Computer Methods and Programs in Biomedicine 257 (2024) 108458

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Multi-modal networks for real-time monitoring of intracranial acoustic field during transcranial focused ultrasound therapy

Minjee Seo [a], Minwoo Shin [a], Gunwoo Noh [b], Seung-Schik Yoo [c], Kyungho Yoon [a,*]

[a] Yonsei University, School of Mathematics and Computing (Computational Science and Engineering), Seoul, 03722, Republic of Korea
[b] Korea University, School of Mechanical Engineering, Seoul, 02841, Republic of Korea
[c] Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, 02115, MA, USA

ARTICLE INFO

Keywords:
Transcranial focused ultrasound
Computer simulation
Deep learning
Convolutional neural network
Swin transformer
Real-time monitoring

ABSTRACT

Background and objective: Transcranial focused ultrasound (tFUS) is an emerging non-invasive therapeutic technology that offers new brain stimulation modality. Precise localization of the acoustic focus to the desired brain target throughout the procedure is needed to ensure the safety and effectiveness of the treatment, but acoustic distortion caused by the skull poses a challenge. Although computational methods can provide the estimated location and shape of the focus, the computation has not reached sufficient speed for real-time inference, which is demanded in real-world clinical situations. Leveraging the advantages of deep learning, we propose multi-modal networks capable of generating intracranial pressure map in real-time.
Methods: The dataset consisted of free-field pressure maps, intracranial pressure maps, medical images, and transducer placements was obtained from 11 human subjects. The free-field and intracranial pressure maps were computed using the *k*-space method. We developed network models based on convolutional neural networks and the Swin Transformer, featuring a multi-modal encoder and a decoder.
Results: Evaluations on foreseen data achieved high focal volume conformity of approximately 93% for both computed tomography (CT) and magnetic resonance (MR) data. For unforeseen data, the networks achieved the focal volume conformity of 88% for CT and 82% for MR. The inference time of the proposed networks was under 0.02 s, indicating the feasibility for real-time simulation.
Conclusions: The results indicate that our networks can effectively and precisely perform real-time simulation of the intracranial pressure map during tFUS applications. Our work will enhance the safety and accuracy of treatments, representing significant progress for low-intensity focused ultrasound (LIFU) therapies.

1. Introduction

Transcranial focused ultrasound (tFUS) is a groundbreaking therapeutic technique, offering a non-invasive approach to precisely deliver concentrated acoustic energy to specific localized brain regions [1–6]. tFUS has gained further attention in the clinical research field as a new modality of non-invasive brain stimulation (NIBS) [7–11]. This technology has opened new avenues for non-pharmacological and non-invasive treatments of neurological disorders [12] and neuropsychiatric conditions [13].

The primary challenge associated with tFUS lies in ensuring the accurate delivery of acoustic focus to the intended brain region while the cranium distorts and obstructs the acoustic waves [14]. The wave propagation characteristics are intricately linked to the acoustic properties of the medium through which it travels. Since the anatomical structure of the skull is highly heterogeneous, acoustic waves are distorted and attenuated inevitably as they pass through the skull. This presents a significant challenge in accurately determining the actual location of the focal region during tFUS treatment [15–18].

To ensure precise sonication, it is crucial to monitor the location and intensity of the acoustic focus throughout the tFUS procedures. To address this need, several tFUS guidance techniques have been developed and adapted for clinical settings. For example, magnetic resonance (MR)-guided FUS (MRgFUS) [18], is a widely employed method that utilizes MR thermal imaging technique to monitor the temperature elevation at the focal region. However, its application to low-intensity FUS (LIFU) treatment is not possible due to the non-thermal nature of LIFU. Thus, recent LIFU studies have incorporated a neuronavigation system that geometrically tracks the position of the tFUS transducer in real-time and overlays it onto pre-acquired medical images [19–22].

* Corresponding author.
E-mail address: yoonkh@yonsei.ac.kr (K. Yoon).

https://doi.org/10.1016/j.cmpb.2024.108458
Received 17 July 2024; Received in revised form 22 September 2024; Accepted 7 October 2024
Available online 15 October 2024

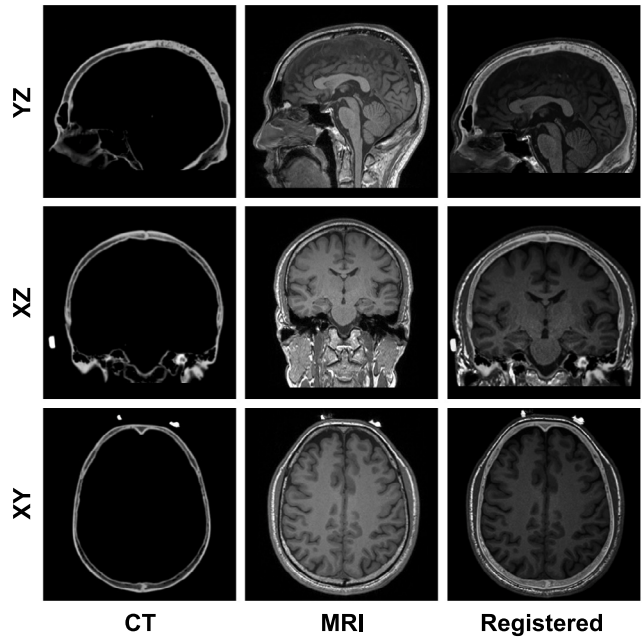


Fig. 1. The example images represent CT, MR, and co-registered CT and MR images from subject 1. During data generation, the sagittal plane was considered as the yz plane, the coronal plane as the xz plane, and the axial plane as the xy plane.

However, the approaches do not account for the changes in the focal location and intensity that accompany the skull transmission.

Computational methods can provide precise simulation outputs by numerically solving the wave equations with inclusion of the skull structure and material properties [23]. Despite their promising utility, high computational cost restricts practical and wide-spread use in clinical settings. To address these limitations, there is a growing demand for advanced simulation techniques capable of providing real-time feedback information while preserving the accuracy [16,24].

Recently, deep learning (DL) methods are gaining increasing attention in the medical domain for various applications including diagnosis [25], medical image analysis [26], prognosis [27], drug discovery [28], and treatment planning [29]. Furthermore, recent studies have applied DL methods to tFUS treatments, particularly providing real-time navigational information for the transducer placement [30] and generating high-resolution tFUS simulation results from low-resolution inputs [31,32]. The fundamental advantage of employing DL methods lies in their data-driven nature: networks automatically learn essential features from the given dataset [33]. This approach enables the network to identify complex structure within the data and uncover the meaningful patterns that are not immediately apparent, thereby enhancing its predictive and analytical effectiveness.

Another benefit of DL is its rapid inference capability, which can be accelerated using GPUs [34]. A representative example is autonomous driving, where it processes and interprets data from the vehicle's sensors in real-time [35]. Building on these advantages, there have been attempts in the medical field to develop real-time diagnostic systems using DL methods [36–38].

Traditional DL methods utilize a single modality dataset, leading to the development of specialized models that are tailored for specific tasks. However, this approach limits the capability of a network that interprets and integrates the complex information presented in real-world scenarios. To address this limitation, recent studies have introduced multi-modal learning methodologies that allow for a single network to handle various data types [39–41].

This approach proves especially beneficial in the medical domain, where the domain knowledge is crucial and often suffers from data scarcity. It enables networks to effectively learn diverse information, contributing to more precise outcomes in medical applications [42]. In addition, recent advancements in wearable devices, data acquisition methods, and omics technology have expanded the data modalities beyond traditional medical data [42], enabling the utilization of diverse array of datasets. Moreover, a single-domain medical data (e.g., medical images) is organically linked with other clinical features (e.g., lab results) due to its nature, making it more advantageous for processing multi-modal data [43].

Studies are being conducted to utilize multi-modal learning across different applications within the medical domain. Several methods have been proposed to enhance the processing of visual representations of medical images using paired medical image–text data, with applications ranging from classification and image retrieval tasks to the automatic generation of medical imaging reports [43,44]. Processing medical image data from different modalities can also serve as an example of multi-modal learning. For instance, several methods for medical image fusion based on convolutional neural network (CNN), generative adversarial network (GAN), and Transformer have been proposed in [45–47]. ECG signal data, primarily utilized for detecting abnormalities in the cardiovascular system, leverages multi-modal fusion technology by integrating with other datasets such as echocardiography to improve disease diagnosis performance [48,49]. Additionally, integration of multi-omics data, including genomics, epigenomics, and proteomics, enhances outcomes in various medical analysis, such as in survival rate prediction associated with cancers [50,51].

Inspired by the exceptional performance of multi-modality learning in medical applications, this study presents multi-modal deep neural networks that enable real-time monitoring of the intracranial acoustic pressure map during tFUS treatment. We employed a dataset obtained from 11 healthy human subjects, which consisted of free-field pressure map, medical images, and transducer placements, all of which were used as inputs to the network. The intracranial pressure field acquired from physics simulations using the *k*-space method was set as the target data. Then, we developed a set of networks, each consisting of (1) a multi-modal feature encoder that extracts and integrates features across various modalities and (2) a decoder that restores the spatial dimensions of the encoded representations. The accuracy of the networks was assessed using foreseen data from 8 subjects, which was used for network training, and subsequently evaluated on unforeseen data from an additional 3 subjects. Our contributions are summarized as follows:

- We have achieved real-time prediction of intracranial pressure field with high accuracy.
- We have developed sets of networks to effectively process multi-modal information.
- We have achieved higher performance by using multi-modal data for tFUS simulation compared to relying on a single modality.
- We have demonstrated the feasibility of predicting intracranial pressure maps using only MR images, eliminating the need for CT that requires exposure to ionizing radiations.

2. Data generation

In this section, we provide the outline of the multi-modal data acquisition methods for network training and evaluation: (1) the acquisition methods for CT and MR images, (2) numerical modeling process for the skull structure, (3) analysis of skull structures among participants, (4) methods for defining the position of the transducer, and (5) numerical simulation method to obtain free-field and intracranial acoustic pressure maps.

