RF-DTR: A Multi-Stage DCT Token Regression Network for Progressive Rib Fracture Mask Refinement

ShouYu Chen[1] , **Liang Hu**[1*] , **JunTao Wang**[1] , **Usman Naseem**[2] , **ZhongYuan Lai**[3] , **Qi Zhang**[1]
[1]Tongji University
[2]Macquarie University
[3]Shanghai Ballsnow Intelligent Technology Co. Ltd

Abstract

Rib fracture patterns are key indicators of trauma severity. Detecting and locating these fractures is a critical yet time-consuming task, especially in 3D imaging, due to their minute size and irregular geometries. Existing voxel-based spatial methods fail to capture frequency-domain variations inherent in imaging and do not replicate the progressive refinement process used by clinicians during manual annotation, leading to suboptimal results. We propose a novel regression network, RF-DTR, incorporating a gated regressor mechanism and operating entirely in the frequency domain to address these challenges. Specifically, we present an innovative spatial-frequency transform applied to volumes and corresponding masks. Furthermore, we introduce a Mahalanobis regularization technique to enhance the model and learn high-frequency DCT components relevant to clinical tasks. Finally, a multi-stage penalty is proposed to improve the confidence of the prediction. Extensive experiments confirm our method's superiority in handling complex, sparsely annotated medical imaging datasets.

1 Introduction

Rib fracture detection presents a significant challenge due to the need to accurately identify small, hollow lesions with intricate geometries in large 3D voxel spaces. The scarcity of positive samples further complicates tasks such as classification [Lindsey *et al.*, 2018; Cheng *et al.*, 2019; Huang *et al.*, 2023], segmentation [Yao *et al.*, 2021; Wu *et al.*, 2021], and object detection [Yao *et al.*, 2021; Yu *et al.*, 2022]. While generative anomaly detection models trained exclusively on healthy samples have demonstrated potential in clinical applications, their reliance on one-class setting often undermines their robustness. These models identify high reconstruction errors from out-of-distribution (OOD) data [Fernando *et al.*, 2021] as pixel-level anomalies. However, recent studies [Lu *et al.*, 2023; You *et al.*, 2022; Zhang *et al.*, 2023] indicate that generative models can inadvertently reconstruct OOD samples with high fidelity, leading to false negatives in anomaly detection. Although semantic segmentation-based methods typically achieve higher accuracy, their ability to capture spatial details remains limited.

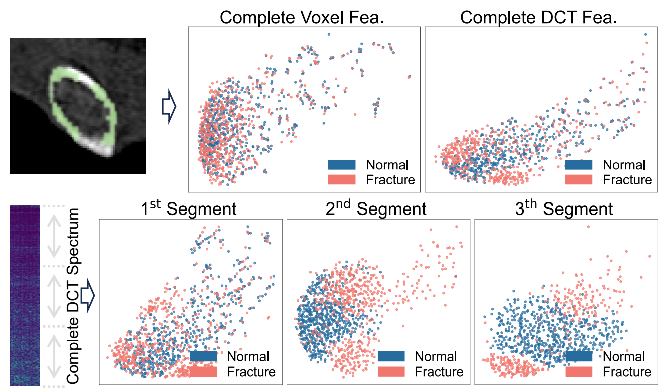

Figure 1: UMAP visualization illustrates that the frequency-domain (DCT) representation better distinguishes CT volumes than the spatial-domain (voxel) representation. The distinct frequency bands within the three sub-spectrums exhibit varying degrees of separability, motivating frequency-aware modeling. Further details are provided in Supplementary Materials §A.

Despite the progress in rib fracture detection, existing methods inadequately leverage the high-frequency characteristics of fractures in the frequency domain. For instance, FracNet [Jin *et al.*, 2020] employs a sliding window strategy with 3D U-Net variants for patch-wise learning, establishing a foundational pipeline for subsequent works [Wu *et al.*, 2021; Yao *et al.*, 2021]. However, as illustrated in Figure 1, our empirical analysis demonstrates that rib fractures exhibit greater discriminability in the frequency domain, particularly in the high-frequency components. This observation underscores the potential advantages of frequency-domain modeling. Clinically, annotating small and hollow rib fractures is an iterative process that relies on human expertise for refinement and quality assurance. While this well-established method produces reliable results, it is often labor-intensive and susceptible to human error. An automated workflow that accurately replicates this process would be highly desirable, offering the dual benefits of ensuring robust quality and maintaining interpretability.

Our study targets two fundamental challenges: (1) detecting small and hollow fractures and (2) designing interpretable

*Corresponding author. Email: lianghu@tongji.edu.cn

models that emulate clinicians' annotation process, ensuring a more transparent and intuitive decision-making framework. Recent research [Xu *et al.*, 2020; Wen *et al.*, 2022] suggests that frequency-domain representations can serve as effective feature embeddings. Since fine structural details in CT images primarily manifest as high-frequency signals, accurately capturing these components is crucial. However, previous studies have demonstrated that deep neural networks (DNNs) [Xu *et al.*, 2019], convolutional neural networks (CNNs) [Xu *et al.*, 2020], and Transformers [Wang *et al.*, 2022; Piao *et al.*, 2024] often exhibit insensitivity to high-frequency information, limiting their ability to learn fine-grained structures. Empirical results in Figure 3(a) further illustrate this challenge in rib fracture detection. To mitigate this issue, we introduce a frequency-domain regularization to enhance high-frequency learning. Last but not least, we propose a multi-stage penalty mechanism that progressively refines predictions, closely mimicking expert annotation workflows.

We formulate rib fracture detection as an instance segmentation task and propose a model that learns from spatial-frequency transformed input images and corresponding output masks. Our model integrates the discrete cosine transform (DCT) [Ahmed *et al.*, 1974] and employs a progressive mask refinement strategy. Inspired by hierarchical designs in computer vision, we introduce frequency-domain regression modules and a conditional penalty term to improve mask prediction. Our key contributions are as follows:

- We conduct a frequency-domain analysis revealing that existing rib fracture detection models suffer from a critical limitation: insufficient high-frequency learning.
- To address this issue, we propose an encoder-only DCT token regression network that operates entirely in the frequency domain, significantly enhancing sensitivity to fine-grained structures.
- We introduce a novel Mahalanobis regularization to enhance high-frequency learning. Moreover, we improve our method's interpretability for a transparent decision-making process by a unified cross-stage penalty.
- We validate our approach through experiments on a public CT benchmark and our curated dataset, demonstrating superior performance in segmentation and detection tasks, surpassing state-of-the-art (SOTA) methods.

