FreMAE

Wang, Wenxuan; Wang, Jing; Chen, Chen; Jiao, Jianbo; Sun, Lichao; Cai, Yuanxiu; Song, Shanshan; Li, Jiangyun

DOI:
10.48550/arXiv.2304.10864

FreMAE: Fourier Transform Meets Masked Autoencoders for Medical Image Segmentation

Wenxuan Wang[1,*] Jing Wang[1,*] Chen Chen[2] Jianbo Jiao[3]
Lichao Sun[4] Yuanxiu Cai[1] Shanshan Song[1] Jiangyun Li[1†]
[1]School of Automation and Electrical Engineering, University of Science and Technology Beijing
[2]Center for Research in Computer Vision, University of Central Florida
[3]University of Birmingham
[4]Lehigh University
{s20200579,m202120718}@xs.ustb.edu.cn, chen.chen@crcv.ucf.edu
jiaojianbo.i@gmail.com,leejy@ustb.edu.cn

Abstract

The research community has witnessed the powerful potential of self-supervised Masked Image Modeling (MIM), which enables the models capable of learning visual representation from unlabeled data. In this paper, to incorporate both the crucial global structural information and local details for dense prediction tasks, we alter the perspective to the frequency domain and present a new MIM-based framework named FreMAE for self-supervised pre-training for medical image segmentation. Based on the observations that the detailed structural information mainly lies in the high-frequency components and the high-level semantics are abundant in the low-frequency counterparts, we further incorporate multi-stage supervision to guide the representation learning during the pre-training phase. Extensive experiments on three benchmark datasets show the superior advantage of our proposed FreMAE over previous state-of-the-art MIM methods. Compared with various baselines trained from scratch, our FreMAE could consistently bring considerable improvements to the model performance. To the best our knowledge, this is the first attempt towards MIM with Fourier Transform in medical image segmentation.

1. Introduction

Since Masked Language Modeling (MLM) obtained great success in the field of Natural Language Processing (NLP) [18], numerous works [25, 50, 40, 4, 49, 12] have transferred this idea to the vision domain, making Mask Image Modeling (MIM) an effective pre-training strategy. One of the most representative approaches is Masked Autoencoders (MAE) [25], which pre-trains the model by masking partial regions within an image and reconstructing them. After the pre-training, the model is fine-tuned on various downstream tasks and achieves state-of-the-art (SOTA) performance. Following-up works mainly focus on improving the accuracy and efficiency by introducing new designs, such as ConvMAE [23] and Siamese Image Modeling [45].

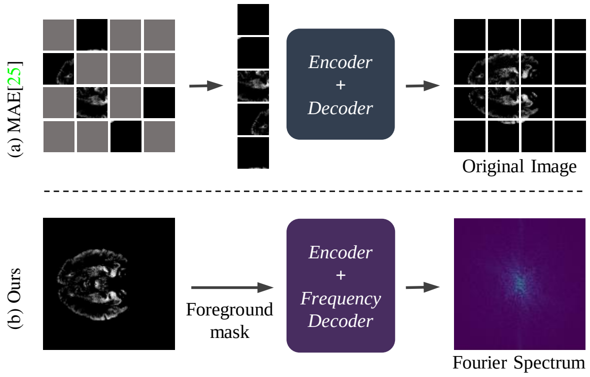

Figure 1. The comparison of key ideas between MAE frameworks and our proposed FreMAE. (a) MAE: randomly masks the patch tokens and reconstruct raw pixels of original image. (b) Our FreMAE: randomly masks the **foreground pixels** and reconstructs the **Fourier spectrum** of original image.

Aiming to propagate the success of MAE, some recent works applied MAE-based methods for medical image analysis [51, 44, 26] and achieved promising results across various benchmark datasets with different modalities, including computed tomography images (CT) [36], magnetic resonance imaging (MRI) [27], to name a few. Despite making methodological advancements and structural innovations, these methods have not essentially solved the key limitations of MAE. Although compared with other self-

[*] Equal Contribution. [†] Corresponding author.

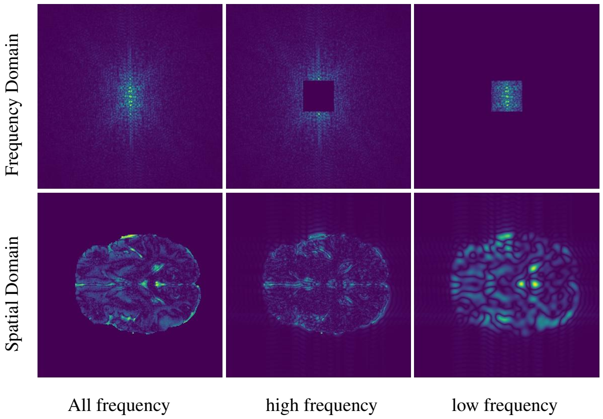

Figure 2. The visualization of the whole Fourier spectrum, high-frequency components, and low-frequency counterparts respectively, the high/low-frequency components of which are acquired by applying the corresponding high/low-pass filters on the whole Fourier spectrum. The inspiration for our FreMAE comes from the observations that local details (like texture and contours) mainly lie in the high-frequency components while the global and smooth structural information is rich in the low-frequency counterparts.

supervised learning frameworks MAE can consistently help the model extract generally useful features even with few training samples (as proven by [30]), to some extent, MAE solely takes raw pixels as reconstruction targets mainly depending on local feature representation rather than fully utilizing the global information. Besides, since the model is expected to possess the ability of extracting features with multiple semantic levels at different stages, only the output from the last stage is fed into the decoder for the reconstruction task, lacking the supervision from other stages to provide multi-scale information. Furthermore, due to high acquisition costs and patients' privacy, the training samples of commonly small-scale medical image datasets are relatively limited, but none of these previous works have taken this unique characteristic of medical image datasets into consideration and made tailored designs. In summary, previous works [25, 4, 48] crucially require a certain trade-off between the local details and contextual semantics.

Therefore, in order to fully exploit the potential of MAE-based methods for medical image segmentation under the circumstance of limited training samples, *how to acquire the global information while preserving the detailed local features as much as possible* has become the key problem. Considering the nature of Fourier Transform in image processing, it might be a possible solution. As studied in lots of previous works [43, 14, 7, 5, 29] and shown in Fig. 2, the detailed texture information mainly lies in the high-frequency components and the low-frequency counterparts carry rich global information. Following this observation, an intuitive solution would be creatively exploring the powerful potential of MIM coupled with Fourier Transform.

