IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 July 2024, accepted 21 August 2024, date of publication 28 August 2024, date of current version 17 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3450961

RESEARCH ARTICLE

SDSL: Spectral Distance Scaling Loss Pretraining SwinUNETR for 3D Medical Image Segmentation

JIN LEE[1], (Graduate Student Member, IEEE), DANG THANH VU[2], GWANGHYUN YU[1], JINSUL KIM[1], (Member, IEEE), KUNYUNG KIM[3], AND JINYOUNG KIM[1], (Member, IEEE)

[1]Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, South Korea
[2]Research Center, AISeed Inc., Gwangju 61186, South Korea
[3]Department of Radiology, Seoul National University Bundang Hospital, Seongnam 13620, South Korea

Corresponding authors: Jinyoung Kim (beyondi@jnu.ac.kr) and Kunyung Kim (kky2kkw@gmail.com)

This work was supported in part by the Innovative Human Resource Development for Local Intellectualization Program through the Institute of Information and Communications Technology Planning and Evaluation (IITP), Korean Government (MSIT) under Grant IITP-2024-00156287, 50; and in part by the Artificial Intelligence Industrial Convergence Cluster Development Project, Ministry of Science and ICT (MSIT), South Korea and Gwangju Metropolitan City.

ABSTRACT Recent approaches utilizing self-supervised learning with masked image modeling (MIM) have demonstrated great performance. However, applying MIM naively to small datasets results in poor generalization to downstream tasks. We hypothesize that capturing detailed anatomical structures can compensate for the limitations posed by the dataset shortage; thus, we introduce Spectral Distance Scaling Loss (SDSL), designed to improve generalization in medical imaging tasks. Unlike traditional pixel-based methods, SDSL incorporates frequency-domain information to enhance encoder representations, ensuring balanced learning of both low and high-frequency details. Furthermore, wavelet multi resolution decomposition was utilized to enable the pretrained model to reconstruct frequency information across multiple stages. The comprehensive experiments demonstrate that SDSL pretraining yields sharper reconstruction results and more accurate segmentation outcomes than existing methods. The proposed approach achieved the highest average Dice scores of 84.17% on the Beyond the Cranial Vault dataset, 98.20% on the Medical Segmentation Decathlon Spleen dataset, and 90.38% on the Multimodality Whole Heart Segmentation dataset. The findings highlight the potential of SDSL in advancing medical imaging techniques by effectively handling spectral variations and improving model generalization.

INDEX TERMS Spectral bias, self-supervised learning, swin U-Net transformer, medical image segmentation.

I. INTRODUCTION

Among segmentation tasks, medical imaging poses unique challenges, especially in the precise delineation of anatomical structures. These structures, which include blood vessels, lesions, and small tumors [1], [2], [3], [4], often have intricate details represented by high-frequency components. Unlike natural images, medical images typically involve analyzing volumetric data, represented as a stack of 2D slices forming a 3D volume (e.g., MRI and computed tomography [CT] scans). Each voxel (3D pixel) contains intensity values corresponding to various tissues, making the data sparse and computationally demanding to process [5], [6], [7], [8], [9], [10]. Additionally, the high-end equipment and labor-intensive annotation process result in excessive costs, further restricting their usage.

A promising method to address dataset scarcity is self-supervised learning (SSL) using masked image modeling (MIM) [11], [12], [13]. This approach employs large amounts of unlabeled data to train models on a pre-text task, which can be fine-tuned for the downstream task of interest. In addition, MIM involves masking part of an image and training the model to predict the hidden content. Models trained this way often generalize well to new data because they learn rich representation during masking. The success of MIM is largely attributed to vision transformers [14], [15], which treat images as sequences of visual tokens, allowing for masking similar to how masked language models

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

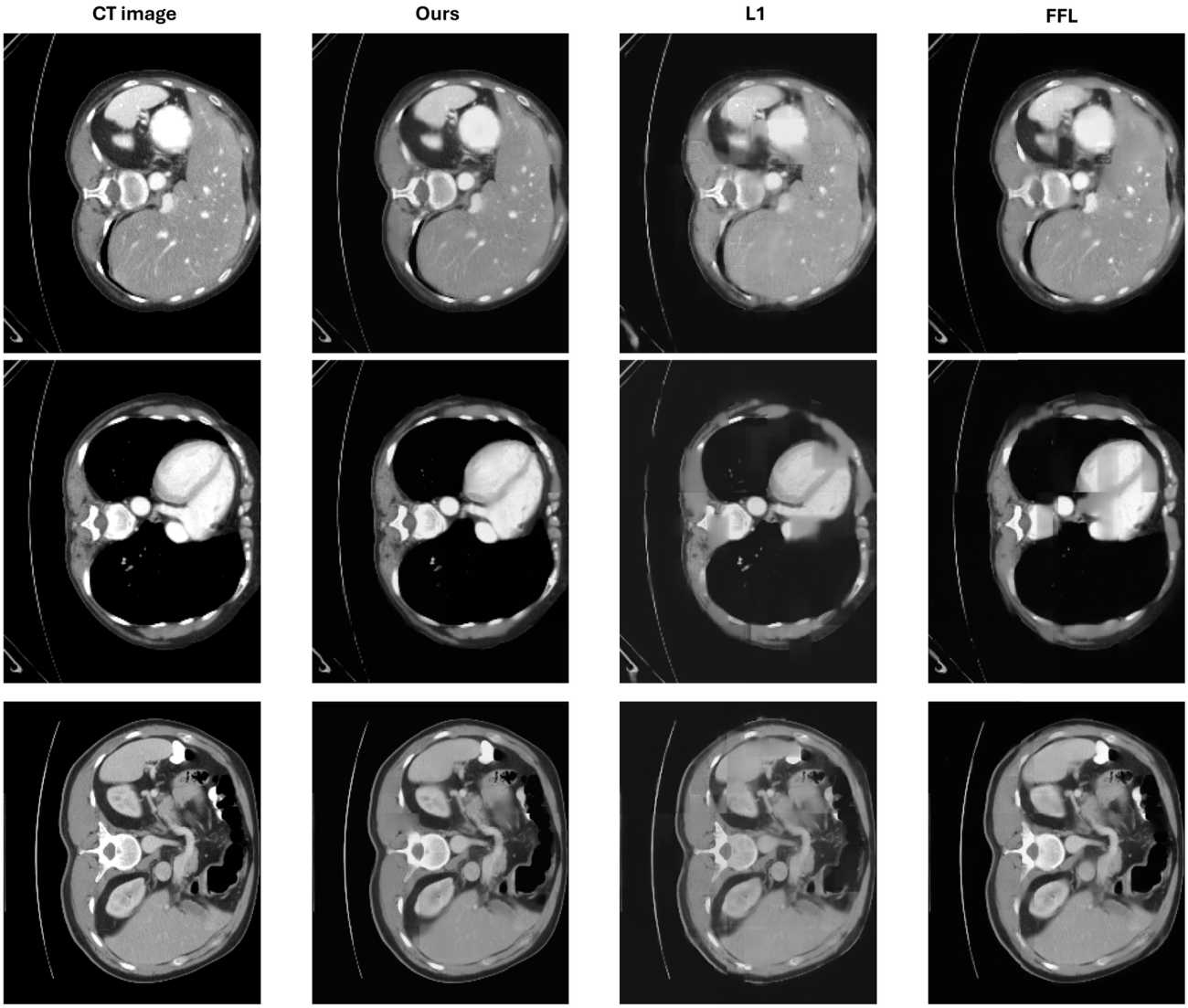


FIGURE 1. Groundtruth (CT image) and reconstruction comparison using the proposed spectral distance scaling loss method (ours), L_1 loss ($L1$), and focal frequency loss (FFL).

manage text. However, a critical difference between MIM and masked language modeling is that, although linguistic tokens have semantic meaning in a predefined vocabulary, visual tokens are merely grids of local pixels with fewer intrinsic constraints.

Despite this complexity, pioneering research has demonstrated that straightforward random masking strategies [16], [17] and predicting RGB values of raw pixels [11], [12], [18] can ensure learning transferable feature representations. Nevertheless, the effectiveness of such approaches often follows the neural scaling principle [19], where the performance of SSL improves with increased training data

Medical datasets are much smaller than natural image datasets, generally comprising thousands to tens of thousands of 3D volumes [20], [21], [22], [23] compared to millions of natural images [11], [13], [16]. Unlike natural images, which are standardized in RGB with values ranging from [0, 255], medical images (e.g., CT scans) are typically single-channel images with values ranging from [-1000, 1000]. Consequently, using pretrained models on natural images is unsuitable for these data because the modality varies. Moreover, although employing only pixel reconstruction loss in pretraining is beneficial for the spatial domain, it results in biased learning in the frequency domain, hindering the ability to acquire enhanced representations.

