FreqMamba-UNet with Dual-Path Fusion and Multi-Scale Feature Interaction: Segmentation of Ground-Glass Opacities and Consolidation in Lung CT

Gang Li†
Taiyuan University of Technology
Taiyuan 030024, China
tx2090@126.com

Yijun Lin†
2024511302@link.tyut.edu.cn

Kairu Zhang
2023521878@link.tyut.edu.cn

Hao Liu
2023511245@link.tyut.edu.cn

Ling Zhang*
zl2090@126.com

Abstract

Ground-glass opacities (GGOs) and consolidations (CONs) are key radiological markers for staging pneumonia, and their accurate segmentation is crucial for disease assessment. However, GGOs and CONs exhibit distinct lesion characteristics: GGOs feature blurred borders and heterogeneous density, while CONs, though denser, closely resemble the gray-scale of normal pulmonary structures. Both lesions are morphologically diverse and heterogeneously distributed, posing significant challenges for high-precision segmentation in pulmonary CT images. To address this, this paper proposes a deep learning segmentation architecture named FreqMamba-UNet. First, a dual-path fusion encoder combines frequency-domain and spatial-domain features. Frequency-domain information enhances sensitivity to the blurred boundaries of GGOs and amplifies texture differences to distinguish CON from normal tissue. Second, a multi-scale feature bridge based on Mamba is designed to model long-range dependencies between diffuse lesions. Finally, a dual-decoder architecture optimizes segmentation strategies for GGOs and CON separately, enhancing adaptability to distinct lesion characteristics. Experiments on public COVID-19 CT datasets demonstrate that the proposed model outperforms mainstream methods including U-Net, U-Net++, Attention U-Net, and Inf-Net, achieving optimal results in metrics such as F1 score and IoU. This exhibits excellent generalization capability and clinical applicability potential.

***Keywords:** Medical image segmentation, Convolutional neural networks, Frequency domain feature fusion.*

*Corresponding author:Ling Zhang (E-mail:zl2090@126.com).† Gang Li and Yijun Lin contributed equally to this work (Co-first authors).

1. Introduction

Pneumonia is an inflammatory lung disease caused by various pathogens including bacteria, viruses, and fungi, posing significant risks to vulnerable populations such as children, the elderly, and immunocompromised individuals [1] . Statistics indicate that over 2.5 million people die annually from lung inflammation. Its high incidence and mortality rates, coupled with the growing issue of antibiotic resistance, collectively constitute a major global health challenge [2, 3] .

Chest computed tomography (CT) demonstrates high sensitivity for detecting early-stage disease and can identify minute lesions, making it widely used for pneumonia diagnosis and prognosis [4]. Pneumonia's CT presentation evolves dynamically with disease progression and individual variations, featuring typical radiographic characteristics including bilateral patchy ground-glass opacities (GGO) and denser consolidation (CON). GGOs appear as areas of increased density with blurred margins, typically presenting as round or oval opacities. CONs manifest as more dense, pure-white areas with indistinct or blurred borders due to inflammatory exudate or hemorrhage, potentially appearing as irregular or patchy regions. The coexistence of both indicates different stages of alveolar injury, serving not only as a key indicator for early diagnosis but also closely correlating with disease severity and prognosis [5, 6, 7, 8]. A single lesion on pulmonary CT often contains multiple pathological changes such as GGOs, CONs, or even fibrosis, leading to overlapping imaging features. Infected areas exhibit low contrast against normal tissue, while CONs, crazy paving sign and non-pulmonary tissue boundaries appear blurred. Additionally, lesions often display irregular morphology and discrete spatial distribution [9, 10]. However, traditional convolutional neural networks (CNNs) extract only spatiotemporal features, prone to losing critical edge details and struggling to accurately identify the blurred borders of GGOs or the subtle transition zones

between CON and lung parenchyma in CT images. While U-Net and its variants demonstrate excellent performance, they still fail to effectively capture the long-range spatial correlations required for diffuse lesions in CT, thereby compromising segmentation accuracy.

To address the aforementioned challenges, this paper proposes a medical image segmentation architecture named FreqMamba-UNet.By integrating frequency-domain and spatial-domain analysis with collaborative modeling of global and local features, it optimizes segmentation for GGO and CON characteristics, thereby enhancing the model's segmentation accuracy and robustness. This architecture comprises three components: First, a dual-path fusion encoder is designed. By concurrently integrating a multi-level wavelet frequency domain branch and a CNN spatial domain branch, it utilizes the Discrete Wavelet Transform(DWT) to preserve high-frequency information. This effectively merges the image's frequency domain information with spatial context, enhancing the ability to capture high-frequency details such as the fuzzy boundaries of GGOs. Second, a multi-scale feature bridge based on Mamba is introduced between the encoder and decoder. This enables cross-scale feature interaction and modeling of global dependencies, addressing the heterogeneity in lesion morphology and distribution. Finally, a dual-decoder parallel architecture is adopted, customizing dedicated segmentation paths for GGOs and CONs. Attention mechanisms or Edge Enhancement Modules(EEM) are selectively integrated with spatial pyramid pooling to enhance the model's adaptability to the characteristics of these two heterogeneous lesions and improve overall segmentation accuracy. Experiments conducted on publicly available datasets including COVID-19 CT segmentation, segmentation dataset nr.2, and Synapse demonstrate superior segmentation performance of the proposed architecture across multi-class segmentation tasks when compared against eight segmentation architectures. The main contributions of this paper are as follows:

(1) Constructing a multi-level wavelet-frequency domain branch based on the Wavelet Frequency Merger(WFM) to fuse frequency-domain and spatial-domain features, thereby enhancing perception of high-frequency information such as edges and textures.

(2) Designing a multi-scale feature bridge based on Mamba to model long-range dependencies among multi-scale features.

(3) A dual-decoder parallel architecture is adopted, with dedicated segmentation paths tailored for the heterogeneous characteristics of GGO and CON, enhancing segmentation accuracy and performance across different target regions.

2. Related Methods

In recent years, medical image segmentation methods have achieved significant progress, evolving from traditional approaches (e.g., edge detection, fuzzy logic, and graph models [11]) to deep learning (DL)-based segmentation algorithms. Among these, convolutional neural networks (CNNs) have demonstrated superior performance in medical image segmentation tasks due to their robust local feature extraction and spatial modeling capabilities [12, 13] . As an early classic CNN-based architecture, the U-Net model established the foundation for the encoder-decoder segmentation paradigm, with its core principles serving as a crucial reference for subsequent improved models. Concurrently, frequency-domain representations have been proven to serve as an effective complementary modality to spatial-domain features due to their sensitivity in capturing global textures and periodic signals . By separating and extracting high-frequency edge information and low-frequency structural information from images, they further enhance the accuracy of medical image segmentation.

2.1. U-Net-Based Segmentation Architecture

The classic U-Net proposed by Ronneberger et al. [14] employs a symmetric encoder-decoder architecture to extract multi-level features and achieve precise localization. Zhu et al.'s [15] CoupleNet captures local and global information through dual branches, laying the foundation for feature collaboration modeling. Zhou et al. [16]introduced U-Net++ with nested dense skip connections to narrow semantic gaps and enhance accuracy; Oktay et al. [17] embedded attention gates into skip connections in Attention U-Net to optimize lesion localization. Fan et al.'s [18]Inf-Net models lesion boundaries using parallel decoders and attention modules; Paluru et al.'s [19]Anam-Net achieves lightweight architecture with AD blocks, retaining feature capacity while reducing parameters to 1/7.8 of U-Net. Zhao et al.'s [20] SCOAT-Net builds upon U-Net++ by incorporating spatio-channel attention to enhance feature selectivity; Bougourzi et al.'s [21]D-TrAttUnet fuses Transformers with CNNs, using dual decoders to enhance robustness in handling diffuse lesions. Dang et al.'s [22]CAD-Unet combines capsule networks with U-Net, employing dual-path coupling to capture lesion spatial orientation while dual decoders focus on segmenting effective lung regions.

