A KL Divergence-Based Loss for In Vivo Ultrafast Ultrasound Image Enhancement with Deep Learning

Roser Viñals and Jean-Philippe Thiran

***Abstract*— Ultrafast ultrasound (US) imaging is a pioneering imaging modality that achieves higher frame rates than traditional US imaging, enabling the visualization and analysis of fast dynamics in tissues and flows. Nevertheless, images resulting from this technique suffer from a low-quality level. Recently, convolutional neural networks (CNN) have demonstrated great potential for reducing image artifacts and recovering speckle patterns without compromising the frame rate. As yet, CNNs have been mostly trained on large datasets of simulated or *in vitro* phantom images, but their performances on *in vivo* images remains suboptimal. In the current study, we present a method to enhance the image quality of single unfocused acquisitions by relying on a CNN. We introduce a training loss function that accounts for the high dynamic range of the radio frequency data and uses the Kullback–Leibler (KL) divergence to preserve the probability distributions of the echogenicity values. We conduct an extensive performance analysis of our approach using a new large *in vivo* dataset of 20,000 images. The predicted images are compared qualitatively to the target images obtained from the coherent compounding of 87 plane waves (PW). The structural similarity index measure, peak signal-to-noise ratio and KL divergence are used to quantitatively analyze the performance of our method. Our results demonstrate significant improvements in image quality of single PW acquisitions, highly reducing artifacts.**

***Index Terms*— Deep learning, Image reconstruction, Quality enhancement, Ultrafast ultrasound imaging**

I. Introduction

Ultrasound (US) imaging is widely used in medical imaging due to its real-time ability to produce high-quality images of soft tissues. In particular, a technique achieving frame rates of multiple kilohertz called ultrafast US has revolutionized US imaging. The high frame rates achieved by ultrafast US can be exploited to study fast changes in the human body and have enabled new imaging modalities such as shear-wave elastography, which analyses the tissues' viscoelasticity, or ultrafast Doppler imaging for flow imaging [1].

Traditional ultrasound uses focused beams to scan the imaging plane line by line, whereas ultrafast US transmits a single unfocused wavefront such as a diverging wave (DW) or a plane wave (PW) [1]. While focused beams concentrate energy in narrow beams, unfocused wavefronts disperse energy across the entire field of view. Consequently, imaging with unfocused beams yields lower-amplitude backscattered echoes and a lower signal-to-noise ratio (SNR), resulting in lower contrast. Contrast is also degraded by artifacts caused by grating lobes (GLs) and side lobes (SLs). Furthermore, ultrafast acquisitions suffer from lower lateral resolution due to broader main lobes of the point spread function, compared to line by line acquisitions.

A technique to improve the image quality of ultrafast US images is coherent plane wave compounding (CPWC). This strategy coherently compounds multiple images obtained from unfocused wavefronts steered at different angles, suffering from a trade-off between image quality, which is enhanced by increasing the number of compounded acquisitions, and frame rate, which is reduced [2]. Furthermore, coherent compounding assumes that during acquisition the region of interest is stationary. Consequently, images acquired on fast-moving areas might suffer from severe motion artifacts.

Several deep learning-based techniques have been proposed to enhance the image quality of ultrafast acquisitions. These approaches are intended to reduce the artifacts caused by GLs and SLs while preserving the speckle patterns, as they comprise positional information of the underlying physical phenomena. While [3]–[8] focus on enhancing the image quality of single PW acquisitions, others intend to improve the quality of the compounding of few PWs [9]–[11] or DWs [12], [13].

Most of these studies use convolutional neural networks (CNNs) that learn the mapping between an input image, acquired with one or a few unfocused acquisitions, and a target image resulting from the compounding of several unfocused acquisitions [4], [8], [9], [11]–[13]. Gasse *et al.* [9] enhanced the contrast ratio and lateral resolution of radio frequency (RF) images resulting from the compounding of 3 PWs by using 31 CPWCs as target images. Similarly, Lu *et al.* [12], [13] trained a neural network that used beamformed images acquired with 3 DWs tilted at different angles as input images,

This work was supported by the Swiss National Science Foundation under Grant 205320-207486. (Corresponding author: Roser Viñals)

Roser Viñals is with the Signal Processing Laboratory 5 (LTS5) at the École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland, (email: roser.vinalsterres@epfl.ch).

Jean-Philippe Thiran is with the Signal Processing Laboratory 5 (LTS5), École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland, also with the Department of Radiology, University Hospital Center (CHUV) and University of Lausanne (UNIL), 1011 Lausanne, Switzerland, and also with the CIBM Center for Biomedical Imaging, 1015 Lausanne, Switzerland, (email: jean-philippe.thiran@epfl.ch).

and the images formed by compounding 31 DWs as target images. RF images were used in [12], while in phase and quadrature (IQ) images were utilized in [13]. Perdios *et al.* [4] trained a CNN using as input and target the RF images simulated considering a single PW and synthetic aperture (SA), respectively. Jansen *et al.* [11] presented a deep learning-based reconstruction method in the Radon domain. Their method successfully improved the image quality of the images acquired with 3 PWs using the compounding of 51 PWs as target images. In [8], the authors used a CNN to improve the beamforming of single unfocused acquisitions by training a CNN using as input the RF images corresponding to single PWs steered at 0° and as target the IQ data resulting from compounding with 3 and 5 PWs.

Using focused acquisitions as target images has also been proposed [5], [10]. Zhou *et al.* [5] used a generative adversarial network (GAN) that used RF images acquired with 1 PW as input. Khan *et al.* [10] implemented a CycleGAN to enhance the B-mode image quality resulting from different numbers of compounded acquisitions: 3, 7, 11, and 31 PWs. Alternatively, recent studies have been conducted on the use of self-supervised learning for enhancing the quality of single unfocused acquisitions without target images [6], [7].

Acquiring large *in vivo* datasets to develop deep learning methods for ultrafast image improvement is a time-consuming and ethically regulated process. Consequently, most of the training datasets combine *in vitro* (phantom) and a limited number of *in vivo* acquisitions. Furthermore, the models are typically evaluated on acquisitions taken on the same phantoms used for training, hindering the assessment of their generalization capabilities [5], [7], [9], [10], [12], [13]. Only a few datasets contain exclusively *in vivo* data, such as the one used in [5] where a GAN was trained, tested, and evaluated using 360 pairs of RF data acquired on different body parts of 30 healthy volunteers.

