images no longer showed noticeable improvements, and by assessing the changes in validation metrics. It is important to note that only images from the validation dataset were monitored and used for model development, ensuring an unbiased final evaluation on the unseen test set.

2.2.6 Ablation study

Ablation experiments were conducted to investigate the impact of different aspects of the proposed model on the quality of the reconstructed images. First, the effect of adding SSIM to the content loss (which in the baseline model comprised only L1) was examined, resulting in the 'L1 + SSIM' model. This model was compared quantitatively to the baseline model, referred to as 'baseline (L1)'. The baseline model's objective is detailed in Equation 2.4, while the objective of the L1 + SSIM model is given in Equation 2.6.

Additionally, for the loss function that achieved the best quantitative results, the effects of pretraining (described in Section 2.2.3) and data augmentation (described in Section 2.2.4) were further evaluated. These models are referred to as 'augmentation' and 'pretraining', respectively. If both data augmentation and pretraining improve the quantitative results, their combined effect will be assessed in a final model.

For the best-performing model, a two-sided paired t-test was conducted to compare its performance metrics with the low-quality input and assess whether it achieved a statistically significant improvement in image quality. Additionally, if applicable, the best-performing model will be compared to the baseline model to determine if the proposed improvements result in a significant quality enhancement. This statistical analysis will be performed for each quantitative performance metric, using a significance level of $p = 0.05$.

2.2.7 Quantitative evaluation metrics

The developed models were assessed quantitatively using both full-reference metrics and no-reference metrics. Full-reference metrics evaluate image quality by comparing a test image to a high-quality reference image, whilst no-reference methods evaluate the image quality without any reference image. The full-reference metrics used in this study are given below, defined for a reference image f and a test image g.

1) Structural similarity index measure (SSIM)[47]: predicts perceived quality by assessing perceptual similarity between paired images, correlating with human visual perception. It models image distortion through loss of correlation, luminance distortion, and contrast distortion. The SSIM ranges between -1 and 1, with a value of 1 indicating perfect correlation, 0 indicating no correlation, and -1 indicating anti-correlation between the images [47–49]. For images f and g, the SSIM is defined as:

$$\mathrm{SSIM}(f,g) = \frac{(2\mu_f\mu_g + C_1)(2\sigma_{fg} + C_2)}{(\mu_f^2 + \mu_g^2 + C_1)(\sigma_f^2 + \sigma_g^2 + C_2)}, \tag{2.7}$$

where μ denotes the mean, σ the (co)variance, and C_1 and C_2 are positive regularization constants used to avoid instability in image regions where the local mean or standard deviation is close to zero. These constants are small non-zero values based on the pixel intensity range (255). Specifically, C_1 is calculated by multiplying this range by a luminance constant (0.01), while C_2 is calculated by multiplying it by a contrast constant (0.03).

2) Peak signal-to-noise ratio (PSNR): reflects pixel-based similarity by measuring the ratio between the maximum possible signal power and the power of noise affecting image quality. A higher PSNR value indicates better image quality and more detail, while a lower value indicates greater differences between images [48, 49]. For images f and g, both of size MxN with maximum intensity MAX_I and the Mean Squared Error MSE, the PSNR is defined as:

$$\mathrm{PSNR}(f,g) = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}(f,g)}\right), \tag{2.8}$$

where MSE represents:

$$\mathrm{MSE}(f,g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(f_{ij} - g_{ij})^2. \tag{2.9}$$

The following no-reference metrics were used:

3) Naturalness image quality evaluator (NIQE)[50]: an advanced image quality metric for natural images that uses statistical features from a spatial domain natural scene statistic model. It effectively reflects the quality of medical images, as verified in [51]. The metric is inversely correlated to perceptual image quality, where lower scores indicate higher perceptual quality and higher scores suggest lower perceptual quality. NIQE assessments were performed using the default model provided in Matlab (MATLAB Research

R2024a; The MathWorks, Inc).

4) Perception-based image quality evaluator (PIQE)[52]: evaluates image quality based on human perception using statistical features of an input image. It is an unsupervised method that does not require a learning model, and has been shown to effectively reflect medical image quality [53]. PIQE values range between 0 and 100, with lower values indicating better perceptual quality. The quality scale is as follows: excellent (0-20), good (21-35), fair (36-50), poor (51-80), and bad (81-100). The PIQE metric was evaluated using MATLAB (MATLAB Research R2024a; the MathWorks, Inc).

2.2.8 Qualitative evaluation

Quantitative metrics have limitations in fully capturing the perceptual quality of images. Therefore, a qualitative assessment was also conducted. The images generated by the best-performing deep learning model were qualitatively compared to the corresponding low- and intermediate-quality images acquired with the low-cost and mid-range handheld devices, respectively.

Six researchers with experience in ultrasound imaging, acquisitions and analysis participated in the evaluation. The model-generated images, along with the low- and intermediate-quality images, were presented side-by-side in random pairs, ensuring that the left/right positioning of the images and the combination of categories (low-quality, medium-quality, and model-generated) were randomized. Simultaneously, the image acquired with the high-quality probe was displayed as a ground truth reference. This setup allowed the reviewers to identify correct structures that should be visible and detect missing structures due to poor quality or issues such as hallucinations produced by the generative model. Each reviewer independently selected the image from the pair that they believed showed the highest quality, best corresponding to the ground truth image. This process was repeated for 45 one-by-one comparisons, each time comparing two of the categories, using images from the test set. An example of such a comparison as presented to the reviewers is shown in Figure 2.11. The results were used to rank the three categories (model-generated, low-quality, and intermediate-quality) based on how frequently they were chosen over the others, expressed as percentages. These percentages were calculated based on the total number of times each category appeared in a comparison with one of the other two categories. A score of 0% indicates that the category was never chosen, while 100% indicates that it was always chosen.

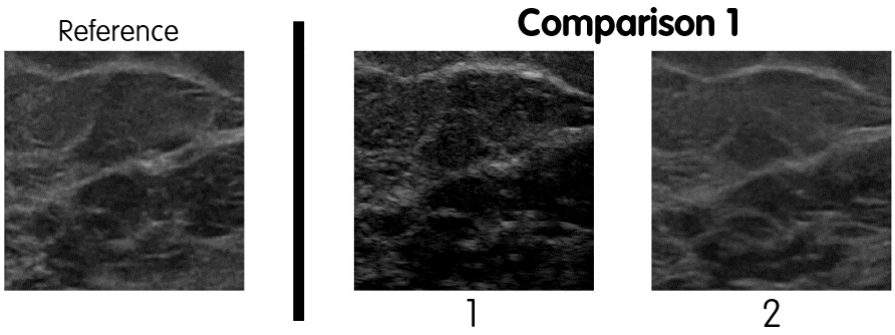

Figure 2.11: Example of a one-by-one comparison as shown to the reviewers in the qualitative evaluation, presented alongside a high-quality reference image (ground truth). In this example, a low-quality image is shown on the left (1) and a model-generated image on the right (2).

3

Results

3.1 Collection of paired ultrasound images

3.1.1 Dataset characteristics

Ultrasound images were acquired using each of the three probes (low-, intermediate-, and high-quality) at various locations for every image target. This resulted in a total of 1299 images per probe, so 1299 sets of paired images. Images were excluded if they predominantly showed water with only a small part of the image target, or if they showed large artefacts caused by air bubbles in the vacuum bag. After exclusion, 1064 image sets remained, including 936 ex vivo and 128 phantom images. An overview of the characteristics of the acquired images can be found in Table 3.1.

Table 3.1: Overview of acquired paired US image dataset, after exclusions.

Target type	**Specifications**	**Number of targets**	**Number of image sets**
Ex vivo	Breast specimen	9	527
	Sarcoma specimen	3	153
	Colorectal specimen	3	256
Phantom	IOUSFAN	1	128
Total	**Ex vivo & phantom**	**16**	**1064**

3.1.2 Registration and preprocessing

After acquisition, the images from the different probes were registered. This was achieved by experimenting with the default Affine parameter map of Elastix until satisfactory registration results were obtained. The following adjustments were made to the parameters: Advanced Normalized Correlation as the image metric, a maximum of 200 iterations, 40 resolutions, and 2000 spatial samples. Figure 3.1-a shows unprocessed example image pairs acquired with the low-quality (Telemed) and high-quality (Philips) ultrasound devices. Figure 3.1-b displays these image pairs after registration and preprocessing.

