

FreqDualNet: frequency-aware vision transformers for tumor segmentation

Jungro Lee^{1,2} and Minhyeok Lee^{2,3}

¹ Department of Computer Science, University College London, University of London, London, United Kingdom

² School of Electrical and Electronics Engineering, Chung-Ang University, Seoul, Republic of South Korea

³ Department of Intelligent Semiconductor Engineering, Chung-Ang University, Seoul, Republic of South Korea

ABSTRACT

Recent advances in deep learning models such as U-Net Transformer (UNETR) and Swin U-Net Transformer (SwinUNETR) have significantly improved tumor segmentation accuracy and local structural sensitivity in medical imaging. Notably, OrgUNETR model demonstrated that predicting both organs and tumors enhances tumor segmentation performance by leveraging their anatomical relationships. However, those models still struggle to capture the complex and irregular shapes of tumors that result from structural constraints or physiological pressure exerted by surrounding organs. These irregular morphologies often manifest as lobulated, or infiltrative margins, containing high frequency edge information that is challenging to represent using a convolutional network or transformer based architectures alone. To address this limitation, we propose FreqDualNet, a frequency aware extension of SwinUNETR and OrgUNETR, which explicitly integrates three types of frequency awareness modules into the network. FreqDualNet introduces three complementary frequency awareness modules: (i) Fourier based spectral feature transformation to enhance high frequency sensitivity, (ii) frequency guided feature fusion between organ and tumor representations, and (iii) high frequency residual refinement to sharpen tumor boundaries. By explicitly modeling frequency domain characteristics associated with tumor shape complexity, the proposed architecture improves sensitivity to fine-grained morphological variations. Experiments on the KiTS19 dataset demonstrate that FreqDualNet achieves a substantial improvement in tumor segmentation performance, increasing the tumor Dice score by 12.75% compared to SwinUNETR (0.5072 vs. 0.4499). On the Prostate158 dataset, FreqDualNet also demonstrated improved tumor segmentation performance, achieving a 6.70% increase in tumor Dice score compared to SwinUNETR (0.2282 vs. 0.2138). By directly utilizing the morphological differences between organs and tumors with frequency aware modules, FreqDualNet improves both segmentation accuracy and generalizability, with strong potential to assist tumor diagnosis, prognosis evaluation, and understanding tumor organ interactions. The code is available at: <https://github.com/JungroLee/FreqDualNet>.

Submitted 5 November 2025

Accepted 9 March 2026

Published 21 April 2026

Corresponding author

Minhyeok Lee, mlee@cau.ac.kr

Academic editor

Ankit Vishnoi

Additional Information and
Declarations can be found on
page 22

DOI 10.7717/peerj-cs.3801

© Copyright

2026 Lee and Lee

Distributed under

Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Computer Vision, Data Mining and Machine Learning, Neural Networks

Keywords Medical imaging, Tumor segmentation, Frequency mapping, Vision transformer

INTRODUCTION

Artificial intelligence (AI) has strongly influenced medical imaging, particularly in the field of tumor segmentation ([Xu et al., 2024](#); [Rayed et al., 2024](#)). Especially, diagnostic processes have been accelerated and AI enabled more precise visualization of tumor boundaries ([Choi et al., 2025](#); [Mildenhall et al., 2021](#); [Wang et al., 2024b](#); [Zou, Lin & Zhu, 2024](#)).

Tumor segmentation, the task of predicting tumor regions from surrounding tissues in medical images, is an important task in diagnosis, treatment planning, and disease monitoring. Over the past decade, as AI develops, various benchmark challenges such as the brain tumor segmentation (BraTS) challenge and the kidney tumor segmentation (KiTS19) competition have been held for more accurate tumor segmentation ([Baid et al., 2021](#); [Li et al., 2024](#); [Abdelrahman & Viriri, 2022](#); [Heller et al., 2019](#)). In BraTS, participants aimed to segment gliomas across multiple MRI modalities accurately, while KiTS19 participants are required to predict kidney tumor from multi-phase CT images.

Early segmentation models were predominantly based on convolutional neural networks (CNNs) ([Yamashita et al., 2018](#)), including architectures like U-Net ([Ronneberger, Fischer & Brox, 2015](#)), GoogleNet ([Szegedy et al., 2015](#)), and ResNet ([He et al., 2016](#)). However, due to their inherently limited receptive fields, convolutional encoders inherently bias representations toward low frequency components due to spatial averaging and downsampling, leading to systematic attenuation of fine-grained boundary cues critical for irregular tumor morphologies ([Hatamizadeh et al., 2022](#)). To overcome these limitations, transformer based models such as UNETR ([Hatamizadeh et al., 2022](#)) and SwinUNETR ([Hatamizadeh et al., 2021](#); [He et al., 2023](#)) were proposed. These architectures integrate the global attention mechanisms of transformers with the spatial localization strengths of U-Net, leading to improved performance in 3D medical image segmentation tasks. Although transformer based models capture long range dependencies, their tokenisation and hierarchical merging implicitly suppress high frequency spatial variations, making them suboptimal for delineating sharply irregular tumor boundaries and there have been new attempts to reflect boundary awareness in segmentation ([Li, Liu & Song, 2026](#)).

Moreover, conventional models typically focus solely on tumor segmentation without explicitly leveraging the anatomical correlation between tumors and surrounding organs. To address this limitation, OrgUNETR model was proposed, with an architecture that predicts both organ and tumor within a unified pipeline ([Choi et al., 2024](#); [Choi, Lee & Lee, 2024](#)). By incorporating anatomical context, OrgUNETR improves tumor segmentation accuracy compared to models that predict tumors only. While OrgUNETR leverages anatomical context by jointly predicting organs and tumors, it remains primarily a semantic alignment framework and does not explicitly model boundary frequency characteristics or geometric irregularity.

Despite these advancements, accurately capturing irregular and high frequency tumor boundaries, such as lobulated or infiltrative morphologies is still challenging ([Agrawal et al., 2009](#); [Hansell et al., 2008](#)). This is because, these complex shapes exhibit high frequency edge patterns that are often smoothed out or poorly represented by conventional feature extraction modules ([Jiang et al., 2021](#)).

To tackle this problem, we propose FreqDualNet, a novel architecture that integrates frequency aware layers into the SwinUNETR and OrgUNETR. Drawing inspiration from recent works in frequency domain networks utilizing frequency spatial representation to capture local texture information (Zhang, Gu & Wang, 2025), we also utilize modules that reflect edge sensitivity and capture shape complexity through Fourier based transformations, high frequency refinement blocks, and hybrid skip connections (Li et al., 2023; Qin et al., 2021; Zou et al., 2023). Those frequency aware modules explicitly preserve high frequency boundary information at multiple hierarchical levels, rather than treating boundary refinement.

We evaluate our model on public medical image datasets, KiTS19 (kidney tumor) and Prostate 158 (prostate tumor), demonstrating that FreqDualNet significantly outperforms SwinUNETR, existing baselines, in tumor segmentation tasks.

The main contributions of this study are as follows:

- (1) We propose a frequency aware segmentation framework that explicitly incorporates frequency domain information to enhance tumor segmentation. To better predict boundary sharpness and structural fidelity, our model deploys a frequency aware module with skip connections, which selectively preserves boundary information during the whole process.
- (2) We demonstrate that the FreqDualNet achieves superior segmentation performance across tumor types characterized by lobulated, and infiltrative morphologies, on various model architectures, such as U-Net, UNETR and SwinUNETR.

RELATED WORKS

Tumor segmentation task

Tumor segmentation is significant in medical diagnosis by predicting the precise anatomical structures. It involves pixel wise classification to distinguish tumor regions from surrounding tissues in medical imaging modalities such as CT, MRI, and PET scans. Among these various imaging modalities, CT and MRI images are commonly used for tumor segmentation tasks due to their ability to provide detailed anatomical information and tissue contrast, which aid in the accurate delineation of tumor boundaries.

In the past, segmentation has been conducted manually by clinicians, which is a time consuming task (McGrath et al., 2020). To address this limitation, AI models have been trained on annotated tumor segmentation datasets. These models have significantly improved the efficiency of the segmentation process by enabling automated identification of tumor regions, thereby supporting clinical decision making and enhancing diagnostic precision.

Numerous deep learning architectures have been proposed for tumor segmentation. For example, in the BraTS challenge, held annually since 2012, early approaches around 2014 primarily employed CNN based and conventional machine learning models (Li et al., 2024). Starting from 2017, U-Net based architectures began to dominate, demonstrating strong performance in medical image segmentation tasks. In recent years, with the advent of large language models (LLMs), hybrid models that integrate vision transformers have

also emerged. Beyond BraTS, other competitions such as the KiTS23 Challenge also focused on kidney tumor segmentation.

Correlation between organ and tumor

Recent advances in medical image segmentation have utilized not only tumor specific features but also the anatomical context provided by surrounding organs ([Fu et al., 2020](#)). It turns out that the morphology of tumors is frequently shaped by the structural constraints imposed by surrounding organs. As tumors interact with adjacent tissues, they often develop in nontrivial geometries, such as lobulated or infiltrative boundaries ([Hansell et al., 2008](#)). These physically constrained growth patterns pose challenges for segmentation models, underscoring the need for organ aware tumor segmentation architectures.

Several segmentation models have leveraged this spatial interdependence to improve performance. For instance, Domain Adaptive Relational Reasoning (DARR) ([Fu et al., 2020](#)) demonstrated that modeling spatial relationships between kidney and tumor enhances segmentation accuracy.