2.2. Frequency-Domain Segmentation Methods

As research progresses, frequency-domain methods demonstrate significant potential in medical image segmentation. They capture high-frequency edges and low-frequency structural features, effectively addressing limitations of traditional spatial-domain methods such as blurred boundary segmentation and missed detection of subtle lesions [23, 24, 25]. This advantage has been validated: Rao

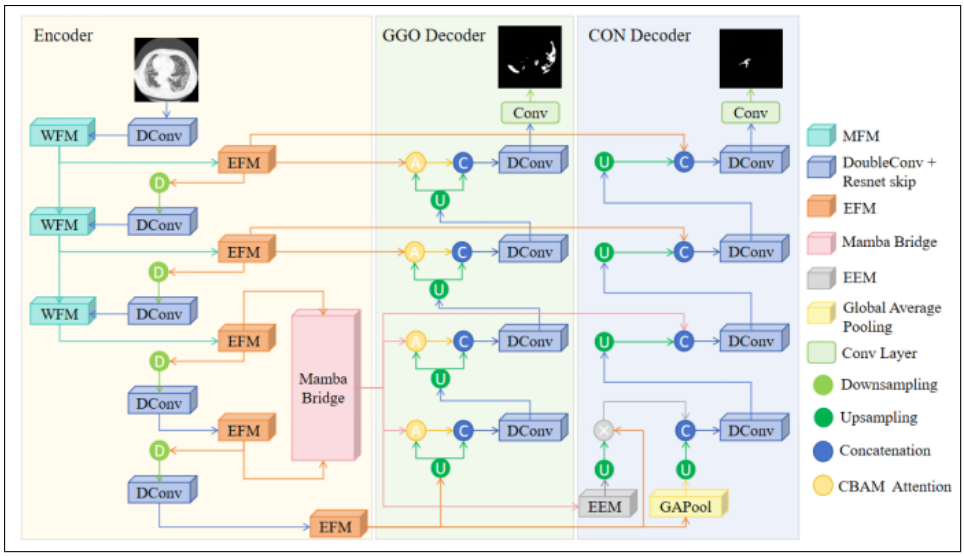

Figure 1: Overall Architecture Diagram of the FreqMamba-UNet Model.

et al.'s GFNet [26] maps spatial features to the frequency domain for enhanced reconstruction, achieving accuracy comparable to ViT and CNN in ImageNet classification; Showrav et al.'s Hi-gMISnet [27] employs wavelet transform frequency decomposition to address multi-scale object segmentation challenges; Resmi et al.'s [28] combined DCT frequency domain analysis with spectral attention to adapt to complex scenarios like low contrast and small targets, offering a frequency domain solution for pathological image analysis.

3. Method

3.1. Overall Network Architecture

This paper proposes a medical image segmentation architecture, FreqMamba-UNet, which employs collaborative modeling for simultaneous GGO and CON lesion segmentation to enhance the model's performance in distinguishing these two lesion types within infected regions. Within this architecture, multi-level wavelet branches are integrated in parallel with the CNN encoder branch. The former captures high-frequency edge information through the Wavelet Fusion Module (WFM), particularly suited for the precise detection of GGO's blurred boundaries. while the latter employs DoubleConv and an Edge Fusion Module (EFM) to extract spatiotemporal texture features. To enable efficient interaction and information transfer of multi-scale features between encoder and decoder, a Mamba-based multi-scale feature bridge is introduced between them. Addressing the morphological heterogeneity of GGO and CON, two independent decoding paths are designed: the GGO decoder focuses on segmenting typical GGO-affected regions, while the CON decoder emphasizes precise identification of minute CON lesions by integrating spatial pyramid pooling with an Edge Enhancement Module (EEM). The overall network architecture is illustrated in Figure 1.

3.2. Dual-Path Fusion Encoder

The dual-path fusion encoder consists of parallel multi-level wavelet branches and CNN encoder paths, achieving cross-path information collaboration through feature fusion. Its specific structure and principles are as follows:

3.2.1 Multi-level Wavelet Branches

The high-frequency components in wavelet transforms (LH, HL, HH subbands, corresponding to horizontal, vertical, and diagonal edge information respectively) can effectively enhance edge gradient features and improve resolution capabilities in low-contrast regions. Therefore, this paper introduces a multi-level wavelet branch based on a frequency domain feature fusion (WFM) module. By extracting high-frequency edge features through frequency domain analysis, it compensates for the deficiency of traditional spatial branches in capturing edge details. This branch comprises three WFM modules employing a multi-level wavelet decomposition mechanism. After each decomposition level, the high-frequency and low-frequency components at the current level undergo channel-dimensional integration and redundant information compression. The processed low-frequency features (LL subband) are then fed into the next WFM module for further decomposition. The final output is a feature map containing multi-scale frequency-domain information, enabling simultaneous ex-

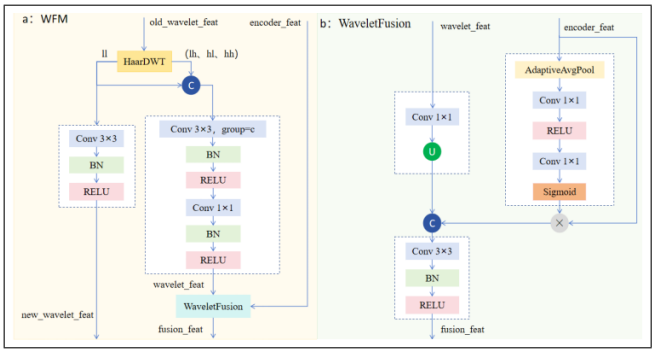

Figure 2: Multi-level wavelet branches. Figure (a) shows the WFM module diagram, while Figure (b) depicts the WaveletFusion module diagram—a feature fusion submodule within WFM—used to fuse frequency-domain features with original CNN encoder features.

traction of high-frequency edge features and low-frequency structural features from the input CT image.

(1) HaarDWT

The predefined HaarDWT comprises four fixed 2x2 filters: W_{ll}, W_{lh}, W_{hl}, and W_{hh}, corresponding to different frequency component extractions. Let the input feature map be x with shape $[B, C, H, W]$, where B is the batch size, C is the number of channels, and H/W is the height/width. Each channel undergoes an independent wavelet transform, calculated as in (1).

$$\begin{aligned} Y = \text{Conv2D}(X, \{W_{ll}, W_{lh}, W_{hl}, W_{hh}\}, \\ \text{stride} = 2, \text{groups} = C) \\ LL = Y_{0:C}, \quad LH = Y_{C:2C} \\ HL = Y_{2C:3C}, \quad HH = Y_{3C:4C} \end{aligned} \tag{1}$$

Where $W = [W^{ll}, W^{lh}, W^{hl}, W^{hh}]^T \in \mathbb{R}^{4\times1\times2\times2}$.

HaarDWT stacks the four filters into a tensor of shape $[4, 1, 2, 2]$. This is then expanded to $[4C, 1, 2, 2]$ via $repeat(C, 1, 1, 1)$, enabling independent filter application per channel. Group convolution decomposes each input image into four subbands: the low-frequency component LL preserves global structure, while the three high-frequency components LH, HL, and HH capture edge details in horizontal, vertical, and diagonal directions.