Alternatively, simulated data has been used to develop these deep learning-based methods for image enhancement [4], [6], [8]. For instance, Zhang *et al.* [6] trained a network exclusively on simulated data generated with the Field II Ultrasound Simulation Program, without evaluating their performance on experimental acquisitions. Similarly, Lu *et al.* [8] trained a CNN using simulated data obtained with the Field II Ultrasound Simulation Program and phantom images. Their approach was evaluated on phantom and *in vivo* images. Finally, Perdios *et al.* [4] trained a CNN on simulated data using a spline-based Spatial Impulse Response (SIR) US simulator. The SIR simulation was used to generate a large dataset containing pairs of single unfocused and SA acquisitions, using simulated phantoms containing random ellipses of constant mean echogenicity. Their CNN-based reconstruction was tested on *in vitro* and a few *in vivo* frames.

Despite the advantages of training on simulated or *in vitro* data, some of these methods lack large-scale validation on *in vivo* data and an objective evaluation on phantoms not included in the training dataset. Therefore, in the current work, we present an improved deep learning image enhancement method for ultrafast ultrasound imaging based on the approach introduced in [4]. A detailed analysis of the method performance when trained and tested on a large collection of *in vivo* images is conducted.

Authors in [4] proposed a CNN-based US image reconstruction method that not only reduces artifacts and restores the speckle patterns of single ultrafast acquisitions but also can be used for displacement estimation [14]. Although this approach showed potential for recovering high-quality images from single unfocused acquisitions using simulated data, the quality improvement dropped significantly when applied to *in vivo* data due to the domain shift between *in vivo* and simulated data [4]. The objective of this work is to improve the performance of this approach on *in vivo* data, by reducing noise and artifacts from single *in vivo* PW acquisitions to achieve an image quality comparable to that of CPWC with 87 PWs. To accomplish this, the CNN described in [4] has been modified and trained on a large *in vivo* dataset.

This work introduces two significant contributions. Firstly, a novel loss function is proposed that effectively handles the high dynamic range of the RF images while preserving the probability distribution function of the echogenicity values. Secondly, a comprehensive *in vivo* dataset comprising 20,000 images is presented. This dataset has been used for training the CNNs and will be made available for public access along with this paper.

II. Methods

A. Dataset Acquisition and Preprocessing

A large dataset of 20,000 *in vivo* images acquired on different body parts has been collected from nine healthy volunteers, as outlined in Table I. The acquisitions were performed with the approval of the Cantonal Commission on Ethics in Human Research (2022-01696, CER-VD, Switzerland). An *in vitro* image was also acquired on the CIRS model 054GS phantom (CIRS, Norfolk, VA, USA) to assess the performance of our method and derive normalization matrices. The acquisitions were collected using the GE 9L-D linear array transducer (GE Healthcare, Chicago, Illinois, USA), a linear array transducer with 192 elements and a center frequency of 5.3 MHz, and the Vantage 256 system (Verasonics, Kirkland, WA, USA).

Each acquisition consisted of 87 PWs steered at different angles acquired at a pulse repetition frequency of 9 kHz. An alternating steering angle sequence [15] with a steering angle spacing of 0.38° was employed. The steering angle spacing and the number of steered acquisitions were determined such that the focusing quality was comparable to that of the optimal multi-focus, as described in [2] and [4], considering an F-number of 1.75. Time gain compensation was applied assuming a tissue attenuation of 0.5 dB/ (cm · MHz).

The ultrasound probe was moved before each measurement to ensure that each acquisition was distinct from the previous one. The maximum frame rate between two acquisitions was restricted to 47.5 Hz, maintaining an Intensity Spatial Peak Temporal Average (ISPTA) below the Food Drug Administration (FDA) recommended threshold of 94 mW/cm^2 [16]. The peak-to-peak voltage was set to 40V to ensure a Mechanical Index (MI) below 0.7, as recommended by the British Medical Ultrasound Society (BMUS) [17]. The imaging configuration and parameters used are specified in Table II.

TABLE I: Number of images, and mean and standard deviation of the echogenicity values of the dataset

	Number of images	Echogenicity (dB)	
		1 PW	87 PWs
Dataset	20,000	$4.65_{\pm 9.93}$	$-3.83_{\pm 12.28}$
Abdomen	6,599	$5.43_{\pm 9.38}$	$-3.08_{\pm 11.45}$
Carotids	3,294	$2.99_{\pm 10.25}$	$-5.77_{\pm 13.35}$
Breast	3,291	$4.28_{\pm 10.50}$	$-4.64_{\pm 12.96}$
Lower limbs	2,616	$6.34_{\pm 9.57}$	$-0.87_{\pm 11.40}$
Upper limbs	2,110	$3.99_{\pm 10.78}$	$-3.93_{\pm 13.03}$
Back	2,090	$3.93_{\pm 9.08}$	$-5.44_{\pm 11.27}$

TABLE II: Imaging configuration and acquisitions' parameters

Parameter	Value
Linear array transducer	GE 9L-D
Center frequency	5.3 MHz
Bandwidth (at -6 dB)	75%
Aperture	43.93 mm
Element number	192
Pitch	230 μm
Element width[1]	207 μm
Element height	6 mm
Elevation focus	28 mm
Transmit frequency	5.208 MHz
Excitation cycles	1
Sampling frequency	20.833 MHz
Number of compounded acquisitions	87
Steering angle spacing	0.38°
Pulse repetition frequency	9 kHz
Peak-to-peak voltage	40 V

[1] Estimated value

Ultrafast US imaging can be formulated as an inverse problem [18]. Let us consider the measurements $\mathbf{y} \in \mathbb{R}^N$, the measurement noise $\boldsymbol{\epsilon} \in \mathbb{R}^N$, the measurement model operator $\mathbf{H}\colon \mathbb{R}^M \to \mathbb{R}^N$, and the vectorized image that we want to estimate $\boldsymbol{\theta} \in \mathbb{R}^M$. Then, the inverse problem can be formulated as finding $\boldsymbol{\theta}$ such that $\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\epsilon}$.

Our reconstruction pipeline relies on the estimation of a solution to this inverse problem. This estimation is obtained following the method described in [4] with a backprojection-based DAS operator that has been implemented using PyUS [19], a GPU-accelerated Python package for US imaging. A $\lambda/8 \times \lambda/8$ grid with a width spanning the probe aperture and a depth from 1 mm to 55 mm has been considered, resulting in images of 1483 $\times$ 1189 pixels.

From each acquisition, we estimated two RF images. The first corresponds to the single unfocused acquisition obtained from the PW measurement steered at 0°, and it is referred to as the input image. The second results from coherently compounding the 87 PWs acquisitions steered at different angles and is referred to as the target or CPWC image.

