TFD-Net: Towards Intelligent Time-Frequency Mode Decomposition with Practical Applications

Pingping Pan[1], **Yunjian Zhang**[2*], **Jinyi Liu**[2]

[1]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Guandong, China
[2]Shenzhen Kaihong Digital Industry Development Co., Ltd, Shenzhen, Guangdong, China
panpingping@gml.ac.cn, zhangyunjian@kaihong.com, liujinyi@kaihong.com

Abstract

Time-frequency analysis (TFA) and mode decomposition for non-stationary signals are research hotspots in the field of signal processing. Current optimization-based decomposition methods require a good initial IF estimate. However, due to the Heisenberg uncertainty, achieving accurate ridge extraction from the time-frequency representation (TFR) necessitates empirical parameter adjustments. In this paper, we propose the TFD-Net framework, which takes time-series signals as inputs and adaptively conducts TFR construction and mode decomposition. Specifically, the framework integrates a physically interpretable TFA encoder and a query-based mode decomposition decoder. The highlights of this study include exploring the mathematical equivalence between deep convolutional operators and classical TFA methods. This enables the extraction of multi-scale features for TFR construction and mode separation in a data-driven manner, eliminating the need for signal-specific manual tuning and enhancing adaptability. Finally, simulated and real-world experiments demonstrate TFD-Net's superior performance over several state-of-the-art methods in complex signal processing.

Introduction

Non-stationary signals in natural scenarios, such as echolocation signals, power system oscillation signals, and mechanical vibration signals, are characterized by time-varying instantaneous frequencies (IF), multi-mode coupling, and strong noise interference. How to extract time-varying features and separate mixed modes has become a research hotspot in signal processing, and the achievements will benefit applications like target identification and tracking, stability monitoring in power grids, channel equalization and fault diagnosis (Matz, Bolcskei, and Hlawatsch 2013; Silva, Oliveira, and Rocha 2025; Tang, Shen, and Lam 2024; Pan et al. 2025; Feng, Ming, and Fulei 2013; Bai et al. 2023).

Existing approaches suffer from several limitations: time-frequency analysis (TFA) techniques are difficult to handle the trade-off between time and frequency resolutions (Auger et al. 2013); Mode decomposition methods suffer from endpoint effects and sensitivity to noise and sampling (Huang et al. 1998; Pan et al. 2021).

*Corresponding author

Recent advances in TFA for improving the resolution of time-frequency representation (TFR) predominantly employ the synchrosqueezing transform (SST) (Thakur and Wu 2010), which sharpens short-time Fourier transform (STFT) energy distributions via phase-derived IF alignment. Subsequent variants like reassignment-synchrosqueezing (RS), synchroextracting transform (SET), and multisynchrosqueezing (MSST) (Yu, Yu, and Xu 2017; Yu, Wang, and Zhao 2019; He et al. 2020) are proposed for TFR refinement. With the development of deep learning, some TFA networks further improve the energy concentration (Pan et al. 2023; Chen et al. 2024; Dai, Xu, and Zhou 2025; Ding et al. 2025). However, these methods neglect mode decomposition capabilities.

In the aspect of mode decomposition, variational nonlinear chirp mode decomposition (VNCMD) (Chen et al. 2017a) reformulates mode decomposition as a signal demodulation optimization problem. When provided with proper initial IF estimates, the method can progressively refine IF measurements and recover all signal components. Therefore, enhanced initialization strategies (e.g., adaptive chirp mode decomposition (ACMD) (Chen et al. 2019), ridge path regrouping with intrinsic chirp component decomposition (RPRG+ICCD) (Chen et al. 2017b), and improved variational generalized nonlinear mode decomposition (IVGNMD) (Wang, Chen, and Zhai 2025)) are proposed. However, these methods suffer from heavy reliance on empirical parameter tuning, limiting practical applicability.

Therefore, we propose TFD-Net, a deep learning-based time-frequency (TF) mode decomposition network. By treating mode decomposition as a segmentation task on TFRs, TFD-Net is designed with a particular emphasis on a physically interpretable module for multi-scale feature extraction from time-series signals, which simultaneously benefits the TFR generation and mode decomposition. The main contributions of our work are as follows:

- Propose a physics-inspired TFA encoder characterized by the multi-resolution TF transform and multi-scale TF feature extraction. Specifically, STFT-inspired multi-kernel convolutional layers are designed to convert time-series signals into coarse TFRs, which are then refined by channel attention and residual U-Net. Furthermore, multi-scale features from the residual U-Net are fed into

a decoder for mode decomposition.

- Reformulate mode decomposition as an IF trajectory segmentation problem and construct a query-based decoder. Different from the classical Mask2Former (Cheng et al. 2021), our decoder directly takes multi-scale features from the TFA encoder as inputs, and learnable queries are updated to segment distinct TF modes through attention-guided mask prediction.
- Conduct experiments to demonstrate the superiority of the proposed method in obtaining more accurate TFRs and TF modes, compared with several state-of-the-art methods.

Methodology

The TFD-Net framework consists of a physics-inspired TFA encoder and a mode decomposition decoder, as depicted in Fig. 1.

Physics-inspired TFA encoder

The physics-inspired TFA encoder includes the multi-resolution TF transform and multi-scale TF feature extraction.

Multi-resolution TF transform In the multi-resolution TF transform, three complex-valued convolutional layers with different kernel sizes are first constructed to transform time-domain signals into coarse TF-domain representations of different resolutions.