Notably, Fourier phase information encapsulates high-level semantics and remains resilient to transfer learning. The low-level spectrum can vary significantly without affecting high-level semantics. This variability, influenced by factors such as the sensor or illuminant, affects the spectral characteristics and necessitates that learning-based models account for these variations [24]. In medical data, fine details such as lesions or small organs are important within the overall volume. Therefore, learning biased toward a certain

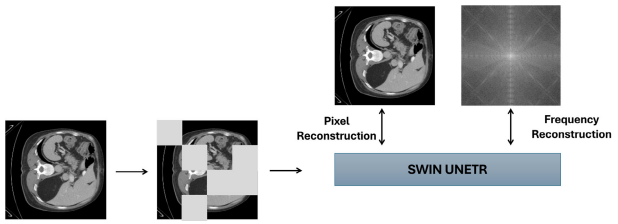


FIGURE 2. Pretraining framework with masked image modeling. The input image is randomly masked, with the visible parts input into the Swin U-Net transformer model to predict the missing parts by minimizing the reconstruction differences.

frequency band can be seen to have a negative impact on down-stream segmentation. To address this limitation, the research community has advanced techniques that enhance a model's ability to predict masked regions more accurately by manipulating frequency components [25], [26], [27], [28], [29]. These efforts aim to compensate for the smaller dataset size by improving the pretrained model to capture detailed information.

Focal frequency loss (FFL) [30] measures the Euclidean distance between spectral vectors and facilitates weighted frequency loss for improved image reconstruction and synthesis. Reducing this difference improves the representation learning capability, enhancing the performance on downstream tasks [25], [27], [29]. However, FFL still has room for further improvement. As illustrated in Fig. 1, using FFL in pretraining yields sharper reconstruction results than using only the pixel loss (L_1). Nevertheless, the contours or edges remain incomplete in some image areas. This problem arises because the magnitude value of the low-frequency band, including the DC component, is significantly large when applying the discrete Fourier transform (DFT) to the reconstructed image. Thus, the spectrum distance has an excessively large dynamic range. The large magnitude dominates the frequency loss value in backpropagation, leading to spectral bias.

The convergence of neural networks exhibits spectral bias, the tendency of deep neural networks to prioritize fitting target functions from low to high frequencies during training [31], [32], [33], [34]. Recent research has presented contradictory yet supportive findings: neural networks tend to emphasize learning high-frequency components for semantic tasks, such as recognition [35], [36], [37], [38]. However, this emphasis shifts to low- or middle-frequency components for synthesis tasks [39], [40], [41], [42]. This phenomenon is evident because models can readily outline coarse (low-frequency) structures, whereas capturing finer (high-frequency) details remains challenging [43], [44], [45], [46].

Thus, this paper addresses these challenges using self-supervised pretraining with an adaptive loss function, featuring the spectral distance scaling loss (SDSL), scaling the spectral distance to ensure balanced frequency learning and enhancing the ability of the model to capture low- and high-frequency details.

Fig. 2 illustrates the overall mechanism of the proposed SDSL framework. This MIM-based SSL approach separates the reconstructed volume into multiresolution high-pass and low-pass bands through a 3D-wavelet transform. Subsequently, the framework is transformed into the frequency domain using a 3D-DFT. The distance between the transformed original and reconstructed volumes is calculated for each frequency, and these distances are scaled using the tanh function to ensure that no particular frequency band disproportionately influences the loss. To verify the performance of SDSL, we pretrained the model on 3647 CT volumes from 11 datasets, including the Abdomen 1k and Word datasets. The evaluation was conducted using Beyond the Cranial Vault (BTCV) [4], Medical Segmentation Decathlon (MSD) Spleen [2], Multimodality Whole Heart Segmentation MM-WHS [47], and CT-ORG [48] datasets. In downstream segmentation tasks, SDSL recorded the highest Dice scores across all datasets compared to the previous MIM-based methods [20], [22], [29].

The contributions of this study are as follows:

- The novel loss function SDSL scales spectral distances to balance frequency learning, improving the ability to capture low- and high-frequency details.
- A training SSL framework for effective representation learning is constructed in the frequency domain based on MIM.
- The effectiveness of the proposed approach is demonstrated by pretraining the model on 3679 CT volumes from 11 datasets, and the evaluation of the model on multiple downstream segmentation tasks reveals that it achieves superior performance compared to existing methods.

The remaining sections are organized as follows: Section II reviews the related work in the field, focusing on 3D medical image segmentation and recent advancements in MIM frameworks. Next, Section III details the proposed SDSL, and Section IV discusses the performance evaluation. Finally, Section V presents the discussion and conclusions.

II. RELATED WORK

A. 3D MEDICAL IMAGE SEGMENTATION

In medical imaging, 3D medical image segmentation is crucial and involves partitioning a 3D medical scan (e.g., MRI and CT) into regions that correspond to various anatomical structures or pathological regions [2], [3], [4]. This process aids in the diagnosis, treatment, and monitoring of disease progression. The state-of-the-art architecture for medical image segmentation employs deep learning models [10], [49], [50], [51], primarily categorized as fully convolutional networks [5], [6], [9] and transformer-based networks [8], [22], [52], [53], [54].

The V-Net is a 3D convolutional neural network (CNN) capable of end-to-end training for MRI image segmentation [55]. A double-pathway 3D-CNN segments brain lesions at multiple scales [6], which are enhanced using a conditional random field to mitigate false positives, improving the computational efficiency of the 3D data [5]. Addressing the

orderliness of volumetric data, a study [7] combined fully convolutional networks with recurrent neural networks to manage intra- and inter-slice components for 3D fungus segmentation

Another cornerstone in this field is the U-Net architecture [51], which was initially designed for 2D images [56] but was extended to 3D volumetric images [57]. As a notable variant, attention gates have been integrated into the U-Net to enhance its ability to capture low-level features efficiently, creating an (*attention U-Net*). Additionally, the H-DenseUNet [58] combines the 2D DenseUNet for intra-slice feature extraction with a 3D component to aggregate the volumetric context explicitly for liver tumor segmentation. Recognizing that manual intervention is tedious; thus, nnU-Net [59] was designed as a self-configuring U-Net that optimizes segmentation tasks using empirical rules and interdependent configurations.

The intrinsic locality of convolutional operations may limit the ability of a model to capture long-range dependencies in an image. These architectures have difficulty integrating global contextual information that is crucial for accurately segmenting large structures. The fixed receptive field of convolutional layers may not be optimal to segment objects with varying sizes and shapes accurately in medical images. Many attempts, such as widening the receptive field with atrous convolution [60], [60], [61] or applying cascaded modules to extract regions of interest [62], [63], [64], have been made to address these problems. As a compromise, vision transformer models [14], [15] have emerged because their self-attention mechanism is adept at capturing the global context.

Ali et al. treated 3D volumetric medical image segmentation as a sequence-to-sequence prediction problem with the introduction of UNETR, which employs direct skip connections at various resolutions [22], [53], [54]. Building on this foundation, subsequent models, such as Swin UNETR [52], Swin-UNet [65], Swin UNETR-V2 [66], and DS-TransUNet [67] replaced the transformer backbone with a hierarchical Swin transformer. These advancements display promising performance on 3D medical benchmark datasets. Inspired by this innovative work integrating transformers into the U-shape architecture, this study applies Swin UNETR as a backbone model, pretraining it on medical datasets to enhance its segmentation performance.

Recent hybrid models combining CNNs and transformers have gained popularity in the field. TransFuse [68] and TransUNet [54] integrate multi level features from transformers and CNN branches in parallel, capturing global and local features for segmentation tasks. In contrast, TransBTS [69] applies a 3D-CNN to extract spatial features before inputting them into a transformer for progressive upsampling to predict the final segmentation mask. In addition, CoTR [8] reduces computational complexity via a deformable self-attention mechanism, enhancing efficiency in large-scale image segmentation. Moreover, MedT [70] employs a gated axial-attention module that adapts to smaller datasets. However, training a model with more data offers better results, as explored below.

B. SELF-SUPERVISED LEARNING FOR 3D MEDICAL IMAGE SEGMENTATION

The SSL method has emerged as a promising approach to medical image segmentation [21], [25], [71], [72], [73], [74], applying numerous unlabeled data to pretrain models that can be fine-tuned on smaller labeled datasets [68], [75], [76]. In SSL, the model learns representations by solving pre-text tasks without manual annotations [18], [77].

The two primary approaches to defining a pre-text task are contrastive learning (CL) [78], [79], [80] and context restoration (CR) [11], [13], [16], [81], [82], [83]. Specifically, CL focuses on learning semantic features, working primarily in the representation space by contrasting positive pairs (augmentations of the same image) against negative pairs (different images). In contrast, CR emphasizes the spatial context, in which the model learns to reconstruct missing (or corrupted) parts of an image.

In the CL framework, the task specificity considerably influences the pretraining phase. In 3D medical imaging, proxy tasks (e.g., cube recovery) involving rearranging, rotating, and masking, encourage networks to learn features invariant to translation and rotation [84], [85]. In modality CL, pre-text tasks are extended by incorporating inpainting alongside contrastive coding and rotation prediction [71]. These advancements underscore the versatility and effectiveness of integrating diverse pre-text tasks within the CL framework for enhancing feature learning.