To this end, aiming at the circumstance of limited training samples in medical image analysis, we propose a new MIM-based framework conducted in the Fourier domain, namely *FreMAE*, which to our knowledge is the first work to explore the potential of MIM with Fourier Transform for 2D medical image segmentation. Specifically, our FreMAE first masks out a portion of randomly selected image pixels and then predicts the corresponding missing frequency spectrum of the input image in the Fourier domain. Since medical images of the same organ essentially correspond to similar features, we conduct difficult cross-domain reconstruction tasks to avoid model learning with shortcuts and achieve strong feature representation capability. Meanwhile, inspired by the previous findings [47] that the detailed structural information mainly lies in the high-frequency components and the high-level semantics are abundant in the low-frequency counterparts, the proposed bilateral aggregation decoder is leveraged to sequentially apply the Fourier Transform on the original image and low/high-pass filters on the converted Fourier spectrum to get the expected reconstruction target. Such a multi-stage supervision approach could better guide the model pre-training, resulting in better representations for segmentation. Besides, we propose an effective foreground masking strategy as the alternative to the original random masking, which is proven to be more suitable for textures and details modeling for medical image segmentation. In summary, the main contributions of this work are as follows:

- We present the first study on exploring the powerful potential of masked image modeling with frequency domain for medical image segmentation. The proposed FreMAE is a generic self-supervised pre-training framework that can be integrated with different model architectures (*e.g.* both CNNs and Transformers).
- By designing a multi-stage supervision scheme coupled with a well-designed bilateral decoder, we propose a new cross-domain mask-reconstruction framework for masked image modeling.
- A simple yet effective masking strategy among foreground pixels is proposed as a better alternative to the original random masking pixels strategy, providing more precise and informative masks for the following self-supervised representation learning.
- Without introducing extra training samples, extensive experiments on three benchmark datasets and three representative 2D baselines prove the superiority of the proposed FreMAE, outperforming other alternative self-supervised SOTA approaches.

2. Related Work

2.1. Masked Image Modeling

As a powerful self-supervised learning paradigm, MIM has attracted increasing interests recently. By reconstructing the masked portion of images, models could learn informative feature representations that are favorable for various downstream tasks.

On Natural Images. Previous works of reconstruction targets could be divided into three categories, including discrete tokens[4, 40], feature maps[48, 50], and raw image pixels[49, 25]. For example, BEiT [4] and BEiTV2 [40] added a classifier to predict masked visual tokens, and it is supervised by the encoded image patches from offline tokenizer. Inspired by the self-distillation paradigm in DINO[9], iBOT [50] adopted a teacher-student framework to perform MIM. The teacher network serves as an online tokenizer to learn visual semantics from all image patches, while the student network, only processes visible patches. Moreover, MaskFeat [48] first explored features as prediction targets. Besides, SimMIM [49] discarded the tokenizer and patch classification, simply employing RGB values of raw pixels as predicted targets. Without masked tokens feeding into encoder, MAE [25] designed a simple decoder to reconstruct image patches, leading to a considerable reduction of computation complexity during pre-training.

On Medical Images. At the same time, various works [51, 44, 26] have explored the effectiveness of MIM pre-training on various medical benchmark datasets. Zhou et al.[51] applied MAE pre-training paradigm for medical image segmentation and significantly improved the results. Huang et al.[26] proposed a manually settled attentive reconstruction loss that pays more attention to the informative regions. Tang et al.[44] explored the hierarchical structure for full extraction of image features and constructed a self-supervised pre-training framework with three proxy tasks. However, random masking strategy of patches utilized previously are rough and may result in computation waste on the useless background. Considering that informative foreground and useless background are discriminate in medical images, we design a masking strategy among foreground pixels to obtain more effective masks, assisting models in better representation learning. Moreover, our method could cast off the reliance of the pre-training paradigm on specific model structures, which is different from previous works (e.g. Swin Transformer and CNN-based models can not be directly integrated with MAE).

2.2. Fourier Transform

Recently, a series of research[41, 52, 28] have performed Fourier Transform on images and leveraged the frequency information to improve model performance and efficiency. For example, [41] utilized Fast Fourier Transform (FFT) as the alternative of self-attention modules in the original Transformer, successfully acquiring global information with low computation costs. [28] designed a novel focal frequency loss for Fourier spectrum supervision to improve popular image generative model performance.

Inspired by these previous researches [43, 14, 7, 5, 29], we randomly mask the original image and reconstruct the Fourier spectrum in the frequency domain to help the model learn more generalized global representation. In addition, multi-stage supervision coupled with leveraged specific characteristics of FFT (i.e. high-pass and low-pass frequency components) is also proposed to better guide the model representation learning among different stages.

3. Methodology

In this section, we first briefly review some key preliminary knowledge for the ease of understanding our proposed method. Then, the overall architecture of our FreMAE is introduced, followed by the detailed elaboration of each component, *i.e.* masking strategy, multi-stage supervision scheme, bilateral aggregation decoder, and reconstruction target. At last, the specific pre-training strategy is presented.

3.1. Preliminaries: Fourier Transform

Since Discrete Fourier Transform (DFT) plays a vital role in our proposed methodWe first give a brief review of the 2D DFT that serves as an indispensable technique for traditional signal analysis. Given a 2D signal $\mathbf{F} \in \mathbb{R}^{W\times H}$, its corresponding 2D-DFT can be defined as:

$$f(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} F(h,w)e^{-j2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)}, \tag{1}$$

where $F(h,w)$ represents the signal located at (h,w) in $\mathbf{F}$, while the u and v are indices of horizontal and vertical spatial frequencies in Fourier spectrum. Correspondingly, the 2D Inverse DFT (2D-IDFT) is formulated as:

$$F(h,w) = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1} f(u,v)e^{j2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)}. \tag{2}$$

Both DFT and IDFT can be accelerated with their fast version, FFT algorithm [37]. For medical images with various modalities, the Fourier Transform is operated on each channel independently. Besides, as already shown in previous works [43, 14, 7, 5, 29], the detailed structural texture information of an image mainly lies in the high-frequency part of the Fourier spectrum while the global information is rich in the low-frequency counterpart. Fig. 2 presents the visualization of this intriguing characteristic.

3.2. The Proposed FreMAE

Overall Architecture. An overview of the proposed self-supervised learning framework FreMAE is presented in Fig. 3. Given an input medical image slice $X \in \mathbb{R}^{C\times H\times W}$ with a spatial resolution of $H \times W$ and C channels (# of modalities), the proposed foreground masking strategy is first employed on the original image slice to generate the

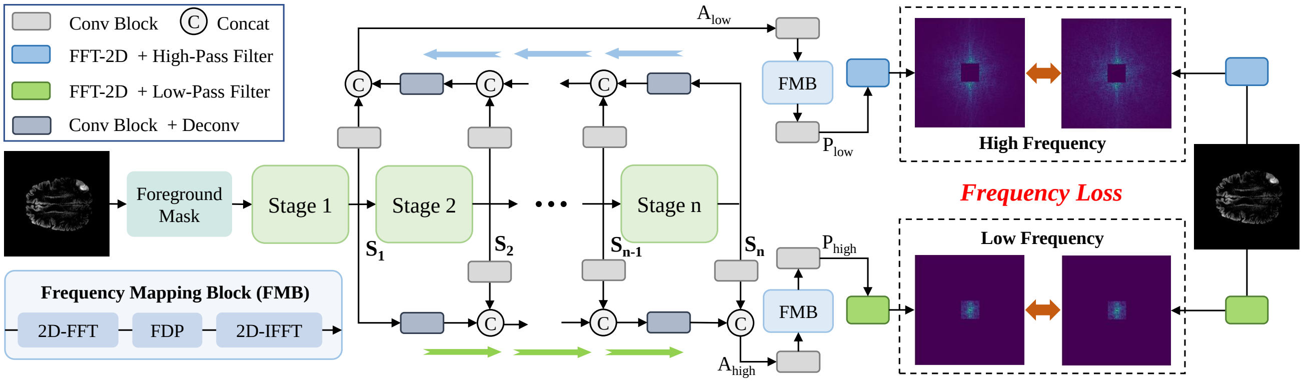

Figure 3. The overall architecture of our proposed FreMAE. At first, the input medical image is corrupted by the foreground masking strategy and then fed into the encoder, which consists of several stages with a hierarchical structure. The captured feature maps at different stages (*i.e.* S_1, S_2, ... S_n) are fused by a bilateral aggregation decoder to generate the aggregated high-level and low-level feature representations (*i.e.* A_{high} and A_{low}). For the fused feature of each semantic level, an FMB is applied respectively to learn its recessive information in the frequency domain, resulting in the acquired P_{low} and P_{high}. Finally, the low-pass and high-pass Fourier spectrum are both adopted as the reconstruction target to better guide the model to capture local details and global information. Note that only the encoder part is retained after the fine-tuning phase.