Herein, we explain why the convolutional layers can be used to achieve the TF transform. For the convolution operation between an input signal of length L and a convolutional layer with the kernel size $N_f \times 1 \times L_w$, it will produce a feature map of size $N_f \times 1 \times L$ using the zero-padding technique. Reshaping it into $1 \times N_f \times L$, the feature map behaves like one TFR. And also, L_w and N_f mathematically correspond to the window length parameter and the number of frequency bins in the classical STFT operation.

Mathematically, a bias-free convolutional layer processes input data through the following mechanism:

$$\mathbf{TFR}_m(k) = \sum_{n=0}^{L_w-1} x_m(n)\mathbf{W}(n,k) \tag{1}$$

where $x_m(\cdot)$ denotes the m-th segment of the input signal, obtained by the sliding window of length L_w. $\mathbf{W}$ denotes the filter kernel with size $L_w \times N_f$.

More generally, the segmented signal in (1) can be rewritten as

$$\mathbf{TFR}(m,k) = \sum_{n=0}^{L_w-1} g(n-m)s(n)\mathbf{W}(n,k) \tag{2}$$

where $g(\cdot)$ represents the window function.

Compared with the classical STFT operation:

$$G(t,\omega) = \int_{-\infty}^{+\infty} g(u-t)x(u)e^{-j\omega u}du \tag{3}$$

It is seen that (2) represents a general transform framework, with the STFT serving as its special case. They are totally same if $\mathbf{W}(n,k)$ is updated to be equal to $\exp(-j2\pi k\Delta f n)$.

However, even though neural networks demonstrate theoretical capacity to approximate Fourier-based kernel functions, it remains insufficient to achieve energy-concentrated TFRs. We hypothesize that gradient-based optimization enables the data-driven discovery of task-specific basis functions, overcoming the rigid harmonic assumptions inherent in conventional Fourier decomposition. Therefore, we expand the number of channels in the weight matrix $\mathbf{W}$ from N_f to CN_f, enabling the original signal to be projected into C distinct subspaces, where each subspace is spanned by N_f basis functions. This leads to enhanced TFRs with richer representational diversity. Additionally, since the kernel size L_w is equivalent to the window length that determines the TF resolution, three convolutional layers with different kernel sizes are used in our TF transform block for multi-resolution TFR generation. Given that each convolutional layer generates C TFRs of size $N_f \times L$, the output feature maps of the multi-resolution TF transform block is of size $3C \times N_f \times L$.

For multi-channel feature maps, using the channel attention mechanism for channel enhancement is a commonly used technique, enabling dynamic modeling of the importance of feature channels. Traditional squeeze-and-excitation network (SENet) (Hu et al. 2020) compresses spatial information through global average pooling to generate channel descriptors, followed by two fully connected layers to learn channel-wise importance. Here, we adopt the efficient channel attention (Wang et al. 2019), which replaces fully connected layers with 1-D convolution, effectively reducing the number of parameters while maintaining performance.

Multi-scale TF feature extraction Our goal is to track and separate the IF trajectories of signal components in the TF domain, thus the features that encode refined TFR are required. However, the aforementioned TF transform block merely obtains coarse TFRs. We therefore built a residual U-Net to filter out noise and achieve refined TFRs. The U-Net is chosen for its proven effectiveness in image super-resolution tasks. Driven by spectral fidelity loss, the U-Net can progressively capture multi-scale features of TFRs, which benefits subsequent TF mode decomposition. Furthermore, as the TF transform block serves as a critical foundation for feature extraction and mode decomposition, residual connections are incorporated into the U-Net to facilitate gradient backpropagation to shallow layers.

Mode decomposition decoder

The mode decomposition in the TF domain is equivalent to the image segmentation task, i.e., separating the IF trajectories of different signal components on a given TFR image. Therefore, a query-based decoder is adopted. Two points are highlighted: On one hand, we avoid constructing a new feature extraction backbone based on the TFA module's TFR output. Instead, the multi-scale features from the U-Net are directly fed into the Mask2Former's pixel decoder, because the U-Net has preserved sufficient semantic and structural

