Multihead Attention U-Net for Magnetic Particle Imaging–Computed Tomography Image Segmentation

*Aniwat Juhong, Bo Li, Yifan Liu, Chia-Wei Yang, Cheng-You Yao, Dalen W. Agnew, Yu Leo Lei, Gary D. Luker, Harvey Bumpers, Xuefei Huang, Wibool Piyawattanametha, and Zhen Qiu**

Magnetic particle imaging (MPI) is an emerging noninvasive molecular imaging modality with high sensitivity and specificity, exceptional linear quantitative ability, and potential for successful applications in clinical settings. Computed tomography (CT) is typically combined with the MPI image to obtain more anatomical information. Herein, a deep learning-based approach for MPI-CT image segmentation is presented. The dataset utilized in training the proposed deep learning model is obtained from a transgenic mouse model of breast cancer following administration of indocyanine green (ICG)-conjugated superparamagnetic iron oxide nanoworms (NWs-ICG) as the tracer. The NWs-ICG particles progressively accumulate in tumors due to the enhanced permeability and retention (EPR) effect. The proposed deep learning model exploits the advantages of the multihead attention mechanism and the U-Net model to perform segmentation on the MPI-CT images, showing superb results. In addition, the model is characterized with a different number of attention heads to explore the optimal number for our custom MPI-CT dataset.

1. Introduction

Magnetic particle imaging (MPI) is a highly sensitive imaging modality initially introduced in 2005.[1–3] Unlike traditional imaging techniques such as magnetic resonance imaging (MRI), sonography, computed tomography (CT), and X-ray, MPI is not employed for structural imaging purposes. Nevertheless, it is a tracer imaging modality akin to positron emission tomography (PET) and single-photon emission computed tomography (SPECT). The concept of MPI is to detect the 3D distribution of superparamagnetic iron-oxide nanoparticles (SPIONs) with extraordinary contrast and sensitivity, allowing us to track and quantify the tracer materials effectively. Biocompatibility is one of the essential features for using biomaterials, particularly MPI tracers (iron oxide particles), for in

A. Juhong, B. Li, Y. Liu, Z. Qiu
Department of Electrical and Computer Engineering
Michigan State University
East Lansing, Michigan 48824, USA
E-mail: qiuzhen@msu.edu

A. Juhong, B. Li, Y. Liu, C.-W. Yang, C.-Y. Yao, X. Huang,
W. Piyawattanametha, Z. Qiu
Institute for Quantitative Health Science and Engineering
Michigan State University
East Lansing, Michigan 48824, USA

C.-W. Yang, X. Huang
Department of Chemistry
Michigan State University
East Lansing, Michigan 48824, USA

C.-Y. Yao, X. Huang, Z. Qiu
Department of Biomedical Engineering
Michigan State University
East Lansing, Michigan 48824, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202400007.

DOI: 10.1002/aisy.202400007

D. W. Agnew
Department of Pathobiology and Diagnostic Investigation
College of Veterinary Medicine
Michigan State University
East Lansing, Michigan 48824, USA

Y. L. Lei
Department of Periodontics and Oral Medicine
University of Michigan
Ann Arbor, Michigan 48109, USA

G. D. Luker
Departments of Radiology and Biomedical Engineering
University of Michigan
Ann Arbor, Michigan 481054, USA

H. Bumpers
Department of Surgery
Michigan State University
East Lansing, Michigan 48823, USA

W. Piyawattanametha
Department of Biomedical Engineering
School of Engineering
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520, Thailand

vivo applications and clinical trials. Nanoworms (NWs) are biocompatible iron oxide particles widely used for biomedical applications. NWs include a considerably lower inflammatory response than spherical iron oxide nanoparticles (NPs),[4] and they are a nanostructure with an elongated assembly of iron oxide (IO).[5] This structure can potentially augment the NPs' capability for circulation and tumor targeting. Due to their nanoscale dimensions, NWs can remain in tumors longer than pure fluorescence contrast agents, also recognized as the enhanced permeability and retention (EPR) effect.[6,7] In addition, MPI signal can only be detected from the administered tracer providing an image without background as well as improving signal-to-noise ratios. Indeed, the development of MPI involved strengthening the existing imaging modalities (MRI, PET, SPECT, etc.). For instance, PET and SPECT tracers typically have half-lives in a range of minutes to hours, whereas the MPI tracer can last for several days to weeks.[8] Therefore, MPI is more eminently suitable for dynamic imaging applications than traditional tracer imaging methods. Numerous prototypes and commercial MPI scanners have demonstrated impressive results in in vivo studies for vascular imaging,[9–11] oncology,[12–14] and cell tracking.[15,16] The MPI system for humans is under development and may become available in the near future.[17] Like PET, an MPI image is frequently combined with a CT image for registering the particle signal (the MPI image) and the anatomical information (the CT image). This will enhance the diagnostic potential by identifying the precise location of functional events in the body.[18] Using MPI-CT images for data analysis, rather than solely relying on MPI, is worthwhile, particularly for in vivo data analysis. Initially, users may accurately and expeditiously determine the location of the MPI signal. Occasionally, the MPI signal of a tumor located close to organs showing a substantially strong signal (such as the liver, kidney, and spleen) could be significantly attenuated. As a result, the users may need to pay more attention to this valuable information. Utilizing CT information together with MPI could enhance the accuracy and meticulousness of data analysis.

As the MPI system is costly and a newly emerging system, the availability of MPI data is limited. Nevertheless, some research groups have employed artificial intelligence or deep learning techniques for MPI applications. Image reconstruction and enhancement datasets do not require ground truth labeling (unsupervised learning), which can directly transform one domain (input) to another domain (ground truth). In addition, the MPI data simulation is undemanding; therefore, the simulated large datasets were utilized in the MPI image reconstruction and resolution enhancement.[19–21] By contrast, the supervised learning problem of MPI image segmentation is challenging to prepare the datasets due to the need for ground truth labeling. Consequently, the MPI image segmentation dataset is restricted, and applying deep learning (typically requires a large dataset) is demanding for this problem. Thus far, only a machine learning technique has been utilized for the MPI image segmentation[22] with limited performance.