2 Related Works

2.1 Classical Methods: Challenges and Advances

Deep learning models achieve high recall but often exhibit higher false positive rates than radiologists [Zhang *et al.*, 2021]. Existing methods [Chen *et al.*, 2017; Jin *et al.*, 2020] struggle to capture fine-grained fracture features, limiting their effectiveness in precise localization. To mitigate this issue, some approaches integrate detection and segmentation techniques. For instance, [Wu *et al.*, 2021] combines 2D Faster R-CNN for detection with 3D U-Net for segmentation, while a three-stage pipeline [Yao *et al.*, 2021] sequentially performs rib segmentation, localization, and fracture classification. Cascade-based framework [Zhang *et al.*, 2021] leverages the Foveal network [Brosch and Saalbach, 2018] and Faster R-CNN to refine rib masks and detect fracture candidates. To better capture rib morphology, SA-FracNet [Cao *et al.*, 2023] employs contrastive learning to address the elongated and inclined rib structure, while CCE-Net [Gao *et al.*, 2022] adopts feature fusion. Additionally, SA-FracNet introduces a shape-aware loss function based on Signed Distance Maps to improve fracture delineation. In contrast, our method employs an encoder-only architecture that effectively captures detailed fracture features by frequency modeling.

2.2 Frequency-Informed Learning

Spatial voxel details correspond to high-frequency components. Recent studies have revealed inherent learning biases in neural networks when analyzed from a frequency perspective. The Frequency Principle [Xu *et al.*, 2019] suggests that DNNs inherently prioritize low-frequency signals. CNNs exhibit a strong preference for low-frequency components [Xu *et al.*, 2020], and similar tendencies have been observed in Transformers [Wang *et al.*, 2022; Park and Kim, 2022; Tian *et al.*, 2023; Guo *et al.*, 2023; Piao *et al.*, 2024]. Fast Fourier Transform (FFT) has gained widespread application. GFNet [Rao *et al.*, 2023] utilizes FFT for global feature extraction, while Frequency-Adaptive Dilated Convolutions [Chen *et al.*, 2024] dynamically adjust dilation rates based on local frequency characteristics. FFT-based token mixers provide a computationally efficient alternative to self-attention [Tatsunami and Taki, 2024], and Fourier regularization mitigates high-frequency artifacts [Xu *et al.*, 2019]. Additionally, MDTNet [Zhao *et al.*, 2024] enforces prediction-ground truth alignment via Fourier constraints, improving tasks such as image reconstruction [Wang *et al.*, 2018; Jiang *et al.*, 2021] and enhancement [Greenspan *et al.*, 2000; Fuoli *et al.*, 2021]. Building on these insights, we propose an efficient DCT-based architecture incorporating a multi-stage penalty mechanism to refine mask predictions hierarchically.

2.3 DCT-Based Frequency-Domain Analysis

DCT is a real-valued frequency transformation that provides greater computational efficiency than FFT [Pan *et al.*, 2022]. It has been integrated into neural networks to replace convolutions with DCT-based perceptrons, reducing computational costs [Pan *et al.*, 2022]. Additionally, a frequency-channel selection method [Xu *et al.*, 2020] eliminates non-salient DCT components without compromising accuracy. In instance segmentation, DCT-Mask [Shen *et al.*, 2021] encodes binary masks into compact vectors, reducing training costs, while PatchDCT [Wen *et al.*, 2022] refines this approach for precise boundary segmentation. In channel attention, DCT frequency analysis [Qin *et al.*, 2021] models channel representation as a compression process. Furthermore, DCTNet [Zhao *et al.*, 2022] reconstructs high-resolution images from low-resolution depth maps by capturing both shared and modality-specific features. Despite these advances, existing methods such as DCT-Mask and PatchDCT [Wen *et al.*, 2022] rely on L1 loss and overlook critical frequency-domain properties such as scale and correlation. To address this, we introduce a metric learning approach based on DCT that de-correlates frequency features while ensuring scale alignment.

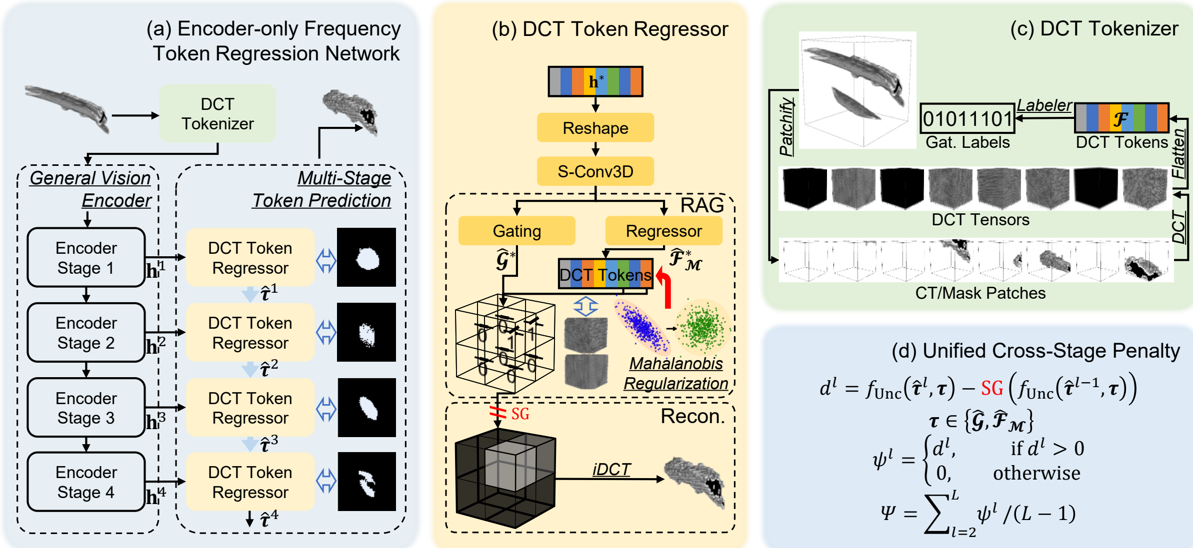

Figure 2: (a) Overview of the proposed encoder-only regression framework, where the Multi-Stage Token Prediction is designed to integrate seamlessly with any General Vision Encoder. (b) The DCT Token Regressor reconstructs hidden features from each stage into a predicted fracture mask. (c) Illustration of transforming a 3D sub-volume into DCT tokens, with an optional gating label generation process. (d) The proposed Unified Cross-Stage Penalty is hierarchically compatible with both the gating and regressor modules. "Gat." denotes the gating module, "RAG" represents the Regressor-After-Gating mechanism, and "SG" refers to the stop-gradient operation.