masked image. Then, the generic encoder (*i.e.* according to different pre-training requirements, both CNNs and Transformers encoder can be integrated into our framework) takes the masked image as input, capturing the masked visual features through the hierarchical structure. After that, the encoded feature representations at different stages are jointly fed into our well-designed bilateral aggregation decoder, gradually producing the reconstructed Fourier spectrum with both low-level detail information and high-level semantic representation. By sequentially applying Fourier Transform on the original image and low/high-pass filters on the converted Fourier spectrum to acquire the expected reconstruction target, the reconstruction loss (*e.g.* l_1 loss) is applied on the similarity between the reconstructed Fourier spectrum and expected low/high-pass spectrum target, realizing the helpful multi-stage supervision scheme with both low-level and high-level feature representations in an end-to-end manner.

Masking Strategy. As experimentally illustrated in several previous works[4, 40, 49, 25, 23, 45], random masking strategy is not only simple but also effective for MIM-based self-supervised learning paradigm on large-scale natural images. However, different from natural images, the distribution of foreground and background pixels in medical images is extremely unbalanced. So randomly selecting spatial positions of a medical image would inevitably cause the generated mask to mostly cover background pixels and too many foreground pixels of the objects are reserved, counting against the model's reconstruction ability. To this end, we propose a simple yet effective foreground masking strategy to address this uneven distribution issue.

Specifically, given a binary mask $M \in \{0,1\}^{H\times W}$ initialized with zeros, its value at each spatial position is determined by whether the corresponding pixel value belongs to the foreground or not. If a pixel belongs to the foreground area, it will be filtered as one of the candidates to be masked during self-supervised pre-training. Since a medical image commonly consists of diverse channels, each one emphasizing a different foreground area, we take the overlapping parts as the final masked regions. The overall foreground masking strategy can be defined as:

$$M_n(x,y) = \begin{cases} 0, & P_n(x,y) = 0 \\ 1, & P_n(x,y) \neq 0 \end{cases}, \tag{3}$$

$$\mathcal{M} = M_1 \cap M_2 \cap M_3 ... \cap M_n, \tag{4}$$

$$X_{\mathcal{M}} = \mathcal{M} \odot X, \tag{5}$$

where $\odot$ is the Hadamard product, $P_n(x,y)$ represents the specific pixel value of the corresponding position (x,y), M_n denotes the generated mask of the specific image modality M_n. $\mathcal{M}$ and $X_{\mathcal{M}}$ respectively indicate the final mask of the original image and the masked image that will be fed into the model for the following reconstruction task.

Generic Eencoder. As for the selection of encoder in our framework, FreMAE is not restricted to any specific kind of structure thanks to our pixel-wise foreground masking strategy. Dislike some previous MIM-based methods can only be incorporated with various Vision Transformers (*e.g.* Due to the random masking strategy of embedded image patches, MAE is only applicable for ViT[19] without the consideration of CNNs or hierarchical Transformer architecture), our FreMAE is a generic and flexible framework, which means both CNN-based and Transformer-based models can be easily integrated with our FreMAE for effective self-supervised pre-training. Taking the aforementioned masked image as input, the network encoder gradually encodes the masked image slice with the hierarchical structure, producing the feature representations with diverse levels (*i.e.* from low-level detail information to high-level semantics). In this paper, three previous SOTA methods for medical image segmentation, *i.e.* the representatives of the CNN-Transformer hybrid architectures and Vision Transformers, are selected as the backbones to validate the effectiveness of our method (more details are in Sec. 4).

Multi-stage Supervision Scheme. Both low-level details and high-level global semantics are crucial, especially for medical image segmentation. The expectation of an effective self-supervised learning paradigm is to guide the visual backbone to learn the required feature representations with different levels through the hierarchical structure. Following this intuition, we propose to design a multi-stage supervision scheme to fully supervise the representation learning of hierarchical stages.

As emphasized in Sec. 1, high-level and low-level information of an image distribute in different frequency bands of the Fourier spectrum. So we propose to separately take advantage of the low-pass and high-pass Fourier spectrum as the supervision signal (*i.e.* **reconstruction target**). One of the most intuitive ways is to utilize the identical high-pass Fourier spectrum to directly supervise multiple low-level stages and vice versa for low-pass counterparts. However, there are mainly two drawbacks for this intuitive manner. On the one hand, this manner is kind of unreasonable and it violates the original intention of model learning at various low-level stages cause the feature representations learned at different low-level stages should be naturally different instead of the same. On the other hand, such a supervision method is too direct and simple, and does not make full use of the correlation between the captured multi-stage features by the hierarchical structure to help the model better perform the MIM pretext task.

With regard to this, a well-designed **bilateral aggregation decoder** is proposed to better solve the reconstruction task in the frequency domain, further helping the encoder to learn a more generalized and more meaningful feature representation. Specifically, inside the proposed bilateral aggregation decoder, the encoded features at different stages are converged to the lowest stage (*i.e.* with maximum spatial resolution) and the highest stage (*i.e.* with minimum spatial resolution) in a bottom-up and top-down manner, respectively. In other words, the bilateral aggregation decoder separately aggregates the feature maps of different stages to the lowest and highest resolution. Specifically, for ViT, the feature maps of layers 4th, 8th, and 12th are upsampled by 8, 4, and 2 times respectively to be fed to the BAD, following the deconvolution module in UNETR. To be clear, the captured features of each adjacent stage will be fed into the convolutional block to achieve the strict alignment of both spatial resolution and channel dimension, which can be expressed as:

$$\mathbf{A}_{\text{low}} = \mathbf{Cat}(\mathbf{C}(S_1), \mathbf{Dc}(..., \mathbf{Cat}(\mathbf{C}(S_{n-1}), \mathbf{Dc}(S_n)))), \quad (6)$$

$$\mathbf{A}_{\text{high}} = \mathbf{Cat}(\mathbf{C}(S_n), \mathbf{Dc}(..., \mathbf{Cat}(\mathbf{C}(S_2), \mathbf{Dc}(S_1)))), \quad (7)$$

where $\mathbf{A}_{\text{high}}$ and $\mathbf{A}_{\text{low}}$ separately denote the bilaterally aggregated high-level and low-level feature representations, $\mathbf{C}$, $\mathbf{Dc}$ and $\mathbf{Cat}$ indicate the convolutional block, deconvolution block, and concatenation respectively, S_i denotes the feature maps output by the stage i.