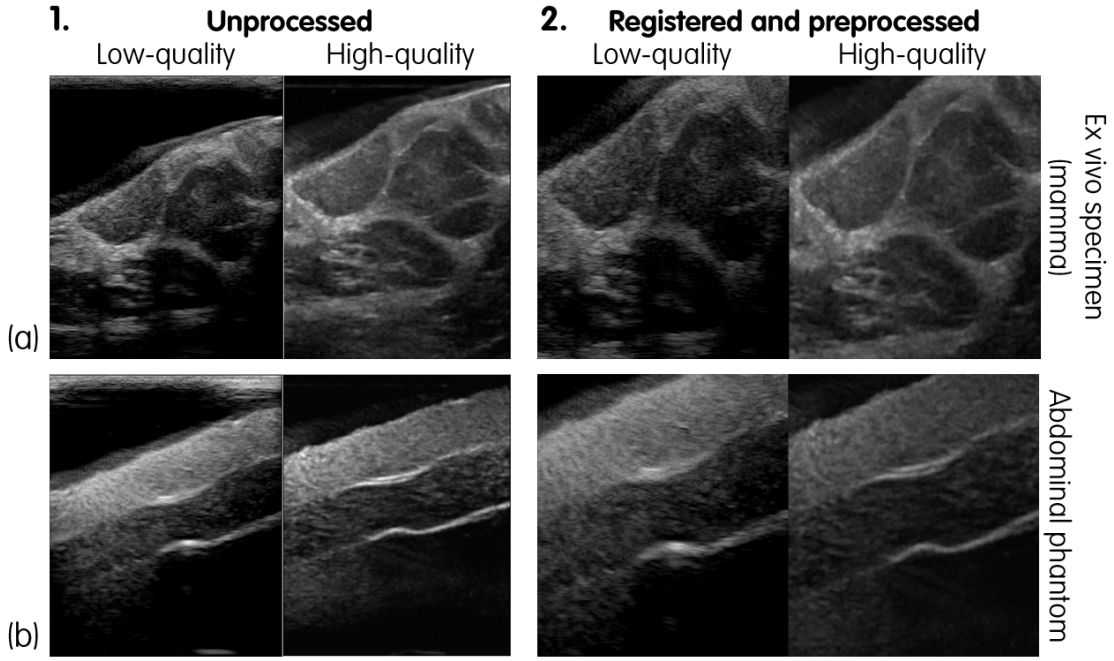

Figure 3.1: Image pairs acquired with low- and high-quality ultrasound devices: (1) unprocessed and (2) after registration and preprocessing, showing examples of (a) an ex vivo specimen (mamma) and (b) an abdominal phantom.

3.2 Development of deep learning network

3.2.1 Model convergence

To determine the appropriate stopping point for training, several indicators were monitored, including the L1 loss curve during training (and SSIM if applicable), graphs of the SSIM and PSNR metrics for the validation set, and the evolution of generated images at specific epochs for the validation set. These are illustrated for the baseline (L1) model in the supplementary materials. From epochs 200 to 300, the L1 loss shows minimal change, as can be seen in Figure 1 in the supplementary materials. The SSIM and PSNR metrics for the validation set plateau before reaching epoch 300 as well (Figure 2 and 3 in the supplementary materials). However, these metrics alone do not fully capture the network's progress, as the generated images from the validation dataset show continued quality improvement up to epoch 300, though the progress diminishes over time. This can be seen by comparing images at epochs 210 and 300 in Figure 4 in the supplementary materials. Based on these findings, training was stopped at epoch 300 for all models.

Additionally, attention should be paid to significant drops in performance metrics of the validation dataset, such as the one observed at epoch 240. When visually assessing the validation images generated by the model stopped at an epoch around this performance drop, the images indeed appear qualitatively less satisfactory and less similar to the high-quality reference images, as shown in Figure 5 in the supplementary materials. For each

model, the validation metrics and generated images were checked to ensure the model was not within such a performance drop when training stopped at epoch 300. Since this was not the case for any of the models, epoch 300 was definitively chosen as stopping point for all models during testing.

3.2.2 Ablation study and quantitative evaluation

Four different models were developed and compared: the baseline (L1), L1+SSIM, augmentation (L1+SSIM), and pretraining model (L1+SSIM). Training for each model took approximately 4.5 hours, while pretraining on the simulation dataset required around 25 hours. The inference time per image was 6.9 ± 1.4 milliseconds.

Quantitative metrics for each model, evaluated on the test set, are detailed in Table 3.2. It can be seen that applying the baseline model increased the SSIM from 0.286 to 0.495 and the PSNR from 19.155 dB to 21.735 dB compared to the low-quality input. Moreover, the NIQE decreased from 7.948 to 5.130 and the PIQE decreased from 31.116 to 28.982. When assessing the added value of adding SSIM in the content loss, it can be seen that the L1+SSIM model outperformed the baseline (L1) model across most quantitative metrics. Therefore, subsequent experiments with data augmentation and pretraining used this combination of SSIM and L1 as the content loss. Among all models, the pretraining (L1+SSIM) model achieved the best performance across all evaluation metrics. The augmentation (L1+SSIM) model, on the other hand, showed comparable or worse performance compared to the other models. As a result, the pretraining and augmentation steps were not combined into a fifth model, and the pretraining model with combined L1 and SSIM content loss was selected as the best-performing final model. This final model achieved a quality enhancement in the SSIM from 0.286 to 0.540 and the PSNR from 19.155 dB to 22.406 dB compared to the low-quality input. Additionally, the NIQE decreased from 7.948 to 4.436 and the PIQE decreased from 31.116 to 19.991, surpassing the scores of the high-quality reference images. Moreover, the paired t-test revealed statistically significant improvements across all image metrics ($p < 0.001$) when comparing the final best-performing model to both the low-quality input and the baseline model.

Example images generated by each model, alongside the low-quality input and high-quality reference images, are presented in Figure 3.2. Overall, the images produced by the best-performing model (pretraining (L1+SSIM)) show satisfactory visual quality and closely resemble the high-quality reference image, effectively enhancing the low-quality input. Consistent with the quantitative metrics, the pretraining model's images in some examples show higher quality and greater similarity to the reference image compared to those generated by other models, for instance in Figure 3.2-a, -b, and -f. Conversely, the augmentation model's images generally exhibit less satisfactory quality in these examples. However, for some test images, such as those in Figures 3.2-c, -d, and -e, the images generated by all models appear quite similar to one another. Notably, the images generated by the deep learning models are often visually more appealing and more similar to the high-quality reference images compared to the low-quality input.

Table 3.2: Full-reference and no-reference image quality metrics.

	SSIM (mean ± std)	**PSNR (dB)** (mean ± std)	**NIQE** (mean ± std)	**PIQE** (mean ± std)
High-quality reference	-	-	5.171 ± 0.600	23.288 ± 6.744
Low-quality input	0.286 ± 0.062	19.155 ± 1.948	7.948 ± 1.772	31.116 ± 5.911
Baseline (L1)	0.495 ± 0.086	21.735 ± 2.005	5.130 ± 0.462	28.982 ± 4.532
L1 + SSIM	0.519 ± 0.073	22.074 ± 2.094	5.329 ± 0.709	28.263 ± 5.832
Augmentation (L1 + SSIM)	0.438 ± 0.087	20.414 ± 2.282	5.562 ± 1.240	34.992 ± 7.640
Pretraining (L1 + SSIM)	0.540 ± 0.082	22.406 ± 2.189	4.436 ± 0.528	19.991 ± 5.722