2.1. Acquisition of CT and MR images

3D CT (Aquilion One, Toshiba, Japan) and MR (Skyra, Siemens, Germany) images of skulls from 11 participants were acquired under the approval of the Institutional Review Board (IRB; Incheon St.Mary's Hospital, The Catholic University of Korea). To attain the spatial alignment of each imaging modality, four fiducial markers (Pinpoint, Beekley Medical, Bristol, CT) were attached onto the subject's head. CT images were scanned with an isovoxel size of $0.5 \times 0.5 \times 0.5$ mm^3 and cropped above the frontal sinus, covering field-of-view of $225 \times 225 \times 150$ mm^3. T1-weighted MR images (3D GRAPPA sequence, acceleration factor = 2, TR/TE = 1900/2.46 ms, Flip angle = 9°) were scanned with an isovoxel size of $1.0 \times 1.0 \times 1.0$ mm^3 covering field-of-view of $240 \times 180 \times 240$ mm^3, and then resampled to an isovoxel size of $0.5 \times 0.5 \times 0.5$ mm^3 with the Lanczos interpolation to match the size of the CT images [52]. The obtained CT and MR images of each subject were co-registered using a point-based rigid body registration method [53]. The example of medical images used for data generation is shown in Fig. 1.

2.2. Skull modeling

For the numerical simulation of ultrasound propagation, geometric and material properties of the skull were modeled using Hounsfield units (HU) $\phi_{i,j,k}$ of each CT image, where i, j, k represent the voxel indices of x-, y-, and z-direction, respectively. The simulation domains were classified by thresholding each voxel with HU values, where regions with $\phi_{i,j,k} \leq 0$ were defined as water, those with $0 < \phi_{i,j,k} < 1000$ as trabecular bone, and regions with $1000 \leq \phi_{i,j,k}$ as cortical bone. Ultrasound velocity ($c_{i,j,k}$), density ($\rho_{i,j,k}$), and attenuation coefficient ($a_{i,j,k}$) of the skull were modeled as [30,54,55]:

$$
\begin{aligned}
c_{i,j,k} &= \begin{cases} 1500 \text{ m/s}, & \text{for } \phi_{i,j,k} \leq 0, \\ 2140 \text{ m/s}, & \text{for } 0 < \phi_{i,j,k} < 1000, \\ 2384 \text{ m/s}, & \text{for } 1000 \leq \phi_{i,j,k}, \end{cases} \\
\rho_{i,j,k} &= \begin{cases} 1000 \text{ kg/m}^3, & \text{for } \phi_{i,j,k} \leq 0, \\ 1000 + 1.19\phi_{i,j,k} \text{ kg/m}^3, & \text{for } 0 < \phi_{i,j,k} < 1000, \\ 2190 \text{ kg/m}^3, & \text{for } 1000 \leq \phi_{i,j,k}, \end{cases} \\
a_{i,j,k} &= 33 \text{ Np/m}, \qquad \text{for } 0 < \phi_{i,j,k}.
\end{aligned}
\tag{1}
$$

2.3. Analysis of skull structure

To examine the variations in skull structure between patients, we conducted a detailed analysis of the modeled skull structures from 11 participants described in Section 2.2, which are denoted as S1–S11. This analysis includes the number of segmented voxels for trabecular and cortical bone, the ratio of trabecular to cortical bone, and the mean and standard deviation of the trabecular bone density. The comparative results of skull structure analysis are presented in Table 1.

2.4. Transducer modeling

Assuming the LIFU treatment condition, we modeled a single-element partial hemisphere-shaped transducer with an operating frequency of 250 kHz, featuring a diameter, radius of curvature, and focal length set to 75 mm, 83 mm, and 83 mm, respectively.

Here, we denote the location and orientation of the transducer as $\mathbf{T}_c = [T_x, T_y, T_z]^T$ and $\mathbf{n}_t = [n_x, n_y, n_z]^T$, where $\mathbf{T}_c$ is the vertex of the transducer and $\mathbf{n}_t$ indicates its normal direction. To define the transducer placement, we sequentially define $\mathbf{C}_{\text{ROI}}$, $-\mathbf{s}_t$, $\mathbf{n}_t$, and $\mathbf{T}_c$. $\mathbf{C}_{\text{ROI}}$, the target area for the focal region, was identified manually by analyzing registered CT-MR images, targeting the deep brain area. Normal vectors of the skull surface node $-\mathbf{s}_t$ were obtained by applying MATLAB built-in functions "triangulation" and "vertexNormal" on the skull surface [56]. The vector $\mathbf{n}_t$ was derived by connecting the $\mathbf{C}_{\text{ROI}}$ to the skull surface node, and $\mathbf{T}_c$ was calculated by extending $\mathbf{n}_t$ until it reaches the condition. The illustration of the defined vectors is shown in Fig. 2.

Table 1
Comparison of measurements of skull structures. #TB denotes the number of voxels segmented as trabecular bone, #CB denotes the number of voxels segmented as cortical bone, and Mean(TB) denotes the mean density of the trabecular bone.

Skull	#TB	#CB	#TB/#CB	Mean (TB)
S1	838 221	906 775	0.9244	1516 ± 166
S2	809 852	1 206 623	0.6712	1772 ± 264
S3	929 632	903 984	1.0284	1729 ± 271
S4	920 120	901 503	1.0207	1659 ± 227
S5	992 816	1 176 011	0.8442	1789 ± 254
S6	947 878	1 367 146	0.6933	1746 ± 257
S7	441 831	1 018 810	0.4337	1715 ± 284
S8	945 858	1 480 966	0.6387	1816 ± 274
S9	894 055	1 077 108	0.8301	1775 ± 255
S10	736 964	1 694 643	0.4349	1737 ± 266
S11	1 140 307	1 294 106	0.8812	1832 ± 256

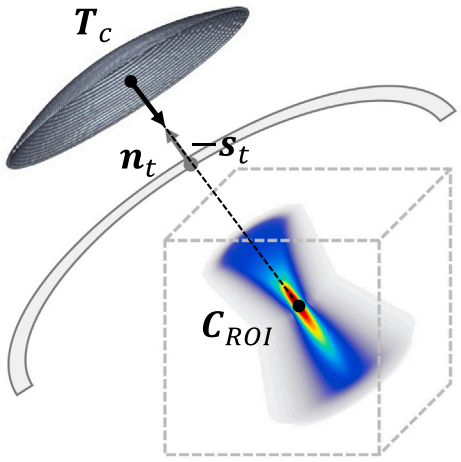


Fig. 2. Illustration of transducer placements modeling. $\mathbf{T}_c$ is the vertex of the partial hemisphere-shaped transducer, $\mathbf{n}_t$ is the normal direction of the transducer, $-\mathbf{s}_t$ is the normal vector of the skull surface pointing outward, and $\mathbf{C}_{\text{ROI}}$ is the target point for the transducer.

A total of 400 transducer placements were randomly selected within each subject's skull with two constraints: (1) ensuring that the distance between $\mathbf{T}_c$ and each corresponding region of interest (ROI) center $\mathbf{C}_{\text{ROI}}$ was 83 mm, and (2) keeping the angle between the normal vector of the skull surface $\mathbf{s}_t$ and $\mathbf{n}_t$ was under 10 degrees to achieve optimal focusing.

2.5. k-space method

The intracranial acoustic pressure map corresponding to each transducer position was obtained using the k-wave MATLAB toolbox [57], which is based on the k-space method [58,59]. The governing equations for ultrasound propagation simulation are stated below:

$$
\begin{aligned}
\frac{\partial \mathbf{u}}{\partial t} &= -\frac{1}{\rho_0} \nabla p, \\
\frac{\partial \rho}{\partial t} &= -\rho_0 \nabla \cdot \mathbf{u} - \mathbf{u} \cdot \nabla \rho_0, \\
p &= c_0^2 \left(\rho + \mathbf{d} \cdot \nabla \rho_0 - L\rho \right),
\end{aligned}
\tag{2}
$$

where $\mathbf{u}$ is the acoustic particle velocity, p is the acoustic pressure, ρ is the acoustic density, ρ_0 is the ambient density, c_0 is the isentropic sound speed, $\mathbf{d}$ is the acoustic particle displacement, and the operator L is a linear integro-differential operator that accounts for acoustic absorption and dispersion that follows a frequency power law [60].

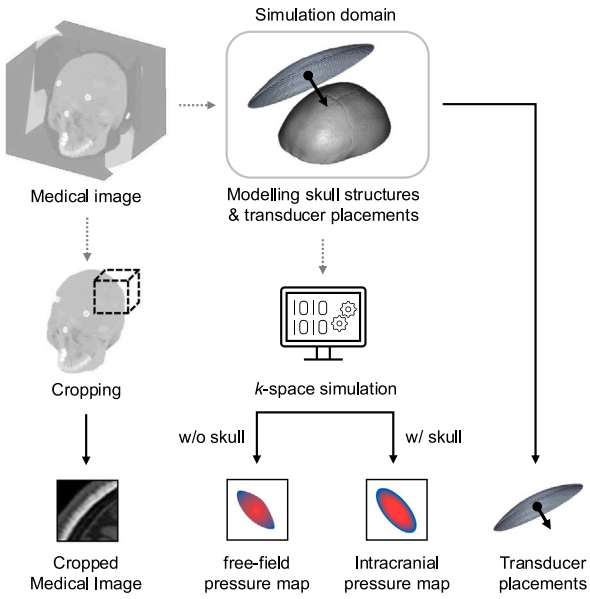


Fig. 3. Construction of the dataset. Skull properties and transducer placements defined using medical images, are utilized to generate pressure map data. Medical images (CT and MR) are cropped at the point where the transducer beamline intersects with the skull surface, acting as an additional input modality. Transducer placement vectors are used as input data that conveys transducer location information to the network.

