Quantitative Assessment of Thyroid Nodules through Ultrasound Imaging Analysis

Young-Min Kim[1], Myeong-Gee Kim[1], Seok-Hwan Oh[1], Guil Jung[1], Hyeon-Jik Lee[1], Sang-Yun Kim[1], Hyuk-Sool Kwon[2], Sang-Il Choi[3], and Hyeon-Min Bae[1]

[1] School of Electrical Engineering Korea Advanced Institute of Science and Technology, Daejeon, South Korea
{youngmin2007,hmbae}@kaist.ac.kr
[2] Department of Emergency Medicine, Seoul National University Bundang Hospital, Seong-nam, South Korea
[3] Department of Radiology, Seoul National University Bundang Hospital, Seong-nam, South Korea

Abstract. Recent studies have proposed quantitative ultrasound (QUS) to extract the acoustic properties of tissues from pulse-echo data obtained through multiple transmissions. In this paper, we introduce a learning-based approach to identify thyroid nodule malignancy by extracting acoustic attenuation and speed of sound from ultrasound imaging. The proposed method employs a neural model that integrates a convolutional neural network (CNN) for detailed local pulse-echo pattern analysis with a Transformer architecture, enhancing the model's ability to capture complex correlations among multiple beam receptions. B-mode images are employed as both an input and label to guarantee robust performance regardless of the complex structures present in the human neck, such as the thyroid, blood vessels, and trachea. In order to train the proposed deep neural model, a simulation phantom mimicking the structure of human muscle, fat layers, and the shape of the thyroid gland has been designed. The effectiveness of the proposed method is evaluated through numerical simulations and clinical tests.

Keywords: Quantitative Ultrasound Imaging · Medical Ultrasound · Deep Neural Network · Transformer · Quantitative Imaging.

1 Introduction

Quantitative ultrasound (QUS) imaging, developed to extract tissue characteristics due to pathological changes, plays a crucial role in early disease detection, potentially improving diagnostic accuracy and patient outcomes. Recent advancements in deep neural networks (DNNs) have shown remarkable potential in generating QUS image, with studies demonstrating significant improvements in image quality and diagnostic reliability [**?**,**?**,**?**]. However, the majority of existing DNN-based QUS relying on fully convolutional neural networks (FCNs) leaves room for improvement since FCNs are known to be ineffective in capturing

global correlations. In addition, most commonly used multi-scanline transmission (MST) scheme that generates a series of focused beams through sequential excitation of grouped transducers makes it even difficult to extract of wide range correlations. In order to improve global correlations, Transformer model based deep neural networks are proposed [**?**,**?**]. Transformer model extracts the global range correlation while maintaining computational efficiency through the use of patch embedding and multi-head attention mechanisms.

Preserving consistent quantitative imaging performance, irrespective of the structural complexity of the target tissue, is a critical area of research in QUS. Oh et al. [**?**] proposed a feed-forward neural style transfer framework that utilizes B-mode image as an additional input, achieving improved image reconstruction by leveraging structurally reliable B-mode information. However, the style transfer framework requires additional supervisory layers to ensure the quantitative maps generated by the network accurately reflects the structural intricacies observed in B-mode images. Hence, the style transfer module trained on simulation phantoms can experience performance degradation when applied in the real-world scenarios.

For training QUS networks in a data-driven manner, it is critical to have a simulation dataset that can encompass a wide range of data distributions and produce the necessary volume of data. For breast [**?**,**?**] and liver [**?**] QUS imaging applications, multiple elliptical objects placed randomly are used as simulation dataset. However, such simple overlapping ellipses cannot adequately represent complex structures of the human neck such as the thyroid gland, trachea, and arteries. Moreover, due to the structural consistency among individuals, there is a need for a simulation phantom that accurately reflects the anatomy.

