Signal Enhancement via Multi-view Dynamic Representation and Alignment-aware Fusion

Zikun Jin[1,2], Yuhua Qian[1,2*], Xinyan Liang[1,2], Jiaqian Zhang[1,2], Jinpeng Yuan[3], Shen Hu[1,2], Haijun Geng[4], Honghong Cheng[5]

[1]Institute of Big Data Science and Industry, Shanxi University, Taiyuan, China
[2]Key Laboratory of Evolutionary Science Intelligence of Shanxi Province, Shanxi University, Taiyuan, China
[3]State Key Laboratory of Quantum Optics Technologies and Devices, Institute of Laser Spectroscopy, Shanxi University, Taiyuan, China
[4]School of Automation and Software Engineering, Shanxi University, Taiyuan, China
[5]School of Information, Shanxi University of Finance and Economics, Taiyuan, China
jinzikun@163.com, jinchengqyh@126.com, liangxinyan48@163.com, zhangjiaqian1@sxu.edu.cn, yjp@sxu.edu.cn, hushen@alu.sxu.edu.cn, ghj123025449@163.com, chhsxdx@163.com

Abstract

Robust signal enhancement under non-stationary and low SNR conditions remains challenging, as methods based on the short-time Fourier transform (STFT) with fixed resolution struggle to represent complex and time–frequency structures. While leveraging the fractional domain as an auxiliary view offers flexibility in modeling time-frequency structures, existing methods typically adopt fixed transform orders and overlook alignment between views, hindering effective integration of complementary representations and leaving frequency domain misalignment unresolved. Therefore, we propose FracFusion, a novel framework that integrates a learnable short-time fractional Fourier Transform (STFrFT) module to generate dynamic auxiliary views, combined with two stage alignment-aware fusion modules: Pearson Channel Fusion for correlation-guided consistency and Efficient Align Fusion for fine-grained, frequency aligned interaction. Experiments on speech and electromagnetic (EM) datasets show that FracFusion consistently outperforms state-of-the-art baselines across diverse noise levels and signal types, demonstrating robust adaptability across domains.

1 Introduction

Signal enhancement (SE) is vital in applications like speech communication (Luo and Mesgarani 2019), wireless signal demodulation (Du et al. 2022), and intelligent sensing (Su et al. 2023). In speech, SE improves intelligibility and recognition in noisy environments (Guo et al. 2024), while in electromagnetic (EM) domains, it aids in recovering modulated signals for tasks like radar detection and secure communications. Both face common challenges non-stationary, low SNR, and complex time-frequency dynamics that complicate signal recovery, underscoring the need for a unified adaptive modeling framework.

SE approaches, particularly STFT based methods, project signals into fixed resolution time–frequency domains, enabling localized spectral analysis (Fu et al. 2021; Shin et al. 2023; Fan et al. 2025; Wang et al. 2025). However, their rigid windowing constrains representation capacity under non-stationary and structurally asymmetric conditions, such as frequency drifting EM modulations and time varying harmonic trajectories in speech (Nareddula, Gorthi, and Gorthi 2021; Dang et al. 2023). To enhance representational capacity, researchers have explored alternative time–frequency transforms such as wavelets and the fractional Fourier transform (FrFT) to construct complementary auxiliary views(Zhou et al. 2023). Yet, these are typically implemented with static parameters and lack the adaptability to different signal types or varying noise conditions. More importantly, fusion across multiple representations is often performed via naive operations like direct summation or concatenation, without accounting for alignment or correlation between views (Michelsanti et al. 2021; Xiong et al. 2022). These limitations highlight the need for multi-view dynamic representations that adaptively capture diverse time–frequency structures (Zeng, Xu, and Wang 2024), as well as alignment-aware fusion strategies to reconcile inconsistencies across-views, laying the foundation for robust modeling in complex, non-stationary signal conditions.

To this end, a robust multi-view SE framework should satisfy two properties: (i) the ability to adaptively adjust its representation based on the non-stationarity of the input signal; (ii) the capacity for alignment-aware fusion to effectively integrate complementary multi-view representations.

In response, we propose FracFusion, a unified deep enhancement framework that incorporates a learnable short-time fractional Fourier Transform (STFrFT) module to generate adaptive spectral views tailored to signal dynamics. To fully exploit multi-view consistency and complementarity, we design a two branch fusion architecture consisting of Efficient Align Fusion for frequency alignment and Pearson Channel Fusion for correlation-aware integration. Our method is effectively generalized across both the speech and EM signal domains.

Extensive experiments on both speech and EM signal benchmarks demonstrate that our method consistently out-

*The corresponding author

performs recent state-of-the-art approaches in terms of objective metrics and perceptual quality. The improvements are particularly significant under challenging low SNR scenarios and in the presence of cross-domain noise, highlighting the robustness and generalizability of our framework.

This paper presents the following key contributions:

- We propose an adaptive STFrFT module that dynamically learns signal-specific fractional orders, enabling the generation of complementary non-stationary spectral views for enriched multi-view representation.
- We propose Pearson Channel Fusion and Efficient Align Fusion, complementary fusion modules that jointly exploit inter-view consistency and feature complementarity: PCF at the coarse channel level, EAF through fine-grained temporal alignment and localized fusion.
- We construct a structure-aware module with frequency-aware and asymmetric residual modeling for joint time-frequency feature extraction.

2 Related Work

2.1 Deep Learning Based Signal Enhancement

Deep learning has emerged as a dominant paradigm for signal enhancement (SE), particularly in speech and electromagnetic (EM) domains. STFT-based approaches estimate spectral masks or magnitudes using convolutional or recurrent architectures (Fu et al. 2019; Défossez, Synnaeve, and Adi 2020; P-J et al. 2023), while end-to-end waveform models operate directly in the time domain using encoder–decoder structures or dilated convolution networks (Phan et al. 2020; Lu et al. 2022; Kong et al. 2022). These methods have shown impressive performance in stationary environments, but often fail under non-stationary and low SNR conditions due to their reliance on fixed-resolution analysis or uniform treatment of temporal and spectral features. In the EM domain, recent work has extended deep enhancement techniques to tasks such as demodulated signal restoration and interference suppression (Fuchs et al. 2021; Du et al. 2022; Su et al. 2023), yet similar structural modeling limitations persist.