Then, the aggregated feature representations at the lowest stage and highest stage will be mapped to the frequency domain through the introduced frequency mapping block (as illustrated in Fig. 3), which are followed by the low-pass and high-pass filters to get the corresponding high-pass and low-pass prediction spectrum for the employed reconstruction loss. Specifically, the frequency mapping block (FMB) consists of a 2D-DFT, a Frequency Domain Perceptron (FDP), and a 2D-IDFT, which can be calculated as:

$$\mathbf{P}_{\text{low}} = IDFT(W \odot DFT(\mathbf{A}_{\text{low}}) + b), \quad (8)$$

$$\mathbf{P}_{\text{high}} = IDFT(W \odot DFT(\mathbf{A}_{\text{high}}) + b), \quad (9)$$

where DFT and $IDFT$ represent the Fast Fourier Transform and Inverse Fast Fourier Transform. W and b are both learnable parameters, $\odot$ is the Hadamard product. In this way, a powerful self-supervised framework for cross-domain reconstruction is built with the benefit of the Fourier Transform's unique characteristics. Although such a cross-domain reconstruction task is more difficult than intra-domain reconstruction, it can also assist the model to learn more robust feature representation, which is fully demonstrated in the following experimental section.

3.3. Pre-training Strategy

Frequency Loss. To alleviate the weight imbalance between different frequency bands spectrums and facilitate the reconstruction of difficult frequency bands, we adopt focal frequency loss [28] as the loss function $\mathcal{L}_{\text{freq}}$ to implement gradient updating of weights for both low and high-frequency mapping, which is defined as:

$$\mathcal{L}_{\text{freq}} = \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1} \omega(u,v) \odot \gamma(f(u,v), \hat{f}(u,v))^2, \quad (10)$$

where $f(u,v)$ is the predicted 2D-DFT of spatial frequency coordinate (u,v) while $\hat{f}(u,v)$ denotes its corresponding Ground Truth value. $\gamma(f,\hat{f})$ calculates the squared Euclidean distance between actual and predicted values as their frequency distance. And ω is the spectrum weight matrix of a given location, which suppresses weights of easy frequencies. The specific calculation formulas are as follows:

$$\omega(u,v) = \gamma(f(u,v), \hat{f}(u,v))^{\beta}, \quad (11)$$

$$\gamma(f,\hat{f}) = \sqrt{(\mathcal{R}-\tilde{\mathcal{R}})^2 + (\mathcal{I}-\tilde{\mathcal{I}})^2}, \quad (12)$$

β is the scaling factor for flexibility (β = 1 in default) .

Overall Loss. During pre-training, our FreMAE learns representation by solving content gestalt from both high-pass and low-pass frequency:

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_{\text{freq}}(\mathbf{F}_H(\mathbf{P}_{\text{low}}), \mathbf{F}_H(\mathbf{T})) \\ &+\alpha\mathcal{L}_{\text{freq}}(\mathbf{F}_L(\mathbf{P}_{\text{high}}), \mathbf{F}_L(\mathbf{T})),\end{aligned} \tag{13}$$

where $\mathbf{F}_H$ and $\mathbf{F}_L$ represent high-pass and low-pass frequency filter respectively. $\mathbf{T}$ indicates the original images. As shown in Fig. 3, $\mathbf{P}_{\text{low}}$ is obtained by highest-stage and $\mathbf{P}_{\text{high}}$ is the opposite. α is the weight of high-level semantic information branches (α = 3 in default).

4. Experiments and Results

In this section, focusing on solely exploiting the given training samples (*i.e.* without introducing any extra data) for 2D medical image segmentation, extensive experiments on three benchmark datasets are conducted to fully verify the effectiveness of FreMAE.

4.1. Experimental Setup

Data and Evaluation Metrics. Our proposed method is evaluated on three benchmark datasets for medical segmentation. The Brain Tumor Segmentation 2019 challenge (**BraTS 2019**) dataset [34, 2, 3] is composed of multi-institutional pre-operative MRI sequences, including 335 patient cases for training and 125 cases for validation. Each sample contains four modalities (FLAIR, T1, T1c, T2) with the size of $240 \times 240 \times 155$, and the corresponding ground truth consists of 4 classes: background (label 0), necrotic and non-enhancing tumor (label 1), peritumoral edema (label 2) and GD-enhancing tumor (label 4). The Dice score and the Hausdorff distance (95%) metrics are used for evaluating the segmentation accuracy of different regions, including enhancing tumor region (ET, label 4), regions of the tumor core (TC, labels 1 and 4), and the whole tumor region (WT, labels 1,2 and 4). The International Skin Imaging Collaboration 2018 (**ISIC 2018**) dataset [46, 16] is a collection of 2594 RGB images of skin lesion for training, around 100 samples for validation, and 1000 samples for testing. Five metrics are specifically employed for the quantitative assessment of model performance, including Dice, Jaccard Index (JI), Accuracy, Recall, and Precision. The Automated Cardiac Diagnosis Challenge 2017 (**ACDC 2017**) dataset [6] is collected from different patient cases using MRI scanners, including 3D cardiac MRI cine for both end-diastolic (ED) and end-systolic (ES) phases instances. The publicly available training dataset consists of 100 patient scans, which are split into 80 training samples and 20 testing samples. The ground truth contains 3 classes: right ventricle (RV), myocardium (Myo) and left ventricle (LV).

Implementation Details. The proposed method is implemented in PyTorch [39] and trained with two NVIDIA Geforce RTX 3090 GPUs. The specific training hyper-parameter configurations of our FreMAE on BraTS 2019, ISIC 2018 and ACDC 2017 can be found in Table 2, 3, 4 respectively.

4.2. Results and Analysis

Comparison with Previous SSL Frameworks. Based on five-fold cross-validation on the BraTS 2019 training set, we perform a fair comparison between our proposed FreMAE and previous self-supervised learning methods on various baselines including TransBTSV2 [31], UNETR [24], and Swin UNETR [44], demonstrating the effectiveness and generalization capability of our FreMAE. For comprehensive comparisons, we select multiple self-supervised learning methods (*i.e.*, MAE [25], SimMIM [49], DINO [9] and Swin UNETR [44]), among which MAE and SimMIM have achieved promising results on natural images, DINO is a representative contrastive learning method, and Swin UNETR has a pretraining method for medical image analysis.

As presented in Table 1, our FreMAE shows great superiority over all three baselines. Compared to training from scratch, the Average Dice scores on three baselines are simultaneously increased by 1.14%, 1.38%, and 0.98% respectively after pre-trained with our proposed MIM-based framework. In comparison with MAE on UNETR and SimMIM on Swin UNETR, our FreMAE greatly improves model performance (*i.e.* ↑ **0.52%** and ↑ **0.36%** on Average Dice) with the benefit of exploiting MIM in the frequency domain for global representation learning. Since contrastive learning methods mainly focus on learning high-level semantics by instance discrimination task, neglecting the fine-grained representation learning results in poor results for UNETR with DINO pre-training. In contrast, FreMAE takes advantage of the smooth structure information of organs and detailed contours and textures as supervision signals, better guiding the model's high-level and low-level representation learning. Additionally, the Swin UNETR pre-training method achieves inferior performance. We believe the reasonable explanation for this phenomenon is that the Swin UNETR pre-training method heavily relies on the number of training samples to acquire the useful prior knowledge (*i.e.* it can not help models to capture the helpful representations as expected under the circumstance of limited pre-training samples). On the contrary, without introducing any extra samples, our FreMAE can greatly boost the model performance compared with random initialization, which exactly proves the effectiveness and data-efficient characteristic of our method. In summary, our FreMAE with the advantages of the frequency domain is a generic and powerful MIM-based framework, which could bring consistent improvement in model performance without introducing extra data.