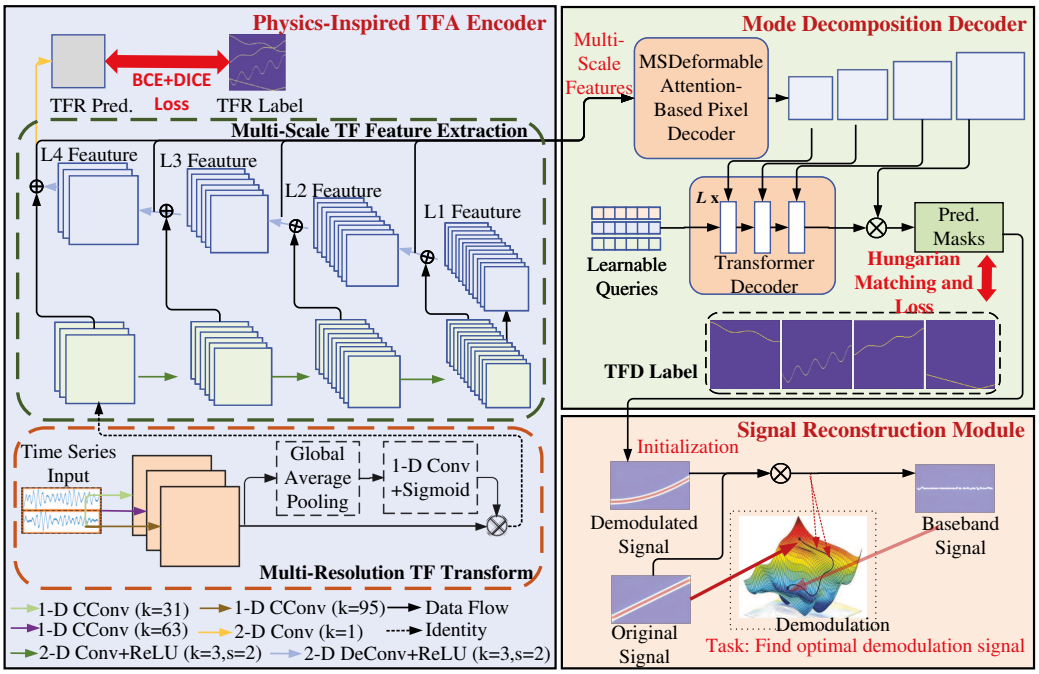

Figure 1: The TFD-Net framework.

features for representing the TFR. On the other hand, existing methods for separating IF trajectories on the TFR image rely on distance clustering or slope metrics of each IF curve, making them threshold-sensitive and ineffective for corner cases. We employ the pixel decoder to map multi-scale features into a high-dimensional space via pixel-wise embedding, where pixels from a same IF trajectory share similar features. Finally, the learnable queries are trained to extract these features for tracking and decomposing IF trajectories, with the predicted masks for TF modes as output.

Signal reconstruction module

To better evaluate the mode decomposition performance, recovering each TF mode into a time-domain waveform is necessary. For the p-th predicted mask for TF mode $\mathbf{TFR}_p(t,f)$, the corresponding IF trajectory can be obtained by locating the peaks along the frequency dimension:

$$\mathrm{IF}_p(t) = \arg\max_f(\mathbf{TFR}_p(t,f)) \tag{4}$$

Thus, the recovered waveform can be achieved based on the signal demodulation principle (Chen et al. 2017a). To be specific, consider a general signal:

$$s(t) = a(t)\exp\left(j2\pi\left(\int_0^t f(v)dv + \phi\right)\right) \tag{5}$$

where $j^2 = -1$. The demodulation operator and modulation can be respectively expressed as

$$\Phi^-(t) = \exp\left(-j2\pi\left(\int_0^t f_d(v)dv - f_c t\right)\right) \tag{6}$$

$$\Phi^+(t) = \exp\left(j2\pi\left(\int_0^t f_d(v)dv - f_c t\right)\right) \tag{7}$$

where $f_d(v)$ denotes a frequency function of operators and $f_c > 0$ denotes a constant frequency. By multiplying $s(t)$ with the demodulation operator, the demodulated signal is obtained:

$$s_d(t) = a(t)\exp\left(j2\pi\left(\int_0^t f(v) - f_d(v)dv - f_c t + \phi\right)\right) \tag{8}$$

It is seen that $s_d(t)$ becomes a single-frequency signal with frequency f_c, when $f(v) - f_d(v) = 0$ is satisfied. That is, if the frequency function of the demodulation operator is updated to approximate the IF function of the original signal, the demodulated signal would have the narrowest bandwidth, and meanwhile, the included signal components would be restored by multiplying the demodulated signal with the modulation operator in (7). Therefore, the optimization problem is formulated as minimizing the bandwidth of demodulated signals, and the constraint condition is that the sum of all the recovered signal components approximates the original signal. For accelerating the algorithm convergence and improving the recovery accuracy, the estimated IFs in (4) from separated TF modes serve as good enough initial frequency functions of the demodulation operator for optimization.

Dataset and Loss function

Modulated signals are generated as the dataset. In order to make our model applicable beyond local parameter ranges, the signal parameter settings should ensure that the signal's

IFs cover the entire unblurred spectrum. The expressions of signals and the corresponding IFs are respectively written as

$$s(n) = \sum_{p=0}^{P-1} A_p \, e^{j2\pi\left(B_p \cos(b_p n+\theta_p)+d_p n+a_p n^2\right)} \quad (9)$$

$$\mathrm{IF}_p(n) = d_p + 2a_p n - B_p b_p \sin\left(b_p n + \theta_p\right) \quad (10)$$

where the component number P follows the uniform distribution $\mathcal{U}(1,6)$. $A_p \in (0.1, 15)$ denotes the intensity of the p-th component. B_p denotes the vibration amplitude of the p-th component, and follows $\mathcal{U}(0.01, 1)$. a_p and b_p are respectively sampled in the range of $[-30, 30)$ and $[-25, 25)$. θ_p follows $\mathcal{U}(0, 2\pi)$. Doppler d_p follows $\mathcal{U}(-128, 128)$. The signal length L and the frequencies of interest N_f are both set to 256.

Noted that due to the exponential signal's multivalued nature, the dataset's fixed sampling rate does not impair generalization, as confirmed by subsequent real-world experiments.