2.2 Extended Time–frequency Representations

To better accommodate non-stationary and dynamically evolving signal structures, alternative time–frequency transforms have been explored. Wavelet transforms and constant-Q transforms (CQT) offer multi-resolution analysis tailored to auditory perception or scale invariance (Wang et al. 2021; Gu et al. 2024). Fractional Fourier transform (FrFT) and its short-time variant (STFrFT) extend this idea by allowing continuous tuning of time–frequency resolution via fractional orders (Koç and Koç 2022; Huang, Zhang, and Tao 2022). STFrFT, in particular, provides a unified view bridging time and frequency domains and has been applied to improve robustness in various signal classification tasks (Li et al. 2024). However, existing uses of FrFT/STFrFT rely on manually selected or static orders and are rarely optimized jointly with downstream tasks.

Their lack of adaptability and end-to-end integration limits their effectiveness in dynamic signal environments. This motivates the need for an adaptive STFrFT module capable of learning signal fractional orders, thereby enhancing dynamic representation within end-to-end enhancement.

2.3 Multi-View Representation Fusion and Alignment

Multi-view learning has emerged as a promising approach to leverage complementary information from different signal perspectives (Liang et al. 2025a,b). Previous works have explored various fusion strategies, including combining STFT with wavelet features (Ventricci, Ribeiro Junior, and Gomes 2024; Ishwarya and Kothandaraman 2025), multi-channel modeling (Xue et al. 2024; Son et al. 2024), and cross-domain representations (Yu et al. 2020; Wang et al. 2022; Jin et al. 2025). However, existing multi-view fusion methods in signal enhancement tasks often face challenges such as semantic mismatches and phase inconsistencies between views, primarily due to reliance on naive concatenation or static weighting strategies (Tian et al. 2024; Wang et al. 2024). To address these limitations, we propose two fusion modules: Pearson channel fusion, a coarse-grained feature level strategy that exploits channel correlation to achieve feature fusion for inter-view channels, and efficient alignment fusion, a fine-grained feature level strategy that exploits lightweight attention to achieve frequency alignment of STFrFT and STFT.

3 Methodology

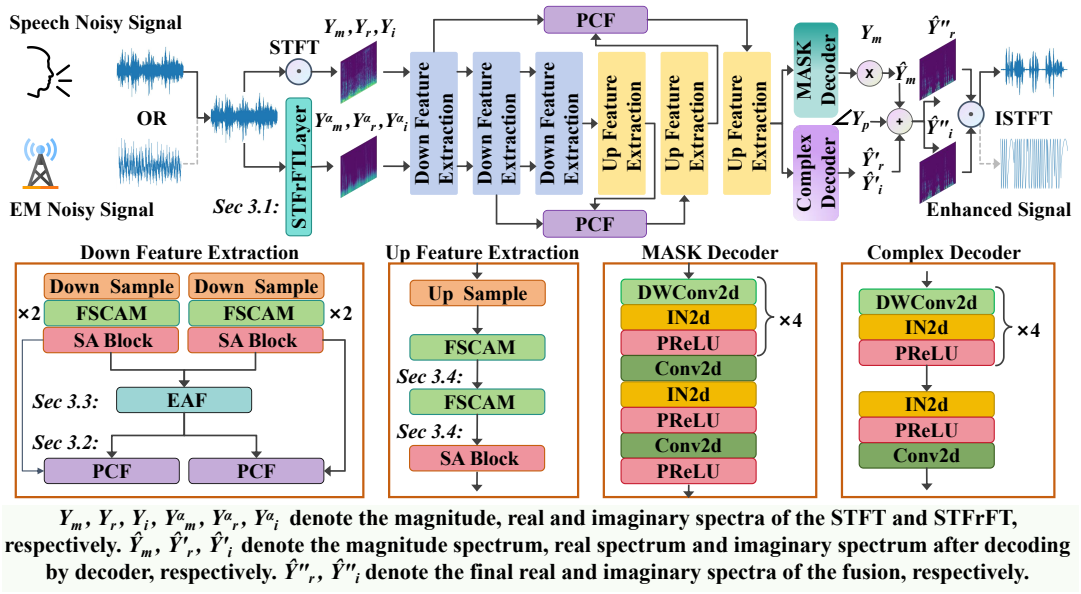
This section details the modular design principles and corresponding loss functions of the FracFusion network components, as depicted in Figure 1.

3.1 Adaptive Short-time Fractional Fourier Transform Module

Non-stationary signals, such as chirps and time-varying harmonics, present significant challenges to conventional short-time Fourier Transform (STFT) due to its fixed basis and rigid time–frequency resolution trade-off, often leading to smeared or misaligned spectral representations in dynamic environments. To address this, we design a learnable short-time fractional Fourier Transform (STFrFT) module that extends STFT by projecting the signal onto a rotated time–frequency plane. Unlike fixed parameter transforms, our STFrFT layer features a trainable fractional order $\alpha \in [0, 1]$, enabling the network to adaptively select the optimal spectral resolution based on the input signal's non-stationarity. This approach not only enhances the model's representational flexibility but also incorporates fractional domain priors into end-to-end learning, allowing the transform to specialize across varying signal conditions.

Let $x(t) \in \mathbb{R}$ be a real-valued discrete signal. The fractional Fourier transform of order $\alpha \in [0, 1]$ is defined as a linear transformation: $\mathcal{F}_\alpha\{x(t)\} = X_\alpha(u)$, where $\alpha = 0$, yields $x(t)$; $\alpha = 1$ yields FFT.

In the discrete setting, we adopt the approximation to implement FrFT efficiently. Given a framed signal segment



Y_m, Y_r, Y_i, $Y^{\alpha}_m, Y^{\alpha}_r, Y^{\alpha}_i$ denote the magnitude, real and imaginary spectra of the STFT and STFrFT, respectively. $\hat{Y}_m$, $\hat{Y}'_r$, $\hat{Y}'_i$ denote the magnitude spectrum, real spectrum and imaginary spectrum after decoding by decoder, respectively. $\hat{Y}''_r$, $\hat{Y}''_i$ denote the final real and imaginary spectra of the fusion, respectively.