Unlike CL, CR maintains spatial relationships and semantic consistency in reconstructed images. There has been significant research progress in the field of medical imaging in this direction, including masked autoencoders to pretrain vision transformer encoders and fine-tune UNETR decoders [22], [23]. The SwinMM [20], [29] employs a mutual learning task to align reconstructions from multiple views. Using a high masking ratio and smaller patch size to predict raw voxel values is effective [21], as opposed to using FreMIM [25], which emphasizes learning high-frequency components.

This work employs the MIM framework to train a pretrained Swin UNETR and fine-tune it for downstream 3D medical image segmentation. We hypothesize that the learning through MIM by the encoder is effective because it prioritizes spatial relationships and semantic consistency over semantic coherence, and abundant evidence supports this hypothesis [11], [16], [83], [86].

C. LEARNING IN THE FREQUENCY DOMAIN

In evaluating the pretrained model, MIM applies the discrepancy between reconstructed images and ground-truth data. In this context, pixel-based losses have well-known limitations that can lead to suboptimal learning outcomes in medical images. For example, these losses are sensitive

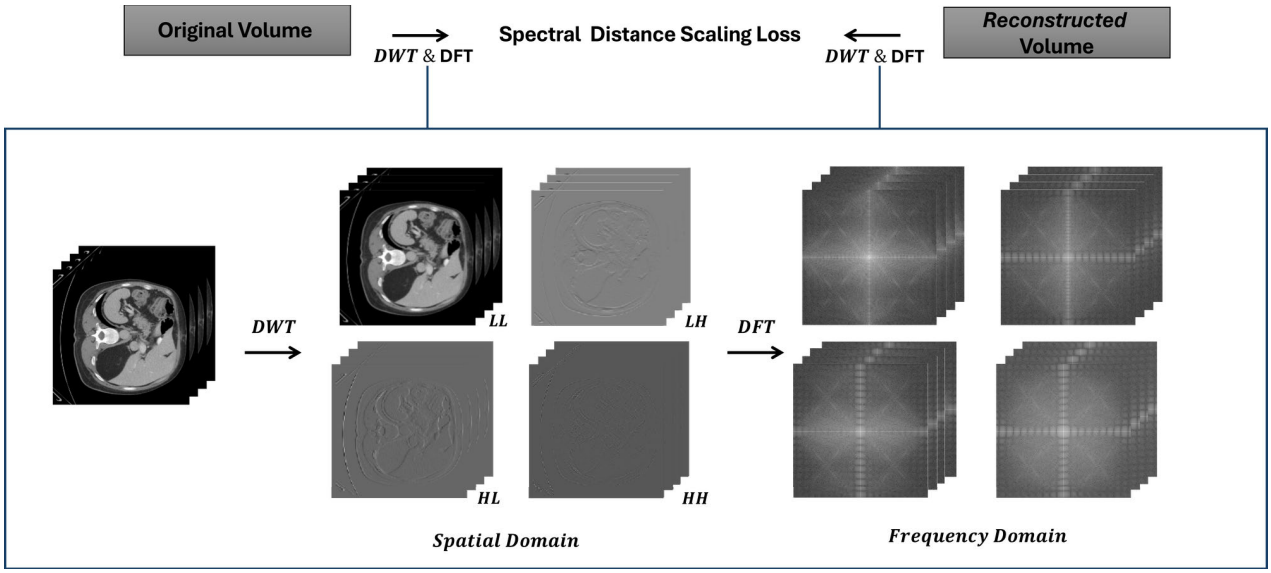


FIGURE 3. Pretraining a model using wavelet multi-resolution decomposition and spectral distance scaling loss.

to noise [87], [88], ignore structural information [89], [90], and inadequately capture high-frequency details. Researchers often combine pixel-based losses with other types, such as frequency-domain losses, to address these problems [30], [39], [42], [91], [92].

Models can preserve the global structure (low frequencies) and fine details (high frequencies) by calculating the discrepancy (e.g., the L2 distance) between the frequency coefficients of the reconstructed image and ground truth. This approach is exemplified by FFL [30], multi-spectral channel attention [93], dynamic spectrum loss [94], GLaMa [95], and spectral distribution aware [96], making the frequency distribution uniform. Similar to Fourier-based losses, wavelet loss calculates the discrepancies in the wavelet domain, encouraging models to learn representations that are robust across scales and frequencies, such as the wavelet-domain high-frequency Loss [39], which enhances FFL by combining log scaling to reduce the influence of low-frequency bands.

This work builds on the frequency spectrum analysis, revealing that models have difficulty maintaining high-frequency information because they tend to prioritize reconstructing low frequencies. To address this problem, we propose SDSL to regularize the frequency throughput during pretraining.

III. METHOD

This section describes the overall mechanism of the proposed framework for effective frequency-domain pretraining and explains the SDSL

A. 3D MASKED IMAGE MODELING

The pretraining phase follows the mechanism of the MIM framework, which pre-learns image representations without labels by masking a portion of the image and restoring it. Unlike existing methods (e.g., DAE [29], SwinMM [52], SSL-Swin UNETR [71]) that alter masking mechanisms and pre-text tasks in the spatial domain, the proposed work addresses spectral bias by employing frequency regression and pixel regression.

The overall process of the proposed framework for frequency-domain representation learning is depicted in Fig. 3. As illustrated, the input volume is reconstructed using the Swin UNETR model, and the reconstructed volume is decomposed into high- and low-pass bands via a wavelet transform, facilitating the multiresolution analysis. This decomposition allows the 3D volume to be examined at various resolutions, allowing a detailed analysis of the volume characteristics at each level. The Haar wavelet served as the foundation for the wavelet transform in this study. After wavelet decomposition, the volume undergoes a transition into the frequency domain by performing a 3D-DFT. The next step involves calculating the magnitude of the distance of spectral vectors between the original volume (serving as the ground truth) and the reconstructed volume. Next, a scaling function is employed for these spectral distances, enhancing the capacity of the model to extract detailed frequency information.

During pretraining, the embedded patch in the input CT volume is masked and reconstructed based on the existing MIM mechanism. The output is used to calculate the original input, and L_{pixel} and L_{SDSL} are calculated as described in Section III-B. Throughout pretraining, the encoder and decoder in Swin UNETR were employed in their original form without a separate decoder. For the downstream task, the encoder and decoder were retained without modification in the pretrained model. In addition to the SDSL, the pixel loss between the ground truth and reconstructed volume was calculated, including the masked patch and all remaining patches, and the total loss is given as follows:

$$L_{\text{total}} = L_{\text{pixel}} + L_{\text{SDSL}} \tag{1}$$

where L_{pixel} is defined by the L_1 loss, and L_{SDSL} is detailed below.

B. SPECTRAL DISTANCE SCALING LOSS

In computer vision tasks, the objective function in the frequency domain often employs FFL. The equation for this function is as follows [30]:

$$L_{\text{FFL}} = \frac{1}{N_x N_y} \sum_{u=0}^{N_x-1} \sum_{v=0}^{N_y-1} w(u,v) \left| \mathcal{F}(u,v) - \mathcal{F}_{\text{pred}}(u,v) \right|^2, \tag{2}$$

where

$$w(u,v) = \left(\left| \mathcal{F}_{\text{gt}}(u,v) - \mathcal{F}_{\text{pred}}(u,v) \right| \right)^{\alpha} \tag{3}$$

where N_{x} and N_{y} denote the width and height of the image, respectively, and $\mathcal{F}_{\text{pred}}$ represents the reconstructed spectrum. In FFL, w represents the spectrum weight, an adaptively generated values in the range of [0, 1] based on the spectrum distance. Compared with previous reconstruction results and downstream tasks that primarily focus on the low-pass band, minimizing the frequency distance through FFL improves the performance.

Although the FFL has a weighted distance, the magnitude of the DC components remains relatively large, significantly influencing low-pass bands during backpropagation. Low-pass bands contain ample semantic information, and fine-grained features in high-pass bands in medical imaging data are also critical for downstream tasks. Therefore, adequately considering information at high spatial frequencies during pretraining is critical. This study introduces an improved scaling function designed to facilitate balanced learning across frequency bands. The proposed method determines the loss value for 3D CT volumes and is formulated based on the 3D-DFT.

