# FUSION: Frequency-Unified Structural Image diffusiON for Low-Dose CT Denoising

**Yudhistira Arief Wibowo**
20256303
`wibowo@kaist.ac.kr`

**Hojoon Ki**
20253052
`joon4366@kaist.ac.kr`

**Minseo Park**
20258072
`joon4366@kaist.ac.kr`

## 1 Motivation

Low-Dose Computed Tomography (LDCT) is crucial for reducing radiation exposure in medical imaging, but it inevitably introduces significant noise and artifacts, potentially compromising diagnostic accuracy. While deep learning methods have shown great promise for LDCT denoising, a fundamental challenge remains: effectively removing noise without sacrificing critical anatomical structures or introducing spurious details.

Diffusion Models have recently emerged as state-of-the-art generative models, capable of producing remarkably high-quality images. Their iterative refinement process allows for sophisticated noise removal. However, standard diffusion models operate primarily in the image (spatial) domain. While powerful, they might inadvertently alter fine anatomical details, especially if the noise characteristics are complex or spatially varying, as is common in CT. There's a risk of the model "hallucinating" or overly smoothing structures during the reverse diffusion process.

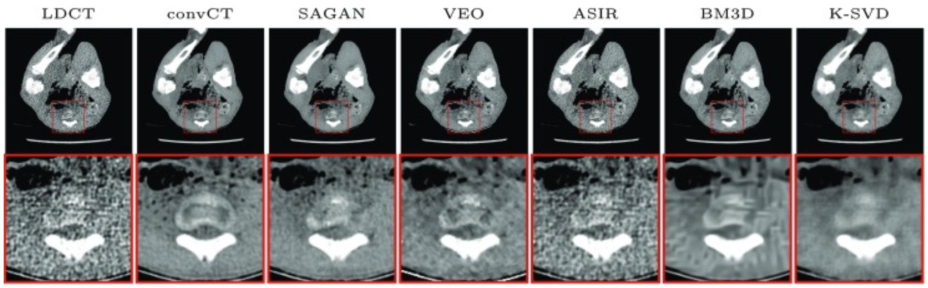

Figure 1: Example of oversmoothing of edges due to structural information loss. [Yi and Babyn, 2018]

Conversely, the Frequency Domain offers a distinct advantage: it naturally separates global structural information (concentrated in low frequencies) from high-frequency noise. This property has been exploited in classical signal processing and more recently in deep learning approaches like the FFM model [Gong and Yu, 2025], which demonstrated superior structure preservation by focusing denoising efforts on specific frequency bands.

However, methods operating solely in the frequency domain might struggle to capture fine local textures and context that are better represented in the image domain. We hypothesize that synergistically combining the strengths of both domains within a diffusion framework can lead to a more robust and accurate denoising model.

Our goal is to develop a model named FUSION (**F**requency-**U**nified **S**tructural **I**mage diffusi**ON**) that leverages structural priors from the frequency domain to explicitly guide the diffusion process in the image domain. This guidance aims to ensure that the model focuses its denoising power primarily on noise components while rigorously preserving essential anatomical contours and details.

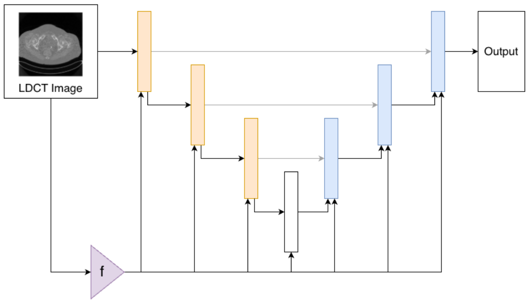

Figure 2: Planned architecture for our proposed model

## 2 Methodology

Our proposed approach integrates frequency-domain structural information into an image-domain diffusion model using a cross-attention mechanism.

### 2.1 Core Denoising Engine: Conditional Diffusion Model

We will employ a conditional Diffusion Model (e.g., based on DDPM or DDIM) as the backbone. The model takes a noisy LDCT image $y$ and learns to iteratively denoise it to predict the corresponding clean NDCT image $x$. The standard architecture involves a U-Net within the diffusion framework to predict the noise (or the clean image) at each timestep $t$.

### 2.2 Frequency Domain Structural Prior Extraction

For a given noisy input image $y$, we compute its 2D Fast Fourier Transform (FFT):

$$Y = \text{FFT}(y)$$

We apply a low-pass filter in the frequency domain to isolate the low-frequency components, which primarily represent the image's structural information:

$$Y_{LF} = \text{LowPassFilter}(Y)$$

This low-frequency map $Y_{LF}$ (or its inverse FFT $y_{LF} = \text{IFFT}(Y_{LF})$) serves as the structural prior. It captures the essential anatomical contours while inherently suppressing high-frequency noise.

### 2.3 Frequency-Guided Cross-Attention Mechanism

The core innovation lies in integrating the structural prior $y_{LF}$ into the U-Net backbone of the diffusion model. We will incorporate Cross-Attention layers at various stages of the U-Net decoder (and potentially the encoder).

**Mechanism:** At a given layer in the U-Net (operating on image-domain features $z_t$ at timestep $t$), the features $z_t$ will serve as the Query. The structural prior $y_{LF}$ (potentially processed through a small feature extractor to match dimensions) will serve as the Key and Value. The cross-attention module computes:

$$\text{Attention}(\text{Query} = z_t, \text{Key} = y_{LF}, \text{Value} = y_{LF})$$

This allows the U-Net features $z_t$ to selectively attend to and incorporate information from the structural prior $y_{LF}$. Essentially, the model asks: "For this region I'm currently denoising, what does the stable, low-frequency structural map tell me should be preserved here?" This attention mechanism guides the diffusion model during the reverse process, encouraging it to generate details consistent with the underlying low-frequency structure while removing inconsistencies (noise) present in the original noisy image $y$.

## References

Guoliang Gong and Man Yu. A denoising framework for real-world ultra-low dose lung ct images based on an image purification strategy, 2025. URL https://arxiv.org/abs/2510.07492.

Xin Yi and Paul Babyn. Sharpness-aware low-dose ct denoising using conditional generative adversarial network. *Journal of digital imaging*, 31(5):655–669, 2018.