ViT and UNETR

Transformers, initially introduced for translation tasks, have deployed self attention mechanisms that capture long range dependencies without relying on recurrent or convolutional structures. It extended to a broad range of natural language processing (NLP) tasks, including text classification, question answering, and language generation ([Vaswani et al., 2017](#)).

Building on this, the Vision Transformer (ViT) was proposed as an application of transformers in computer vision ([Khan et al., 2023](#)). Unlike CNNs, which utilize local feature extraction, ViT partitions an image into fixed size patches, linearly embeds them, and applies stacked transformer layers to capture global contextual relationships across the entire image.

Inspired by the success of ViT, UNETR was also developed to bring transformer based global context modeling into 3D medical image segmentation ([Hatamizadeh et al., 2022](#)). UNETR integrates a ViT style encoder within a U-Net architecture, effectively combining the strengths of transformer driven feature extraction with the spatial localization capabilities of CNN based architecture. In this design, a 3D medical volume is divided into uniform patches, projected into embedding and being processed through a series of encoder transformer layers composed of multi head self attention (MHSA) and multi layer perceptron (MLPs). The decoder reconstructs the segmentation map through a sequence of 3D deconvolutional layers. Moreover, skip connection between encoder and decoder stages allows the model to leverage both coarse global information and fine grained local features. These features, which capture long range dependencies and preserve local resolution, enable UNETR to excel in various vision tasks such as image classification, object detection and semantic segmentation.

OrgUNETR

The spatial relationship between organs and tumors has been widely recognized, and several prior studies have already attempted to leverage this correlation to improve tumor localization. Building on this insight, OrgUNETR was developed to explicitly incorporate organ context into deep learning based tumor segmentation ([Choi, Lee & Lee, 2024](#)). In this model, the original UNETR architecture is modified to simultaneously predict both organs and tumors as the final output. This joint prediction allows the model to implicitly reflect the spatial and structural relationship between organ and tumor during the training process, as both convolutional and attention layers are updated with shared contextual information.

In addition, to improve computational efficiency, OrgUNETR also replaces the computationally heavy multi head self attention (MHSA) layers with lightweight squeeze and excitation (SE) blocks, which perform channel wise attention in a more efficient manner ([Hu, Shen & Sun, 2018](#)). This modification reduces computational costs while achieving superior performance. Experimental results demonstrate the effectiveness of the proposed approach, OrgUNETR achieved a 40.86% improvement in dice score over the UNETR on the KiTS19 (kidney tumor) dataset and a 32.69% improvement on the Prostate158 dataset. Additionally, the SE blocks led to a 13.9% reduction in computational load.

SwinUNETR

Building on the success of UNETR, SwinUNETR introduces a hierarchical encoder design by adopting Swin Transformers ([Hatamizadeh et al., 2021](#)). Unlike UNETR's fixed resolution attention, Swin Transformers utilize a shifted window mechanism that enables local self attention within non overlapping windows while gradually expanding the receptive field across stages. This hierarchical structure allows SwinUNETR to effectively model both local and global context at multiple scales, improving its ability to handle tumors of varying sizes and appearances ([Capellán-Martín et al., 2023](#)). Additionally, the patch merging layers within SwinUNETR maintain computational efficiency while preserving spatial information.

According to experimental results, SwinUNETR achieved higher dice scores compared to established models such as nnU-Net, SegResNet, and even ViT based models like TransBTS at the BraTS 2021 challenge. To be specific, SwinUNETR showed a 0.5% average improvement across tumor sub regions (ET, WT, and TC) and demonstrated superior generalization on both validation and test sets.

Frequency aware layers

In recent years, frequency aware networks have gained attention in deep learning as a means to capture fine grained details, particularly in tasks involving high frequency patterns ([Qin et al., 2021](#)). Since conventional CNNs tend to focus more on low frequency components due to their local receptive fields and pooling operations, they often lead to

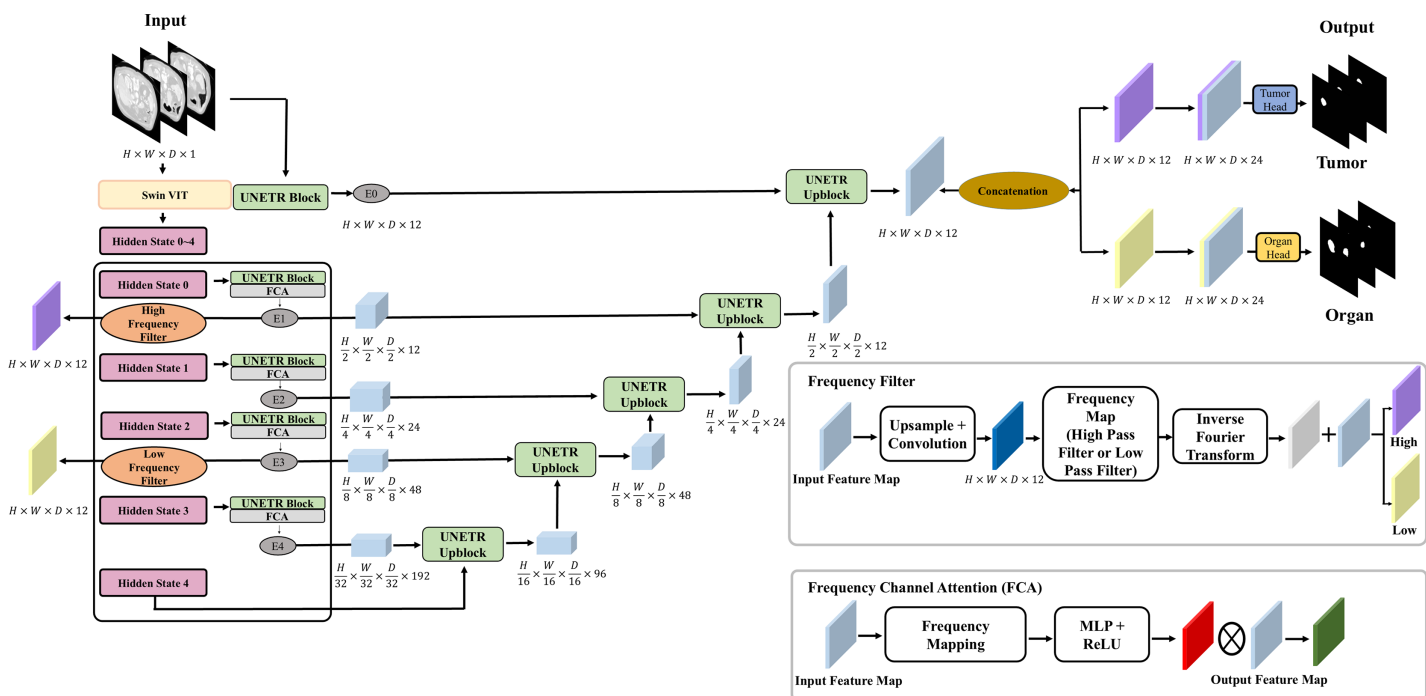


Figure 1 Overall architecture of FreqDualNet. The input 3D CT volume is first processed through a SwinViT encoder, which generates hierarchical hidden states at multiple resolutions. These hidden states are then fed into the UNETR decoder blocks to produce encoded features at each stage. The Feature Channel Attention (FCA) module extracts refined feature maps from the encoded features. Subsequently, the feature maps are progressively upsampled through deconvolution blocks. To enhance segmentation accuracy, the first and third encoded features are processed through High-Frequency Filtering (HFF) and Low-Frequency Filtering (LFF) modules, respectively, and concatenated with the final decoded feature map. The network outputs two separate predictions for organ segmentation and tumor segmentation. [Full-size](#) [DOI: 10.7717/peerj-cs.3801/fig-1](#)

blurred or over smoothed segmentation outputs. To address this, frequency domain processing techniques have been integrated into AI models, by transforming spatial features into the frequency domain. For instance, the frequency channel attention (FCA) block computes channel attention in the frequency domain to better preserve texture and edge information (Zou et al., 2023).

Also, wavelet based U-Net incorporates discrete wavelet transform layers that apply high and low pass filters to extract corresponding frequency components (Li & Shen, 2022). These are then reversed through inverse transformation to reconstruct image features with frequency aware representations.

MATERIALS AND METHODS

Model's architecture

Figure 1 shows the overall architecture of FreqDualNet. The input 3D CT volume is first processed through a SwinViT encoder, which generates hierarchical hidden states at multiple resolutions. These hidden states are then fed into the UNETR decoder blocks to produce encoded features at each stage. The Feature Channel Attention (FCA) module extracts refined feature maps from the encoded features. Subsequently, the feature maps are progressively upsampled through deconvolution blocks. To enhance segmentation

accuracy, the first and third encoded features are processed through High-Frequency Filtering (HFF) and Low-Frequency Filtering (LFF) modules, respectively, and concatenated with the final decoded feature map. The network outputs two separate predictions for organ segmentation and tumor segmentation.

Figure 1 illustrates the comprehensive architecture of the FreqDualNet model, detailing its various layers and connections. A 3D CT scan is processed in the SwinViT block to create hidden states 0~4, and those hidden states are applied to the UNETR block and give encoded features. These encoded features went through FCA block to extract feature maps. These feature maps are then upsampled using deconvolution blocks. Finally, first encoded features and third encoded features after passing High Frequency Filtering and Low Frequency Filtering, are concatenated with final decoded feature map to predict organ and tumor separately.