We introduce the 3D-DFT formula (4), transforming volumetric data into the frequency domain. Compared with the 2D-DFT, the 3D-DFT functions with an additional dimension denoted by z, corresponding to the frequency dimension ϕ, where N_z indicates the volume depth:

$$\mathcal{F}(u,v,w) = \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} \sum_{z=0}^{N_z-1} f(x,y,z) e^{-j2\pi \left(\frac{ux}{N_x} + \frac{vy}{N_y} + \frac{wz}{N_z} \right)} \tag{4}$$

The SDSL method applies a scaling function to modify the spectral distance distribution in the frequency domain in the SSL framework. In the calculation process, the distance between the reconstructed volume and ground truth is calculated after applying the 3D-DFT. Then, the scaling function is applied to the spectral distance distribution. Using 3D frequency components, SDSL (5) is defined as the arithmetic average of the scaling distance between the predictive frequency $\mathcal{F}_{pred}$ and its ground-truth counterparts $\mathcal{F}_{gt}$, where $\rho : [0, +\infty) \rightarrow [0, 1]$ represents a scaling function,

$$L_{\text{SDSL}} = \frac{1}{N_u N_v N_W} \times \sum_{u=0}^{N_u-1} \sum_{v=0}^{N_v-1} \sum_{w=0}^{N_w-1} \rho(|\mathcal{F}_{\text{gt}}(u,v,w) - \mathcal{F}_{\text{pred}}(u,v,w)|) \tag{5}$$

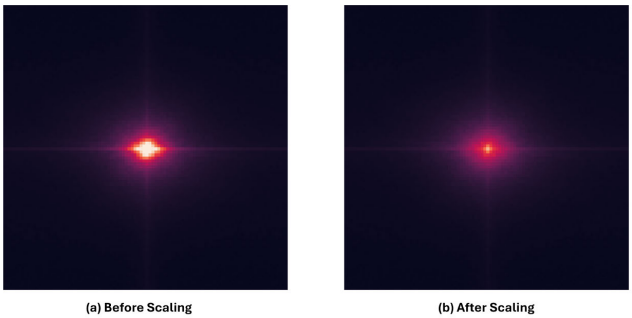


FIGURE 4. Spectral magnitude scaling. (a) Before scaling. (b) After scaling with the tanh function.

Typically, before scaling, the residual between $\mathcal{F}_{\text{pred}}$ and $\mathcal{F}_{\text{gt}}$ at low frequencies is relatively large. Conversely, the magnitude of high spatial frequency components is very small, so the value difference between $\mathcal{F}_{pred}$ and $\mathcal{F}_{gt}$ is proportionally suppressed. In L_{SDSL}, the scaling function ρ serves to reduce these distance differences in the spectral coordinates. Through the scaling function, which saturates at large values, the spectral distance is normalized to the range [0, 1], and the disparity between the distance values of the spectral coordinates is adjusted. In this study, ρ is implemented as the tanh function, and ablation study will dive more into the behavior of chosen scaling functions.

Large gaps among spectral values adversely influence pretraining. Before scaling, low-frequency components with large values cause the representation to be learned with a focus on these bands during backpropagation. This phenomenon facilitates the learning of large anatomical structures, such as the overall shape and liver in the CT volume, but is relatively insufficient for learning fine details (e.g., edges, texture, and contour). The SDSL mitigates this problem by preventing the spatial frequency of a specific band from excessively influencing the learning process during pretraining, promoting balanced spatial frequency learning.

Fig. 4 illustrates the effect of the scaling function operation. The left panel depicts the average spectral magnitude after applying DFT to the BTCV validation dataset, whereas the right panel displays the result after scaling using the tanh function. On the left, the DC component values and certain low spatial frequencies in the center are very large. Applying the scaling function (right side) noticeably reduces the relative magnitude difference in the spatial frequency.

C. SCALING AS A ROBUST ESTIMATOR

This section clarifies the concept behind scaling the objective function. We reformulated the problem of spectral bias as

a problem of hard-sample mining, where the magnitude of low-frequency components dominates the loss despite comprising only a small portion of it, and the high-frequency components contribute insignificantly to the loss, making them difficult to optimize. A robust estimator can minimize the influence of training samples with unusually large errors on the training procedure by down-weighting them in the loss function minimization.

The objective function can be written as a sum of the scaled L_1 spectrum distance between ground truth and prediction, where $\epsilon_i = \mathcal{F}_{\text{gt}}(u, v, w) - \mathcal{F}_{\text{pred}}(u, v, w)$,

$$\sum_{i=0}^{N} \rho(|\epsilon_i|) \tag{6}$$

where $i = 0 \ldots N (= N_u \times N_v \times N_w)$. The unscaled version of the squared error $\sum_{i=0}^{N} \epsilon_i^2$ and focal version $\sum_{i=0}^{N} |\epsilon_i|^{\alpha} \epsilon_i^2$ have previously been employed. This method generalizes the focal loss using a scaling function ρ to normalize the large dynamic range of $\{\epsilon_i\}_N$. Notably, 6 is a generalization of the minimum log-likelihood, approximating an M-estimator. For instance, the tanh function naturally down-weights extreme values, minimizing their influence on the overall loss. This behavior is similar to other robust loss functions in M-estimators, such as the Huber loss or Tukey's biweight function [97].

Instead of using L_2 distance, we consider using L_1 due to its more favorable characteristics in learning within the frequency domain. To be precise, if the learning is conducted properly, $\mathcal{F}_{\text{pred}}$ and $\mathcal{F}_{\text{gt}}$ will each follow a Gaussian distribution, implying $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Hence the distribution of ϵ^2will be a chi-square distribution, $\epsilon^2 \sim \chi^2(1)$, while $|\epsilon|$ will follow a half-normal distribution with a relatively lower variance.

IV. EXPERIMENTS

This chapter describes the datasets used for pretraining and the downstream task. We compare the performance of existing methods on the downstream segmentation task using the SwinUNETR model pretrained with SDSL. The baselines include models without SSL pretraining (e.g., SwinUNETR [52] and UNETR [22]) and models with SSL pretraining (e.g., SSL-SwinUNETR, DAE [29], and SwinMM [20]). All five models are explicitly designed for medical segmentation tasks and have been publicly evaluated on the datasets presented in this manuscript, providing fair baselines to demonstrate our performance. We fine-tuned the SSL pretrained weights obtained from previous work and compared the segmentation results under identical conditions to ensure a fair comparison, and qualitative comparisons are provided for each segmentation task.

All training implementations employed the open-source MONAI[1], and the experiments were conducted on devices equipped with four Nvidia A100 GPUs.[1]

[1]https://monai.io

A. TRAINING MASKED IMAGE MODELING

The pretraining dataset combines 11 datasets, including the Abdomen-1k [98], TCIA Covid19 [99], MSD [2] (lung, heart, liver, hippocampus, pancreas, spleen, and colon), TCIA LIDC, and TCIA Colon datasets, encompassing 3399 CT volumes and 280 brain MRI volumes. Among these, the MSD Spleen dataset was also employed for downstream segmentation tasks. The pretraining dataset includes CT scans of the abdomen, chest and MRI scans of the brain and heart. Tables 1 and 2 list the number of training and validation datasets, respectively.

TABLE 1. Training datasets and details used in pretraining.

Training Dataset	Modality	Number of Data
Abdomen 1k	CT	1062
MSD Lung	CT	63
MSD Heart	MRI	20
MSD Liver	CT	131
MSD Pancreas	CT	281
MSD Hepatic Vessel	CT	303
MSD Spleen	CT	41
MSD Colon	CT	126
MSD Hippocampus	MRI	260
TCIA Covid19	CT	722
TCIA LIDC	CT	670

TABLE 2. Validation datasets and details in the pretraining phase.

Validation Dataset	Modality	Number of Data
MSD Pancreas	CT	139
MSD Liver	CT	49
TCIA Covid19	CT	70

This work applied Swin UNETR in pretraining and downstream tasks due to its high segmentation performance in previous studies. The pretraining settings for the proposed model include 500K iterations. The spatial resolution was maintained at a spacing of [2.0, 2.0, 2.0]. The normalization parameters were set to values ranging from -160 to 250 for a and 0 to 1 for b. Random spatial cropping was performed on the scans with a region of interest size of $96 \times 96 \times 96$. The model was optimized using the AdamW optimizer, with an initial learning rate of $4e^{-4}$. The learning rate was adjusted using a step decay schedule, with a warm-up phase implemented for the first 500 iterations. The weight decay was set to $2e^{-5}$, and the drop path rate was 0.1. Additionally, an exponential moving average with a decay rate of 0.9999 was applied to stabilize training.

B. 3D MEDICAL IMAGE DOWNSTREAM SEGMENTATION

We evaluated the pretrained model on downstream segmentation tasks on the BTCV, MSD Spleen, and MM-WHS datasets. The configurations for each dataset differ based primarily on prior studies conducted on these datasets and the environmental constraints affecting them. Table 3 details the configurations.

1) BEYOND THE CRANIAL VAULT DATASET

The BTCV abdomen dataset challenge comprises 30 abdominal scans, with 13 organs labeled under the supervision

TABLE 3. Down-stream configuration details for the BTCV, MSD Spleen, and MM-WHS datasets.