Besides, the dataset consists of a training set of 30,000 signals, a validation set of 1,000 signals and a test set of 200 signals. Meanwhile, all these signals are mixed with the additive white Gaussian noise (AWGN) with SNRs of $[-5, 40]$ dB. The training parameters include the initial learning rate of 0.0003 and batch size of 8. The Adam algorithm (Kingma and Ba 2015) is adopted to optimize the model. Besides, a machine with one NVIDIA RTX 6000 GPU with memory of 48 GB and the 14-core Xeon(R) Gold 6330 with memory of 62 GB is used for the model training.

The loss function consists of a TFR loss $Loss_{\mathrm{tfr}}$ and a mode decomposition loss $Loss_{\mathrm{tfd}}$. Specifically, $Loss_{\mathrm{tfr}}$ includes the DICE loss and BCE loss between the predicted and ground-truth TFRs in the TFA module. Similarly, $Loss_{\mathrm{tfd}}$ includes the DICE loss and BCE loss between the matched and ground-truth TF modes in the TF mode decomposition module. It is noted that the matched TF modes in $Loss_{\mathrm{tfd}}$ are chosen by the Hungarian matching algorithm (Mills-Tettey, Stentz, and Dias 2007). Besides, since the IF trajectories (positive samples) are far less than the background (negative samples) in the TFR, the positive sample pixels are weighted. Mathematically, the loss function is expressed as

$$Loss = Loss_{\mathrm{tfr}} + Loss_{\mathrm{tfd}} \quad (11)$$

$$Loss_{\mathrm{tfr}} = \alpha \mathrm{BCE}(P_{\mathrm{tfr}}, G_{\mathrm{tfr}}) \cdot \omega + (1-\alpha)\mathrm{DICE}(P_{\mathrm{tfr}}, G_{\mathrm{tfr}}) \quad (12)$$

$$Loss_{\mathrm{tfd}} = \alpha \mathrm{BCE}(P_{\mathrm{tfd}}, G_{\mathrm{tfd}}) + (1-\alpha)\mathrm{DICE}(P_{\mathrm{tfd}}, G_{\mathrm{tfd}}) \quad (13)$$

where $P_{\mathrm{tfr/tfd}}$ denotes the predicted results by the TFA encoder/mode decomposition decoder, and $G_{\mathrm{tfr/tfd}}$ denotes the ground-truth labels. ω denotes the weights for positive pixels, and α is a parameter to balance the DICE and BCE loss. It is noted that since a masked Transformer is used in mode decomposition modules, no weights are given to positive-sample pixels in (13). In our model, α and ω are set to 0.7 and 5, respectively. Besides, since the mode decomposition focuses on segmenting IF trajectories with distinct boundaries, the weight α for the BCE term is set to a higher value of 0.7, while the DICE term is assigned a weight of 0.3 for both $Loss_{\mathrm{tfr}}$ and $Loss_{\mathrm{tfd}}$.

Simulated Experiments

In this section, we evaluate the performance of the proposed TFD-Net framework and compare it with state-of-the-art methods. Then, we conduct ablation studies to investigate important components of our model and clarify the criterion for parameter selection.

Performance comparison

To evaluate the performance of our proposed model, experiments are conducted and the comparison methods include VNCMD (Chen et al. 2017a), RPRP+ICCD (Chen et al. 2017b), SET (Yu, Yu, and Xu 2017), ACMD (Chen et al. 2019), MSST (Yu, Wang, and Zhao 2019), TFA-Net (Pan et al. 2023) and IVGNMD (Wang, Chen, and Zhai 2025). Among them, SET, MSST and TFA-Net are TFA methods, while the others are mode decomposition methods.

First, a test signal with four components is processed to show the intuitive performance comparison. The ground-truth IF trajectories and the corresponding STFT results with different window lengths are respectively shown in Fig. 2. It is found that different window lengths lead to the TFRs of different resolutions. Specifically, a window function of small length leads to poor frequency-domain resolution, while a long time-domain window will degrade the time-domain resolution.

Since the signal components are closely located, it is seen that the STFT results fail to clearly distinguish them. TFRs produced by the proposed TFD-Net and the mentioned comparison methods are given in Fig. 3. It is found that for the energy-reassignment methods, i.e., SST and SET in Figs. 3 (e)–(f), they can well capture the independent and non-crossing IF trajectory in the lower part of the TFR images, and the energy concentration is improved. However, because they reassign TF energy based on the STFT result, a phase-disrupted STFT result caused by the mutual interference of multiple components inevitably leads to the failure of energy reassignment. For the mode decomposition methods in Figs. 3 (a)–(d), they are also post-processing methods based on the STFT, and pixel-by-pixel track each IF curve in the STFT result by using morphological filtering or logical judgment. Therefore, the tracking results are not satisfactory due to the low-resolution STFT results. Moreover, it is assumed that the IF trajectory changes slowly, resulting in errors in the tracking results. For example, due to the more similar IF curve slopes for the yellow and orange trajectories around the TF coordinate (0.45 s, 82 Hz) as shown in Fig. 2 (a), the first IF curve (the yellow one) is tracked and connected to the second IF curve (the orange one) for the IVGNMD method, as shown in Fig. 3 (a). The TFA-Net produces sharper and more accurate TFR result, confirming the superiority of automatic feature extraction by deep learning. However, the convolution kernel of a single scale is used for the TF transform, which indicates that it can only obtain the single-resolution TFR, leading to limited TF feature extraction. Meanwhile, the TFA-Net emphasizes energy concentration and neglects mode decomposition. Compared with TFA-Net, the proposed TFD-Net effectively generates and integrates multi-resolution TFRs, further improving the