Using 1,000 speckle image pairs acquired on the CIRS model 054G phantom, we computed two normalization matrices: one for the input and the other for the target images. We first beamformed the speckle images. Afterward, we detected the envelope and log-compressed the resulting images to generate the B-mode images. These B-mode speckle images were averaged, giving rise to a matrix of 1483 $\times$ 1189 values. By converting the B-mode average matrices to linear scale, we obtained the normalization matrices. These normalization matrices were applied to normalize all the RF images by dividing the RF images by them. The vectorized normalized RF image corresponding to the single unfocused acquisition is denoted as $\mathbf{x}_{\mathrm{1PW}} \in \mathbb{R}^M$, while the one corresponding to the target image is denoted as $\mathbf{x} \in \mathbb{R}^M$.

To evaluate the diversity of our datasets, the probability distributions of the B-mode values of the normalized images, $\mathbf{x}_{\mathrm{1PW}}$ and $\mathbf{x}$, have been analyzed. The mean and standard deviation of these distributions for both imaging modalities are presented in Table I. We observe that our images span a high dynamic range, which significantly varies across different imaged body areas. Furthermore, the single unfocused images tend to have higher echogenicity and a narrower range compared to the target images, leading to reduced contrast.

B. CNN Architecture and Training

Our CNN architecture is based on the U-Net architecture described in [4]. This architecture has previously demonstrated success in enhancing ultrafast ultrasound images by effectively mitigating artifacts from GLs and SLs when trained on simulated data. The main modification from the architecture presented in [4] is the replacement of the rectified linear unit activation functions with the Scaled Exponential Linear Unit (SELU) activation functions [20]. This activation function accelerates the convergence of the network. The resulting architecture is illustrated in Figure 1.

The network aims to learn the mapping $f\colon \mathbb{R}^M \to \mathbb{R}^M$ between $\mathbf{x}_{\mathrm{1PW}}$ and $\mathbf{x}$ in order to estimate higher-quality images, $\hat{\mathbf{x}}$, from the PWs steered at 0°: $\hat{\mathbf{x}} = f(\mathbf{x}_{\mathrm{1PW}})$. Thus, the CNN has been trained using as input images the estimated RF images corresponding to the PWs steered at 0°, $\mathbf{x}_{\mathrm{1PW}}$, and as target images, the estimated RF images resulting from the 87 PWs compounded acquisitions, $\mathbf{x}$.

To prevent the inclusion of similar images from the same volunteer in both the training and validation or test sets, a volunteer-based split is performed. Out of the 9 volunteers, 6 have been used for training, 1 for validation and 2 for testing. This split corresponds to 16,077 image pairs for training, 1,826 for validation and 2,097 for testing.

The network has been trained for 20 epochs using 16 channels and Adam optimizer [21] with a learning rate of 0.0003 and a weight decay of 0.005. The training batch size has been set to 16 and a random shuffle has been applied on every epoch. All these parameters' values have been optimized using Optuna [22], a software that implements a Bayesian optimization algorithm for hyperparameter tuning.

C. Training Losses

Due to the high dynamic range of our data, traditional losses such as mean absolute error and mean squared error are not suitable. To address this issue, authors in [4] introduced a log-compressed loss named MSLAE that showed a great potential to train networks with RF simulated images of high dynamic range. This loss can be expressed as follows:

$$L_{\mathrm{MSLAE}}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} ||g_\alpha(\mathbf{x}) - g_\alpha(\hat{\mathbf{x}})||_1, \qquad (1)$$

with

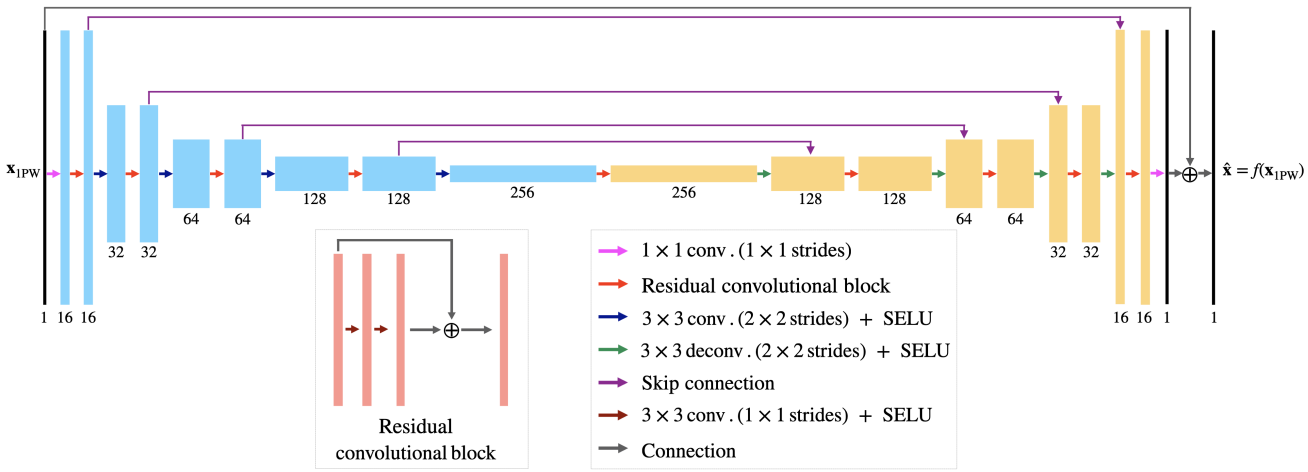

Fig. 1: Convolutional neural network architecture and the residual convolutional blocks considered. Arrows represent network layers and operations, while rectangles represent tensors with the number of channels specified below them.

$$g_\alpha(x_m) = \text{sign}(x_m) \log_\alpha \left(\frac{\alpha}{\max(\alpha, |x_m|)} \right) \tag{2}$$

where x_m denotes the pixel m of the vectorized image $\mathbf{x}$ and $\alpha \in (0, 1)$.

When using this loss with our *in vivo* dataset, the network tends to widen the echogenicities distribution and shift them to lower echogenicities.