	BTCV	MSD Spleen	MM-WHS
Spacing	[1.5, 1.5, 2.0]	[1.5, 1.5, 2.0]	[1.5, 1.5, 1.5]
Norm a	[-175, 250]	[-175, 250]	[0, 1700]
Norm b	[0, 1]	[0, 1]	[0, 1]
Training epochs	2500	1000	160
Batch size	2	1	2
Sw batch size	4	4	4
Inference	sliding window	sliding window	sliding window
Augmentation	RandFlip, RandRotate, RandShiftIntensity	RandFlip, RandRotate, RandShiftIntensity	RandFlip, RandRotate, RandShiftIntensity
RoI size	$96 \times 96 \times 96$	$96 \times 96 \times 96$	$64 \times 64 \times 64$
Optimizer	AdamW	AdamW	AdamW
Learning rate(LR)	$1e^{-3}$, (UNETR $1e^{-4}$)	$1e^{-3}$, (UNETR $1e^{-4}$)	$1e^{-3}$, (UNETR $1e^{-4}$)
Lr scheduler	step scheduler	warmup cosine	step scheduler
Warmup iteration	500	50	500
Weight decay	$2e^{-5}$	$5e^{-2}$	$5e^{-2}$
Drop path rate	0.1	-	0.1
Exponential moving average	0.9999	-	0.9999

of radiologists at Vanderbilt University Medical Center. In the BTCV segmentation task, the proposed method achieved a Dice score of 84.17%, surpassing that of DAE [29], the current state-of-the-art model on the BTCV challenge leaderboard. The proposed method demonstrated superior performance compared with previous MIM-based SSL methods. Notably, the Dice score for the stomach was 2.78% higher than that achieved by Swin UNETR SSL [71], which previously held the highest score among comparable methods. Furthermore, the proposed approach attained the highest Dice scores for the segmentation of the gallbladder, inferior vena cava, and left adrenal gland. The split between the training and validation datasets followed the protocol in [71]. Table 4 details the performance.

2) MSD DATASET

The MSD Spleen dataset is in the pretraining dataset, and the MSD spleen segmentation task distinguishes the spleen from the background. The original dataset did not offer a separate training and validation split; hence, the performance was verified using five-fold cross-validation, where each fold selection was set to the same seed. This comparison revealed that the proposed method outperformed the others with a Dice score of 98.20%. Table 5 presents the performance details.

3) MULTIMODALITY WHOLE HEART SEGMENTATION DATASET

The MM-WHS dataset has 14 CT volumes and 4 validation volumes and is employed for heart segmentation, focusing on eight distinct heart regions. The proposed method exhibited the highest mean Dice score of 90.38% and the approach excelled at segmenting the right atrium (96.07%) and demonstrated strong performance across all structures, particularly in the left atrium (88.03%) and pulmonary artery (82.98%). Table 6 presents the performance details.

4) SEGMENTATION QUALITY ASSESSMENT

Fig. 5 comparatively visualizes the abdominal CT scan (BTCV) segmentation results from various methods, with each row representing a different CT scan slice. The first column depicts the ground truth, consisting of expert manual annotations serving as the reference standard. The second column displays the results from the proposed segmentation method, followed by those from the Swin UNETR, DAE, SSL-Swin UNETR, and FFL methods. The comparison highlights the accuracy, with color-coded regions indicating the differences in the segmentation of various organs and tissues. The segmentations under the proposed approach align with the ground truth, demonstrating well-delineated outputs for all sub regions.

Additionally, Fig. 6 visualizes the segmentation results on the MM-WHS dataset. In the first row, in the region highlighted in the red square, the proposed model accurately separates the two segmented regions (in yellow and pink), whereas the other methods confuse the boundary between these regions. Similarly, the proposed model successfully delineates the object (in purple) compared to other methods that miss some areas near the boundary. This assessment demonstrates that the proposed method outperforms previous methods in terms of multimodality whole heart segmentation.

C. ABLATION STUDIES

We conducted further experiments to clarify the effects of scaling functions, masking ratios, loss functions, and pretraining iterations. These ablation studies were conducted on the BTCV dataset.

1) SCALING FUNCTION

In this experiment, the tanh, Gaussian, and Laplace scaling functions were applied and compared with no scaling. These scaling functions normalize the spectral distances to a range of [0, 1]. For the complement Laplace, $\text{Laplace}(\text{LFD}|\mu,\sigma) = 1 - \frac{1}{2\sigma}\exp\left(-\frac{|\text{LFD}-\mu|}{\sigma}\right)$, we applied $\mu = 0, \sigma = 2$, whereas for the complement Gaussian, $\text{Gaussian}(\text{LFD}|\mu,\sigma) = 1 - \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\text{LFD}-\mu)^2}{2\sigma^2}\right)$, we applied $\mu = 0, \sigma = 0.5$. The random variable x represents the distance between the spectral vector of the reconstructed volume and ground truth, and the shape of these functions is visualized in Fig. 8. As depicted in Fig. 7(a), the highest segmentation Dice score of 84.17% was reached when using tanh scaling. The lowest Dice score of 83.72% was recorded when the SDSL was applied without any scaling.

2) MASKING RATIO

In the MIM framework, such methods as MAE [16], SimMIM [11], and A^2MIM [37], typically adopt a high masking ratio. However, in this work, a high masking ratio led to inferior downstream segmentation results. This

TABLE 4. Dice score comparison (%) for the assessed methods by organ in the beyond the cranial vault dataset.

Methods	Spleen	Rkid	Lkid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	Rag	Lag	AVG
UNETR [53]	94.23	93.81	93.71	60.50	70.65	95.91	82.06	88.80	82.48	70.09	78.14	66.55	61.09	79.85
SwinUNETR [52]	96.26	94.69	94.58	64.39	72.50	97.01	83.71	89.86	86.30	76.70	82.94	70.66	67.52	82.86
SSL-SwinUNETR [71]	96.48	95.01	94.43	63.31	77.42	97.04	86.32	90.62	85.91	75.37	83.54	71.38	68.34	83.47
SwinMM [20]	96.48	94.93	94.69	61.89	76.29	97.07	85.10	90.24	86.27	76.19	84.18	72.86	70.38	83.58
DAE [29]	96.41	94.93	94.75	65.79	75.68	97.10	85.57	91.02	86.19	75.72	85.17	71.07	68.02	83.65
Proposed	96.47	94.64	94.56	**66.23**	**78.02**	96.93	**89.10**	90.04	**87.36**	76.34	82.69	71.14	70.67	**84.17**

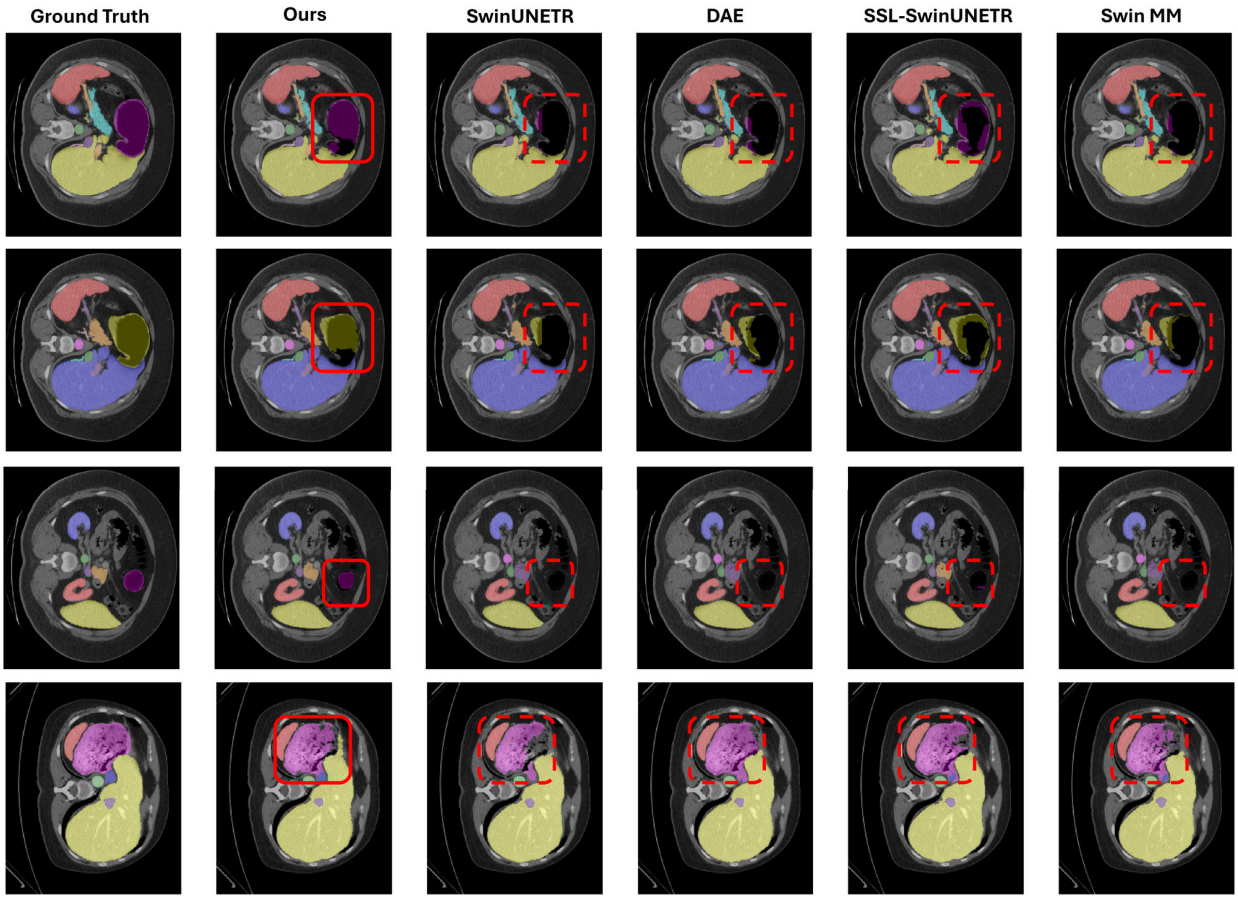


FIGURE 5. Qualitative comparison of segmentation results on the Beyond the Cranial Vault (BTCV) dataset between the proposed and existing methods. Red square regions demonstrate how the proposed model outperforms other methods.