To ensure the stability and convergence of the numerical simulation, we selected an appropriate time step size that conforms to the Courant–Friedrichs–Lewy (CFL) criterion as follows [58,61]:

$$\Delta t = C_{\mathrm{CFL}} \frac{\Delta x}{c_{\max}} \tag{3}$$

where $c_{\max}$ is the maximum speed of sound in the medium with $C_{\mathrm{CFL}} = 0.3$. All simulations were performed with the simulation time of 120 μs, and the ultrasound signal emitting from the source was modeled as a tone burst signal with a cycle count of 5.

2.6. Construction of dataset

We constructed a dataset consisting of three different modalities: (1) acoustic pressure maps representing the physical information of ultrasound propagation, (2) medical images conveying geometric information about the biological structure, and (3) transducer location vectors which indicate the position and orientation of the transducer. Fig. 3 is a schematic diagram of the dataset generation process.

2.6.1. Acoustic pressure maps

In this study, two types of pressure maps were computed with the *k*-space method for network training: the free-field pressure map and the intracranial pressure map. The free-field pressure map served as the network input, with the intracranial pressure map as the target 'ground-truth'. Here, the free-field pressure map refers to the pressure map generated when ultrasound propagates through a homogeneous medium (in our case, water), while the intracranial pressure map refers to the pressure map generated when ultrasound propagates through the skull.

Using the transducer locations and the skull geometry obtained previously, corresponding intracranial pressure maps were computed via *k*-space method and cropped to a field-of-view of $56 \times 56 \times 56$ mm^3 centered at $\mathbf{C}_{\mathrm{ROI}}$, resulting in an image matrix of $112 \times 112 \times 112$.

The free-field pressure map can be calculated by ignoring the modeled skull geometry in the simulation domain and assuming all acoustic properties as water. To generate free-field pressure map immediately during the actual clinical applications, we used a pre-calculated reference free-field pressure map with the $\mathbf{n}_t$ set to $[1, 0, 0]$. By rotating the reference pressure map with the intrinsic rotation matrix according to the $\mathbf{n}_t$ which was used for intracranial simulation, we obtained the corresponding rotated free-field pressure maps.

To enhance the efficiency of network training, the free-field pressure map was normalized to a range between 0 to 1 using Min–Max Scaling. The intracranial pressure map was also normalized with the same min–max range as the free-field pressure map, thereby representing the transmission rate of the peak pressure considering the attenuation caused by the skull.

2.6.2. Medical images

To enable the network to effectively capture the geometric information of the skull structure, an appropriate medical image modality should be provided. To address this additional need, we utilized CT and MR images as the network input. The registered medical images were cropped to an image matrix of $112 \times 112 \times 112$ at the intersection of the transducer beamline with the skull. To highlight essential information in each imaging modality, CT and MR images were thresholded to have intensity values in the ranges of (0,2000) and (0,1000), respectively. Values outside these ranges were assigned to the nearest minimum or maximum value within the range.

2.6.3. Transducer placements

To provide additional information about the transducer placement, we utilized the transducer's location and orientation vector as a network input. The transducer placement vector ($\mathbf{v}$) was obtained by concatenating $\mathbf{T}_c$ and $\mathbf{n}_t$ along the row, resulting a dimension of 1×6. For $\mathbf{T}_c$, its values in $x-, y-, z-$ directions were normalized through Min–Max Scaling with the simulation domain's field-of-view: an image matrix of $450 \times 450 \times 300$.

2.6.4. Splitting data for training and evaluation

We obtained datasets from total of 11 skulls, each with 400 data points. Data from 8 skulls (S2, S4, S5, S7–S11) of these were used for training, with 320 data points per skull for training and the remaining 80 data points for validation on foreseen data. The remaining 3 skulls (S1, S3, S6) provided 1200 data points used to validate network performance on unforeseen data. Detailed descriptions of the foreseen and unforeseen sets are provided in Section 4.4.

3. Multi-modal networks for tFUS simulation

This section introduces the architecture of our network. Since the input data comprises information from various modalities, each of the features should be processed and integrated. To handle this task, we developed three network structures specifically for multi-modal data processing based on convolutional neural network (CNN) and Swin-transformer.

3.1. CNN-based networks

Here, we propose CNN-based network architectures consisting of an encoder for fusing multi-modal feature and a decoder for generating intracranial pressure maps. The entire structure of CNN networks is illustrated in Fig. 4.

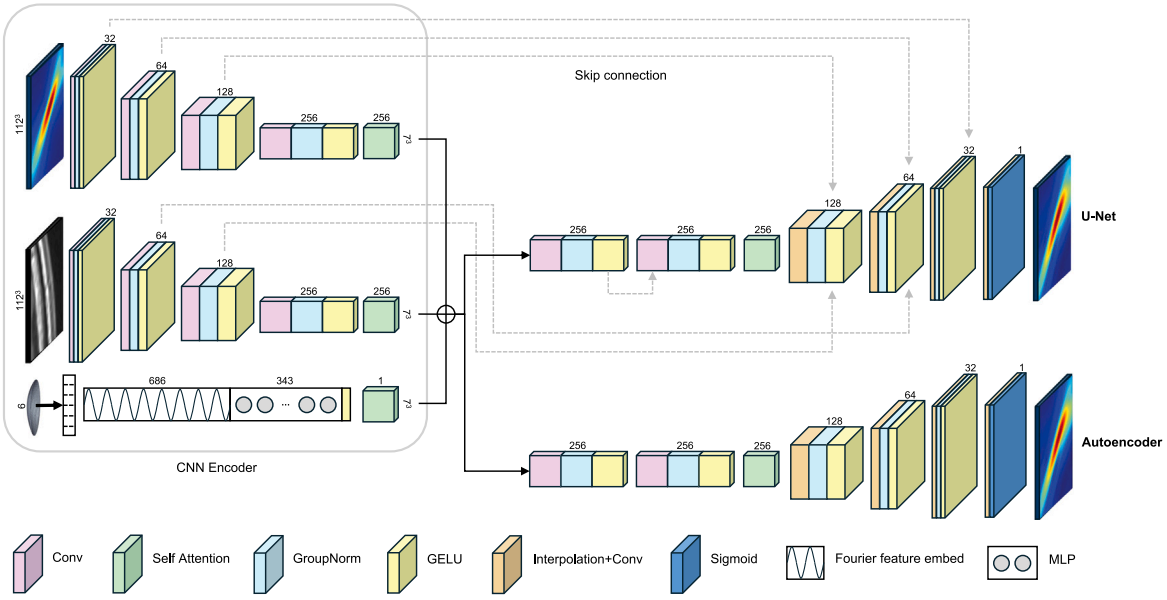


Fig. 4. Architectures of CNN-based networks. The basic configuration of the two networks is the same, but their structures diverge based on the presence of skip connections. Note that the final skip connection for medical image features was excluded in the U-Net structure.

3.1.1. CNN-based multi-modal feature encoder and deep feature extraction layers

In the encoder, features from free-field pressure maps and medical images (CT or MR) were respectively extracted using serially-connected four convolution blocks. Each convolution block comprises a 3D convolution with a kernel size of $4 \times 4 \times 4$, stride of 2 and zero padding of 1, which reduces the spatial dimension of the data by half. Group normalization and GELU is applied after convolution operation, and dropout with $p = 0.1$ was applied to prevent overfitting. To enhance training efficiency, instance normalization is applied to the pressure map and medical image before feeding them into the network. The size of the final output becomes $256 \times 7 \times 7 \times 7$.

In the case of transducer placement vectors, the Fourier feature mapping [62] was applied before passing into the network to effectively capture the spatial information :

$$\gamma(\mathbf{v}) = [\cos(2\pi \mathbf{B}\mathbf{v}), \sin(2\pi \mathbf{B}\mathbf{v})^{\mathrm{T}}] \tag{4}$$

where $\mathbf{B}$ is a random Gaussian matrix and $\mathbf{v}$ is the transducer placement vector.

An MLP layer with GELU activation function is then applied to provide additional nonlinear mapping and to transform the dimension to 343×1. Output of the layer is then reshaped into $1 \times 7 \times 7 \times 7$, aligning the spatial dimensions with the 3D data.

Multi-head self-attention (MHA) [63] stated below was respectively applied to each of the processed features to highlight the essential information :

$$\begin{aligned} &\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W^O, \\ &\mathrm{head}_i = \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \\ &\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right) V \end{aligned} \tag{5}$$

where Q is the query matrix, K is the key matrix, V is the value matrix, d is the dimension of the data, h is the number of attention heads, i is the index of each head, and W is the weight matrix.

MHA with $h = 4$ was applied to the channel direction of free-field pressure maps and medical images (i.e., CT or MR), while transducer placements were processed with a single attention head. Computed MHA values were added to the input feature to enhance feature processing. Features from each modality were merged with an element-wise sum in the channel direction after self-attention. Note that the information regarding the transducer placement is distributed across all channels, given its single-channel dimension.

After encoding and merging each feature, deep feature extraction was performed using two convolution blocks with a kernel size of $3 \times 3 \times 3$, a stride of 1 and zero padding of 1, maintaining the spatial dimension of the merged feature map. Group normalization, GELU activation function, and dropout with $p = 0.5$ were applied after each convolution operation. Finally, MHA with $h = 4$ was applied, and computed attention values were added to the feature.

3.1.2. CNN-based decoder for generation of intracranial pressure map

The decoder serves to restore the reduced spatial dimensions during the encoding process. We adopted two different decoder structures: Autoencoder (AE) and U-Net.

An AE [64] is a network structure that consists of an encoder and decoder. The encoder processes and downsamples the feature information into the latent vector, while the decoder reconstructs and expands the compressed feature into a representative reconstruction of the original input data. The decoder of AE was implemented by upsampling the encoded features in spatial dimensions through trilinear interpolation, followed by 3D convolution operations with a kernel size of $3 \times 3 \times 3$, stride of 1, and zero padding of 1.