Then, the network produces four output channels, $H \times W \times D \times 4$, corresponding to two binary segmentation heads: an organ head (background vs. kidney) and a tumor head (background vs. tumor). DiceCE loss was computed independently for each head, and the final loss was defined as a weighted sum of organ and tumor losses, with weights of 0.68 and 0.32, respectively, to balance anatomical localization and tumor delineation performance.

The proposed model, FreqDualNet, is a 3D segmentation AI model designed to simultaneously segment both organ and tumor from volumetric CT scans. It builds upon the SwinUNETR architecture by inserting anatomical context and frequency domain modules to improve tumor localization accuracy, particularly in cases involving small or irregularly shaped lesions.

As described in Fig. 1, the model takes a 3D CT volume as an input ($\mathbb{R}^{H \times W \times D}$) and processes it through a Swin Transformer backbone, which extracts multi scale contextual features using hierarchical self attention blocks. Simultaneously, a parallel convolutional encoder is applied to the input, extracting features with local spatial consistency.

At each encoder stage, feature map's frequency domain information is integrated through a FCA module to extract frequency rich channel. To further leverage frequency information, two fast Fourier transform (FFT) based filtering, high frequency filter and low frequency filter, are applied to two feature maps that are extracted from first and third encoders. Therefore, high frequency filter module extracts tumor related feature map and low frequency filter modules results in an organ related feature map.

The decoder path follows the standard UNETR structure, gradually upsampling the encoded features using a deconvolution-based blocks. To integrate context between the encoders and decoders, organ related and tumor related features are concatenated with the shared decoder output before prediction.

Then, the final features are branched into two separate segmentation heads, one for organ and the other for tumor. The organ head processes feature fused between final decoder output and organ related features, while the tumor head focuses on tumor related features. The model finally outputs four channels, two for organ segmentation (background and organ) and two for tumor segmentation (background and tumor).

This architecture enables the network to explicitly model the interaction between organ and tumor, while also integrating frequency domain attributes to improve performance on regions with irregular features.

Preprocessing method and patch embedding

To improve generalization, we applied a series of data augmentations. These included random 3D flipping along each spatial axis, Gaussian noise injection, random intensity scaling, and random intensity shifting (Shorten & Khoshgoftaar, 2019). These methods represent real world anatomical variability to mitigate overfitting problems.

We then fed these images into FreqDualNet architecture, which begins with a patch embedding module. This module partitions the input 3D volume into non overlapping cubic patches, each of which is flattened and linearly projected through a convolutional embedding block to form a sequence of patch tokens. These tokens are then fed into a hierarchical Swin Transformer block, which applies Layer Normalization and dropout for regularization.

Frequency comparison

To quantitatively compare the spatial frequency characteristics of the organ and tumor regions, we computed a high to low frequency energy ratio using a radial frequency based analysis (Van Heel & Schatz, 2005).

For each 3D segmentation mask $x(r)$,

$$x(r) \in \mathbb{R}^{H \times W \times D}, \quad r = (x, y, z). \quad (1)$$

we applied a 3D FFT to obtain the frequency domain magnitude representation, $\hat{x}(f)$.

$$\hat{x}(f), f = (f_x, f_y, f_z), \quad |\hat{x}(f)|. \quad (2)$$

Then, the radial frequency was defined as $r(f)$.

$$r(f) = \sqrt{f_x^2 + f_y^2 + f_z^2}. \quad (3)$$

We then defined the high and low frequency regions by setting the radial frequency's threshold at half its maximum,

$$r_{\text{thresh}} = 0.5 * \max r(f). \quad (4)$$

The average spectral energy in the high and low frequency domains was computed as:

$$E_{\text{high}} = \frac{1}{|\Omega_{\text{high}}|} * \sum_{f \in \Omega_{\text{high}}} |\hat{x}(f)|, \quad \Omega_{\text{high}} = \{f | r(f) \geq r_{\text{thresh}}\} \quad (5)$$

$$E_{\text{low}} = \frac{1}{|\Omega_{\text{low}}|} * \sum_{f \in \Omega_{\text{low}}} |\hat{x}(f)|, \quad \Omega_{\text{low}} = \{f | r(f) < r_{\text{thresh}}\}. \quad (6)$$

Finally, the high to low frequency energy ratio was defined as,

$$R = \frac{E_{\text{high}}}{E_{\text{low}} + \varepsilon}, \quad \varepsilon = 10^{-6}. \quad (7)$$

We utilize this ratio to serve as a measure of high frequency richness within each region. Applying this method to 210 paired samples of organ and tumor segmentations, we found that in 181 cases, the tumor region exhibited a higher frequency energy ratio than in the organ region, whereas the opposite was observed in 29 cases. Statistical significance testing confirmed that the observed difference is non random. A paired Wilcoxon signed-rank test showed a highly significant increase in frequency energy ratio in tumor regions compared to organ regions ($p = 1.44 \times 10^{-25}$), with large effect sizes (Cliff's $\delta = 0.724$, Cohen's $d_p = 0.948$), indicating a robust and substantial difference (Thiese, Ronna & Ott, 2016).

Frequency aware modules

To enhance the segmentation ability of our network, we inserted a frequency aware module within the SwinUNETR architecture.

Firstly, we integrate a FCA module after each encoder stage. Given a 3D feature map, $x \in \mathbb{R}^{B \times C \times H \times W \times D}$, we compute its magnitude spectrum per channel through a 3D FFT, $F(x) = |FFT_{3D}(x)|$. A global average pooling operation is then applied over the spatial dimensions to produce a frequency summary for each channel.

$$s_c = \frac{1}{HWD} \sum_{i,j,k} F(x_{c,i,j,k}). \quad (8)$$

These summaries are passed through a lightweight MLP with sigmoid activation to produce channel wise attention weights, α .

$$\alpha = \sigma(W_2 * ReLU(W_1 * s_c)). W_1, W_2 : \text{Weights}. \quad (9)$$

The original feature map is then reweighed as: $x' = x \odot \alpha$. This module enables the network to selectively emphasize frequency rich channels, allowing the model to catch the boundary features.

Secondly, we apply frequency filtering module to further enhance spectral discrimination (Cai et al., 2021; Wang et al., 2024a). After first and third encoders, feature map $x(r)$'s ($x(r) \in \mathbb{R}^{H \times W \times D}$, $r = (x, y, z)$), low and high frequency components are figured out using radial masks in the frequency domain ($f = (f_x, f_y, f_z)$). For the first step, we defined the normalized radial frequency, $r(f)$.

$$r(f) = \sqrt{f_x^2 + f_y^2 + f_z^2}, f_x, f_y, f_z \in [-1, 1]. \quad (10)$$

Then, a binary high pass mask is applied in the frequency domain to isolate boundary enhancing signals (r_{high} and r_{low} are tuned through hyperparameter tuning).

$$1_{high}(f) = \begin{cases} 1, & \text{if } r(f) \geq r_{high} \\ 0, & \text{otherwise} \end{cases}, r_{high} = 0.4. \quad (11)$$

Filtered features are then reconstructed via inverse FFT:

$$x^{high} = Re[IFFT_{3D}(FFT_{3D}(x) * 1_{high})]. \quad (12)$$

Likewise, a low pass filter is used to retain smooth structural context.

$$1_{low}(f) = \begin{cases} 1, & \text{if } r(f) < r_{low} \\ 0, & \text{otherwise} \end{cases}, r_{low} = 0.25 \quad (13)$$

$$x^{low} = \text{Re}[IFFT_{3D}(FFT_{3D}(x) * 1_{low})]. \quad (14)$$

Finally, the filtered components, x^{high} , x^{low} are linearly combined with the original features:

$$\tilde{x}_{tumor} = x(r) + \alpha \times x^{high}, \quad \tilde{x}_{organ} = x(r) + \beta \times x^{low}, \\ \alpha, \beta = 1.5 \text{ (arbitrary value)}. \quad (15)$$

These filtered representations, $\tilde{x}_{tumor}$ and $\tilde{x}_{organ}$ are later concatenated with final decoder outputs and passed through organ and tumor specific segmentation heads.

Metrics

To train and evaluate the performance of the FreqDualNet model, we adopted a composite loss function combining the dice coefficient and cross entropy (CE) loss ([Sudre et al., 2017](#)). The Dice coefficient calculates the spatial overlap between the predicted segmentation, p_i and the ground truth, g_i and is defined as:

$$Dice(p, g) = \frac{2 \times \sum_i p_i g_i}{\sum_i p_i + \sum_i g_i + \varepsilon}, \quad \varepsilon = 10^{-5}, \quad DiceLoss = 1 - Dice(p, g). \quad (16)$$

Dice scores range from 0 to 1, with higher values indicating better segmentation performance. We also incorporated the CE loss, which measures the divergence between predicted class probabilities and the true labels. It is computed as,

$$CE(T, P) = - \sum_{c=1}^C T_c \times \log(P_c) \quad (17)$$

where T is the one hot encoded true label, c indexes the class, and P_c is the predicted probability for class c obtained *via* the softmax function.

This equation quantifies the negative log likelihood of the correct class being predicted, with lower values indicating better segmentation performance.

To leverage the strengths of both metrics, we utilized a DiceCE Loss, L_{DiceCE} , which combines the dice and CE losses as a weighted sum:

$$L_{DiceCE} = \beta \times DiceLoss + (1 - \beta) \times CE \text{ Loss}. \quad (18)$$

Here, β is tunable parameter that controls the relative influence of each term during training.