Recently, image processing based on deep learning has become a promising approach for medical applications due to the rapid development of computation technologies for image classification,[23–25] regression,[26–28] reconstruction,[29–31] and segmentation.[32–36] Deep learning models contain a large number of function approximators. As a result, the models without further modifications tend to neglect essential parts of the input and focus on others. The use of the attention mechanism[37] is one of the practical approaches to remedy this problem. The attention mechanism is an ingenious and powerful technique, allowing neural networks to focus on meaningful parts of an input tensor. This mechanism is the key innovation behind numerous successful deep learning architectures, such as TransUnet,[38] BRET,[39] and Swin transformer.[40] Multiplicative attention (Luong attention)[41] and additive attention (Bahanau attention)[42] are two initial instances of attention sparking the revolution. Since multiplicative attention implements matrix multiplication for calculating the output, it is more memory efficient in practice and faster than additive attention. However, additive attention can be superior to multiplicative attention for large dimensional input features.[43] The U-Net architecture[44] is a widely recognized convolutional neural network (CNN) that has achieved prominence in the field of medical image segmentation due to its simplicity and remarkable performance. The original U-Net architecture contains two main components: an encoder and a decoder. The skip-connection (SC) mechanism is added to the same dimensional encoder and decoder. Essentially, it combines spatial information from the downsampling path (encoder) with the upsampling path (decoder) to retain marvelous spatial information. In addition, the SC mechanism allows the gradient to readily propagate back to update the weights (learnable parameters). However, the SC mechanism brings along the poor feature representation from the encoder path. The attention U-Net architecture[45,46] can tackle this problem by implementing the attention mechanism at the SC, allowing the model to actively suppress actions at irrelevant features. This reduces the computational resources wasted on irrelevant activations and provides superior network generalization. The attention mechanism applied in the attention U-Net is called the attention gates (AGs)[45,46] based on additive attention. The CNN model with AGs can be easily trained from scratch and boost the model's performance by automatically learning to focus on some crucial features without additional supervision. Available MPI data are remarkably limited for a computational study of robust MPI image quantification. Herein, we propose a multihead attention U-Net model for MPI-CT image segmentation. The MPI-CT images acquired from mice with breast tumors were manually labeled as the ground truths for training the model. The attention U-Net model[45,46] inspires the proposed model. Still, we apply the attention mechanism in parallel (multihead attention) to step up the model capability for focusing on noteworthy features.

2. Experimental Section

An extensive overview of the workflow involved in training the proposed multihead attention U-Net model is shown in **Figure 1**. First, NWs were synthesized by the coprecipitation method of Fe^{2+} and Fe^{3+} salts with the polysaccharide dextran coating, as depicted in Figure 1a1, the particles were then conjugated with indocyanine green (ICG), resulting in the formation of conjugated superparamagnetic iron oxide nanoworms referred to as NWs-ICG.[47]

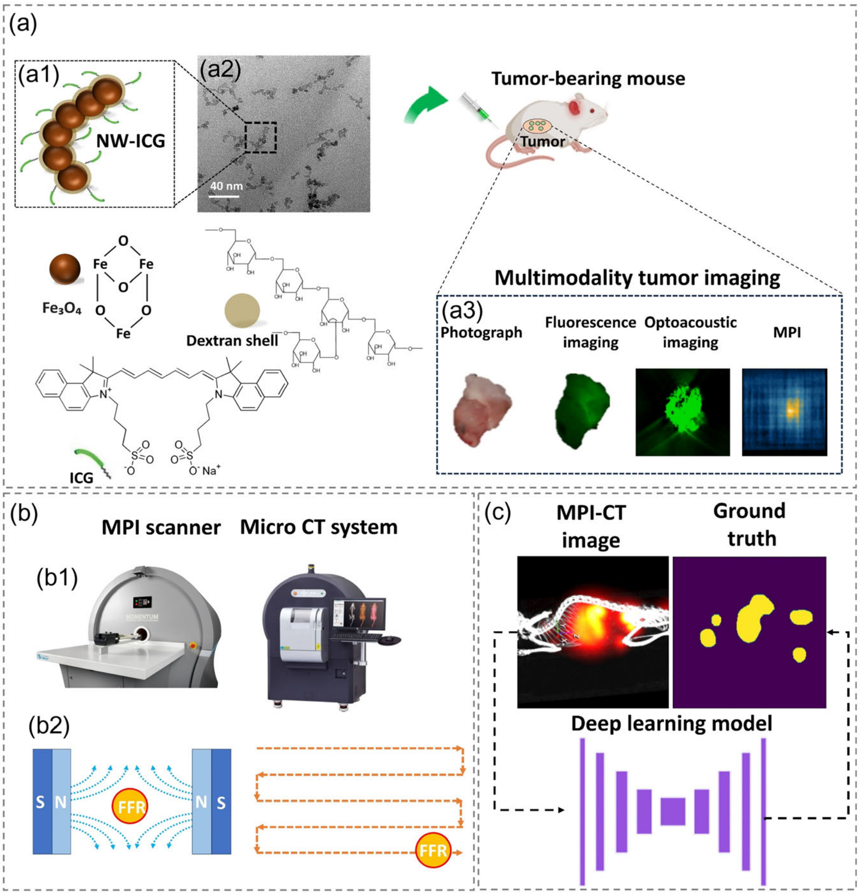

Figure 1. Overview of MPI-CT image segmentation using the custom dataset. a) An injected NWs-ICG breast tumor mouse: a1) the chemical structure of NWs-ICG; a2) TEM image of NWs-ICG particles with a scale bar of 40 nm; a3) the multimodality imaging (fluorescence, optoacoustic, and MPI) of the tumor dissected from the NWs-ICG-injected mouse. b) MPI-CT image acquisition: b1) the MPI scanner and the micro-CT imaging system; b2) illustration of the MPI principle. c) Ground truth labeling in MPI-CT image segmentation.

In addition, we also acquired a transmission electron microscopy transmission (TEM) image of NWs-ICG particles as shown in Figure 1a2. With this structure, the detection of NWs-ICG can be achieved by fluorescence imaging and optoacoustic imaging, in addition to the use of MPI as shown in Figure 1a3. Thus, this offers captivating prospects for a multimodal imaging study.

It is essential to clarify that the greater the concentration of nanoparticles (NPs), the higher the intensity of MPI will be. This isdifferent from fluorescence imaging, particularly the ICG fluorescence dye. Fluorescence intensity is not related to an ICG concentration in a linear manner. It could increase in a low concentration range of the ICG.[48] In addition, for an in vivo experiment, fluorescence imaging is prone to photobleaching, which is challenging for prolonged imaging. Therefore, in this work, we only focus on MPI as the main modality, which can easily be further quantitatively analyzed by the concentration of NWs-ICG. Apart from the potential capability to accumulate in the tumor of NWs-ICG, the main benefit of conjugating NWs with ICG is to confirm that the signal we obtain from the MPI is genuinely from NWs-ICG as multimodality can be used to acquire the data of the same sample.

In this work, a mouse with breast tumors was injected with NWs-ICG through the intravenous administration injection method, followed by MPI-CT image acquisition. Figure 1b1 shows the MPI and micro-CT systems used in this work. The fundamental concept of MPI is illustrated in Figure 1b2. In short, an intense magnetic field is generated by two permanent magnets, and the inside of this magnetic field contains a small area with low magnetic field intensity known as the field-free region (FFR). By rapidly moving the FFR across the imaging volume, the magnetization of SPIONs passing through the FFR induces a signal (oscillating changes in magnetization) in the imager's receive coil. In other words, SPIONs not passing through the FFR donot generate a signal in the receiver coil due to a strong magnetic field outside the FFR inhibiting SPIONs from rotating. Lastly, the MPI-CT images were manually labeled as the ground truths for training the deep learning model as shown in Figure 1c.