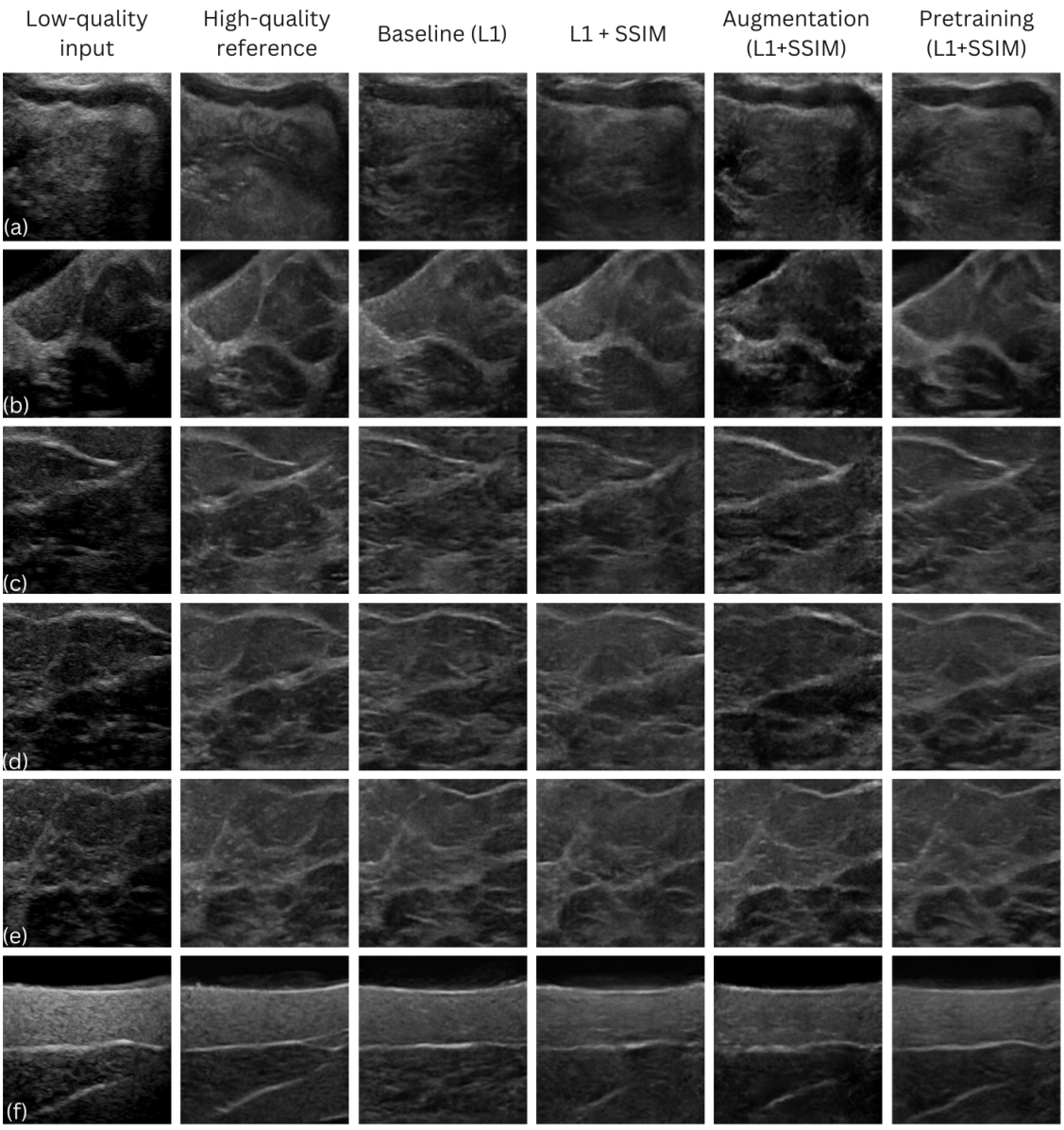

Figure 3.2: Example images from the test set (a-f), displaying the low-quality input, high-quality reference, and the generated images by each of the four models: Baseline, L1 + SSIM, Augmentation, and Pretraining.

3.2.3 Qualitative evaluation

The final model was also qualitatively compared to the low- and intermediate-quality images, with the results summarized in Table 3.3. Note that the high-quality reference images are not listed here, as they served as the ground truth reference during the comparison of the other three categories. This table shows the win percentages for each category, reflecting the proportion of times each category was selected (blindly) by the expert reviewers as having the highest quality over the other two categories. The images generated by the proposed model, pretrained on the simulation dataset, achieved the highest win percentage, being selected 62.2% of the time. The images acquired with the intermediate-quality handheld probe were chosen 58.6% of the time, while the images acquired with the low-quality and low-cost handheld probe were selected only 30.6% of the time.

Table 3.3: Win percentages for each category in the qualitative image quality assessment.

	Win percentage (%)
Low-quality images	30.6
Intermediate-quality images	58.6
Model-generated images	62.2

4

Discussion

4.1 Summary of results

This thesis developed a novel, accurately paired dataset of images acquired with POCUS devices and a high-end ultrasound device. This dataset is crucial for advancing post-image processing quality enhancement for POCUS and achieving reliable real-time methods. To the best of our knowledge, this is the first POCUS dataset accurately paired with high-quality reference images without relying on artificial data. This was achieved by developing a motorized and automated setup that acquired images at consistent positions using two POCUS devices (low-cost and mid-range) and a high-end ultrasound device. Additionally, a comprehensive registration step was implemented to enhance the similarity between paired images. The collected dataset contained 1064 ex vivo and abdominal phantom images for each of the three ultrasound devices.

Additionally, a deep learning network was developed for POCUS image quality enhancement using this paired dataset. The proposed method consisted of a cGAN framework with a U-Net generator, pretrained on a simulation dataset. The proposed deep learning method achieved significant improvements in image quality over the original low-quality input. Specifically, the SSIM increased from 0.286 ± 0.062 to 0.540 ± 0.082 (+88.8%), and the PSNR improved from 19.155 ± 1.948 dB to 22.406 ± 2.189 dB (+17.0%). Additionally, the method reduced the NIQE and PIQE scores from 7.948 ± 1.772 to 4.436 ± 0.528 (-44.2%) and 31.116 ± 5.911 to 19.991 ± 5.722 (-35.8%), respectively, where lower scores indicate higher quality. Notably, the improvements even surpass the NIQE and PIQE values of the high-quality reference images, which showed a NIQE of 5.171 ± 0.600 and a PIQE of 23.288 ± 6.744. Furthermore, the qualitative assessment revealed that reviewers rated the enhanced model-generated images as superior to those from both the original low-quality images acquired with the cost-efficient handheld device and the intermediate-quality images obtained with the handheld ultrasound device from a higher price segment.

4.2 Comparison to literature

When visually comparing this dataset to other paired POCUS and high-end ultrasound datasets reported in literature, such as those from Zhou et al. [6, 18] and the open-source

dataset from the Ultrasound Image Enhancement Challenge (USenhance) [27], it is evident that the paired images in this thesis show greater similarity. The dataset collected in this thesis, comprising over 1000 image pairs, is competitive compared to the 1500 pairs in the USenhance dataset [27] and 846 pairs in the clinical dataset by Zhou et al [6]. However, the number of different ex vivo specimens (n=15) is relatively low compared to the 109 patients and 47 healthy volunteers in the USenhance dataset [27] and the dataset by Zhou et al. [6], respectively.

A previously conducted literature review by the authors [26] on deep learning methods for POCUS image quality enhancement revealed substantial disparities in low-quality input performance and variability in performance gains across datasets. These differences are likely due to heterogeneity in dataset types and ultrasound devices used across studies. Additionally, an interesting observation from this review was that simulated datasets generally exhibited higher performance gains compared to in vivo and phantom datasets, indicating that simulation results may not fully represent clinical scenarios. Therefore, our results can best be compared to studies with clinical in vivo or ex vivo datasets and specifically focusing on POCUS.

The quantitative results obtained in this thesis for the enhanced POCUS images generated by the proposed model surpass those reported in similar studies in literature. An overview of the results obtained in these previous studies focused on enhancing POCUS image quality using clinical data, along with the results from this thesis, is provided in Table 4.1. Guo et al. [54] demonstrated an improvement in the image quality of handheld ultrasound with their LG-Unet method, with PSNR increasing from 16.04 to 18.94 (+18.1%). Zhou et al. [6] proposed a two-stage GAN for handheld ultrasound image quality improvement. When considering only clinical data, this method improved the SSIM of low-quality input images from 0.18 ± 0.04 to 0.41 ± 0.05 (+127.8%) and the PSNR from 8.65 ± 1.32 to 18.08 ± 1.57 (+109%). In a subsequent study by Zhou et al. [18], focused on handheld ultrasound video quality reconstruction using a low-rank representation multipathway GAN, the reported metrics for low-quality input using a clinical dataset of healthy volunteers were SSIM = 0.24 ± 0.06 and PSNR = 12.68 ± 3.45. These metrics were improved to SSIM = 0.45 ± 0.06 (+87.5%) and PSNR = 19.95 ± 3.24 (+57.3%).

A noteworthy observation when comparing the results from this thesis to similar studies is that the full-reference metrics for low-quality input images were generally lower in literature. Specifically, the PSNR ranged from 8.65 to 16.04 in literature, compared to 19.16 in this thesis, and the SSIM ranged from 0.18 to 0.24 in literature, compared to 0.29 in this thesis. This low baseline quality results in relatively large performance gains reported in some of these previous papers. Nevertheless, the improvements achieved by the proposed model in this thesis achieved higher performance in absolute terms for the model-generated images compared to those reported in these studies. For instance, while other studies reported final PSNR values of 18.08 to 19.95, this thesis achieved a value of 22.41. Similarly, the SSIM improved to a range of 0.41–0.45 in other studies, compared to 0.54 in this thesis. A possible reason for these disparities in baseline quality metrics

Table 4.1: Overview of image quality metrics reported in previous studies for POCUS image enhancement using clinical data, compared to the proposed model in this thesis.