3 Method

3.1 Problem Description

We aim to detect and segment rib fractures in CT volumes using a sliding window approach, framing the problem as a 3D instance segmentation task. Formally, the training dataset is defined as $\mathcal{D} = \{(\mathcal{V}_i, \mathcal{M}_i)\}_{i=1}^{N}$, where each image window $\mathcal{V}_i \in \mathbb{R}^{D\times H\times W}$ represents a cropped 3D sub-volume extracted from the CT volume, and its corresponding ground truth mask $\mathcal{M}_i \in \{0,1\}^{D\times H\times W}$ indicates the presence of rib fractures at a voxel level. During inference, the model takes $\mathcal{V}_i$ as input and predicts a segmentation mask $\hat{\mathcal{M}}_i$, where each of its voxels represents the probability of belonging to a fractured region, as illustrated in Figure 2(a).

3.2 DCT Tokenizer

Spatial-Frequency Transformation. Inspired by the JPEG compression standard [Wallace, 1992] and related methods [Shen *et al.*, 2021; Wen *et al.*, 2022] in 2D computer vision, we design a DCT Tokenizer tailored for volumetric $\mathcal{V}$ and $\mathcal{M}$, as illustrated in Figure 2(c). The tokenizer projects spatial patches into the frequency ones. For $\mathbf{t} \in \{\mathcal{V}, \mathcal{M}\}$, this process is defined as $\mathcal{F}_{\mathbf{t}} = \text{Tokenizer}(\mathbf{t})$. Specifically, the data is first split into non-overlapping patches along all dimensions:

$$\mathcal{B}_{\mathbf{t}} = \text{Patchify}(\mathbf{t}) \in \mathbb{R}^{\frac{D}{B}\times\frac{H}{B}\times\frac{W}{B}\times B\times B\times B}. \quad (1)$$

Here, B denotes the patch size, set to 8 to align with the JPEG standard. The terms $\frac{D}{B}, \frac{H}{B}, \frac{W}{B}$ represent the number of patches along each axis. Each patch undergoes DCT-II encoding as $\mathcal{F}_{\mathbf{t}} = \text{DCT-II}(\mathcal{B}_{\mathbf{t}})$ (details in Supplementary Materials §B). Notably, $B = 8$ aligns with the configuration of the 3D ViT encoder, allowing $\mathcal{F}_{\mathcal{V}}$ to be used seamlessly as input to the ViT's projection layer without any modifications.

The task is then reformulated as patch-level regression, targeting $\mathcal{F}_{\mathcal{M}} = \{\mathcal{F}_{\mathcal{M},p}\}_p$, $\mathcal{F}_{\mathcal{M},p}$ denotes a DCT mask patch, and $p = (x, y, z)$ specifies the patch indices along each axis: $x \in \{1, \cdots, \frac{D}{B}\}, y \in \{1, \cdots, \frac{H}{B}\}, z \in \{1, \cdots, \frac{W}{B}\}$.

Gating Label Definition. Each patch is labeled as state 0, 1, or 2, representing non-fractured, partially fractured, and fully fractured samples. Following PatchDCT [Wen *et al.*, 2022], the label is determined by the DCT tensor's direct current component (DCC), which reflects its overall intensity: a DCC of 0 corresponds to state 0, $\frac{B^2}{\sqrt{2}}$ to state 2, and otherwise to state 1. The states for all patches of current $\mathcal{M}$ are collectively denoted as $\mathcal{G} = \{\mathcal{G}_p\}_p$.

3.3 Multi-Stage Token Prediction

The ViT encoder produces hierarchical hidden features represented as $\mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3, \mathbf{h}^4 = \text{ViT}(\mathcal{F}_{\mathcal{V}})$. For each hierarchical level $l \in \{1, 2, 3, 4\}$, $\mathbf{h}^l \in \mathbb{R}^{B^3\times d_{\text{model}}}$, d_{model} is the dimension of the ViT patch vector. RF-DTR is composed of three core components: (1) a DCT Token Regressor with Regressor-After-Geting (RAG) mechanism, (2) a frequency-domain regularization, and (3) a multi-stage conditional penalty. Patch indices are omitted in the subsequent descriptions for clarity.

DCT Token Regressor. As illustrated in Figure 2(b), the output $\mathbf{h}^l$ from the l-th ViT stage is reshaped first to restore its spatial dimensions:

$$\mathbf{h}^l = \text{Reshape}(\mathbf{h}^l) \in \mathbb{R}^{d_{\text{model}}\times B\times B\times B}. \quad (2)$$

We present the RAG mechanism, designed to successively predict the gating state and the DCT token for each patch, denoted as $\hat{\mathcal{G}}^l, \hat{\mathcal{F}}^l_{\mathcal{M}} = \text{RAG}(\mathbf{h}^l)$. The module begins by applying stacked 3D convolutions (S-Conv3D) for feature extraction, resulting in $\mathbf{h}^l_{\text{share}} = \text{S-Conv3D}(\mathbf{h}^l)$. Subsequently, the gating and regression tasks are executed upon $\mathbf{h}^l_{\text{share}}$:

$$\hat{\mathcal{G}}^l = \text{Conv3D}(\mathbf{h}^l_{\text{share}}) \in \mathbb{R}^{\frac{D}{B} \times \frac{H}{B} \times \frac{W}{B}}, \tag{3}$$

$$\hat{\mathcal{F}}^l_{\mathcal{M}} = \text{Conv3D}(\mathbf{h}^l_{\text{share}}) \in \mathbb{R}^{\frac{D}{B} \times \frac{H}{B} \times \frac{W}{B} \times B^3}. \tag{4}$$