However, since DiceCE loss combines overlap based and probabilistic objectives, improvements in Dice score may not always lead to lower DiceCE loss, particularly for small or ambiguous tumor regions.

Given the dual target nature of our segmentation task, we further extended this formulation by applying specific weighting to the loss components. The total loss used for model optimization was defined as,

$$L_{total} = 0.65 \times L_{organ} + 0.35 \times L_{tumor} \quad (19)$$

where L_{organ} , L_{tumor} represent the DiceCE Loss applied to organ and tumor predictions, respectively. The weights (65% for organ and 35% for tumor) reflect the model's design principle of leveraging organ segmentation as contextual guidance for tumor localization.

KITS19

To evaluate the FreqDualNet on kidney tumor segmentation, we utilized the publicly available KiTS19 dataset, a benchmark dataset released as part of the kidney tumor Segmentation Challenge (<https://kits19.grand-challenge.org/>). The dataset was organized in a conjunction with MICCAI 2019 and is jointly managed by the University of Minnesota and the University of Melbourne, with data collected from patients who underwent partial or radical nephrectomy at the University of Minnesota Medical Center between 2010 and 2018. The KiTS19 dataset is publicly accessible through the official challenge website, distributed under the CC BY-NC-SA license, and does not have a single assigned DOI. A detailed description of the dataset is provided in the official challenge article by *Heller et al. (2019)* (MICCAI 2019) (DOI: [10.1016/j.media.2020.101821](https://doi.org/10.1016/j.media.2020.101821)).

A total of 300 cases were retrospectively collected from which 210 cases with complete imaging and expert validated annotations were released as training data. Cases were included only if a pre-operative arterial phase abdominal CT scan was available, and cases lacking such imaging were excluded.

All CT volumes were provided in anonymized NIFTI format (*Heller et al., 2019*) with an original in-plane resolution of 512×512 pixels, while the number of axial slices varied substantially across patients, ranging from 29 to 1,059 slices. Slice thicknesses ranged from 1 to 5 mm, reflecting real-world clinical variability. All scans were acquired with patients in the supine position.

The KiTS19 dataset provides separate voxel wise annotations for kidney and tumor, with tumor labels strictly constrained within the kidney region to ensure anatomical consistency. These annotations were generated through a semi manual process: trained medical students delineated kidney and tumor contours on selected axial slices under clinical supervision, and missing slices were completed *via* linear interpolation. All cases were subsequently reviewed and corrected by experienced researchers with access to radiology reports and surgical pathology findings to ensure accurate tumor localization.

For our preprocessing, we first exclude test datasets which do not contain ground truth of organ and tumor images. Then, we normalize the value of images for generalizing the model's ability, not to react sharply to outliers. Lastly, we reshape the dataset into fixed spatial size to ensure uniform input dimensions.

In order to normalize, voxel intensities were clipped to a range defined by two standard deviations around the mean, $[x_{min}, x_{max}] = [\mu - 2\sigma, \mu + 2\sigma]$, where μ is the mean intensity of CT volume and σ is the standard deviation. Then, the clipped intensities were normalized to the $[0, 1]$ range using the min max normalization.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}, \quad x_{norm} \in [0, 1]. \quad (20)$$

After normalization, to reduce computational cost and ensure uniform input dimensions for the model, we first reshaped all CT volumes to a fixed spatial size of $224 \times 224 \times 224$ using trilinear interpolation.

Prostate158

In addition to the KiTS19 dataset, we also utilized the Prostate158 dataset to evaluate the performance and generalizability of our model on prostate tumor segmentation tasks (Adams *et al.*, 2022).

The Prostate158 dataset was created and curated by Keno Bressem, Lisa Adams, and Günther Engel, with contributions from multiple researchers across Charité—Universitätsmedizin Berlin, Technische Universität München, Universitätsmedizin Göttingen, and Radboud University. All imaging data and annotations were generated under institutional oversight, with annotations provided by experienced radiologists. The dataset is publicly accessible *via* Zenodo (<https://zenodo.org/records/6481141>), where detailed metadata and licensing information are provided.

It includes voxel wise annotations for both prostate anatomy and tumor regions, enabling simultaneous learning of organ structure and tumor localization. In total, Prostate158 contains 139 training samples and 19 validation samples. Each MRI scan has a native spatial resolution of $270 \times 270 \times 24$ voxels. Due to a different data shape, all volumes were resized to a uniform spatial resolution of $224 \times 224 \times 224$ using linear interpolation. During this process, samples in which tumor voxels were completely lost were excluded, resulting in the removal of 72 cases to ensure sufficient tumor representation in the training data.

For preprocessing, as in the KiTS19 dataset preprocessing, min max normalization was applied to each MRI scan to scale intensity values into a consistent range. This step reduces variability caused by different acquisition settings and improves training stability.

Computing infrastructure

All experiments were conducted on a Linux workstation running Ubuntu 24.04.2 LTS equipped with two NVIDIA GeForce RTX 3090 GPUs (24 GB VRAM each). The model was trained primarily on a single GPU, utilizing approximately 21,360 MiB of VRAM during execution. The environment was configured with Python 3.13.5, PyTorch 2.6.0 + CUDA 12.4.

RESULTS

Experimental details

Training was performed with a batch size of 1 due to GPU memory constraints, using gradient accumulation over four iterations to simulate a larger effective batch size. Gradients were clipped with a maximum norm of 1.0 to stabilize optimization. Model checkpoints were selected based on the highest validation dice score for tumor segmentation after epoch 100.

To evaluate FreqDualNet, we used two publicly available datasets: KiTS19 (210 CT volumes with expert-annotated kidney and tumor masks) and Prostate158 (72 annotated volumetric scans). The dataset was randomly split into training (75%), validation (20%), and test (5%) subsets at the patient level using a fixed random seed (seed = 55) to ensure reproducibility. The model was trained using the AdamW optimizer, with an initial learning rate of 1×10^{-3} and a weight decay of 5×10^{-5} (Loshchilov & Hutter, 2017). For the loss function, we employed DiceCE loss, a weighted combination of Dice loss and CE loss, which jointly optimizes overlap based and pixel wise classification objectives.

We did not adopt an early stopping strategy and instead trained all models for 600 epochs to ensure a fair and consistent comparison between FreqDualNet and the baseline models. Using a fixed training length avoids discrepancies that may arise from model dependent stopping criteria. The training was terminated at 600 epochs, corresponding to the stage at which the tumor prediction accuracy on the training set reached approximately 0.6–0.7 across models, allowing direct performance comparison under comparable optimization conditions.

In addition, the Swin Transformer backbone was configured with four hierarchical stages, with transformer depths of (3, 6, 12, 24) and attention head counts of (1, 2, 1, 2) across the stages. A base channel width of 12 was used throughout the network.

Moreover, to reduce GPU memory consumption and improve training efficiency, we adopted automatic mixed precision (AMP), which casts computations to float16 (Micikevicius et al., 2017).

To maintain generalizability, scaler function scales gradients during back propagation and adjusts the scale factor every four steps. This setup enabled us to use smaller batch sizes and optimize GPU utilization without compromising model accuracy (Ott et al., 2018).

Computational efficiency

We evaluated the computational efficiency of FreqDualNet in comparison with the baseline SwinUNETR under identical experimental settings (batch size = 1, input size = 224^3 , AMP enabled). The two models have nearly identical parameter counts (6.822M for FreqDualNet vs. 6.819M for the baseline, +0.04%), indicating that the additional computational cost is not attributable to increased model size. In inference, FreqDualNet exhibited a latency of 481.32 ± 0.72 ms per iteration, compared to 332.61 ± 0.79 ms for the baseline, corresponding to a 44.7% increase in runtime. The peak GPU memory allocation during inference increased from 6.33 GB to 8.83 GB (+39.4%). A similar trend was observed during training, where the end to end training step (forward, loss computation, backward pass, and parameter update) showed a 39.2% increase in latency and a 76.2% increase in peak memory usage. These overheads arise from the additional spectral analysis modules and frequency-aware processing in FreqDualNet, reflecting an explicit trade off between computational efficiency and enhanced frequency based feature modeling.

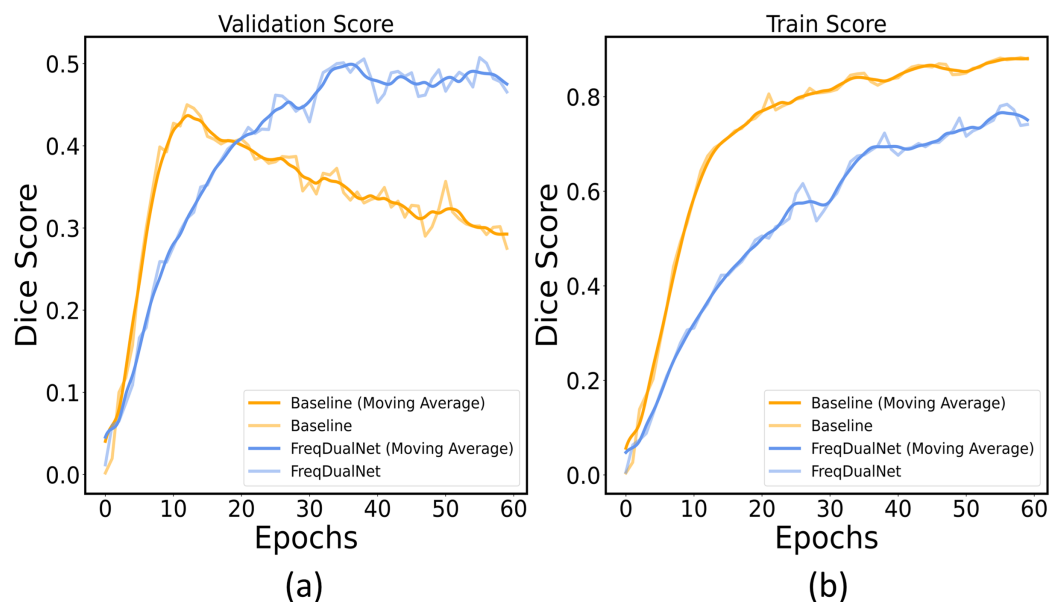


Figure 2 Comparison of tumor segmentation performance between FreqDualNet and the baseline model on the KiTS19 dataset. Dice score trajectories are shown for (A) validation and (B) training datasets, demonstrating more stable and consistently higher performance of FreqDualNet across training epochs.