2.1. Dataset Preparation

To acquire a custom MPI-CT image dataset, MMTV-PyMT transgenic mice with breast cancer were intravenously injected with NWs-ICG at the concentration and volume of 2 mg mL^{-1} and 400 μL, respectively. All procedures performed on animals were approved by the Institutional Animal Care & Use Committee (IACUC) of Michigan State University (Protocol #: 2021000095). The Momentum MPI scanner (Magnetic Insight, Inc., Alameda, CA, USA) was employed to acquire the 3D MPI images of the NWs-ICG-injected mice. The scanner was configured with the following parameters: 3D scan mode, Z FOV 10.0 cm, number of projections 21, and selection field gradient 5.7 T m^{-1}. The micro-CT system (PerkinElmer, Inc., Hopkinton, MA, USA) with a speed scan mode and voltage of 90 kV was then used to acquire the corresponding CT images. Finally, 3D MPI-CT images were reconstructed using VivoQuant software (Magnetic Insight, Inc., Alameda, CA, USA). The imaging was performed at four different time points: 1, 24, 48, and 72 h after injection. Therefore, with one mouse, we can obtain 3D datasets at these four different time points. However, we only focus on 2D images in this work. To obtain the 2D image dataset, the 3D images were rotated with random angles for capturing the 2D images, and we had to ensure that the perspectives or rotation angles were not the same (0 or 180 degrees from the existing images) for the data cleaning purpose. **Figure 2**a shows the MPI-CT images of the NWs-ICG-injected mouse 1–72 h postinjection. MPI signal areas from MPI-CT images were manually labeled as the ground truths for training the segmentation deep learning model. There were 104 2D MPI-CT images and their corresponding ground truths from four different mice used for this study (91 images for a training dataset, 4 images for a validation dataset, and 9 images for a testing dataset). To affirm that there were NWs-ICG particles in the tumor tissues, after acquiring MPI-CT images, the tissues were dissected from the mice and preserved in a solution of 10% neutral buffered formalin (NBF). These NBF-fixed tissues were embedded in paraffin, followed by sectioning with a thickness of 5 µm and staining with Prussian blue to detect ferric from iron and hematoxylin and eosin (H&E). All histological procedures were carried out by the Michigan State University investigative histopathology laboratory. Figure 2c,d shows the Prussian blue stained histology image of one of the dissected tumors from the NWs-ICG-injected mouse acquired by a commercially available microscope (Nikon Eclipse Ci, Nikon Inc, Tokyo, Japan).

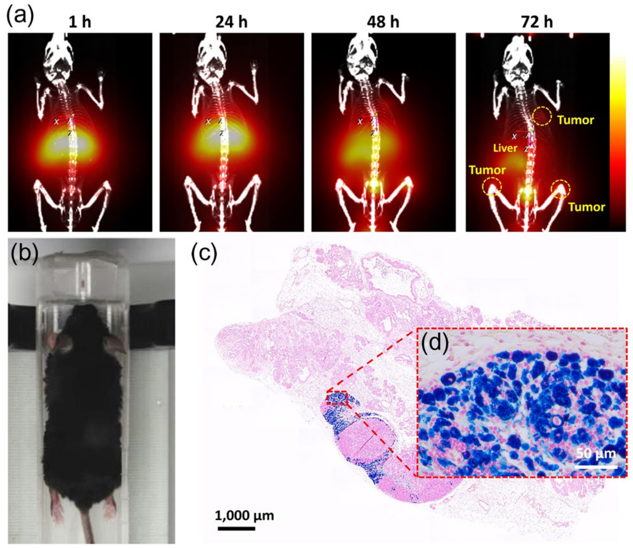

Figure 2. a) MPI-CT images of the NWs-ICG-injected mouse acquired from 1 to 72 h postinjection. The yellow-dashed circles (MPI-CT image at 72 h) show the MPI signal of NWs-ICG from the tumors. b) Photograph of the NWs-ICG-injected mouse. c,d) Prussian blue-stained histological image of the breast tumor dissected from the NWs-ICG-injected mouse acquired by 10× and 40× magnifications, respectively.

2.2. Multihead Attention U-Net

The U-Net architecture was originally designed for semantic segmentation tasks with a "U-shaped" encoder–decoder network associated with the use of a concatenating feature map from encoder to decoder. Attention is a mechanism that helps a neural network to highlight meaningful and relevant features. In other words, it helps the neural network to enhance a generalization capability by weighting different areas of the input image. Using the attention mechanism, high-relevant areas will be multiplied with large weights, whereas low-relevant areas will be multiplied with small weights. These weights are learnable parameters that are updated during the training process. U-Net employs SC to avoid imprecisely generating spatial information during upsampling. However, this includes numerous redundant low-level features (poor feature representation). This problem can be remedied by adding the attention mechanism, so-called AGs, at the SCs, as shown in **Figure 3**a to suppress activation in irrelevant areas; thus, it can reduce the number of redundant features brought across. The proposed model replaced a single AG in the original attention U-Net with parallel AGs in each SC. This modification allows the model to collect and incorporate more salient information effectively. In addition, employing parallel AGs enables the model to simultaneously process input from distinct representation subspaces at numerous locations.[49]

The first part of the proposed model is the encoder (the left side of Figure 3a). The input image is progressively filtered and downsampled by applying a convolution block, then a rectified linear unit (ReLU), and max-pooling 2 × 2 filters with a stride of 2. Furthermore, the number of feature channels is doubled at each downsampling step. The second part is multihead attention gates (MH-AGs). The features propagated through the SCs are filtered by exploiting these MH-AGs, which can help the model localize and focus on relevant features without cropping regions of interest. The third part is the decoder (the right side of Figure 3a). It consists of a concatenation of the attention weights from the MH-AG layer, a convolution block with the ReLU activation function, and a feature map upsampling followed by a 2 × 2 upconvolution resulting in a reduction of the number of feature channels by half. Figure 3b shows the MH-AG architecture employed between the encoder and decoder of the U-Net in Figure 3a. MH-AG is a parallel mechanism block that minimized the need for training a significant number of weights (learnable parameters) to enhance the performance of the U-Net model. Moreover, the MH-AG adopts the same transformation in all branches to minimize the need to adjust hyperparameters in each branch manually. The output of each branch in MH-AG is obtained by performing elementwise multiplication between