In this paper, we introduce QIT-net, a Quantitative Imaging technique for Thyroid assessment. QIT-net quantifies acoustic attenuation (ATT) and speed of sound (SOS) to identify the malignancy of a thyroid nodule. We propose a CNN-Transformer hybrid architecture that combines the strengths of both CNNs and Transformers. The CNN interprets RF signals locally, while the Transformer identifies correlations among multiple beam patterns using a hierarchical encoder. Additionally, the network utilizes B-mode image and QUS maps to enhance its accuracy in reconstructing ATT and SOS of tissues. To train QIT-net, we implemented two types of datasets: an elliptical object simulation phantom representing general soft tissue and a simulation phantom representing the shape and acoustic properties of the human thyroid gland.

2 Methods

The proposed system is modeled based on the Vantage 64LE (Verasonics Inc.), operated in conjunction with a 5MHz linear array probe (Humanscan Inc.). The probe consists of 128 transducer elements with a pitch size of 0.30mm. Ultrasound RF signals are obtained using MST beamforming with the transmission window size of 32, focused at 30mm depth.

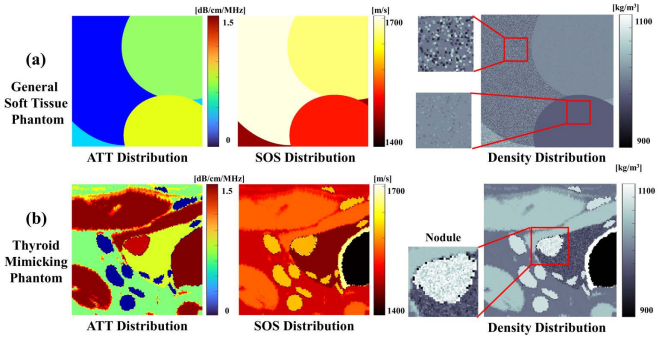

Fig. 1. Two types of training dataset

2.1 Simulation Phantoms

For data-driven training, we utilized the strengths of two distinct datasets as shown in Fig. 1. The general soft tissue phantom is design to replicate broad diversity of the acoustic properties of the typical tissue (see Fig. 1-(a)). The thyroid-mimicking phantom represents the unique anatomy and acoustic characteristics of the human thyroid, thereby enhancing the accuracy of thyroid nodule quantification (see Fig. 1-(b)). Both simulation phantoms are modeled with dimensions of 45mm $\times$ 45mm and discretized on an 1800 $\times$ 1800 grid. For simulations, in-silico model is created using an ultrasound simulation MATLAB toolbox, k-wave [**?**].

General Soft Tissue Phantom General soft tissue phantom comprises up to five ellipses with radii ranging from 2 to 30mm, randomly positioned within the Region of Interest (ROI), as shown in Fig. 1 (a). The acoustic properties of the soft tissue are modeled as follows [**?**]: the acoustic attenuation coefficient varies from 0 dB/cm/MHz to 1.5 dB/cm/MHz, the speed of sound ranges from 1400 m/s to 1700 m/s, and the density varies from 0.9 kg/m^3 to 1.1 kg/m^3. Additionally, a square speckle, ranging in size from 25 to 150 μm, is added with a random concentration of 0-10 /wavelength^2. In total, 14k general soft tissue phantoms are simulated for training the QIT-net.

Thyroid-Mimicking Phantom The overall procedure for designing the thyroid-mimicking phantom is depicted in Fig. 2. The anatomical structure of the human neck is obtained from the Visible Human Project dataset [**?**], which includes cross-sectional images captured from real human bodies (see Fig. 2-1). Manual segmentation of the thyroid gland, cartilage, trachea, and blood vessels is performed on 91 cross-sectional images. Muscle and fat layers are categorized into five groups based on their brightness levels (see Fig. 2-2). Elliptical objects representing thyroid nodules are then inserted within the thyroid gland (see Fig. 2-3).