Study	**Data**	**SSIM** (mean ± std)	**PSNR (dB)** (mean ± std)
Guo et al. [54]	Low-quality input	-	16.04
	Reconstructed images	-	18.94
Zhou et al. [6]	Low-quality input	0.18 ± 0.04	8.65 ± 1.32
	Reconstructed images	0.41 ± 0.05	18.08 ± 1.57
Zhou et al. [18]	Low-quality input	0.24 ± 0.06	12.68 ± 3.45
	Reconstructed images	0.45 ± 0.06	19.95 ± 3.24
Proposed Model	Low-quality input	**0.29 ± 0.06**	**19.16 ± 1.95**
	Reconstructed images	**0.54 ± 0.08**	**22.41 ± 2.19**

might be attributed to locational differences in the unpaired datasets used in other studies. Additionally, it is not always clear in these studies if the baseline quality metrics for the low-quality input images are calculated before or after preprocessing, which includes the registration step. This might also significantly impact the full-reference performance metrics, which rely on the similarity between low-quality input images and the high-quality reference, and, consequently, the relative improvement achieved. This makes it challenging to determine if the observed relatively large improvements in full-reference metrics are solely due to quality enhancement or whether they are also influenced by the minimization of locational differences, which is not the primary aim of an image enhancement model. In contrast, the image quality improvements observed in this thesis can be predominantly attributed to the quality enhancement achieved by the proposed model, given the paired nature of the dataset used.

The challenge of reliably validating quality improvements underscores the primary contribution of this thesis to current literature: the development of a paired POCUS dataset. This dataset is crucial for advancing post-image enhancement methods for POCUS, as it not only facilitates model development but also enables reliable validation using paired and accurately registered reference images. The successful development of a deep learning network in this thesis, which effectively competes with state-of-the-art methods reported in literature, further demonstrates the potential of this dataset.

4.3 Limitations

This study acknowledges several limitations. First, while the dataset size in terms of the number of images was reasonable, including a greater variety of ex vivo specimens would have enhanced the diversity of the dataset. However, due to time constraints, further expansion was not feasible for this thesis. Additionally, the automated setup acquired images at specified locations, which were randomly selected relative to the ex vivo specimen. This sometimes resulted in images lacking clear anatomical structures, unlike manually acquired images where locations with distinct anatomical features are intentionally chosen. Addressing this limitation by acquiring more data and excluding images lacking anatomical detail could further enhance the dataset's value and potentially improve the proposed model's

performance.

Additionally, the only normalization step performed in this pipeline was re-scaling the pixel intensities in each image to a range of -1 to 1, as done in the original pix2pix paper. Experiments with gray-level stretching were also performed but did not yield the expected results and even reduced model performance. This was likely due to noise in the low-quality images acquired with the low-cost handheld probe. Using the 0.1th and 99.9th percentiles instead of the minimum and maximum values for gray level stretching—intended to limit the effect of noise—also failed to improve the results. Normalization methods such as standard normal variate (SNV), calculated based on the mean and standard deviation of each image, were not incorporated because they posed challenges in transforming the images back to their original representation. However, since many normalization techniques exist and their effectiveness depends on the specific problem, further experimentation with different methods could potentially enhance model performance.

Training instability, a common issue in GANs, was also observed in this study. Although GANs are among the most popular models for image generation, training instability remains an open problem for all GAN-based algorithms [55–57]. Stabilizing GAN training is an intensively researched area, with numerous methods and strategies proposed in literature, focusing on aspects such as loss functions, regularization, normalization techniques, training algorithms, and model architectures [56]. Future work should explore strategies to improve the training stability of the model proposed in this thesis. Additionally, incorporating cross-validation in the evaluation stage could enhance result reproducibility, since this was not yet implemented in the current study due to time constraints and the extensive training time required for deep learning models. Moreover, the differences seen in quantitative metrics between the proposed models in the ablation study were minimal. These small differences, combined with the performance fluctuations due to training instability, make it difficult to determine which of the models is the most suitable. However, the performance gain of the proposed models compared to the low-quality input is evident.

Due to time constraints in this thesis, promising options regarding the deep learning approach have remained unexplored. For example, incorporating a perceptual loss function alongside the current pixel- and structural-based loss functions could help capture visual feature-based information more effectively. A perceptual loss function based on the pre-trained VGG network, as utilized in previous ultrasound quality enhancement methods [6, 58, 59], or an ultrasound-specific perceptual loss function, as used by Zhou et al. [18], could be promising directions to explore. Moreover, while a computationally feasible network was selected for this thesis, other network architectures might be better suited for addressing high-detail, context-rich problems like ultrasound image enhancement. For instance, the two-stage GAN architecture proposed by Zhou et al. [6] utilizes a residual network with multiple residual blocks and avoids downsampling and upsampling, which is particularly useful for tasks requiring the maintenance of spatial resolution. However, this approach is computationally expensive; training the two-stage GAN required around 700 hours, which was not feasible within the scope of this thesis.

Multiple quantitative metrics were employed to assess the baseline image quality and determine the model's effectiveness in quality enhancement. By using two full-reference and two no-reference metrics, the goal was to provide a comprehensive evaluation of image quality in relation to both the high-quality reference images and overall quality. However, the test set was relatively small, containing only 90 images, as a significant portion of the data was required for training. Since quantitative metrics alone cannot fully capture the perceived quality of an ultrasound image, a qualitative assessment by end-users was included as well, again based on a limited number of images. It is also important to note that the dataset used in this study consists of ex vivo and abdominal phantom images, while the intended application would likely involve in vivo settings. Despite these limitations, the assessments provided valuable insights and facilitated comparisons with similar methods in literature.

4.4 Future recommendations

Beyond the results achieved by the proposed model in this work, this thesis lays the groundwork for fully exploiting this methodology in future research. The automated data collection setup developed in this thesis facilitates expansion of the dataset in the future, potentially leading to even better results. Additionally, exploring different deep learning approaches could further improve the outcomes obtained in this thesis.

It is important to realize that quality enhancement is not a goal in itself; it rather serves as a tool to improve the effectiveness of POCUS in specific tasks. These can include its direct clinical use or its application in downstream tasks, such as algorithms designed to support particular medical applications. To this end, more extensive validation is necessary to fully assess the model's applicability in clinical practice and its impact on algorithm performance.

When looking at the use of POCUS in daily clinical practice, it is primarily used as a supplementary tool rather than a replacement for comprehensive ultrasound. It provides physicians with immediate imaging access for rapid diagnosis and efficient patient workup and treatment, often complementing or broadening the physical examination [60]. It is important to note that the validation in this thesis is retrospective, relying on previously acquired images rather than real-time assessment of enhanced POCUS's sufficiency for specific clinical applications. Future validation should therefore focus on determining whether enhanced POCUS is sufficient for accurate diagnoses and whether it can assist in interventions through image-guided approaches. Validation should be application-specific, as different clinical settings—such as general practitioner offices or emergency rooms [7, 12, 13]—and different medical specialties—such as gastroenterology [61] or neurology [62]—each have unique requirements and considerations.

Additionally, an enhancement model for POCUS could be valuable for downstream tasks,

which involve algorithms leveraging ultrasound image data for specific medical applications. Currently, these algorithms are often developed using images acquired with high-end ultrasound devices. However, with enhanced quality, POCUS could possibly offer a more affordable and portable alternative, though its impact on these algorithms must be validated. Its potential could be explored in various ultrasound-based techniques, such as prostate volume prediction in general practitioner offices [63]. Enhanced POCUS could also be beneficial in tasks like tumor detection, segmentation, and margin assessment [64–67], aiding surgeons in intraoperative settings. Furthermore, it could support navigated surgery by assisting with vessel segmentation [68, 69] or bone registration [70], enabling accurate registration. While some of these techniques are still in early stages of development, others are already in clinical use. By offering a more compact and cost-effective option, enhanced POCUS could facilitate the widespread adoption of these ultrasound-based applications.

Looking to the future, handheld ultrasound devices are expected to be increasingly adopted by physicians as bedside tools, with their use expanding beyond the radiology department. Moreover, enhanced POCUS could be valuable in specialized tasks that currently rely on high-end ultrasound imaging, such as intraoperative guidance and navigation, facilitating their broader adoption. Enhancing POCUS image quality will therefore become increasingly important while maintaining its advantages of low cost and portability. Advancements in post-image processing models will be crucial to achieve this. This work demonstrates the potential for reliable quality enhancement methods for POCUS while preserving its cost and portability benefits, ultimately increasing its value and impact across different medical fields.

5

Conclusion

This thesis presents the first accurately paired dataset of POCUS and high-end ultrasound images, crucial for advancing post-image enhancement methods for POCUS. This dataset facilitates model development and enables reliable validation with paired, accurately registered reference images. Additionally, a cGAN was developed for POCUS image quality enhancement using this dataset. The proposed model achieved significant improvements in image quality over low-quality input images, as shown by both quantitative metrics and a qualitative assessment. Furthermore, the quantitative results surpass those reported in similar studies. The successful development of this deep learning network further highlights the methodology's potential and lays the groundwork for future research. However, further validation, particularly in real-time clinical settings, is needed to fully assess the model's clinical applicability. Nonetheless, this thesis represents a significant step toward developing reliable quality enhancement methods for POCUS, while preserving its cost and portability benefits, ultimately increasing its value and impact in the medical field.