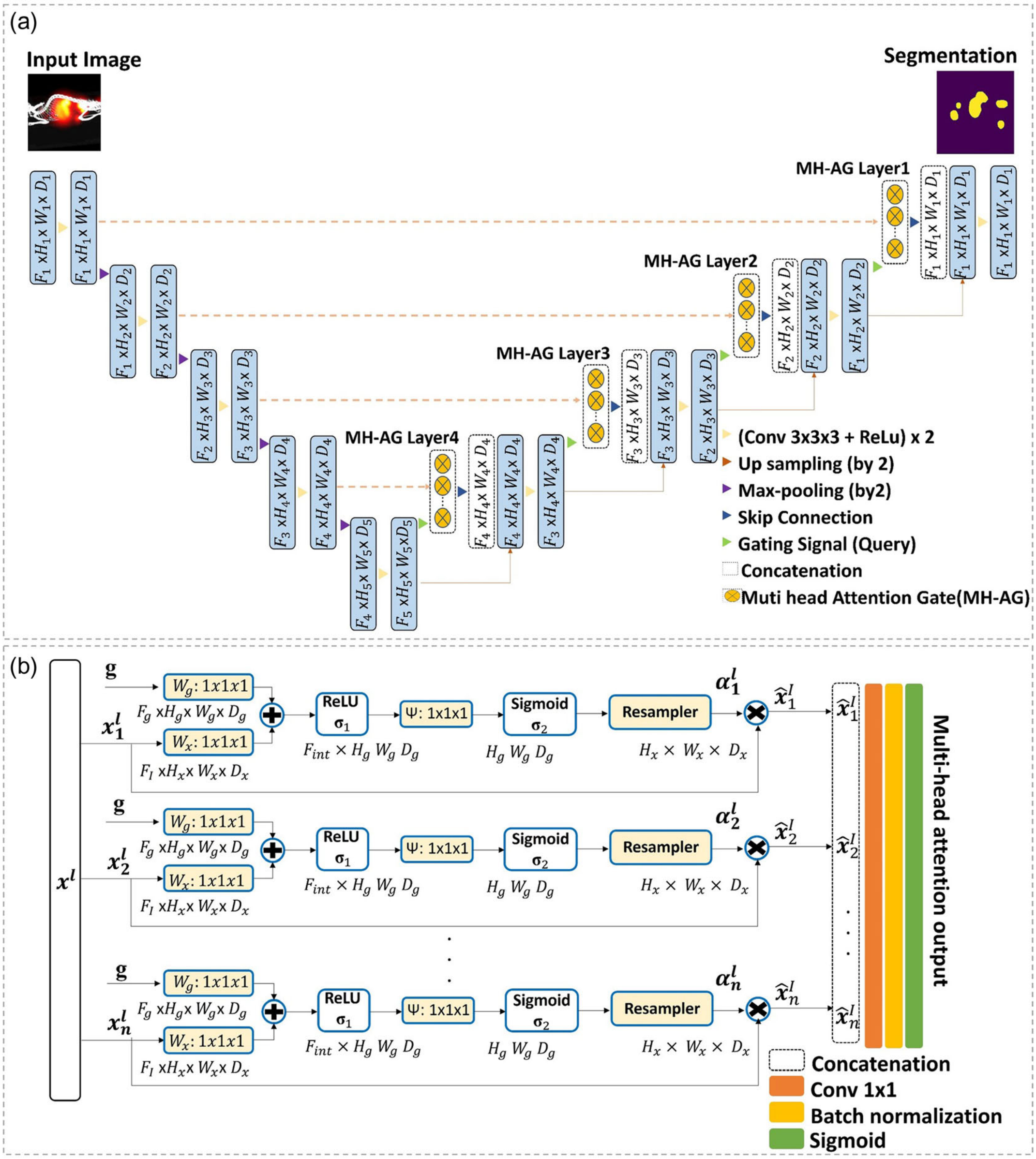

Figure 3. Schematic of the multihead attention U-Net (the proposed model) for MPI-CT image segmentation. a) The left side of the schematic represents the encoder blocks; the tensor is progressively downsampled by a factor of 2 (e.g., $H_1 = H_5/16$); the right side represents the decoder blocks; the tensor is gradually upsampled by a factor of 2. The MH-AGs are applied between the encoder and decoder to assign weights (learnable parameters) to noteworthy features. b) MH-AG architecture (n is the number of attention heads). Input features (x_n^{l}) are scaled with attention coefficients (α_n^{l}) computed in each branch of MH-AG. The gating signal (g) collected from a coarser scale provides activations and contextual information, which is applied to determine spatial regions. The output of each branch is then concatenated before feeding to the convolutional layer, batch normalization, and sigmoid function to compute the final result of MH-AG.

the input feature maps and attention coefficients ($\hat{x}_n^{\mathrm{l}} = x_i^{\mathrm{l}} \cdot \alpha_i^{\mathrm{l}}$), allowing the model to identify salient information. A gating vector g_i vector is taken from the next lowest layer of the network (better feature presentation) that encompasses contextual information, which can be used to suppress lower-level feature response electively.

To identify focus areas, g_i is assigned to each pixel by performing summed elementwise with x_i^{l}. As a result, aligned weights become larger, whereas unaligned weights become smaller. The gating coefficient is derived through the utilization of additive attention mathematically represented as follows

$$q_{\mathrm{att}}^{\mathrm{l}} = \psi^{\mathrm{T}}\left(\sigma_1\left(W_x^{\mathrm{T}} x_i^{\mathrm{l}} + W_{\mathrm{g}}^{\mathrm{T}} g_i + b_{\mathrm{g}}\right) + b_{\psi}\right) \tag{1}$$

$$\sigma_i^{\mathrm{l}} = \sigma_2\left(q_{\mathrm{att}}^{\mathrm{l}}\left(x_i^{\mathrm{l}}, g_i; \Theta_{\mathrm{att}}\right)\right) \tag{2}$$

where $\sigma_2(x_i) = \frac{1}{1+\exp(-x_i)}$ represents the sigmoid activation function and Θ_{att} represents a group of parameters that comprises linear transformation $w_x \in R^{\mathrm{F_l} \times \mathrm{F_{int}}}$, ${}_{\mathrm{g}} \in R^{\mathrm{F_g} \times \mathrm{F_{int}}}$, $\psi \in R^{\mathrm{F_l} \times \mathrm{F_{int}}}$, and bias terms $b_\psi \in R$, and $b_{\mathrm{g}} \in R^{\mathrm{F_g} \times \mathrm{F_{int}}}$. Channel-wise $1 \times 1 \times 1$ convolution for the input tensor is employed for computing the linear transformations.