For $\hat{\mathcal{F}}^l_{\mathcal{M}}$, the last dimension corresponds to a flattened cubic structure encoding DCT tokens. As shown in Table 4, the proposed RAG is an effective and important design. However, relying solely on $\mathbf{h}^4$ may lead to blurred predictions, particularly in complex boundaries. To compensate for this limitation, we propose a multi-stage design optimized using the gating loss $\mathcal{L}_{\text{Gat}}$ and regression loss $\mathcal{L}_{\text{Reg}}$. The hierarchical loss function at a single stage l is defined as follows:

$$\mathcal{L}^l = \frac{\sum_p \mathcal{L}_{\text{Gat}}(\hat{\mathcal{G}}^l_p, \mathcal{G}_p)}{|\mathcal{G}|} + \frac{\sum_{p \in \hat{\mathcal{C}}_{\text{PF}}} \mathcal{L}_{\text{Reg}}(\hat{\boldsymbol{v}}^l_p, \boldsymbol{v}_p)}{|\hat{\mathcal{G}}_{\text{PF}}|}. \tag{5}$$

Here, $\hat{\boldsymbol{v}}^l_p = \text{vec}(\hat{\mathcal{F}}^l_p)$ and $\boldsymbol{v}^l_p = \text{vec}(\mathcal{F}_p)$ represent the flattened predicted and ground truth DCT tokens at p in $\hat{\mathcal{F}}_{\mathcal{M}}$ and $\mathcal{F}_{\mathcal{M}}$. The set $\hat{\mathcal{G}}_{\text{PF}} = \{p \mid \hat{\mathcal{G}}^l_p = 1, \hat{\mathcal{G}}^l_p \in \hat{\mathcal{G}}^l\}$ identifies the indices of patches classified as partially-fractured. We employ shared weights for RAG modules across all stages, supported by the fact that the hierarchical features $\mathbf{h}^l$ maintain identical spatial scale and target. Under this design, deeper features $\mathbf{h}^l$ can be viewed as conditional encodings of $\mathbf{h}^{l-1}$, capturing progressively refined representations.

Mahalanobis-Regularized Regression Loss. Figure 4(a) empirically reveals that the naive hybrid Transformer-CNN architecture is disproportionately sensitive to low-frequency signals, leading to the loss mainly concentrated in high-frequency components. To mitigate this imbalance, we improve the L1 regression loss by introducing the Mahalanobis regularization, enabling scale normalization and feature decorrelation across different frequencies:

$$D_M(\hat{\boldsymbol{v}}^l_p, \boldsymbol{v}_p) = \sqrt{(\hat{\boldsymbol{v}}^l_p - \boldsymbol{v}_p)^\top \Sigma^{-1} (\hat{\boldsymbol{v}}^l_p - \boldsymbol{v}_p)}. \tag{6}$$

$\mathbf{\Sigma} \in \mathbb{R}^{B^3 \times B^3}$ is a learnable symmetric and positive definite matrix, and its inverse $\mathbf{\Sigma}^{-1}$ can be factorized as:

$$\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1} = \mathbf{\Sigma}^{-1}, \tag{7}$$

$\mathbf{Q} \in \mathbb{R}^{B^3 \times B^3}$ is an orthogonal matrix representing rotation, and $\mathbf{\Lambda}$ is a diagonal matrix representing scaling. The DCT token discrepancy between the prediction and ground truth, projected onto the positive definite cone, is defined as $\Delta \boldsymbol{F} = \Delta \boldsymbol{v}^\top \mathbf{Q}$, where $\Delta \boldsymbol{v} = \hat{\boldsymbol{v}}^l_p - \boldsymbol{v}_p$. The self-adaptive weight factor is subsequently noted as follows:

$$\boldsymbol{w}_i = \texttt{sg}\left(\frac{|\Delta \boldsymbol{F}_i|}{\max(\Delta \boldsymbol{F})}\right), \quad i \in [1, B^3], \tag{8}$$

$\texttt{sg}$ denotes the stop-gradient operation. This formulation leads to the Mahalanobis-regularized loss, as shown below, l and p in Equation 6 are omitted for simplicity:

$$\mathcal{L}_{\text{M}}(\hat{\boldsymbol{v}}, \boldsymbol{v}) = \sqrt{\Delta \boldsymbol{v}^\top \mathbf{Q}(\boldsymbol{w}^\top \mathbf{\Lambda})\mathbf{Q}^{-1} \Delta \boldsymbol{v}}, \tag{9}$$

Unified Cross-Stage Penalty. We propose a cross-stage penalty to encourage deeper stages to generate progressively more accurate prediction masks. An uncertainty function, f_{Unc}, quantifies the discrepancy between prediction and the shared ground truth at each stage. The function is applicable for both gating and regressor:

$$d^l_p = f_{\text{Unc}}(\hat{\tau}^l_p, \tau_p), \tau \in \{\mathcal{G}, \mathcal{F}_{\mathcal{M}}\}, \tag{10}$$

where f_{Unc} is instantiated by D_{M} in Equation 6 for the regressor and cross-entropy for gating module, ensuring alignment with the primary loss functions. The penalty that exists between consecutive stages l and $l-1$ at position p is formally defined as follows:

$$\psi^l_p = \begin{cases} d^l_p - \texttt{sg}(d^{l-1}_p), & \text{if } d^l_p > d^{l-1}_p, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

For stage $l \geq 2$, the penalty term is computed as:

$$\psi^l_{\mathcal{M}} = \frac{\sum_{p \in \mathcal{G}_+} \psi^l_{\mathcal{M},p}}{|\mathcal{G}_+|}, \tag{12}$$

where $|\mathcal{G}_+|$ denotes the cardinality of valid positions that are predicted as partially fractured by both consecutive stages. Similarly, $\psi^l_{\mathcal{G}}$ for the gating module can be computed by applying cross-entropy as f_{Unc} over all patches. The overall multi-stage penalty term across all stages is then defined as:

$$\Psi = \frac{1}{L-1} \sum_{l=2}^{L} \left(\psi^l_{\mathcal{G}} + \psi^l_{\mathcal{M}}\right). \tag{13}$$

This formulation ensures a progressive contribution from each stage, encouraging the outputs of deeper stages to achieve consistently greater accuracy than their shallower counterparts. The overall loss function incorporates the regularization and multi-stage penalty as follows:

$$\mathcal{L}_{\text{Overall}} = \mathcal{L}_{\text{Gat}} + \alpha \mathcal{L}_{\text{M}} + \beta \Psi, \tag{14}$$

where $\mathcal{L}_{\text{Gat}}$ denotes the cross-entropy loss for the gating module, and $\mathcal{L}_{\text{M}}$ represents the Mahalanobis-regularized regression loss. The weighting factors α and β control the relative contributions of these terms. This formulation enables multi-stage refinement by mitigating the inherent limitations of L1 loss in frequency-sensitive tasks, enhancing both robustness and precision—particularly in scenarios demanding fine-grained recognition.