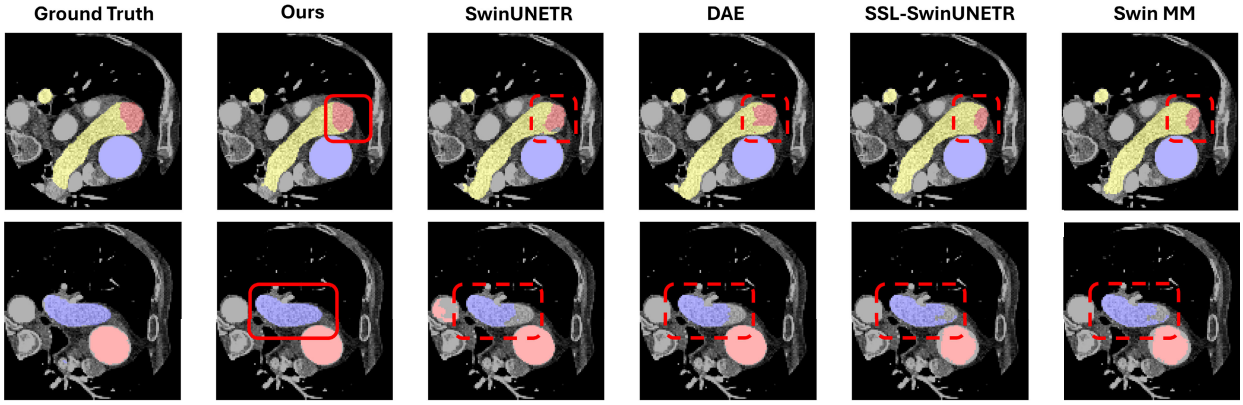


FIGURE 6. Qualitative comparison of the segmentation results on the MM-WHS dataset between the proposed and existing methods. Red square regions demonstrate how the proposed model outperforms other methods.

difference is attributed to the distinction between the previous 2D-based methods and the proposed 3D-based approach. Reconstructing a masked 3D volume is challenging, and a high masking ratio is assumed to affect downstream

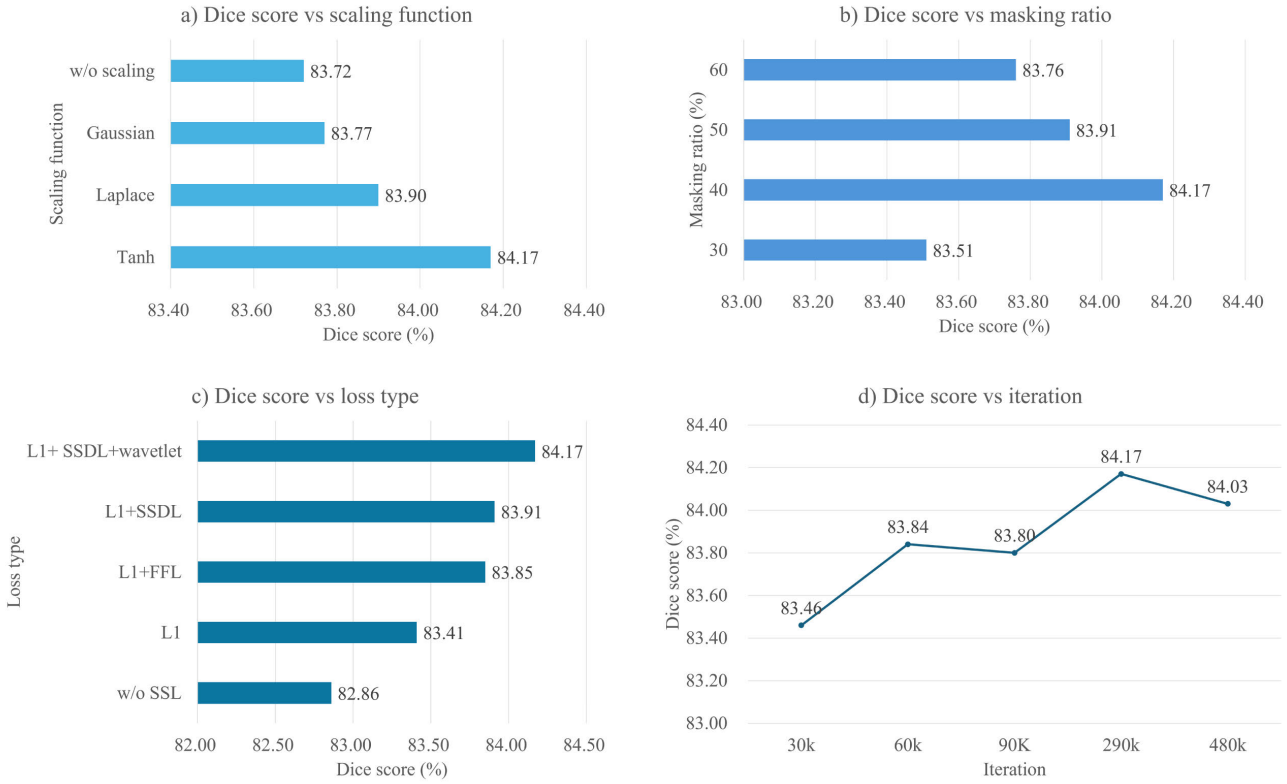


FIGURE 7. Ablation study results reported as the Dice score on the downstream segmentation on the BTCV dataset.

TABLE 5. Dice score comparison (%) on the spleen MSD dataset using five-fold cross-validation.

Spleen	Fold1	Fold2	Fold3	Fold4	Fold5	AVG
UNETR [53]	96.90	97.40	95.45	97.79	97.04	96.91
SwinUNETR [52]	97.69	98.43	97.42	98.68	97.87	98.01
SSL-SwinUNETR [71]	97.70	98.50	97.60	98.59	98.10	98.10
SwinMM [20]	97.98	97.98	98.53	97.55	98.78	98.16
DAE [29]	98.02	98.05	97.57	98.73	97.99	98.07
Proposed	98.00	98.44	97.69	98.77	98.11	**98.20**

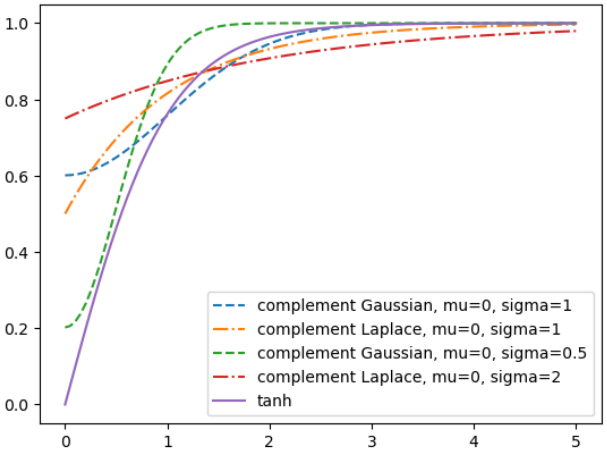


FIGURE 8. Scaling functions. Gaussian (- -) and Laplace (-.) functions have different σ parameters to control the scaling degree but apply the same mean $\mu = 0$ to ensure symmetry over the y-axis.

segmentation training adversely. The masking ratio comparison is presented in Fig. 7(b).

3) LOSS COMPARISON

We evaluated the performance of each component in the proposed frequency-domain training framework. Performance was assessed when the frequency loss was substituted with FFL. The experiment revealed that the Dice score was 0.32% higher with SDSL compared to FFL. Furthermore, when evaluating the performance improvement of each element, SDSL made the most significant contribution to the proposed framework. Notably, if the wavelet transform is not applied, the DFT is applied without separating the volume into low and high bands. Fig. 7(c) presents the loss type comparison.