(2) Wavelet Feature Fusion Module (WFM)

The module structure of WFM is. shown in Figure 2(a) This module recursively applies Haar wavelet transforms to achieve multi-level wavelet decomposition, fusing frequency-domain features extracted from wavelet branches with spatial-domain features extracted from the encoder. Specifically, the first level decomposes the feature map output from the first DoubleConv block in the CNN encoder. Each subsequent level decomposes the low-frequency component (LL) passed from the previous level and transmits it to the next level. Assuming the input feature map is x(l) with shape $[B, C, H, W]$, the formula for extracting frequency-domain features is shown in (2).

$$\begin{aligned} LL^{(l)}, \{LH^{(l)}, HL^{(l)}, HH^{(l)}\} = \text{Haar-DWT}(X^{(l)}) \\ LL^{(l)}_{\text{proc}} = \text{ReLU}\big(\text{BN}\big(\text{Conv}_{3\times3}(LL^{(l)})\big)\big) \\ F^{(l)}_{\text{combined}} = \text{Concat}[LL^{(l)}, LH^{(l)}, HL^{(l)}, HH^{(l)}] \\ F^{(l)}_{\text{wavelet}} = \Gamma^{(l)}_{\text{wavelet}}(F^{(l)}_{\text{combined}}) \\ \Gamma^{(l)}_{\text{wavelet}} = \text{Conv}_{1\times1}\big(\text{ReLU}\big(\text{BN}\big(\text{Conv}^{\text{group}}_{3\times3}(\cdot)\big)\big)\big) \end{aligned} \tag{2}$$

Here, $l \in \{1, 2, 3\}$ denotes the decomposition level; $LL^{(l)}_{\text{proc}}$ represents the input to the next level of decomposition; $F^{(l)}_{\text{wavelet}}$ denotes the frequency domain features extracted at level l; $\text{Conv}_{3\times3}(\cdot)$ denotes a 3×3 convolution; $\text{BN}(\cdot)$ denotes batch normalization; $\text{ReLU}(\cdot)$ denotes the ReLU activation operation; and $\text{Concat}(\cdot)$ denotes the channel concatenation operation.

The WaveletFusion structure within the WFM module is illustrated in Figure 2(b). The formula for calculating the fused features is shown in Equation (3). It employs a channel attention mechanism to enhance critical information in the encoder features, fuses frequency-domain and spatial-domain features through feature concatenation and convolution, and performs adaptive upsampling to match the original feature dimensions.

$$\begin{aligned} W_{\text{red}} = \text{Conv}_{1\times1}(W_{\text{in}}) \\ \alpha = \sigma\big(\text{Conv}_{1\times1}\big(\text{ReLU}\big(\text{Conv}_{1\times1}\big(\text{GAP}(F_{\text{enc}})\big)\big)\big)\big) \\ F_{\text{fused}} = \Gamma_{\text{fuse}}\big(\text{Concat}\big[\alpha \odot F_{\text{enc}}, W_{\text{red}}\big]\big) \\ \Gamma_{\text{fuse}} = \text{ReLU}\big(\text{BN}\big(\text{Conv}_{3\times3}(\cdot)\big)\big) \end{aligned} \tag{3}$$

Here, $W_{\text{in}} \in \mathbb{R}^{B\times3\times H\times W}$ denotes the frequency-domain features extracted by the wavelet branches, $F_{\text{enc}} \in \mathbb{R}^{B\times C\times H\times W}$ represents the spatial-domain features extracted by the encoder, $\text{GAP}(\cdot)$ indicates global average pooling, $\sigma(\cdot)$ denotes the Sigmoid function, and $\odot$ signifies channel-wise multiplication.

3.2.2 CNN Encoder Path

The CNN encoder path consists of five DoubleConv modules. Each module employs a ”double convolution + residual connection” design to enhance feature depth while preventing gradient vanishing, progressively extracting spatiotemporal features. Each convolutional block is followed by an Edge Fusion Module (EFM). This module extracts edge features via separable convolutions and dynamically weights edge and texture features using lightweight channel attention, thereby enhancing local responses at lesion

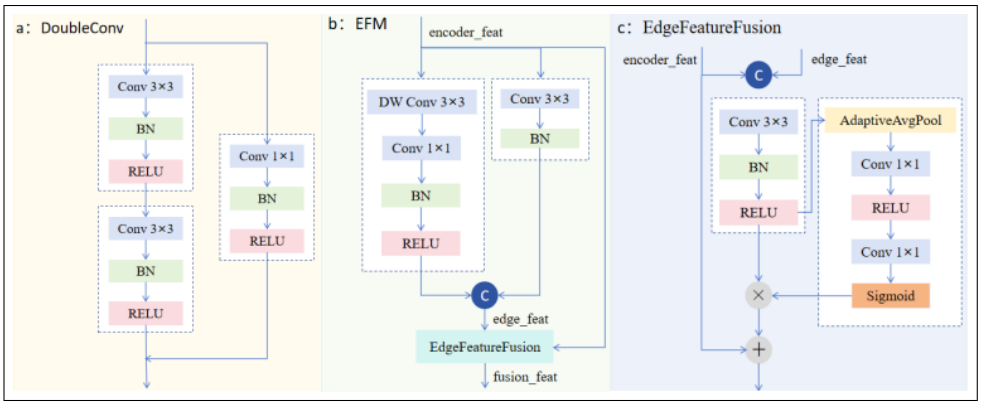

Figure 3: Schematic of CNN encoder modules: (a) DoubleConv module with residual connections; (b) EFM module; (c) EdgeFeatureFusion submodule within EFM.

boundaries . The module structure diagram forpath this is shown in Figure 3.

(1) DoubleConv

DoubleConv employs the classic "Convolution-BN-ReLU" combination, consecutively applying two 3×3 convolutional layers to increase the receptive field. Residual connections adjust channel dimensions via 1×1 convolutions before summing with the main branch output, preserving low-level features while learning high-level feature differences. The module structure is shown in Figure 3(a). Let the input feature be x; the module calculation formula is shown in (4).

$$\begin{aligned} F_{\text{conv}} &= \text{ReLU}\big(\text{BN}\big(\text{Conv}_{3\times3}\big(\\ &\quad \text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(x))))\big) \\ F_a &= \text{ReLU}\big(\text{BN}\big(\text{Conv}_{1\times1}(x)\big)\big) \\ \text{DoubleConv}(x) &= F_{\text{conv}} + F_{\text{aux}} \end{aligned} \tag{4}$$

(2) Edge Fusion Module (EFM)

The EFM module structure is shown in Figure 3(b). This module extracts image edge and contour features by employing depth-separable convolutions to reduce computational load. Predefined directional convolution kernels specifically capture edge information in horizontal, vertical, and other orientations. Given input features x, both basic edge feature extraction and directional edge feature extraction are performed. The results are concatenated channel-wise and fed as edge features to the EdgeFeatureFusion module, as calculated by Equation (5).

$$\begin{aligned} F_{\text{base\_edge}} &= \text{ReLU}\big(\text{BN}\big(C_{1\times1}(\text{Conv}_{3\times3,\text{depthwise}}(x))\big)\big) \\ F_{\text{direction\_edge}} &= \text{ReLU}\big(\text{Conv}_{3\times3,\text{direction}}(x)\big) \\ \text{EdgeEnhancement}(x) &= \text{Concat}(F_{\text{base\_edge}}, \\ &\quad F_{\text{direction\_edge}}) \end{aligned} \tag{5}$$

Additionally, the EdgeFeatureFusion in EFM combines original features with edge features. Through a channel attention mechanism, it autonomously focuses on important features while suppressing redundant information, enhancing the effectiveness of feature representation. The calculation formula is shown in (6).

$$\begin{aligned} F_{\text{edge}} &= \text{EdgeEnhancement}(x) \\ F_{\text{fused\_init}} &= \text{ReLU}\big(\text{BN}\big(C_{1\times1}(\text{Concat}(x, F_{\text{edge}}))\big)\big) \\ \alpha &= \text{Sigmoid}\big(C_{1\times1}\big(\text{ReLU}\big(\\ &\quad C_{1\times1}\big(\text{AvgPool}(F_{\text{fused\_init}})\big)\big)\big)\big) \\ \text{FeatureFusion}(x) &= x + \alpha \odot F_{\text{fused\_init}} \end{aligned} \tag{6}$$

3.3. Multi-scale Bridge Based on Mamba

This paper designs a multi-scale feature bridge based on Mamba as the core module connecting the encoder and decoder, with its module diagram shown in Figure 4. Its robust sequence modeling capability efficiently captures cross-region feature dependencies, addressing the limitation of insufficient long-range spatial correlation capture. Simultaneously, Mamba's context aggregation capability narrows the semantic gap between low-level and high-level features in the encoder through global information exchange. FeatureGate dynamically balances the weights of raw features and bridged features, preventing fine-grained edge features from being diluted by high-level semantic features.