4 Experiments and Results

4.1 Dataset and Preprocessing

We conducted experiments using the publicly available *RibFrac* Challenge dataset [Jin *et al.*, 2020]. In our sliding-window framework, the patch size was set to $D \times H \times W = 64 \times 64 \times 64$ voxels, following [Jin *et al.*, 2020]. To enhance rib visibility, bone window normalization was applied using a window width of 1200 Hounsfield units (HU) and a window level of 400 HU. The CT intensities were then linearly

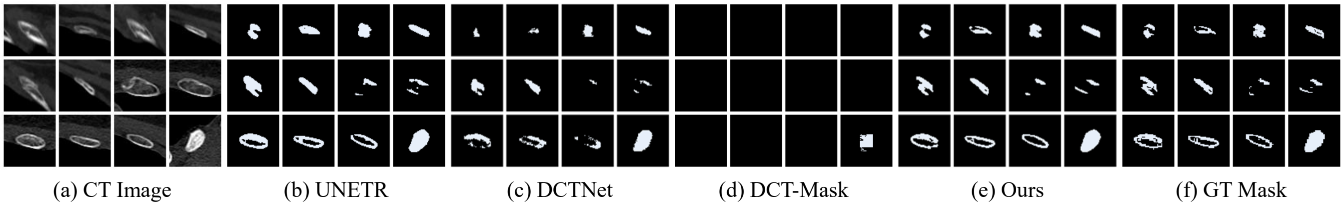

Figure 3: Segmentation results on the *RibFrac-GTRefined* dataset. Our method performs better in identifying small and hollow rib fractures, which pose significant challenges due to their subtle appearance and complex anatomical structures. “GT” denotes the ground truth.

Family	Method	*RibFrac*			*RibFrac-GTRefined*		
		Se. (↑)	Sp. (↑)	F1 (↑)	Se. (↑)	Sp. (↑)	F1 (↑)
$\mathcal{S}$	FracNet [Jin *et al.*, 2020]	0.820	0.901	0.580	0.780	0.892	0.473
	UNETR [Hatamizadeh *et al.*, 2022]	0.861	0.969	0.610	0.853	0.971	0.513
	UNETR++ [Shaker *et al.*, 2024]	<u>0.864</u>	0.903	0.609	0.831	0.912	0.526
$\mathcal{F}$	DCTNet [Xu *et al.*, 2020]	0.860	0.973	0.614	<u>0.855</u>	0.930	0.516
	DCT-Mask [Shen *et al.*, 2021]	0.579	**0.999**	0.312	0.580	**0.998**	0.301
	PatchDCT [Wen *et al.*, 2022]	0.855	0.967	<u>0.626</u>	0.841	0.932	<u>0.630</u>
	Ours	**0.876**	<u>0.981</u>	**0.699**	**0.881**	<u>0.982</u>	**0.714**

Table 1: Quantitative comparison of our method against SOTA approaches for rib fracture detection on both the official and refined *RibFrac* datasets. “$\mathcal{S}$” denotes spatial-domain models, while “$\mathcal{F}$” represents frequency-domain ones. The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>. “Se.” and “Sp.” correspond to sensitivity and specificity, respectively.

Method	# Param. (↓)	FLOPs (↓)
FracNet	**1.40**	94.87
UNETR	19.50	27.62
UNETR++	<u>9.06</u>	**17.49**
DCTNet	19.50	27.62
PatchDCT	20.60	27.76
Ours	21.50	<u>17.50</u>

Table 2: Comparison of our method with SOTA approaches in model size (M) and computational complexity, measured by FLOPs (B).

scaled to $[-1, 1]$ via min-max normalization. Spatial augmentations were applied to training patches, including random perturbation, flipping, and axis permutation. To standardize pixel spacing across different CT scanners, we set the spacing parameter to $(0.6, 0.6, 0.6)$, corresponding to a voxel size of 0.6^3mm^3. Figure S1 in Supplementary Materials illustrates the coarse nature of rib fracture annotations in *RibFrac*, which do not fully capture the complexity of real-world fractures. To address this limitation, we refined *RibFrac* using rib annotations from *RibSeg v2* [Jin *et al.*, 2023], producing a higher-quality dataset with reduced label noise. Detailed information on the refinement process is provided in Supplementary Materials §D.

4.2 Settings

Performance Metrics. We evaluate the model's performance in fracture detection and instance segmentation. Following the FracNet workflow [Jin *et al.*, 2020], we report sensitivity, specificity, and F1-score for detection. For segmentation, we use Intersection over Union (IoU) and Dice coefficient, consistent with previous studies [Jin *et al.*, 2020; Yu *et al.*, 2022; Zhao *et al.*, 2021; Wu *et al.*, 2021]. To assess computational efficiency, we report floating point operations (FLOPs).

Network Configuration and Training Protocol. Our model consists of a DCT token regressor and a four-stage ViT encoder, resulting in a compact and computationally efficient design. The model is randomly initialized and trained for 200 epochs using the AdamW optimizer [Loshchilov, 2017]. The learning rate is set to 1×10^{-4}, with a batch size of 4 and a gradient accumulation factor of 2. The loss function incorporates weighting factors $\alpha = 0.3$ and $\beta = 0.1$. We sample 8 patches per volume to ensure class balance, evenly split between fractured and healthy patches. Fracture patches are extracted around the centroid of each lesion, while healthy patches are selected from symmetrical regions corresponding to the fracture locations and from the spine. During validation and testing, all patches are sampled using a sliding window with a fixed stride. For a fair comparison, we implemented 3D adaptations of the competitor methods DCTNet, DCT-Mask, and PatchDCT, initially designed for 2D inputs. Additional architectural details and implementation specifics are provided in Supplementary Materials §C.