Baseline	Backbone	Pre-train Method	Dice Score (%) ↑			
			ET	WT	TC	Average
TransBTSV2 [31]	CNN-Transformer	-	77.11	90.32	82.90	83.44
TransBTSV2 [31]	CNN-Transformer	FreMAE	**79.65** (+2.54)	**90.80** (+0.48)	**83.33** (+0.43)	**84.59** (+1.15)
UNETR [24]	ViT-B/16 [20]	-	75.28	88.42	76.33	80.01
UNETR [24]	ViT-B/16 [20]	MAE [25]	75.18 (-0.10)	**88.95** (+0.53)	78.47 (+2.14)	80.87 (+0.86)
UNETR [24]	ViT-B/16 [20]	DINO [9]	75.22 (-0.06)	88.33 (-0.09)	75.89 (-0.44)	79.81 (-0.20)
UNETR [24]	ViT-B/16 [20]	FreMAE	**76.50** (+1.22)	88.86 (+0.44)	**78.82** (+2.49)	**81.39** (+1.38)
Swin UNETR [44]	Swin-B [33]	-	76.68	89.89	79.98	82.18
Swin UNETR [44]	Swin-B [33]	SimMIM [49]	77.59 (+0.91)	**90.47** (+0.58)	80.34 (+0.36)	82.80 (+0.62)
Swin UNETR [44]	Swin-B [33]	Swin UNETR [44]	77.85 (+1.17)	89.63 (-0.26)	78.65 (-1.33)	82.04 (-0.14)
Swin UNETR [44]	Swin-B [33]	FreMAE	**78.38** (+1.70)	90.06 (+0.17)	**81.05** (+1.07)	**83.16** (+0.98)

Table 1. Comparison with previous self-supervised learning frameworks. '-' represents training from scratch. Without introducing any extra samples, our FreMAE can consistently boost the model performance by a large margin compared with randomly initialized baselines.

Config	Pre-training	Fine-tuning
optimizer	Adam	Adam
base learning rate	10^{-4}	10^{-4}
weight decay	10^{-5}	10^{-5}
batch size	64	64
lr decay schedule	cosine decay	cosine decay
training epochs	250	500

Table 2. Training settings on BraTS 2019 dataset.

Config	Pre-training	Fine-tuning
optimizer	SGD	SGD
base learning rate	10^{-3}	5×10^{-4}
weight decay	10^{-8}	10^{-8}
batch size	12	12
lr decay schedule	poly	poly
training epochs	125	300

Table 3. Training settings on ISIC 2018 dataset.

Config	Pre-training	Fine-tuning
optimizer	SGD	SGD
base learning rate	10^{-2}	10^{-2}
weight decay	10^{-4}	10^{-4}
batch size	16	16
lr decay schedule	poly	poly
training epochs	300	1200

Table 4. Training settings on ACDC 2017 dataset.

(a) BraTS 2019

Method	Dice Score (%) ↑			Hausdorff Dist. (mm) ↓		
	ET	WT	TC	ET	WT	TC
3D U-Net [15]	70.86	87.38	72.48	5.062	9.432	8.719
V-Net [35]	73.89	88.73	76.56	6.131	6.256	8.705
Attention U-Net [38]	75.96	88.81	77.20	5.202	7.756	8.258
Chen et al. [13]	74.16	**90.26**	79.25	4.575	**4.378**	7.954
Li et al. [32]	77.10	88.60	**81.30**	6.033	6.232	7.409
Frey et al. [22]	78.70	89.60	80.00	6.005	8.171	8.241
TransUNet [10]	78.17	89.48	78.91	4.832	6.667	7.365
Swin-UNet [8]	78.49	89.38	78.75	6.925	7.505	9.260
TransBTSV2 [31]	78.63	90.09	80.23	3.729	6.194	7.725
TransBTSV2	**79.74**	90.23	81.25	**3.209**	5.875	**6.934**
+FreMAE	**+1.11**	**+0.14**	**+1.02**	-0.520	-0.319	**-0.791**

(b) ISIC 2018

Method	JI	Dice	Accuracy	Recall	Precision
U-Net [42]	81.69	88.81	95.68	88.58	91.31
U-Net++ [53]	81.87	88.93	95.68	89.10	90.98
AttU-Net [38]	81.99	89.03	95.77	88.98	91.26
DeepLabv3+ [11]	82.32	89.26	95.87	89.74	90.87
CPF-Net [21]	82.92	89.63	96.02	**90.62**	90.71
BCDU-Net [1]	80.84	88.33	95.48	89.12	89.68
Ms RED [17]	83.45	89.99	96.19	90.49	91.47
TransBTSV2 [31]	81.96	92.56	95.88	90.21	90.78
TransBTSV2	**83.53**	**93.39**	**96.44**	90.18	**92.61**
+FreMAE	**+1.57**	**+0.83**	**+0.56**	-0.03	**+1.83**

(c) ACDC 2017

Method	RV	Myo	LV	Average
U-Net [42]	86.91	87.17	90.65	88.25
AttU-Net [38]	86.78	86.93	91.84	88.52
Swin-UNet [8]	86.62	88.72	92.44	89.26
TransUNet [10]	87.04	88.51	**92.85**	89.47
TransBTSV2 [31]	86.80	87.76	91.87	88.81
TransBTSV2 [31]	**87.12**	**88.87**	92.69	**89.56**
+FreMAE	**+0.32**	**+1.11**	**+0.82**	**+0.75**

Table 5. Performance comparisons on BraTS 2019, ISIC 2018 and ACDC 2017 datasets.

Evaluation on Brain Tumor Segmentation. Comparative experiments are also conducted on BraTS 2019 validation set. As shown in Table 5 (a), our FreMAE achieves superior performance than previous methods with the Dice scores of 79.74%, 90.23%, and 81.25% on ET, WT, and TC respectively. In addition, it is notable that our method realizes a considerable decrease of Hausdorff distance on TC, reaching $6.934mm$. These quantitative results are powerful evidence of the availability and effectiveness of using our method on MRI benchmarks.

Evaluation on Skin Lesion Segmentation. We also verified the generality of FreMAE on RGB images dataset namely ISIC 2018 compared with the other seven well-performed algorithms. It could be seen from Table 5 (b) that, with the informative feature representations obtained from pre-training stages, our method could reach great performance on ISIC 2018 the five-fold cross-validation. Specifically, compared with previous SOTA methods, our results are higher on both JI, Dice, Accuracy, and Precision metrics. It is worth noting that our method promotes **1.57%**

and **1.83%** on Dice score and Precision compared to training from scratch, demonstrating that FreMAE also presents strong capability on skin lesion segmentation.