References

1. Hashim, A., Tahir, M. J., Ullah, I., Asghar, M. S., Siddiqi, H. & Yousaf, Z. The utility of point of care ultrasonography (POCUS). *Ann Med Surg (Lond)* **71,** 102982 (2021).
2. Riley, A. *et al.* Utility of hand-held echocardiography in outpatient pediatric cardiology management. *Pediatr Cardiol* **35,** 1379–86. ISSN: 0172-0643 (2014).
3. Gilbertson, E. A., Hatton, N. D. & Ryan, J. J. Point of care ultrasound: the next evolution of medical education. *Annals of translational medicine* **8** (2020).
4. Stock, K. F., Klein, B., Steubl, D., Lersch, C., Heemann, U., Wagenpfeil, S., Eyer, F. & Clevert, D. A. Comparison of a pocket-size ultrasound device with a premium ultrasound machine: diagnostic value and time required in bedside ultrasound examination. *Abdom Imaging* **40,** 2861–6. ISSN: 0942-8925 (2015).
5. Han, P. J., Tsai, B. T., Martin, J. W., Keen, W. D., Waalen, J. & Kimura, B. J. Evidence basis for a point-of-care ultrasound examination to refine referral for outpatient echocardiography. *The American Journal of Medicine* **132,** 227–233. ISSN: 0002-9343 (2019).
6. Zhou, Z., Wang, Y., Guo, Y., Qi, Y. & Yu, J. Image Quality Improvement of Hand-Held Ultrasound Devices With a Two-Stage Generative Adversarial Network. *IEEE Transactions on Biomedical Engineering* **67,** 298–311 (2020).
7. Nelson, B. & Sanghvi, A. Out of hospital point of care ultrasound: current use models and future directions. *European Journal of Trauma and Emergency Surgery* **42,** 139–150 (2016).
8. Kolbe, N., Killu, K., Coba, V., Neri, L., Garcia, K. M., McCulloch, M., Spreafico, A. & Dulchavsky, S. Point of care ultrasound (POCUS) telemedicine project in rural Nicaragua and its impact on patient management. *Journal of ultrasound* **18,** 179–185 (2015).
9. Stewart, K. A., Navarro, S. M., Kambala, S., Tan, G., Poondla, R., Lederman, S., Barbour, K. & Lavy, C. Trends in ultrasound use in low and middle income countries: a systematic review. *International Journal of Maternal and Child Health and AIDS* **9,** 103 (2020).
10. Becker, D. M., Tafoya, C. A., Becker, S. L., Kruger, G. H., Tafoya, M. J. & Becker, T. K. The use of portable ultrasound devices in low-and middle-income countries: a systematic review of the literature. *Tropical Medicine & International Health* **21,** 294–311 (2016).
11. McBeth, P. B., Hamilton, T. & Kirkpatrick, A. W. Cost-effective remote iPhone-teathered telementored trauma telesonography. *Journal of Trauma and Acute Care Surgery* **69,** 1597–1599 (2010).
12. Evangelista, A. *et al.* Hand-held cardiac ultrasound screening performed by family doctors with remote expert support interpretation. *Heart* **102,** 376–382 (2016).
13. Osterwalder, J., Polyzogopoulou, E. & Hoffmann, B. Point-of-care ultrasound—history, current and evolving clinical concepts in emergency medicine. *Medicina* **59,** 2179 (2023).
14. Henderson, R. & Murphy, S. *Patent.* Google Patents (2017).

15. Ahn, S., Kang, J., Kim, P., Lee, G., Jeong, E., Jung, W., Park, M. & Song, T.-k. *Smartphone-based portable ultrasound imaging system: Prototype implementation and evaluation* in *2015 IEEE International Ultrasonics Symposium (IUS)* (2015), 1–4.
16. Salimi, N. *et al.* Ultrasound image quality comparison between a handheld ultrasound transducer and mid-range ultrasound machine. *POCUS journal* **7,** 154 (2022).
17. Khan, S., Huh, J. & Ye, J. C. in, 158–167 (Sept. 2021). ISBN: 978-3-030-87721-7.
18. Zhou, Z., Guo, Y. & Wang, Y. Handheld Ultrasound Video High-Quality Reconstruction Using a Low-Rank Representation Multipathway Generative Adversarial Network. *IEEE Transactions on Neural Networks and Learning Systems* **32,** 575–588 (2021).
19. Jafari, M. H. *et al.* Cardiac point-of-care to cart-based ultrasound translation using constrained CycleGAN. *International journal of computer assisted radiology and surgery* **15,** 877–886 (2020).
20. Lockwood, G. R., Talman, J. R. & Brunke, S. S. Real-time 3-D ultrasound imaging using sparse synthetic aperture beamforming. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control* **45,** 980–988 (1998).
21. Matrone, G., Savoia, A. S., Caliano, G. & Magenes, G. The delay multiply and sum beamforming algorithm in ultrasound B-mode medical imaging. *IEEE transactions on medical imaging* **34,** 940–949. ISSN: 0278-0062 (2014).
22. Contreras Ortiz, S. H., Chiu, T. & Fox, M. D. Ultrasound image enhancement: A review. *Biomedical Signal Processing and Control* **7,** 419–428. ISSN: 1746-8094. https://www.sciencedirect.com/science/article/pii/S1746809412000183 (2012).
23. Anaya-Isaza, A., Mera-Jiménez, L. & Zequera-Diaz, M. An overview of deep learning in medical imaging. *Informatics in Medicine Unlocked* **26,** 100723. ISSN: 2352-9148. https://www.sciencedirect.com/science/article/pii/S2352914821002033 (2021).
24. Zhang, H.-M. & Dong, B. A review on deep learning in medical image reconstruction. *Journal of the Operations Research Society of China* **8,** 311–340 (2020).
25. Liu, J., Li, K., Dong, H., Han, Y. & Li, R. Medical Image Processing based on Generative Adversarial Networks: A Systematic Review. *Current medical imaging.* ISSN: 1573-4056. https://doi.org/10.2174/0115734056258198230920042358 (Oct. 2023).
26. Van der Pol, H. G. A., van Karnenbeek, L. M., Wijkhuizen, M., Geldof, F. & Dashtbozorg, B. Deep Learning for Point-of-Care Ultrasound Image Quality Enhancement: A Review. *Applied Sciences* **14.** ISSN: 2076-3417. https://www.mdpi.com/2076-3417/14/16/7132 (2024).
27. Guo, Y., Zhou, S., Shi, J. & Wang, Y. *Ultrasound Image Enhancement challenge 2023* Apr. 2023. https://doi.org/10.5281/zenodo.7841250.
28. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. *Generative Adversarial Networks* 2014. arXiv: 1406.2661 [stat.ML]. https://arxiv.org/abs/1406.2661.
29. Islam, S., Aziz, M. T., Nabil, H. R., Jim, J. R., Mridha, M. F., Kabir, M. M., Asai, N. & Shin, J. Generative Adversarial Networks (GANs) in Medical Imaging: Advancements, Applications, and Challenges. *IEEE Access* **12,** 35728–35753 (2024).
30. Nayak, A. A., Venugopala, P. & Ashwini, B. A Systematic Review on Generative Adversarial Network (GAN): Challenges and Future Directions. *Archives of Computational Methods in Engineering,* 1–34 (2024).
31. Jeong, J. J., Tariq, A., Adejumo, T., Trivedi, H., Gichoya, J. W. & Banerjee, I. Systematic review of generative adversarial networks (GANs) for medical image classification and segmentation. *Journal of Digital Imaging* **35,** 137–152 (2022).

32. Isola, P., Zhu, J., Zhou, T. & Efros, A. A. Image-to-Image Translation with Conditional Adversarial Networks. *CoRR* **abs/1611.07004.** arXiv: 1611.07004. http://arxiv.org/abs/1611.07004 (2016).

33. Klein, S., Staring, M., Murphy, K., Viergever, M. A. & Pluim, J. P. W. elastix: A Toolbox for Intensity-Based Medical Image Registration. *IEEE Transactions on Medical Imaging* **29,** 196–205 (2010).

34. Ntatsis, K., Dekker, N., van der Valk, V., Birdsong, T., Zukić, D., Klein, S., Staring, M. & McCormick, M. *itk-elastix: Medical image registration in Python* in *Proceedings of the 22nd Python in Science Conference* (eds Agarwal, M., Calloway, C. & Niederhut, D.) (2023), 101–105.

35. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* 2015. arXiv: 1505.04597 [cs.CV]. https://arxiv.org/abs/1505.04597.

36. Gatys, L. A., Ecker, A. S. & Bethge, M. *Texture Synthesis Using Convolutional Neural Networks* 2015. arXiv: 1505.07376 [cs.CV]. https://arxiv.org/abs/1505.07376.

37. Gatys, L. A., Ecker, A. S. & Bethge, M. *Image Style Transfer Using Convolutional Neural Networks* in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 2414–2423.

38. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J. & Wang, L. *A new dataset and a baseline model for breast lesion detection in ultrasound videos* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2022), 614–623.

39. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data in brief* **28,** 104863 (2020).

40. Ardakani, A. A., Mohammadi, A., Mirza-Aghazadeh-Attari, M. & Acharya, U. R. An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. *Computers in Biology and Medicine* **152,** 106438 (2023).

41. Gómez-Flores, W., Gregorio-Calas, M. J. & Coelho de Albuquerque Pereira, W. BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics* **51,** 3110–3123 (2024).

42. Huang, J. *et al. BUSI_ WHU: Breast Cancer Ultrasound Image Dataset* version V1 (Mendeley Data, 2023). https://doi.org/10.17632/k6cpmwybk3.1.

43. Vallez, N., Bueno, G., Deniz, O., Rienda, M. A. & Pastor, C. *BUS-UCLM: Breast ultrasound lesion segmentation dataset* version V1 (Mendeley Data, 2024). https://doi.org/10.17632/7fvgj4jsp7.1.

44. Pawłowska, A. *et al.* Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data* **11,** 148 (2024).

45. Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M. & Kalinin, A. A. Albumentations: Fast and Flexible Image Augmentations. *Information* **11.** ISSN: 2078-2489. https://www.mdpi.com/2078-2489/11/2/125 (2020).

46. Goodfellow, I. J. NIPS 2016 Tutorial: Generative Adversarial Networks. *ArXiv* **abs/1701.00160.** https://api.semanticscholar.org/CorpusID:5216145 (2016).

47. Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **13,** 600–612 (2004).

48. Horé, A. & Ziou, D. *Image Quality Metrics: PSNR vs. SSIM* in *2010 20th International Conference on Pattern Recognition* (2010), 2366–2369.

49. Sara, U., Akter, M. & Uddin, M. S. Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study. *Journal of Computer and Communications* **07,** 8–18 (Jan. 2019).

50. Mittal, A., Soundararajan, R. & Bovik, A. C. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters* **20,** 209–212. https://api.semanticscholar.org/CorpusID:16892725 (2013).

51. Zhang, Z., Dai, G., Liang, X., Yu, S., Li, L. & Xie, Y. Can Signal-to-Noise Ratio Perform as a Baseline Indicator for Medical Image Quality Assessment. *IEEE Access* **PP,** 1–1 (Jan. 2018).

52. Venkatanath, F., Praneeth, Bh., M. C., Channappayya, S. S. & Medasani, S. S. Blind image quality evaluation using perception based features. *2015 Twenty First National Conference on Communications (NCC),* 1–6. https://api.semanticscholar.org/CorpusID:6917137 (2015).

53. Pandey, A., Yadav, D., Sharma, A., Sonker, D., Patel, C., Bal, C. & Kumar, R. Evaluation of Perception based Image Quality Evaluator (PIQE) no-reference image quality score for 99mTc-MDP bone scan images. **61,** 1415–1415 (2020).

54. Guo, B., Zhang, B., Ma, Z., Li, N., Bao, Y. & Yu, D. *High-Quality Plane Wave Compounding Using Deep Learning for Hand-Held Ultrasound Devices* in *Advanced Data Mining and Applications: 16th International Conference, ADMA 2020, Foshan, China, November 12–14, 2020, Proceedings* (Springer-Verlag, Foshan, China, 2020), 547–559. ISBN: 978-3-030-65389-7. https://doi.org/10.1007/978-3-030-65390-3_41.

55. Ahmad, Z., Jaffri, Z. u. A., Chen, M. & Bao, S. Understanding GANs: fundamentals, variants, training challenges, applications, and open problems. *Multimedia Tools and Applications,* 1–77 (2024).

56. Li, Z., Xia, P., Tao, R., Niu, H. & Li, B. A New Perspective on Stabilizing GANs Training: Direct Adversarial Training. *IEEE Transactions on Emerging Topics in Computational Intelligence* **7,** 178–189 (2023).

57. Mescheder, L., Geiger, A. & Nowozin, S. *Which training methods for GANs do actually converge?* in *International conference on machine learning* (2018), 3481–3490.

58. Lyu, Y., Jiang, X., Xu, Y., Hou, J., Zhao, X. & Zhu, X. ARU-GAN: U-shaped GAN based on Attention and Residual connection for super-resolution reconstruction. *Computers in Biology and Medicine* **164,** 107316. ISSN: 0010-4825. https://www.sciencedirect.com/science/article/pii/S0010482523007813 (2023).

59. Tang, J., Zou, B., Li, C., Feng, S. & Peng, H. Plane-Wave Image Reconstruction via Generative Adversarial Network and Attention Mechanism. *IEEE Transactions on Instrumentation and Measurement* **70,** 1–15 (2021).

60. Smallwood, N. & Dachsel, M. Point-of-care ultrasound (POCUS): unnecessary gadgetry or evidence-based medicine? *Clinical Medicine* **18,** 219–224 (2018).

61. Gilja, O. H. & Nylund, K. Point-of-care Ultrasound of the Gastrointestinal Tract. *Journal of Medical Ultrasound* **31,** 1–7 (2023).

62. Sigman, E. J., Laghari, F. J. & Sarwal, A. *Neuro POCUS* in *Seminars in Ultrasound, CT and MRI* (2023).

63. Dekalo, S., Savin, Z., Schreter, E., Marom, R., Bar-Yosef, Y., Mano, R., Yossepowitch, O. & Sofer, M. Novel ultrasound-based volume estimation of prostatic benign enlargement to improve decision-making on surgical approach. *Therapeutic Advances in Urology* **13.** PMID: 33633800, 1756287221993301. https://doi.org/10.1177/1756287221993301 (2021).

64. Geldof, F., Pruijssers, C. W. A., Jong, L. S., Veluponnar, D., Ruers, T. J. M. & Dashtbozorg, B. Tumor Segmentation in Colorectal Ultrasound Images Using an Ensemble Transfer Learning Model: Towards Intra-Operative Margin Assessment. *Diagnostics (Basel)* **13.** ISSN: 2075-4418 (Print) 2075-4418 (2023).

65. Natali, T. *et al.* Automatic hepatic tumor segmentation in intra-operative ultrasound: a supervised deep-learning approach. *Journal of Medical Imaging* **11,** 024501. https://doi.org/10.1117/1.JMI.11.2.024501 (2024).

66. Wijkhuizen, M., van Karnenbeek, L., Geldof, F., Ruers, T. J. & Dashtbozorg, B. *Ultrasound tumor detection using an adapted Mask-RCNN with a continuous objectness score* in *Medical Imaging with Deep Learning* (2024). https://openreview.net/forum?id=IHmvNgX34A.

67. Veluponnar, D. *et al.* Toward Intraoperative Margin Assessment Using a Deep Learning-Based Approach for Automatic Tumor Segmentation in Breast Lumpectomy Ultrasound Images. *Cancers (Basel)* **15.** ISSN: 2072-6694 (Print) 2072-6694 (2023).

68. Smit, J. N., Kuhlmann, K. F., Thomson, B. R., Kok, N. F., Ruers, T. J. & Fusaglia, M. Ultrasound guidance in navigated liver surgery: toward deep-learning enhanced compensation of deformation and organ motion. *International journal of computer assisted radiology and surgery* **19,** 1–9 (2024).

69. Thomson, B. R. *et al. Hepatic vessel segmentation using a reduced filter 3D U-Net in ultrasound imaging* 2019. arXiv: 1907.12109 [eess.IV]. https://arxiv.org/abs/1907.12109.

70. Hiep, M. A., Heerink, W. J., Groen, H. C. & Ruers, T. J. M. Feasibility of tracked ultrasound registration for pelvic–abdominal tumor navigation: a patient study. *International journal of computer assisted radiology and surgery* **18,** 1725–1734 (2023).