3.3.1 Mamba

Mamba is a novel sequence model designed to address the computational inefficiency of Transformers when processing long sequences while enabling content-aware reasoning. It builds upon structured state space models (SSMs) through a series of enhancements, achieving

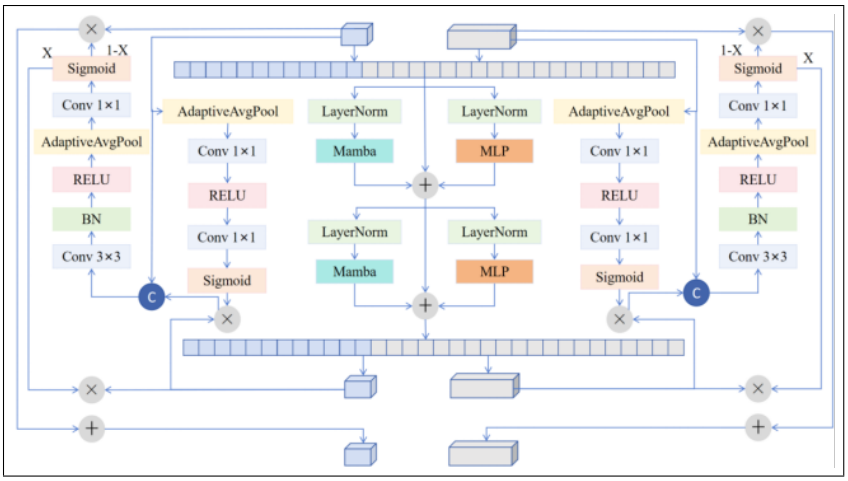

Figure 4: MambaBridge Module Diagram.

content-sensitive sequence processing via a selective mechanism while maintaining linear time complexity [29] . Its principle involves transforming the fixed parameters of traditional SSMs into input-dependent dynamic parameters, enabling the model to adaptively adjust information propagation based on current input. Specifically, Mamba implements this mechanism through the calculation formula (7):

$$\begin{aligned} x'_t &= W_{\text{in}} x_t \in \mathbb{R}^{2ED} \\ [A, B, C, D] &= \text{split}(x'_t) \\ s_t &= \exp(-\Delta_t) \odot s_{t-1} + B_t \odot x_t \\ y_t &= W_{\text{out}}(C_t \odot s_t) \in \mathbb{R}^D \end{aligned} \tag{7}$$

where $x_t \in \mathbb{R}^D$ is the input vector at time t, W_{in} and W_{out} are the input and output projection matrices, respectively, E is the scaling factor, N is the state dimension, $A \in \mathbb{R}^N B, C \in \mathbb{R}^N \Delta \in \mathbb{R}^N$ is the input-dependent dynamic parameter, $s_t \in \mathbb{R}^N$ is the state vector at time t, $\odot$ denotes element-wise multiplication, and $\text{softplus}(x) = \log(1 + \exp(x))$ ensures the non-negativity of Δ.

3.3.2 Feature Serialization and Mamba Processing

The Mamba module converts the multi-scale features output by the encoder into a sequence format ($B \times N \times C$, where N is the flattened length of the spatial dimension). Through selective state updates and gated convolutions, it captures long-range dependencies within the feature sequence while preserving edge details of small-sized CON lesions. By stacking two layers of Mamba processing with MLP mapping, it enhances semantic correlations across scales.

(1) Feature Sequentialization

Let the input feature map be $X \in \mathbb{R}^{B \times C \times H \times W}$ (where B is the batch size, C is the number of channels, and H, W are the spatial dimensions). The calculation formula for the serialization operation is shown in (8).

$$\begin{aligned} X_m = \text{Reshape}\big(&\text{Permute}(X, [0, 2, 3, 1]), \\ &[B, H \times W, C]\big) \end{aligned} \tag{8}$$

Here, $\text{permute}(X, [0, 2, 3, 1])$ moves the channel dimension to the end, while $\text{reshape}(\cdot)$ flattens the spatial dimensions into sequence length $H \times W$.

(2) Mamba Processing

Mamba Bridge is a module integrating the Mamba sequence model with feature fusion mechanisms, comprising LayerNorm, Mamba, and MLP residual connections. The calculation formula is shown in (9).

$$\begin{aligned} \hat{x} &= \text{LayerNorm}(x) \\ x_{\text{mamba}} &= \text{Mamba}(\hat{x}) \\ x_{\text{mlp}} &= \text{MLP}(\hat{x}) = W_2 \cdot \text{GELU}(W_1 \cdot \hat{x}) \\ x_{\text{out}} &= x + x_{\text{mamba}} + x_{\text{mlp}} \end{aligned} \tag{9}$$

Here, $W_1 \in \mathbb{R}^{C \times C}$ and $W_2 \in \mathbb{R}^{C \times C}$ represent the MLP layer weights.

3.3.3 Multi-scale Feature Reconstruction and Optimization

Sequence features processed by Mamba are restructured into spatial features matching the input scale, corresponding to fine-grained and coarse-grained features in the encoder. The FeatureGate module dynamically fuses original encoder features with Mamba-processed features, filtering key information from the bridged features to output an

optimized combination of "original features + bridged features."

The calculation formula for the feature gating mechanism used to adaptively fuse original and bridged features is shown in (10).

$$\begin{aligned} \text{retention} &= \sigma(\text{Conv}(\text{GlobalAvgPool}(X_{\text{pridged}}))) \\ \text{gate} &= \sigma(\text{Conv}(\text{GlobalAvgPool}(\\ &\quad \text{Conv}(\text{Concat}(X_{\text{original}}, X_{\text{bridged}} \circ \text{retention}))))) \\ X_{\text{out}} &= \text{gate} \circ X_{\text{pridged}} + (1 - \text{gate}) \circ X_{\text{original}} \end{aligned} \tag{10}$$

Here, x_{original} denotes the original features, x_{bridged} represents the Mamba-processed features, retention evaluates the importance of bridged features, and gate is the final fusion weight.

Finally, Mamba outputs sequences converted back into feature map form. This process is the inverse of the serialization operation, remapping sequences into two-dimensional feature maps to complete multi-scale feature reconstruction.

3.4. Dual Decoder

In CT scans, ground-glass opacities (GGOs) exhibit low contrast and blurred boundaries, while consolidation (CON) features high density and relatively clear borders, presenting significant differences. Traditional single decoders often misclassify normal lung tissue as GGOs or fibrotic regions as CON. The dual decoder shares features from the encoder and Mamba bridge, employing distinct network designs to adapt to the segmentation requirements of different lesions.

3.4.1 GGO Decoder

Given the blurred boundaries and susceptibility to small lesion misdetection characteristic of GGO lesions in imaging, the decoder must possess robust feature retention and detail capture capabilities. It should supplement contextual information for blurred boundaries through multi-scale feature fusion, thereby preventing misdetection due to insufficient information.

This decoder embeds a CBAM attention module at each decoding layer. Channel attention filters key feature channels, while spatial attention enhances lesion region responses. Simultaneously, a feature fusion pathway based on CBAM attention achieves refined integration of encoder features and bridged features. The decoding process adopts the classic " + DoubleConv" architecture, progressively mapping high-dimensional semantic features back to the image domain. The final segmentation result for GGO is output via a 1×1 convolution.

3.4.2 CON Decoder

Addressing the imaging characteristics of CON lesions—high density in , relatively clear boundaries, yet prone to confusion with fibrosis—the decoder requires enhanced edge localization and global morphological perception capabilities.

This decoder specifically incorporates an Edge Enhancement Module (EEM) and Spatial Pyramid Pooling (SPP). The EEM extracts high-frequency edge features via 3×3 convolutions, then generates an edge weight map through Sigmoid activation. This weighting enhances feature responses along CON lesion boundaries. Spatial Pyramid Pooling (SPP) captures the overall distribution characteristics of CON lesions through adaptive global pooling, compensating for segmentation biases in large CON regions caused by local features. The decoding path continues to employ a "dual convolution + upsampling" strategy to progressively restore feature map resolution. Simultaneously, by combining features from the Mamba bridge with low-level encoder features, high-resolution details are prioritized to precisely locate CON boundaries.