A well-known measure to quantify the similarity between two probability distributions is the Kullback-Leibler (KL) divergence. It is a non-symmetric measure of the difference between two distributions. Let us consider two probability distributions $p(\mathbf{z})\colon \mathbb{R}^M \to \mathbb{R}^K$ and $q(\hat{\mathbf{z}})\colon \mathbb{R}^M \to \mathbb{R}^K$, with M and K denoting the number of samples and bins, respectively. Then, the KL divergence of $q(\hat{\mathbf{z}})$ from $p(\mathbf{z})$ is defined as

$$D_{KL}(p(\mathbf{z})||q(\hat{\mathbf{z}})) = \sum_{k=0}^{K} p(\mathbf{z})_k \ln \frac{p(\mathbf{z})_k}{q(\hat{\mathbf{z}})_k}, \tag{3}$$

where $p(\mathbf{z})_k$ and $q(\hat{\mathbf{z}})_k$ are the probability estimates of the k-th bin. To improve the performance of the image enhancement method, we introduce a new loss named KLD-MSLAE that aims to reduce diffraction artifacts while preserving the echogenicity distribution probabilities by combining MSLAE with the KL divergence. It is defined as follows:

$$\begin{aligned} L_{\text{KLD}-\text{MSLAE}}(\mathbf{x}, \hat{\mathbf{x}}) = L_{\text{MSLAE}}(\mathbf{x}, \hat{\mathbf{x}}) \\ + \beta D_{KL}(p(\mathbf{z})||q(\hat{\mathbf{z}})), \end{aligned} \tag{4}$$

where $\beta \in \mathbb{R}$ is a weighting factor, and $p(\mathbf{z})$ and $q(\hat{\mathbf{z}})$ denote the estimated probability distributions of $\mathbf{z} = 20\log_{10}(\max(\alpha, |\mathbf{x}|))$ and $\hat{\mathbf{z}} = 20\log_{10}(\max(\alpha, |\hat{\mathbf{x}}|))$, respectively.

The probability distributions $p(\mathbf{z})$ and $q(\hat{\mathbf{z}})$ have to be estimated so that the estimates are differentiable. We consider that our probability distributions span over the range $[-\alpha_{\text{dB}}, \alpha_{\text{dB}}]$, where $\alpha_{\text{dB}} = 20\log_{10}(\alpha)$, and we set the number of bins to K. Each bin k has a width of $\delta = 2\alpha_{\text{dB}}/K$ and is centered at $c_k = -\alpha_{\text{dB}} + (k + 0.5)\delta$, with $k = 0, ..., K$. Then, we can define $\Delta_{m,k} = z_m - c_k$. The probability distribution on the k-th bin, $p(\mathbf{z})_k$ can be approximated by

$$p(\mathbf{z})_k = \frac{\sum_{m=1}^{M} (s_\lambda(\Delta_{m,k} + \frac{\delta}{2}) - s_\lambda(\Delta_{m,k} - \frac{\delta}{2}))}{\sum_{n=1}^{N}\sum_{m=1}^{M} (s_\lambda(\Delta_{m,k} + \frac{\delta}{2}) - s_\lambda(\Delta_{m,k} - \frac{\delta}{2}))}, \tag{5}$$

with $s_\lambda(x) = 1/(1+e^{-\lambda x})$ denoting the logistic function with a growth rate of λ.

The accuracy of this probability distribution estimation increases with the number of bins K and the logistic function steepness λ. As we increase λ, the logistic function will approach a Heaviside step function, becoming less differentiable. In this work, we have set the number of bins K to 40, the logistic growth λ to 0.5 and the weighting factor β to 1.

In both components of the loss, the parameter α plays a key role. The components $g_\alpha(x_k)$ are zero for any RF value that satisfies $|x_k| < \alpha$. This threshold enables us to use a logarithmic loss without facing the vertical asymptote of the logarithmic function in 0 and prevents the network to learn from lower echogenicities than α. Similarly, the probability distributions estimated to compute the divergence only consider the range $[-\alpha_{\text{dB}}, \alpha_{\text{dB}}]$. Different α_{dB} values were used to train the network. By visually assessing the resulting images, we observed the best results are obtained with α_{dB} =-60 dB.

D. Performance Evaluation and Metrics

To evaluate the performance of our method, we first compare the output of the CNN to the target test images acquired with 87 PWs. Three metrics are considered: the structural similarity index measure (SSIM), the peak signal-to-noise ratio (PSNR), and the KL divergence. These metrics are computed between the B-mode images within the range of [-40 dB, 40 dB]. Furthermore, we calculate the means and standard deviations of the resulting echogenicity values.

The contrast (C) and the speckle patterns are assessed in selected regions of two test images. The contrast between two image areas is calculated on the envelop-detected images

following [23]. Specifically, the contrast between two designated areas, denoted as A and B, is computed in decibels as $C = 20 \cdot \log_{10}(\overline{s_A}/\overline{s_B})$. Here, $\overline{s_A}$ and $\overline{s_B}$ represent the mean values of the envelop-detected images in regions A and B, respectively.

For the assessment of speckle patterns, the signal-to-noise ratio (SNR) is calculated in selected homogeneous areas. The SNR is computed as the ratio of the mean value to the standard deviation: $\mathrm{SNR} = \overline{s_A}/\sigma_{s_A}$, where $\overline{s_A}$ and σ_{s_A} denote the mean and standard deviation of the amplitude of the envelop-detected image in the region A. For an ideal Rayleigh distribution, the expected SNR is 1.91 [24]. To further evaluate speckle patterns and their resolution, the FWHM of the axial and lateral dimensions of the 2-D autocovariance function (ACF) is computed within the same areas containing the speckle patterns [4], [25].

Our reconstruction method is also evaluated on an *in vitro* image taken on the CIRS model 054GS phantom. This image contains three inclusions with different contrasts: one anechoic inclusion and two low-echogenic inclusions with a contrast of -6 dB and -3 dB. All three inclusions are located at a depth of 40 mm and have a diameter of 8 mm. As with the two *in vivo* images, we compute the contrasts of these inclusions. We also assess the speckle patterns by computing the SNR and the lateral and axial FWHM of the 2-D ACF. This assessment is performed within selected areas exclusively containing speckle patterns.

III. Results

A. In Vivo

Our CNN has been trained using the MSLAE and KLD-MSLAE losses. Figure 2 shows the input, target, and output images of two acquisitions. The first row shows a carotid artery of one of the volunteers of the test set, while the second row shows an acquisition taken on the back of the other test volunteer.

The improvement in terms of the reduction of artifacts is noticeable using both losses. Particularly, this improvement can be clearly observed in the area outlined in yellow in the carotid images, where a large artifact is highly visible in the input image (Figure 2a), and the area delimited in red in the back image (Figure 2e). When zooming in on both areas, we can observe that the artifacts have been reduced and that some speckle patterns hidden or modified by the artifacts have been restored. To evaluate the restoration of speckle patterns, the SNR and the axial and lateral FWHM of the 2-D ACF have been computed in the areas delimited by yellow and red dotted lines. The resulting values are specified in Table III.

It is important to acknowledge that the target images might also be affected by artifacts, such as the SLs present in the region highlighted in magenta (Figure 2b). These SLs are partially attenuated but not entirely removed by the CNN, as shown in the magenta areas of Figure 2c and Figure 2d.