Full-size DOI: [10.7717/peerj-cs.3801/fig-2](https://doi.org/10.7717/peerj-cs.3801/fig-2)

Quantitative results

For our baseline model, we employed the SwinUNETR model from the MONAI library, modified to output four channels for simultaneous segmentation of organs and tumors without separate prediction heads. Also, U-Net and UNETR are also modified in a similar way for a comparison.

In terms of hyperparameter tuning, we experimented with different learning rates and found that both the baseline and FreqDualNet achieved their best tumor prediction performance at a learning rate of 1×10^{-3} .

As shown in Fig. 2, the baseline model initially achieved about 0.1 higher score on the training set. However, it exhibited reducing performance on a certain number of epochs at the test set, which indicates the overfitting problem. In contrast, FreqDualNet maintained a stable dice score of around 0.5 on the test dataset, indicating stronger generalization capability.

To be specific, as shown in Table 1, the best tumor prediction dice score on the test set was 50.72% for the FreqDualNet, compared to 44.99% for the baseline, indicating a 12.75% improvement.

A similar trend was observed in the DiceCE Loss curves, as shown in Fig. 3. While the baseline model showed lower training loss initially, its test loss began to increase after a certain number of epochs, further indicating overfitting and limited generalizability. In contrast, FreqDualNet maintained a stable DiceCE Loss of around 0.2 on the test dataset.

Specifically, as in Table 1, FreqDualNet achieved a lower DiceCE Loss on the test set (0.19) than the baseline (0.21), representing an 8.49% reduction.

Table 1 Model Comparison with KITS19. Evaluation on the FreqDualNet and other segmentation models. Bold indicates the best value in each metric column.

Model	DiceCE loss	Dice score
FreqDualNet	0.1995	0.5072
Baseline	0.2180	0.4499
U-Net	0.2552	0.3734
UNETR	0.3190	0.2902
OrgUNETR	0.2840	0.3282

Note:

Kits19 dataset evaluation on the FreqDualNet and other segmentation models.

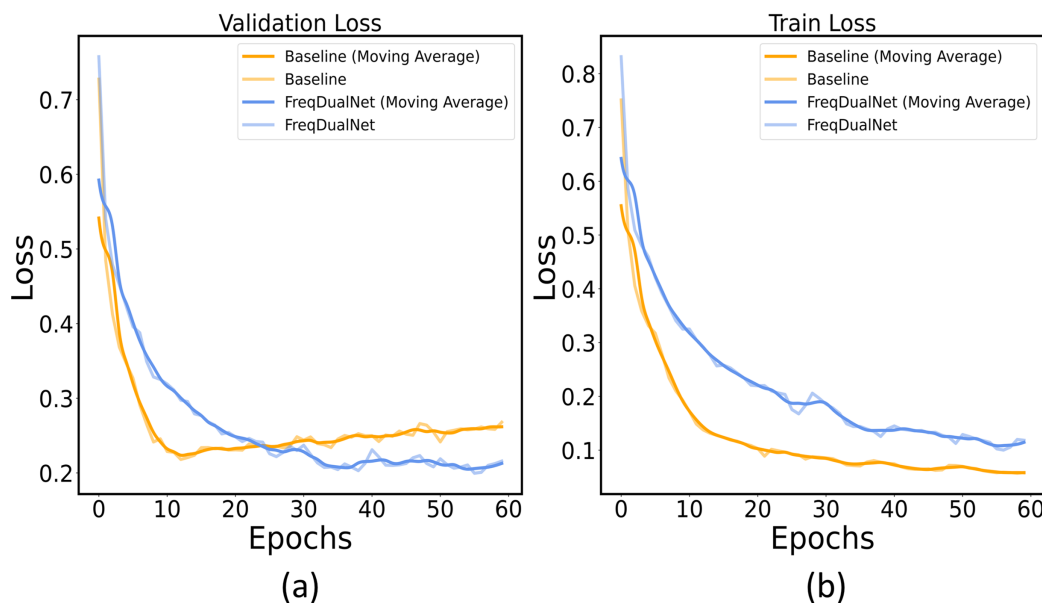


Figure 3 Comparison of tumor segmentation performance between FreqDualNet and the baseline model on the KiTS19 dataset. Dice Loss trajectories are shown for (A) validation and (B) training datasets, demonstrating more stable and consistently higher performance of FreqDualNet across training epochs.

Full-size DOI: [10.7717/peerj-cs.3801/fig-3](https://doi.org/10.7717/peerj-cs.3801/fig-3)

Additionally, as shown in Table 1, FreqDualNet achieves a 35.83% higher Dice score than U-Net (0.3734), a 74.77% improvement over UNETR (0.2902), and a 54.53% improvement over OrgUNETR (0.3282). Similarly, in terms of DiceCE Loss, FreqDualNet achieves reductions of 27.91% compared to U-Net (0.2552), 59.89% compared to UNETR (0.3190), and 42.35% compared to OrgUNETR (0.2840). These results indicate that FreqDualNet not only demonstrates stronger generalization but also delivers superior quantitative performance across both evaluation metrics compared to existing models.

We further evaluated the generalizability of FreqDualNet on the Prostate158 dataset.

As shown in Fig. 4, FreqDualNet achieved a maximum tumor Dice score of 22.82%, compared to 21.38% for the baseline model. More importantly, when averaged across training epochs, FreqDualNet demonstrated a 24.84% improvement in mean tumor Dice

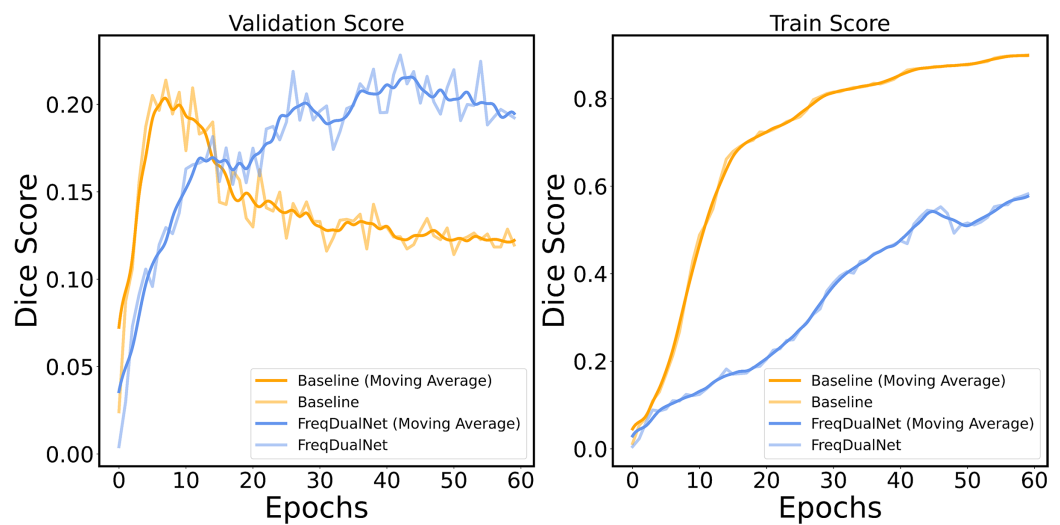


Figure 4 Comparison of tumor segmentation performance between FreqDualNet and the baseline model on the Prostate 158 dataset. Dice score trajectories are shown for (A) validation and (B) training datasets, demonstrating more stable and consistently higher performance of FreqDualNet across training epochs.