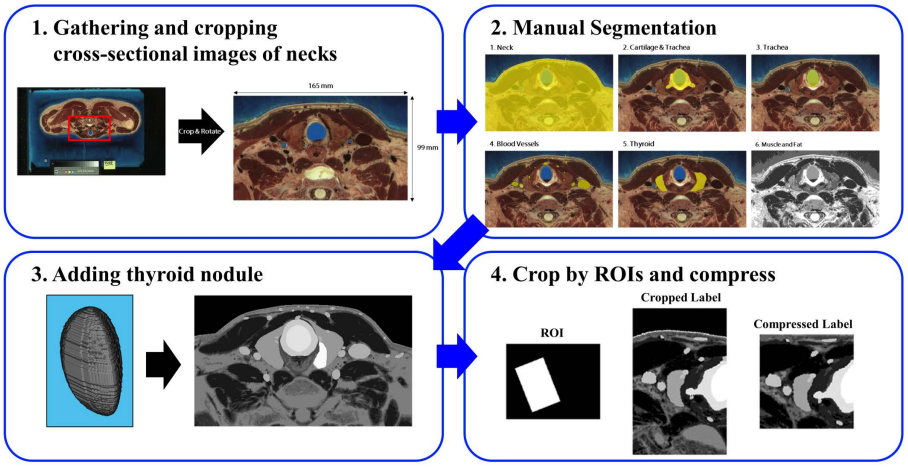

Fig. 2. Overall procedure for thyroid-mimicking phantom design

Table 1. Acoustic properties of tissues in thyroid-mimicking phantom

	Speed of Sound [m/s]	Attenuation Coefficient [dB/cm/MHz]	Scatter Diameter [μm]	Scatter Concentration [#/lambda^2]
Blood	1550 - 1600	0.03 - 0.2	25 - 100	0 - 2
Fat	1400 - 1500	0.3 - 0.8	25 - 125	1.44 - 25
Muscle	1530 - 1650	0.8 - 1.5	50 - 150	2 - 36
Thyroid	1450 - 1550	0.8 - 1.5	200 - 375	4 - 10
Nodule	1400 - 1700	0.03 - 1.6	25 - 375	1 - 10
Trachea	340 - 360	5 - 10	25 - 150	0
Cartilage	1600 - 1700	5 - 10	200 - 375	4 - 10

The probe is positioned along the neck skin layer, and the ROI is defined accordingly. To simulate tissue deformation caused by the pressure of the linear probe, the label is deformed in the orthogonal axis to represent the compressed skin layer (see Fig. 2-4). The acoustic properties of the tissues are modeled as demonstrated in Table 1 [**?**]. In total, 3k thyroid-mimicking phantoms are simulated for training the QIT-net.

2.2 Network Architecture

In this section, the architecture of the proposed QIT-net that reconstructs quantitative image including ATT and SOS from RF signals is presented. Following the multi-scanline transmission (MST) beamforming method, the RF signal is acquired N_{SL} times by sequentially exciting groups of 32 out of 128 aligned transducer elements. As illustrated in Fig. 3, the QIT-net is composed of four components: convolutional RF encoder, hierarchical Transformer encoder, B-mode morphology encoder, and quantitative image synthesizer.

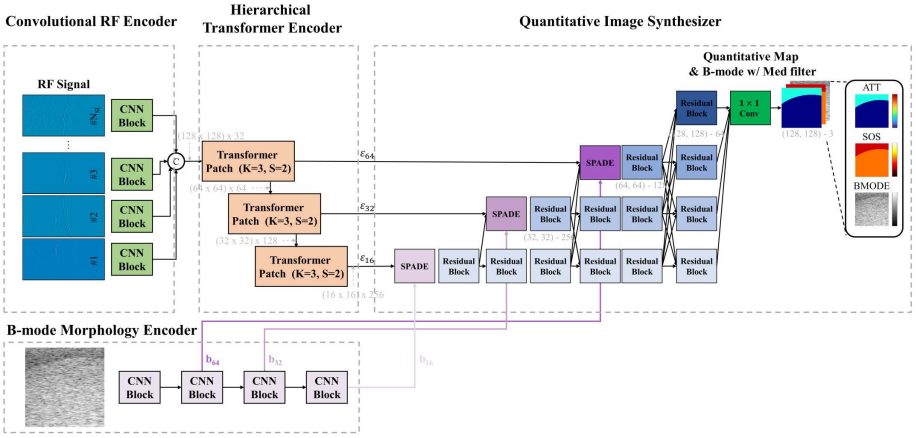

Fig. 3. QIT-net architecture

Convolutional RF Encoder RF signal is obtained from multiple ultrasound transmissions, and is configured to $x \in \mathbb{R}^{N_{SL} \times N_{ch} \times N_{Samp}}$ where N_{SL}, N_{ch}, and N_{Samp} represent the number of scanlines, receiver channels, and the time samples, respectively. Each beam-formed RF signal is processed through a convolutional operation to extract local signal features. Since MST generates an identical TX beam pattern across the scanlines, the convolutional RF encoder is designed to share CNN parameters among scanlines, thus facilitating RF pattern recognition.