When using the MSLAE loss, the images exhibit increased contrast. Particularly, there is an over-attenuation of the low-echogenic areas, which is evident in the deeper area of Figure 2d. In contrast, the KLD-MSLAE loss attains a comparable contrast to the target images. To quantify this, the contrasts between the upper and lower areas delimited in magenta and blue dotted lines have been computed and are presented in Table III. To further analyze the discrepancies arising from training with the two different losses, Figure 3 presents the probability distributions of B-mode values for the input, target, and CNN's output images of the test set. It is evident that the CNN trained with the MSLAE loss causes the echogenicity distribution to widen and shift toward lower values. Conversely, training with the KLD-MSLAE loss enables the CNN to achieve a distribution of echogenicity closer to that of the target images.

The reconstructed B-mode images have been compared to the target images using the metrics SSIM, PSNR and KL divergence. Table IV presents the mean and standard deviation of these metrics across all test set acquisitions, along with the mean and standard deviation of the resulting echogenicity values. From these results, it is evident that the CNN, when trained with the KLD-MSLAE loss, enhances both the PSNR and SSIM with respect to the target images, in comparison to the CNN trained with the MSLAE loss. Furthermore, the KL divergence between the output and target images is also highly improved. A lower KL divergence indicates a higher similarity in echogenicity distributions and, consequently, a closer resemblance in contrast to the target images. The resemblance in echogenicity distributions can also be observed by analyzing the mean and standard deviation of the resulting echogenicity values. The CNN trained with KLD-MSLAE presents a mean and standard deviation closer to the target echogenicity values. In contrast, when trained with MSLAE, the resulting echogenicity values have a mean shifted towards lower values and a higher standard deviation compared to the target values.

B. In Vitro

The network trained on *in vivo* data has been applied to an *in vitro* phantom acquisition. Figure 4 shows the input, target, and CNN output images using the two losses. The regions where the contrasts have been calculated are marked with multiple concentric circles. The contrasts are calculated between the inner part of the smaller circles and the background areas between the two outer circles. The two low-echogenic inclusions with a contrast of -3 dB and -6 dB with respect to the background are highlighted in magenta and green, and the anechoic inclusion is indicated in blue. The speckle patterns are assessed in three regions highlighted in yellow by computing the SNR and the FWHM of the axial and lateral dimensions of the 2-D ACF. Table V summarizes the resulting metrics.

IV. Discussion

Our deep learning-based ultrafast ultrasound image enhancement method has proven to successfully reduce artifacts, leading to an improvement in the image quality of single unfocused acquisitions. The two *in vivo* examples demonstrate the CNN's capability to effectively mitigate artifacts on different body parts. The network achieves higher PSNR and

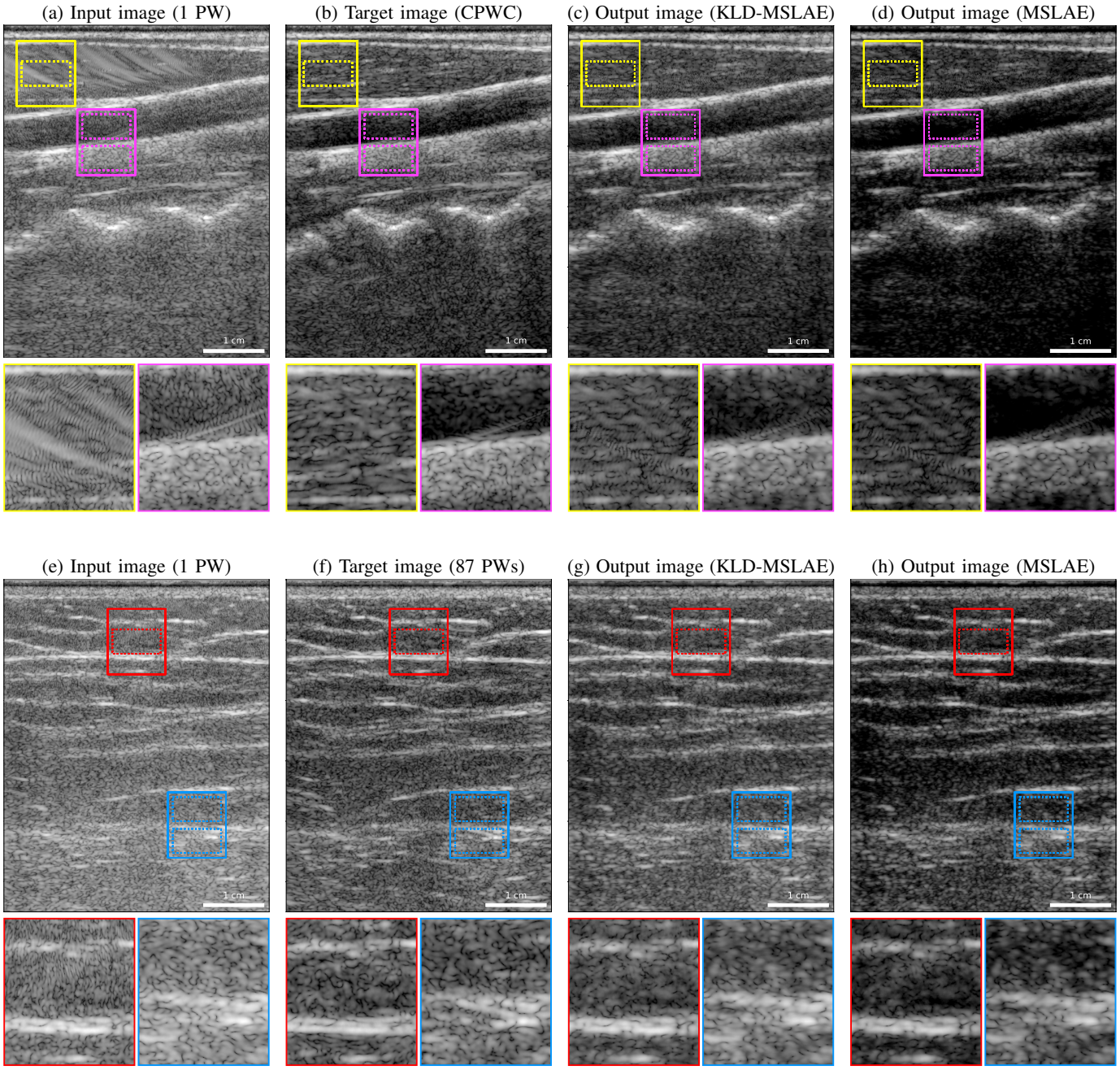

Fig. 2: B-mode images with a dynamic range of 65 dB (-25 to 40 dB) of the carotid (top row) and back (bottom row) of two test volunteers: (a) and (e) input images acquired with one plane wave (PW); (b) and (f) target images obtained from the coherent compounding with 87 PWs; (c) and (g) resulting images from the convolutional neural network (CNN) trained with the mean signed logarithmic absolute error (MSLAE) combined with the Kullback-Leibler divergence (KLD-MSLAE) loss; (d) and (h) resulting images from the CNN trained with the MSLAE loss.