Evaluation on Cardiac sSegmentation. To evaluate the generalization ability of our proposed FreMAE, we also conduct experiments of cardiac segmentation on MRI scans utilizing the ACDC 2017 dataset [6]. Since the official evaluation is supported by the online evaluation platform, the five-fold cross-validation is performed on ACDC 2017 training set. The quantitative results on ACDC 2017 training set are presented in Table 5 (c). It is obvious that with boosted model performance in comparison with baseline, our framework once again achieves comparable or even higher Dice scores than previous SOTA methods.

5. Visual Comparison for Qualitative Analysis

Segmentation Results. Firstly, the skin lesion segmentation results on ISIC 2018 dataset is presented in Fig. 4. It can be obviously seen that the model can generate much more accurate and fine-grained segmentation masks compared with baseline with the benefit of employing our proposed FreMAE. Simultaneously, we compare the segmentation performance of different self-supervised methods, including MAE, DINO, and FreMAE on the BraTS 2019 dataset with visualization results. As shown in Fig. 5, our method promotes the detailed pixel delineation of brain tumors and obtains more accurate predictions.

Reconstruction Results. To convincingly prove the superiority of our FreMAE, we further supplement more visual comparison of reconstruction results on BraTS 2019 dataset for qualitative analysis. As is shown in Fig. 6, our method can nicely achieve the reconstruction task of Fourier spectrum and generate the corresponding reconstruction spectrum approximately the same as original image. To be mentioned, for each image slice, the first row is the original image and the second row is our reconstruction results of the Fourier spectrum.

5.1. Ablation Studies

We conduct extensive ablation experiments to prove the effectiveness of our FreMAE and validate its design rationale based on five-fold cross-validation on the BraTS 2019 training set, while TransBTSV2[31] is selected as our baseline for ablation studies.

low-level target	high-level target	Dice Score (%) ↑			
		ET	WT	TC	Average
-	-	77.11	90.32	82.90	83.44
high-pass	-	77.82	90.60	**83.60**	84.01(+0.57)
-	low-pass	77.44	90.12	82.89	83.48(+0.04)
original image	original image	79.33	90.23	81.95	83.83(+0.39)
all frequency	all frequency	79.12	**90.80**	82.58	84.17(+0.73)
low-pass	high-pass	79.01	90.41	83.00	84.14(+0.70)
high-pass	low-pass	**79.65**	**90.80**	83.33	**84.59**(**+1.15**)

Table 6. Ablation study on the reconstruction target and supervision scheme.

Masking strategy	Dice Score (%) ↑			
	ET	WT	TC	Average
baseline	77.11	90.32	82.90	83.44
random mask	79.07	90.64	83.19	84.30(+0.86)
block wise mask	79.03	90.00	82.11	83.71(+0.27)
foreground mask	**79.65**	**90.80**	**83.33**	**84.59**(**+1.15**)

Table 7. Ablation study on the masking strategy.

Reconstruction Target and Supervision Scheme. Firstly, we explore the effect of different types of reconstruction targets and verify the effectiveness of our introduced multi-stage supervision scheme. The quantitative results are presented in Table 6. In comparison with random initialization in the first row, introducing either high-pass Fourier spectrum or low-pass counterpart as the reconstruction target at the corresponding low-level or high-level stage both lead to better segmentation performance to some extent. On the basis of this kind of single-level supervision manner, we further explore the effectiveness of a multi-level supervision scheme. As can be clearly seen in Table 6 below the dividing line, simultaneously taking advantage of high-pass and low-pass frequency components, that carry abundant local details and global structural information, results in the best segmentation accuracy with the highest Average Dice Score of 84.59%, fully demonstrating the powerful potential and rationale design of our FreMAE. No matter whether the reconstruction target is adjusted to the original image, the whole Fourier spectrum, or exchanged low/high-level target, it will all lead to a considerable decrease in model performance, which once again proves the strong theoretical rationale of exploiting FFT with the proposed FreMAE.

Masking Strategy. Then we investigate the influence of different masking strategies to prove the effectiveness of the proposed foreground masking strategy. Table 7 shows the performance comparison of our FreMAE with different masking strategies. It can be clearly seen in Table 7 that the original random masking leads to an accuracy increase ↑0.86% on Average Dice score from 83.44% to 84.30%, which is really promising. However, by replacing the vanilla random masking with our simple yet powerful foreground masking strategy, the model performance on segmentation tasks can be further boosted by a considerable margin, which shows the great superiority of selecting masked pixel candidates solely among foreground over conventional masking strategy.

Masking Ratio. After investigating the influence of various masking strategies, we further conduct experiments to seek the optimal masking ratio for our current framework. As presented in Table 8, our FreMAE with a masking ra-

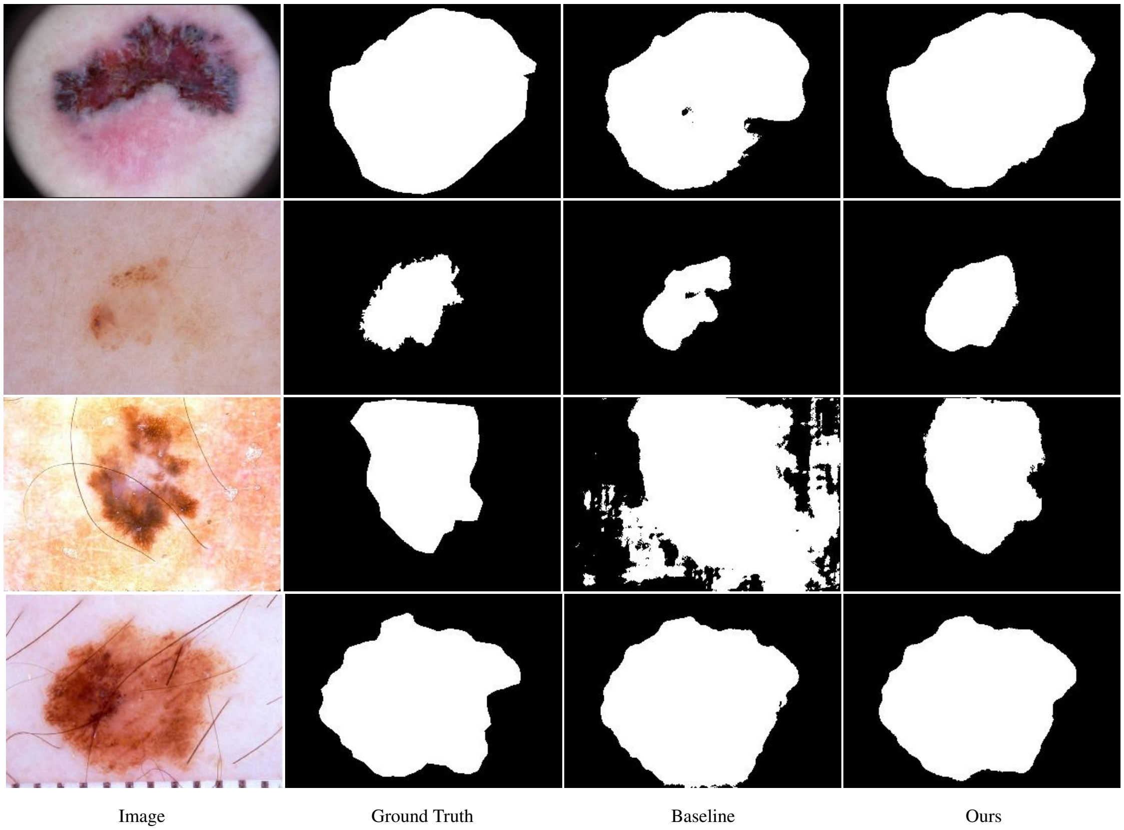

Figure 4. The visual comparison of skin lesion segmentation results on ISIC 2018 dataset with TransBTSV2 as the baseline.