Hierarchical Transformer Encoder (HTE) HTE is proposed to capture inter-correlations among multiple beam receptions while gradually reducing the feature resolution from 128×128 to 16×16 through repeated patch merging. Multi-scale feature representation, $\varepsilon_{16}, \varepsilon_{32}$, and ε_{64}, is generated and fed to the decoder for synthesizing the QUS image.

The standard multi-head attention (MHA) with the length of the sequence N has computational complexity of $O(N^2)$ [**?**], which implies that the execution time and memory requirements are compromised. Since the real-time QUS task is sensitive to inference speed, we incorporated spatial reduction attention (SRA) [**?**] to enhance computational efficiency. The hierarchical architecture is implemented using overlapped patch merging inspired by SegFormer [**?**].

B-mode Morphology Encoder (BME) B-mode image is generated through a conventional model-based numerical algorithm known as Delay-and-Sum (DAS) using the received ultrasound signals [**?**]. BME generates a feature map containing structural information of the target tissue, utilizing repeated convolutional and pooling layers. The extracted semantic features (b_{16}, b_{32},and b_{64}) are then

passed to the image synthesizer to provide positional markers for quantitative image inference.

Quantitative Image Synthesizer The quantitative image synthesizer serves two primary functions. Firstly, it incorporates acoustic information extracted from RF signal based on the morphological structure obtained from B-mode input. Secondly, it synthesizes the final quantitative acoustic image using the multi-scale hierarchical structure.

The encoded features from B-mode (b_{16}, b_{32}, and b_{64}) and RF signal (ε_{16}, ε_{32}, and ε_{64}) are fused to quantitative image using spatially adaptive demodulation (SPADE) followed by a series of residual convolution blocks [**?**]. The SPADE is denoted as

$$SPADE(b_r, \varepsilon_r) = \gamma_{x,y,ch}(\varepsilon_r)\frac{b_r - \mu_{ch}(b_r)}{\sigma_{ch}(b_r)} + \beta_{x,y,ch}(\varepsilon_r), \tag{1}$$

where morphological feature b_r is channel-wise normalized, and $\gamma_{x,y,ch}(\varepsilon_r)$ and $\beta_{x,y,ch}(\varepsilon_r)$ are learnable denormalization parameters. The SPADE operation provides the neural network with a detailed understanding of how quantitative value distribution correlates with the observed target structure.

The proposed multi-scale image synthesis module is inspired by U-Net [**?**] and HR-Net [**?**]. The decoder generates a higher resolution quantitative image, $G(x) \in \mathbb{R}^{128\times128}$, starting with the modulated image feature (SPADE(b_r, ε_r)). Unlike the conventional successive up-sampling schemes [**?**], the proposed decoder consists of a parallel multi-resolution subnetwork. Each unit subnetworks are designed with four residual blocks to enhance the stability of the training [**?**].

The low-resolution output images G_{16}, G_{32}, and G_{64}, and the final prediction map G_{128} are obtained using 1×1 convolution on feature map produced in each stage of the decoder. The 1×1 convolutional layer is an intuitive module that produces output with high spatial similarity among channels. Therefore, we defined the network's output as the concatenation of the ATT map, SOS map, and median-filtered B-mode image to supervise the inferred QUS maps to ensure morphological consistency with the input B-mode image.