3.5. Loss Function

In lung CT lesion segmentation tasks, the distinct characteristics of GGO and CON lesions pose challenges for accurate segmentation. GGOs exhibit blurred boundaries and low contrast against normal tissue, while CONs feature high density and may be accompanied by complex fibrotic structures. This model employs tailored loss functions for each lesion type, utilizing a dual-lesion binary cross-entropy loss function to guide network training. By separately constraining the prediction results for both lesion types, it achieves precise segmentation of lesions with distinct characteristics.

The loss function formula is shown in (11), comprising two core components: GGO segmentation loss, which constrains the prediction results for GGO lesions to optimize segmentation accuracy in blurred boundary regions; and CON segmentation loss, which constrains the prediction results for CON lesions to enhance segmentation accuracy in high-density solid areas.

$$\begin{aligned} L_{\text{Total}} &= L_{\text{GGO}} + L_{\text{CON}} \\ L_{\text{GGO}} &= -\sum_{m=1}^{B}\sum_{i=1}^{W\cdot H}\big[y_{G,i}^{m}\log(p_{G,i}^{m}) \\ &\quad +(1-y_{G,i}^{m})\log(1-p_{G,i}^{m})\big] \\ L_{\text{CON}} &= -\sum_{m=1}^{B}\sum_{i=1}^{W\cdot H}\big[y_{C,i}^{m}\log(p_{C,i}^{m}) \\ &\quad +(1-y_{C,i}^{m})\log(1-p_{C,i}^{m})\big] \end{aligned} \tag{11}$$

Where B denotes the batch size, W and H represent the width and height of the predicted mask, respectively, $y_{\text{G},i}^{m} \in$

$\{0,1\}$ and $y^m_{C,i} \in \{0,1\}$ denote the ground truth labels (GGO and CON) for pixel i in the m-th sample, respectively, and $p^m_{G,i} \in [0,1]$ and $p^m_{C,i} \in [0,1]$ denote the model's predicted probabilities for pixel i being GGO and CON in the m-th sample, respectively.

4. Experiments and Results

4.1. Datasets

The proposed algorithm was evaluated on three publicly available datasets: , including COVID-19 CT segmentation, segmentation dataset nr.2 [30]—most existing public datasets with complete, fine-grained annotations for both GGO and CON lesions are COVID-19-related, and these two are representative ones—and the Synapse multi-organ segmentation dataset (Synapse). Table 1 provides an overview of the datasets used in this paper.

Name	Dataset 1	Dataset 2	Dataset 3
Dataset	COVID-19 CT Segmentation	Segmentation Dataset Nr.2	Synapse
CT-Scans	40	9	30
Slices	100	829	3779
Used Slices	100	829	2211

Table 1: Dataset Overview.

Due to limited data availability, this study combined Dataset 1 and Dataset 2—two core datasets—to evaluate model performance in multi-class segmentation tasks. The COVID-19 CT segmentation dataset comprises 100 axial CT slices collected from over 40 confirmed patients. Annotated by professional radiologists, it is specifically designed for multi-class segmentation tasks, providing detailed differentiation of lesion types. Segmentation Dataset No. 2 comprises 829 slices from 9 sets of 3D CT scans, with 373 slices featuring dual-label annotations (simultaneously containing both binary and multi-class annotations) by radiologists. This makes it a comprehensive dataset suitable for both types of segmentation tasks. Dataset 3 was used to validate model generalization. Synapse comprises abdominal CT scans from 30 patients (3,779 axial slices), annotated by specialists for eight abdominal organs: aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach. In the experimental setup, to simulate the limited data scenario encountered in clinical practice, only 18 training cases from this dataset were used, further divided into an 8:2 training-to-test split.

This study applied multimodal data augmentation techniques, including random rotation (-35° to +35°), elastic deformation, Gaussian blurring, and random occlusion. Augmentation intensity was balanced with data fidelity through hierarchical probability settings (e.g., geometric transformation p=0.8, color transformation p=0.5) to prevent excessive distortion of key pathological features and enhance model generalization.

4.2. Evaluation Metrics

The F1 score is the harmonic mean of precision and recall, reflecting the balance between true and false positives in positive class predictions. The Jaccard index(IoU) calculates the ratio of the intersection to the union of predicted and ground-truth regions, commonly used in segmentation tasks. Recall (sensitivity/true positive rate) quantifies a model's ability to cover positive samples.Specificity measures the model's ability to identify negative samples. Sensitivity measures the proportion of correctly identified positive samples (target class) out of all true positive samples.The calculation formula is shown in Formula (12).

$$\begin{aligned} F1 &= 100 \times \tfrac{2TP}{2TP+FP+FN}, \\ IoU &= 100 \times \tfrac{TP}{TP+FP+FN}, \\ Rec &= 100 \times \tfrac{TP}{TP+FN}, \\ Spec &= 100 \times \tfrac{TN}{FP+TN}, \\ Sensitivity &= \tfrac{TP}{TP+FN} \end{aligned} \tag{12}$$

Where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

4.3. Experimental Setup

The primary parameter configurations for this study are shown in Table 2:

Category	Configuration
GPU	NVIDIA GeForce RTX 4090
PyTorch Version	1.12.1
CUDA Version	12.2
Programming Language	Python 3.10.13
Image size	224
Epoch	120
Batch size	6
Optimizer	Adam
Learning rate	epoch $\leq$ 50: 1e-4 50 $<$ epoch $\leq$ 90: 1e-5 90 $<$ epoch: 1e-6

Table 2: Experiment Configuration Table

Model	GGO					CON				
	F1	IoU	Rec	Spec	Sens	F1	IoU	Rec	Spec	Sens
U-Net	61.22	44.11	50.43	99.20	50.43	55.04	37.96	53.88	99.19	53.88
U-Net++	64.00	47.06	52.78	99.19	53.88	46.83	30.57	45.63	99.11	44.62
Attention-Unet	58.22	41.06	45.37	99.42	45.37	50.58	33.85	50.72	99.03	50.72
Inf-Net	68.17	51.71	59.81	99.13	59.81	52.97	36.02	51.67	99.13	51.67
AnamNet	70.34	54.25	66.58	98.74	66.58	51.14	34.36	48.53	99.20	48.53
SCOAT-Net	61.12	44.01	49.17	99.35	49.17	54.82	37.76	57.08	99.01	57.08
D-TrAttUnet	63.93	46.98	49.17	99.24	53.37	53.98	36.96	57.08	99.01	55.92
CAD-Unet	70.71	54.69	65.43	99.04	68.28	59.61	42.47	65.74	98.85	65.74
Ours	72.68	57.08	76.28	98.22	75.35	63.29	46.30	69.32	98.95	72.99

Table 3: COVID-19 Dataset Comparison Experiment Results

4.4. Experimental Results

4.4.1 Comparative Experiments on COVID-19 Datasets

On the merged dataset combining Dataset 1 and Dataset 2, the proposed FreqMamba-UNet model is compared with U-Net, U-Net++, Attention-UNet, Inf-Net, AnamNet, SCOAT-Net, D-TrAttUnet, and CAD-Unet. Given that efficient model design is crucial for practical clinical applications in medical image segmentation, this paper also provides the number of parameters and computational cost of the models when processing batches of 6 images. The comparative experimental results are shown in Table 3.

Model	Params (M)	FLOPs (G)
U-Net	7.85	10.80
U-Net++	36.63	106.17
Attention-Unet	34.88	51.02
Inf-Net	17.27	30.77
AnamNet	4.63	19.42
SCOAT-Net	10.21	30.10
D-TrAttUnet	103.78	41.24
CAD-Unet	15.06	22.68
Ours	17.34	22.87

Table 4: COVID-19 Dataset Resource Consumption Results

Experimental results demonstrate that our model achieves outstanding performance in the COVID-19 CT multi-lesion segmentation task. It ranks first in both F_1 score (72.68) and recall (75.35) for ground-glass opacity (GGO) segmentation, surpassing the second-best model by over 2.34 points. For CON segmentation, it achieved an F_1 score of 63.29 and an IoU of 46.30, surpassing the second-best model by 3.68 and 3.83 points respectively, demonstrating significantly improved boundary segmentation accuracy. Moreover,as shown in Table 4, our model's parameter count (17.34M) is only 47% of U-Net++ and 17% of D-TrAttUnet, while its computational cost (22.87G FLOPs) is substantially lower than most comparison models, achieving a balance between high performance and lightweight design. To showcase our model's performance in COVID-19 CT lesion segmentation, Figure 5 compares predicted masks and ground truth (GT) for GGO/CON regions.