4.3 Quantitative Comparison with SOTAs

Table 1 compares our method against SOTA approaches in spatial and frequency domains. The comparison includes CNN-based architectures and Transformer-CNN hybrid models for the detection task. Our results demonstrate consistent superiority over existing methods on the *RibFrac* dataset. Notably, the performance gap widens on the more challenging *RibFrac-GTRefined* dataset, which better reflects real-world clinical scenarios, highlighting our model's robustness in handling complex anatomical structures. Additionally, frequency-domain-based methods generally achieve

Family	Method	*RibFrac*		*RibFrac-GTRefined*	
		IoU (↑)	Dice Coefficient (↑)	IoU (↑)	Dice Coefficient (↑)
$\mathcal{S}$	FracNet [Jin *et al.*, 2020]	0.532	0.695	0.531	0.694
	UNETR [Hatamizadeh *et al.*, 2022]	0.583	0.737	0.576	0.731
	UNETR++ [Shaker *et al.*, 2024]	0.584	0.738	0.581	0.735
$\mathcal{F}$	DCTNet [Xu *et al.*, 2020]	<u>0.589</u>	<u>0.741</u>	0.592	0.743
	DCT-Mask [Shen *et al.*, 2021]	0.204	0.339	0.205	0.340
	PatchDCT [Wen *et al.*, 2022]	0.531	0.694	<u>0.622</u>	<u>0.767</u>
	Ours	**0.613**	**0.760**	**0.649**	**0.787**

Table 3: Quantitative comparison of our method against SOTA approaches for rib fracture segmentation. "$\mathcal{S}$" denotes spatial-domain models, while "$\mathcal{F}$" represents frequency-domain models. The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>.

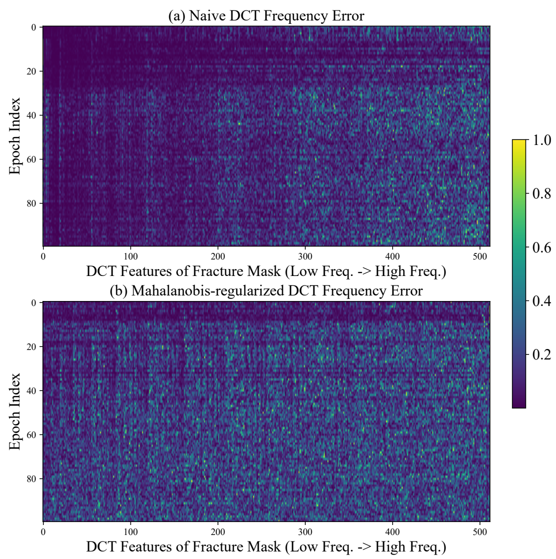

Figure 4: Impact of Mahalanobis regularization in the DCT domain. (a) Without regularization, the model progressively concentrates loss on high-frequency components during training. (b) With regularization, loss distribution remains balanced across all frequency components, improving stability

Input	Output	F1 (↑)	Dice (↑)
$\mathcal{S}$	$\mathcal{S}$	0.513	0.524
$\mathcal{F}$-8	$\mathcal{S}$	0.511	0.532
$\mathcal{F}$-16	$\mathcal{S}$	0.515	0.533
$\mathcal{F}$-8-Norm	$\mathcal{S}$	0.535	0.554
$\mathcal{F}$-16-Norm	$\mathcal{S}$	0.531	0.539
$\mathcal{S}$	$\mathcal{F}$-8	0.303	0.302
$\mathcal{S}$	$\mathcal{F}$-16	0.304	0.323
$\mathcal{S}$	$\mathcal{F}$-8-RAG	0.653	0.691
$\mathcal{S}$	$\mathcal{F}$-16-RAG	0.643	0.680
$\mathcal{F}$-8-Norm	$\mathcal{F}$-8-RAG	<u>0.671</u>	**0.702**
$\mathcal{F}$-16-Norm	$\mathcal{F}$-8-RAG	0.669	<u>0.701</u>
$\mathcal{F}$-8-Norm	$\mathcal{F}$-16-RAG	**0.673**	**0.702**
$\mathcal{F}$-16-Norm	$\mathcal{F}$-16-RAG	0.652	0.609

Table 4: Effect of input and output domains on the validation set. Here, "$\mathcal{S}$" denotes the spatial domain, while "$\mathcal{F}$" represents the frequency domain. "Norm" refers to batch normalization applied after DCT tokens, and "RAG" denotes the Regressor-After-Gating mechanism. The numbers 8 and 16 indicate different patch sizes.

Loss Function	F1 (↑)	Dice (↑)
L1	0.674	0.701
L1$_{\mathcal{M}}$ (Min-Max Norm)	**0.693**	**0.746**
L1$_{\mathcal{M}}$ (Softmax Norm)	<u>0.681</u>	<u>0.742</u>

Table 5: Comparison of the proposed Mahalanobis regularization ($\mathcal{M}$) with the standard L1 loss. Two weighting strategies are evaluated for Mahalanobis regularization.

higher F1 scores than their spatial-domain counterparts, except DCT-Mask, underscoring the advantage of frequency-domain modeling in capturing high-frequency details. Table 3 further corroborates this trend in segmentation results, where our method consistently outperforms others. The superior performance on *RibFrac-GTRefined* reinforces the effectiveness of our approach in tackling challenging clinical segmentation tasks. Beyond accuracy, our method demonstrates superior computational efficiency, as summarized in Table 2. By adopting an encoder-only architecture with a shared DTR module, our design significantly reduces the overall parameter count while preserving segmentation accuracy. This architectural choice optimizes the trade-off between accuracy and computational cost, making our model more feasible for real-world medical applications.

4.4 Visualization

Figure 3 provides a qualitative comparison of segmentation results between our method and SOTA approaches. Our method exhibits superior segmentation accuracy, particularly in detecting small, hollow structures that pose significant challenges for existing models. Figure 5 illustrates our proposed multi-stage refinement strategy, where penalty terms between consecutive stages enforce consistency and guide deeper stages toward more confident predictions. This design mirrors the iterative refinement process employed by clinicians during manual annotation, enhancing both segmentation reliability and interpretability.