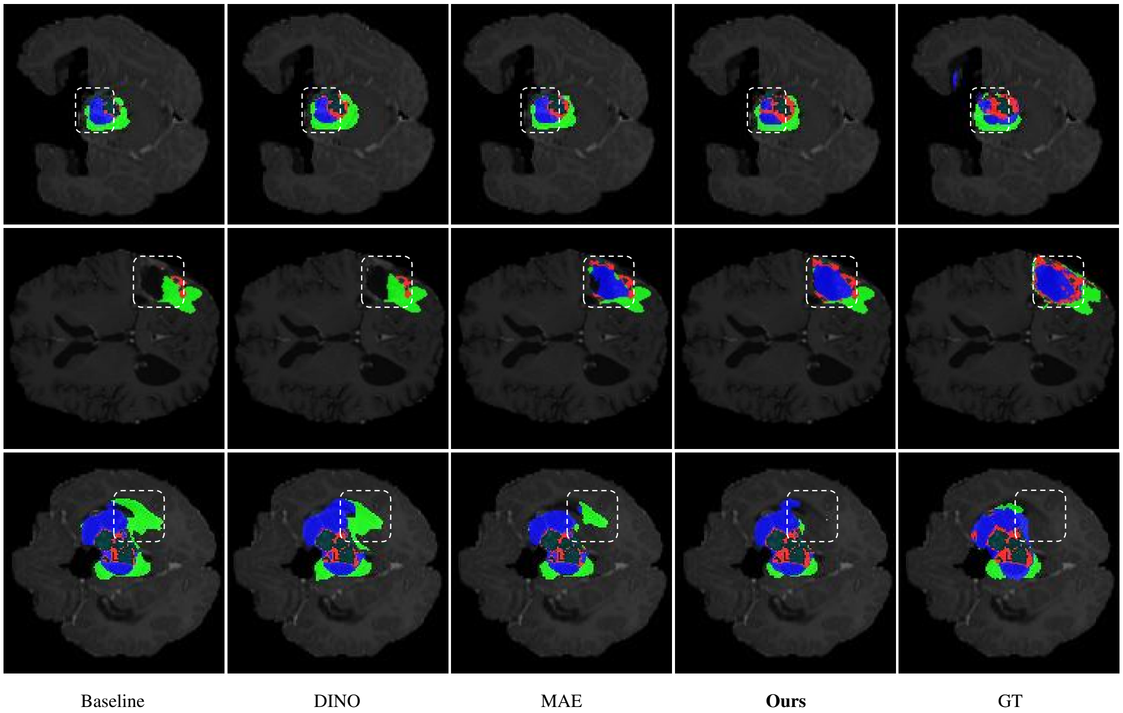

Figure 5. The visual comparison of MRI brain tumor segmentation results with UNETR as baseline. The blue regions denote the enhancing tumors, the red regions denote the non-enhancing tumors and the green ones denote the peritumoral edema.

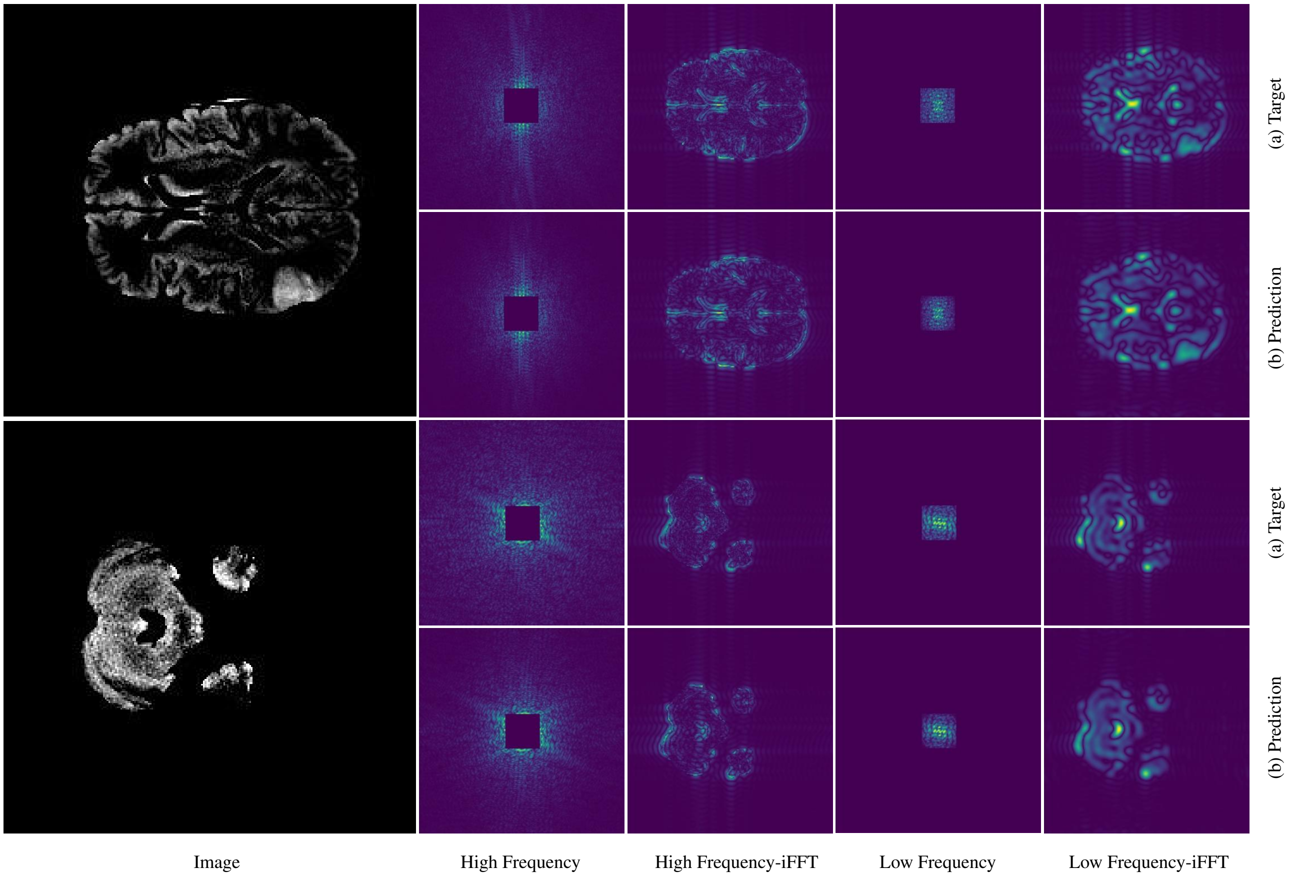

Figure 6. The visualization of reconstruction results by our FreMAE in the frequency domain.