To highlight the core comparative analysis and ensure image readability, Figure 6 presents the performance curves of four state-of-the-art benchmark models (AnamNet, SCOAT-Net, D-TriAttUnet, CAD-Unet) alongside our proposed model for medical image segmentation tasks. These curves track the evolution of F1 scores and IoU scores for GGO and CON during training iterations. Results indicate that all models demonstrate performance improvements with increasing training iterations, converging after 80–100 iterations. Our model not only achieves superior final performance but also exhibits faster learning rates and smoother performance curves during the initial training phase (first 20 iterations), reflecting exceptional optimization characteristics and training stability.

4.4.2 Synapse Dataset Comparison

To validate the model's generalization capability, this study conducted experiments on the publicly available Synapse multi-organ segmentation dataset despite limited data scale (restricted annotated data for GGO and CON). Specifically, the aorta and pancreas structures from Synapse were selected for validation. The aorta typically appears as a homogeneous soft tissue density structure in CT images,exhibiting low contrast between its boundaries and surrounding tissues. This resembles the blurred boundaries and weak contrast features trained by the GGO decoder. To validate whether the GGO decoder can generalize to other non-nodular anatomical structures with similar imaging characteristics, the aorta was analyzed using the GGO decoder. The pancreas typically presents as a sharply defined anatomical structure in CT scans,exhibiting local con-

Model	Aorta					Spleen				
	F1	IoU	Rec	Spec	Sens	F1	IoU	Rec	Spec	Sens
U-Net	73.00	57.49	67.14	99.97	67.14	82.14	69.70	80.16	99.91	80.16
U-Net++	77.90	63.80	75.01	99.97	75.01	84.04	72.48	82.94	99.92	82.94
Attention-Unet	71.85	56.06	65.94	99.97	65.94	82.49	70.20	81.01	99.91	81.01
Inf-Net	73.70	58.35	67.97	99.97	67.97	82.56	70.30	78.44	99.93	78.44
AnamNet	51.34	34.53	37.97	99.98	37.97	65.30	48.48	65.05	99.80	65.05
SCOAT-Net	76.70	62.21	71.60	99.98	71.60	85.41	74.54	82.38	99.94	82.38
D-TrAttUnet	67.83	51.32	57.99	99.98	57.99	78.44	64.53	70.05	99.95	70.05
CAD-Unet	72.86	57.31	64.37	99.98	64.37	82.60	70.36	78.81	99.93	78.81
Ours	79.95	66.60	83.41	99.96	83.41	86.43	76.10	89.48	99.88	90.59

Table 5: Synapse Dataset Comparison Experiment Results

trast characteristics similar to the high contrast and well-defined boundaries used to train the CON decoder. To test the CON decoder's segmentation capability on non-tumorous, normally structured organs with clear boundaries, the pancreas was segmented using the CON decoder. Comparative experimental results are shown in Table 5.

Experimental results demonstrate that the Ours model achieves outstanding performance in multi-class segmentation tasks for the aorta and spleen. Specifically, for aortic segmentation, our model achieves an F1 score of 79.95, surpassing the second-best model SCOAT-Net (76.70) by 3.25 points. Its recall rate (83.41) significantly outperforms all comparison models, demonstrating robust aortic lesion detection capabilities. For spleen segmentation, our model's F1 score (86.43) and IoU (76.10) surpassed the second-place SCOAT-Net by 1.02 and 1.56 points respectively, while its recall (89.48) demonstrated a significant advantage, reflecting breakthroughs in spleen boundary segmentation accuracy. Overall, this approach enhances sensitivity and segmentation accuracy for aortic and splenic lesion detection while maintaining high specificity through optimized network architecture and training strategies. To visually demonstrate the model's multi-class segmentation performance on the Synapse dataset, Figure 7 presents comparisons between predicted masks and ground truth (GT) labels for selected aortic and splenic segmentation tasks.

4.4.3 Ablation Studies

To evaluate the effectiveness of the proposed model and its components, ablation experiments were conducted to investigate the impact of different components on model performance. Following advanced research methodologies, we established "GD+CD" (dual-decoder architecture, where GD and CD correspond to GGO and CON decoders, respectively) as the baseline. We then progressively introduced EFM, WFM, and Mamba bridges, examining performance under scenarios such as single-decoder and cross-decoder task adaptation. We conducted multi-class segmentation task tests on a dataset combining Dataset 1 and Dataset 2 to clarify the value of each component. The experimental results are shown in Table 6.

Using "GD+CD" as the baseline, we validated the dual-decoder architecture's foundational adaptability for both tasks. After introducing the edge modules (EFM) ("GD+CD+EFM"), the GGO task's F1 score improved to 70.74, indicating that (EFM) captures edge details to enhance (GGO) feature representation. However, the CON task's F1 score decreased to 60.40, reflecting EFM's interference with CON feature adaptation, necessitating further optimization for synergy. Further adding WFM frequency domain branch ("GD+CD+WFM+EFM"), GGO task F1 reached 71.07, demonstrating frequency domain information enhances feature dimensions. However, CON task F1 dropped to 57.98, IoU was 40.82, Sens was 64.16 , indicating insufficient feature adaptability between WFM and CON decoders, highlighting the need for component-specific collaborative design. Introducing the Mamba bridge ("GD+CD+Mamba+EFM") yields CON task F1 of 59.97, IoU of 42.83, and Sens of 69.80. While metrics show some fluctuation, this validates the bridge's enhancement of feature interaction.

Using only the GGO decoder ("GD+Mamba+WFM +EFM"), the GGO task achieved an F1 score of 71.32, IoU of 55.43, and Sensitivity of 73.47. This demonstrates that when a single decoder focuses on a specific task, it can deeply adapt to target features alongside supporting modules. However, when the CON decoder operates independently ("CD+Mamba+WFM+EFM"), its F1 score drops to 59.02, IoU to 41.87, and Sensitivity to 64.38—all significantly lower than the dual-decoder architecture. These results not only validate the advantage of the dual decoders' "shared feature foundation + task-specific customization" coupling mechanism but also confirm the core reason for insufficient cross-decoder adaptation: the significant radio-