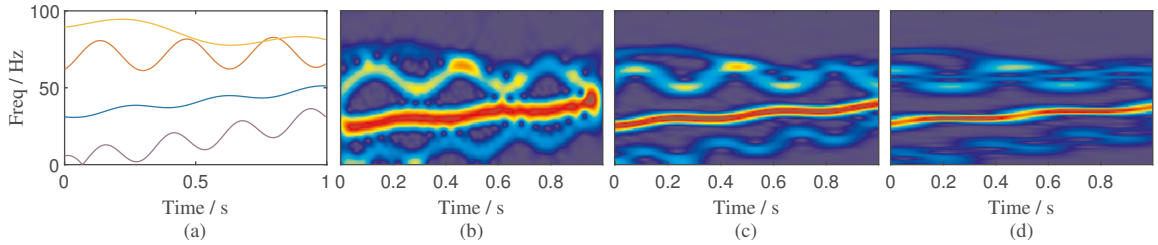

Figure 2: IF trajectories and STFT results of a test signal. (a) True IF trajectories, (b) STFT with window length 32, (c) STFT with window length 64, (d) STFT with window length 96.

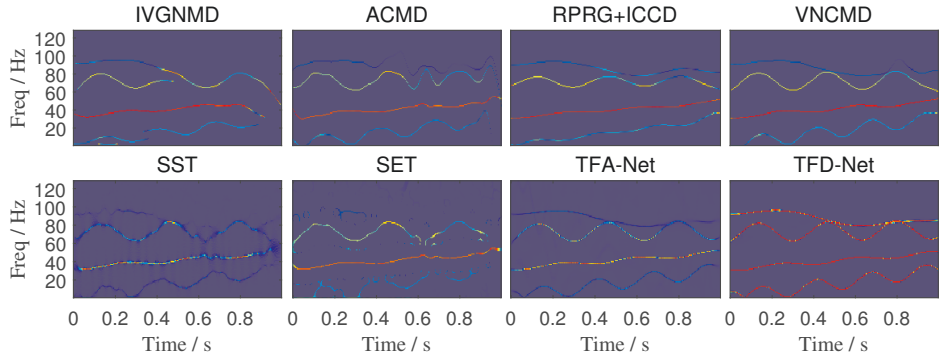

Figure 3: Comparisons of TFRs produced by different methods

location accuracy of the closely placed IFs. For example, for the intersection area of the upper two IF trajectories, the TFD-Net exhibits better tracking performance, as shown in Fig. 3 (h).

Furthermore, the separated TF modes and recovered signal components by the proposed TFD-Net are respectively shown in the first and second row of Fig. 4. It is found that the separated TF modes are highly in accord with the true IF trajectories shown in Fig. 2 (a). Also, with the accurate IF initialization, the reconstruction error is greatly reduced.

Additionally, the quantitative performance of the test set is statistically analyzed. Specifically, the DICE losses for predicted TFR and separated TF modes, and the MSE metric for the reconstructed components concerning SNRs are calculated in Table 1.

As the SST, SET and TFA-Net cannot realize the mode decomposition and signal recovery, they are not involved in the comparison here. It is seen that the proposed TFD-Net outperforms the other mode decomposition methods under various SNRs.

Ablation study

To validate the effectiveness of our physics-inspired TFA module, we design three ablation experiments as follows:

- TFD-Net-L: This experiment removes the constraint of TFR loss in the TFA module, with all other modules being the same as in TFD-Net.
- TFD-Net-E: This experiment removes the enhanced channel attention in the TF transform block of the TFA module, with all other modules being the same as in TFD-Net.
- TFD-Net-R: This experiment removes the residual connection in the U-Net block of the TFA module, with all other modules being the same as in TFD-Net.

The main comparative results are presented in Fig. 5. The results from the ablation experiments demonstrate that each component positively influences the overall performance. For the TFD-Net-R and TFD-Net-L models, by removing the TFR loss or the residual connection in the TFA module, the models hardly work. These results confirm that the multi-resolution TF transform block serves as a critical foundation for downstream multi-scale feature extraction and mode decomposition. By imposing a TFR loss and back-propagating gradients through residual connections, the convolutional kernels within the TF transform block are effectively constrained and optimized to approximate physically meaningful short-time filters. Besides, the performance of the TFD-Net-E model also degrades compared to the complete TFD-Net due to the absence of enhanced channel attention, verifying that the channel attention mechanism plays a certain role in highlighting more important features among multi-resolution TFRs. In summary, these ablation experiments not only validate the effectiveness of the proposed TFA module but also showcase its significant role in improving the mode decomposition and signal reconstruction.