Supplementary materials

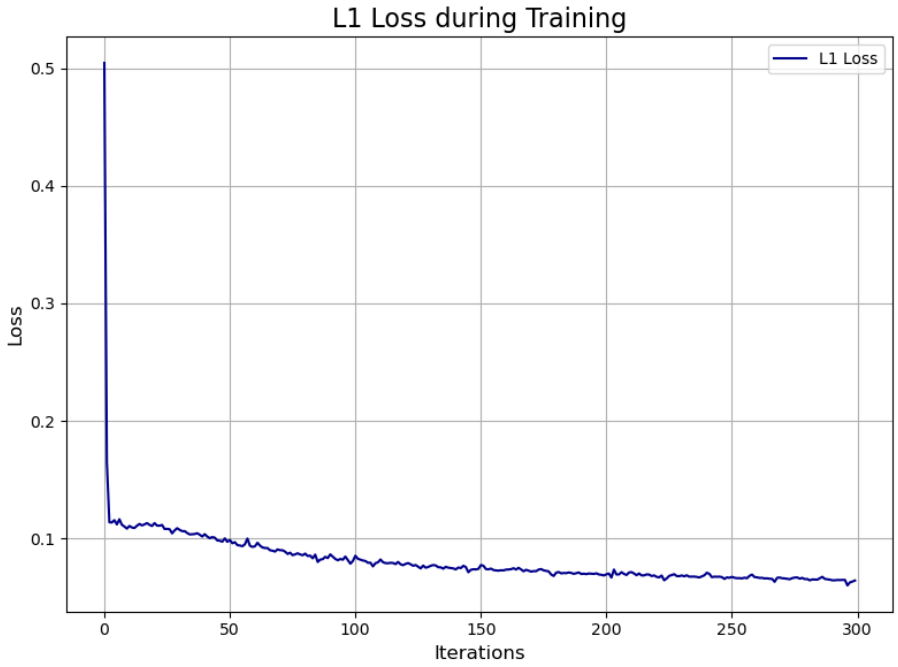

Figure 1: L1 loss during training of the baseline model.

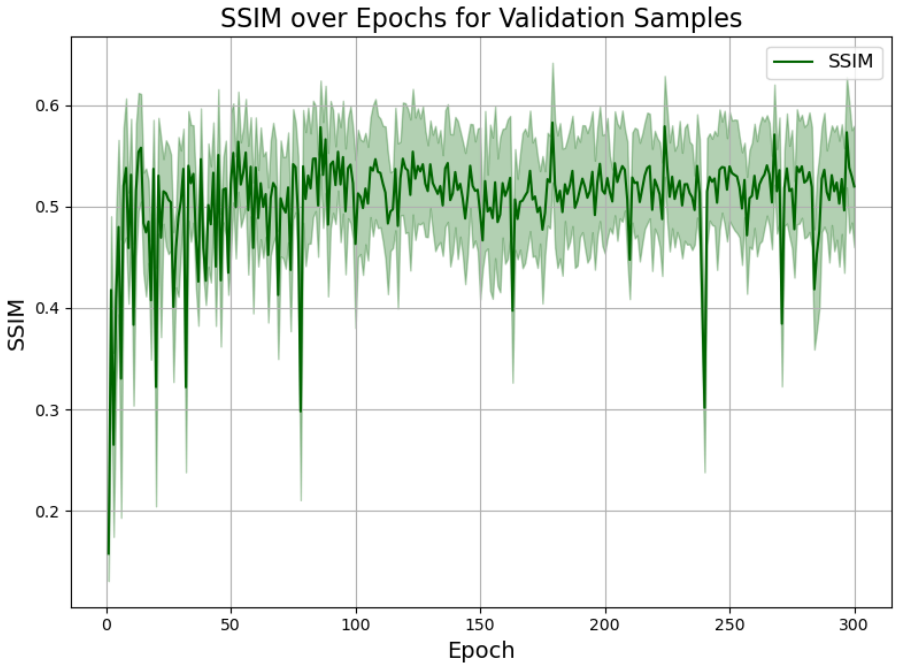

Figure 2: SSIM metric evaluated on the validation set and calculated for each epoch throughout the training process.

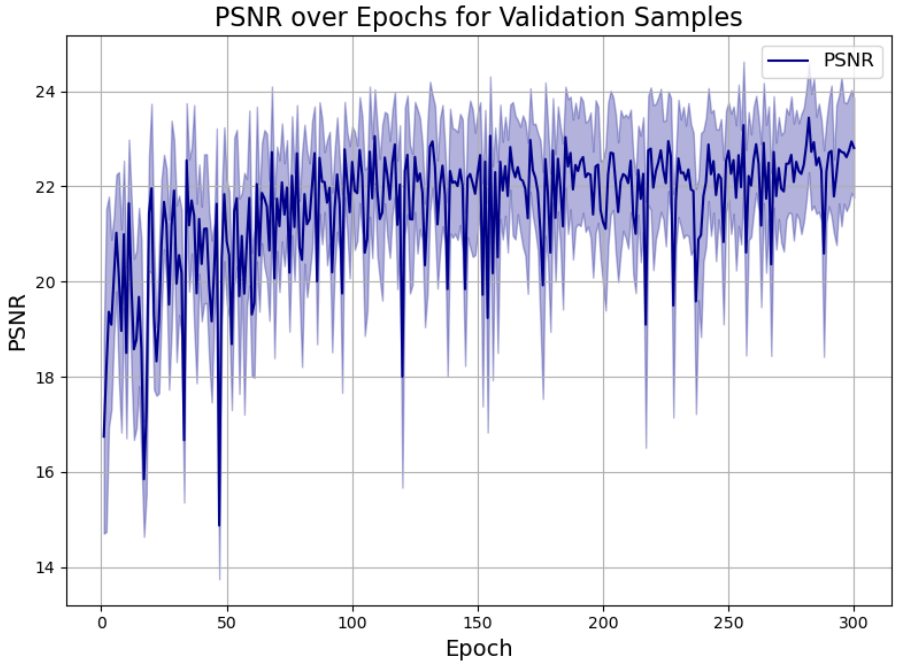

Figure 3: PSNR metric evaluated on the validation set and calculated for each epoch throughout the training process.

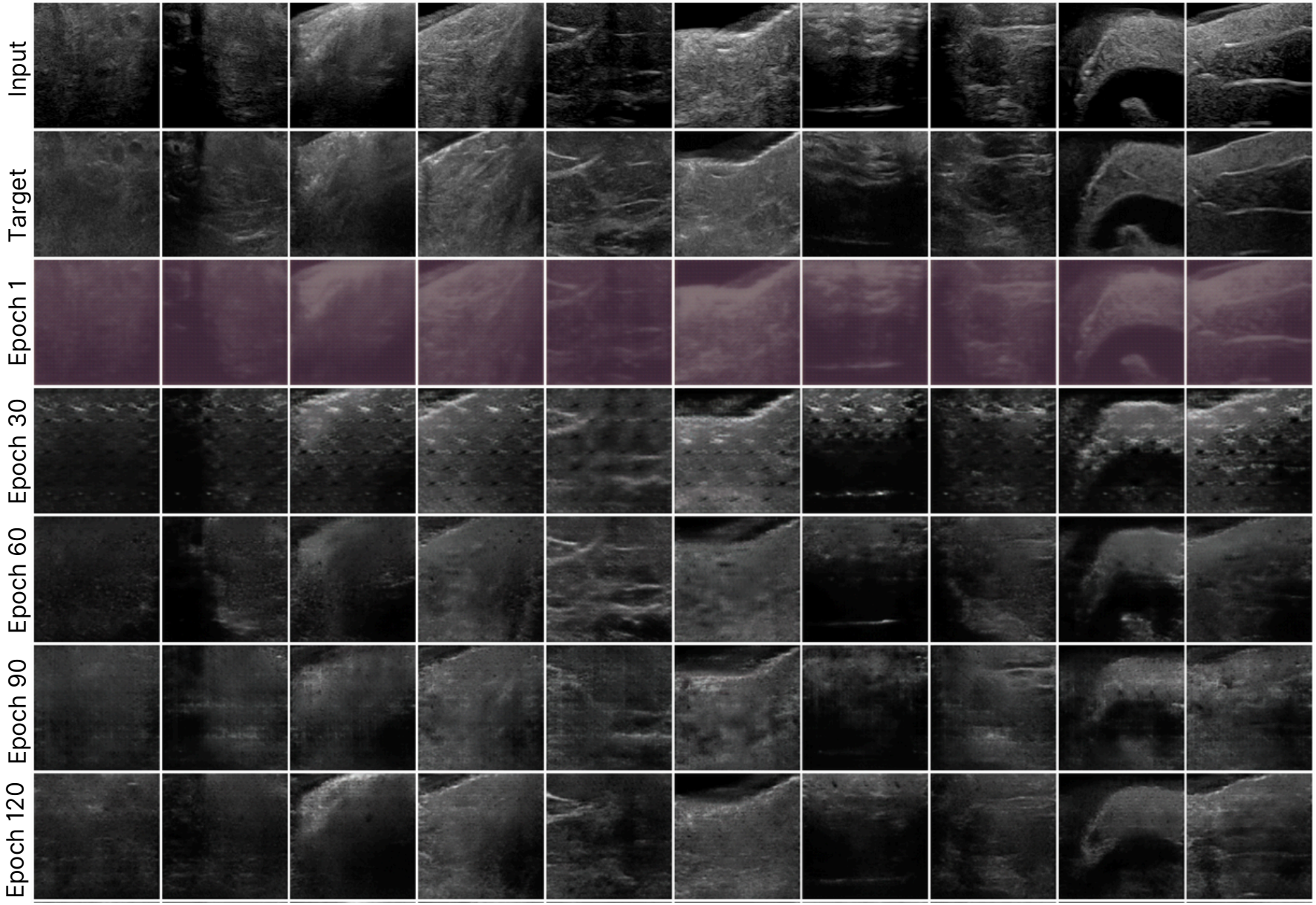

Figure 4: Generated validation images shown over different epochs.