2.3 Training Details

The objective function of the QIT-Net is defined as follows:

$$G^* = \arg\min_G E_{(x,y)}\|y - G(x)\|^2 + L_{SUB}, \tag{2}$$

$$\text{where } L_{SUB} = E_{(x,y_r)} \sum_{r\in\{16,32,64\}} \|y_r - G_r(x)\|^2. \tag{3}$$

y is the ground truth quantitative image with full resolution, and y_{16}, y_{32}, and y_{64} are the down-sampled images of y with dimension $\mathbb{R}^{16\times16}$, $\mathbb{R}^{32\times32}$,and $\mathbb{R}^{64\times64}$, respectively. The network G^* is trained to minimize the mean square difference

Table 2. Quantitative assessment results

Baseline	B-mode input	B-mode Concatenated label	ATT		SOS	
			PSNR [dB]	MAE [dB/cm/MHz]	PSNR [dB]	MAE [m/s]
FCN-based	X	X	25.52	0.056	22.16	20.7
	O	X	25.76	0.054	22.96	20.4
	O	O	26.26	0.053	23.53	17.4
CNN + Transformer Hybrid	X	X	26.57	0.047	22.81	19.8
	O	X	26.61	0.048	23.67	17.7
	O	O	27.86	0.041	24.60	15.6

between the ground truth y and the synthesized image $G(x)$. Furthermore, L_{sub} regularizes each subnetwork to progressively synthesize the corresponding resolution of quantitative images y_{16}, y_{32},and y_{64}. The QIT-net is optimized using Adam [**?**] with a learning rate of 10^{-4}, β_1 of 0.9 and β_2 of 0.999. For every convolutional operation, a dropout with a retention probability of 0.2 is applied to improve the network's generalization.

3 Experiments

3.1 Numerical Simulation

Performance Evaluation We evaluated the performance of QIT-net using 710 representative simulated test samples and assessed it based on the peak signal-to-noise ratio (PSNR) and the mean absolute error (MAE). Table 2 presents the results of ablation studies evaluating the effectiveness of additive B-mode input and the proposed B-mode concatenated label against the baseline FCN and CNN-Transformer hybrid encoder. The models with Transformers achieve PSNR values of 27.01 dB and 23.69 dB for ATT and SOS reconstruction on average, respectively, representing improvements of 1.16 dB and 0.82 dB over the FCN-based model. Furthermore, while the additional B-mode input alone yielded an average performance improvement of 0.49 dB, incorporating the B-mode concatenated label led to an additional enhancement of 0.81 dB in PSNR by reinforcement of morphological consistency.

3.2 Clinical Test

In-vivo thyroid measurements were obtained from 37 neck subjects with benign masses (n=29) and malignant tumors (n=8), confirmed through aspiration. To ensure the representativeness of the data and increase the sample size, each malignant tumor was measured twice, once along the transversal and longitudinal axes. Quantitative values for each lesion were measured 30 times, and the average values are recorded, as depicted in Fig. 4 (a)-(d).

In Fig. 4 (a), the reconstructed quantitative images from the clinical tests are presented. Networks trained solely on general soft tissue phantoms, which