Parameter selection

In this section, we investigate the impact of parameter selection on the model performance. One of the important parameters is the number of convolutional filters used for the multi-resolution TF transform. As mentioned above, convolution operations with various kernel sizes approximate the

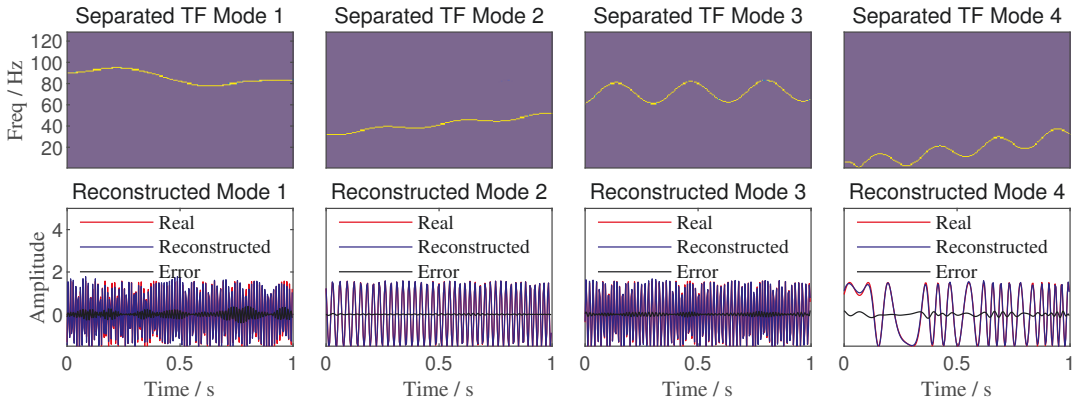

Figure 4: Results of mode decomposition and signal reconstruction by the proposed TFD-Net

Methods	SNR = 20dB			SNR = 15dB			SNR = 10dB			SNR = 5dB		
	TFR DICE	TF modes DICE	Recon. MSE	TFR DICE	TF modes DICE	Recon. MSE	TFR DICE	TF modes DICE	Recon. MSE	TFR DICE	TF modes DICE	Recon. MSE
IVGNMD	0.753	0.769	0.295	0.758	0.769	0.304	0.786	0.781	0.324	0.887	0.864	0.416
ACMD	0.775	0.765	0.271	0.812	0.793	0.289	0.862	0.838	0.318	0.908	0.885	0.368
RPRG+ICCD	0.781	0.770	0.319	0.785	0.773	0.325	0.803	0.787	0.347	0.839	0.819	0.383
VNCMD	0.761	0.736	0.342	0.778	0.751	0.336	0.812	0.783	0.342	0.859	0.834	0.369
TFD-Net	**0.276**	**0.458**	**0.079**	**0.329**	**0.488**	**0.095**	**0.404**	**0.530**	**0.121**	**0.510**	**0.587**	**0.173**

Table 1: DICE (↓) and MSE (↓) metrics for TFR and separated/reconstructed modes under different SNRs.

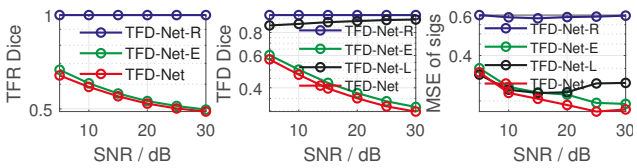

Figure 5: Ablation experiment results. Left: TFR DICE loss; Middle: Mean DICE loss for separated TF modes; Right: MSE for reconstructed signals.

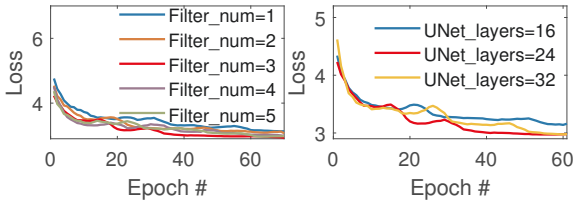

Figure 6: The influence of hyperparameters. Left: The number of filters for TF transform; Right: The number of convolutional layers in the U-Net block.

STFT operations with different window lengths, which can provide different combinations of TF resolutions. However, more convolutional filters might lead to a larger model that could cause over-fitting, high computational complexity and other problems. Therefore, we conduct experiments to determine the optimal parameter selection. The validation losses for the cases including one filter with size 31, two filters with size 31 and 63, three filters with sizes 31, 63 and 95, four filters with sizes 31, 63, 95 and 127, and five filters with sizes 31, 63, 95, 127 and 159 are displayed in Fig. 6 (a). It can be seen that as the number of filters increases from one to three, the performance of the model gradually improves. Then the number of filters is increased, and the performance of the method reaches saturation and no longer shows a significant improvement. Therefore, in our model, three convolutional filters with sizes 31, 63 and 95 are used for multi-resolution TF transform.

Another important module is the residual U-Net-shaped

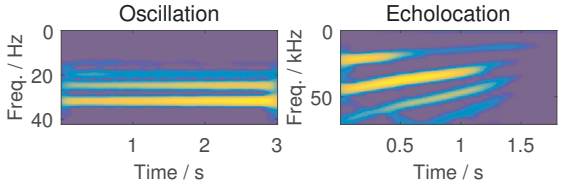

Figure 7: STFT results of two real-world data.