Figure 1: Overview of the FracFusion architecture. The input signals are transformed into STFT and adaptive STFrFT views, which are first encoded through separate down feature extraction layers. They are then initially decoded using a shared up feature extraction, followed by decoding with the magnitude mask and complex spectrum decoders. The enhanced waveforms are reconstructed in the time domain.

$x \in \mathbb{C}^N$, the approximate discrete FrFT is:

$$\phi(n) = e^{-j\pi n^2 \tan(\theta/2)/N}, \tag{1}$$

$$X_\alpha = \frac{1}{\sqrt{\sin(\theta)}} \cdot [\mathrm{FFT}(x \cdot \phi(n)) \cdot \phi(n)]\,. \tag{2}$$

where $\theta = \frac{\pi}{2} \cdot \alpha$ is the FrFT rotation angle; $n = 0, 1, \ldots, N-1$ is the discrete time index; The exponential terms represent chirp modulations.

The adaptive fractional order α is modeled as: $\alpha = 0.5 \cdot (\tanh(\alpha_{\text{raw}}) + 1)$, where $\alpha_{\text{raw}} \in \mathbb{R}$ is a trainable parameter, ensuring $\alpha \in (0, 1)$ and allowing gradient-based learning. The short-time FrFT (STFrFT) is computed by applying the above FrFT to overlapping windowed segments of the signal: $\mathbf{X}_\alpha(b, \omega) = \mathcal{F}_\alpha\{x(t) \cdot w(t-b)\}, \quad \forall b \in$ frame positions, where $w(t)$ is a Hamming window and b is the frame center. Only the positive frequency spectrum is retained: $\mathbf{X}^{\text{half}}_\alpha \in \mathbb{C}^{B \times F_{\text{half}} \times T_{\text{frames}}}, \quad F_{\text{half}} = W/2 + 1$.

The proposed STFrFT layer enhances time–frequency modeling of non-stationary signals by enabling data driven fractional domain selection within neural networks.

3.2 Pearson Channel Fusion

In signal enhancement tasks, simultaneously incorporating multiple time–frequency views, such as STFT and STFrFT, can provide richer representations by exploiting their complementary properties. STFT is well-suited for capturing stable harmonic structures, while STFrFT offers enhanced flexibility for modeling non-stationary components, such as linearly frequency-modulated (FM) signals. However, traditional fusion strategies, such as uniform weighting or direct feature concatenation, treat all channels equally and ignore the local consistency or structural disparity between views, leading to suboptimal integration.

To address this, we propose Pearson Channel Fusion (PCF), a coarse-grained, channel-wise fusion mechanism that adaptively selects and combines information based on inter-view correlation. Specifically, we compute the Pearson correlation coefficient between corresponding channels of STFT and STFrFT feature maps to estimate their statistical similarity. Channels with high correlation are considered redundant, and the primary view (STFT) is retained; conversely, low-correlation channels are treated as complementary and enhanced by incorporating information from the auxiliary view (STFrFT). This selective aggregation improves robustness and avoids feature interference across heterogeneous regions. The specific process of Pearson channel fusion can be represented in the following form.

Let the input feature maps from two distinct views be denoted as $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{B \times C \times F \times T}$, where B is the batch size, C is the number of channels, and F and T represent the frequency and time dimensions, respectively. To compute the channel-wise similarity between the two views, we first flatten each feature map along the spatial dimensions for each channel $c \in \{1, \ldots, C\}$:

$$\mathbf{x}_1^{(c)} = \mathrm{vec}\left(\mathbf{X}_1^{(c)}\right) \in \mathbb{R}^{F \cdot T}, \quad \mathbf{x}_2^{(c)} = \mathrm{vec}\left(\mathbf{X}_2^{(c)}\right). \tag{3}$$

Each flattened feature vector is then standardized by subtracting the mean and dividing by the standard deviation:

$$\tilde{\mathbf{x}}_i^{(c)} = \frac{\mathbf{x}_i^{(c)} - \mu\left(\mathbf{x}_i^{(c)}\right)}{\sigma\left(\mathbf{x}_i^{(c)}\right) + \epsilon}, \quad i = 1, 2 \tag{4}$$

Algorithm 1: STFrFT Layer

Input: 1D signal $x \in \mathbb{R}^{B\times T}$

Parameter: Window size W, hop size H; Learnable fractional order $\alpha \in (0,1)$.

Note: α is implemented via a reparameterized tanh transformation to ensure differentiability and boundedness: $\alpha = 0.5 \cdot (\tanh(\alpha_{\text{raw}}) + 1)$.

Output: $X_{stfrft} \in \mathbb{C}^{B\times F_{half}\times T_{frames}}$.

Pad input signal x with H samples on both sides .

Split x into overlapping frames of size W with hop size H: $x_{frames} \leftarrow \text{Frame}(x, W, H) \in \mathbb{R}^{B\times T_{frames}\times W}$

Apply window function w to each frame: $x_{frames} \leftarrow x_{frames} \cdot w$

Reshape and cast to complex: $x_{flat} \leftarrow \text{Reshape}(x_{frames}) \in \mathbb{C}^{(B\cdot T_{frames})\times W}$

Compute fractional angle: $\theta = \alpha \cdot \frac{\pi}{2}$

For each frame:

$x' = x_{flat} \cdot \exp\left(-j\pi \tan\left(\frac{\theta}{2}\right) \cdot \frac{n^2}{W}\right)$

$X = \text{FFT}(x')$

$X' = X \cdot \exp\left(-j\pi \tan\left(\frac{\theta}{2}\right) \cdot \frac{n^2}{W}\right)$

Normalization: $X_{\text{frft}} = X'/\sqrt{\sin(\theta)}$

Retain half-spectrum: $X_{stfrft} \leftarrow X_{frft}[:, 0 : F_{half}]$

Reshape to output format: $X_{stfrft} \in \mathbb{C}^{B\times F_{half}\times T_{frames}}$

return X_{stfrft}

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation computed along the spatial dimension, and ϵ is a small constant for numerical stability.