TABLE III: Evaluation metrics computed on the highlighted areas of the two *in vivo* acquisitions

Metric	**C (dB)**	**C (dB)**	**SNR**	**FWHM$_{\mathrm{ACF_A}}$ (μm)**	**FWHM$_{\mathrm{ACF_L}}$ (μm)**	**SNR**	**FWHM$_{\mathrm{ACF_A}}$ (μm)**	**FWHM$_{\mathrm{ACF_L}}$ (μm)**
Target	-21.90	-15.83	1.261	254.15	542.29	0.838	302.12	580.19
Input	-15.28	-11.95	1.451	446.78	1120.27	1.134	260.01	222.95
KLD-MSLAE	-20.16	-16.49	1.436	242.43	474.97	0.789	296.63	764.88
MSLAE	-23.74	-18.88	1.377	248.29	524.63	0.654	337.65	1248.30

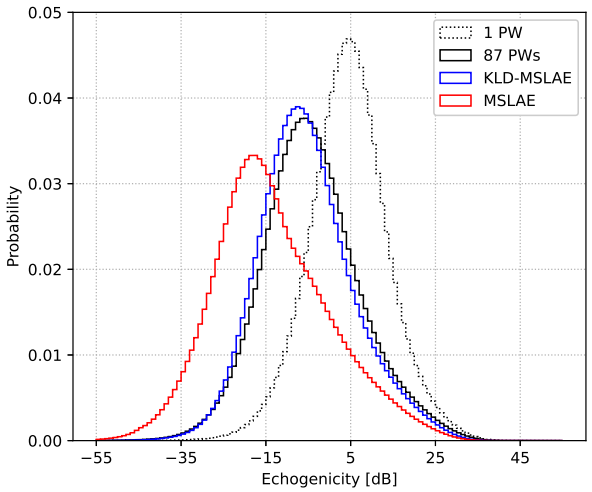

Fig. 3: Probability distributions of echogenicity values in the test set for input, target, and output images of the convolutional neural network (CNN). The CNN was trained using both the mean signed logarithmic absolute error (MSLAE) combined with Kullback-Leibler divergence (KLD-MSLAE) loss, and the standalone MSLAE loss.

TABLE IV: Evaluation metrics computed on the *in vivo* test set

Metric	PSNR	SSIM	KL divergence	Echogenicity (dB)
Target	-	-	-	-4.18±11.64
Input	16.466±0.801	0.105±0.060	16.466±0.090	4.48±9.44
KLD-MSLAE	20.292±0.307	0.272±0.040	20.292±0.015	-5.41±11.25
MSLAE	16.196±1.008	0.179±0.036	16.196±0.092	-13.65±13.16

SSIM between the output and target images, compared to those between the input and target images.

Furthermore, by adopting the KLD-MSLAE loss, we achieve an overall enhancement in terms of SSIM and PSNR. The KL divergence component of the loss helps to attain a contrast and echogenicity distribution similar to the target images. In contrast, the MSLAE loss shifts the echogenicity values to lower levels and spans them to a wider range. This induces a higher contrast, specially visible in anechoic regions and greater depths. The fact that MSLAE achieves higher contrast than KLD-MSLAE is further corroborated by analyzing the computed contrasts within the highlighted magenta and blue regions. In both areas, the CNN trained with the MSLAE loss consistently achieves higher contrasts than when trained with the KLD-MSLAE loss, surpassing the intended target contrasts. Note that both losses yield contrasts closer to those of the target images than the single unfocused input images.

Two specific regions, highlighted in yellow and red, that exhibit artifacts that hide or alter the speckle patterns have been analyzed. Upon visual assessment, we can observe that the CNN recovers speckle patterns that resemble more to those in the target images when contrasted with the original regions on the single PW images. In the area indicated in yellow of the carotid image, the achieved SNR and the FWHM of the 2-D ACF in both dimensions closely approach the target values, being the FWHM slightly lower. Nevertheless, within the red region of the back image, the FWHM of the ACF in the lateral dimension significantly exceeds the target value. This disparity is particularly evident when the CNN is trained with the MSLAE loss. Note that the speckle patterns of this specific region of the input image are highly altered by artifacts, leading to an increase in their resolution and rendering them significantly distinct from the speckle patterns in the target image. Despite the increase in the lateral FWHM of the ACF, the region restored by the CNN is much similar to the target one than those in the input image. It is worth mentioning that in both regions and in both dimensions, training with the KLD-MSLAE loss results in a lower FWHM of the 2-D ACF compared to training with MSLAE.

While there is a clear improvement in *in vivo* data in terms of contrast and artifacts removal, this improvement does not extend to the *in vitro* phantom image. This disparity could arise from the domain gap between the *in vitro* data and the training dataset, which comprises vastly different structures and artifacts compared to those present in the *in vitro* image.

When visually assessing the *in vitro* image, we can observe that the CNN produces images of lower echogenicity, specially when trained with the MSLAE loss. Regardless of the loss used for training, the CNN produces higher contrasts on the two low-echogenic inclusions, surpassing those observed in the target images. Conversely, the contrast of the anechoic inclusion is lower than the target value when the CNN is applied, although representing an enhancement compared to the input image. As observed in the *in vivo* images, the contrasts in the CNN's output images, trained with the MSLAE loss, exceed those achieved when trained with the KLD-MSLAE loss. This fact can be attributed to the widening effect observed in the echogenicity distribution when training with MSLAE.

To assess the preservation of speckle patterns, the SNR and the FWHM in both axial and lateral dimensions of the ACF have been computed for three areas containing only speckle patterns. In terms of SNR, when trained with the KLD-MSLAE loss, the CNN slightly improved the SNR. By contrast, training with MSLAE led to a significantly lower SNR compared to the target. Furthermore, regardless of the loss used, the FWHMs of the ACF, especially in the lateral dimension, exceed the desired values, indicating that the resolution of the speckle patterns in the phantom image is penalized. Notably, the KLD-MSLAE achieves lower FWHM in both dimensions compared to MSLAE, suggesting a better speckle preservation.

Despite the promising results, our approach has two main limitations that need to be addressed. These limitations arise from training the CNN exclusively using *in vivo* data. Firstly, artifacts similar to those present in single unfocused acquisitions also appear in the CPWC target images. This restricts the overall quality improvement that the CNN can achieve, as the target images themselves have inherent limitations. Therefore, while our network successfully reduces artifacts, complete eliminate remains challenging.