The key idea of the U-Net [65] is in the use of skip connections, which allows for direct transfer of high-resolution features extracted by the encoder to the decoder, preventing the loss of spatial information in the data. In our implementation, based on the AE structure, features extracted from encoding blocks were added to the upsampled features, and the output of each block in deep feature extraction layers was merged with the output of the previous block. Here, we empirically excluded the skip connection of medical image features from the first encoding block to generate higher quality outputs.

3.2. Swin transformer-based networks

In this section, we propose Swin Transformer [66] based network architecture comprising Swin Transformer encoder and the CNN decoder. The core concept of the Swin Transformer encoder is to divide the input data into patches and compute the relation between patches using shifted window attention mechanism. To reduce the significant computational load inherent in the Transformer structure, we adopted the CNN architecture for the decoder.

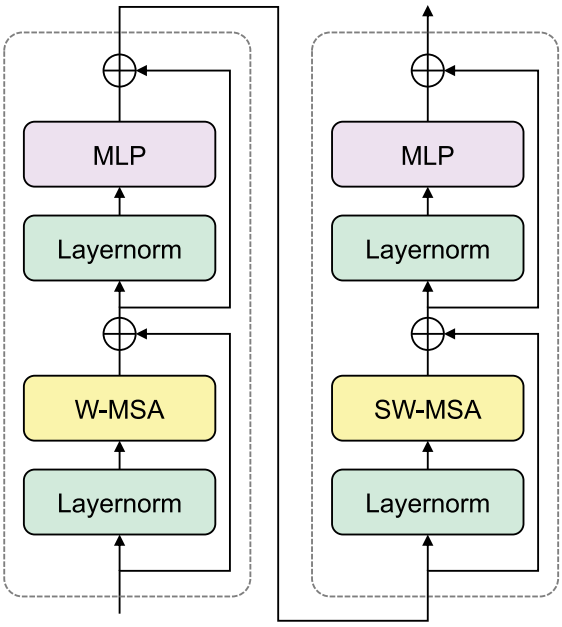


Fig. 5. Illustration of two Swin Transformer blocks. Each of the block shares the same structure, except for using different attention mechanisms.

3.2.1. Swin transformer-based encoder

The Swin Transformer-based encoder consists of patch processing blocks and Swin Transformer blocks. For the pressure map and medical image data, the initial patch partition process divides the whole image map into $28 \times 28 \times 28$ patches of size $4 \times 4 \times 4$ with 3D convolution. Subsequently, each patch size was increased to 96 through linear embedding, and fed into four Swin Transformer blocks.

The odd-numbered Swin Transformer blocks consist of a sequence of layer normalization, window attention, layer normalization, and an MLP layer, with each data being merged with a skip connection before layer normalization. In the even-numbered Swin Transformer block, shifted window attention was applied instead of window attention, enabling the computation of attention scores between windows. The structure of two Swin Transformer blocks is represented in Fig. 5.

After passing through the Swin Transformer blocks, adjacent patches are merged, reducing the number of patches by half. Subsequently, each patch was embedded to a size of 192 and passes through another Swin Transformer block. Following the same process, the dimension of the data after the final Swin Transformer block became 7×7×7×384. For the transducer placement vectors, the dimension was expanded to a size of $7 \times 7 \times 7 \times 1$ using the same feature encoding process as in CNNs, and then merged with the encoded 3D data.

3.2.2. Swin transformer-based decoder for pressure map generation

To reduce the computational burden, we adopted the CNN-based U-Net shaped decoder architecture similar to the CNN decoder block structure described in 3.1.2, but increased the decoder's complexity by performing convolution twice after the interpolation. After passing the decoder block, the number of patches doubles while the size of each patch was halved. Skip connections were added between feature maps from previous encoding blocks and feature maps from current decoding blocks in the channel direction. The structure of Swin Transformer-based U-Net (Swin U-Net) is illustrated in Fig. 6.

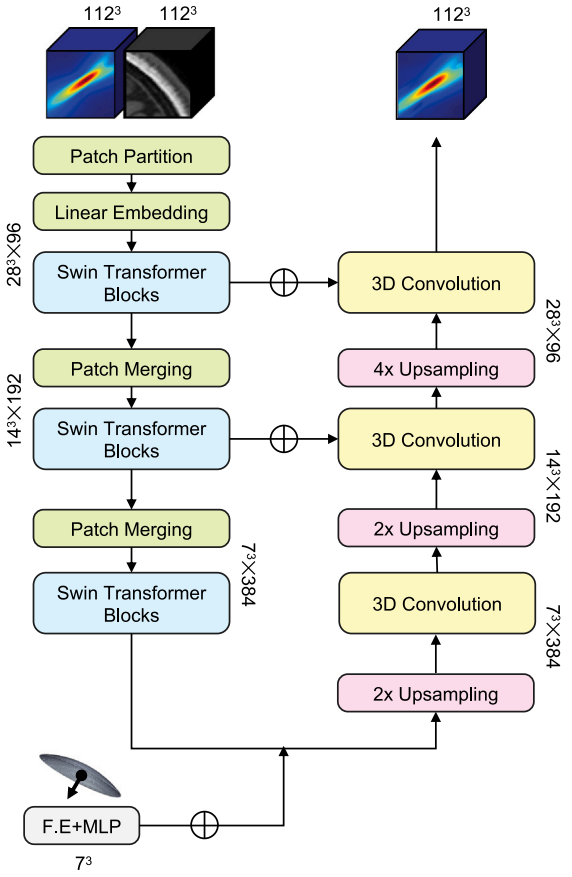


Fig. 6. Network architecture of Swin Transformer-based U-Net. The initial patch partition block employs 3D convolution, while patch merging blocks use MLP layers to reduce the number of patches. Decoding procedures were performed with interpolation and 3D convolution to conserve computational resources.

3.3. Optimization and hyperparameter settings

The similarity between the model output and the target data was computed using the mean-squared error (MSE), aiming to minimize the following loss function $\mathcal{L}$:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}[(p_{\text{pred}}^{(i)} - p_{\text{true}}^{(i)})^2] \tag{6}$$

where N is the number of the training data, p_{pred} is the intracranial pressure map output from the network, p_{true} is the ground truth intracranial pressure map, and i is the index of the training data. The implementation and training of the network structure were developed with Pytorch. Batch size for training processes was set to 8 for CNN-based networks and 4 for the Swin U-Net. The weights of the convolution layers, MLP layers, and group normalization layers were initialized to follow a truncated normal distribution with std = 0.02. The weights and biases of layer normalization were initialized to constant 1 and 0, respectively. The loss function was optimized using the AdamW optimizer($b_1 = 0.5$, $b_2 = 0.999$, weight_decay = 0.05). The training was set for a total of 200 epochs. For the first 100 epochs, learning rate was set to 0.0002 for the CNN networks and 0.0005 for the Swin Transformer. For the subsequent 100 epochs, a learning rate scheduling was applied using the following rule:

$$lr = 1.0 - \frac{\mathbf{max}(0, \mathbf{e} - 99)}{100}, \text{for } \mathbf{e} \geq 100 \tag{7}$$

where $\mathbf{e}$ is the current training epoch.

Table 2
The number of parameters and FLOPs of the networks.

Networks	num. params	FLOPs
CNN AE, CNN U-Net	11,636,522	30,891,114,737
Swin U-Net	38,869,586	78,433,351,409

Table 3
Comparison of performance across three different network architectures.

Networks	Dice (%)	D_p (mm)	Δ_p (%)
CNN AE	87.49 ± 3.30	1.52 ± 1.04	1.69 ± 1.13
CNN U-Net	93.09 ± 2.95	1.36 ± 1.11	0.71 ± 0.59
Swin U-Net	91.26 ± 3.70	1.68 ± 1.38	1.28 ± 1.14

Table 4
Comparison of performance depending on different input modalities for training.

Networks	Dice (%)	D_p (mm)	Δ_p (%)
Pressure map			
CNN AE	78.60 ± 10.18	2.68 ± 1.81	4.10 ± 3.34
CNN U-Net	80.46 ± 10.63	2.71 ± 1.94	3.23 ± 2.83
Swin U-Net	83.42 ± 6.66	2.61 ± 1.66	2.88 ± 2.31
Pressure map + CT			
CNN AE	87.33 ± 4.05	1.43 ± 0.92	2.73 ± 1.48
CNN U-Net	94.79 ± 2.17	1.22 ± 0.95	0.63 ± 0.55
Swin U-Net	93.13 ± 3.52	1.65 ± 1.26	1.39 ± 1.15
Pressure map + CT + transducer placements			
CNN AE	87.49 ± 3.30	1.52 ± 1.04	1.69 ± 1.13
CNN U-Net	93.73 ± 2.40	1.36 ± 1.02	0.71 ± 0.59
Swin U-Net	91.26 ± 3.70	1.68 ± 1.38	1.28 ± 1.14

4. Results

In this section, we assess the performance of our network using three different evaluation metrics. To assess the robustness of our network, we also performed evaluations on unforeseen data. All of the data generation and network training processes were conducted with an AMD Ryzen 9 5950X 16-Core Processor, 128.0 GB-RAM, and a single NVIDIA RTX 3090 Ti GPU.

4.1. Evaluation metrics

The primary performance evaluation metric was the degree of similarities in intracranial pressure focal volume predicted by the network model compared to the ground truth. Therefore, we calculated the Dice score [67] between the full-width at half-maximum (FWHM) region of the network output and the ground truth. Additionally, we measured the Euclidean distance between the peak pressure points, denoted by D_p. To compare the error in transmission rates of the acoustic pressure through the skull, we defined Δ_p as the difference in peak pressure values. The formulations for the evaluation metrics are as follows:

$$\begin{aligned} \text{Dice} &= \frac{2|\text{FWHM}(p_{\text{pred}}) \cap \text{FWHM}(p_{\text{true}})|}{|\text{FWHM}(p_{\text{pred}})| + |\text{FWHM}(p_{\text{true}})|}, \\ D_p &= \sqrt{\sum_{i=1}^{3} |(\mathbf{argmax}_{x,y,z}(p_{\text{pred}}) - \mathbf{argmax}_{x,y,z}(p_{\text{true}}))_i|^2}, \\ \Delta_p &= |\mathbf{max}(p_{\text{pred}}) - \mathbf{max}(p_{\text{true}})| \end{aligned} \tag{8}$$

where i is the spatial indices of the pressure map.