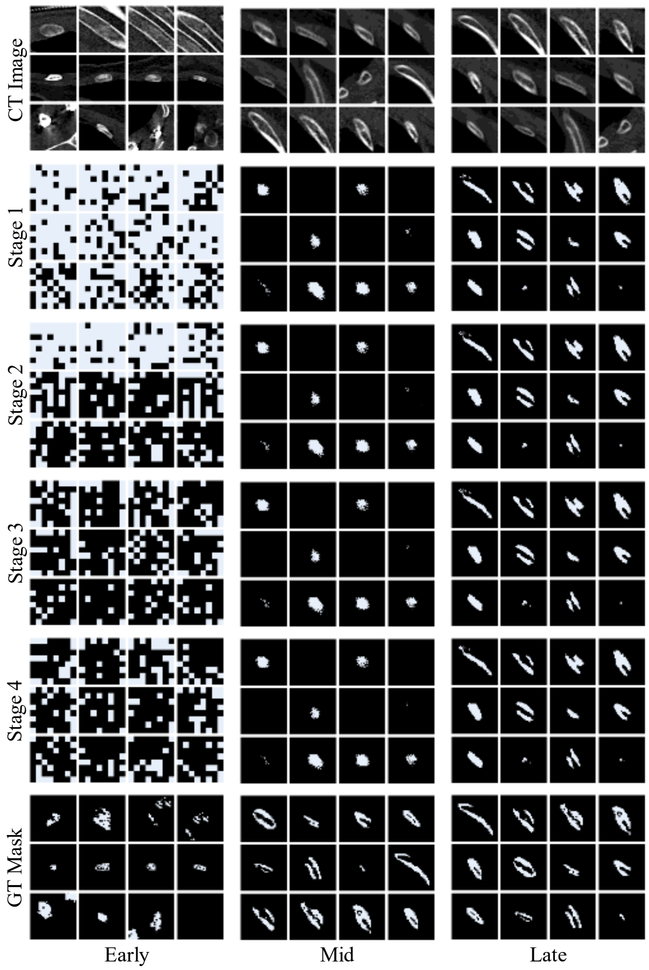

Figure 5: Progressive visualization of our hierarchical refinement strategy, mimicking human annotators' stepwise delineation of fracture masks. Predicted segmentations at early, mid, and late training epochs are shown alongside corresponding CT and ground-truth masks, demonstrating the model's capacity for iterative refinement

Stages	F1 (↑)	Dice (↑)
4	0.681	0.740
3,4	0.687	0.744
2,3,4	**0.693**	**0.751**
1,2,3,4	0.690	0.747

Table 6: Performance comparison across varying hierarchical depths. Models with 1 to 4 hierarchical stages are evaluated, demonstrating that hierarchical architectures consistently outperform non-hierarchical counterparts.

4.5 Ablation Study

Different Input and Output Domains. We explore the effect of varying the data domains for model inputs and outputs. As shown in Table 4, replacing voxel representations with DCT tokens in the input space consistently improves model performance, with larger patch sizes yielding better results. However, substituting the segmentation mask with DCT tokens and using L1 loss on the output side significantly degrades performance. This is likely due to the dominance of healthy voxels, which overshadow the gradients and impair the model's ability to capture fracture features effectively. In contrast, the proposed RAG mechanism substantially enhances model performance, demonstrating its effectiveness in modeling small and hollow targets.

Cls.	Reg.	F1 (↑)	Dice (↑)
		0.693	0.751
✓		0.696	0.753
	✓	0.702	0.749
✓	✓	**0.708**	**0.770**

Table 7: Ablation study on the effect of penalizing the gating and regressor modules in the hierarchical model with four stages.

Mahalanobis Regularization. Table 5 presents a quantitative comparison between the L1 loss and the proposed Mahalanobis-regularized loss. We evaluate two self-adaptive weighting strategies for Mahalanobis loss, and the results consistently show its superiority in improving model performance. This improvement is further supported by Figure 4, which illustrates that Mahalanobis regularization alleviates the known deficiency of L1 loss in capturing high-frequency signals, leading to more robust feature learning.

Number of Stages. We investigate the effect of varying the number of stages, as shown in Table 6. While a hierarchical design intuitively suggests performance gains, our results reveal a non-monotonic trend as the number of stages increases from 1 to 4. Performance initially improves with additional stages but declines when the number of stages exceeds three. Nevertheless, hierarchical architectures consistently outperform their non-hierarchical counterparts, emphasizing their effectiveness in structured feature learning.

Unified Multi-Stage Penalty. Table 7 evaluates the impact of the proposed hierarchical penalty, which imposes constraints on both the gating module and the regressor, either independently or in combination. The results indicate that penalizing both components simultaneously yields the highest performance, highlighting the importance of jointly optimizing these modules for optimal results.

5 Conclusions

We propose a novel approach for rib fracture analysis by developing a full-frequency model, which integrates Mahalanobis-regularized frequency loss and a unified cross-stage penalty within the RAG module. This method improves interpretability and achieves high-performance fracture detection, bridging the gap between clinical annotation workflows and deep learning model design. Future iterations are expected to improve performance by strategically selecting signals from input DCT tokens and developing a more task-oriented frequency-domain model.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Granted No. 62276190).

References

[Ahmed *et al.*, 1974] Nasir Ahmed, T₋ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

[Brosch and Saalbach, 2018] Tom Brosch and Axel Saalbach. Foveal fully convolutional nets for multi-organ segmentation. In *Medical imaging 2018: Image processing*, volume 10574, pages 198–206. SPIE, 2018.

[Cao *et al.*, 2023] Zheng Cao, Liming Xu, Danny Z Chen, Honghao Gao, and Jian Wu. A robust shape-aware rib fracture detection and segmentation framework with contrastive learning. *IEEE Transactions on Multimedia*, 25:1584–1591, 2023.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[Chen *et al.*, 2024] Linwei Chen, Lin Gu, Dezhi Zheng, and Ying Fu. Frequency-adaptive dilated convolution for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3414–3425, 2024.

[Cheng *et al.*, 2019] Chi-Tung Cheng, Tsung-Ying Ho, Tao-Yi Lee, Chih-Chen Chang, Ching-Cheng Chou, Chih-Chi Chen, I Chung, Chien-Hung Liao, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. *European radiology*, 29(10):5469–5477, 2019.

[Fernando *et al.*, 2021] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021.

[Fuoli *et al.*, 2021] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2360–2369, 2021.

[Gao *et al.*, 2022] Yuan Gao, Hongzhi Liu, Liang Jiang, Chunfeng Yang, Xindao Yin, Jean-Louis Coatrieux, and Yang Chen. Cce-net: A rib fracture diagnosis network based on contralateral, contextual, and edge enhanced modules. *Biomedical Signal Processing and Control*, 75:103620, 2022.