Secondly, part of our dataset consists of data acquired from body parts with a shallow depth, where deep regions contain only noise. In addition, our echogenicity values follow a Gaussian-shaped distribution, containing only a few samples

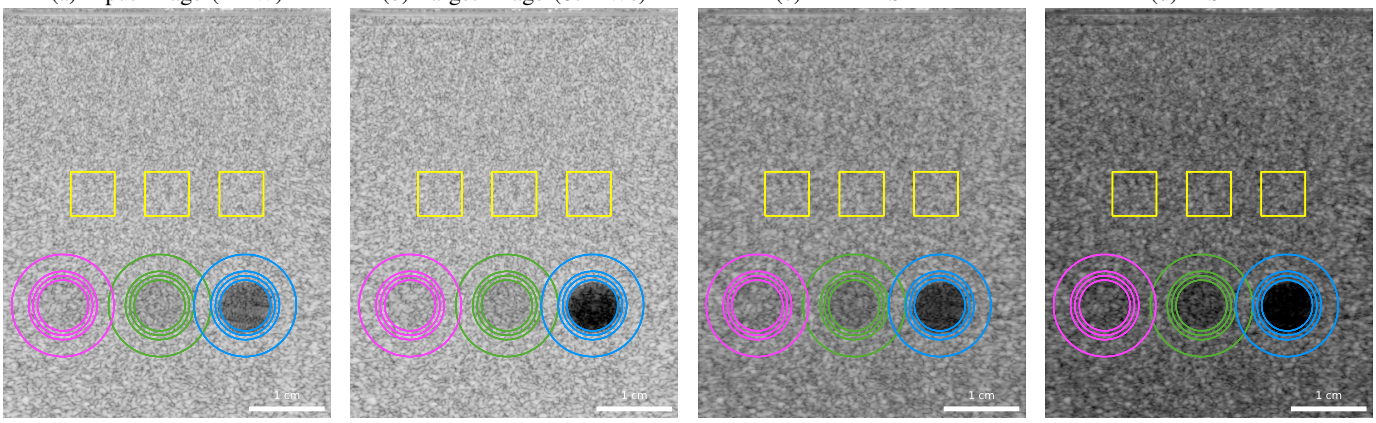

Fig. 4: B-mode images with a dynamic range of 65 dB (-45 to 20 dB) of an *in vitro* acquisition containing two low-echogenic inclusions and an anechoic inclusion: (a) input image acquired with one plane wave (PW); (b) target image obtained from the coherent compounding with 87 PWs; (c) resulting image from the convolutional neural network (CNN) trained with the mean signed logarithmic absolute error (MSLAE) combined with the Kullback-Leibler divergence (KLD-MSLAE) loss; (d) resulting image from the CNN trained with the MSLAE loss.

TABLE V: Evaluation metrics computed on the *in vitro* acquisition

Metric	**C (dB)**	**C (dB)**	**C (dB)**	**SNR**	**FWHM$_{\mathbf{ACF_A}}$ (μm)**	**FWHM$_{\mathbf{ACF_L}}$ (μm)**
Target	-3.00	-6.27	-28.35	$1.884_{\pm 0.039}$	$244.38_{\pm 5.00}$	$235.93_{\pm 8.45}$
Input	-3.18	-6.00	-18.33	$1.911_{\pm 0.024}$	$239.97_{\pm 4.24}$	$239.85_{\pm 8.05}$
KLD-MSLAE	-3.60	-7.38	-20.31	$1.895_{\pm 0.009}$	$287.23_{\pm 7.63}$	$365.10_{\pm 6.68}$
MSLAE	-4.69	-9.13	-24.04	$1.658_{\pm 0.005}$	$297.14_{\pm 6.70}$	$405.14_{\pm 7.08}$

for very low or very high echogenicities. Consequently, the network encounters challenges in learning from the extreme echogenicity values.

In contrast, these limitations were not present when using simulated data, as shown in [4]. Firstly, some of their target images were obtained after oversampling the transducer aperture, resulting in images with reduced GLs and higher quality target images compared to ours. Secondly, their dataset was simulated with phantoms containing random ellipsoidal inclusions of uniformly distributed mean echogenicity in the range of -50dB and +30dB with respect to the background, resulting in a wider range of echogenicities with a more uniform distribution. Therefore, all echogenicities were better represented in their simulated dataset.

To tackle these constraints, future studies could explore using transfer learning from simulated to *in vivo* data. This could help the network to generalize from simulated to *in vivo* data, leading to enhanced image quality and a reduction of the number of *in vivo* acquisitions required to train the network.

V. Conclusion

Ultrafast ultrasound achieves high frame rates but at the expense of image quality. Training a CNN on a large dataset of simulated images has been previously proposed to enhance image quality. However, the domain shift between *in vivo* and simulated images hindered CNN performances in practice.

To overcome this challenge, we developed a deep learning-based method for enhancing single unfocused acquisitions. This method was trained and tested on a large *in vivo* dataset. To further enhance the performance of the method, we introduced a novel loss function named KLD-MSLAE. This loss outperforms MSLAE and accounts both for the high dynamic range of RF images and the echogenicity's distribution.

Our approach significantly yielded a substantial enhancement in image contrast and highly reduced artifacts in single unfocused *in vivo* acquisitions acquired in different body parts. The CNN resulted in higher PSNR and SSIM between the output and target images. Further enhancement in image quality was achieved through the adoption of the KLD-MSLAE loss, resulting in a contrast and echogenicity distribution similar to the target images. Nevertheless, the image quality enhancement was not observed when applied to the *in vitro* image.

Despite our artifact-reduction method showing promising results, its performance remains constrained by the quality of the target images and the distribution of values within the dataset. Previous studies using simulated data have encountered fewer limitations, attributed to the availability of higher-quality target images and datasets with more diverse echogenicity values. Therefore, future studies could consider using transfer learning from simulated to *in vivo* data, thereby offering the potential for further advancements in the quality enhancement of ultrafast US images.

Acknowledgment

The authors would like to thank Dimitris Perdios for his insightful comments on the manuscript and Paolo Motta for his contributions and study on the loss.

REFERENCES

[1] M. Tanter and M. Fink, "Ultrafast imaging in biomedical ultrasound." *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 61, no. 1, pp. 102–119, jan 2014.

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

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

[4] D. Perdios, M. Vonlanthen, F. Martinez, M. Arditi, and J.-P. Thiran, "CNN-Based Image Reconstruction Method for Ultrafast Ultrasound Imaging," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 4, pp. 1154–1168, 2022.