4.2. Complexity of the network

Prior to comparing the performance of CNN-based networks and Swin U-Net, we evaluated the complexity of each network by calculating their respective numbers of network parameters and floating point operations (FLOPs). Only parameters requiring gradient calculations were included for the network parameters, and FLOPs were computed for self-attention, MLP, and convolution operations.

Table 2 shows the number of parameters and FLOPs used in CNN AE, CNN U-Net, and Swin U-Net. Swin U-Net had approximately three times more parameters than the CNN-based networks, and its FLOPs were about 2.6 times higher. Based on these results, we proceed with the analysis under assumption that Swin U-Net is significantly more complex than the CNN-based networks.

4.3. Ablation studies

We conducted ablation studies to evaluate the impacts of different network structures, number of input modalities, type of input medical images, and loss functions.

4.3.1. Comparison among CNN-based and Swin-transformer-based networks

We compared the performance of three different network architectures. Table 3 exhibits the comparisons in terms of the accuracy of generated intracranial pressure maps among CNN AE, CNN U-Net, and Swin U-Net in terms of Dice, D_p, and Δ_p. CNN U-Net demonstrated the best performance. We observed high focal volume conformity of 93.73%, with minimal differences in D_p and Δ_p. Swin U-Net showed the next highest focal volume conformity and low Δ_p, however, displayed the greatest variance in Dice scores and the highest D_p among all networks. CNN AE showed the lowest focal volume conformity and the highest Δ_p.

4.3.2. Comparison between using single modality and multiple modalities

The objective of this section is to examine how the input modalities affect the performance of the network models. Table 4 shows the mean and standard deviation values of evaluation metrics depending on the number of input modalities used in the network. All network structures using a single modality have exhibited the worst performance: the lowest mean Dice score and the highest standard deviation. The mean distance and peak pressure ratio was also the highest. Based on Dice score and Δ_p, the performance was best in the order of Swin U-Net, CNN U-Net, and then CNN AE, while D_p showed no significant difference among all networks. When CT image was added as an input modality, significant performance improvement was seen across all network architectures. However, when the transducer input vector was used as an additional input modality, the Dice score improved for CNN AE, but the performance declined slightly in CNN U-Net and Swin U-Net.

4.3.3. Comparison between different medical image modalities

Here we conducted comparative experiments to determine if the network can still effectively extract skull features and maintain its performance when using MR images instead of CT images. Table 5 shows the mean and standard deviation values of evaluation metrics when using CT or MR images as network inputs. The use of MR image data displayed similar performance compared to that achieved with CT image modality. In particular, CNN AE showed higher Dice score, while CNN U-Net and Swin U-Net exhibited a slight decline in performance. For D_p, a slight decrease was shown in CNN AE, with no significant changes observed in the other networks. For Δ_p, a slight increase was noted in both CNN AE and CNN U-Net, while Swin U-Net showed a slight decrease.

Table 5
Comparison of performance depending on different medical image modality for training.

Networks	Dice (%)	D_p (mm)	Δ_p (%)
CT			
CNN AE	87.49 ± 3.30	1.52 ± 1.04	1.69 ± 1.13
CNN U-Net	93.73 ± 2.40	1.36 ± 1.02	0.71 ± 0.59
Swin U-Net	92.10 ± 3.39	1.70 ± 1.32	1.28 ± 1.14
MRI			
CNN AE	89.57 ± 3.33	1.40 ± 0.96	1.87 ± 1.05
CNN U-Net	93.09 ± 2.95	1.36 ± 1.11	0.75 ± 0.67
Swin U-Net	91.26 ± 3.70	1.68 ± 1.38	1.25 ± 1.18

Table 6
Comparison of performance using different loss functions for network optimization.

Loss	Dice (%)	D_p (mm)	Δ_p (%)
MAE	93.95 ± 2.45	1.36 ± 1.05	0.73 ± 0.67
MSE	93.73 ± 2.40	1.36 ± 1.02	0.71 ± 0.59
SSIM	93.74 ± 2.58	1.37 ± 1.12	0.94 ± 0.79

Table 7
Comparison of performance using foreseen and unforeseen data.

Networks	Dice (%)	D_p (mm)	Δ_p (%)
Foreseen with CT			
CNN AE	87.49 ± 3.30	1.52 ± 1.04	1.69 ± 1.13
CNN U-Net	93.73 ± 2.40	1.36 ± 1.02	0.71 ± 0.59
Swin U-Net	92.10 ± 3.39	1.70 ± 1.32	1.28 ± 1.14
Unforeseen with CT			
CNN AE	86.00 ± 3.48	1.62 ± 1.00	4.44 ± 2.11
CNN U-Net	88.64 ± 3.71	1.90 ± 1.29	5.29 ± 2.41
Swin U-Net	84.33 ± 4.63	2.18 ± 1.60	6.66 ± 3.11
Foreseen with MRI			
CNN AE	89.57 ± 3.33	1.40 ± 0.96	1.87 ± 1.05
CNN U-Net	93.09 ± 2.95	1.36 ± 1.11	0.75 ± 0.67
Swin U-Net	91.26 ± 3.70	1.68 ± 1.38	1.25 ± 1.18
Unforeseen with MRI			
CNN AE	78.01 ± 7.68	2.31 ± 1.39	5.99 ± 3.51
CNN U-Net	79.48 ± 7.12	2.83 ± 1.78	6.02 ± 3.55
Swin U-Net	75.71 ± 8.04	3.51 ± 2.10	4.83 ± 3.29

4.3.4. Comparison between different loss functions

Since our objective involves generating 3D images, we compared the performance of various loss functions commonly used in image processing [68], including mean absolute error (MAE), mean squared error (MSE), and structural similarity index measure (SSIM). Experiments were conducted based on a CNN U-Net using CT as an input medical image modality. Table 6 shows the mean and standard deviation values of evaluation metrics when using MAE, MSE, and SSIM as loss functions during network training. Despite employing different loss functions, there were no significant differences in the results. However, SSIM required additional computations, doubling the training time compared to MAE and MSE. The mean results were similar when using MAE and MSE, but the variance slightly increased when using the MAE. Therefore, we selected MSE as the loss function for the network training.

4.4. Evaluation with foreseen/unforeseen data

Foreseen data refers to the skull data that was included in the training process, while unforeseen data denotes skull data that was not used for training. To assess the robustness of our proposed network for new subject's data, we here presented performance evaluation using unforeseen dataset.

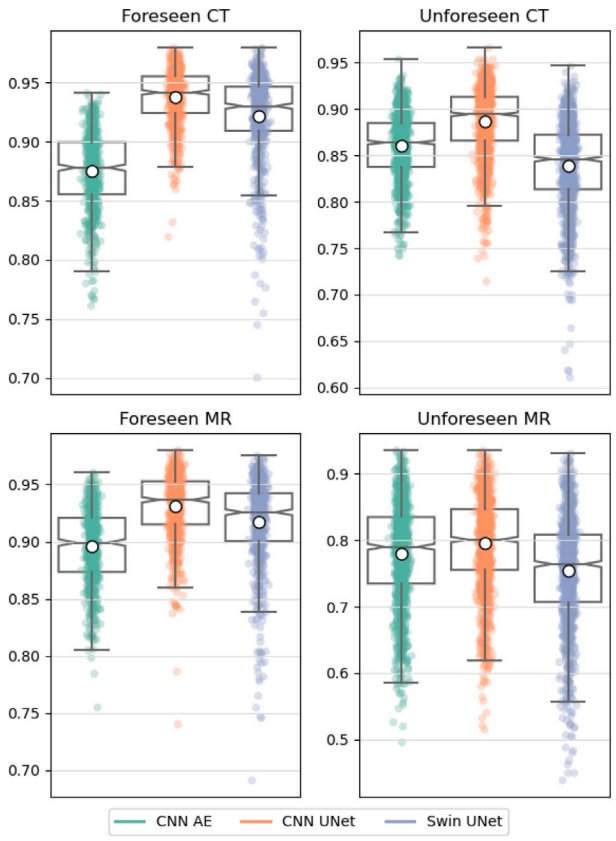


Fig. 7. Box plot of Dice scores for each network architecture and foreseen/unforeseen test data conditions. The horizontal line and circle in boxes respectively represent the median and mean values, while the box indicates the interquartile range (IQR: Spanning Q1 to Q3). The whiskers denote the lower (Q1-1.5IQR) and upper (Q3+1.5IQR) fences.

Table 8
Comparison of performance between transfer learning network and the network trained with single medical image modality (MR) on unforeseen data.

Modality	Dice (%)	D_p (mm)	Δ_p (%)
MR	79.48 ± 7.12	2.83 ± 1.78	6.02 ± 3.55
CT⇒MR	82.60 ± 5.83	2.27 ± 1.42	2.34 ± 1.80

Table 7 and Fig. 7 respectively shows the mean/standard deviation values of evaluation metrics and the box plots of Dice score when evaluated with foreseen and unforeseen test data. All network architectures experienced performance declines with unforeseen data. In the case of using CT images, the CNN AE proved most robust to variations in input data, though it had the drawback of low average performance. Performance of CNN U-Net with unforeseen data decreased, yet it still recorded a high Dice score of 88.64%. Swin Transformer recorded the lowest score with the most significant performance degradation. When using MR images as an input modality, all networks showed a significant drop in performance for unforeseen data. This performance decline was particularly pronounced in the Dice score and Δ_p. Fig. 8 illustrates exemplar intracranial pressure maps and the corresponding FWHM regions in the central yz-, xz-, xy-planes of the target data and those generated by the proposed network model.