[Full-size](#) DOI: [10.7717/peerj-cs.3801/fig-4](https://doi.org/10.7717/peerj-cs.3801/fig-4)

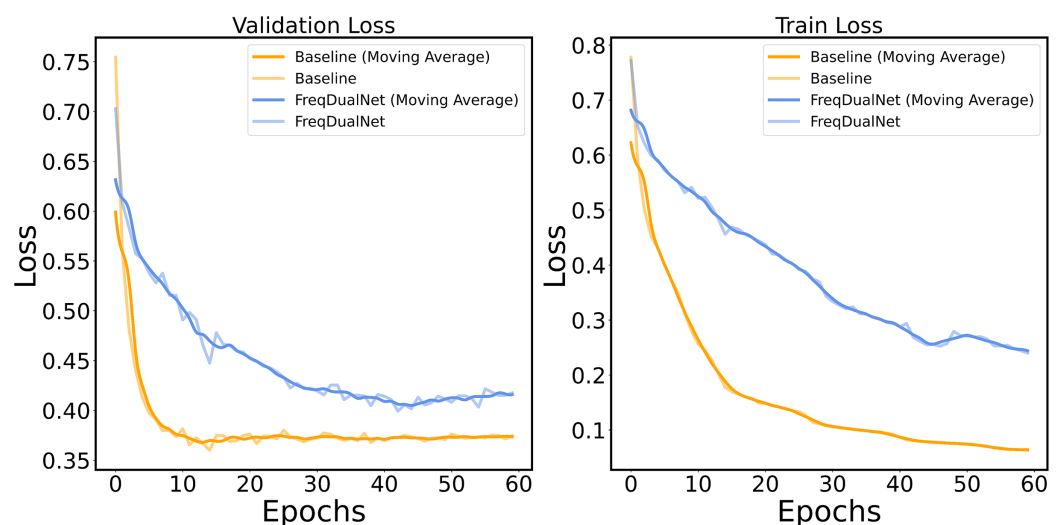


Figure 5 Comparison of tumor segmentation performance between FreqDualNet and the baseline model on the Prostate158 dataset. Dice Loss trajectories are shown for (A) validation and (B) training datasets, demonstrating more stable and consistently higher performance of FreqDualNet across training epochs.

[Full-size](#) DOI: [10.7717/peerj-cs.3801/fig-5](https://doi.org/10.7717/peerj-cs.3801/fig-5)

score over the baseline, indicating more stable and consistent tumor segmentation performance.

In contrast, as in Fig. 5, the DiceCE Loss on the Prostate158 dataset showed an opposite trend: FreqDualNet exhibited a higher loss value (0.3992) than the baseline model (0.3600), corresponding to a 10.88% increase. This suggests that while FreqDualNet improves overlap-based segmentation accuracy, its optimization behavior under DiceCE

Table 2 Model comparison with Prostate158. Evaluation on the FreqDualNet and other segmentation models. Bold indicates the best value in each metric column.

Model	DiceCE loss	Dice score
<i>FreqDualNet</i>	0.3992	0.2282
<i>Baseline</i>	0.3600	0.2138
<i>U-Net</i>	0.3917	0.2244
<i>UNETR</i>	0.7257	0.1461
<i>OrgUNETR</i>	0.5111	0.1893

Note:

Prostate 158 dataset's evaluation on the FreqDualNet and other segmentation models.

Loss may differ across datasets. This can be explained by a trade-off which is explained at Metrics part.

Additionally, as shown in Table 2, FreqDualNet achieves a 1.69% higher Dice score than U-Net (0.2244), a 56.19% improvement over UNETR (0.1461), and a 20.54% improvement over OrgUNETR. In terms of DiceCE Loss, FreqDualNet shows a marginal difference compared to U-Net (1.94% higher loss than 0.3917), but substantially outperforms UNETR and OrgUNETR, achieving reductions of 50.39% (0.7257) and 21.89% (0.5111), respectively.

Qualitative results

In addition to quantitative evaluation, we conducted a qualitative analysis to assess how well the predicted tumors resemble the ground truth in terms of shape, location, and boundary detail. As illustrated in Fig. 6, the first row shows the ground truth of kidney tumor, the second row displays the tumor predictions from our FreqDualNet, and the third row shows predictions from the baseline model.

Across all seven cases, the tumors predicted by the FreqDualNet more closely resemble the ground truth in size, location, and overall morphology compared to the baseline. For instance, in the second column, the ground truth tumor has a circular shape, which is accurately captured by the FreqDualNet. Conversely, for the baseline model, the upper right region is recognizably distorted resulting in an irregular and inaccurate boundary. Furthermore, in the third and fourth columns, FreqDualNet successfully identifies tumor regions at the correct locations, whereas the baseline model fails to detect any tumors. Additionally, in the seventh column, the actual ground truth tumor is very small and not separated. However, the baseline model overestimates its size and incorrectly predicts two separate tumors, while the FreqDualNet provides an accurate prediction.

In Fig. 7, we further present qualitative results on the Prostate158 test dataset to visually assess the extent to which FreqDualNet improves tumor segmentation performance. In columns 1, 2, 4, and 5, FreqDualNet more accurately delineates tumor boundaries and scales compared to the baseline model. However, in certain cases, such as the third column, both FreqDualNet and the baseline model fail to detect very small tumor regions, highlighting the inherent difficulty of segmenting small-scale tumors in this dataset. These visual comparisons highlight the FreqDualNet's superior ability to localize and delineate tumors with greater precision.

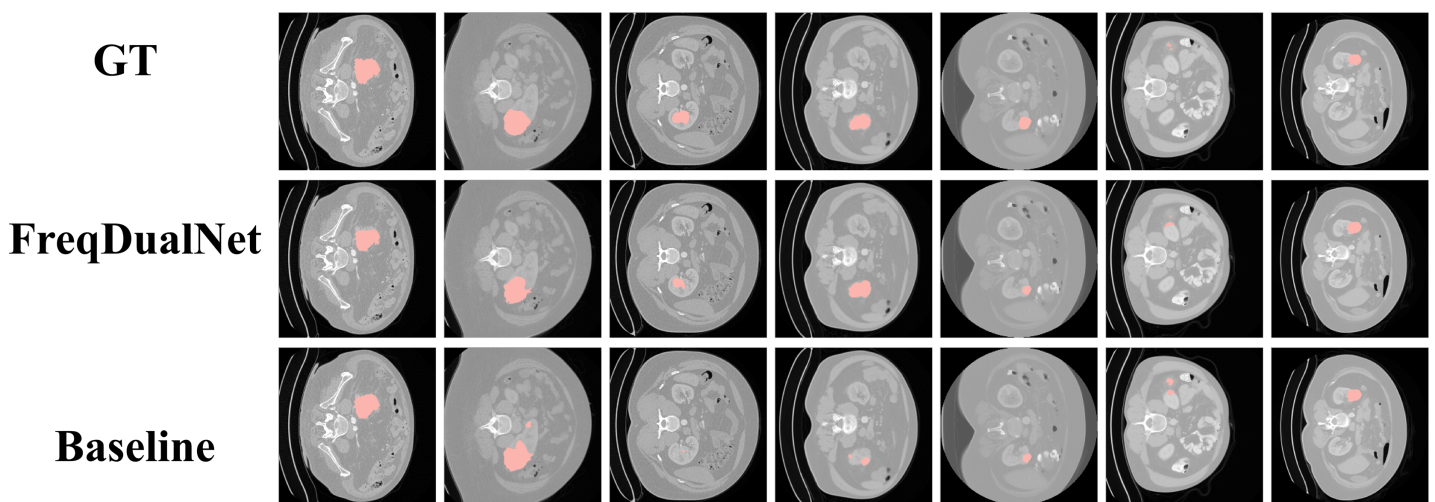


Figure 6 Qualitative comparison of tumor segmentation results on the KiTS19 dataset. From top to bottom: ground truth, FreqDualNet prediction, and baseline model prediction. The pink pixels throughout the image represents the tumor pixels, whereas the greyscale pixels indicate the background. [Full-size](#) DOI: 10.7717/peerj-cs.3801/fig-6

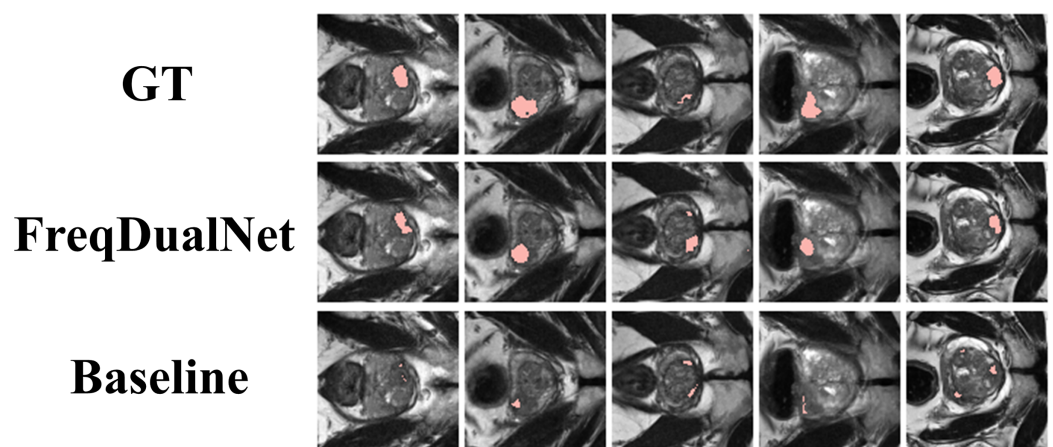


Figure 7 Qualitative comparison of tumor segmentation results on the Prostate 158 dataset. From top to bottom: ground truth, FreqDualNet prediction, and baseline model prediction. The pink pixels throughout the image represents the tumor pixels, whereas the greyscale pixels indicate the background. [Full-size](#) DOI: 10.7717/peerj-cs.3801/fig-7

Additional experiments

Figure 8. Validation score of the additional experiments. The red, orange, and blue curves denote the Dice scores of the full FreqDualNet, the variant without the FCA layer, and the variant without frequency filtering, respectively.