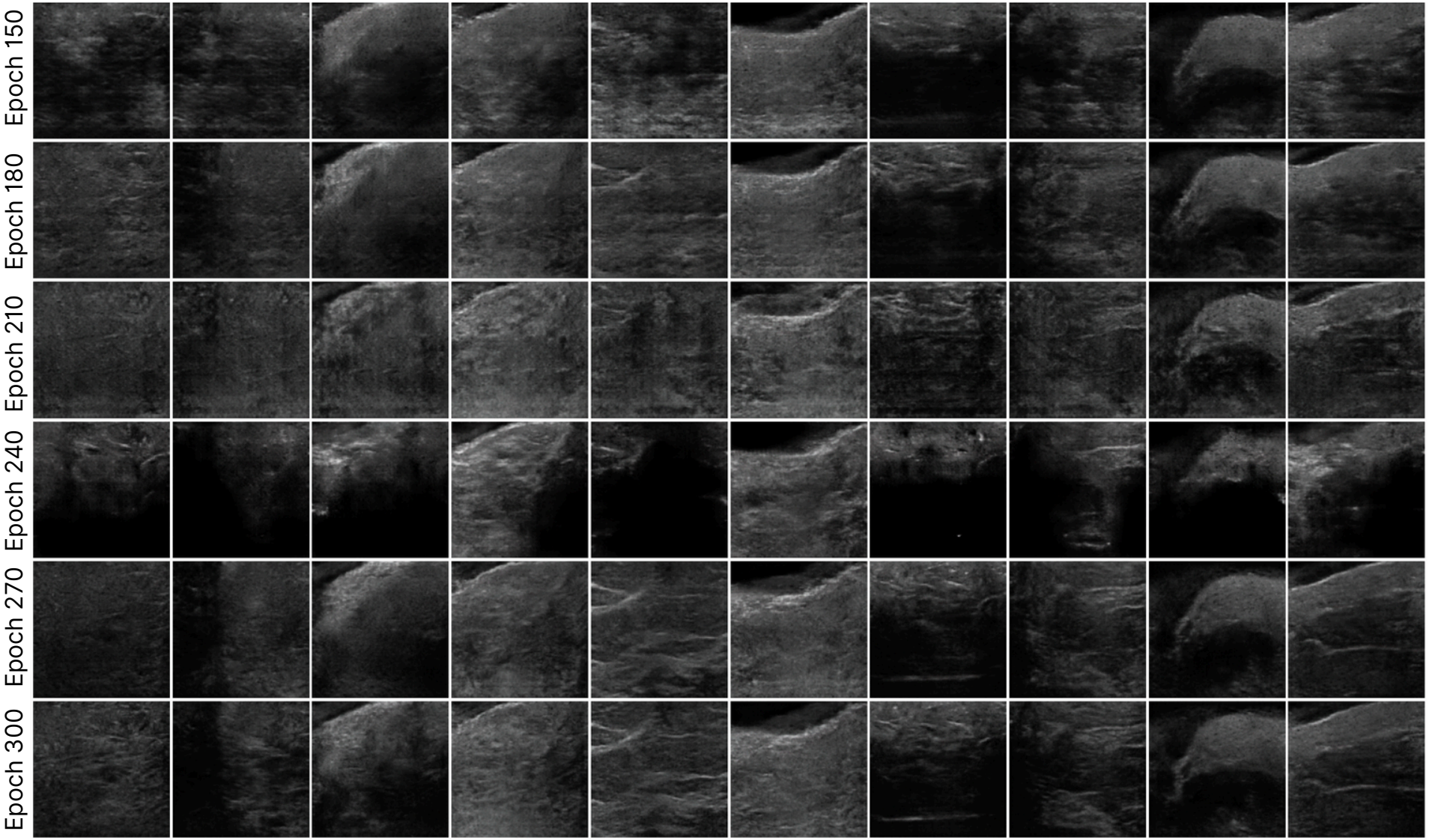

Figure 4: *Continued.*

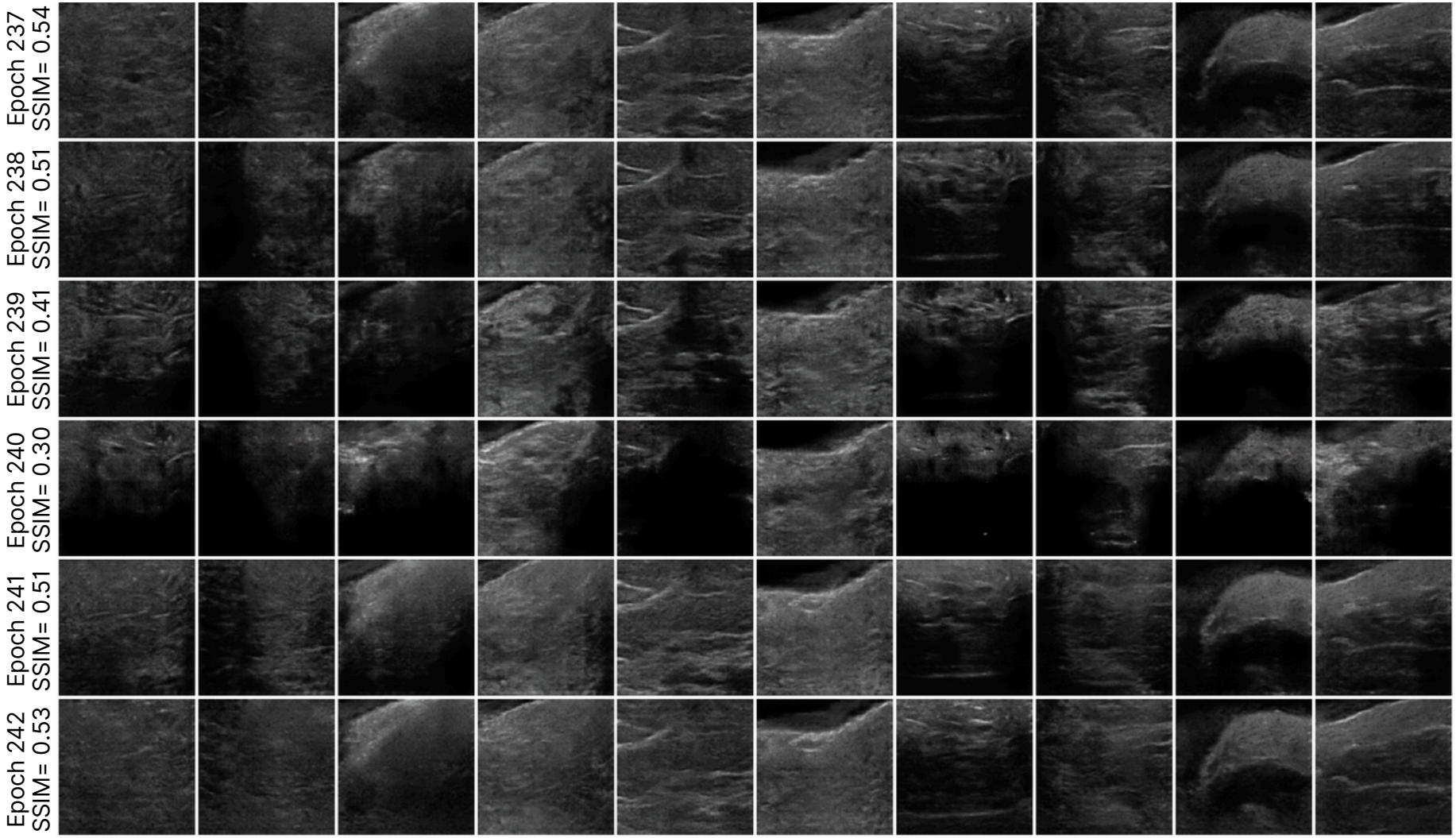

Figure 5: Generated validation images around performance drop seen in validation SSIM metrics at epoch 240.