4.5. Using transfer learning to enhance unforeseen score

Based on the results in Section 4.4, we found that using MR images as the input modality leads to a performance drop on unforeseen data. In contrast, using CT images as the input modality significantly reduces the performance decline compared to the use of MR images. Therefore,

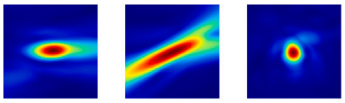
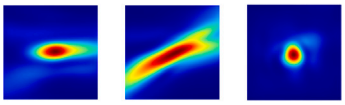
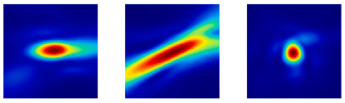
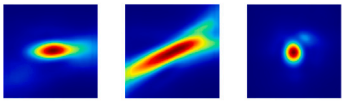
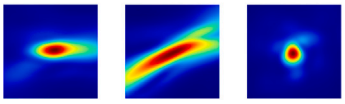
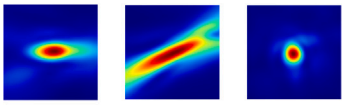
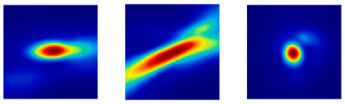
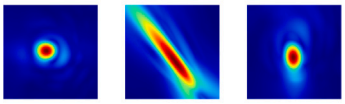
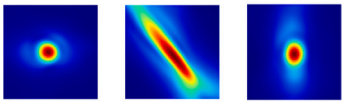
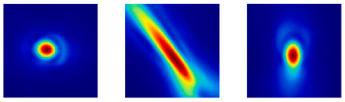
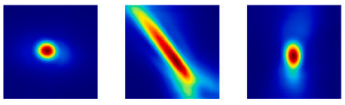
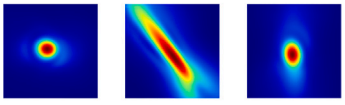
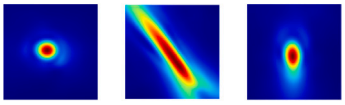
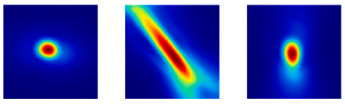
data	networks	pressure map			FWHM			Dice	D_p	Δ_p
		yz	xz	xy	yz	xz	xy			
	target							-	-	-
Foreseen CT	CNN AE							87.49	1.52	1.69
	CNN U-Net							93.73	1.36	0.71
	Swin U-Net							92.10	1.70	1.28
Foreseen MR	CNN AE							89.57	1.40	1.87
	CNN U-Net							93.09	1.36	0.75
	Swin U-Net							91.26	1.68	1.25
	target							-	-	-
Unforeseen CT	CNN AE							86.00	1.62	4.44
	CNN U-Net							88.64	1.90	5.29
	Swin U-Net							84.33	2.18	6.66
Unforeseen MR	CNN AE							78.01	2.31	5.99
	CNN U-Net							79.48	2.83	6.02
	Swin U-Net							75.51	3.51	4.83

Fig. 8. An example of target and generated pressure maps. Columns 3 to 5 in the table display three sectional views of target and network-generated intracranial pressure maps, while columns 6 to 8 show overlap of their FWHM regions with the target. In columns 6 to 8, the green-colored region represents the prediction's FWHM, and the black-colored edge indicates the overlap between the two.

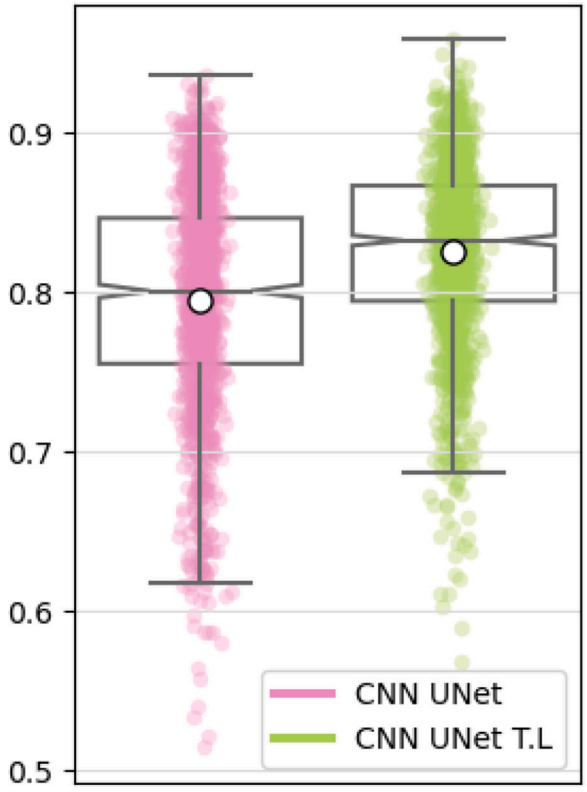


Fig. 9. Box plot of Dice scores for unforeseen subject in the MR-trained network (pink-colored plot) and the CT-MR transfer learning network (green-colored plot). The horizontal line and circle in boxes respectively represent the median and mean values, while the box indicates the interquartile range (IQR: Spanning Q1 to Q3). The whiskers denote the lower (Q1-1.5IQR) and upper (Q3+1.5IQR) fences.

Table 9
Comparison of simulation time for generating intracranial pressure map.

Methods	Time (s/data)
Conventional method	308.033
CNN AE	0.00227
CNN U-Net	0.00230
Swin U-Net	0.0232

we were motivated to test how the network trained with both CT and MR images (as the input modality) performs. Experiments were conducted based on the CNN U-Net structure, which showed the best performance on unforeseen MR data. We conducted the additional training using the same initial learning rate, learning rate scheduler, and epochs as the initial training.

Table 8 and Fig. 9 respectively show the mean/standard deviation values of evaluation metrics and the box plots of Dice score for the MR-trained network and the CT-MR transfer learning network. With respect to unforeseen MR data, we observed improved performance when the model trained with CT images was further trained with MR images as the input.

4.6. Evaluation of inference time

To verify the feasibility of real-time data generation, we measured the inference time of proposed networks. Based on the hardware specifications mentioned in Section 4, we averaged the time taken to generate a single pressure map from the inference process of 1200 unforeseen data samples. We also compared the obtained inference time with the conventional simulation time of the *k*-space method.

Table 9 shows the time each network takes to generate intracranial pressure map. CNN AE and CNN U-Net share essentially the same structure, differing only in the presence or absence of addition operations in skip connections, which resulted in negligible differences in inference time. In contrast, the Swin U-Net required a larger number of parameters for MLP and attention operations, resulting in inference times approximately 10 times longer than those of CNN-based networks. Nevertheless, the inference time of approximately 0.02 s suggests the feasibility of real-time operation. Moreover, compared to the conventional *k*-space method, our networks achieved over 13,000 times faster simulation time per data, resulting in significant computational time savings.

5. Discussion

For smart healthcare system, computer simulations have become essential tools for analyzing complex medical data and on-site information to achieve personalized precision treatment. However, the current computer simulations are impractical for clinical use due to their significant computational costs. To address this issue, our study took advantage of the rapid inference capabilities of deep learning to analyze complex medical information with various modalities, achieving real-time simulation of intracranial tFUS propagation. By integrating information from various modalities such as free-field pressure map, medical image, and transducer placement data into the network, we attained highly accurate prediction of intracranial acoustic pressure maps. Notably, by using transfer learning, we were able to predict intracranial pressure map with an 82% focal volume conformity using only MR images, eliminating the need for CT images when using the network.

We proposed three network structures, CNN AE, CNN U-Net, and Swin U-Net. Upon evaluating with a foreseen dataset, CNN U-Net exhibited the best performance across all evaluation metrics. Comparing with CNN AE with the same CNN structure, we found that the skip connections significantly impacts the network's ability to generate accurate intracranial pressure maps. When compared to Swin U-Net, the slightly lower performance associated with Swin U-Net suggests that network complexity does not always correlate with performance: selecting a network structure suitable for the given task is more critical than the network's complexity.

To assess the impact of multi-modal data on network performance, we conducted an ablation study by incrementally increasing the number of input modalities. We demonstrated that the performance of all networks improved when CT images were used as an additional input modality compared to using only the free-field pressure map. Since the skull structure is the critical factor in forming the intracranial pressure map, the provision of CT images enables the network to better understand the skull structure due to their accurate characterization of the skull macrostructure. The addition of the transducer placement vector did not significantly impact the performance, suggesting that the free-field pressure map already conveyed sufficient information about the transducer placements.

Predictive methods via numerical simulation have been limited by the obligatory use of CT imaging, as CT can clearly characterize the structures of hard tissues. However, CT scans can expose patients to unwanted radiation. Therefore, performing simulation solely with MR images would reduce the burden of extra radiation exposure while allowing for sufficient numerical modeling. We assessed the performance when MR data was used as the input medical image in our network, and found that MR data could achieve comparable accuracy to CT data. The MR data used in this study is T1-weighted, and we anticipate that utilizing T2-weighted MR data may further improve performance through better structural characterization through future investigation.

To demonstrate the robustness of the proposed networks, we conducted validation on an unforeseen dataset to determine whether the networks can extract generalized features from the skull data not seen

during training. For CT inputs, while the performance on unforeseen data slightly dropped compared to foreseen results, it still maintained a high focal volume conformity of over 88% in CNN U-Net. This indicates that our proposed network architecture can consistently extract features from the individual skull structures, based on the given input modalities to a certain extent.

Conversely, when MR images were used as the input modality, all networks exhibited a greater decline in performance for unforeseen data compared to using CT images. This implies that extracting a consistent and appropriate representation of skull information from MR images alone is still challenging. This is further evidenced by subsequent transfer learning experiments, where a network pre-trained on CT images showed improved performance on unforeseen data after additional training on MR images. This indicates that the network can achieve a degree of generalization with MR images when it pre-learns a clearer representation of the skull structure through CT images.