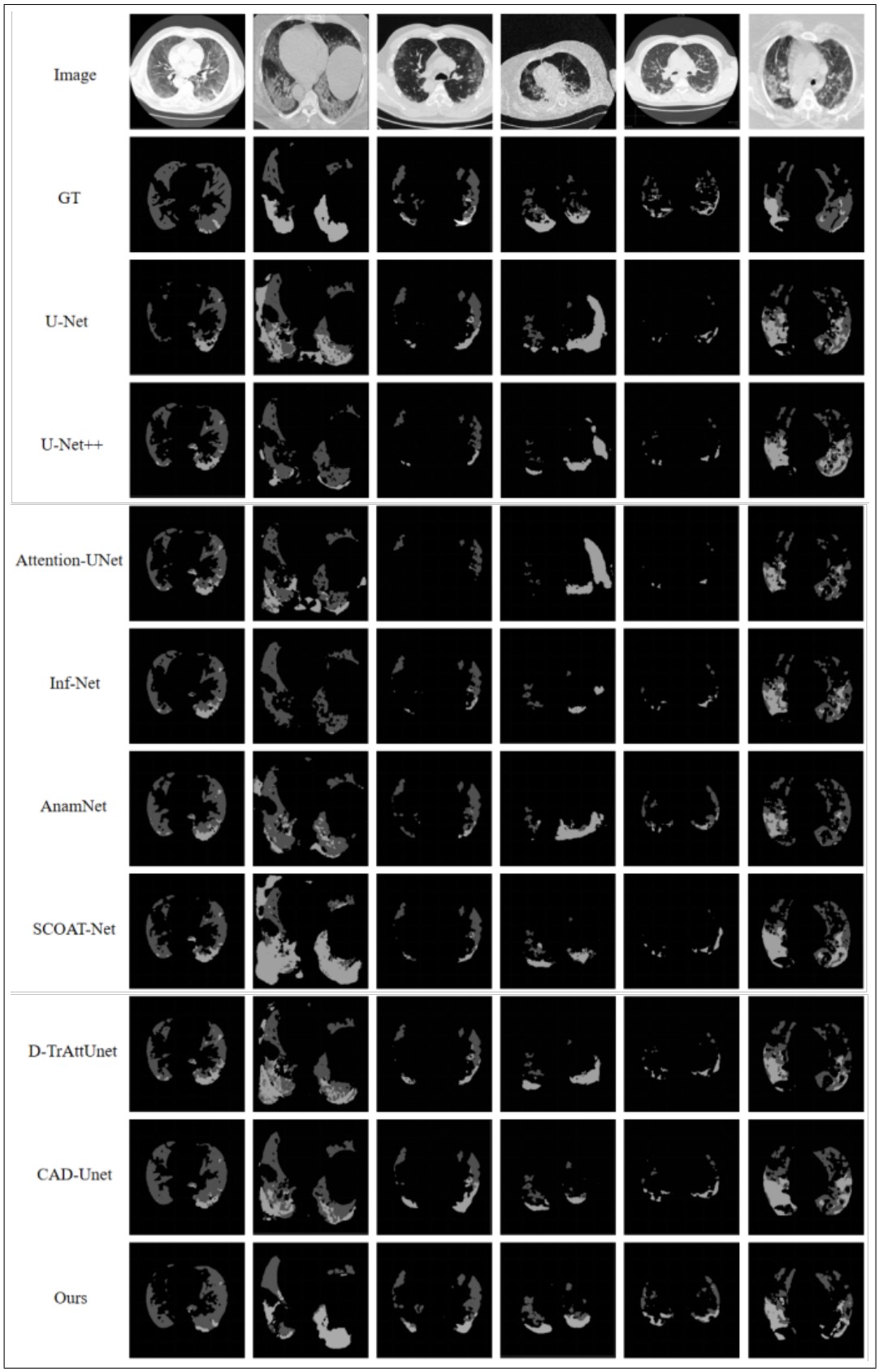

Figure 5: Experimental comparison on the COVID-19 dataset.

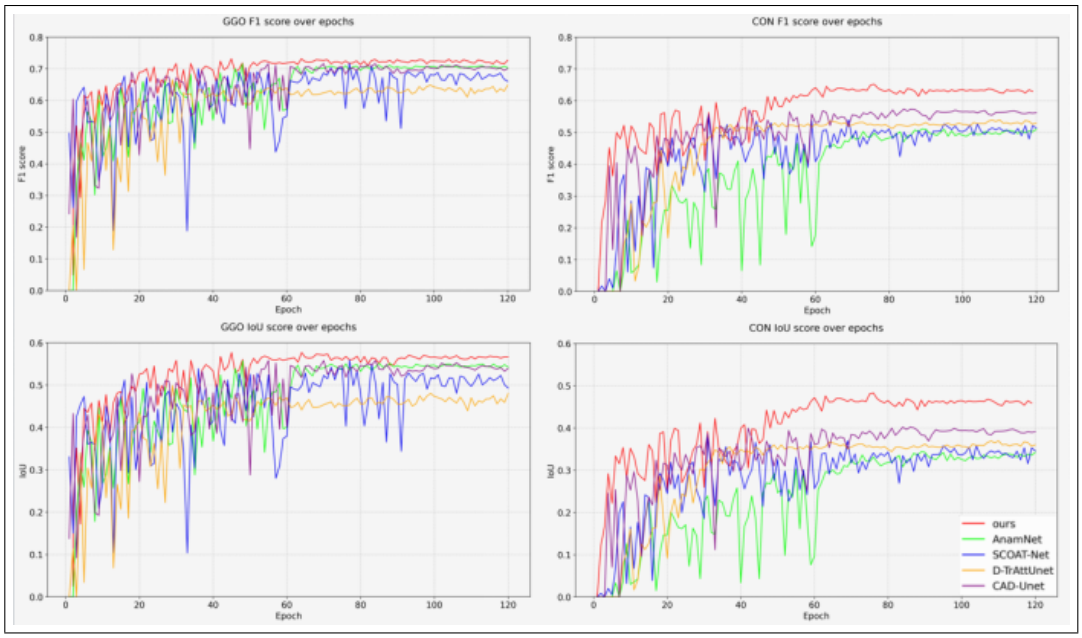

Figure 6: Performance Curve Comparison for COVID-19 Dataset Experiments.

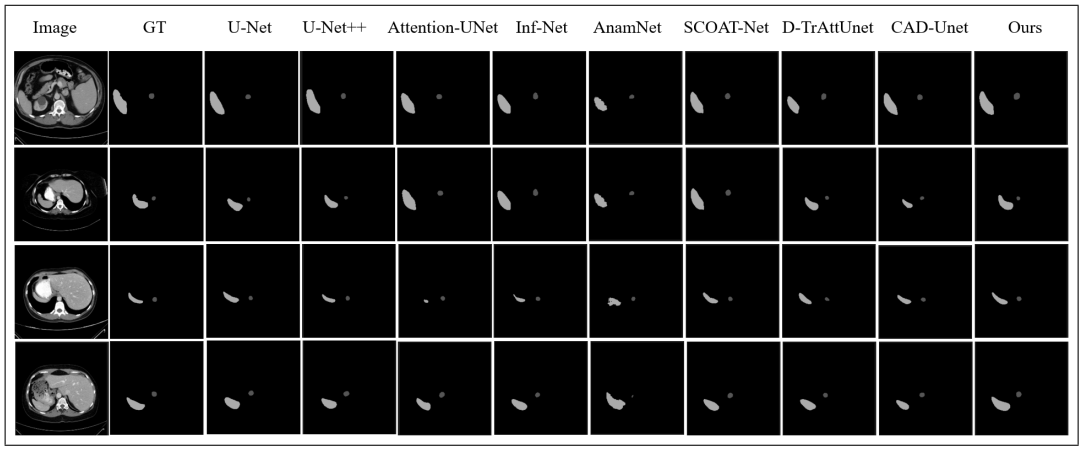

Figure 7: Experimental Comparison on the Synapse Dataset.

logical differences between GGO and CON make it difficult for a single decoder to accommodate both heterogeneous features simultaneously. In contrast, the dual-decoder design achieves precise adaptation through collaboration, ensuring robust multi-task segmentation performance.

Attempting ”cross-decoder tasks”—using the GGO decoder to segment CON and the CON decoder to segment GGO (”CD+GD+Mamba+WFM+EFM”)—yielded GGO task F1 of 72.43, IoU of 56.77, and Sens of 74.74 . However, for the CON task , CON resulted in F1 dropping to 60.55, IoU of 43.42, Sensitivity to 61.20 . Performance significantly deteriorated, indicating strong decoder-task coupling. Cross-task deployment compromises feature adaptability, further validating the value of dual decoders.

5. Conclusions and Discussion

The FreqMamba-UNet architecture proposed in this study achieves a performance breakthrough in the segmentation task of GGO and CON lesions in COVID-19 CT images through frequency-domain-spatial-domain feature fusion, global-local feature collaborative modeling, and a dual-decoder customized segmentation strategy. Experimental results demonstrate that this architecture outperforms classical models such as U-Net and Attention U-Net, as well as state-of-the-art architectures like CAD-Unet, in multi-class segmentation tasks .