2.3. Loss Function

Dice loss is widely used for medical image segmentation by comparing the similarity of two binary images (ground truth segmentation and predicted segmentation). Since our custom MPI-CT image dataset was limited and we wanted to prove the concept that multihead attention can potentially enhance the model performance for MPI-CT image segmentation, the dice loss was simply used to train all models for a performance comparison purpose. Equation (3) shows the dice loss function

$$\mathrm{Dice\,loss}(y, \bar{y}) = 1 - \frac{(2y\bar{y} + 1)}{(y + \bar{y} + 1)} \tag{3}$$

where y represents the ground truth and $\bar{y}$ represents the predicted segmentation generated by a deep learning model. After assembling all the parts for building the models, the MPI-CT images and their corresponding segmentation masks were then utilized to train the models as inputs and ground truths, respectively with the following hyperparameters: an Adam optimizer[50] with a learningrate of 5×10^{-4}, a batch size of 8, and 60 epochs. All the models in this study were trained on a personal computer equipped with an 11th Gen Intel core i7-11700k CPU, 64 GB of RAM, and an NVIDIA RTX 3090 graphic card.

3. Experimental Results

3.1. Gradient-Weighted Class Activation Maps

Gradient-weighted class activation mapping (Grad-CAM)[51] is a class-discriminative localization technique. It can generate a visual representation of a CNN-based model without altering the model itself. Grad-CAM leverages the gradient information flowing through a specific convolutional layer to assign crucial weights to each neuron to determine a particular decision of interest. This gradient information is then used to calculate the localization map visualized as a heat map image. In short, the intuitive interpretation of Grad-CAM is based on the concept that the model must observe some pixels and decide what object is present in the image, which can be interpreted as a gradient in mathematical terms. To compute Grad-CAM, the equations below are applied. Equation (4) is used to calculate the neuron's important weight (α_k^c) by calculating the global average pooling of the gradient from backpropagation. α_k^c is then employed to calculate the localization map Grad-CAM as shown in Equation (5)

$$\alpha_k^c = \frac{1}{Z}\left(\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}\right) \tag{4}$$

$$L_{\mathrm{Grad-CAM}}^{c} = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{5}$$

where $\frac{\partial y^c}{\partial A_{ij}^k}$ is the gradient from backpropagation, A^k is the feature map activation of a convolutional layer, i and j represent the width and height dimensions of the input tensor, Z is the total number of elements ($i \times j$), α_k^c is the neuron import weight, and $L_{\mathrm{Grad-CAM}}^{c}$ is the localization map Grad-CAM (coarse heat map).

Grad-CAM is applied to each multihead attention layer (MH-AG layer 1–4) output in order to characterize and understand the multihead attention U-Net model behavior. The attention weights of different MH-AG layers (the SC outputs) are visualized as shown in **Figure 4**. Figure 4a shows the input image, ground truth, and the segmentation results of 6-head, 4-head, and 2-head attention U-Net models, as well as the result of the original U-Net model (without an attention head). Figure 4b shows the Grad-CAM results of the corresponding attention U-Net models and the original U-Net model.

According to these Grad-CAM results and final segmentation outputs, the 4-head attention U-Net model can exceptionally perform MPI-CT image segmentation and surpass 6-head and 2-head attention U-Net models since it can focus on more meaningful features and predict a more accurate result. It is interesting to note that each SC layer output of the 4-head

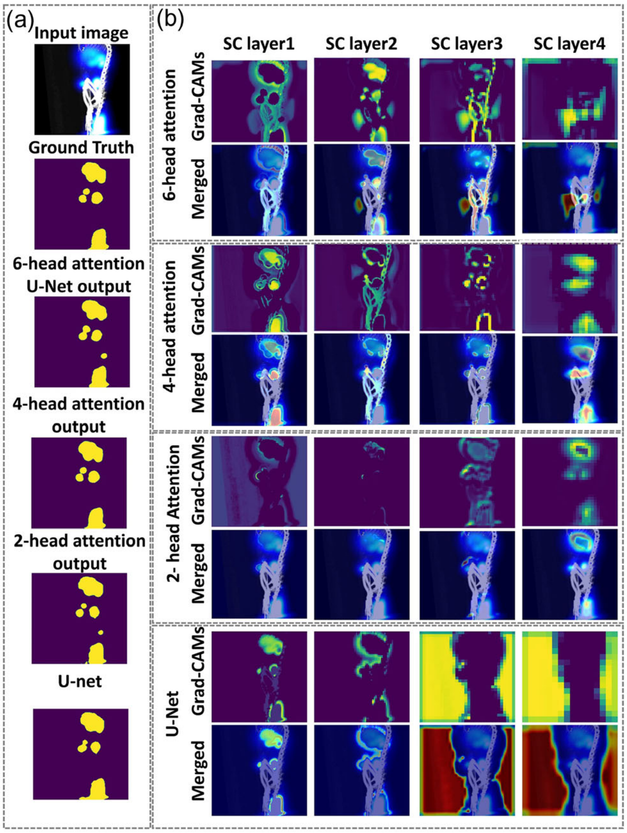

Figure 4. A comparison of Grad-CAMs results of the 2-head attention, 4-head attention, 6-head attention, and original U-Net models. a) Input MPI-CT image, segmentation ground truth, and outputs of each architecture. b) The Grad-CAM results of the SC outputs at different layers; SC outputs of multihead attention architectures are obtained by using MH-AG, whereas SC outputs of traditional U-Net are directly from the encoder blocks without applying MH-AG.

attention U-Net model pays attention to different meaningful features, the SC layer 4 pays attention to the overall boundary of the MPI signal, the SC layer 3 focuses on the increasingly precise boundary of the MPI signal, the SC layer 2 changes the focus from the boundary of the MPI signal to the skeleton (bone structure, i.e., CT image), and the SC layer 1 entirely focuses on the real target MPI signal. With these different meaningful features, the learnable parameters of the model can be assigned to pay attention to the relevant features and circumvent irrelevant features for the final prediction. However, the 2-head and 6-head attention U-Net models behave in different ways. The SC layers 4 and 3 of the 2-head attention U-Net focus on somewhat the same features (the boundary of MPI signal areas) and the SC layers 2 and 1 poorly focus on essential features. Although the SC layers 1, 2, and 3 of the 6-head attention U-Net can perform better than the 2-head attention model, the SC layer 4 pays attention to partially relevant features. Indeed, the optimal number of attention heads depends on the tasks we desire to train the deep learning model and the data features. If there are a larger number of important features, the higher number of attention heads could potentially help the model perform better by capturing more essential information. Nevertheless, the excessive number of attention heads could lead to less impressive performance, according to the Grad-CAM results illustrated in Figure 4 and our quantitative experiment discussed in the next section. Furthermore, we also compare Grad-CAM of the original U-Net (without MH-AGs) to the proposed models. At deeper layers (SC 3-4), the U-Net model cannot focus on relevant features. This clearly demonstrates that integrating MH-AGs with U-Net provides more generalization and assists the model to capture the important features to enhance the model's performance.