We then compute the channel Pearson correlation to quantify the linear correlation between the two views:

$$\rho^{(c)} = \frac{1}{F\cdot T}\sum_{j=1}^{F\cdot T} \tilde{x}_1^{(c)}(j)\cdot \tilde{x}_2^{(c)}(j). \tag{5}$$

To ensure a bounded weighting scheme, we rescale the correlation coefficient to the range $[0, 1]$ as follows:

$$w^{(c)} = \frac{\rho^{(c)}+1}{2}, \quad w^{(c)} \in [0,1]. \tag{6}$$

Finally, the fused output for each channel is obtained via a correlation-guided weighted sum of the two input views:

$$\mathbf{Z}^{(c)} = w^{(c)}\cdot \mathbf{X}_1^{(c)} + \left(1 - w^{(c)}\right)\cdot \mathbf{X}_2^{(c)}. \tag{7}$$

This strategy enables the network to adaptively preserve information from the more reliable source (as determined by correlation), while simultaneously injecting complementary cues from the less correlated view. PCF provides a global alignment before the channel level, laying the foundation for a more refined fusion. The fine-grained alignment-aware fusion mechanisms are detailed in the next section.

3.3 Efficient Align Fusion

Due to the different transform domain and phase characteristics of STFT and STFrFT, structural misalignment often occurs. Direct fusion of the misalignments can lead to information conflicts and performance degradation, especially under non-stationary or complex signal conditions.

To address this, we propose Efficient Align Fusion (EAF), a lightweight, fine-grained alignment-and-fusion module designed for misaligned spectral views. EAF first employs compressed temporal attention to align the STFrFT features with the STFT domain to ensure same-frame consistency. A subsequent gated fusion mechanism then adaptively integrates the aligned representations based on local spatiotemporal relevance.

EAF offers the following advantages for signal enhancement: (1) Efficiency: It employs low dimensional projections and restricts attention computation to the time axis, reducing overhead relative to full spatial attention. (2) Alignment: Temporal attention explicitly compensates for cross-view discrepancies, such as phase shifts and time–frequency localization mismatches. (3) Adaptivity: Gated fusion dynamically balances contributions from STFT and STFrFT based on per-location relevance.

As the fine-grained counterpart to coarse-grained Pearson Channel Fusion, EAF enables localized, context-aware integration of heterogeneous time–frequency features, jointly forming a complementary dual stage fusion framework for robust signal enhancement.

Let $\mathbf{X}_{\text{S}}, \mathbf{X}_{\alpha} \in \mathbb{R}^{B\times C\times F\times T}$ denote the input feature maps from the STFT and STFrFT branches, respectively. To perform efficient alignment, we first compress the channel dimension from C to $C_r = C/4$ using 1×1 convolutions:

$$\mathbf{Q} = \text{Mean}_F(\mathbf{W}q * \mathbf{X}_{\text{S}}) \in \mathbb{R}^{B\times C_r\times T}, \tag{8}$$

$$\mathbf{K} = \text{Mean}_F(\mathbf{W}k * \mathbf{X}_{\alpha}) \in \mathbb{R}^{B\times C_r\times T}, \tag{9}$$

$$\mathbf{V} = \text{Mean}_F(\mathbf{W}v * \mathbf{X}_{\alpha}) \in \mathbb{R}^{B\times C_r\times T}, \tag{10}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{C_r\times C\times 1\times 1}$ are 1×1 convolutions applied independently to compress channel dimensions, and $\text{Mean}_F(\cdot)$ denotes frequency-wise average pooling across the F axis.

We then compute the scaled dot-product temporal attention matrix:

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}^{\top}\mathbf{K}}{\sqrt{C_r}}\right) \in \mathbb{R}^{B\times T\times T}, \tag{11}$$

where $\mathbf{Q}^{\top} \in \mathbb{R}^{B\times T\times C_r}$ and $\mathbf{K} \in \mathbb{R}^{B\times C_r\times T}$.

The attention output is then obtained by:

$$\mathbf{V}_{\text{aligned}} = \mathbf{A}\cdot \mathbf{V}^{\top} \in \mathbb{R}^{B\times T\times C_r}. \tag{12}$$

We reshape this aligned representation back to the original 4D feature map via a learned projection and expansion:

$$\tilde{\mathbf{X}}_{\alpha} = \mathbf{W}_{\text{out}} * \text{Reshape}(\mathbf{V}_{\text{aligned}}) \in \mathbb{R}^{B\times C\times F\times T}, \tag{13}$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{C\times C_r\times 1\times 1}$ is a 1×1 convolution, and Reshape restores the frequency dimension F via broadcasting or repetition along F (depending on implementation).

To fuse the aligned $\tilde{\mathbf{X}}_{\alpha}$ with the original $\mathbf{X}_{\text{S}}$, we first concatenate them along the channel dimension and generate a fusion gate:

$$\mathbf{G} = \sigma\left(\mathbf{W}g * [\mathbf{X}_{\text{S}}|\tilde{\mathbf{X}}_{\alpha}]\right) \in \mathbb{R}^{B\times C\times F\times T}, \tag{14}$$

where $\mathbf{W}_g \in \mathbb{R}^{C\times 2C\times 1\times 1}$ is a 1×1 convolution, $[\cdot|\cdot]$ denotes channel-wise concatenation, and $\sigma(\cdot)$ is the sigmoid activation.

The final fused output is computed as a convex combination of the two inputs:

$$\mathbf{Z} = \mathbf{G}\odot\mathbf{X}_{\mathrm{S}} + (1-\mathbf{G})\odot\tilde{\mathbf{X}}_{\alpha} \in \mathbb{R}^{B\times C\times F\times T}, \quad (15)$$

where $\odot$ denotes element-wise multiplication.

The EAF module achieves content-adaptive, lightweight alignment across views, followed by gated fusion to ensure stable and coherent information integration. This mechanism is particularly effective when heterogeneous time–frequency representations display structural discrepancies while retaining complementary signal characteristics.

3.4 Structure-aware Modeling and Joint Loss Optimization

To effectively capture the structured, asymmetric patterns of real world signals in the time–frequency domain, we introduce the Fused Structure-aware Convolution and Attention Module (FSCAM) as the core feature extractor. FSCAM integrates depthwise separable convolutions for efficiency, asymmetric temporal–spectral modeling for structure preservation, and lightweight attention mechanisms (channel, spatial, residual) to emphasize salient patterns. A structure-aware extension (SA Block) further enhances frequency sensitivity and long range temporal context. Through this design, FSCAM effectively characterizes heterogeneous signal patterns, such as speech formants and RF bursts, while maintaining a lightweight computational profile.