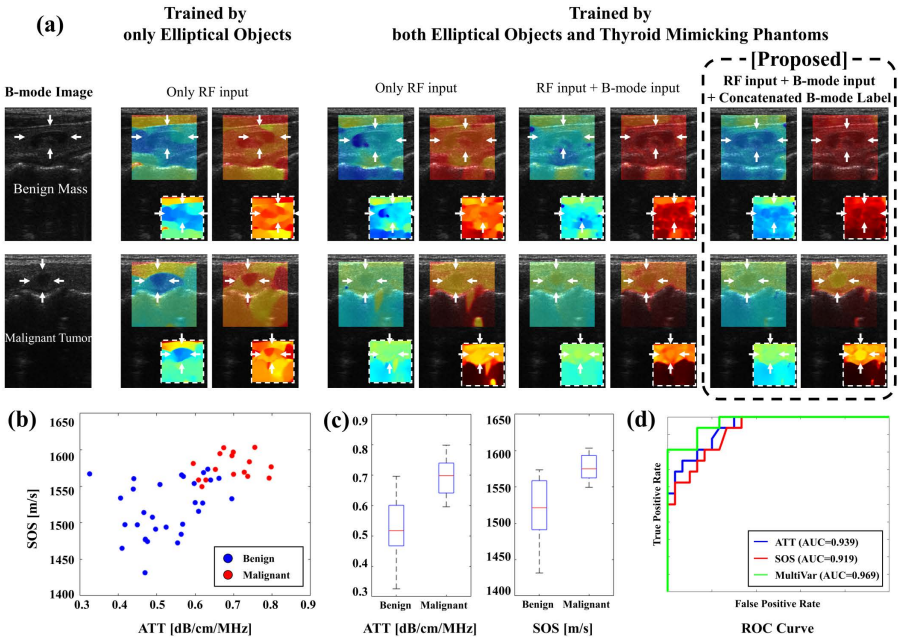

Fig. 4. In-vivo thyroid nodule measurements

consist of multiple elliptical objects, tend to produce oversimplified structures, lacking fidelity to delineate thyroid glands. In contrast, networks trained on both elliptical objects and thyroid-mimicking phantoms are capable of demonstrating detailed tissue structure. Notably, incorporating additional B-mode input and concatenated B-mode labels significantly improves the structural reconstruction accuracy and robustness.

As illustrated in Fig. 4 (b) and (c), benign masses exhibit lower acoustic attenuation and speed of sound compared to malignant tumors. In Fig. 4 (d), the receiver-operating characteristic (ROC) curve identifies malignancy using QIT with an AUC of 0.939 for ATT and 0.919 for SOS. Furthermore, the ROC curve defined by the sum of normalized ATT and SOS as a new variable is represented by the green line, demonstrating an AUC of 0.969.

4 Conclusions

In this paper, we propose a single-probe US system for quantifying acoustic attenuation and sound speed to identify thyroid nodules' malignancy. In numerical simulations, the CNN-Transformer hybrid model demonstrates a PSNR improvement of 0.99 dB compared to a CNN only approach. Additionally, the use of B-mode concatenated labels enhances the network's performance, resulting in an improvement of 0.81 dB. Clinical test demonstrates that the proposed

QUS system extracting ATT and SOS is able to determine the malignancy of thyroid nodule with an AUC of 0.969. Furthermore, the proposed simulation phantoms representing the actual structure and acoustic characteristics of the thyroid gland are proven effective in achieving structural accuracy in in-vivo settings. The proposed system can easily be employed in standard US systems and has great potential for clinical use, particularly in the area of non-invasive screening and differential diagnosis of cancer.

Acknowledgement. This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: RS2020-KD000007) and BK21 FOUR(Connected AI Education & Research Program for Industry and Society Innovation, KAIST EE, No. 4120200113769).

Disclosure of Interests. Y-M. Kim and S-H. Oh are under contracts with Ministry of Korea National Defense. Y-M. Kim, H-J. Lee and S-Y. Kim are under scholarship from the Korea Government Scholarship Program.

References

1. Ackerman, M.J.: The visible human project. Proceedings of the IEEE **86**(3), 504–511 (1998)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Feigin, M., Freedman, D., Anthony, B.W.: A deep learning framework for single-sided sound speed inversion in medical ultrasound. IEEE Transactions on Biomedical Engineering **67**(4), 1142–1151 (2019)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. IT'IS Foundation, Zurich, S.: Tissue properties database v4.0 (2018), https://itis.swiss/virtual-population/tissue-properties/downloads/database-v4-0/
6. Kim, M.G., Oh, S., Kim, Y., Kwon, H., Bae, H.M.: Learning-based attenuation quantification in abdominal ultrasound. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24. pp. 14–23. Springer (2021)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Oh, S., Kim, M.G., Kim, Y., Bae, H.M.: A learned representation for multi-variable ultrasonic lesion quantification. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1177–1181. IEEE (2021)

9. Oh, S., Kim, M.G., Kim, Y., Kwon, H., Bae, H.M.: A neural framework for multi-variable lesion quantification through b-mode style transfer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp. 222–231. Springer (2021)
10. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019)
11. Perrot, V., Polichetti, M., Varray, F., Garcia, D.: So you think you can das? a viewpoint on delay-and-sum beamforming. Ultrasonics **111**, 106309 (2021)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
13. Treeby, B.E., Cox, B.T.: k-wave: Matlab toolbox for the simulation and reconstruction of photoacoustic wave fields. Journal of biomedical optics **15**(2), 021314–021314 (2010)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
15. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020)
16. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
17. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)