3.2. Implementation and Evaluation Metrics

Intersection over Union (IoU) is commonly used to evaluate the similarity between a predicted segmentation area and its ground truth.[33] The concept of IoU is to quantify the common area of the ground truth and prediction mask (intersection) divided by the entire number of pixels present across both the prediction mask and ground truth (union) as shown in the following equation

$$\mathrm{IoU} = \frac{\text{ground truth} \cap \text{prediction}}{\text{ground truth} \cup \text{prediction}} \tag{6}$$

The IoU ranges from 0 to 1 (0 to 100%), with 0 indicating no overlapping area, whereas 1 indicates impeccably overlapping area.

The dice similarity coefficient (DSC) is another well-known parameter used to evaluate the similarity between the predicted area (our output) and ground truth.[32] The DSC can be calculated following the equation

$$\mathrm{DSC} = \frac{2|\text{ground truth } \cap \text{prediction}|}{|\text{ground truth}| + |\text{prediction}|} \tag{7}$$

Precision is defined as the ratio of true-positive results to the total number of positive results, which is the summation of true positive and false positive as shown in Equation (8)

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{8}$$

Sensitivity, also known as recall, is the number of true-positive results over the summation of true-positive and false-negative results as shown in Equation (9)

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

Accuracy, also known as the Rand index, is the number of correct predictions divided by the total number of predictions as shown in Equation (10)

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{10}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

As previously stated, if the number of attention heads is excessive, the performance of a deep learning model based on the attention heads could deteriorate. Thus, we characterized the number of attention heads and employed Dice and IoU as the representative benchmarks. **Figure 5** illustrates the characterization results of the U-Net based on the different number of attention heads. With regard to the plot of Dice/IoU scores vs the number of attention heads, it begins at 0.889/0.804 with the 1-head attention architecture, and it gradually increases and then reaches the highest score at 0.909/0.835 with the 4-head attention architecture before declining progressively to 0.906/0.829 and 0.901/0.822 with 5 and 6 attention heads, respectively. Therefore, the multihead attention U-Net with 4 heads is the optimal model providing the best result for the MPI-CT image segmentation.

Table 1 shows the comprehensive characterization results of MPI-CT image segmentation of deep learning models with different architectures. Apart from using Dice and IoU scores as model evaluation metrics, we also characterized the performance of each model using accuracy, precision, and recall. Overall, the 4-head attention U-Net model can outperform other multihead attention U-Net models including the original U-Net model as

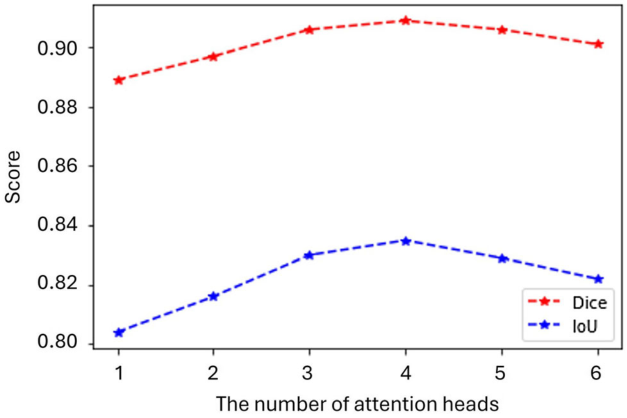

Figure 5. The performance of the multihead attention U-Net models with the different number of attention heads (Dice/IoU scores vs the number of attention heads plot).

Table 1. Quantitative evaluation (average ± standard deviation of each metric) of the different deep learning architectures for MPI-CT image segmentation.

Methods	Accuracy	Precision	Recall	Dice	IoU
U-Net[44]	0.983 ± 0.004	0.891 ± 0.074	0.879 ± 0.076	0.883 ± 0.059	0.794 ± 0.089
Attention U-Net[46]	0.984 ± 0.005	0.892 ± 0.068	0.891 ± 0.069	0.889 ± 0.052	0.804 ± 0.083
Transformer U-Net[52]	0.985 ± 0.005	0.909 ± 0.057	0.878 ± 0.069	0.892 ± 0.053	0.809 ± 0.083
Recurrent U-Net[53]	0.986 ± 0.005	0.918 ± 0.063	0.904 ± 0.062	0.909 ± 0.069	0.835 ± 0.054
Double U-Net[54]	0.985 ± 0.005	0.892 ± 0.053	0.9103 ± 0.057	0.898 ± 0.036	0.818 ± 0.060
2-head attention U-Net	0.985 ± 0.004	0.888 ± 0.063	0.911 ± 0.057	0.897 ± 0.041	0.816 ± 0.052
3-head attention U-Net	0.987 ± 0.005	0.926 ± 0.038	0.890 ± 0.065	0.906 ± 0.039	0.830 ± 0.063
4-head attention U-Net	0.987 ± 0.005	0.920 ± 0.040	0.902 ± 0.058	0.909 ± 0.036	0.835 ± 0.060
5-head attention U-Net	0.986 ± 0.004	0.913 ± 0.049	0.903 ± 0.060	0.906 ± 0.030	0.830 ± 0.050
6-head attention U-Net	0.985 ± 0.005	0.894 ± 0.074	0.912 ± 0.053	0.901 ± 0.043	0.822 ± 0.070

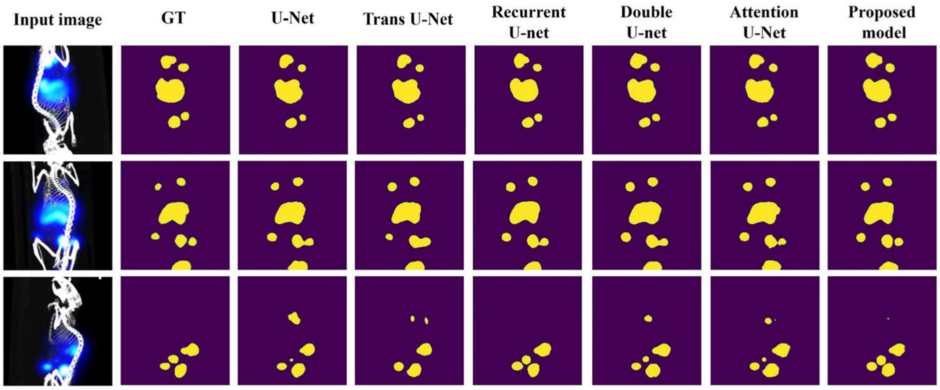

Figure 6. Visualization semantic segmentation results of the proposed model compared to other state-of-the-art U-Net models. From left to right, input MPI-CT images, the ground truth images, the segmentation results generated by U-Net, Trans-U-Net, Recurrent U-Net, Double U-Net, original attention U-Net, and our proposed model (4-head attention, which is the optimal number of attention heads for our MPI-CT dataset), respectively.

well as other state-of-the-art models: Transformer U-Net, Recurrent U-Net, and Double U-Net. The representative visualization MPI-CT image segmentation results, together with the corresponding input images and ground truths, are illustrated in **Figure 6**. All the evaluation results were obtained by using the testing dataset, which is not used for training the models.