For training, we adopt a multi-objective loss function that jointly supervises magnitude, complex spectrum, and time-domain reconstruction. The overall loss is defined as: $\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{Mag}} + (1-\alpha)\mathcal{L}_{\text{RI}} + \beta\mathcal{L}_{\text{Time}}$, where $\alpha = 0.9$ and $\beta = 0.2$ are weighting coefficients. We define the magnitude loss to measure the discrepancy between the estimated and ground-truth magnitude spectra: $\mathcal{L}_{\text{Mag}} = \mathbb{E}\left[\left|Y_m - \hat{Y}_m\right|^2\right]$, where $\hat{Y}_m = \sqrt{\hat{Y}_r^2 + \hat{Y}_i^2}$ is computed from the predicted real and imaginary components. To enforce accurate reconstruction in the complex domain, we compute the real-imaginary (RI) loss as: $\mathcal{L}_{\text{RI}} = \mathbb{E}\left[\left|Y_r - \hat{Y}_r\right|^2\right] + \mathbb{E}\left[\left|Y_i - \hat{Y}_i\right|^2\right]$. Finally, to improve temporal consistency and perceptual quality, we incorporate an $\mathcal{L}_1$ loss in the waveform domain: $\mathcal{L}_{\text{Time}} = \mathbb{E}\left[|x-\hat{x}|_1\right]$, where $\hat{x} = \text{ISTFT}\left(\hat{Y}_r, \hat{Y}_i\right)$ denotes the inverse transform of the predicted spectrum.

4 Experiments

We evaluated FracFusion on five signal enhancement datasets: (i) Speech Enhancement Datasets: VoiceBank+DEMAND dataset; EARS-WHAM! dataset (Richter et al. 2024). (ii) Electromagnetic (EM) Signal Enhancement Datasets: Radio_0dB dataset; Radio_-10dB dataset; Modulated_-10dB dataset. We conducted a brief analysis and provided the experimental parameter settings.

4.1 Speech Signal Enhancement Datasets

VoiceBank+DEMAND dataset is a widely used benchmark for speech enhancement. It comprises 11,572 clean utterances from 28 training speakers and 824 utterances from 2 unseen test speakers. Noisy mixtures are generated at four SNRs (0, 5, 10, 15 dB) for training and four mismatched SNRs (2.5, 7.5, 12.5, 17.5 dB) for testing. The noise conditions span diverse real-world environments, including public venues, domestic settings, and transportation scenes. All audio is resampled to 16 kHz for consistency.

EARS-WHAM! is a recent high-quality speech enhancement benchmark comprising 100 hours of clean, anechoic speech mixed with real-world noise. It includes 32,475 training utterances from 99 speakers and 886 test utterances from 7 unseen speakers. Compared to VoiceBank, EARS covers more diverse speaking styles (e.g., emotional, conversational, non-verbal). Noise is sourced from WHAM!, recorded in public venues around the San Francisco Bay Area. Mixtures are generated at SNRs uniformly sampled from $[-2.5, 17.5]$dB and scaled using LKFS. All audio is resampled to 16kHz for consistency.

4.2 Electromagnetic Signal Enhancement Dataset

To comprehensively evaluate FracFusion under diverse EM environments, we construct three benchmark datasets that span a wide range of signal types, frequencies, sampling settings, and noise scenarios. These datasets simulate real world low SNR conditions with representative signal forms and interference patterns, ensuring the generalizability and robustness of model learning.

We simulate a multi-frequency electromagnetic (EM) environment inspired by Rydberg atom-based sensing systems, which leverage the extreme sensitivity of highly excited atoms (with large principal quantum numbers) to enable high-precision, wide-bandwidth, and quantum-intrinsic passive detection of radio-frequency and microwave signals. Our dataset spans 12 carrier frequencies from 20MHz to 29.51GHz, each paired with practical sampling rates (0.16–160GHz) and acquisition durations reflecting real world sensing constraints. For each frequency point, we construct a diverse set of baseband signals, including linear frequency modulated (LFM), sinusoidal, square, and triangular waveforms, and augment them with four distinct noise sources: Gaussian, Poisson, pink, and empirically measured Rydberg EM background noise. Noise is added under predefined SNR conditions to simulate realistic low SNR EM environments. Specifically, the Radio_0dB dataset samples SNR values uniformly from -20 dB to 20 dB, while the Radio_-10dB dataset samples uniformly from -20 dB to 0 dB. To benchmark performance under complex modulated transmissions, we construct a synthetic dataset consisting of 28 canonical analog and digital modulation types, including AM, FM, PM, CW, BPSK, QPSK, 8PSK, 16/64/128/256-QAM, MSK, GMSK, OFDM, and various FSK and ASK variants. Each modulated signal is generated with a carrier frequency of 1 kHz and sampled at 48 kHz over a duration of 3 seconds, resulting in 48,000 time-domain samples per instance. Random bitstreams are modulated into base-

Methods	Pub/Year	Input	PESQ ↑	CSIG ↑	CBAK ↑	COVL ↑	SSNR ↑	STOI ↑
Noisy	-	-	1.97	3.3	2.44	2.63	1.68	0.91
TFT-Net	IJCAI/2020	Time+STFT	2.75	3.93	3.44	3.34	-	-
MetricGAN+	Interspeech/2021	STFT	3.15	4.14	3.16	3.64	-	-
CDiffuSE	ICASSP/2022	Time	2.52	3.72	2.91	3.10	5.28	0.91
MetricGAN-OKDv2	ICML/2023	STFT	3.12	4.17	3.13	3.64	-	-
DR-DifuSE	AAAI/2023	STFT	3.09	4.38	3.57	3.76	9.52	0.95
DOSE	NeurIPS/2023	STFT	2.56	3.83	3.27	3.19	-	0.94
S4ND-UNet	Interspeech/2023	STFT	2.99	4.37	3.56	3.70	-	-
VPIDM	TASLP/2024	STFT	3.16	4.23	3.53	3.70	-	-
Dual-S4D	TASLP/2024	STFT	2.55	3.94	3.00	3.23	-	0.93
OFIF-Net	ICASSP/2025	Time	3.06	4.27	3.51	3.67	-	-
GALD-SE	SPL/2025	STFT	3.19	4.23	3.61	3.72	-	0.88
BSDBNet(256)	AAAI/2025	STFT	3.11	4.33	3.58	3.73	-	0.95
MFSE	IJCAI/2025	STFT+STFrFT	3.17	4.46	3.79	3.89	**10.63**	0.95
Ours	-	STFT+STFrFT	**3.27**	**4.56**	**3.84**	**4.02**	10.52	**0.96**

Table 1: Comparing with state-of-the-art models on VoiceBank+DEMAND dataset for PESQ, CSIG, CBAK, COVL, SSNR, STOI metrics, our approach achieves well results. "-" denotes that the result is not provided in the original paper.