To evaluate the ability to learn frequency information, we conducted an experiment to reconstruct unseen samples after pretraining and compared its frequency distance with the ground truth. Fig 9 illustrates the distribution of Log Frequency Distance (LFD) of the BTCV dataset using using the proposed framework versus FFL. Here, only the frequency diagonal components were extracted and visualized. LFD is calculated as in [30]:

$$\text{LFD} = \log\left(\sum_{u=0}^{N_x-1}\sum_{v=0}^{N_y-1}\sum_{w=0}^{N_z-1}\left(|\mathcal{F}_{\text{gt}}(u,v,w) - \mathcal{F}_{\text{pred}}(u,v,w)|^2 + 1\right)\right) \quad (7)$$

Fig. 9 shows that the mean distance is lower when using our framework compared to FFL. Both distributions exhibit a heavy-tailed Gaussian shape. However, when learning with

TABLE 6. Dice score comparison (%) for methods across anatomical structures in the MM-WHS dataset.

Methods	LV	W Aorta	R Ventricle	L Atrium	MLV	R Atrium	P Artery	AVG
UNETR [53]	88.80	88.03	93.07	80.85	84.82	92.21	75.77	86.22
SwinUNETR [52]	90.20	90.47	94.17	85.37	88.39	93.70	80.42	88.96
SSL-SwinUNETR [71]	90.11	89.60	94.20	86.07	87.40	96.00	81.41	89.25
SwinMM [20]	90.57	89.06	94.00	87.06	88.94	95.14	82.54	89.62
DAE [29]	90.00	89.91	93.84	87.05	89.55	95.39	81.38	89.59
Proposed	90.84	90.14	94.30	88.03	90.27	96.07	82.98	**90.38**

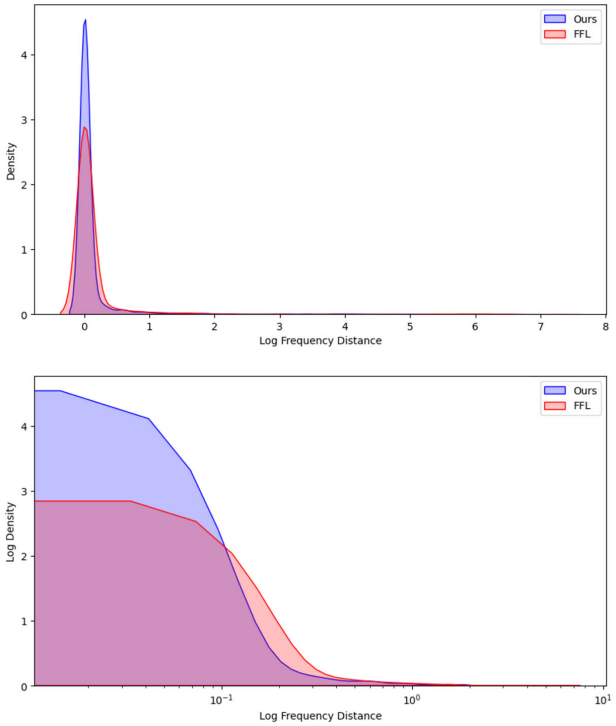


FIGURE 9. Log frequency distance comparsion: (top) kernel density estimation of the probability density function of the frequency distance, (bottom) horizontal log-scale conversion of (top).

our method, the distance from the ground truth is lower, forming a Gaussian distribution with a smaller standard deviation. One reason for this may be that the frequencies of various bands, scaled by SDSL during the pretraining process, contribute equitably to the loss function. Hence, the proposed method allows the model to capture anatomical details accurately by learning fine-grained features in the frequency domain alongside the coarse features in the spatial domain.

4) PRETRAIN ITERATION COMPARISON

Finally, we conducted experiments to optimize the number of pretraining iterations. During pretraining, validation was evaluated once per epoch. Each point on the x-axis in Fig. 7(d) represents a saved checkpoint iteration. Fine-tuning the SSL weights saved at each checkpoint caused the downstream segmentation performance to improve with increased pretraining iterations up to an extent. However, after 290K iterations, performance deteriorated even when weights with lower loss values were assigned. Fig. 7(d) illustrates the iteration comparison.

These results may stem from the disparity between the reconstruction task and representation learning. The MIM -based SSL method learns the image representation by inferring the unmasked patches. Reconciling with the findings of MAE [16], although the reconstruction results improved when random masking was replaced with uniform grid masking, the downstream task performance decreased. Improvements in reconstruction task results led to better representation learning but may not be equivalent due to the differences in the tasks.

V. CONCLUSION

This study introduced the SDSL as a novel objective for the self-supervised Swin UNETR to enhance 3D medical image segmentation. The experiments demonstrated that SDSL improves representation learning by incorporating frequency-domain information, resulting in superior Dice scores compared with recent approaches across several public datasets. A critical insight is that balanced frequency learning via the scaling function allows the model to capture fine-grained and coarse features, enhancing reconstruction and segmentation outcomes.

Despite these advancements, this study is not without limitations. The generalizability of SDSL across medical imaging modalities and datasets requires further investigation. Additionally, the computational cost of frequency-domain transformations and wavelet decompositions could hinder broader applications. Future work could address these challenges by incorporating imputation, contrastive learning, and expanding the approach to other medical imaging modalities.

ACKNOWLEDGMENT

(Jin Lee and Dang Thanh Vu contributed equally to this work.)

REFERENCES

[1] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[2] M. Antonelli, "The medical segmentation decathlon," *Nature Commun.*, vol. 13, no. 1, p. 4128, 2022.

[3] B. H. Menze, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[4] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Automatic multi-organ segmentation on abdominal CT with dense V-networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 8, pp. 1822–1834, Aug. 2018.

[5] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.*, vol. 36, pp. 61–78, Feb. 2017.

[6] K. Kamnitsas, L. Chen, C. Ledig, D. Rueckert, and B. Glocker, "Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI," *Ischemic Stroke Lesion Segmentation*, vol. 13, p. 46, Oct. 2015.

[7] J. Chen, L. Yang, Y. Zhang, M. Alber, and D. Z. Chen, "Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1–11.

[8] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 171–180.

[9] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3D deeply supervised network for automated segmentation of volumetric medical images," *Med. Image Anal.*, vol. 41, pp. 40–54, Oct. 2017.

[10] J. Dolz, C. Desrosiers, and I. Ben Ayed, "3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study," *NeuroImage*, vol. 170, pp. 456–470, Apr. 2018.

[11] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9643–9653.

[12] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "A unified view of masked image modeling," 2022, *arXiv:2210.10615*.

[13] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[17] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16133–16142.

[18] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

[19] Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma, "Explaining neural scaling laws," 2021, *arXiv:2102.06701*.

[20] Y. Wang, Z. Li, J. Mei, Z. Wei, L. Liu, C. Wang, S. Sang, A. L. Yuille, C. Xie, and Y. Zhou, "SwinMM: Masked multi-view with Swin transformers for 3D medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2023, pp. 486–496.

[21] Z. Chen, D. Agarwal, K. Aggarwal, W. Safta, M. M. Balan, and K. Brown, "Masked image modeling advances 3D medical image analysis," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1969–1979.

[22] A. Hatamizadeh, Z. Xu, D. Yang, W. Li, H. Roth, and D. Xu, "UNetFormer: A unified vision transformer model and pre-training framework for 3D medical image segmentation," 2022, *arXiv:2204.00631*.

[23] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna, "Self pre-training with masked autoencoders for medical image classification and segmentation," in *Proc. IEEE 20th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2023, pp. 1–6.

[24] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4084–4094.

[25] W. Wang, J. Wang, C. Chen, J. Jiao, Y. Cai, S. Song, and J. Li, "FreMIM: Fourier transform meets masked image modeling for medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 7860–7870.

[26] J. Xie, W. Li, X. Zhan, Z. Liu, Y. S. Ong, and C. C. Loy, "Masked frequency modeling for self-supervised visual pre-training," 2022, *arXiv:2206.07706*.

[27] H. Liu, X. Jiang, X. Li, A. Guo, Y. Hu, D. Jiang, and B. Ren, "The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 1649–1656.

[28] Y. Liu, S. Zhang, J. Chen, K. Chen, and D. Lin, "PixMIM: Rethinking pixel reconstruction in masked image modeling," 2023, *arXiv:2303.02416*.

[29] J. M. J. Valanarasu, Y. Tang, D. Yang, Z. Xu, C. Zhao, W. Li, V. M. Patel, B. Landman, D. Xu, Y. He, and V. Nath, "Disruptive autoencoders: Leveraging low-level features for 3D medical image pre-training," 2023, *arXiv:2307.16896*.

[30] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13899–13909.

[31] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5301–5310.

[32] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7537–7547.

[33] Z.-Q. J. Xu, Y. Zhang, and Y. Xiao, "Training behavior of deep neural network in frequency domain," in *Proc. Int. Conf. Neural Inf. Process.*, 2019, pp. 264–274.

[34] Z.-Q. John Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," 2019, *arXiv:1901.06523*.

[35] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8681–8691.