4. Discussion

CNNs and vision transformer (ViT) are two different architectures commonly used for computer vision tasks. ViT can perform better than CNNs when global dependencies and contextual information are crucial. Furthermore, it is important to note that ViT-based models rely upon the information from the inputs and the previous hidden stage to generate the current hidden stage allowing the network to capture dependencies in long sequential data (embedded image patches). To effectively learn these long sequential data, ViT needs to be trained on a large dataset to outperform CNN-based models. Nevertheless, CNNs can surpass ViT in tasks involving spatial hierarchies, local pattern extraction, and limited training datasets. In this study, we compare the proposed method with five other state-of-the-art models: original U-Net, Transformer U-Net, original attention U-Net (1 attention head), Recurrent U-Net, and Double U-Net. The Transformer U-Net model leverages both ViT and CNNs to enhance the model's capability in terms of local and global pattern extraction. Although it is an exceptional hybrid neural network for image segmentation problems, it is fairly complicated and requires a somewhat large dataset. In general, the majority of CNN-based models are comparable with Transformer U-Net. Overall, the proposed model outperforms both CNN- and transformer-based models. Indeed, the proposed method extends the number of attention heads, which can be easily integrated with a wide range of existing models to boost the models' performance. This aspect could be potentially investigated in future work. With the multi-head attention mechanism, the model has a better ability to gather diverse representations of features, resulting in a more comprehensive recognition of the features. The optimal number of attention heads for each task is different. In fact, it depends on the context of the image. For instance, segmentation or classification problems for a small object are favorable for applying the small number of attention heads, whereas if the object is large, the larger number of attention heads is eminently suitable. When the number of heads is either excessively large or small, it will impede the model's capability to generalize the data, resulting in a decrease in model performance. In other words, the number of attention heads can be considered as the hyperparameter. Therefore, obtaining the desirable number of attention heads requires trial and error for each task; the MPI-CT image segmentation has to do likewise. The restricted dataset due to implementing a novel and costly MPI system posed challenges in fine tuning all the hyperparameters to achieve an immensely robust model and the overfitting problem. To alleviate these problems, the early stop schedule (stop training if there is no improvement in the validation loss) was applied in training the models. In the future, we anticipate generating a larger dataset by collecting all data from various experiments, as well as applying deep learning techniques to generate data for data augmentation. By utilizing the extensive dataset, the model will exhibit substantially greater robustness and achieve superior performance.

5. Conclusion

This work demonstrates the multihead attention U-Net model, an efficient end-to-end deep learning based on U-Net architecture and multihead attention mechanism, for MPI-CT image segmentation. The proposed model was trained using a custom MPI-CT image dataset collected from transgenic mice with breast tumors injected with a promising MPI tracer for tumor imaging, namely NWs-ICG. To examine the concept of multihead attention, a simple convolution block is employed as the backbone structure of the U-Net architecture to minimize the influence of other factors. Genuinely, the performance of the U-Net architecture can also be improved by using more efficient convolution blocks as the backbone. The optimal number of attention heads was experimentally observed in this study. Although an increase in the number of attention heads can potentially boost the model's capability, the excessive number of attention heads results in a decline in capability. Our study shows that the attention U-Net with 4 heads is the most favorable architecture for MPI-CT image segmentation.

Acknowledgements

This work was supported by the National Science Foundation (NSF) (grant nos. 1808436, 1918074, 2306708, and 2237142-CAREER) and the Department of Energy (DOE) (grant no. 234402) and partially supported by the National Research Council of Thailand, Grant/Award Number: FRB650025/0258, NRCT.MHESRI/505/2563-65, RE-KRIS- FF65-14/FF65-38, and RE-KRIS-FF66-63/ FF66-64.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Zhen Qiu: Conceptualization (lead); Funding acquisition (lead); Resources (lead); Supervision (lead); Writing—original draft (lead); Writing—review & editing (lead). **Aniwat Juhong**: Data curation (lead); Investigation (lead); Methodology (lead); Software (lead); Visualization (lead); Writing—original draft (lead); Writing—review & editing (lead). **Bo Li**: Data curation (supporting). **Yifan Liu**: Data curation (supporting). **Chia-Wei Yang**: Data curation (supporting). **Cheng-You Yao**: Data curation (supporting). **Dalen W. Agnew**: Methodology (supporting); Writing—review & editing (supporting). **Yu Leo Lei**: Methodology (supporting); Writing—review & editing (supporting). **Gary D. Luker**: Methodology (supporting); Writing—review & editing (supporting). **Harvey Bumpers**: Methodology (supporting); Writing—review & editing (supporting). **Xuefei Huang**: Methodology (supporting); Writing—review & editing (supporting). **Wibool Piyawattanametha**: Methodology (supporting); Writing—review & editing (supporting).

Data Availability Statement

Our implementation and pre-trained models are available at https://github.com/AniwatJuhongNACK/Multi-head-attention-U-Net-for-MPI-CT-image-segmentation.git. Additional data and information are available from the corresponding author upon reasonable request.

Keywords

deep learning, image segmentation, magnetic iron oxide nanoworms, magnetic particle imaging, magnetic particle imaging–computed tomography

Received: January 3, 2024
Revised: August 7, 2024
Published online: September 9, 2024