Methods	Input	PESQ ↑	CSIG ↑	CBAK ↑	COVL ↑	SSNR ↑	STOI ↑
Noisy	-	1.24	2.75	2.10	2.02	-0.80	0.82
Conv-TasNet	Time	1.83	3.21	2.87	2.57	5.90	0.88
CDiffuSE	Time	1.45	2.97	2.29	2.26	-0.13	0.82
DEMUCS	Time	1.87	3.22	3.02	2.60	4.34	0.89
MetricGAN	STFT	2.05	3.52	2.87	2.84	4.34	0.88
DOSE	STFT	1.64	2.50	2.74	1.87	6.69	0.85
S4ND-UNet	STFT	2.35	4.04	3.17	3.25	6.79	0.91
VPIDM	STFT	2.39	3.51	2.98	2.92	7.21	0.85
GALD-SE	STFT	2.38	3.33	3.00	2.83	7.32	0.82
Ours	STFT+STFrFT	**2.79**	**4.21**	**3.53**	**3.56**	**9.04**	**0.94**

Table 2: The performance of our approach with state-of-the-art open source models on the EARS-WHAM! dataset.

band waveforms and upconverted to passband using coherent mixing. Each signal instance is then corrupted with four types of noise (Gaussian, Poisson, pink, and Rydberg background) under a uniformly sampled SNR level from –20 dB to 0 dB. Both noisy and clean versions are stored as paired arrays, creating a rich training ground for denoising models under complex modulation and low SNR EM conditions.

Together, the **Radio_0dB**, **Radio_-10dB**, and **Modulated_-10dB** datasets form a comprehensive testbed for training and benchmarking models in both unmodulated sensing and modulated EM communication environments.

4.3 Evaluation Metrics

In the speech enhancement task, we adopt standard evaluation metrics following prior works, including: Perceptual Evaluation of Speech Quality (PESQ), prediction of signal distortion (CSIG), background intrusiveness (CBAK), and overall speech quality (COVL), as well as Segmental Signal-to-Noise Ratio (SSNR) and short-time Objective Intelligibility (STOI), to comprehensively assess the effectiveness of our method. We further assess the proposed method on the electromagnetic (EM) signal dataset using two complementary metrics: the Structural Similarity Index Measure (SSIM), which evaluates the perceptual and structural consistency between enhanced and clean signals, and the Signal-to-Noise Ratio (SNR), which reflects the overall improvement in signal energy quality.

4.4 Training Setup

During training, all signal waveforms are segmented into a fixed length sequence of 51,040 samples. If there were insufficient segments, the previous signal content was used to fill in. The time-frequency transform uses an FFT size and window length of 510, with a sample rate of 16 kHz for speech signals, 48 kHz for EM signals, and a hop size of 160. The model was trained for up to 120 epochs using the AdamW optimizer, with a batch size of 4. The initial learning rate was set to 5e-4 and decayed by a factor of 0.5 every 30 hours of training. An early stopping mechanism was adopted to prevent overfitting. All experiments were conducted on a single NVIDIA L40 GPU with 48 GB of memory.

4.5 Ablation Experiments

The ablation study results, as shown in Figure 2, highlight the effectiveness of our proposed FracFusion framework. It outperforms all baseline methods across all evaluation metrics, achieving the highest scores for PESQ, CSIG, CBAK, COVL, SSNR, and STOI. Notably, when maintaining a two

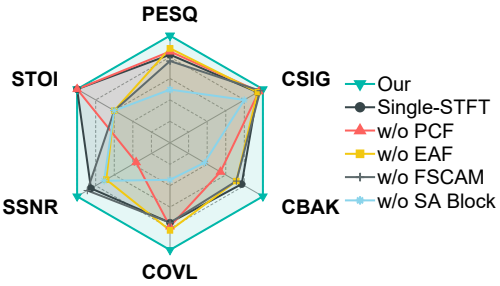


Figure 2: Performance of ablation on the VoiceBank+DEMAND dataset for the main innovations.

branch architecture using only a single STFT view, out of STOI and CSIG, the rest of the metrics slipped badly. The absence of Pearson Channel Fusion leads to a slight degradation in performance, particularly in PESQ and CSIG, indicating the importance of correlation-aware fusion for capturing complementary spectral features. Removing Efficient Align Fusion reduces SSNR, demonstrating the significance of temporal alignment for preserving structural coherence across-views. The absence of the FSCAM and Structure Aware Block results in a modest drop in performance, further emphasizing the need for structure-aware and context-preserving feature extraction in complex signal environments. Overall, the results validate that each component of FracFusion contributes to improving both perceptual quality and signal intelligibility.

4.6 Performance on Signal Datasets

Performance on Speech Datasets: The experimental results on both the VoiceBank+DEMAND and EARS-WHAM! datasets demonstrate that FracFusion significantly outperforms existing state-of-the-art models across all evaluation metrics. FracFusion achieves the highest scores for PESQ, CSIG, CBAK, COVL, SSNR, and STOI, highlighting its superior performance in perceptual quality, background intrusiveness, and signal intelligibility. On the VoiceBank+DEMAND dataset, as shown in Table 1, our method surpasses the best baseline, DR-DifuSE, in PESQ (+0.18), CSIG (+0.18), CBAK (+0.27), and COVL (+0.26), while excelling in SSNR (+1.54), demonstrating its strong denoising capability. On the EARS-WHAM! dataset, as shown in Table 2, FracFusion outperforms the suboptimal model in PESQ (+0.40), CBAK (+0.36), and COVL (+0.31), while also achieving a notable SSNR of 9.04 and STOI of 0.94. These results validate the effectiveness of combining STFT with adaptive STFrFT representations and leveraging advanced multi-view fusion mechanisms for robust speech enhancement under diverse, real world noise conditions.