To ensure the best performance of the FreqDualNet, we compared three models: (1) FreqDualNet, (2) FreqDualNet without the FCA module, and (3) FreqDualNet without the frequency filtering module. As shown in [Fig. 8](#), FreqDualNet (marked in red) achieved the highest Dice score of 50.72%. FreqDualNet without the FCA (orange) reached a peak score of 45.53%, while the model without the frequency filtering module (blue) attained

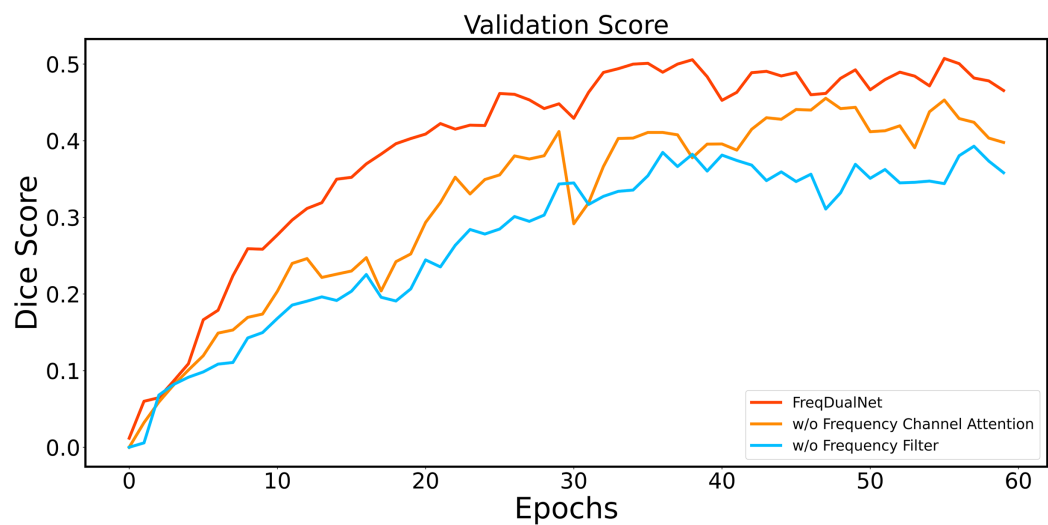


Figure 8 Validation score of the additional experiments. The red, orange, and blue curves denote the Dice scores of the full FreqDualNet, the variant without the FCA layer, and the variant without frequency filtering, respectively. [Full-size !\[\]\(2154c9f569edab9d15ee6b3ae0aa8398_img.jpg\) DOI: 10.7717/peerj-cs.3801/fig-8](https://doi.org/10.7717/peerj-cs.3801/fig-8)

only 39.28%. These results indicate that the FreqDualNet outperformed the FreqDualNet without the FCA module by 11.41%, and FreqDualNet without the frequency filtering module by 29.14%. It demonstrates the effectiveness of frequency based modules in enhancing segmentation performance.

Secondly, to investigate the contribution of frequency domain operations, we performed a hyperparameter tuning on two key components, the reduction ratio in the FCA modules, and the radius ratio in the frequency domain filtering blocks (Wang et al., 2024a; Zou et al., 2023).

In the FCA, the reduction ratio controls the dimensionality compression in the channel wise MLP, thereby regulating how selectively the model can assign importance weights to individual frequency channels. We varied this ratio across (4, 8), to explore the model's efficiency and effectiveness.

In the frequency filters, the radius ratio defines the cutoff boundary in the normalized spatial frequency domain for the high pass operations. A larger radius ratio in the high pass filter preserves more high frequency components (edges, fine textures), whereas a smaller ratio suppresses them.

Conversely, in the low pass filter, a smaller radius ratio retains a lower frequency structure while attenuating high frequency noise.

We tested high pass radius ratio values of (0.25, 0.4, 0.5) for tumor branch, while applying same low pass filtering in the organ branch. According to Table 3, the best performance was achieved with a reduction ratio of 4 and a high pass radius ratio of 0.4, yielding a Dice score of 0.5072. Notably, except for the combination of a 0.4 radius ratio with a reduction ratio of 8, all configurations outperformed the baseline model (Dice score: 0.4499), demonstrating the robustness of our frequency aware design.

Table 3 Dice score by tuning frequency modules. Dice scores of the FreqDualNet based on different frequency modules' hyperparameter. Bold indicates the best-performing hyperparameter setting and corresponding Dice score.

High-pass radius ratio	Reduction rate	Dice score
Baseline		0.4499
0.25	4	0.4597
0.25	8	0.4612
0.4	4	0.5072
0.4	8	0.4079
0.5	4	0.4785
0.5	8	0.4720

Note:

Dice Scores of the FreqDualNet based on different frequency modules' hyperparameter.

LIMITATIONS DISCUSSION

Although our model achieves superior performance compared to previous approaches and demonstrates the potential utility of modeling the relationship between organs and tumors, several limitations remain.

The first limitation lies in its relatively low Dice score for tumor segmentation, which poses challenges for clinical applicability. The FreqDualNet model achieves a Dice score of 50.72% for tumor segmentation, indicating room for improvement. Although the Dice score is comparatively lower than that reported by other models on the benchmark dataset, this difference can be largely attributed to discrepancies in the evaluation protocol. Most prior studies reported results on the official KiTS19 external test set and were trained with a larger number of samples (Heller et al., 2021). In contrast, our evaluation was conducted under more constrained training and testing conditions, which partly explains the observed performance gap. To address this issue, the model should be trained and validated on larger and more diverse datasets.

The second limitation concerns the scarcity of datasets that contain both organ and tumor annotations. Most publicly available datasets include either tumor regions or organ structures, but not both, which restricts our ability to comprehensively evaluate the model's generalizability and to further substantiate the effectiveness of frequency domain modules in capturing inherent organ and tumor relationships.

Third, for the Prostate158 dataset, FreqDualNet demonstrated a consistent improvement in tumor Dice score over the baseline model across most test cases. Although the paired Wilcoxon signed-rank test did not reach conventional statistical significance ($p = 0.0938$), the observed effect sizes were moderate to large (Cohen's $d_z = 0.759$; paired Cliff's delta = 0.600), indicating a practically meaningful performance gain. Most cases (four out of five) showed higher Dice scores for FreqDualNet, suggesting a consistent improvement trend despite the limited sample size.

However, the limitation lies in the statistical validation of performance comparisons on the KiTS19 test dataset. Paired statistical tests conducted on the held-out test split yielded no statistically significant difference in tumor Dice score between FreqDualNet and the

baseline model (Wilcoxon signed-rank test, $p > 0.05$), accompanied by small effect sizes. This is largely attributable to the limited number of test samples and the high variance of Dice scores, particularly in small or low-contrast tumor cases. As a result, the statistical power of test-set comparisons on KiTS19 is inherently constrained.

These limitations highlight the need for broader data acquisition and integration efforts to facilitate the real-world deployment of AI based tumor segmentation systems.

CONCLUSIONS

In this study, we proposed FreqDualNet, a novel model that integrates frequency aware modules into the SwinUNETR architecture for simultaneous organ and tumor segmentation.

This model not only effectively captured high frequency morphological features, such as irregular tumor boundaries, but also demonstrated substantial improvements in overall segmentation performance.

On the KiTS19 dataset, FreqDualNet achieved faster convergence and more stable performance compared to the baseline. While the baseline model exhibited a sharp decline in test performance after approximately 120 epochs, our FreqDualNet maintained a stable Dice score around 0.5 throughout 600 epochs. Specifically, FreqDualNet achieved a Dice score of 50.72%, compared to 44.99% from the baseline, indicating a 12.75% improvement. Similarly, in terms of DiceCE Loss, our model yielded a DiceCE Loss of 0.19, showing an 8.49% reduction compared to the baseline loss of 0.21, further supporting its performance.

On the Prostate158 dataset, FreqDualNet exhibited more stable convergence, maintaining a higher Dice score of approximately 0.22 throughout 600 training epochs. Specifically, FreqDualNet achieved a Dice score of 0.2282, compared to 0.2138 for the baseline model, representing a 6.70% improvement. Moreover, the baseline model showed a strong tendency toward overfitting, with training Dice scores exceeding 0.8, whereas FreqDualNet demonstrated reduced overfitting behavior, indicating improved generalization capability.

Visual inspection also revealed that FreqDualNet more accurately captured meaningful high frequency tumor margins, such as lobulated and infiltrative patterns. For instance, in cases with rounded tumor shapes, the baseline model produced distorted predictions, while our model preserved the correct shape. Additionally, in samples where the baseline failed to detect the tumor, FreqDualNet correctly localized the tumor in terms of position and size.

Ablation studies confirmed the contribution of each frequency aware module. Removing either the FCA or frequency based filtering layers led to significant performance drops, 11.41% and 29.14%, stressing the usefulness of combining both mechanisms for KiTS19 dataset.

Overall, FreqDualNet demonstrated not only improved accuracy but also enhanced sensitivity to anatomically relevant structures. These findings suggest strong potential for the model's application in future clinical decision support systems.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Research Foundation of Korea (NRF) through the Korean Government (MSIT) under Grant RS-2024-00337250. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Research Foundation of Korea (NRF) through the Korean Government (MSIT): RS-2024-00337250.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Jungro Lee conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Minhyeok Lee performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available in the [Supplemental File](#).

The data is available at <https://kits19.grand-challenge.org>.

The code is available at GitHub and Zenodo:

- <https://github.com/JungroLee/FreqDualNet>.

- Lee, M., & Lee, J. (2026). Python Code for the article “FreqDualNet: Frequency Aware Vision Transformers for Tumor Segmentation”. Zenodo. <https://doi.org/10.5281/zenodo.18434819>.