The limitations of this research include the following aspects. The first limitation is that additional training would be required for various transducer geometries and fundamental frequencies, as the data used for this study was generated from a single transducer geometry and frequency. Further research should focus on building datasets that account for diverse conditions, and optimizing the network accordingly. The second limitation is that the physical properties of the brain tissue inside the skull were assumed to be same as those of water, for simplicity. However, more accurate simulation would require considering the complex biological structures (e.g., white matter, gray matter, and cerebrospinal fluid) within the skull. Future research should incorporate simulations that account for intracranial tissue structures, potentially utilizing the information from the MR images. As the modeling of ultrasound propagation becomes more intricate, more advanced network architectures may be required to reflect this complexity.

6. Conclusion

In conclusion, our study presented multi-modal networks for real-time simulation of intracranial pressure map during tFUS treatment. Using dataset from 11 subjects, we demonstrated that our network can effectively and swiftly predict the physical phenomena occurring during tFUS therapy. The results present a future potential for ensuring safer and more accurate treatment, which will make significant progress toward LIFU therapy, where traditional monitoring methods have been limited. Our future research focus on developing networks that consider a wider range of transducer geometries and fundamental frequencies, aiming for broader applicability to diverse real-world treatment scenarios.

Code availability

The source code and description are available at https://github.com/Minjee-Seo/tFUS-Multimodal.

CRediT authorship contribution statement

Minjee Seo: Writing – original draft, Validation, Methodology, Formal analysis, Data curation. **Minwoo Shin:** Writing – review & editing, Methodology, Data curation. **Gunwoo Noh:** Data curation, Conceptualization. **Seung-Schik Yoo:** Writing – review & editing, Data curation. **Kyungho Yoon:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was supported by the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) under Grants RS-2024-00335185 and RS-2023-00220762.

References

[1] O. Naor, S. Krupa, S. Shoham, Ultrasonic neuromodulation, J. Neural Eng. 13 (2016) 031003, http://dx.doi.org/10.1088/1741-2560/13/3/031003.
[2] D. Coluccia, J. Fandino, L. Schwyzer, R. O'Gorman, L. Remonda, J. Anon, E. Martin, B. Werner, First non-invasive thermal ablation of a brain tumor with MR guided focused ultrasound, J. Ther. Ultrasound 2 (2014) 17, http://dx.doi.org/10.1186/2050-5736-2-17.
[3] K. Yoon, W. Lee, E. Chen, J.E. Lee, P. Croce, A. Cammalleri, L. Foley, S.-S. Yoo, Localized blood-brain barrier opening in ovine model using image-guided transcranial focused ultrasound, Ultrasound Med. Biol. 45 (2019) 2391–2404, http://dx.doi.org/10.1016/j.ultrasmedbio.2019.05.023.
[4] L. Xu, W. Lee, A. Rotenberg, M. Böhlke, K. Yoon, S.-S. Yoo, Localized disruption of blood albumin-phenytoin binding using transcranial focused ultrasound, Ultrasound Med. Biol. 46 (2020) 1986–1997, http://dx.doi.org/10.1016/j.ultrasmedbio.2020.04.011.
[5] N. Lipsman, M.L. Schwartz, Y. Huang, L. Lee, T. Sankar, M. Chapman, K. Hynynen, A.M. Lozano, MR-guided focused ultrasound thalamotomy for essential tremor: a proof-of-concept study, Lancet Neurol. 12 (5) (2013) 462–468, http://dx.doi.org/10.1016/S1474-4422(13)70048-6.
[6] V. Krishna, F. Sammartino, A. Rezai, A Review of the Current Therapies, Challenges, and Future Directions of Transcranial Focused Ultrasound Technology: Advances in Diagnosis and Treatment, JAMA Neurol. 75 (2) (2018) 246–254, http://dx.doi.org/10.1001/jamaneurol.2017.3129.
[7] S.-S. Yoo, A. Bystritsky, J.-H. Lee, Y. Zhang, K. Fischer, B.-K. Min, N.J. McDannold, A. Pascual-Leone, F.A. Jolesz, Focused ultrasound modulates region-specific brain activity, NeuroImage 56 (3) (2011) 1267–1275, http://dx.doi.org/10.1016/j.neuroimage.2011.02.058.
[8] W. Legon, T.F. Sato, A. Opitz, J. Mueller, A. Barbour, A. Williams, W.J. Tyler, Transcranial focused ultrasound modulates the activity of primary somatosensory cortex in humans, Nat. Neurosci. 17 (2) (2014) 322–329, http://dx.doi.org/10.1038/nn.3620.
[9] W. Lee, H.-C. Kim, Y. Jung, Y.A. Chung, I.-U. Song, J.-H. Lee, S.-S. Yoo, Transcranial focused ultrasound stimulation of human primary visual cortex, Sci. Rep. 6 (1) (2016) 34026, http://dx.doi.org/10.1038/srep34026.
[10] G. Darmani, T. Bergmann, K. Butts Pauly, C. Caskey, L. de Lecea, A. Fomenko, E. Fouragnan, W. Legon, K. Murphy, T. Nandi, M. Phipps, G. Pinton, H. Ramezanpour, J. Sallet, S. Yaakub, S. Yoo, R. Chen, Non-invasive transcranial ultrasound stimulation for neuromodulation, Clin. Neurophysiol. 135 (2022) 51–73, http://dx.doi.org/10.1016/j.clinph.2021.12.010.
[11] K. Yoon, W. Lee, J.E. Lee, L. Xu, P. Croce, L. Foley, S.-S. Yoo, Effects of sonication parameters on transcranial focused ultrasound brain stimulation in an ovine model, PLoS One 14 (10) (2019) e0224311, http://dx.doi.org/10.1371/journal.pone.0224311.
[12] F. Fregni, A. Pascual-Leone, Technology insight: noninvasive brain stimulation in neurology-perspectives on the therapeutic potential of rTMS and tDCS, Nat. Clin. Pract. Neurol. 3 (2007) 383–393, http://dx.doi.org/10.1038/ncpneuro0530.
[13] K. Hoy, P. Fitzgerald, Brain stimulation in psychiatry and its effects on cognition, Nat. Rev. Neurol. 6 (5) (2010) 267–275, http://dx.doi.org/10.1038/nrneurol.2010.30.
[14] C. Pasquinelli, L.G. Hanson, H.R. Siebner, H.J. Lee, A. Thielscher, Safety of transcranial focused ultrasound stimulation: A systematic review of the state of knowledge from both human and animal studies, Brain Stimul. 12 (6) (2019) 1367–1380, http://dx.doi.org/10.1016/j.brs.2019.07.024.
[15] C.W. Connor, K. Hynynen, Patterns of thermal deposition in the skull during transcranial focused ultrasound surgery, IEEE Trans. Biomed. Eng. 51 (10) (2004) 1693–1706, http://dx.doi.org/10.1109/TBME.2004.831516.
[16] K. Yoon, W. Lee, P. Croce, A. Cammalleri, S.-S. Yoo, Multi-resolution simulation of focused ultrasound propagation through ovine skull from a single-element transducer, Phys. Med. Biol. 63 (2018) 105001, http://dx.doi.org/10.1088/1361-6560/aabe37.
[17] Z. Wang, T. Komatsu, H. Mitsumura, N. Nakata, T. Ogawa, Y. Iguchi, M. Yokoyama, An uncovered risk factor of sonothrombolysis: Substantial fluctuation of ultrasound transmittance through the human skull, Ultrasonics 77 (2017) 168–175, http://dx.doi.org/10.1016/j.ultras.2017.02.012.
[18] P. Ghanouni, K.B. Pauly, W.J. Elias, J. Henderson, J. Sheehan, S. Monteith, M. Wintermark, Transcranial MRI-guided focused ultrasound: A review of the technologic and neurologic applications, AJR Am. J. Roentgenol. 205 (1) (2015) 150–159, http://dx.doi.org/10.2214/AJR.14.13632.
[19] A.N. Pouliopoulos, S.-Y. Wu, M.T. Burgess, M.E. Karakatsani, H.A. Kamimura, E.E. Konofagou, A clinical system for non-invasive blood–brain barrier opening using a neuronavigation-guided single-element focused ultrasound transducer, Ultrasound Med. Biol. 46 (1) (2020) 73–89, http://dx.doi.org/10.1016/j.ultrasmedbio.2019.09.010.

[20] H. Kim, A. Chiu, S. Park, S.-S. Yoo, Image-guided navigation of single-element focused ultrasound transducer, Int. J. Imaging Syst. Technol. 22 (3) (2012) 177–184, http://dx.doi.org/10.1002/ima.22020.

[21] W. Lee, H. Kim, Y. Jung, I.-U. Song, Y.A. Chung, S.-S. Yoo, Image-guided transcranial focused ultrasound stimulates human primary somatosensory cortex, Sci. Rep. 5 (2015) 8743, http://dx.doi.org/10.1038/srep08743.

[22] S. Brinker, P. Balchandani, A. Seifert, H.-J. Kim, K. Yoon, Feasibility of upper cranial nerve sonication in human application via neuronavigated single-element pulsed focused ultrasound, Ultrasound Med. Biol. 48 (2022) 1045–1057, http://dx.doi.org/10.1016/j.ultrasmedbio.2022.01.022.

[23] Y. Huang, P. Wen, B. Song, Y. Li, Numerical investigation of the energy distribution of Low-intensity transcranial focused ultrasound neuromodulation for hippocampus, Ultrasonics 124 (2022) 106724, http://dx.doi.org/10.1016/j.ultras.2022.106724.

[24] S. Leung, T. Webb, R.R. Bitton, P. Ghanouni, K.B. Pauly, A rapid beam simulation framework for transcranial focused ultrasound, Sci. Rep. 9 (2019) 7965, http://dx.doi.org/10.1038/s41598-019-43775-6.

[25] M. Bakator, D. Radosav, Deep learning and medical diagnosis: A review of literature, Multimodal Technol. Interact. 2 (2018) 47, http://dx.doi.org/10.3390/mti2030047.

[26] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (1) (2017) 221–248, http://dx.doi.org/10.1146/annurev-bioeng-071516-044442.