Performance on EM Datasets: The experimental results on the Radio_0dB, Radio_-10dB, and Modulated_-10dB datasets demonstrate that our method, combining STFT and STFrFT, significantly outperforms existing state-of-the-art models across both SNR (shown in Figure 3) and SSIM (shown in Figure 4) metrics. In the Radio_0dB dataset, our approach achieves a notable SNR of 34.49 and SSIM of 0.99, surpassing the best baseline, VPIDM, in both metrics.

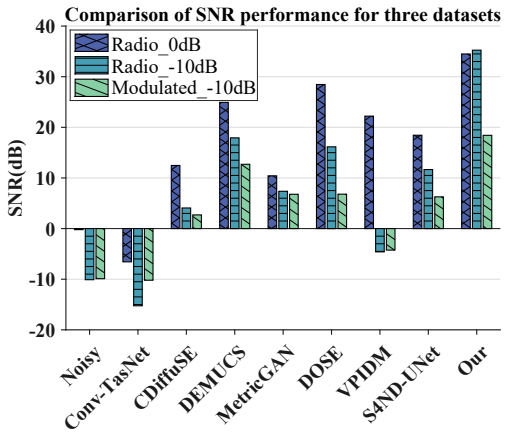


Figure 3: Performance of our method with state-of-the-art open source models for SNR metrics on three EM datasets.

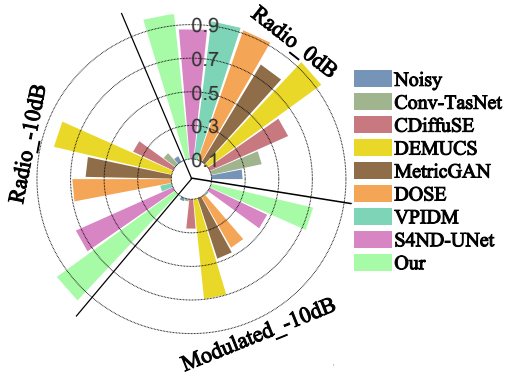


Figure 4: Performance of our method with state-of-the-art open source models for SSIM metrics on three EM datasets.

Similarly, on the Radio_-10dB dataset, we achieve 35.24 in SNR and 0.99 in SSIM, outperforming all compared models, including DEMUCS and S4ND-UNet. On the Modulated_-10dB dataset, we achieve 18.62 in SNR and 0.72 in SSIM, again outperforming multiple baselines, particularly in SNR. These results highlight the effectiveness of our method in handling both unmodulated and modulated EM signals, demonstrating its strong denoising capability across diverse EM environments.

5 Conclusion

We present a unified signal enhancement framework that combines complementary spectral views and structure-aware modeling to address non-stationary, low SNR challenges in speech and EM domains. Our dual stage fusion strategy, comprising PCF for coarse-grained consistency and EAF for fine-grained alignment, is paired with the FSCAM for capturing asymmetric structures. Our method achieves state-of-the-art results on VoiceBank+DEMAND, EARS-WHAM!, and simulated Rydberg datasets, demonstrating strong generalization across domains and enabling robust enhancement in complex environments.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Nos. T2495250, T2495251, T2495253, 62306171, 62472267, 62506217, 62572290), the Science and Technology Major Project of Shanxi (No. 202201020101006), the Fundamental Research Program of Shanxi Province (Nos. 202303021211023, 202303021221075, 202203021222183) and Open Project Foundation of Key Laboratory of Evolutionary Science Intelligence of Shanxi Province.

References

Dang, F.; Hu, Q.; Zhang, P.; and Yan, Y. 2023. Forknet: simultaneous time and time-frequency domain modeling for speech enhancement. *arXiv preprint arXiv:2305.08292*.

Défossez, A.; Synnaeve, G.; and Adi, Y. 2020. Real Time Speech Enhancement in the Waveform Domain. In *Interspeech*, 3291–3295.

Du, M.; Zhong, P.; Cai, X.; and Bi, D. 2022. DNCNet: Deep radar signal denoising and recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 58(4): 3549–3562.

Fan, C.; Liu, E.; Li, A.; Tao, J.; Zhou, J.; Li, J.; Zheng, C.; and Lv, Z. 2025. BSDB-Net: Band-Split Dual-Branch Network with Selective State Spaces Mechanism for Monaural Speech Enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23850–23858.

Fu, S.-W.; Liao, C.-F.; Tsao, Y.; and Lin, S.-D. 2019. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*, 2031–2041. Long Beach, California: PmLR.

Fu, S.-W.; Yu, C.; Hsieh, T.-A.; Plantinga, P.; Ravanelli, M.; Lu, X.; and Tsao, Y. 2021. MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement. In *Interspeech*, 201–205. ISCA.

Fuchs, A.; Rock, J.; Toth, M.; Meissner, P.; and Pernkopf, F. 2021. Complex-valued convolutional neural networks for enhanced radar signal denoising and interference mitigation. In *2021 IEEE Radar Conference (RadarConf21)*, 1–6. IEEE.

Gu, Y.; Zhang, X.; Xue, L.; and Wu, Z. 2024. Multi-scale sub-band constant-q transform discriminator for high-fidelity vocoder. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10616–10620. IEEE.

Guo, Z.; Wang, Q.; Du, J.; Pan, J.; Liu, Q.-F.; and Lee, C.-H. 2024. A Variance-Preserving Interpolation Approach for Diffusion Models With Applications to Single Channel Speech Enhancement and Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 3025–3038.

Huang, G.; Zhang, F.; and Tao, R. 2022. Sliding short-time fractional Fourier transform. *IEEE Signal Processing Letters*, 29: 1823–1827.

Ishwarya, V. S.; and Kothandaraman, M. 2025. A Novel Feature-Fusion-Based Sparse Masked Attention Network for Acoustic Echo Cancellation Using Wavelet and STFT Synergies. *Circuits, Systems, and Signal Processing*, 44(4): 2882–2901.