[5] Z. Zhou, Y. Wang, Y. Guo, X. Jiang, and Y. Qi, "Ultrafast Plane Wave Imaging with Line-Scan-Quality Using an Ultrasound-Transfer Generative Adversarial Network," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 4, pp. 943–956, apr 2020.

[6] J. Zhang, Q. He, Y. Xiao, H. Zheng, C. Wang, and J. Luo, "Self-supervised learning of a deep neural network for ultrafast ultrasound imaging as an inverse problem," in *2020 IEEE International Ultrasonics Symposium (IUS)*. IEEE, 2020, pp. 1–4.

[7] ——, "Ultrasound image reconstruction from plane wave radio-frequency data by self-supervised deep neural network," *Medical Image Analysis*, vol. 70, may 2021.

[8] J. Y. Lu, P. Y. Lee, and C. C. Huang, "Improving Image Quality for Single-Angle Plane Wave Ultrasound Imaging With Convolutional Neural Network Beamformer," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 4, pp. 1326–1336, apr 2022.

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

[10] S. Khan, J. Huh, and J. C. Ye, "Variational Formulation of Unsupervised Deep Learning for Ultrasound Image Artifact Removal," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 6, pp. 2086–2100, jun 2021.

[11] G. Jansen, N. Awasthi, H. M. Schwab, and R. Lopata, "Enhanced Radon Domain Beamforming Using Deep-Learning-Based Plane Wave Compounding," *IEEE International Ultrasonics Symposium, IUS*, 2021.

[12] J. Lu, F. Millioz, D. Garcia, S. Salles, and D. Friboulet, "Fast Diverging Wave Imaging Using Deep-Learning-Based Compounding," *IEEE International Ultrasonics Symposium, IUS*, vol. 2019-Octob, pp. 2341–2344, oct 2019.

[13] J. Lu, F. Millioz, D. Garcia, S. Salles, D. Ye, and D. Friboulet, "Complex Convolutional Neural Networks for Ultrafast Ultrasound Imaging Reconstruction From In-Phase/Quadrature Signal," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 2, pp. 592–603, feb 2022.

[14] D. Perdios, M. Vonlanthen, F. Martinez, M. Arditi, and J.-P. Thiran, "CNN-Based Ultrasound Image Reconstruction for Ultrafast Displacement Tracking," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, p. 1078–1089, Mar 2021.

[15] B. Denarie, T. A. Tangen, I. K. Ekroll, N. Rolim, H. Torp, T. Bjåstad, and L. Lovstakken, "Coherent plane wave compounding for very high frame rate ultrasonography of rapidly moving targets," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1265–1276, 2013.

[16] Food, Drug Administration, U.S. Department of Health, and Human Services, "Marketing clearance of diagnostic ultrasound systems and transducers," feb 2023. [Online]. Available: https://www.regulations.gov/docket/FDA-2017-D-5372

[17] Safety Group of the British Medical Ultrasound Society, "Guidelines for the safe use of diagnostic ultrasound equipment," *Ultrasound*, vol. 18, no. 2, pp. 52–59, 2010.

[18] A. Besson, D. Perdios, F. Martinez, Z. Chen, R. E. Carrillo, M. Arditi, Y. Wiaux, and J.-P. Thiran, "Ultrafast ultrasound imaging as an inverse problem: Matrix-free sparse image reconstruction," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 65, no. 3, pp. 339–355, 2017.

[19] "Pyus: a gpu-accelerated python package for ultrasound imaging," https://gitlab.com/pyus/pyus, 2020.

[20] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

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

[22] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[23] S. W. Smith, H. Lopez, and W. J. Bodine, "Frequency independent ultrasound contrast-detail analysis," *Ultrasound in Medicine I& Biology*, vol. 11, no. 3, pp. 467–477, 1985.

[24] C. B. Burckhardt, "Speckle in ultrasound b-mode scans," *IEEE Transactions on Sonics and Ultrasonics*, vol. 25, no. 1, pp. 1–6, 1978.

[25] R. Wagner, S. Smith, J. Sandrik, and H. Lopez, "Statistics of speckle in ultrasound b-scans," *IEEE Transactions on Sonics and Ultrasonics*, vol. 30, no. 3, pp. 156–163, 1983.

Roser Viñals (Member, IEEE) received the B.Sc. degree in Telecommunication Engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2016. In 2020, she received an M.Sc. degree in Communication Systems from the École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. Since 2021, she is pursuing a Ph.D. degree in Electrical Engineering at the Signal Processing Laboratory 5 (LTS5) at EPFL, under the supervision of Prof. Jean-Philippe Thiran.

She was a visiting Researcher at the Broadband Wireless Networking Lab, Georgia Institute of Technology (Georgia Tech), Atlanta, United States, and at the Cardiovascular Magnetic Resonance group, Eidgenössische Technische Hochschule Zürich (ETHZ), Switzerland, in 2016 and 2020, respectively. She worked as a Research Intern at Telefónica, Barcelona, Spain, in 2017, and at Philips Research Suresnes, Paris, France, in 2019. From 2020 to 2021, she worked as Research Engineer at LTS5, EPFL. Her current research focuses on ultrasound image reconstruction, inverse models, and deep learning.

Jean-Philippe Thiran (Senior Member, IEEE) was born in Namur, Belgium, in 1970. He received the M.Sc. degree in electrical engineering and the Ph.D. degree from the Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium, in 1993 and 1997, respectively. He joined the École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 1998. He is currently a Full Professor at EPFL and Director of the Signal Processing Laboratory (LTS5).

Since June 2021, he has been the Director of the Institute of Electrical and Micro Engineering, EPFL. His research field is computational imaging, with applications in many domains, including medical image analysis (diffusion magnetic resonance imaging (MRI), ultrasound imaging, and digital pathology) and computer vision. He is currently a part-time Associate Professor with the Department of Radiology, University Hospital Center (CHUV) and University of Lausanne (UNIL), Lausanne. He is the author or the coauthor of one book, nine book chapters, 280 journal articles, and more than 290 peer-reviewed papers published in the proceedings of international conferences. He holds 12 international patents.

Dr. Thiran is a fellow of the European Association for Signal Processing (EURASIP). Among many other duties, he has been the General Chairperson of the 2008 European Signal Processing Conference (EUSIPCO 2008) and the Technical Co-Chair of the 2015 IEEE International Conference on Image Processing (IEEE ICIP 2015). From 2001 to 2005, he was Co-Editor-in-Chief of the Signal Processing international journal (published by Elsevier Science). He has been an Associate Editor of the IEEE Transactions on Image Processing.