Masking Ratio	Dice Score (%) ↑			
	ET	WT	TC	Average
baseline	77.11	90.32	82.90	83.44
0.75	78.99	90.42	83.03	84.15(+0.71)
0.50	79.19	**90.80**	83.18	84.39(+0.95)
0.25	**79.65**	**90.80**	**83.33**	**84.59(+1.15)**
0.15	79.37	90.23	82.88	84.16(+0.72)
0.15, 0.20, 0.25	78.99	90.63	**83.33**	84.32(+0.88)
0.25, 0.50, 0.75	79.23	90.62	82.88	84.24(+0.80)

Table 8. Ablation study on the masking ratio.

tio of 0.25 achieves the best model performance. Once the masking ratio is either too low or too high, the reconstruction task in the frequency domain would be too easy or too hard, which may hinder the model from expected representation learning during self-supervised pre-training. Besides, trying to take a step further, we also attempt to introduce a novel dynamic masking strategy (*i.e.* the masking ratio gradually increases from the lowest to the highest during pre-training) for better guidance of the expected feature representation learning, which endows the SSL with easiest-to-hardest reconstruction level. However, none of these attempts bring further accuracy improvements. Thus, the static masking strategy with masking ratio of 0.25 is selected as our default setting.

Training samples	Dice Score (%) ↑			
	ET	WT	TC	Average
baseline	77.11	90.32	82.90	83.44
0.3%(*i.e.* 1 sample)	79.05	90.60	82.51	84.05(+0.61)
10%	79.06	90.41	**83.43**	84.30(+0.86)
100%	**79.65**	**90.80**	83.33	**84.59(+1.15)**

Table 9. Ablation study on the number of samples for self-supervised pre-training.

Number of Pre-training Samples. Specifically, we further investigate the effect of different percentages of training samples used for our proposed FreMAE. The quantitative results are presented in Table 9. It is clear in Table 9 that the model performance is consistently improved with more and more employed training samples for the proposed FreMAE. Besides, it is also surprising that by solely introducing 1 sample for pre-training our FreMAE can boost the model performance by a large margin (*i.e.* ↑ 0.61% on Average Dice score) compared with the randomly initialized baseline, demonstrating that our method is a data-efficient self-supervised learning paradigm.

6. Conclusion

In this paper, we presented the first study on exploring the powerful potential of MIM with frequency domain on pre-training deep learning models for medical image segmentation tasks. We focus on 2D medical image segmentation and propose a new framework FreMAE taking advantage of both the rich global information and local details in the Fourier spectrum. Deviating from the conventional paradigm as previous MIM methods, realizing reconstruction in the frequency domain empowers the frak with stronger representation learning capability. Besides, by fully exploiting the specific characteristics contained in different frequency bands, the multi-stage supervision scheme can greatly boost the segmentation performance. Comprehensive experiments on three benchmark datasets quantitatively and qualitatively demonstrate that our FreMAE significantly improves the segmentation performance of baselines trained from scratch and shows great superiority over previous self-supervised SOTA approaches.

References

[1] Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. Bi-directional convlstm u-net with densley connected convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 7

[2] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4:170117, 2017. 6

[3] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018. 6

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 2, 3, 4

[5] Moshe Bar. A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of cognitive neuroscience*, 15(4):600–609, 2003. 2, 3

[6] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 6, 8

[7] Jean Bullier. Integrated model of visual processing. *Brain research reviews*, 36(2-3):96–107, 2001. 2, 3

[8] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021. 7

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 3, 6, 7

[10] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 7

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 7

[12] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 1

[13] Minglin Chen, Yaozu Wu, and Jianhuang Wu. Aggregating multi-scale prediction based on 3d u-net in brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 142–152. Springer, 2019. 7

[14] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3435–3444, 2019. 2, 3

[15] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 7

[16] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 6

[17] Duwei Dai, Caixia Dong, Songhua Xu, Qingsen Yan, Zongfang Li, Chunyan Zhang, and Nana Luo. Ms red: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. *Medical Image Analysis*, 75:102293, 2022. 7

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[21] Shuanglang Feng, Heming Zhao, Fei Shi, Xuena Cheng, Meng Wang, Yuhui Ma, Dehui Xiang, Weifang Zhu, and Xinjian Chen. Cpfnet: Context pyramid fusion network for medical image segmentation. *IEEE transactions on medical imaging*, 39(10):3008–3018, 2020. 7

[22] Markus Frey and Matthias Nau. Memory efficient brain tumor segmentation using an autoencoder-regularized u-net. In *International MICCAI Brainlesion Workshop*, pages 388–396. Springer, 2019. 7

[23] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 1, 4

[24] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022. 6, 7

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1, 2, 3, 4, 6, 7

[26] Junjia Huang, Haofeng Li, Guanbin Li, and Xiang Wan. Attentive symmetric autoencoder for brain mri segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 203–213. Springer, 2022. 1, 3

[27] Yuankai Huo, Jiaqi Liu, Zhoubing Xu, Robert L Harrigan, Albert Assad, Richard G Abramson, and Bennett A Landman. Robust multicontrast mri spleen segmentation for splenomegaly using multi-atlas segmentation. *IEEE Transactions on Biomedical Engineering*, 65(2):336–343, 2017. 1

[28] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13919–13929, 2021. 3, 5

[29] Louise Kauffmann, Stephen Ramanoël, and Carole Peyrin. The neural bases of spatial frequency processing during scene perception. *Frontiers in integrative neuroscience*, 8:37, 2014. 2, 3

[30] Xiangwen Kong and Xiangyu Zhang. Understanding masked image modeling via learning occlusion invariant feature. *ArXiv*, abs/2208.04164, 2022. 2

[31] Jiangyun Li, Wenxuan Wang, Chen Chen, Tianxiang Zhang, Sen Zha, Hong Yu, and Jing Wang. Transbtsv2: Wider instead of deeper transformer for medical image segmentation. *arXiv preprint arXiv:2201.12785*, 2022. 6, 7, 8

[32] Xiangyu Li, Gongning Luo, and Kuanquan Wang. Multi-step cascaded networks for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 163–173. Springer, 2019. 7

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 7

[34] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 6

[35] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 7

[36] Huu-Giao Nguyen, Celine Fouard, and Jocelyne Troccaz. Segmentation, separation and pose estimation of prostate brachytherapy seeds in ct images. *IEEE Transactions on Biomedical Engineering*, 62(8):2012–2024, 2015. 1

[37] Henri J Nussbaumer. The fast fourier transform. In *Fast Fourier Transform and Convolution Algorithms*, pages 80–111. Springer, 1981. 3

[38] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 7

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[40] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 1, 3, 4

[41] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021. 3

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 7

[43] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. *arXiv preprint arXiv:2205.12956*, 2022. 2, 3

[44] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20730–20740, 2022. 1, 3, 6, 7

[45] Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022. 1, 4

[46] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 6

[47] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020. 2

[48] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 2, 3

[49] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 1, 3, 4, 6, 7

[50] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 1, 3

[51] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image analysis. *arXiv preprint arXiv:2203.05573*, 2022. 1, 3

[52] Man Zhou, Hu Yu, Jie Huang, Feng Zhao, Jinwei Gu, Chen Change Loy, Deyu Meng, and Chongyi Li. Deep fourier up-sampling. *arXiv preprint arXiv:2210.05171*, 2022. 3

[53] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018. 7