TF feature extraction block. We conduct experiments to determine the number of convolutional layers for obtaining each scale of features. The cases (two convolutional layers for each scale and 16 layers in total, three convolutional layers for each scale and 24 in total, four convolutional layers for each scale and 32 in total) are considered in the experiment, and the validation losses are exhibited in Fig. 6 (b). It is seen that when the number of convolutional layers is

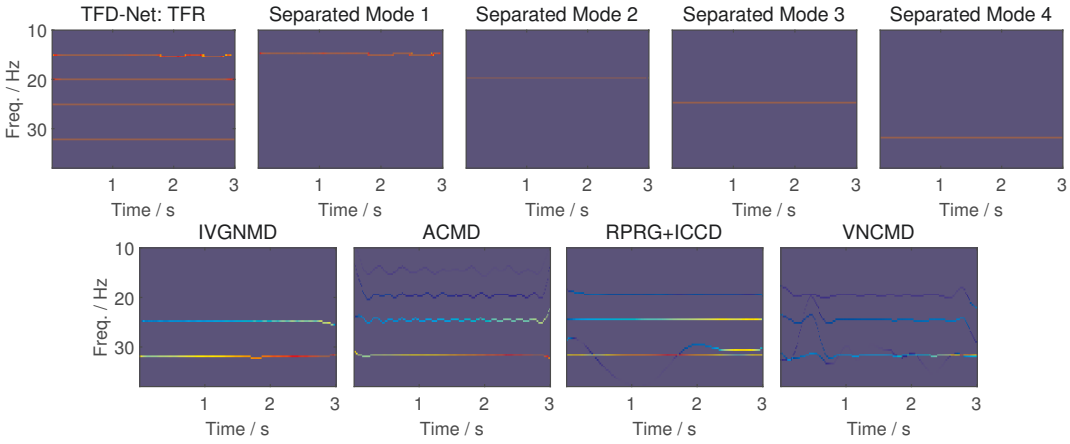

Figure 8: Comparison of mode decomposition for oscillation data. Upper: proposed method. Lower: comparison group.

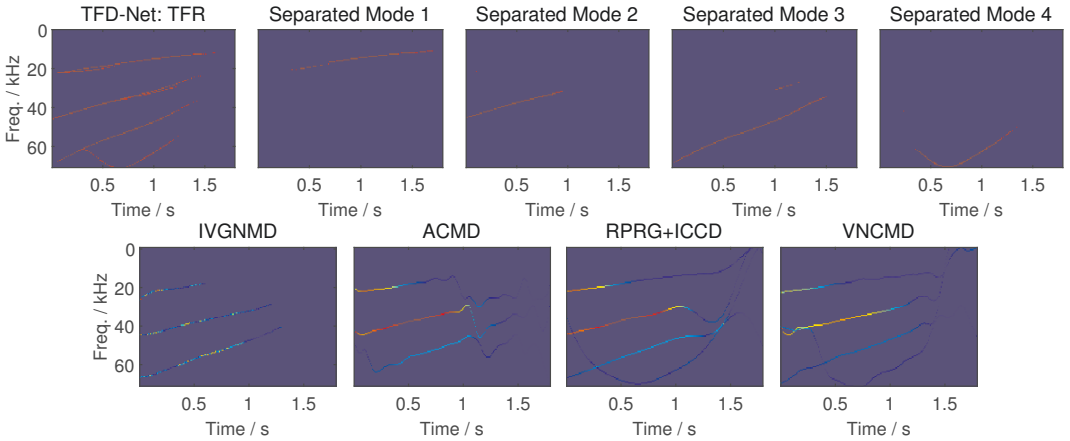

Figure 9: Comparison of mode decomposition for echolocation signal. Upper: proposed method. Lower: comparison group.

increased to 24, the model performance has not improved significantly. Therefore, the number of convolutional layers in the U-Net block is set to 24 in our model.

Applications

In this section, real-world experiments are conducted to verify the effectiveness of our proposed method. Two real data are processed without model retraining, and they can be both directly loaded from the Matlab time-frequency gallery. One is the subsynchronous oscillation data of a power system, and the other is the echolocation signal collected by Rice University. The sampling rates of these two signals are respectively 1 kHz and 142.86 kHz, and their STFT results are shown in Fig. 7.

Fig. 8 and Fig. 9 show the mode decomposition results from different methods, respectively. For the proposed method, the TFR result from the TFA encoder and the separated modes from the mode decomposition decoder are all provided in the upper row of these two figures. It is seen that the proposed method better achieves TF mode decomposition and greatly improves the energy concentration of the weak-intensity components, e.g., the first component (15 Hz) from the oscillation data and the fourth component from the echolocation signal.

Conclusion

This paper proposes TFD-Net for TFA and mode decomposition, with its core contribution being the use of physics-inspired convolutional layers with channel attention to generate multi-resolution TFRs. These layers are mathematically equivalent to STFTs, enabling the network to learn physically interpretable TFRs. Based on this foundation, a residual U-Net extracts multi-scale TF features for TFR refinement and mode decomposition. A query-based Transformer decoder then projects these features into high-dimensional embeddings, where learnable queries are iteratively updated to capture pixel-level similarities along IF trajectories, enabling accurate IF tracking. Finally, the separated modes yield initial IF estimates, which are iteratively refined to reconstruct individual signal components. Numerical and real-world experiments validate the effectiveness, robustness, and generalization of TFD-Net. Future work may explore integrating an unfolded reconstruction algorithm to achieve a fully neural network-based pipeline.

Acknowledgments

This work was supported by the Natural Science Foundation of Guangdong Province, China.

References

Auger, F.; Flandrin, P.; Lin, Y.-T.; McLaughlin, S.; Meignen, S.; Oberlin, T.; and Wu, H.-T. 2013. Time-Frequency Reassignment and Synchrosqueezing: An Overview. *IEEE Signal Processing Magazine*, 30(6): 32–41.

Bai, Y.; Cheng, W.; Wen, W.; and Liu, Y. 2023. Application of Time-Frequency Analysis in Rotating Machinery Fault Diagnosis. *Shock and Vibration*, 2023(1): 9878228.