Structure	GGO			CON		
	F1	IoU	Sens	F1	IoU	Sens
GD+CD	70.38	54.29	72.23	61.70	44.61	70.31
GD+CD+EFM	70.74	54.73	70.37	60.40	43.27	69.02
GD+CD+WFM+EFM	71.07	55.12	73.39	57.98	40.82	64.16
GD+CD+Mamba+EFM	70.08	53.94	73.81	59.97	42.83	69.80
GD+Mamba+WFM+EFM	71.32	55.43	73.47	60.65	43.53	70.17
CD+Mamba+WFM+EFM	71.47	55.61	71.60	59.02	41.87	64.38
CD+GD+Mamba+WFM+EFM	72.43	56.77	74.74	60.55	43.42	61.20
Ours	72.68	57.08	75.35	63.29	46.30	72.99

Table 6: Ablation Experiment Results

Specifically, the parallel design of the multi-level wavelet branches and the CNN encoder enables synergistic extraction of high-frequency edge information and spatial texture features in CT images. The Mamba-based multi-scale feature bridge overcomes the limitation of traditional skip connections, which only capture local correlations at the same level. Through sequential modeling capabilities, it effectively captures long-range spatial dependencies in diffuse lesions while narrowing the semantic gap between high- and low-level features. Addressing the radiological differences between the two lesion types, the GGO decoder enhances boundary details via CBAM attention, while the CON decoder improves localization accuracy in high-density regions through an Edge Enhancement Module (EEM) and Spatial Pyramid Pooling (SSP), resolving the inadequacy of a single decoder for heterogeneous lesions.

Despite improvements in segmentation metrics achieved by FreqMamba-UNet, limitations remain. The study relies on publicly available COVID-19 CT datasets, which feature limited anatomical regions and static slices, resulting in insufficient generalization to complex comorbid scenarios. The dual-path parallel design increases parameters, leading to slower inference speeds compared to lightweight models. Furthermore, the absence of temporal CT data prevents capturing lesion evolution dynamics for disease progression monitoring.

Future work will expand data diversity to enhance robustness in complex scenarios; reduce computational complexity through model pruning and knowledge distillation to compress parameters; design spatiotemporal fusion variants incorporating temporal CT to capture lesion dynamics; and explore multimodal fusion with chest X-rays and clinical indicators to improve synergy between segmentation and disease assessment.

Acknowledgement

This work was supported by Central Leading Local Science and Technology Development Fund(Grant No.YDZISX20231C004). Special thanks to both the team leader and members.

References

[1] M. R. Pereira. Use of covid-19 vaccines for persons aged 6 months: Recommendations of the advisory committee on immunization practices — united states, 2024–2025. *American Journal of Transplantation*, 24(12):2146–2147, 2024. 1

[2] N. R. Miranda, E. A. Simmons, Z. W. M. Li, et al. Cost-utility analysis of covid-19 vaccination strategies for endemic sars-cov-2. *JAMA Network Open*, 8(6):e2515534, 2025. 1

[3] M. Churruca, E. Martínez-Besteiro, F. Couñago, et al. Covid-19 pneumonia: A review of typical radiological characteristics. *World Journal of Radiology*, 13(10):327, 2021. 1

[4] Ayturk Keles, Mustafa Berk Keles, and Ali Keles. Cov19-cnnet and cov19-resnet: Diagnostic inference engines for early detection of covid-19. *Cognitive Computation*, 16(4):1612–1622, 2024. 1

[5] M. Daud, S. N. Goldberg, D. Cohen, et al. Optimal window settings for detection and characterization of ground-glass opacities on computed tomography in covid-19 patients using a simplex algorithm-based approach. *The Israel Medical Association Journal: IMAJ*, 27(5):283–289, 2025. 1

[6] M. S. Kim. Long-term autoimmune inflammatory rheumatic outcomes of covid-19 (vol 177, pg 291, 2024). *Annals of Internal Medicine*, 2024. 1

[7] F. Necker, K. Petkov, K. Engel, et al. Evolution of an acute covid-19 pulmonary infection. *Radiology*, 310(2):e232644, 2024. 1

[8] A. A. Alrashidi, M. Salim, S. Alharthi, et al. Ultrasensitive first derivative synchronous spectrofluorimetric approach for the concurrent quantification of covid-19-2024 treatment combination dextromethorphan and guaifenesin in different matrices: Compliance with greenness and practicality metrics. *Luminescence: Journal of Biological Chemical Luminescence*, 40(4), 2025. 1

[9] Y. Kataoka, N. Tanabe, M. Shirata, et al. Artificial intelligence-based analysis of the spatial distribution of ab-

normal computed tomography patterns in sars-cov-2 pneumonia: association with disease severity. *Respiratory Research*, 25(1):24, 2024. 1

[10] Y. S. Chung, C. Y. Lam, P. H. Tan, et al. Comprehensive review of covid-19: Epidemiology, pathogenesis, advancement in diagnostic and detection techniques, and post-pandemic treatment strategies. *International Journal of Molecular Sciences*, 25(15), 2024. 1

[11] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu. Advances in medical image segmentation: a comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering*, 11(10):1034, 2024. 2

[12] F. Bougourzi, C. Distante, F. Dornaika, et al. Pdatt-unet: Pyramid dual-decoder attention unet for covid-19 infection segmentation from ct-scans. *Medical Image Analysis*, 86:102797, 2023. 2

[13] M. E. Rayed, S. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. F. Mridha. Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Informatics in Medicine Unlocked*, page 101504, 2024. 2

[14] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, Cham, 2015. Springer International Publishing. 2

[15] Y. Zhu, C. Zhao, J. Wang, et al. Couplenet: Coupling global structure with local parts for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4126–4134, 2017. 2

[16] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, et al. Unet++: A nested u-net architecture for medical image segmentation. In *International Workshop on Deep Learning in Medical Image Analysis*, pages 3–11, Cham, 2018. Springer International Publishing. 2

[17] O. Oktay, J. Schlemper, L. L. Folgoc, et al. Attention u-net: Learning where to look for the pancreas, 2018. 2

[18] D. P. Fan, T. Zhou, G. P. Ji, et al. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637, 2020. 2

[19] P. Naveen, D. Aveen, H. J. Bjorke, et al. Anam-net: Anamorphic depth embedding-based lightweight cnn for segmentation of anomalies in covid-19 chest ct images. *IEEE Transactions on Neural Networks and Learning Systems*, 32(3):932–946, 2021. 2

[20] S. Zhao, Z. Li, Y. Chen, et al. Scoat-net: A novel network for segmenting covid-19 lung opacification from ct images. *Pattern Recognition*, 119:108109, 2021. 2

[21] F. Bougourzi, C. Distante, F. Dornaika, et al. D-trattunet: dual-decoder transformer-based attention unet architecture for binary and multi-classes covid-19 infection segmentation, 2023. 2

[22] Y. Dang, W. Ma, X. Luo, et al. Cad-unet: A capsule network-enhanced unet architecture for accurate segmentation of covid-19 lung infections from ct images. *Medical Image Analysis*, page 103583, 2025. 2

[23] T. T. Showrav and M. K. Hasan. Hi-gmisnet: generalized medical image segmentation using dwt based multilayer fusion and dual mode attention into high resolution pgan. *Physics in Medicine Biology*, 69(11):115019, 2024. 2

[24] S. Resmi, R. P. Singh, and K. Palaniappan. Automated cervical cytology image cell segmentation using enhanced multiresunet with dct and spectral domain attention mechanisms. *IEEE Access*, 2024. 2

[25] W. Wang, J. Wang, C. Chen, et al. Fremim: Fourier transform meets masked image modeling for medical image segmentation–supplementary material, 2024. 2

[26] Y. Rao, W. Zhao, Z. Zhu, et al. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021. 3

[27] L. Jiang, B. Dai, W. Wu, et al. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13919–13929, 2021. 3

[28] L. Liu, J. Liu, S. Yuan, et al. Wavelet-based dual-branch network for image demoiréing. In *European Conference on Computer Vision*, pages 86–102, Cham, 2020. Springer International Publishing. 3

[29] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023. 6

[30] Radiologists. Covid-19 ct-scans segmentation datasets, 2019. Available at: http://medicalsegmentation.com/covid19/. Last visited: August 18, 2021. 8