[Greenspan *et al.*, 2000] Hayit Greenspan, Charles H Anderson, and Sofia Akber. Image enhancement by nonlinear extrapolation in frequency space. *IEEE Transactions on Image Processing*, 9(6):1035–1048, 2000.

[Guo *et al.*, 2023] Xiaojun Guo, Yifei Wang, Tianqi Du, and Yisen Wang. Contranorm: A contrastive learning perspective on oversmoothing and beyond. *arXiv preprint arXiv:2303.06562*, 2023.

[Hatamizadeh *et al.*, 2022] 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.

[Huang *et al.*, 2023] Shu-Tien Huang, Liong-Rung Liu, Hung-Wen Chiu, Ming-Yuan Huang, and Ming-Feng Tsai. Deep convolutional neural network for rib fracture recognition on chest radiographs. *Frontiers in Medicine*, 10, 2023.

[Jiang *et al.*, 2021] 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.

[Jin *et al.*, 2020] Liang Jin, Jiancheng Yang, Kaiming Kuang, Bingbing Ni, Yiyi Gao, Yingli Sun, Pan Gao, Weiling Ma, Mingyu Tan, Hui Kang, et al. Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet. *EBioMedicine*, 62, 2020.

[Jin *et al.*, 2023] Liang Jin, Shixuan Gu, Donglai Wei, Jason Ken Adhinarta, Kaiming Kuang, Yongjie Jessica Zhang, Hanspeter Pfister, Bingbing Ni, Jiancheng Yang, and Ming Li. Ribseg v2: A large-scale benchmark for rib labeling and anatomical centerline extraction. *IEEE Transactions on Medical Imaging*, 2023.

[Lindsey *et al.*, 2018] Robert Lindsey, Aaron Daluiski, Sumit Chopra, Alexander Lachapelle, Michael Mozer, Serge Sicular, Douglas Hanel, Michael Gardner, Anurag Gupta, Robert Hotchkiss, et al. Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, 115(45):11591–11596, 2018.

[Loshchilov, 2017] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Lu *et al.*, 2023] Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 36:8487–8500, 2023.

[Pan *et al.*, 2022] Hongyi Pan, Xin Zhu, Salih Atici, and Ahmet Enis Cetin. Dct perceptron layer: A transform domain approach for convolution layer. *arXiv preprint arXiv:2211.08577*, 2022.

[Park and Kim, 2022] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

[Piao *et al.*, 2024] Xihao Piao, Zheng Chen, Taichi Murayama, Yasuko Matsubara, and Yasushi Sakurai. Fredformer: Frequency debiased transformer for time series forecasting. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2400–2410, 2024.

[Qin *et al.*, 2021] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021.

[Rao *et al.*, 2023] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jie Zhou, and Jiwen Lu. Gfnet: Global filter networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10960–10973, 2023.

[Shaker *et al.*, 2024] Abdelrahman M Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024.

[Shen *et al.*, 2021] Xing Shen, Jirui Yang, Chunbo Wei, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Xiaoliang Cheng, and Kewei Liang. Dct-mask: Discrete cosine transform mask representation for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8720–8729, 2021.

[Tatsunami and Taki, 2024] Yuki Tatsunami and Masato Taki. Fft-based dynamic token mixer for vision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15328–15336, 2024.

[Tian *et al.*, 2023] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in Neural Information Processing Systems*, 36:71911–71947, 2023.

[Wallace, 1992] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992.

[Wang *et al.*, 2018] Lizhi Wang, Tao Zhang, Ying Fu, and Hua Huang. Hyperreconnet: Joint coded aperture optimization and image reconstruction for compressive hyperspectral imaging. *IEEE Transactions on Image Processing*, 28(5):2257–2270, 2018.

[Wang *et al.*, 2022] Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*, 2022.

[Wen *et al.*, 2022] Qinrou Wen, Jirui Yang, Xue Yang, and Kewei Liang. Patchdct: Patch refinement for high quality instance segmentation. In *The Eleventh International Conference on Learning Representations*, 2022.

[Wu *et al.*, 2021] Mingxiang Wu, Zhizhong Chai, Guangwu Qian, Huangjing Lin, Qiong Wang, Liansheng Wang, and Hao Chen. Development and evaluation of a deep learning algorithm for rib segmentation and fracture detection from multicenter chest ct images. *Radiology: Artificial Intelligence*, 3(5):e200248, 2021.

[Xu *et al.*, 2019] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*, pages 264–274. Springer, 2019.

[Xu *et al.*, 2020] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1740–1749, 2020.

[Yao *et al.*, 2021] Liding Yao, Xiaojun Guan, Xiaowei Song, Yanbin Tan, Chun Wang, Chaohui Jin, Ming Chen, Huogen Wang, and Minming Zhang. Rib fracture detection system based on deep learning. *Scientific reports*, 11(1):23513, 2021.

[You *et al.*, 2022] Zhiyuan You, Kai Yang, Wenhan Luo, Lei Cui, Yu Zheng, and Xinyi Le. Adtr: Anomaly detection transformer with feature reconstruction. In *International Conference on Neural Information Processing*, pages 298–310. Springer, 2022.

[Yu *et al.*, 2022] Hui Yu, Yongzheng Ding, Jinglai Sun, Xuyao Yu, Fengling Zhu, Huanming Li, Xiaoyun Liang, and Dagong Jia. Deep-learning algorithm for rib fracture detection in low-dose chest ct images. In *Proceedings of the 2022 9th International Conference on Biomedical and Bioinformatics Engineering*, pages 55–60, 2022.

[Zhang *et al.*, 2021] Bin Zhang, Chunxue Jia, Runze Wu, Baotao Lv, Beibei Li, Fuzhou Li, Guijin Du, Zhenchao Sun, and Xiaodong Li. Improving rib fracture detection accuracy and reading efficiency with deep learning-based detection software: a clinical evaluation. *The British journal of radiology*, 94(1118):20200870, 2021.

[Zhang *et al.*, 2023] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6782–6791, 2023.

[Zhao *et al.*, 2021] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[Zhao *et al.*, 2022] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5697–5707, 2022.

[Zhao *et al.*, 2024] Zewei Zhao, Xichao Dong, Yupei Wang, Jianping Wang, Yubao Chen, and Cheng Hu. Mdtnet: Multi-scale deformable transformer network with fourier space losses toward fine-scale spatiotemporal precipitation nowcasting. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.