Chen, Q.; Dong, X.; Tu, G.; Wang, D.; Cheng, C.; Zhao, B.; and Peng, Z. 2024. TFN: An Interpretable Neural Network with Time-Frequency Transform Embedded for Intelligent Fault Diagnosis. *Mechanical Systems and Signal Processing*, 207: 110952.

Chen, S.; Dong, X.; Peng, Z.; Zhang, W.; and Meng, G. 2017a. Nonlinear Chirp Mode Decomposition: A Variational Method. *IEEE Transactions on Signal Processing*, 65(22): 6024–6037.

Chen, S.; Dong, X.; Xing, G.; Peng, Z.; Zhang, W.; and Meng, G. 2017b. Separation of Overlapped Non-Stationary Signals by Ridge Path Regrouping and Intrinsic Chirp Component Decomposition. *IEEE Sensors Journal*, 1–1.

Chen, S.; Yang, Y.; Peng, Z.; Dong, X.; Zhang, W.; and Meng, G. 2019. Adaptive Chirp Mode Pursuit: Algorithm and Applications. *Mechanical Systems and Signal Processing*, 116: 566–584.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2021. Masked-Attention Mask Transformer for Universal Image Segmentation. *arXiv*.

Dai, S.; Xu, J.; and Zhou, H. 2025. Radio Signal Modulation Pattern Recognition Based on Time-Frequency Adaptive Decomposition and Hybrid Neural Network. *IEEE Internet of Things Journal*.

Ding, M.; Ding, Y.; Dongye, G.; and Lv, P. 2025. Enhanced End-to-End and Consistent Time-Frequency Analysis for Tracking. *IEEE Internet of Things Journal*, 1–1.

Feng, Z.; Ming, L.; and Fulei, C. 2013. Recent Advances in Time-Frequency Analysis Methods for Machinery Fault Diagnosis: A Review with Application Examples. *Mechanical Systems and Signal Processing*, 38(1): 165–205. Condition monitoring of machines in non-stationary operations.

He, Z.; Tu, X.; Bao, W.; Hu, Y.; and Li, F. 2020. Gaussian-Modulated Linear Group Delay Model: Application to Second-Order Time-Reassigned Synchrosqueezing Transform. *Signal Processing*, 167: 107275.

Hu, J.; Shen, L.; Albanie, S.; Sun, G.; and Wu, E. 2020. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8): 2011–2023.

Huang, N. E.; Shen, Z.; Long, S. R.; Wu, M. C.; Shih, H. H.; Zheng, Q.; Yen, N.-C.; Tung, C. C.; and Liu, H. H. 1998. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, 454(1971): 903–995.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *International Conference for Learning Representations*, 1–15.

Matz, G.; Bolcskei, H.; and Hlawatsch, F. 2013. Time-Frequency Foundations of Communications: Concepts and Tools. *IEEE Signal Processing Magazine*, 30(6): 87–96.

Mills-Tettey, G. A.; Stentz, A.; and Dias, M. B. 2007. The Dynamic Hungarian Algorithm for the Assignment Problem with Changing Costs. *Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27*.

Pan, H.; Zheng, J.; Yang, Y.; and Cheng, J. 2021. Nonlinear Sparse Mode Decomposition and Its Application in Planetary Gearbox Fault Diagnosis. *Mechanism and Machine Theory*, 155: 104082.

Pan, P.; Zhang, Y.; Deng, Z.; Fan, S.; and Huang, X. 2023. TFA-Net: A Deep Learning-Based Time-Frequency Analysis Tool. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11): 9274–9286.

Pan, P.; Zhang, Y.; Li, Y.; Ye, Y.; He, W.; Zhu, Y.; and Guo, R. 2025. Interpretable Optimization-Inspired Deep Network for Off-Grid Frequency Estimation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8): 14816–14828.

Silva, M.; Oliveira, R.; and Rocha, F. 2025. Enhancing Time-Frequency Signal Analysis: Integrating Windowing with the Fractional Fourier Transform for Modern Applications. *IEEE Transactions on Aerospace and Electronic Systems*, 1–22.

Tang, Z.; Shen, H.; and Lam, C.-T. 2024. Multi-scale TFT-Net Time-Frequency Representation for Multi-component Radar Signal Recognition. In *International Conference on Parallel and Distributed Computing: Applications and Technologies*, 292–303.

Thakur, G.; and Wu, H. T. 2010. Synchrosqueezing-Based Recovery of Instantaneous Frequency from Nonuniform Samples. *Siam Journal on Mathematical Analysis*, 43(5): 2078–2095.

Wang, H.; Chen, S.; and Zhai, W. 2025. Improved Variational Generalized Nonlinear Mode Decomposition for Separating Crossed Chirp Modes and Dispersive Modes of Non-Stationary Signals in Mechanical Systems. *Mechanical Systems and Signal Processing*, 227: 112407.

Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; and Hu, Q. 2019. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks.

Yu, G.; Wang, Z.; and Zhao, P. 2019. Multisynchrosqueezing Transform. *IEEE Transactions on Industrial Electronics*, 66(7): 5441–5455.

Yu, G.; Yu, M.; and Xu, C. 2017. Synchroextracting Transform. *IEEE Transactions on Industrial Electronics*, 64(10): 8042–8054.