Jin, Z.; Qian, Y.; Liang, X.; and Geng, H. 2025. A Multi-view Fusion Approach for Enhancing Speech Signals via Short-time Fractional Fourier Transform. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 5508–5516.

Koç, E.; and Koç, A. 2022. Fractional Fourier transform in time series prediction. *IEEE Signal Processing Letters*, 29: 2542–2546.

Kong, Z.; Ping, W.; Dantrey, A.; and Catanzaro, B. 2022. Speech denoising in the waveform domain with self-attention. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7867–7871.

Li, Z.; Gao, Z.; Chen, L.; Gao, J.; and Xu, Z. 2024. The synchrosqueezed method and its theory-analysis-based novel short-time fractional Fourier transform for chirp signals. *Remote Sensing*, 16(7): 1173.

Liang, X.; Lv, L.; Guo, Q.; Jiang, B.; Li, F.; Du, L.; and Chen, L. 2025a. View-Association-Guided Dynamic Multi-View Classification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 5680–5688.

Liang, X.; Wang, S.; Qian, Y.; Guo, Q.; Du, L.; Jiang, B.; Luo, T.; and Li, F. 2025b. Trusted Multi-View Classification with Expert Knowledge Constraints. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, 37409–37426.

Lu, Y.-J.; Wang, Z.-Q.; Watanabe, S.; Richard, A.; Yu, C.; and Tsao, Y. 2022. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7402–7406.

Luo, Y.; and Mesgarani, N. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8): 1256–1266.

Michelsanti, D.; Tan, Z.-H.; Zhang, S.-X.; Xu, Y.; Yu, M.; Yu, D.; and Jensen, J. 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1368–1396.

Nareddula, S. K. R.; Gorthi, S.; and Gorthi, R. K. S. S. 2021. Fusion-Net: Time-Frequency Information Fusion Y-Network for Speech Enhancement. In *Interspeech*, 3360–3364. Brno, Czechia: ISCA.

P-J, K.; Yang, C.; Siniscalchi, S.; et al. 2023. A Multi-dimensional Deep Structured State Space Approach to Speech Enhancement Using Small-footprint Models. In *Interspeech*, 2453–2457. ISCA.

Phan, H.; McLoughlin, I. V.; Pham, L.; Chén, O. Y.; Koch, P.; De Vos, M.; and Mertins, A. 2020. Improving GANs for speech enhancement. *IEEE Signal Processing Letters*, 27: 1700–1704.

Richter, J.; Wu, Y.-C.; Krenn, S.; Welker, S.; Lay, B.; Watanabe, S.; Richard, A.; and Gerkmann, T. 2024. EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation. In *Proc. Interspeech 2024*, 4873–4877.

Shin, W.; Lee, B. H.; Kim, J. S.; Park, H. J.; and Han, S. W. 2023. MetricGAN-OKD: Multi-Metric Optimization of MetricGAN via Online Knowledge Distillation for Speech Enhancement. In *International Conference on Machine Learning*, 31521–31538. PMLR.

Son, H.; Shin, M.-j.; Cho, M.; Kim, J.; Yun, K.-j.; and Kang, S.-J. 2024. CMVDE: Consistent Multi-View Video Depth Estimation via Geometric-Temporal Coupling Approach. *IEEE Transactions on Multimedia*, 26: 9710–9721.

Su, Z.; Teh, K. C.; Xie, Y.; Razul, S. G.; and Kot, A. C. 2023. Signal enhancement aided end-to-end deep learning approach for joint denoising and spectrum sensing. *IEEE Transactions on Vehicular Technology*, 73(3): 4424–4428.

Tian, Y.; Wang, Z.; Sun, J.; and Zhang, L. 2024. Time-Frequency Domain Fusion Enhancement for Audio Super-Resolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2879–2887. Melbourne, Australia: Association for Computing Machinery.

Ventricci, L.; Ribeiro Junior, R. F.; and Gomes, G. F. 2024. Motor fault classification using hybrid short-time Fourier transform and wavelet transform with vibration signal and convolutional neural network. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 46(6): 337.

Wang, L.; Zhang, H.; Qiu, Y.; Jiang, Y.; Dong, H.; and Guo, P. 2024. Improved Speech Separation via Dual-Domain Joint Encoder in Time-Domain Networks. In *2024 International Conference on Electronic Engineering and Information Systems (EEISS)*, 233–239. Changsha, China: IEEE.

Wang, L.; Zheng, W.; Ma, X.; and Lin, S. 2021. Denoising speech based on deep learning and wavelet decomposition. *Scientific Programming*, 2021(1): 8677043.

Wang, M.; Chen, J.; Zhang, X.; Huang, Z.; and Rahardja, S. 2022. Multi-modal speech enhancement with bone-conducted speech in time domain. *Applied Acoustics*, 200: 109058.

Wang, S.; Guo, Q.; Chen, L.; Du, L.; Jin, Z.; Yuan, Z.; and Liang, X. 2025. An Association-based Fusion Method for Speech Enhancement. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 6406–6414.

Xiong, J.; Zhou, Y.; Zhang, P.; Xie, L.; Huang, W.; and Zha, Y. 2022. Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement. *IEEE Transactions on Multimedia*, 25: 5800–5812.

Xue, Y.; Jin, G.; Shen, T.; Tan, L.; Wang, N.; Gao, J.; and Wang, L. 2024. Consistent Representation Mining for Multi-Drone Single Object Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11): 10845–10859.

Yu, C.; Hung, K.-H.; Wang, S.-S.; Tsao, Y.; and Hung, J.-w. 2020. Time-domain multi-modal bone/air conducted speech enhancement. *IEEE Signal Processing Letters*, 27: 1035–1039.

Zeng, X.; Xu, S.; and Wang, M. 2024. A time-frequency fusion model for multi-channel speech enhancement. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1): 47.

Zhou, Z.; Yu, Y.; Song, C.; Liu, Z.; Shi, M.; and Zhang, J. 2023. Multi-view weighted feature fusion with wavelet transform and CNN for enhanced CT image recognition. *Journal of Intelligent & Fuzzy Systems*, 45(6): 12167–12183.