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj-cs.3801#supplemental-information>.

REFERENCES

- Abdelrahman A, Viriri S. 2022.** Kidney tumor semantic segmentation using deep learning: a survey of state-of-the-art. *Journal of Imaging* **8**(3):55 DOI [10.3390/jimaging8030055](https://doi.org/10.3390/jimaging8030055).
- Adams LC, Makowski MR, Engel G, Rattunde M, Busch F, Asbach P, Niehues SM, Vinayahalingam S, van Ginneken B, Litjens G. 2022.** Prostate158-an expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. *Computers in Biology and Medicine* **148**:105817 DOI [10.1016/j.compbiomed.2022.105817](https://doi.org/10.1016/j.compbiomed.2022.105817).

- Agrawal G, Su MY, Nalcioğlu O, Feig SA, Chen JH. 2009.** Significance of breast lesion descriptors in the ACR BI-RADS MRI lexicon. *Cancer: Interdisciplinary International Journal of the American Cancer Society* 115(7):1363–1380 DOI 10.1002/cncr.24156.
- Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E, Colak E, Farahani K, Kalpathy-Cramer J, Kitamura FC, Pati S. 2021.** The RSNA-ASNR-MICCAI brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *ArXiv* DOI 10.48550/arXiv.2107.02314.
- Cai M, Zhang H, Huang H, Geng Q, Li Y, Huang G. 2021.** Frequency domain image translation: more photo-realistic, better identity-preserving. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE, 13930–13940.
- Capellán-Martín D, Jiang Z, Parida A, Liu X, Lam V, Nisar H, Tapp A, Elsharkawi S, Ledesma-Carbayo MJ, Anwar SM. 2023.** Model ensemble for brain tumor segmentation in magnetic resonance imaging. In: *International Challenge on Cross-Modality Domain Adaptation for Medical Image Segmentation*. Cham: Springer, 221–232 DOI 10.1007/978-3-031-76163-8_20.
- Choi SR, Gu C, Baek SJ, Lee M. 2025.** Rendering 3D CT scans through 3D gaussian splatting initialized with points sampled by cube-based neural radiance fields. In: *2025 International Conference on Artificial Intelligence in Information and Communication (ICAIC)*. Piscataway: IEEE, 0510–0513.
- Choi SR, Ko K, Baek SJ, Lee S, Lee J, Lee M. 2024.** Enhanced kidney tumor segmentation in CT scans using a simplified UNETR with organ information. In: *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIC)*. Piscataway: IEEE.
- Choi SR, Lee J, Lee M. 2024.** OrgUNETR: utilizing organ information and squeeze and excitation block for improved tumor segmentation. *IEEE Access* 12:84122–84133 DOI 10.1109/access.2024.3413717.
- Fu S, Lu Y, Wang Y, Zhou Y, Shen W, Fishman E, Yuille A. 2020.** Domain adaptive relational reasoning for 3D multi-organ segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer, 656–666.
- Hansell DM, Bankier AA, MacMahon H, McCloud TC, Muller NL, Remy J. 2008.** Fleischner society: glossary of terms for thoracic imaging. *Radiology* 246(3):697–722 DOI 10.1148/radiol.2462070712.
- Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. 2021.** Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: *International MICCAI Brainlesion Workshop*. Cham: Springer, 272–284.
- Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, Xu D. 2022.** UNETR: transformers for 3D medical image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 574–584.
- He Y, Nath V, Yang D, Tang Y, Myronenko A, Xu D. 2023.** SwinUNETR-V2: stronger swin transformers with stagewise convolutions for 3D medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer, 416–426.
- He K, Zhang X, Ren S, Sun J. 2016.** Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 770–778.
- Heller N, Isensee F, Maier-Hein KH, Hou X, Xie C, Li F, Nan Y, Mu G, Lin Z, Han M. 2021.** The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the KiTS19 challenge. *Medical Image Analysis* 67(1):101821 DOI 10.1016/j.media.2020.101821.

- Heller N, Sathianathen N, Kalapara A, Walczak E, Moore K, Kaluzniak H, Rosenberg J, Blake P, Rengel Z, Oestreich M. 2019. The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. ArXiv DOI 10.48550/arXiv.1904.00445.
- Hu J, Shen L, Sun G. 2018. Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 7132–7141.
- Jiang L, Dai B, Wu W, Loy CC. 2021. Focal frequency loss for image reconstruction and synthesis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE, 13919–13929.
- Khan A, Rauf Z, Khan AR, Rathore S, Khan SH, Shah NS, Farooq U, Asif H, Asif A, Zahoora U. 2023. A recent survey of vision transformers for medical image segmentation. ArXiv DOI 10.48550/arXiv.2312.00634.
- Li HB, Conte GM, Hu Q, Anwar SM, Kofler F, Ezhov I, van Leemput K, Piraud M, Diaz M, Cole B. 2024. The brain tumor segmentation (BraTS) challenge 2023: brain MR image synthesis for tumor segmentation (BraSyn). ArXiv DOI 10.48550/arXiv.2305.09011.
- Li C, Liu Q, Song J. 2026. Boundary-enhanced sparse transformer for generalizable and accurate medical image segmentation. *Scientific Reports* 16:2191 DOI 10.1038/s41598-025-33686-0.
- Li Q, Shen L. 2022. WaveSNet: wavelet integrated deep networks for image segmentation. In: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Cham: Springer, 325–337.
- Li P, Zhou R, He J, Zhao S, Tian Y. 2023. A global-frequency-domain network for medical image segmentation. *Computers in Biology and Medicine* 164(5):107290 DOI 10.1016/j.compbiomed.2023.107290.
- Loshchilov I, Hutter F. 2017. Decoupled weight decay regularization. ArXiv DOI 10.48550/arXiv.1711.05101.
- McGrath H, Li P, Dorent R, Bradford R, Saeed S, Bisdas S, Ourselin S, Shapey J, Vercauteren T. 2020. Manual segmentation versus semi-automated segmentation for quantifying vestibular schwannoma volume on MRI. *International Journal of Computer Assisted Radiology and Surgery* 15(9):1445–1455 DOI 10.1007/s11548-020-02222-y.
- Mickevicius P, Narang S, Alben J, Diamos G, Elsen E, Garcia D, Ginsburg B, Houston M, Kuchaiev O, Venkatesh G. 2017. Mixed precision training. ArXiv DOI 10.48550/arXiv.1710.03740.
- Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. 2021. NeRF: representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* 65(1):99–106 DOI 10.1145/3503250.
- Ott M, Edunov S, Grangier D, Auli M. 2018. Scaling neural machine translation. ArXiv DOI 10.48550/arXiv.1806.00187.
- Qin Z, Zhang P, Wu F, Li X. 2021. FcaNet: frequency channel attention networks. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE, 783–792.
- Rayed ME, Islam SS, Niha SI, Jim JR, Kabir MM, Mridha M. 2024. Deep learning for medical image segmentation: state-of-the-art advancements and challenges. *Informatics in Medicine Unlocked* 47(7):101504 DOI 10.1016/j.imu.2024.101504.
- Ronneberger O, Fischer P, Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer, 234–241.
- Shorten C, Khoshgoftaar TM. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data* 6(1):1–48 DOI 10.1186/s40537-019-0197-0.

- Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: *International Workshop on Deep Learning in Medical Image Analysis*. Cham: Springer, 240–248.
- Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. 2015. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE.
- Thiese MS, Ronna B, Ott U. 2016. P value interpretations and considerations. *Journal of Thoracic Disease* 8(9):E928 DOI 10.21037/jtd.2016.08.16.
- Van Heel M, Schatz M. 2005. Fourier shell correlation threshold criteria. *Journal of Structural Biology* 151(3):250–262 DOI 10.1016/j.jsb.2005.05.009.
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30(NIPS 2017):5998–6008 DOI 10.48550/arXiv.1706.03762.
- Wang X, Chen Y, Hu S, Fan H, Zhu H, Li X. 2024b. Neural radiance fields in medical imaging: a survey. ArXiv DOI 10.48550/arXiv.2402.17797.
- Wang W, Wang J, Chen C, Jiao J, Cai Y, Song S, Li J. 2024a. FreMIM: fourier transform meets masked image modeling for medical image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Piscataway: IEEE, 7860–7870.
- Xu Y, Quan R, Xu W, Huang Y, Chen X, Liu F. 2024. Advances in medical image segmentation: a comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering* 11(10):1034 DOI 10.3390/bioengineering11101034.
- Yamashita R, Nishio M, Do RKG, Togashi K. 2018. Convolutional neural networks: an overview and application in radiology. *Insights Into Imaging* 9(4):611–629 DOI 10.1007/s13244-018-0639-9.
- Zhang F, Gu Z, Wang H. 2025. Decoding with structured awareness: integrating directional, frequency-spatial, and structural attention for medical image segmentation. ArXiv DOI 10.48550/arXiv.2512.05494.
- Zou Y, Lin Y, Zhu Q. 2024. PA-NeRF, a neural radiance field model for 3D photoacoustic tomography reconstruction from limited Bscan data. *Biomedical Optics Express* 15(3):1651–1667 DOI 10.1364/boe.511807.
- Zou F, Liu Y, Chen Z, Zhanghao K, Jin D. 2023. Fourier channel attention powered lightweight network for image segmentation. *IEEE Journal of Translational Engineering in Health and Medicine* 11:252–260 DOI 10.1109/jtehm.2023.3262841.