[1] J. W. Bulte, *Adv. Drug Delivery Rev.* **2019**, *138*, 293.
[2] B. Gleich, J. Weizenecker, *Nature* **2005**, *435*, 1214.
[3] L. Scarfe, N. Brillant, J. D. Kumar, N. Ali, A. Alrumayh, M. Amali, S. Barbellion, V. Jones, M. Niemeijer, S. Potdevin, *npj Regener. Med.* **2017**, *2*, 28.
[4] S. Hossaini Nasr, A. Tonson, M. H. El-Dakdouki, D. C. Zhu, D. Agnew, R. Wiseman, C. Qian, X. Huang, *ACS Appl. Mater. Interfaces* **2018**, *10*, 11495.
[5] J. H. Park, G. von Maltzahn, L. Zhang, M. P. Schwartz, E. Ruoslahti, S. N. Bhatia, M. J. Sailor, *Adv. Mater.* **2008**, *20*, 1630.
[6] A. K. Iyer, G. Khaled, J. Fang, H. Maeda, *Drug Discovery Today* **2006**, *11*, 812.
[7] H. Kobayashi, R. Watanabe, P. L. Choyke, *Theranostics* **2014**, *4*, 81.
[8] B. Zheng, T. Vazin, P. W. Goodwill, A. Conway, A. Verma, E. Ulku Saritas, D. Schaffer, S. M. Conolly, *Sci. Rep.* **2015**, *5*, 14055.
[9] J. Rahmer, B. Gleich, J. Weizenecker, J. Borgert, in *Proc. of the Int. Society for Magnetic Resonance in Medicine*, Stockholm, Sweden **2010**.
[10] P. Ludewig, N. Gdaniec, J. Sedlacik, N. D. Forkert, P. Szwargulski, M. Graeser, G. Adam, M. G. Kaul, K. M. Krishnan, R. M. Ferguson, *ACS Nano* **2017**, *11*, 10480.
[11] R. Orendorff, K. Keselman, S. Conolly, in *13th European Molecular Imaging Meeting*, San Sebastián, Spain **2018**.
[12] A. Fu, R. J. Wilson, B. R. Smith, J. Mullenix, C. Earhart, D. Akin, S. Guccione, S. X. Wang, S. S. Gambhir, *ACS Nano* **2012**, *6*, 6862.
[13] A. Tomitaka, H. Arami, S. Gandhi, K. M. Krishnan, *Nanoscale* **2015**, *7*, 16890.
[14] D. Finas, K. Baumann, L. Sydow, K. Heinrich, K. Gräfe, A. Rody, K. Lüdtke-Buzug, T. Buzug, *Biomed. Eng.* **2013**, *58*, 10151520134262.
[15] G. Song, M. Chen, Y. Zhang, L. Cui, H. Qu, X. Zheng, M. Wintermark, Z. Liu, J. Rao, *Nano Lett.* **2018**, *18*, 182.

[16] B. Zheng, M. P. von See, E. Yu, B. Gunel, K. Lu, T. Vazin, D. V. Schaffer, P. W. Goodwill, S. M. Conolly, *Theranostics* **2016**, *6*, 291.

[17] L. C. Wu, Y. Zhang, G. Steinberg, H. Qu, S. Huang, M. Cheng, T. Bliss, F. Du, J. Rao, G. Song, *AJNR* **2019**, *40*, 206.

[18] S. Herz, P. Vogel, P. Dietrich, T. Kampf, M. A. Rückert, R. Kickuth, V. C. Behr, T. A. Bley, *CVIR* **2018**, *41*, 1100.

[19] S. Sun, Y. Chen, M. Zhao, L. Xu, J. Zhong, *IEEE Trans. Instrum. Meas.* **2024**, *73*, 3381661.

[20] Y. Shang, J. Liu, L. Zhang, X. Wu, P. Zhang, L. Yin, H. Hui, J. Tian, *Phys. Med. Biol.* **2022**, *67*, 125012.

[21] Z. Peng, L. Yin, Z. Sun, Q. Liang, X. Ma, Y. An, J. Tian, Y. Du, *Phys. Med. Biol.* **2023**, *69*, 015002.

[22] A. Sun, H. Hayat, S. Liu, E. Tull, J. O. Bishop, B. F. Dwan, M. Gudi, N. Talebloo, J. R. Dizon, W. Li, J. Gaudet, A. Alessio, A. Aguirre, P. Wang, *Front. Cell Dev. Biol.* **2021**, *9*, 704483.

[23] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, M. Chen, in *Proc. of the 13th IEEE Int. Conf. on Control Automation Robotics & Vision (ICARCV)*, Singapore **2014**.

[24] Q. Liu, L. Yu, L. Luo, Q. Dou, P. A. Heng, *IEEE Trans. Med. Imaging* **2020**, *39*, 3429.

[25] S. Deepa, B. A. Devi, *Indian J. Sci. Technol.* **2011**, *4*, 1583.

[26] B. A. Goldstein, A. M. Navar, R. E. Carter, *EHJ* **2017**, *38*, 1805.

[27] D. Maulud, A. M. Abdulazeez, *JASTT* **2020**, *1*, 140.

[28] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, B. Van Calster, *J. Clin. Epidemiol.* **2019**, *110*, 12.

[29] S. Zhang, G. Liang, S. Pan, L. Zheng, *IEEE Access* **2018**, *7*, 12319.

[30] G. Wang, J. C. Ye, K. Mueller, J. A. Fessler, *IEEE Trans. Med. Imaging* **2018**, *37*, 1289.

[31] A. S. Lundervold, A. Lundervold, *Z. Med. Phys.* **2019**, *29*, 102.

[32] X. Fang, P. Yan, *IEEE Trans. Med. Imaging* **2020**, *39*, 3619.

[33] M. H. Hesamian, W. Jia, X. He, P. Kennedy, *J. Digital Imaging* **2019**, *32*, 582.

[34] A. Maier, C. Syben, T. Lasser, C. Riess, *Z. Med. Phys.* **2019**, *29*, 86.

[35] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, X. Ding, *Med. Image Anal.* **2020**, *63*, 101693.

[36] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, A. K. Nandi, *IET Image Process.* **2022**, *16*, 1243.

[37] Z. Niu, G. Zhong, H. Yu, *Neurocomputing* **2021**, *452*, 48.

[38] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, arXiv:2102.04306 **2021**.

[39] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, arXiv:1810.04805 **2018**.

[40] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, R. Socher, *npj Digital Med.* **2021**, *4*, 5.

[41] M.-T. Luong, H. Pham, C. D. Manning, arXiv:1508.04025 **2015**.

[42] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, in *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, December **2015**.

[43] D. Britz, A. Goldie, M.-T. Luong, Q. Le, arXiv:1703.03906 **2017**.

[44] O. Ronneberger, P. Fischer, T. Brox, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th Int. Conf.*, Munich, Germany **2015**.

[45] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, D. Rueckert, *Med. Image Anal.* **2019**, *53*, 197.

[46] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, arXiv:1804.03999 **2018**.

[47] C.-W. Yang, K. Liu, C.-Y. Yao, B. Li, A. Juhong, Z. Qiu, X. Huang, *ACS Appl. Mater.* **2022**, *5*, 18912.

[48] F. Belia, A. Biondi, A. Agnes, P. Santocchi, A. Laurino, L. Lorenzon, R. Pezzuto, F. Tirelli, L. Ferri, D. D'Ugo, *Front. Surg.* **2022**, *9*, 880773.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, in *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, December **2017**.

[50] D. P. Kingma, J. Ba, arXiv:1412.6980, **2014**.

[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy **2017**.

[52] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, L. Soler, in *Machine Learning in Medical Imaging: 12th Int. Workshop*, MLMI, Strasbourg, France **2021**.

[53] W. Wang, K. Yu, J. Hugonot, P. Fua, M. Salzmann, in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, Republic of Korea **2019**.

[54] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, H. D. Johansen, in *2020 IEEE 33rd Int. Symposium on Computer-Based Medical Systems (CBMS)*, Virtual Event **2020**.