[27] L. Zhang, J. Lin, B. Liu, Z. Zhang, X. Yan, M. Wei, A review on deep learning applications in prognostics and health management, Ieee Access 7 (2019) 162415–162438, http://dx.doi.org/10.1109/ACCESS.2019.2950985.

[28] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, The rise of deep learning in drug discovery, Drug Discov. Today 23 (6) (2018) 1241–1250, http://dx.doi.org/10.1016/j.drudis.2018.01.039.

[29] M. Wang, Q. Zhang, S. Lam, J. Cai, R. Yang, A review on application of deep learning algorithms in external beam radiotherapy automated treatment planning, Front. Oncol. 10 (2020) 580919, http://dx.doi.org/10.3389/fonc.2020.580919.

[30] M. Choi, M. Jang, S.-S. Yoo, G. Noh, K. Yoon, Deep neural network for navigation of a single-element transducer during transcranial focused ultrasound therapy: Proof of concept, IEEE J. Biomed. Health Inform. 26 (11) (2022) 5653–5664, http://dx.doi.org/10.1109/JBHI.2022.3198650.

[31] M. Shin, Z. Peng, H.-J. Kim, S.-S. Yoo, K. Yoon, Multivariable-incorporating super-resolution residual network for transcranial focused ultrasound simulation, Comput. Methods Programs Biomed. 237 (2023) 107591, http://dx.doi.org/10.1016/j.cmpb.2023.107591.

[32] M. Shin, M. Seo, S.-S. Yoo, K. Yoon, TFUSFormer: Physics-guided super-resolution transformer for simulation of transcranial focused ultrasound propagation in brain stimulation, IEEE J. Biomed. Health Inform. (2024) 1–12, http://dx.doi.org/10.1109/JBHI.2024.3389708.

[33] C. Shen, D. Nguyen, Z. Zhou, S.B. Jiang, B. Dong, X. Jia, An introduction to deep learning in medical physics: advantages, potential, and challenges, Phys. Med. Biol. 65 (5) (2020) 05TR01, http://dx.doi.org/10.1088/1361-6560/ab6f51.

[34] Z. Chen, J. Wang, H. He, X. Huang, A fast deep learning system using GPU, in: 2014 IEEE International Symposium on Circuits and Systems, ISCAS, 2014, pp. 1552–1555, http://dx.doi.org/10.1109/ISCAS.2014.6865444.

[35] S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, A survey of deep learning techniques for autonomous driving, J. Field Robot. 37 (3) (2020) 362–386, http://dx.doi.org/10.48550/arXiv.1910.07738.

[36] M. Yamada, Y. Saito, H. Imaoka, M. Saiko, S. Yamada, H. Kondo, H. Takamaru, T. Sakamoto, J. Sese, A. Kuchiba, T. Shibata, R. Hamamoto, Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy, Sci. Rep. 9 (1) (2019) 14465, http://dx.doi.org/10.1038/s41598-019-50567-5.

[37] L. Guo, X. Xiao, C. Wu, X. Zeng, Y. Zhang, J. Du, S. Bai, J. Xie, Z. Zhang, Y. Li, X. Wang, O. Cheung, M. Sharma, J. Liu, B. Hu, Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos), Gastrointest Endosc. 91 (1) (2020) 41–51, http://dx.doi.org/10.1016/j.gie.2019.08.018.

[38] R.S. Andersen, A. Peimankar, S. Puthusserypady, A deep learning approach for real-time detection of atrial fibrillation, Expert Syst. Appl. 115 (2019) 465–473, http://dx.doi.org/10.1016/j.eswa.2018.08.011.

[39] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 689–696.

[40] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443, http://dx.doi.org/10.1109/TPAMI.2018.2798607.

[41] C. Zhang, Z. Yang, X. He, L. Deng, Multimodal intelligence: Representation learning, information fusion, and applications, IEEE J. Sel. Topics Signal Process. 14 (3) (2020) 478–493, http://dx.doi.org/10.1109/JSTSP.2020.2987728.

[42] J.N. Acosta, G.J. Falcone, P. Rajpurkar, E.J. Topol, Multimodal biomedical AI, Nature Med. 28 (9) (2022) 1773–1784, http://dx.doi.org/10.1038/s41591-022-01981-2.

[43] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, C.P. Langlotz, Contrastive learning of medical visual representations from paired images and text, in: Z. Lipton, R. Ranganath, M. Sendak, M. Sjoding, S. Yeung (Eds.), Proceedings of the 7th Machine Learning for Healthcare Conference, in: Proceedings of Machine Learning Research, Vol. 182, PMLR, 2022, pp. 2–25, http://dx.doi.org/10.48550/arXiv.2010.00747.

[44] B. Jing, P. Xie, E. Xing, On the automatic generation of medical imaging reports, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2577–2586, http://dx.doi.org/10.18653/v1/P18-1240.

[45] Y. Li, J. Zhao, Z. Lv, J. Li, Medical image fusion method by deep learning, Int. J. Cogn. Comput. Eng. 2 (2021) 21–29, http://dx.doi.org/10.1016/j.ijcce.2020.12.004.

[46] C. Wang, G. Yang, G. Papanastasiou, S.A. Tsaftaris, D.E. Newby, C. Gray, G. Macnaught, T.J. MacGillivray, DiCyc: GAN-based deformation invariant cross-domain information fusion for medical image synthesis, Inf. Fusion 67 (2021) 147–160, http://dx.doi.org/10.1016/j.inffus.2020.10.015.

[47] W. Tang, F. He, Y. Liu, Y. Duan, MATR: Multimodal medical image fusion via multiscale adaptive transformer, IEEE Trans. Image Process. 31 (2022) 5134–5149, http://dx.doi.org/10.1109/TIP.2022.3193288.

[48] Z. Ahmad, A. Tabassum, L. Guan, N.M. Khan, ECG heartbeat classification using multimodal fusion, IEEE Access 9 (2021) 100615–100626, http://dx.doi.org/10.1109/ACCESS.2021.3097614.

[49] J.T. Soto, J. Weston Hughes, P.A. Sanchez, M. Perez, D. Ouyang, E.A. Ashley, Multimodal deep learning enhances diagnostic precision in left ventricular hypertrophy, Eur. Heart J., Digit. Health 3 (3) (2022) 380–389, http://dx.doi.org/10.1093/ehjdh/ztac033.

[50] K. Chaudhary, O.B. Poirion, L. Lu, L.X. Garmire, Deep learning–based multi-omics integration robustly predicts survival in liver cancer, Clin. Cancer Res. 24 (6) (2018) 1248–1259, http://dx.doi.org/10.1158/1078-0432.CCR-17-0853.

[51] M. Kang, E. Ko, T.B. Mersha, A roadmap for multi-omics data integration using deep learning, Brief. Bioinform. 23 (1) (2022) http://dx.doi.org/10.1093/bib/bbab454.

[52] W. Burger, M.J. Burge, M.J. Burge, M.J. Burge, Principles of digital image processing, vol. 111, Springer, 2009.

[53] M.H. Moghari, P. Abolmaesumi, Point-based rigid-body registration using an unscented kalman filter, IEEE Trans. Med. Imaging 26 (12) (2007) 1708–1728, http://dx.doi.org/10.1109/tmi.2007.901984.

[54] T. Deffieux, E. Konofagou, Numerical study of a simple transcranial focused ultrasound system applied to blood-brain barrier opening, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 57 (2010) 2637–2653, http://dx.doi.org/10.1109/TUFFC.2010.1738.

[55] S. Pichardo, V. Sin, K. Hynynen, Multi-frequency characterization of the speed of sound and attenuation coefficient for longitudinal transmission of freshly excised human skulls, Phys. Med. Biol. 56 (2011) 219–250, http://dx.doi.org/10.1088/0031-9155/56/1/014.

[56] A. Gilat, MATLAB: An introduction with applications, John Wiley & Sons, 2017.

[57] B.E. Treeby, B.T. Cox, k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields, J. Biomed. Opt. 15 (2) (2010) 021314, http://dx.doi.org/10.1117/1.3360308.

[58] M. Tabei, T.D. Mast, R. Waag, A k-space method for coupled first-order acoustic propagation equations, J. Acoust. Soc. Am. 111 (2002) 53–63, http://dx.doi.org/10.1121/1.1421344.

[59] T. Mast, L. Souriau, D.-L. Liu, M. Tabei, A. Nachman, R. Waag, A k-space method for large-scale models of wave propagation in tissue, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 48 (2) (2001) 341–354, http://dx.doi.org/10.1109/58.911717.

[60] W. Chen, S. Holm, Fractional Laplacian time-space models for linear and nonlinear lossy media exhibiting arbitrary frequency power-law dependency, J. Acoust. Soc. Am. 115 (2004) 1424–1430, http://dx.doi.org/10.1121/1.1646399.

[61] B.E. Treeby, J. Jaros, A.P. Rendell, B.T. Cox, Modeling nonlinear ultrasound propagation in heterogeneous media with power law absorption using a k-space pseudospectral method, J. Acoust. Soc. Am. 131 (6) (2012) 4324–4336, http://dx.doi.org/10.1121/1.4712021.

[62] M. Tancik, P.P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J.T. Barron, R. Ng, Fourier features let networks learn high frequency functions in low dimensional domains, 2020, http://dx.doi.org/10.48550/arXiv.2006.10739.

[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017, http://dx.doi.org/10.48550/arXiv.1706.03762, CoRR, abs/1706.03762.

[64] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507, http://dx.doi.org/10.1126/science.1127647.

[65] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, 2015, http://dx.doi.org/10.48550/arXiv.1505.04597, CoRR, abs/1505.04597.

[66] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical vision transformer using shifted windows, 2021, http://dx.doi.org/10.48550/arXiv.2103.14030.

[67] L.R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302, http://dx.doi.org/10.2307/1932409.

[68] H. Zhao, O. Gallo, I. Frosio, J. Kautz, Loss functions for image restoration with neural networks, IEEE Trans. Comput. Imaging 3 (1) (2017) 47–57, http://dx.doi.org/10.1109/TCI.2016.2644865.