[36] A. Dziedzic, J. Paparrizos, S. Krishnan, A. Elmore, and M. Franklin, "Band-limited training and inference for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 1745–1754.

[37] A. A. Abello, R. Hirata, and Z. Wang, "Dissecting the high-frequency bias in convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 863–871.

[38] J. Bai, L. Yuan, S.-T. Xia, S. Yan, Z. Li, and W. Liu, "Improving vision transformers by revisiting high-frequency components," in *Proc. 17th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 1–18.

[39] M. Woo Kim and N. Ik Cho, "WHFL: Wavelet-domain high frequency loss for Sketch-to-Image translation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 744–754.

[40] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1737–1746.

[41] S. Li, D. Wu, F. Wu, Z. Zang, and S. Z. Li, "Architecture-agnostic masked image modeling—From ViT back to CNN," 2022, *arXiv:2205.13943*.

[42] M. Cai, H. Zhang, H. Huang, Q. Geng, Y. Li, and G. Huang, "Frequency domain image translation: More photo-realistic, better identity-preserving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13910–13920.

[43] L. Shan, X. Li, and W. Wang, "Decouple the high-frequency and low-frequency information of images for semantic segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1805–1809.

[44] R. Azad, I. A. Bozorgpour, M. Asadi-Aghbolaghi, D. Merhof, and S. Escalera, "Deep frequency re-calibration U-Net for medical image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3267–3276.

[45] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7887–7896.

[46] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot…for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8692–8701.

[47] X. Zhuang, "Multivariate mixture model for myocardial segmentation combining multi-source images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2933–2946, Dec. 2019.

[48] B. Rister, D. Yi, K. Shivakumar, T. Nobashi, and D. L. Rubin, "CT-ORG, a new dataset for multiple organ segmentation in computed tomography," *Sci. Data*, vol. 7, no. 1, p. 381, Nov. 2020.

[49] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[50] S Suganyadevi, V Seethalakshmi, and K Balasamy, "A review on deep learning in medical image analysis," *Int. J. Multimedia Inf. Retr.*, vol. 11, no. 1, pp. 19–38, 2022.

[51] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[52] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Proc. Int. MICCAI Brainlesion Workshop*, 2021, pp. 272–284.

[53] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1748–1758.

[54] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[55] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[56] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[57] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Athens, Greece. Cham, Switzerland: Springer, 2016, pp. 424–432.

[58] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[59] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[60] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[61] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[62] M. Khened, V. A. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers," *Med. Image Anal.*, vol. 51, pp. 21–45, Jan. 2019.

[63] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers, "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *Med. Image Anal.*, vol. 45, pp. 94–107, Apr. 2018.

[64] H. R. Roth, H. Oda, Y. Hayashi, M. Oda, N. Shimizu, M. Fujiwara, K. Misawa, and K. Mori, "Hierarchical 3D fully convolutional networks for multi-organ segmentation," 2017, *arXiv:1704.06382*.

[65] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 205–218.

[66] Y. He, V. Nath, D. Yang, Y. Tang, A. Myronenko, and D. Xu, "SwinUNETR-V2: Stronger Swin transformers with stagewise convolutions for 3D medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2023, pp. 416–426.

[67] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual Swin transformer U-Net for medical image segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[68] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

[69] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. 24th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 109–119.

[70] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2021, pp. 36–46.

[71] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of Swin transformers for 3D medical image analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20698–20708.

[72] P. Zhai, H. Cong, E. Zhu, G. Zhao, Y. Yu, and J. Li, "MVCNet: Multiview contrastive network for unsupervised representation learning for 3-D CT lesions," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7376–7390, Jun. 2024.

[73] Y. Xia, D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, and H. Roth, "Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101766.

[74] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, "3D self-supervised methods for medical imaging," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18158–18172.

[75] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," *Int. J. Comput. Vis.*, vol. 132, no. 1, pp. 208–223, Jan. 2024.

[76] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A Fourier-based framework for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14378–14387.

[77] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9620–9629.

[78] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[79] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 776–794.

[80] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[81] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[82] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1422–1430.

[83] M. Assran, "Masked Siamese networks for label-efficient learning," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*. Tel Aviv, Israel: Cham, Switzerland: Springer, Oct. 2022, pp. 456–473.

[84] J. Zhu, Y. Li, Y. Hu, K. Ma, S. K. Zhou, and Y. Zheng, "Rubik's Cube+: A self-supervised feature learning framework for 3D medical image analysis," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101746.

[85] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, "Self-supervised feature learning for 3D medical images by playing a Rubik's cube," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2019, pp. 420–428.

[86] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101539.

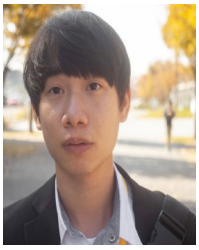
[87] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[88] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[89] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 694–711.

[90] S. Schreiber, J. Geldenhuys, and H. de Villiers, "Texture synthesis using convolutional neural networks with long-range consistency and spectral constraints," in *Proc. Pattern Recognit. Assoc. South Afr. Robot. Mechatronics Int. Conf. (PRASA-RobMech)*, Nov. 2016, pp. 1–6.
[91] Y. Yu, F. Zhan, S. Lu, J. Pan, F. Ma, X. Xie, and C. Miao, "WaveFill: A wavelet-based generation network for image inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14094–14103.
[92] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-SRNet: A wavelet-based CNN for multi-scale face super resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1698–1706.
[93] S. A. Magid, Y. Zhang, D. Wei, W.-D. Jang, Z. Lin, Y. Fu, and H. Pfister, "Dynamic high-pass filtering and multi-spectral attention for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4268–4277.
[94] X. Lin, Y. Li, J. Hsiao, C. Ho, and Y. Kong, "Catch missing details: Image reconstruction with frequency augmented variational autoencoder," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1736–1745.
[95] Z. Lu, J. Jiang, J. Huang, G. Wu, and X. Liu, "GLaMa: Joint spatial and frequency loss for general image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1300–1309.
[96] S. Jung and M. Keuper, "Spectral distribution aware image generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 1734–1742.
[97] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2830–2838.
[98] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, S. Cao, Q. Zhang, S. Liu, Y. Wang, Y. Li, J. He, and X. Yang, "AbdomenCT-1K: Is abdominal organ segmentation a solved problem?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6695–6714, Oct. 2022.
[99] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013.

JIN LEE (Graduate Student Member, IEEE) received the B.S. degree in electronics engineering from Chonnam National University, South Korea, in 2022, where he is currently pursuing the M.S. degree with the Department of Intelligent Electronics and Computer Engineering. His research interests include medical image analysis, semantic segmentation, image classification, and image recognition.

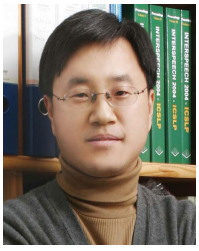


DANG THANH VU received the B.S. degree in mathematics and computer science from Ho Chi Minh University of Science, Vietnam, in 2019, and the Ph.D. degree in ICT convergence engineering systems from Chonnam National University, South Korea, in 2024. Currently, he is working as a Researcher with AISeed Inc., South Korea. His research interests include deep learning, computer vision, and capsule networks.

GWANGHYUN YU received the M.S. and Ph.D. degrees in ICT convergence engineering systems from Chonnam National University, South Korea, in 2017 and 2023, respectively. He currently working as an Artifical Intelligence Researcher with Chonnam National University. His research interests include deep learning, computer vision, smart farm the Internet of Things, and medical imagery.

JINSUL KIM (Member, IEEE) received the B.S. degree in computer science from the University of Utah, Salt Lake City, Utah, USA, in 1998, and the M.S. and Ph.D. degrees in digital media engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2004 and 2008, respectively. Previously, he worked as a Researcher at the Broadcasting/Telecommunications Convergence Research Division, Electronics and Telecommunications Research Institute (ETRI), Daejeon, from 2004 to 2009, and was a Professor at Korea Nazarene University, Cheonan, South Korea, from 2009 to 2011. He is a currently working as a Professor with Chonnam National University, Gwangju, South Korea. He is also a Co-Research Director of the Artificial Intelligence Innovation Hub Research and Development Project hosted by Korea University and the Director of the G5-AICT Research Center. He has participated in various national research projects and domestic and international standardization activities. He is a member of Korean National Delegation for ITU-T SG13 International Standardization.

KUNYUNG KIM received the M.D., M.S., and Ph.D. degrees from Jeonbuk National University, South Korea, in 2008, 2014, and 2018, respectively. He is currently working as an Associate Professor with the Department of Radiology, Seoul National University Bundang Hospital, South Korea. His research interests include deep learning applications for interventional radiology and translational research.



JINYOUNG KIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in engineering from Seoul National University, Seoul, South Korea, in 1986, 1988, and 1994, respectively. Since 1995, he has been a Professor with the Department of Intelligent Electronic and Computer Engineering, Chonnam National University, South Korea.